Spatial methods for improved estimates of abundance indices from preferentially or systematically sampled data

Craig Allan Marsh

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in Statistics,
The University of Auckland, 2022.

Abstract

Fisheries managers often rely on quantitative stock assessments when managing fisheries and setting catch limits. An important input is an index of abundance that is estimated from raw data and tracks abundance over time. This thesis examined the properties of two geostatistical models for estimating indices of abundance and their associated variance, each applied to a different source of data. The first geostatistical model was developed to account for preferential sampling in fishery-dependent catch and effort data, and the second was developed to estimate abundance and its associated variance from two-dimensional systematic surveys.

Fishery-dependent catch and effort data typically have spatial sampling bias due to fishers choosing to optimise catch composition and catch rates, known as preferential sampling. This dependence between sample locations and catch rates is often ignored in conventional geostatistical models (conventional models) which can lead to biased inference of spatial abundance. A geostatistical model that accounts for preferentially sampled data (PS model) was developed and its properties explored using simulations and data from a major New Zealand trawl fishery. An innovative agent-based model (ABM) was developed and applied to simulate fish stock dynamics and complex fishing interactions.

Simulations based on the ABM showed the PS model had similar or better predictive power compared to conventional models by accounting for spatial bias. However, when the PS model was applied to the New Zealand trawl fishery data set, trends in estimated abundance indices were not significantly different between the PS model and conventional models.

Two-dimensional systematic surveys are a common survey design applied in forestry, ecology, and fisheries. However, the most widely used variant consists of just a single primary sampling unit and consequently, there is no applicable design-based variance estimator for the estimated population total. An R package was developed which applied a geostatistical model-based estimator for the population total and its associated variance and employed simulations to compare it with existing methods, which include semiparametric and approximated design-based variance estimators. Results showed that while no estimator achieved the desired confidence interval coverage across all simulations, the geostatistical model-based and semiparametric estimators had better nominal coverage of confidence intervals compared to approximated design-based estimators. Although the geostatistical model-based variance

estimator did not completely resolve the variance estimation issue, we recommend its use for its ability to produce spatial distribution maps and moreover it can include covariates and be applied to irregular spatial domains.

Acknowledgements

I would like to thank my supervisor Professor Russell Millar for his help and advice during this PhD. I appreciate all of your input and the freedom you allowed me to develop my own research topics and approach to addressing them.

I also wish to thank various people for their contribution to this project; Alistair Dunn and Dr Ian Doonan, for their valuable technical support on this project. Helen Smyth and Gary Ferguson for their editing help, and late-night feedback. Additionally, this endeavour would not have been possible without the generous support from the National Institute of Atmosphere and Water (NIWA), who part financed my research. I would also like to thank Fisheries New Zealand for supplying the hoki catch and effort data set.

Lastly, I would be remiss in not mentioning my friends and family who have supported me on this journey. Words cannot express my gratitude to Isabella Rose for her invaluable support and patience throughout this ordeal!

Contents

Chapter 1

Introduction

Fishing is a significant economic, environmental and cultural activity, that provides protein for billions of people globally (FAO 2020). Fisheries management aims to maintain fish populations at sustainable levels while accounting for other environmental and socio-economic factors (Hilborn & Walters 1992). Fisheries management for a species typically occurs at the stock level, where a stock is defined as a management unit of self-reproducing fish (MacLean & Evans 1981, Begg & Waldman 1999). Management of fish stocks relies on statistical models and analysis, with quantitative stock assessments being the main framework (Haddon 2010).

Quantitative stock assessments collate all available information and data to construct statistical models for inference on stock abundance and other important management metrics (Gabriel & Mace 1999). A key input for most quantitative stock assessments are estimated indices that track changes in stock abundance over time (Hilborn & Walters 1992, Haddon 2010). Indices of stock abundance can be either applied directly to inform management such as with harvest control rules (Little et al. 2011) or used as an input to more complex integrated models (Fournier & Archibald 1982, Bull et al. 2012, Methot Jr & Wetzel 2013). The latter is preferred as it allows for additional data such as age-composition and catch to inform estimates of stock dynamics and fishing pressure (Maunder & Punt 2013). In both applications, abundance indices provide direct information on key stock assessment questions: What is the current stock abundance relative to historical levels? Is stock abundance trending up or down? (Francis 2011, 2017).

This thesis examined the properties of two geostatistical models for estimating these indices of abundance and their associated variance, each applied to a different

source of data. The first geostatistical model was developed to account for preferential sampling bias in fishery-dependent catch and effort data (Thorson et al. 2020, Ducharme-Barth et al. 2022) and the second was developed to estimate the variance for estimated total abundance from two-dimensional systematic surveys (Millar & Olsen 1995, Fewster 2011). Both data sources have known challenging properties that are commonly ignored.

Estimators employed for making inference on populations from sampled data include model-based, design-based and model-assisted. Geostatistical models are a model-based estimator that describe spatial and temporal variability in the variable of interest. Model-based estimators treat the sampled population as a realization from a stochastic system, which is expressed via statistical models (Bartolucci & Montanari 2006, Brus 2021). In contrast, design-based estimators make no assumptions on the data generation process and instead use sampling probabilities to estimate population totals and associated variance (Lohr 2009). Hybrid estimators or model-assisted estimators utilize both of the two previous concepts. Design-based methods are often advocated over model-based approaches due to the lack of assumptions required (Wolter 2007, Sterba 2009). However, recent applications of geostatistical model-based estimators to spatial survey data (Shelton et al. 2014, Thorson et al. 2015) have illustrated advantages over design-based estimators when the distribution of individuals is highly variable through space. This thesis explored geostatistical model-based estimators because of the ability to describe complex spatial patterns exhibited by populations coupled with the fact that these estimators can be applied to sampled data with unknown sampling probabilities (Johnson et al. 2010).

Recent advancements in spatial Gaussian Field (GF) methods (Lindgren et al. 2011, Shelton et al. 2014) and statistical software for efficient inference of complex hierarchical models (Kristensen et al. 2016, Krainski et al. 2018) provide another reason for exploring geostatistical models. These innovations have removed many barriers for practically applying geostatistical models because there is now efficient and easy to use software available. This is reflected in a significant uptake of these models in the fisheries literature (Thorson et al. 2015, Grüss & Thorson 2019, Grüss et al. 2019, Pennino et al. 2019, Xu et al. 2019, Maunder et al. 2020, Thorson et al. 2020). Geostatistical models, also known as spatio-temporal models (Thorson et al. 2015, Cressie 2015, Thorson et al. 2020) require spatially referenced observations

to capture spatial structures in the data. The key premise regarding geostatistical models applied in this thesis relates to Tobler's law of geography; "everything is related to everything else, but near things are more related than distant things" Tobler (1970).

Catch and effort data reported by fishers, termed fishery-dependent catch and effort data, is required by most jurisdictions around the world. This makes it commonly accessible, and in some instances, the only available data set for deriving indices of abundance. Alternative data sources include scientific surveys (Stevens et al. 2021) and tag-recapture experiments (Goethel et al. 2011, Tenningen et al. 2011, Bravington et al. 2016). These alternative data sources benefit from robust survey designs and availability of design-based estimators (Sterba 2009, Smith 1990). However, they are often very costly and so are generally limited to high value fisheries. The opportunistic nature of fishery-dependent data means it lacks survey design and sample randomisation, thus limiting the statistical methods available for inference (Johnson et al. 2010).

Indices of abundance derived from fishery-dependent catch and effort data are based on catch rates or Catch Per Unit Effort (CPUE) over time. Abundance indices derived using CPUE assume catch is proportional to fishing effort and stock abundance at small spatial scales. For most fisheries, a unit of fishing effort will yield different catches based on fishing factors such as vessel size, skipper experience, location and time. The main objective of models applied to fishery-dependent catch and effort data is to remove the impacts of these factors on estimated abundance indices, termed CPUE standardisation (Maunder & Punt 2004). The resulting standardised CPUE is then assumed to be proportional to stock abundance.

There have been many recorded pitfalls in analysing fishery-dependent catch and effort data when estimating abundance indices (Gulland 1974, Hilborn 1985, Harley et al. 2001*a*, Maunder et al. 2006). This thesis focuses on the issue of spatial sampling biases due to fishers choosing to optimise catch composition and catch rates, known as preferential sampling (PS) (Diggle et al. 2010, Pennino et al. 2019, Ducharme-Barth et al. 2022). Specifically this thesis focuses on fisheries whose catch composition is predominately made up of a single stock, termed target fisheries rather than by-catch, or mixed fisheries whose catch is a mixture of stocks (Poos et al. 2010). Target fisheries are of interest because it is assumed fishing decisions such as location and gear configurations are made to optimise catch rates for the species of interest,

and thus more likely to satisfy assumptions made by preferential sampling models. Conventional models used in CPUE standardisation generally treat the location of fishing events as ancillary and ignore this sampling bias. Studies have shown that this can result in biased parameter estimates and spatial predictions (Diggle et al. 2010, Dinsdale 2018, Pennino et al. 2019) and is an often commented concern in studies that apply geostatistical models to fishery-dependent catch and effort data (Williams et al. 2018, Xu et al. 2019, Maunder et al. 2020, Thorson et al. 2020).

Geostatistical models that explicitly account for preferentially sampled data, hereinafter simply referred to as PS models, are a commonly applied model in other fields, such as mineral exploration, air quality and soil sciences (Diggle et al. 2010, Pati et al. 2011, Diggle et al. 2013). The application of PS models to fishery-dependent catch and effort data has been limited. We believe this is due to the large sample sizes typical of catch and effort data, PS model application relies on recently developed and complex software (Kristensen et al. 2016, Krainski et al. 2018) and previous applications encountering model instability issues (Conn et al. 2017). Previous studies have explored the effect preferential sampled data has on the performance of conventional geostatistical models to catch and effort data (Grüss & Thorson 2019, Ducharme-Barth et al. 2022) but limited application of PS models to account for it. An exception to this were recent studies by Rufener et al. (2021) and Pennino et al. (2019) but these were not carried out in the context of CPUE standardisation, which is the focus of PS models here. We anticipated that for target fisheries, PS models will improve spatial abundance predictions, and thus indices of abundance compared to conventional geostatistical models. This hypothesis is tested using simulations and an application to a major New Zealand trawl fishery data set (Chapter 4). Additional simulations are also conducted to explore how robust PS models are to violated model assumptions.

The second application of geostatistical models is for estimating indices of population abundance and associated variance from data collected from two-dimensional systematic surveys. Systematics surveys are a survey design that samples a geographic area in a grid-like fashion. Systematic surveys are an important survey design for population inference in forestry, ecology and fisheries (Millar & Olsen 1995, Fewster 2011, Hyun & Seo 2018). This is because they can be more practical to implement and in simulation studies the estimated population density has been shown to have lower variance than random designs under many conditions, (e.g. McGar-

vey et al. 2016). However, the most common systematic survey design consists of just a single primary sampling unit (PSU) and consequently there is no applicable design-based estimator of variance for the sample total. This has led to the development and application of several approximations based on variations of design-based variance estimators for random samples (Wolter 1984, McGarvey et al. 2016) and model-assisted variance estimators (Fewster 2011).

Recent reviews and studies of variance estimators for two-dimensional systematic surveys with a single PSU have lacked model-based estimator methods (D'Orazio 2003, McGarvey et al. 2016, Strand 2017). Exceptions include Simmonds & Fryer (1996), Aune-Lundberg & Strand (2014), who explore an estimator based on the concept of semi-variance (kriging) and Bartolucci & Montanari (2006), who also explore a model-based estimator that assumes the super-population follows a linear trend with homoscedastic and uncorrelated errors. One reason provided for the lack of model-based estimators is they "are more difficult to code and implement than the simpler design based systematic survey variance estimators" (McGarvey et al. 2016, pg 244). We examine properties of a geostatistical model-based variance estimator for the population total. We hypothesised it would perform better with respect to confidence interval coverage when compared to a range of currently applied methods. We also address the barrier for applying model-based estimators by incorporating all estimators into a generalised and easy to use R package.

Simulations are a key framework of this thesis to explore properties of models of interest. Given models are abstract simplifications of the true system, the objective is not to identify the correct model but rather to identify a useful model (Box 1979). A "useful" model in the context of this thesis should provide unbiased estimates of stock abundance and accurately represent uncertainty, whilst being robust to a range of plausible model misspecifications. Simulations consist of two models; an operating model (OM) and an estimation model (EM). The OM represents the hypothesised "true" system, and is responsible for generating synthetic observations. The EM is the model under investigation, which conducts inference using synthetic data from the OM. Key parameters and model quantities are compared between those estimated in the EM and true values assumed in the OM to evaluate model precision, bias, and robustness to a range of assumption violations.

An element of this thesis involved developing a generalised OM that could be used to test hypotheses laid out both in this thesis and beyond this work. An Agent Based

Model (ABM) was chosen for this task. An ABM is a generalisation of an Individual Based Model (IBM), where an agent represents a homogeneous group of individuals with identical characteristics also termed super-individuals. ABMs are an intuitive framework for emulating stock dynamics as individuals are the core components of the system (Grimm & Railsback 2013, Cao et al. 2016). Assumptions are made at the agent level, regarding growth, mortality, movement, and interactions with fisheries. When summaries are made across all agents, stock dynamics emerge that can be used to generate synthetic data; thus, creating a powerful OM for investigating hypothesis laid out in this research. ABMs are flexible OMs that allow for realistic mechanisms for simulating fish stocks.

1.1 Thesis outline

Chapter 2 describes methods that are typically employed to estimate indices of abundance from fishery-dependent catch and effort data. Chapter 3 explores a geostatistical model that can account for preferential sampled data (PS model). Simulations are conducted to illustrate biases in conventional geostatistical model inference when data is preferentially sampled. Proof of concept simulations are performed to show the PS model has identifiable parameters. Chapter 4 applies the PS model to a major New Zealand fishery, the Chatham Rise hoki (*Macruronus novaezelandiae*) trawl fishery. Chapter 5 describes the innovative ABM program CABM that was created to emulate the spatial distribution and productivity assumptions of hoki on the Chatham Rise. A simulation study is conducted using CABM in Chapter 6 to further explore the PS model and provide insight into convergence issues encountered in the case study. Chapter 7 explores a range of estimators from two-dimensional systematic survey data. Simulations are conducted to compare a geostatistical model-based estimator with existing methods. Finally, we summarise our results and outline future research in chapter 8.

Chapter 2

Catch Per Unit Effort (CPUE) standardisation

2.1 Overview

Catch Per Unit Effort (CPUE) standardisation is primarily a regression-based framework for estimating an index of relative abundance from catch and effort data. It is a fundamental method for informing fisheries management globally (Harley et al. 2001*b*, Hinton & Maunder 2003, Campbell 2015). The primary aim of CPUE standardisation is to remove trends and variation in catch rates from factors other than changes to the underlying stock abundance. This chapter describes models typically used in CPUE standardisation and how they are used to estimate an index of relative abundance.

2.2 Conventional CPUE models

In certain cases, only total catch and total effort are known for a fishery, termed nominal catch and effort data (Maunder et al. 2006). Due to the lack of information associated with nominal catch and effort data the analysis is restricted and often assumes total catch during time period t is proportional to the product of total effort and stock abundance,

$$C_t = qE_tN_t, \tag{2.1}$$

where C_t denotes total catch, E_t denotes total effort, N_t denotes stock abundance, and q is the proportion of stock which is captured for one unit of effort, often termed the catchability coefficient (Campbell 2015). This is often rearranged to

$$\frac{C_t}{E_t} = I_t = qN_t,$$
$$I_t \propto N_t,$$

where, I_t denotes the relative abundance at time t. An issue with nominal CPUE is that a unit of effort is expected to vary based on factors such as vessel, skipper experience, market conditions etc. If there are systematic trends in these factors over time, a proportionality assumption with constant q can be dangerous when using nominal CPUE for fisheries management (Hilborn 1985, Maunder et al. 2006).

More often than not, fishery-dependent catch and effort data have a range of covariates associated with each fishing event. These are either reported by the fishers themselves - such as location, targeted species, time of day, fishing gear configuration etc. - or derived from other sources such as satellite records (Reynolds et al. 2007) and weather stations. These additional variables can have substantial impacts on estimates of I_t, and the main task of CPUE standardisation is to identify factors that are practically significant and remove trends and variability caused by them.

When multiple covariates are available the natural extension from nominal CPUE (Equation 2.1) is to apply the Generalised Linear Model (GLM) (McCullagh & Nelder 2019, Gavaris 1980) to account for practically significant factors. The Generalised Additive Model (GAM) (Wood 2017) and Generalised Linear Mixed Model (GLMM) (Bolker et al. 2009) are extensions of the GLM framework and are commonly applied in CPUE analysis (Grüss et al. 2019). In order to apply these regression techniques, data resolution needs to be available at the fishing event (a single trawl, long-line or set net) or fishing trip (collection of fishing events) level. All three regression frameworks (GLM, GAM, and GLMM) have the same structural components:

1. Response variable ($\boldsymbol{y}$)
2. Systematic component ($\boldsymbol{X\beta}$)
3. Link function ($g()$)

but vary in their make up of the systematic components ($\boldsymbol{X\beta}$). GAMs allow non-linear relationships of explanatory covariates through the use of spline-based smoother functions, and GAMM are hierarchical models that include random-effect variables.

Catch rates have been modelled many ways in CPUE standardisations (Maunder & Punt 2004). The main approaches include treating the response variable as presence/absence (i.e. a Bernoulli random variable), strictly positive and right skewed (i.e. a Gamma or Lognormal random variable), or both. Common approaches for dealing with zeros in addition to right skewed data include zero inflated models (for count data) or the Tweedie distribution for semi-continuous (Thorson 2018). If these are not available/possible, another common approach is to use the hurdle-model or delta-model approach (Cragg 1971, Thorson 2018). This is where presence/absence is modelled as a Bernoulli random variable, and the positive values are modelled as a strictly positive response variable (i.e., Lognormal or Gamma) conditional on a presence. CPUE methods in this thesis focused on target fisheries which have high encounter rates (low frequency of fishing events with zero catch) and so focused on catch and effort datasets and models with a strictly positive response variable.

The response variable for each fishing event (observation) is the catch of a given stock denoted by $\boldsymbol{y} = (y_1, \ldots, y_n)^T$. All fishing events are spatially referenced where $\boldsymbol{s} = (s_1, \ldots, s_n)^T$ denotes a location within the spatial domain ($s_i = \{x_i, y_i\}$) and all fishing events have a temporal reference denoted by $\boldsymbol{t} = (t_1, \ldots, t_n)^T$. For certain fishing methods such as trawling there can also be an area-swept covariate denoted by $\boldsymbol{a} = (a_1, \ldots, a_n)^T$, in addition to a range of explanatory variables used to describe trends and variability in $\boldsymbol{y}$ denoted by $\boldsymbol{X_i} = (X_{i,1}, \ldots, X_{i,p})$ containing p covariates. The general form of the GLM model is

$$\begin{aligned} y_i|\eta_i, \boldsymbol{\phi} &\sim f(\eta_i; \boldsymbol{\phi}) \\ \eta_i &= g^{-1}\left(\boldsymbol{X_i\beta}\right), \quad i = 1, \ldots, n \end{aligned} \tag{2.2}$$

where, $g()$ is the link function, $f()$ is a density function with expectation η_i and dispersion parameter ϕ. $\boldsymbol{X_i}$ is the i^{th} row of the model matrix for the fixed effect coefficients, with estimable fixed effect coefficients $\boldsymbol{\beta}$. In the case where $\boldsymbol{y}$ is strictly positive, a commonly assumed form of $g()$ is the natural logarithm (Gavaris 1980, Maunder & Punt 2004). If this is assumed, Equation 2.2 can be extended to include

an offset covariate (often a covariate that expresses fishing effort) which changes the response variable from catch to catch per unit of the offset covariate. In the following models, the offset covariate is area

$$\begin{aligned} y_i|\eta_i, \boldsymbol{\phi} &\sim f(\eta_i; \boldsymbol{\phi}) \ , \\ \eta_i &= a_i \exp\left(\boldsymbol{X_i}\boldsymbol{\beta}\right), \quad i = 1, \ldots, n \ . \end{aligned}$$

An index of relative abundance is derived from the above model by including a temporal covariate denoted by $\boldsymbol{\beta^t}$, whether it was significant or not ($\boldsymbol{\beta^t} \in \boldsymbol{\beta}$, $\boldsymbol{\beta^t} = (\beta_1^t, \ldots, \beta_{n_t}^t)^T$). Temporal covariates were represented as a factor so there is an estimable coefficient for each level (time-period). For example, if an annual index was of interest there would be an estimable coefficient for each year. The temporal coefficients were used to derive an index of relative abundance. For the case when $g()$ is the natural log, the relative abundance at time t (I_t) is

$$I_t = \exp(\beta^t) \ .$$

Often the index of relative abundance reported is a scaled index whereby all elements of I_t are divided by the geometric or arithmetic mean (Francis et al. 2001, Campbell 2015) over estimated temporal effects in the time series,

$$\tilde{I}_t = \frac{I_t}{\left(\prod_{j=1}^{n_t} I_j\right)^{1/n_t}} \ . \tag{2.3}$$

Standard errors for $\tilde{I}_t$ are provided by Francis (1999). Interaction terms can also be included with $\boldsymbol{\beta^t}$. For this case, an index of relative abundance is extracted for all levels of the interacting covariate and $\boldsymbol{\beta^t}$.

A natural extension from the standard GLM is allowing nonlinear relationships between the explanatory variables and response variable. This is often implemented using the GAM framework (Wood 2017, Hastie & Tibshirani 1990) where spline-based smoothers are added into the systematic component,

$$y_i|\eta_i, \boldsymbol{\phi} \sim f(\eta_i; \boldsymbol{\phi}) \ , \tag{2.4}$$

$$\eta_i = g^{-1}\left(\boldsymbol{X_i}\boldsymbol{\beta} + f_1(x_{1,1}) + \cdots + f_p(x_{i,p})\right), \quad i = 1, \ldots, n \ , \tag{2.5}$$

where $f_p(.)$ is a spline-based smoother for covariate p with value $x_{i,p}$. This allows greater flexibility between continuous covariates and the response variable. It is still

assumed that $\boldsymbol{\beta^t} \in \boldsymbol{\beta}$, and an index of relative abundance is derived in the same manner as the GLM (Equation 2.3).

Spline-based smoothers split the covariate space into segments defined by "knots". Each knot location in the covariate space is denoted by $\boldsymbol{x^*} = (x_1^*, \ldots, x_{p_n}^*)^T$ and has an associated basis function denoted by $b_j(., x_j^*)$. The linear combination of basis functions along all knots make up the spline-based smoother $f(.)$. Spline-based smoothers are used in Section 4.4 which have the form,

$$f(x_1) = \sum_{j=1}^{p_n} \beta_j \times b_j \left(x_1, x_j^*\right) \ ,$$

where, j is one of p_n knots used with estimable coefficient β_j. When estimating β_j, the following penalty is added to the likelihood,

$$\lambda \int f''(x)^2 dx \ ,$$

where, f'' is the second order derivative of the spline-based smoothing function. Large values of the integral indicate high non-linearity of $f(.)$, where as values close to zero resemble a straight line. Hence, λ is a smoothing parameter (also termed the shrinkage parameter) that penalises for non-linearity. Large values of λ will amplify the penalty which will encourage the spline-based smoother towards a straight line. Small values will result in a lower penalty in the likelihood which can lead to a considerably non-linear form of the spline-based smoother. The above penalty can be reparameterisation into matrix form as

$$\lambda \int f''(x)^2 dx = \lambda \boldsymbol{\beta^T S \beta} \ ,$$

where, $\boldsymbol{S}$ is the penalty coefficient matrix (see Section 5.3.4 of Wood (2017)). This penalty is implemented using the following improper Gaussian prior

$$\boldsymbol{\beta} \sim \ \mathcal{N}\left(\mathbf{0}, \lambda \boldsymbol{S}\right) \ ,$$

where both λ and $\boldsymbol{\beta}$ are estimated and $\boldsymbol{\beta}$ is treated as a random-effect variable (see the following paragraph for details on random-effect variables). Section 4.4 employs tensor spline-based smoothing functions. The R package `mgcv` (Wood 2017)) is used to calculate the penalty matrix $\boldsymbol{S}$ and model matrix for basis functions. These matrices are passed to Template Model Builder (TMB) models (Kristensen et al. 2016)

which is the package used to evaluate likelihood functions and parameter gradients (see Section 2.4 for more information on inference with TMB). The effective degrees of freedom are reported for models with splines, which account for the number of coefficients and the level of shrinkage see Section 6.1 of (Wood 2017).

The last non geostatistical model covered here are GLMMs. These are hierarchical models which include random-effect variables in addition to fixed-effect parameters in the systematic component,

$$\begin{aligned} y_i|\eta_i, \boldsymbol{\phi} &\sim f(\eta_i; \boldsymbol{\phi}), \\ \eta_i &= g^{-1}\left(\boldsymbol{X_i\beta} + \boldsymbol{Z_i\gamma}\right), \quad i = 1, \ldots, n, \\ \boldsymbol{\gamma}|\boldsymbol{\psi} &\sim \pi(\boldsymbol{\psi}), \ . \end{aligned} \tag{2.6}$$

Where $\boldsymbol{Z_i}$ represents the i^{th} row of the model matrix for the random-effect variables, $\boldsymbol{\gamma}$ are estimable random-effect variables with hyperdistribution $\pi()$ and estimable hyperparameters $\boldsymbol{\psi}$. Hierarchical models allow pooling of information between effects within a factor in addition to allowing for inference on the hyper-distribution $\pi()$.

An important element of CPUE standardisation, as in all regression analysis is variable selection. Catch and effort data are often large data sets (i.e., Section 4.3). Variable selection and hypothesis testing with large data sets can encounter the "P-value problem" Chatfield (1995, pg 70). This is where model effects are classified as statistically significant due to decreasing standard errors as sample sizes increase. Chatfield (1995) state that for large sample inference, the aim is to identify terms of practical significance based on the magnitude of the effect. In this work, practically significant effects were identified using step-wise variable selection methods, and using stopping rules in variable selection procedures using percent deviance explained (section 4.3). This removes the issue of including statistically significant terms that are not accounting for practically significant trends or variation in catch rates.

2.3 Geostatistical models

Geostatistical models in this thesis consider the stock of interest to be estimated within a well-defined spatial region denoted by $\mathcal{D} \subset \mathbb{R}^2$ with known area A. Geostatistical models define $\mathcal{U}(\boldsymbol{s})$ as a continuous Gaussian process over the domain $\mathcal{D}$ $\{\mathcal{U}(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$. Two realisations of this process are represented in models throughout

this thesis. Firstly, $\omega(s_i)$ represents a time-invariant realisation and $\epsilon(t_i, s_i)$ represents a spatio-temporal realisation which describes the process at location s_i at time t_i.

The geostatistical model (also termed conventional geostatistical model throughout this thesis) extends conventional models (Equation 2.6) by incorporating these Gaussian processes as random-effect variables into the systematic component as follows

$$\begin{aligned} y_i|\eta_i, \boldsymbol{\phi} &\sim f(\eta_i; \boldsymbol{\phi}) \\ \eta_i &= g^{-1}\left(\boldsymbol{X_i\beta} + \boldsymbol{Z_i\gamma} + \omega(s_i) + \epsilon(t_i, s_i)\right) \\ \boldsymbol{\gamma}|\boldsymbol{\psi} &\sim \pi(\boldsymbol{\psi}) \\ \boldsymbol{\omega}|\boldsymbol{\theta_\omega} &\sim GF(\mathbf{0}, \Sigma_\omega) \\ \boldsymbol{\epsilon_t}|\boldsymbol{\theta_\epsilon} &\sim GF(\mathbf{0}, \Sigma_\epsilon) \qquad \text{for } t = 1 \\ &\sim GF(\boldsymbol{\xi\epsilon_{t-1}}, \Sigma_\epsilon) \quad \text{for } t > 1 \end{aligned} \tag{2.7}$$

where, ξ is the estimable parameter for an autoregressive process of order one. Both Σ_ϵ and Σ_ω are assumed to be stationary and have isotropic Matérn correlation structures (Equation 2.8) with estimable hyperparameters $\boldsymbol{\theta_\epsilon} = (\tilde{\tau}_\epsilon, \kappa_\epsilon)$ and $\boldsymbol{\theta_\omega} = (\tilde{\tau}_\omega, \kappa_\omega)$ respectively (see below for interpretations of these hyperparameters).

Some applications of spatio-temporal models have included spatial and temporal correlations in Σ_ϵ, as in Nottingham (2021). However, we found for applications with large data sets (i.e. Chapter 4), computational limitations meant this was not possible. For this reason, in this thesis Σ_ϵ only included spatial correlations. Temporal effects were considered by allowing the mean of $\boldsymbol{\epsilon_t}$ to follow an autoregressive process using the parameter ξ in Equation 2.7.

Isotropic correlations represent spatial dependence as a function of distance only (equal in all directions) and do not take into account relative position. The Matérn correlation function is defined as

$$Cor_M\left(U(s_i), U(s_j)\right) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\kappa\|s_i - s_j\|\right)^\nu K_\nu\left(\kappa\|s_i - s_j\|\right) \tag{2.8}$$

where, $\|.\|$ denotes the Euclidean distance and K_ν is the modified Bessel function of the second kind. The correlation is scaled by the marginal variance σ_M^2 to obtain the covariance matrix.

Parameterisation of the Matérn covariance implemented in this work has constants $d = 2$ and $\nu = 1$ and estimable parameters $\tilde{\tau}$ and κ. This differs from the parameterisation in Fuglstad et al. (2019), with the relationship $\tau = 1/\tilde{\tau}$

$$\rho = \sqrt{8\nu}/\kappa \qquad \text{and} \qquad \sigma_M = \tilde{\tau}^{-1}\kappa^{-\nu}\left(\sqrt{\frac{\Gamma(\nu + d/2)(4\pi)^{d/2}}{\Gamma(\nu)}}\right)^{-1}$$

where ρ is interpreted as the distance at which the correlation between two points is 0.1. Throughout this thesis, parameterisation for models can interchange between σ_M, ρ, κ and $\tilde{\tau}$. When configuring operating models often σ_M and ρ are used because they are more intuitive, but estimable parameters in estimation models are generally expressed with κ and $\tilde{\tau}$.

The choice of the Matérn correlation structure is based on recent advancements in geostatistical models made by Lindgren et al. (2011). They show that a GF with Matérn covariance denoted by $U(\boldsymbol{s})$ is the solution to the following stochastic partial differentiation equation (SPDE):

$$\left(\kappa^2 - \nabla\right)^{\alpha/2} u(\boldsymbol{s}) = \boldsymbol{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathbb{R}^d, \alpha = \nu + d/2, \kappa > 0, \nu > 0 \tag{2.9}$$

with ∇ being the Laplacian operator and $\boldsymbol{W}(\boldsymbol{s})$ a Gaussian process with unit variance.

As well as proposing the SPDE approach for GF applications, Lindgren et al. (2011) also extended results to Gaussian Markov Random Fields (GMRF) as solutions to SPDE and proposed the Finite Element Method (FEM) for applying these methods to irregular grids (Bakka 2018). GMRF are defined as a special case of the GF where only neighbourhood points in the covariance matrix include non-zero entries

$$U(s_i|\boldsymbol{s_{-i}}) = U(s_i|\boldsymbol{s_j} : j \in NB_i) \tag{2.10}$$

NB_i are the neighbourhood locations to s_i. This representation of GF has major computational benefits because the covariance matrix becomes sparse (contains many zeros).

To utilise this approach the spatial domain is partitioned into many (user constrained) non-intersecting triangles using Delaunay triangulation (Figure 2.1).

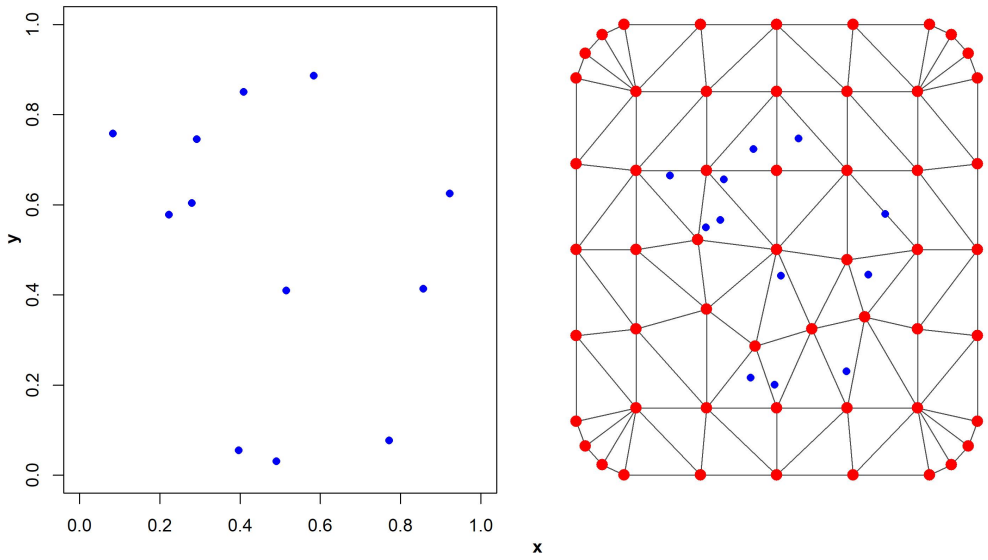


Figure 2.1: An illustration of an irregular grid used to represent a GF.

Each triangle has three vertices (red dots right panel Figure 2.1). Using this irregular grid, hereinafter referred to as the mesh, the GF at location s_i is interpolated over the spatial domain as

$$\omega(s_i) = \sum_{k=1}^{m} \psi_k(s_i)\delta_k \ , \tag{2.11}$$

where ψ_k are basis functions, δ_k are estimable Gaussian weights, $k = (1, \ldots, m)$ where m is the total number of vertices in the mesh. The choice of basis functions requires careful consideration to preserve the weak SPDE solution (Bakka 2018). Two basis functions investigated in this thesis were: the piece-wise linear basis function and the nearest neighbour function. These are illustrated for the one-dimensional case in Figure 2.2. The piece-wise linear provides the best approximation to the GF. However, when large sample data sets were analysed with the piece-wise linear basis function, computational limitations were encountered due to the large memory overhead. This was due to the sparse matrix arithmetic, which restricted its usage. Catch and effort data sets are often quite large (i.e., section 4.3) and can include tens of thousands of observations. The nearest neighbour approach uses considerably less memory than the piece-wise linear basis function, making it a more practical approach for large datasets. The approach taken in this thesis is to use linear basis

functions unless computational barriers were reached. When this happened we used the nearest neighbour approach. The nearest neighbour approach is also implemented in other spatio-temporal software such as the Vector Autoregressive Spatio-Temporal (VAST) model (Thorson 2019).

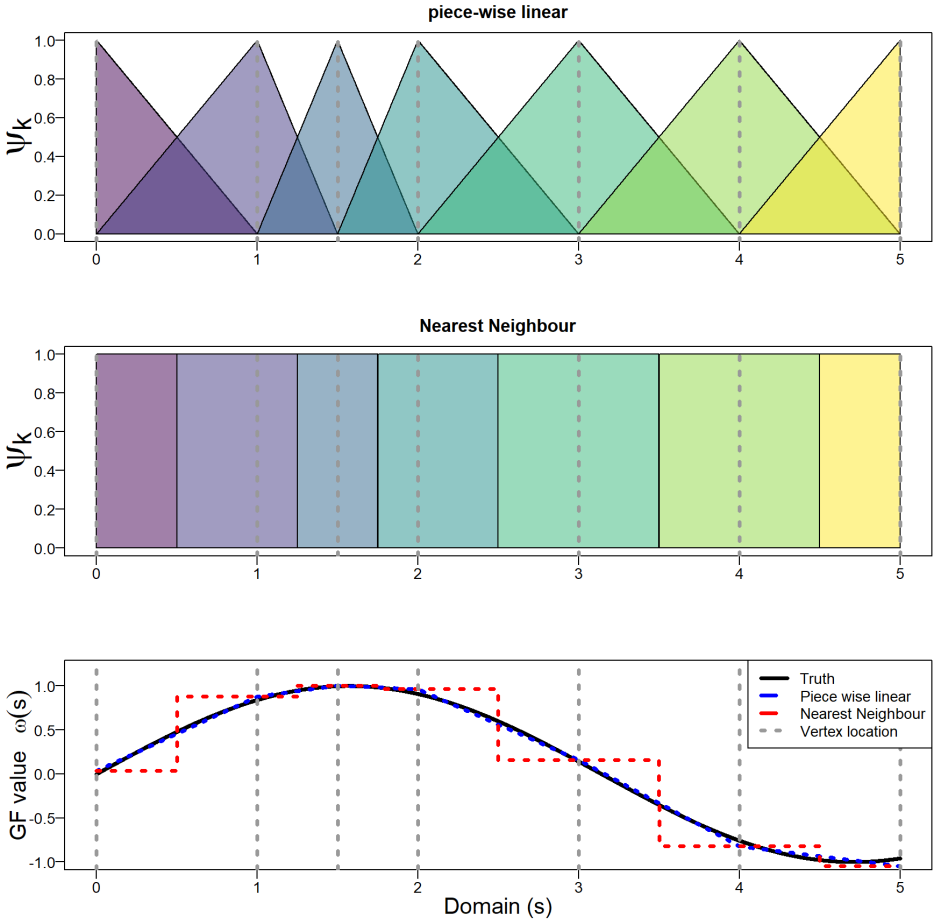


Figure 2.2: Illustration of one-dimensional GF with different basis function approximations explored.

Configuration of the mesh (Figure 2.1) is a subjective decision. Users need to consider attributes such as the number and location of vertices, and whether to include a boundary which can effect the GF approximation (Lindgren 2012). In this thesis, mesh configuration decisions were based on recommendations made in (Blangiardo & Cameletti 2015, Chapter 6.7). These include

1. Have triangles as regular as possible in size and shape

2. Consider convex or concave meshes for irregular domain shapes

3. Use outer extensions to account for inflated variance at the spatial boundaries (Lindgren 2012)

For large sample datasets, compromises were required between the number of vertices within a mesh (spatial resolution) and the model run time. Large sample data sets coupled with fine spatial resolution meshes could lead to hours, and sometimes days, for models to converge.

The above describes the SPDE approach used to implement Gaussian Fields in regression models. There are alternative geostatistical models that use kriging equations (Cressie 2015). These are not explored during this thesis and we recommend Chapter 7 of Gómez-Rubio (2020) and Chang et al. (2015) for a more detailed comparison of these two approaches. The SPDE approach was chosen in this thesis for its flexibility in response variable distributions, it can allow spatio-temporal GFs, ease of including covariates and availability of convenient software for its implementation, i.e. INLA (Lindgren & Rue 2015).

To derive an index of relative abundance from geostatistical models, the model is predicted over the entire spatial domain at each time step. The domain is partitioned into n_s discrete cells. Values for all terms in $\boldsymbol{\beta}$ are required for all cells and are expressed as the projection model matrix $\tilde{\boldsymbol{X}}^{\boldsymbol{t}}$.

$$I_t = \sum_{j=1}^{n_s} a_j g^{-1}\left(\tilde{\boldsymbol{X}}_{\boldsymbol{j}}^{\boldsymbol{t}}\boldsymbol{\beta} + \omega(s_j^*) + \epsilon(s_j^*, t)\right) \tag{2.12}$$

where $\tilde{\boldsymbol{X}}_{\boldsymbol{j}}$ is the j^{th} row corresponding to cell j with midpoint s_j^* and a_j is the area of that cell. This formulation has ignored random-effect variables $\boldsymbol{\gamma}$. Practically significant environmental and physical covariates assumed to influence spatial abundance should have known values for all cells in the partitioned domain. This will improve the models extrapolation over space. Values for other covariates in $\tilde{\boldsymbol{X}}^{\boldsymbol{t}}$ that are vessel or fishing event specific such as skipper experience, vessel size, sea conditions, time of day, etc. should be held constant over time and space as they are assumed to only influence catchability. Often the index I_t is scaled by the geometric or arithmetic mean as in Equation 2.3.

All geostatistical model analysis was conducted in the R statistical language (R Core Team 2020). Mesh and basis functions were generated using the R INLA

package (Lindgren & Rue 2015, Krainski et al. 2018) and model estimation conducted using the TMB R package (Kristensen et al. 2016).

2.4 Inference and model validation

Hierarchical models also termed mixed-effect models and latent variable models contain both fixed-effect parameters denoted by $\boldsymbol{\theta}$, hereinafter simply referred to as parameters, and random variables denoted by $\boldsymbol{u}$. The joint likelihood function denoted by $l(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{u})$ is a function that takes specific values of parameters and random variables, and returns a measure of the relative support conditional on observed data and distributional assumptions for the model

$$l(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{u}) = f(\boldsymbol{y}, \boldsymbol{u}; \boldsymbol{\theta}) = f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\theta}) f(\boldsymbol{u}; \boldsymbol{\theta}) \ .$$

Here, $f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\theta})$ is the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{U}$ and $f(\boldsymbol{u}; \boldsymbol{\theta})$ is the marginal density function for the random variables. The marginal likelihood function of $\boldsymbol{\theta}$ arising from data $\boldsymbol{y}$ requires integration of the joint likelihood with respect to random variables $\boldsymbol{u}$

$$l(\boldsymbol{\theta}; \boldsymbol{y}) = f(\boldsymbol{y}; \boldsymbol{\theta}) = \int f(\boldsymbol{y}, \boldsymbol{u}; \boldsymbol{\theta}) d\boldsymbol{u}. \tag{2.13}$$

In most cases, this integral cannot be expressed in closed form, and so approximations are required (Section 2.4.2).

Maximum likelihood estimates (MLE) for parameters $\boldsymbol{\theta}$ are evaluated by maximising the marginal likelihood,

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \left(l(\boldsymbol{\theta}; \boldsymbol{y}) \right) \tag{2.14}$$

Empirical Bayes estimates (Equation 2.15) are derived from the joint distribution and used as point estimates for $\boldsymbol{u}$, as well as model derived quantities.

$$\widehat{\boldsymbol{u}} = \arg\max_{\boldsymbol{u}} \left(f(\boldsymbol{y}, \boldsymbol{u}; \widehat{\boldsymbol{\theta}}) \right) \tag{2.15}$$

In practice we work with the log-likelihood instead of the likelihood.

The Fisher information matrix denoted by $\boldsymbol{I}(\boldsymbol{\theta})$ is used to derive the covariance of parameters,

$$\boldsymbol{I}(\boldsymbol{\theta}) = -\mathbb{H}(\boldsymbol{\theta}) = -\frac{\partial^2}{\partial \theta_i \partial \theta_j} l(\boldsymbol{\theta}; \boldsymbol{y}), \qquad 1 \leq i, j \leq p.$$

with $-\mathbb{H}(\boldsymbol{\theta})$ denoting the negative hessian matrix, and p is the total number of elements in $\boldsymbol{\theta}$.

The covariance for parameters $\boldsymbol{\theta}$ are derived by

$$\widehat{V}(\boldsymbol{\theta}) = \boldsymbol{I}(\boldsymbol{\theta})^{-1} \ .$$

For non-linear functions of $\boldsymbol{\theta}$ a Taylor series expansion, which is also known as the generalised delta-method (Fournier et al. 2012, Millar 2011), is used to calculate their approximate covariance matrix. Assuming $\boldsymbol{\theta}$ is approximately normal with mean $\boldsymbol{\theta}_0$, then from the generalized delta-method, the approximate mean and variance for $\phi(\widehat{\boldsymbol{\theta}})$ is

$$\phi(\widehat{\boldsymbol{\theta}}) \sim \mathcal{N}\left(\phi(\boldsymbol{\theta}_0), \phi'(\boldsymbol{\theta}_0) V(\boldsymbol{\theta}_0) \phi'(\boldsymbol{\theta}_0)^T\right) ,$$

where, $\phi'(\boldsymbol{\theta})$ is the derivative of the function with respect to the parameters, also termed the Jacobian. In practice $\widehat{V}(\boldsymbol{\theta})$ is substituted for $V(\boldsymbol{\theta}_0)$ and similarly $\phi'(\boldsymbol{\theta}_0)$ for $\phi'(\widehat{\boldsymbol{\theta}})$. The delta-method can be generalised to include the joint covariance of both parameters and random variables (Kristensen et al. 2016), and non-linear functions of both.

When applying maximum likelihood estimators to "non-standard" models that have no closed form, optimization is required to solve Equations 2.14 and 2.15. Fors inference to be valid, an assessment of whether the solution is local or global is necessary. For multidimensional parameter problems, this can be very difficult if not impossible to definitively conclude. Approaches used in this thesis to assess this criterion include multiple optimisation attempts with "jittered" starting parameter values within the parameter space, likelihood profiles. As well as assessing the gradient of the likelihood with respect to each parameter and checking the Fisher information matrix is inevitable. During initial model development eigenvector decomposition was also done on the Hessian matrix (Equation 2.4) to identify ill-posed models and parameterisations (Brown & Sethna 2003, Gábor & Banga 2015).

To evaluate a models goodness of fit, randomised quantile residuals were employed (Dunn & Smyth 1996, Scudilio & Pereira 2020)

$$r_i = \Phi\left(F\left(y_i; \widehat{\eta}_i, \widehat{\phi}\right)\right) \ , \tag{2.16}$$

were $\Phi(.)$ is cumulative distribution function for the standard normal distribution, $F(.)$ is the cumulative distribution function for the chosen density function $f(.)$ in

Equation 2.2. The estimated model expected value is denoted by $\widehat{\eta}_i$ and estimated dispersion parameter $\widehat{\phi}$.

All estimation models in this thesis were implemented in the TMB R package (Kristensen et al. 2016). TMB employs Automatic differentiation (AD) (Bischof & Griewank 1992, Fournier et al. 2012), the generalised delta-method for standard errors and the Laplace approximation for approximating the integral in Equation 2.13 (Thygesen et al. 2017). Optimisation of the marginal likelihood (Equation 2.13) uses R's inbuilt optimiser `nlminb` (Gay 1990).

2.4.1 Automatic Differentiation

Automatic differentiation (AD) is heavily used in the ecological literature, in part due to modelling complex non-linear biological processes (Bolker et al. 2009). There is also the use of integrated analysis, which incorporate multiple data sets and thus multiple likelihood functions for inference (Maunder & Punt 2013). There are generally no closed form solutions of maximum likelihood estimators for these cases, and AD is an approach for evaluating exact derivatives (within machine precision) of parameters given a set of candidate values. The derivatives are used in the optimisation routines to solve Equations 2.14 and 2.15, which are far superior to non-gradient based optimisers (Bertsekas 1997). AD has been shown to be more efficient and stable compared to historical gradient methods such as the finite differences approach and symbolic differentiation (Higham & Higham 2016). The later method can provide accurate derivatives but can be slow to evaluate for complex models (Fournier et al. 2012).

Automatic differentiation breaks down complex mathematical equations to root operations and corresponding derivatives. Then, it applies the chain rule to the sequence of root operations. Below is the least squares example from Fournier et al. (2012), which illustrates how AD is implemented. For the model

$$y_i \sim a + bx_i + \epsilon_i \tag{2.17}$$

parameter estimates for a and b are found in the least squares procedure by optimising the following function

$$S = -\sum_{i=1}^{n} (y_i - (a + bx_i))^2 \tag{2.18}$$

this is achieved when derivatives with respect to the parameters of interest are equal to 0. This is done for a and b by solving

$$\frac{\partial S}{\partial a} = \sum_{i=1}^{n} 2\left(y_i - (a + bx_i)\right) = 0,$$
$$\frac{\partial S}{\partial b} = \sum_{i=1}^{n} 2x_i\left(y_i - (a + bx_i)\right) = 0$$

AD breaks down Equation 2.18 into the root expressions as demonstrated in Table 2.1.

Table 2.1: Example of sequence of derivative evaluations for the least squares example

step	operation	value	partial derivatives
1	$t_1 = bx_i$	bx_i	$\frac{\partial t_1}{\partial b} = x_i,\ \frac{\partial t_1}{\partial x_i} = b$
2	$t_2 = a + t_1$	$a + bx_i$	$\frac{\partial t_2}{\partial a} = 1,\ \frac{\partial t_2}{\partial t_1} = 1$
3	$t_3 = y_i - t_2$	$y_i - (a + bx_i)$	$\frac{\partial t_3}{\partial y_i} = 1,\ \frac{\partial t_3}{\partial t_2} = -1$
4	$t_4 = t_3^2$	$(y_i - (a + bx_i))^2$	$\frac{\partial t_4}{\partial t_3} = 2t_3$
5	$S_i = -t_4$	$(y_i - (a + bx_i))^2$	$\frac{\partial S_i}{\partial t_4} = 1$

The partial derivatives from Table 2.1 are combined using the chain rule as follows,

$$\frac{\partial S_i}{\partial a} = \frac{\partial t_2}{\partial a}\frac{\partial t_3}{\partial t_2}\frac{\partial t_4}{\partial t_3}\frac{\partial S_i}{\partial t_4} = 2\left(y_i - (a + bx_i)\right)$$
$$\frac{\partial S_i}{\partial b} = \frac{\partial t_1}{\partial b}\frac{\partial t_2}{\partial t_1}\frac{\partial t_3}{\partial t_2}\frac{\partial t_4}{\partial t_3}\frac{\partial S_i}{\partial t_4} = 2x_i\left(y_i - (a + bx_i)\right)$$

2.4.2 The Laplace approximation

The Laplace approximation was used to approximate the complex integrals in Equation 2.13. TMB automates the Laplace approximation and is one of the attributes we used this software (Kristensen et al. 2016). The Laplace approximation is a cornerstone method used in frequentist inference for hierarchical models (Bolker et al. 2009). The general idea is to approximate the multivariate normal density with a multidimensional quadratic function. For the expression $g(u) = f(\boldsymbol{y}, u; \boldsymbol{\theta})$, where u is univariate, the objective is to approximate the following integral,

$$\int g(u)du\ .$$

Let $h(u) = \log(g(u))$, then

$$l(\boldsymbol{\theta}; \boldsymbol{y}) = \int \exp\left(h(u)\right) du$$

The approximation applies a 2^{nd}-order Taylor series expansion of $h(u)$ around u_0, the value which maximises $h(u)$,

$$\int \exp\left(h(u)\right) du \approx \int \exp\left(h(u_0) + h'(u_0)(u - u_0) + \frac{1}{2}h''(u_0)(u - u_0)^2\right) du$$

Assuming the log-likelihood surface is quadratic with respect to u, the first derivative at the maximum will equal zero ($h'(u_0) = 0$), simplifying to

$$l(\boldsymbol{\theta}; \boldsymbol{y}) \approx \int \exp\left(h(u_0) + \frac{1}{2}h''(u_0)(u - u_0)^2\right) du$$

given $h(u_0)$ is a constant which doesn't depend on u, it can be moved out of the integral, and with some rearrangement,

$$l(\boldsymbol{\theta}; \boldsymbol{y}) \approx \exp\left(h(u_0)\right) \int \exp\left(-\frac{1}{2}\frac{(u - u_0)^2}{-h''(u_0)^{-1}}\right) du$$

This is proportional by the constant $\sqrt{2\pi/-h''(u_0)^{-1}}$ to a normal density with mean u_0 and variance $-h''(u_0)^{-1}$. Assuming the integral in Equation 2.4.2 is from $-\infty$ to ∞ then the integral in Equation 2.4.2 will equal unity, leaving,

$$\begin{aligned} l(\theta; \boldsymbol{y}) &\approx \exp\left(h(u_0)\right) \sqrt{\frac{2\pi}{-h''(u_0)^{-1}}} \\ &\approx f(\boldsymbol{y}, u_0; \boldsymbol{\theta}) \sqrt{\frac{2\pi}{-h''(u_0)^{-1}}} \end{aligned}$$

This changes an integration problem to an optimisation problem to identify u_0 and evaluate the second order derivative, which is often faster than other forms of integration. This result can be extended to the multivariate case,

$$l(\boldsymbol{\theta}; \boldsymbol{y}) \approx f(\boldsymbol{y}, \widehat{\boldsymbol{u}}_{\boldsymbol{\theta}}; \boldsymbol{\theta}) \sqrt[q]{2\pi} det(\mathbb{H}(\widehat{\boldsymbol{u}}_{\boldsymbol{\theta}}))^{-1/2},$$

where, $\mathbb{H}(\widehat{\boldsymbol{u}}_{\boldsymbol{\theta}})$ is the Hessian matrix of second derivatives

$$\mathbb{H}(\widehat{\boldsymbol{u}}_{\boldsymbol{\theta}}) = \frac{\partial^2}{\partial \boldsymbol{u}^2} h(\boldsymbol{u}) \bigg|_{\boldsymbol{u} = \widehat{\boldsymbol{u}}_{\boldsymbol{\theta}}}.$$

the subscript $\boldsymbol{\theta}$ being added to show the dependency of $\boldsymbol{u}$ on $\boldsymbol{\theta}$. For each iteration during the optimisation in Equation 2.14 with candidate values denoted by $\boldsymbol{\theta_i}$, an additional optimisation is required to to find $\widehat{\boldsymbol{u}}_{\boldsymbol{\theta_i}}$ and evaluate $\mathbb{H}(\widehat{\boldsymbol{u}}_{\boldsymbol{\theta_i}})$ in order to apply the Laplace approximation.

Chapter 3

Preferential sampling models for CPUE standardisation

3.1 Overview

This chapter describes and explores a geostatistical model that accounts for preferentially sampled data (PS model) for CPUE standardisation. Given that, by nature, fishing is targeted and the management of many important global fisheries relies on fishery-dependent catch and effort data, exploring methods that account for preferential sampling is crucial for robust fisheries management.

Section 3.4 describes the PS model and Section 3.5 conducts two simulation studies. The first illustrates how preferentially sampled data can bias geostatistical models and how the PS model can alleviate this bias. The second explores properties of the PS model when the preferential sampling mechanism changes over time.

Finally, we explore a metric for indicating when data is preferentially sampled in Section 3.6.

3.2 Introduction

Research on fleet dynamics and fishing behaviour is an extensive field that crosses economic and social sciences (Van Putten et al. 2012). A key aim of this field is to understand significant behavioural drivers of fishing choices (Girardin et al. 2017). Key drivers found to influence fishing decisions include: expected catch rates (Vignaux 1996, Hutton et al. 2004, Salas & Gaertner 2004), management restrictions

(Pascoe et al. 2009), weather, market conditions (Smith 2005), distance from ports (Sampson 1992) and other economic factors (Sampson 1991, Hilborn 1985, Branch et al. 2006, Van Putten et al. 2012). This chapter proposes and explores a model that assumes fishing location choices are only associated with expected catch rates. Sensitivity analyses are also conducted to evaluate model robustness to deviations from this assumption.

The problem with applying spatial models to preferentially sampled fishery-dependent data is highlighted by Maunder et al. (2020, pg 16): "areas with abundant data (preferentially sampled areas) will inform undersampled and unsampled areas, which may lead to substantial bias when extrapolating over large spatial areas". This can be seen in Equation 2.7, where spatial random-effects are assumed to have a central tendency of zero. As the GFs are extrapolated into locations with little or no data, the predicted spatial random-effects will tend towards zero. This will result in predictions being closer to the sample average. Preferential sampling results in higher concentrations of sampling in areas with higher catch rates, resulting in a sample average that will be larger than the population average. This can result in inflated abundance estimates and positively biased abundance indices. This bias is illustrated in Section 3.5.1, and has also been highlighted in previous geostatistical studies (Diggle et al. 2010, Dinsdale 2018, Conn et al. 2017).

A method that has been applied in spatial analysis to account for preferential sampling (PS), is kriging (Montero et al. 2015, Cressie 2015). Kriging is a spatial modelling approach for estimating a variable over a continuous spatial domain. Kriging uses spatial data to estimate parameters for a variogram and then uses the estimated variogram to calculate weights when predicting at locations in the domain. The geostatistical models described earlier (Section 2.3) differ from kriging in their ability to interpolate variables with non-normal error assumptions (presence/absence), inclusion of covariates and easy application to spatio-temporal data. Simulations from previous studies have shown kriging could reduce bias in parameters and spatial predictions when data is preferentially sampled but could not alleviate the bias (Dinsdale 2018). Due to this conclusion and previously stated advantages of the SPDE approach (Section 2.3). The geostatistical model described earlier in Equation 2.7 was extended to include a model component that describes spatial distributions of fishing locations to account for the preferential sampling.

Spatially referenced observations are required when applying geostatistical models. The proposed model (the PS model) treats spatial locations of observed data, denoted by $\boldsymbol{s} = (s_1, \ldots, s_n)$, as a realisation from a 2-dimensional point process denoted by $\Lambda(\boldsymbol{s}), \boldsymbol{s} \in \mathcal{D}$. Preferential sampling models assume components describing the spatial distribution of $\boldsymbol{s}$ are also in the systematic component describing the spatial process of $\boldsymbol{y}$ (Diggle et al. 2010, Dinsdale 2018). In the context of a CPUE standardisation, areas of high abundance are assumed to have high sampling intensity (high frequency of fishing events) and high catch rates. We anticipated that for target fisheries, the main factor for spatial fishing locations and sampling intensity was expected catch rates. So, CPUE methods that incorporated this assumption could potentially improve model fits, predictions and estimated indices of abundance. This was informed from empirical studies that demonstrated similarities of vessel spatial distribution to that of the underlying stock of interest (Gillis & Peterman 1998).

The use of spatial models accounting for PS is more prevalent in mineral exploration, air quality and soil sciences (Diggle et al. 2010, Pati et al. 2011) than in fisheries research. Cases that have been applied in fisheries literature focused on spatial predictions (Pennino et al. 2019), were time-invariant (Conn et al. 2017) or included fishery-independent data in the analysis (Rufener et al. 2021). Our research is the first of our knowledge to explore properties of the PS model in the context of a CPUE standardisation, with a focus on estimated indices of abundance over time.

Recently, spatial studies using fishery-dependent data have commented that PS data may bias results and conclusions (Williams et al. 2018, Xu et al. 2019, Thorson et al. 2020). Other studies have simulation tested conventional geostatistical models with preferentially sampled fishery-dependent catch and effort data (Ducharme-Barth et al. 2022, Grüss & Thorson 2019, Carruthers et al. 2011) but we are not aware of any that have tried to explicitly account for PS in a CPUE standardisation analysis. The following chapters extend upon this research.

Geostatistical models generally assume sample locations ($\boldsymbol{s}$) are fixed and thus ancillary. In this instance, inferences on $\boldsymbol{\theta}$ are based on the conditional distribution of $\boldsymbol{y}$. PS models assume observed locations are dependent on the spatial stochastic processes $\mathcal{U}(\boldsymbol{s})$ and spatial covariates. The marginal likelihood for geostatistical

models including PS is extended from Equation 2.13 to

$$\begin{aligned} l(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{s}) = & f(\boldsymbol{y}, \boldsymbol{s}; \boldsymbol{\theta}) = \int f(\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{s}; \boldsymbol{\theta}) d\boldsymbol{u} \ , \\ & = \int f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{s}; \boldsymbol{\theta}) f(\boldsymbol{s}|\boldsymbol{u}; \boldsymbol{\theta}) f(\boldsymbol{u}; \boldsymbol{\theta}) d\boldsymbol{u} \ , \end{aligned} \tag{3.1}$$

where, $\boldsymbol{u}$ includes realisations of $\mathcal{U}(\boldsymbol{s})$ ($\omega(\boldsymbol{s})$ and $\epsilon(\boldsymbol{s}, t)$), $f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{s}; \boldsymbol{\theta})$ represents the conditional density of the data generating process, $f(\boldsymbol{s}|\boldsymbol{u}; \boldsymbol{\theta})$ represents the density of the spatial distribution of sampling locations given the spatial GF variables and spatial fixed-effects, and $f(\boldsymbol{u}; \boldsymbol{\theta})$ denotes the density of GF variables. We have excluded distributional statements with respect to random-effect variables $\boldsymbol{\gamma}$ (Equation 2.7). The joint model could easily be extended to include these, but this structure is not explored in the proceeding chapters.

3.3 Spatial point process

Two general approaches for describing $f(\boldsymbol{s}|\boldsymbol{u}; \boldsymbol{\theta})$ were found in the PS literature. The first is to use a continuous intensity function denoted by $\lambda(s_i)$, which denotes an inhomogeneous point process at location s_i (Diggle et al. 2010, Dinsdale 2018, Pati et al. 2011). The second approach partitioned the spatial domain into discrete non overlapping sampling units, and used a Bernoulli random variable to indicate whether a sampling unit was sampled or not (Conn et al. 2017). The approach assumed in this thesis applied the inhomogeneous point process with continuous intensity function. This choice was informed by Conn et al. (2017) who proposed the discrete partition approach and showed through simulations that the model was unstable (did not converge or estimated parameters at boundary constraints) for highly preferentially sampled data. Another consideration for choosing the inhomogeneous point process is its suitability to highly concentrated spatial sampling distributions. Fishery-dependent data can be highly concentrated in space (i.e., Section 4.3). In this case, spatial sampling units would need to be very fine, or aggregate observations repeated within a single spatial sampling unit. The spatial distribution of data in the case study (Section 4.3) highlighted the need for a model to account for high concentrations of samples across small spatial scales.

Let inhomogeneous point process $\Lambda(s)$ on $\mathcal{D} \subseteq \mathbb{R}^2$ have non-constant intensity function $\lambda(.)$. The intensity function is scaled to calculate the density function. For

location s_i on $\mathcal{D}$ the density is

$$f(s_i|\Lambda(\boldsymbol{s})) = \frac{\lambda(s_i)}{\int_{\mathcal{D}} \lambda(s)\, ds} .$$

The integral in the denominator is generally intractable and so approximations are required. The approximation used in this thesis partitioned the spatial domain into a fine grid with n_s cells where each cell has midpoint s_j^* and area a_j. The integral is approximated by summing over the grid, which can be computationally expensive (Simpson et al. 2016)

$$f(s_i|\Lambda(\boldsymbol{s})) \approx \frac{\lambda(s_i)}{\sum_{j=1}^{n_s} a_j \lambda\left(s_j^*\right)} . \tag{3.2}$$

3.4 Geostatistical models for preferential sampled data

To describe the PS model used in CPUE standardisation, the previously described conventional geostatistical model (Equation 2.7) was modified with respect to terms defining the systematic component η_i. These modifications were required so that both η_i and $\lambda(s_i)$ could be defined with common terms. The first modification dropped non spatial random-effect variables ($\boldsymbol{\gamma}$) from the systematic component. Although these can remain, we do not explore models with these components. The second modification expressed fixed-effect coefficients $\boldsymbol{\beta}$ as three separate elements: temporal $\boldsymbol{\beta^t}$, spatial $\boldsymbol{\beta^s}$ and catchability $\boldsymbol{\beta^q}$, with $\boldsymbol{\beta} = (\boldsymbol{\beta^t}, \boldsymbol{\beta^s}, \boldsymbol{\beta^q})^T$. The temporal coefficients contain a coefficient for each time-step in the data i.e., seasonal, annual. Spatial coefficients represent environmental and physical covariates that influence underlying spatial abundance, i.e., temperature and bathymetry. Catchability coefficients relate to other factors that affect catch rates, such as time of day, search time, tow duration etc. This separation of terms is needed for the intensity component in Equation 3.4 and when deriving abundance indices (Equation 3.5). The systematic component for the proposed geostatistical PS model follows

$$\eta_i = g^{-1}\left(\boldsymbol{X_i^t}\boldsymbol{\beta^t} + \boldsymbol{X_i^s}\boldsymbol{\beta^s} + \boldsymbol{X_i^q}\boldsymbol{\beta^q} + \omega(s_i) + \epsilon(s_i, t_i)\right) \tag{3.3}$$

where $\boldsymbol{X_i^t}$ is the i^{th} row for temporal model matrix, $\boldsymbol{X_i^s}$ spatial model matrix and $\boldsymbol{X_i^q}$ catchability model matrix. In order to split fixed effect coefficients into the separate

elements, constraints are required to ensure they are identifiable. A distinction in notation is made between an estimable constrained coefficient denoted by $\tilde{\boldsymbol{\beta}}^{\boldsymbol{t}}$ and $\boldsymbol{\beta}^{\boldsymbol{t}}$, which is the derived parameter applied in Equation 3.3. Temporal coefficients have the constraint

$$\sum_{k=1}^{n_t} \beta_k^t = 0 \ ,$$

where n_t is the number of temporal coefficients. This constraint results in $n_t - 1$ estimable parameters i.e, $\tilde{\boldsymbol{\beta}}^{\boldsymbol{t}} = (\tilde{\beta}_1^t, \ldots, \tilde{\beta}_{n_t-1}^t)^T$. These are transformed to $\boldsymbol{\beta}^{\boldsymbol{t}}$ as,

$$\beta_i^t = \tilde{\beta}_i^t \ , \quad i \in (1, \ldots, n_t - 1)$$

and the last temporal coefficient is defined as

$$\beta_{n_t}^t = -\sum_{k=1}^{n_t-1} \tilde{\beta}_k^t \ .$$

Levels within a categorical or discrete covariate in $\boldsymbol{\beta}^{\boldsymbol{s}}$ are constrained the same as the temporal coefficients. The catchability terms are parametrised similar to conventional GLM models, with an intercept term, contrast coefficients for categorical or discrete covariates and slope parameters for continuous covariates (Gelman & Hill 2006).

The intensity function for the point process is extended to have a temporal dimension. For observation i at location s_i in time period t_i

$$\begin{aligned} \lambda(s_i, t_i) &= \exp\{\beta^{pref} H(s_i, t_i)\}, \\ H(s_i, t) &= \boldsymbol{X}_{\boldsymbol{i}}^{\boldsymbol{s}} \boldsymbol{\beta}^{\boldsymbol{s}} + \omega(s_i) + \epsilon(s_i, t_i), \end{aligned} \tag{3.4}$$

where β^{pref} is an estimable coefficient that defines the strength of preferential sampling. $H(s_i, t_i)$ is the component that describes spatial distribution of catch rates and sampling intensity. When $\beta^{pref} > 0$ higher sampling intensity is associated with larger catch rates and vice versa, demonstrated in Figure 3.1.

The density function in Equation 3.2 is extended to incorporate the temporal dimension

$$f(s_i | \Lambda(\boldsymbol{s}, t_i)) = \frac{\lambda\left(s_i, t_i\right)}{\sum_{j=1}^{n_s} a_j \lambda\left(s_{j,i}^*\right)} \ .$$

The index of abundance from the PS model is derived by predicting spatial and temporal terms over the partitioned domain with n_s cells

$$I_t = \sum_{j=1}^{n_s} a_j g^{-1}\left(\boldsymbol{\beta}^{\boldsymbol{t}}(t) + \tilde{\boldsymbol{X}}_{\boldsymbol{j}}^{\boldsymbol{s},\boldsymbol{t}} \boldsymbol{\beta}^{\boldsymbol{s}} + \omega(s_j^*) + \epsilon(s_j^*, t)\right) \ , \tag{3.5}$$

where $\tilde{\boldsymbol{X}}^{\boldsymbol{s,t}}$ is the matrix containing spatial covariate values for all cells of the partitioned domain for time t, $\tilde{\boldsymbol{X}}_{\boldsymbol{j}}^{\boldsymbol{s,t}}$ is the j^{th} row of the matrix corresponding to cell j with mid point s_j^*, area a_j and $\boldsymbol{\beta^t}(t)$ is the temporal effect for time-step t.

3.5 Simulations

Simulations were utilised to illustrate the consequences of preferentially sampled data with conventional geostatistical model inference in addition to demonstrating that the proposed PS model had identifiable parameters. Three simulations were conducted in the remaining parts of this chapter. The first simulation ignored time and illustrated the consequence of PS with a simple spatial OM (Section 3.5.1). The second simulation added a temporal dimension which is more akin to a CPUE standardisation analysis. This simulation also investigated temporal changes to the preferential sampling mechanism for the point process (Section 3.5.2). The last simulation explored a correlation metric for identifying PS data (Section 3.6). The simulation investigated robustness of the PS model when assumptions regarding the point process were violated.

3.5.1 Illustrating preferential sampling

Methods

To illustrate the effects of preferentially sampled data on geostatistical model inference, a non-temporal spatial OM and two EMs were configured. Both OM and EMs were configured with identical model assumptions regarding the response variable

$$\begin{aligned} y_i &\sim \mathcal{N}\left(\eta_i, \sigma_y^2\right) \ , \\ \eta_i &= \beta_0 + \omega(s_i) \ , \\ \omega(\boldsymbol{s})|\tau, \kappa &\sim \mathcal{GF}\left(\mathbf{0}, \boldsymbol{\Sigma}_\omega\right) \ . \end{aligned} \tag{3.6}$$

No spatial covariates were assumed, with only an intercept denoted by β_0 and GF denoted by $\omega(\boldsymbol{s})$. $\boldsymbol{\Sigma}_\omega$ was assumed to have an isotropic Matérn correlation structure. Parameter values for the OM are given in Table 3.1.

The intensity function assumed for the OM and one of the EMs was

$$\lambda(s) = \exp\left(\beta^{pref}\left(\beta_0 + \omega(s)\right)\right), \quad \forall\ s\ \in \mathcal{D}\ .$$

The simulation explored different sampling strategies, where a sampling strategy is defined by different values of β^{pref}. These sampling strategies emulate avoidance behaviour ($\beta^{pref} = -1.5$), simple random sampling ($\beta^{pref} = 0$) and targeting behaviour ($\beta^{pref} = (1.5, 3.0)$), illustrated in Figure 3.1.

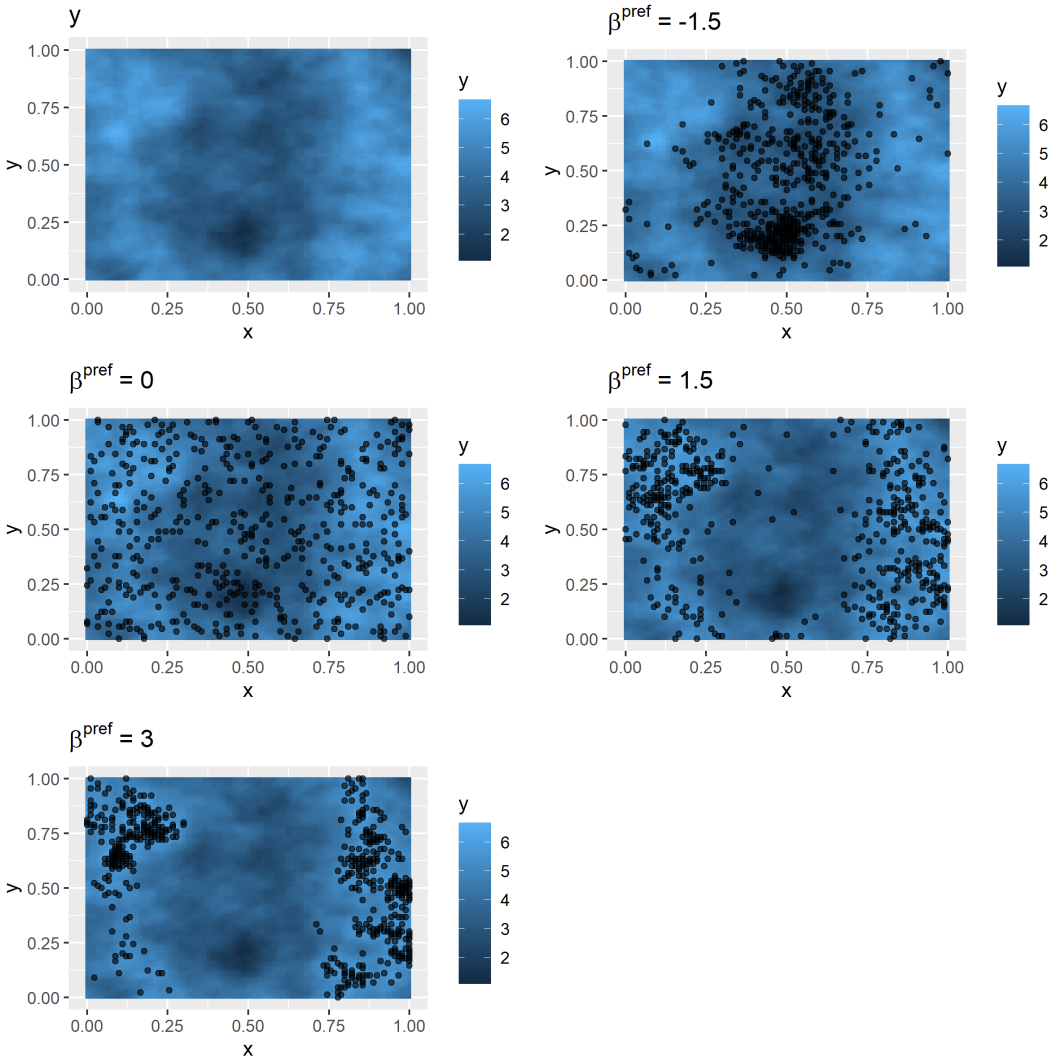


Figure 3.1: Sample locations (black dots) from a single simulation realisation across the four sampling strategies (defined by β^{pref}). The underlying response variable was constant among the realisations, displayed in the top left panel.

In addition to simulating data for a range of β^{pref} a range of observation variance values denoted by σ_y^2 was explored to simulate data. This was to investigate the consequence of observation error for each sampling strategy and inform if observation

error effected estimates of β^{pref}. In total, there were 20 OM scenarios made up of four levels of β^{pref} with five levels of σ_y^2 nested within each level of β^{pref}.

Table 3.1: Model parameters assumed for the operating model.

parameter	values
β_0	4
σ_y^2	0.1, 0.6, 1.1, 1.6, 2.1
σ_M^2	1.5
κ (ρ)	18.86 (0.15)
β^{pref}	-1.5, 0, 1.5, 3

Two EMs were applied to each set of simulated data. Both EMs had the same model equations describing $\boldsymbol{y}$, which were identical to the OM (Equation 3.6), but varied in whether they included the point process component in the joint likelihood (Table 3.2). $\mathcal{M}_0$ denotes the conventional geostatistical model which treats the observation as ancillary (no preferential sampling component), and $\mathcal{M}_1$ denotes the PS model that includes sampling intensity.

Table 3.2: EM model assumptions.

Model label	Systematic component	Intensity function (PS)
$\mathcal{M}_0$	$\eta_i = \beta_0 + \omega(s_i)$	-
$\mathcal{M}_1$	$\eta_i = \beta_0 + \omega(s_i)$	$\lambda(s_i) = \exp\{\beta^{pref}\,(\beta_0 + \omega(s_i))\}$

The OM approximated the continuous domain by partitioning it into a fine spatial lattice with equal area cells when simulating the GF and response variable. The algorithm for generating a simulated data set for a given OM scenario followed:

1. Simulate the GF over the entire spatial domain $\omega(\boldsymbol{s}) \sim \mathcal{GF}\,(\mathbf{0}, \boldsymbol{\Sigma}_\omega)$.

2. Simulate response variable over the entire spatial domain using the cell midpoints (s_j^*) of the partitioned domain,

$$y_j \sim \mathcal{N}\left(\beta_0 + \omega(s_j^*), \sigma_y^2\right) \ .$$

3. Generate $n = 100$ sample locations $s_i \sim f(\lambda\,(\boldsymbol{s}))$ and use the corresponding y_j at each sample location for the response variable.

The above algorithm was repeated to generate 100 simulated data sets for each scenario.

The target quantity compared between the OM and EMs was the total population denoted by N. This was approximated by summing over the grid

$$N = \sum_{j=1}^{n_s} a_j y_j \ ,$$

where, a_j is the area of cell j and n_s is the number of cells in the partitioned domain. This quantity is analogous with how an index of abundance is derived from geostatistical models (Equation 3.5).

Results

Figure 3.2 shows the relative error in $\widehat{N}$ between EMs and OM. When data was sampled preferentially, the EM that accounted for this (EM = $\mathcal{M}_1$) produced less biased estimates $\widehat{N}$. This figure also highlighted that this conclusion was consistent across varying levels of observation errors.

The bias from EM $\mathcal{M}_0$ in Figure 3.2 was due to the model's poor extrapolation in areas that had little to no data. This is highlighted in Figure 3.3, which illustrates how both EMs extrapolate expected values into these spatial regions. The effect was most pronounced for scenario $\beta^{pref} = 3$ (bottom row in Figure 3.3). Both EMs had very few observations in the centre of the spatial domain. $\mathcal{M}_0$ extrapolated close to the model expected average. Due to the preferential sampling selecting locations in regions of high values, the model average was inflated compared to the true average over the entire spatial domain, thus overestimating the total population. This is in contrast to EM $\mathcal{M}_1$ which is the correctly specified model linking low sampling intensity with lower values of $\boldsymbol{y}$, thus generating less biased estimates.

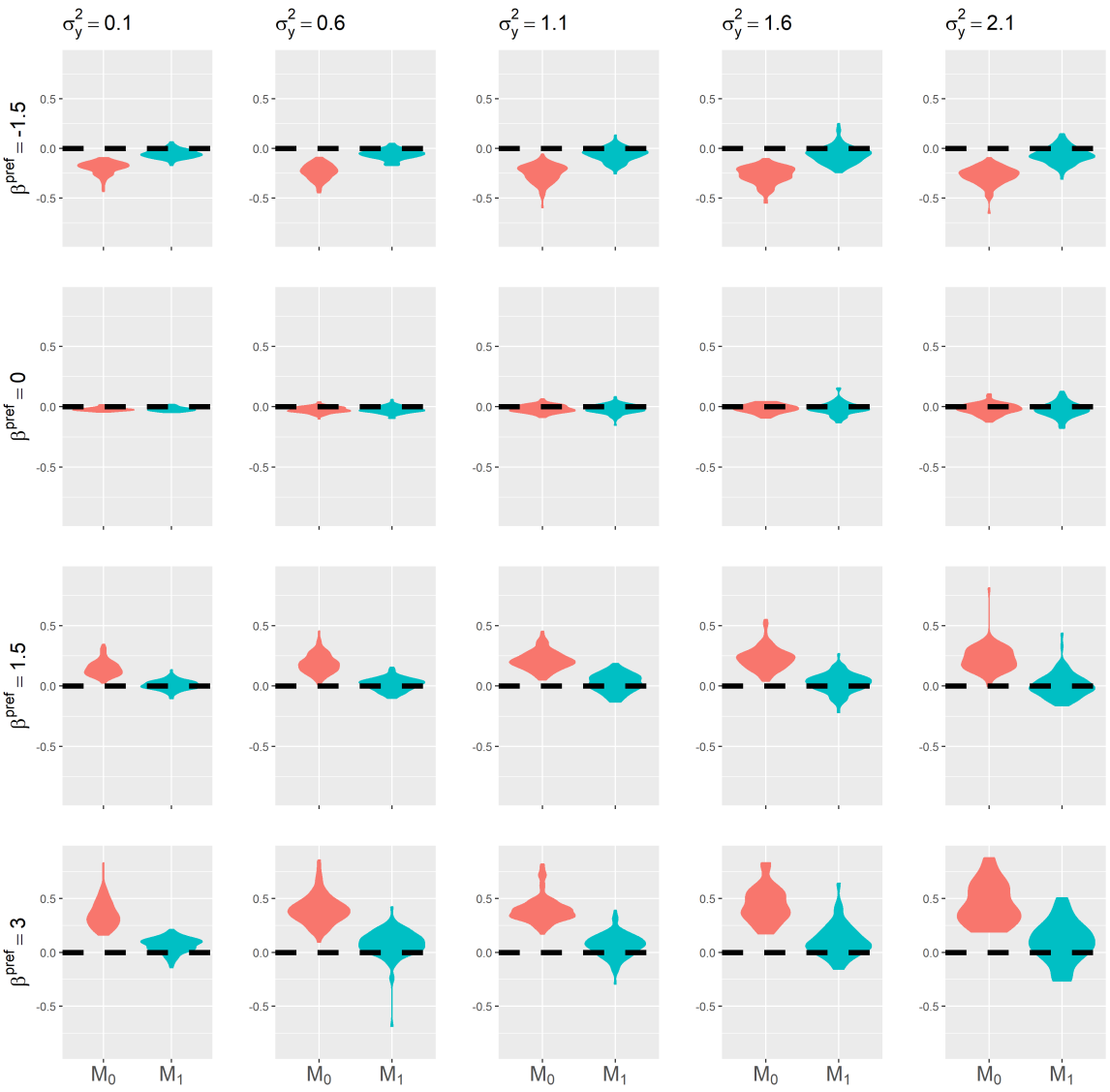


Figure 3.2: Relative error of estimated total population ($\widehat{N}$) from 100 simulation across all OM scenarios.

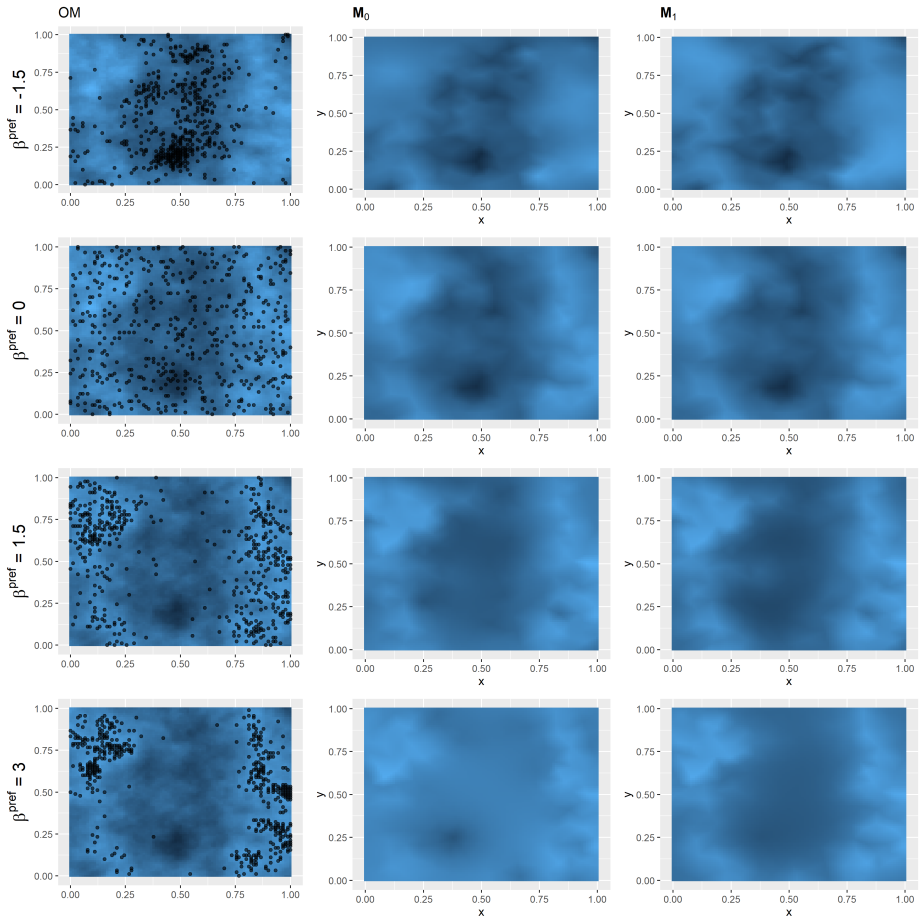


Figure 3.3: Response variable from the OM with sample locations (left panels) and model expected values from the two EMs for one realisation.

3.5.2 Spatio-temporal simulation

The previous simulation illustrated the effects of preferential sampled data on inference using conventional geostatistical models but was overly simplistic. To explore biases that pertain closer to a CPUE standardisation, a simulation with a temporal dimension was conducted. The main objective of a CPUE standardisation is to estimate an index of abundance over time, so it is important to understand how the bias demonstrated in the previous simulation changes with time, and if β^{pref} remains identifiable with spatio-temporal data.

The following simulation served two purposes; firstly to validate that β^{pref} was identifiable when a temporal dimension was included and secondly, to investigate consequences when β^{pref} changed over-time. The simulation study by Ducharme-Barth

et al. (2022) explored biases in estimated indices of abundance with preferentially sampled data but the preferential sampling process was constant over the time series. If indices of abundance are scaled, that is when each element of the index is divided by the mean of all values in the index (Equation 2.3), the bias demonstrated in the previous simple simulation (Section 3.5.1) could be inconsequential assuming the PS process was consistent over time.

The motivation behind exploring trends in the PS process over time was based on hypothesised shifts in spatial fishing effort and behaviour due to management actions such as area closures (Branch et al. 2006, Langley 2020), spatial gear restrictions (Hannah 2003), observer effects (Grüss et al. 2019), and move-on rules (Dunn et al. 2014). Move-on rules are temporary spatial closures triggered by catch rate or by-catch threshold limits being met, and are becoming more prevalent in fisheries management (Geange et al. 2020).

Methods

The OM and four EMs developed had consistent assumptions regarding $\boldsymbol{y}$. They assumed three covariates: time-step denoted in the systematic component as $\boldsymbol{X^t\beta^t}$, spatial region as $\boldsymbol{X^s\beta^s}$ (Figure 3.6), fleet as $\boldsymbol{X^q\beta^q}$, and a time-varying GF denoted by $\epsilon(s_i, t_i)$, for location s_i at time t_i. These covariates were chosen because they are commonly available/applied in CPUE standardisations (Hoyle & Langley 2020, Bentley et al. 2012). The catch rate model followed

$$\begin{aligned} y_i &\sim \Gamma\left(k, \theta_i\right) \\ \theta_i &= \frac{\eta_i}{k} \\ \eta_i &= a_i \exp(X_i^t\boldsymbol{\beta^t} + X_i^s\boldsymbol{\beta^s} + X_i^q\boldsymbol{\beta^q} + \epsilon(s_i, t_i)) \\ \epsilon(s_i, t_i) &\sim \mathcal{GF}\left(\mathbf{0}, \boldsymbol{\Sigma_\epsilon}\right) \ . \end{aligned} \tag{3.7}$$

Where a_i is the area for fishing event i, Γ denotes the Gamma density with shape parameter denoted by k, θ_i is the scale parameter, and the link function is the log link. The temporal varying GF was assumed to have an uncorrelated mean with an isotropic Matérn correlation structure. The Gamma density was chosen for its ability to describe right skewed data, which is commonly observed in fishery-dependent catch and effort data (Maunder & Punt 2004) and in the New Zealand case-study (Section 4.3).

The intensity function in the OM had a time-varying PS coefficient denoted by $\beta^{pref}(t)$

$$\lambda(\boldsymbol{s}, t) = \exp\left(\beta^{pref}(t)\left(\boldsymbol{X}^{\boldsymbol{s}}\boldsymbol{\beta}^{\boldsymbol{s}} + \epsilon(\boldsymbol{s}, t)\right)\right) . \tag{3.8}$$

The assumption of a time-varying PS coefficient was not assumed in any EMs (Table 3.4). The PS model EMs did contain a time-invariant PS coefficient denoted by β^{pref}.

Four OM scenarios were explored. Each had a different trend in $\beta^{pref}(t)$, illustrated in Figure 3.4. The trends were; constant, a step change increasing, step change decreasing and normally distributed values based on $\beta^{pref}(t) \sim \mathcal{N}(1.5, 0.6^2)$. Parameter values assumed in the OM are supplied in Table 3.3. All four OM scenarios simulated 100 synthetic data sets and the four EMs defined in Table 3.4 were fitted to each simulated data set. To investigate identifiability of β^{pref} in the PS EMs ($\mathcal{M}_1$ and $\mathcal{M}_3$), log-likelihood profiles were conducted for the constant OM scenario. Likelihood profiles are a typical diagnostic for parameter identifiability (Millar 2011, Maunder & Piner 2014).

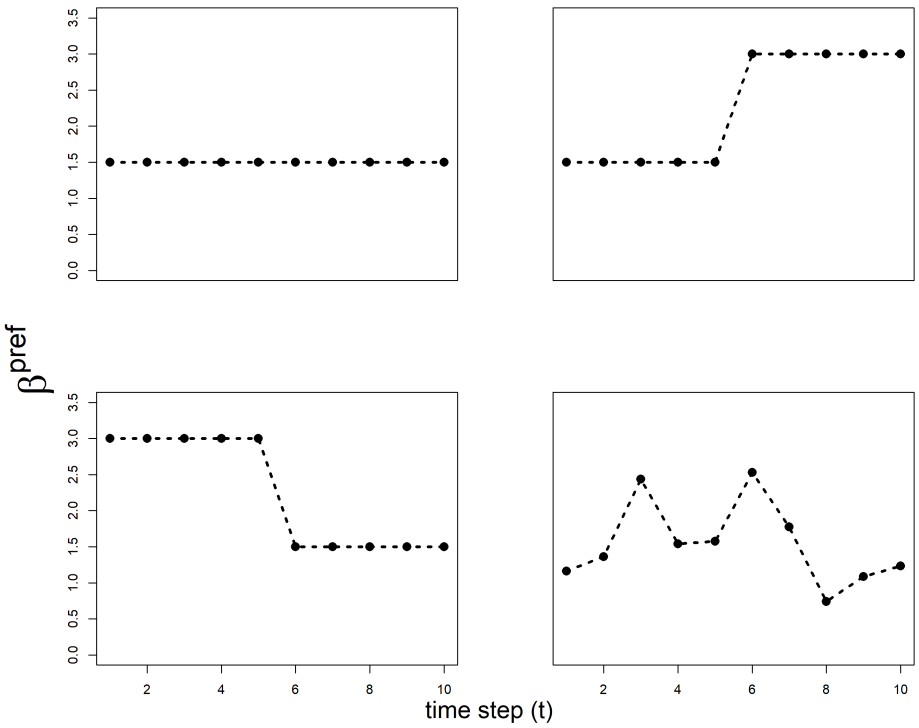


Figure 3.4: The four preference sampling trends assumed in each OM scenario.

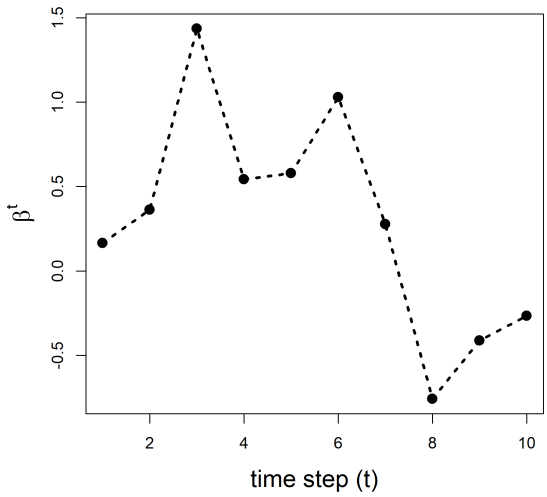


Figure 3.5: Time step coefficients (β^t) assumed for the OM.

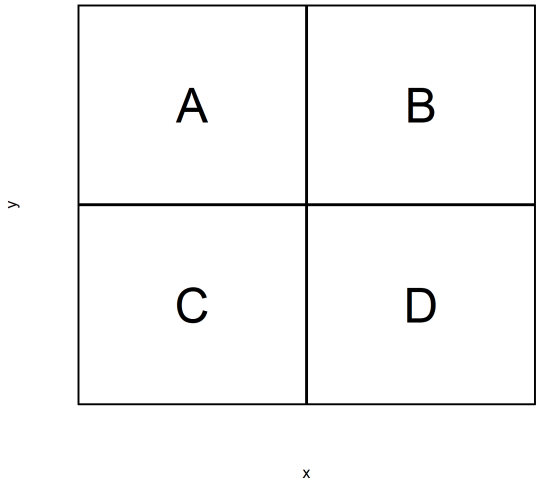


Figure 3.6: Four discrete regions were assumed in the operating model. This spatial covariate is often available for CPUE standardisation in the form of spatial reporting regions.

Table 3.3: List of parameter values set for the operating model (Equations 3.7 & 3.8).

Parameter	value
Time ($\boldsymbol{\beta^t}$)	See Figure 3.5
Region ($\boldsymbol{\beta^s}$)	$\beta^s_A = 0.73, \beta^s_B = -1.02, \beta^s_C = 0.24, \beta^s_D = 0.07$
Fleet ($\boldsymbol{\beta^c}$)	$\beta^c_1 = -0.70, \beta^c_2 = 0.31, \beta^c_3 = -0.20$
Gamma dispersion	$k = 10$ roughly equates to coefficient of variation (CV ≈ 0.3)
σ^2_M	1
κ (ρ)	10.6 (30)
PS Coefficient $\beta^{pref}(t)$	See Figure 3.4

The algorithm for generating a simulated data set was similar to the simple simulations (Section 3.5) and followed:

1. Simulate the GF over the entire domain for each time step $\epsilon(\boldsymbol{s}, t) \sim \mathcal{GF}(\mathbf{0}, \boldsymbol{\Sigma_\epsilon})$
2. Calculate η_i over the entire spatial domain based on Equation 3.7 and parameter values in Table 3.3
3. Simulate response variable with Equation 3.7 using values outlined in Table 3.3
4. Simulate 1000 sample location in each time step $\boldsymbol{s} \sim f(\lambda(\boldsymbol{s}, t))$

Table 3.4: Estimation models (EMs) explored in the simulations.

Model label	Systematic component	Intensity function (PS)
$\mathcal{M}_0$	$X^t_i\boldsymbol{\beta^t} + X^s_i\boldsymbol{\beta^s} + X^q_i\boldsymbol{\beta^q} + \omega(s_i)$	-
$\mathcal{M}_1$	$X^t_i\boldsymbol{\beta^t} + X^s_i\boldsymbol{\beta^s} + X^q_i\boldsymbol{\beta^q} + \omega(s_i)$	$\lambda(s_i) = \exp\{\beta^{pref}(X^s_i\boldsymbol{\beta^s} + \omega(s_i))\}$
$\mathcal{M}_2$	$X^t_i\boldsymbol{\beta^t} + X^s_i\boldsymbol{\beta^s} + X^q_i\boldsymbol{\beta^q} + \epsilon(s_i, t)$	-
$\mathcal{M}_3$	$X^t_i\boldsymbol{\beta^t} + X^s_i\boldsymbol{\beta^s} + X^q_i\boldsymbol{\beta^q} + \epsilon(s_i, t)$	$\lambda(s_i, t_i) = \exp\{\beta^{pref}(X^s_i\boldsymbol{\beta^s} + \epsilon(s_i, t_i))\}$

The geometrically scaled index of relative abundance denoted as $\tilde{I}_t$ (described in Section 2.2) was compared between OM and EM for each scenario and simulation set:

$$\tilde{I}_t = \frac{I_t}{\left(\prod_{j=1}^{n_t} I_j\right)^{1/n_t}}, \tag{3.9}$$

where,

$$I_t = \sum_{j=1}^{n_s} a_j \exp\left(\boldsymbol{\beta^t}(t) + \tilde{\boldsymbol{X}}_{\boldsymbol{j}}^{\boldsymbol{s}}\boldsymbol{\beta^s} + \epsilon(s_j^*, t)\right),$$

where, a_j is the area for cell j with midpoint s_j^* in the domain containing n_s cells, $\boldsymbol{\beta^t}(t)$ is the temporal fixed-effect for time t and $\tilde{\boldsymbol{X}}_{\boldsymbol{j}}^{\boldsymbol{s}}$ is the spatial projection matrix for the region spatial covariate.

Results

Log-likelihood profiles from five randomly chosen simulations for the constant β^{pref} OM scenario (top left panel Figure 3.4) demonstrated that β^{pref} was identifiable for the PS EMs ($\mathcal{M}_1$ and $\mathcal{M}_3$). This was concluded by the quadratic likelihood surface and estimated minima near the true value. One of the likelihood profiles is presented in Figure 3.7 for EM $\mathcal{M}_3$. This was a necessary validation step, before further model exploration (via the second objective) could be achieved.

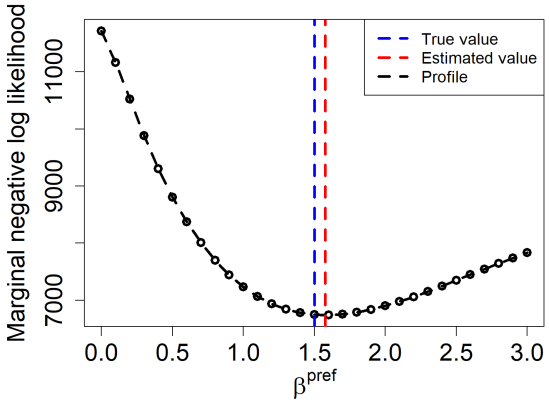


Figure 3.7: Negative log-likelihood profile for β^{pref} from one realisation of the simulation.

Relative error for the geometrically scaled index of relative abundance (Equation 3.9) is shown in Figure 3.8. This showed when preferential sampling was constant through time expressed in the OM as a constant value of $\beta^{pref}(t)$ (simulation scenario label constant), then there was negligible bias between in the EMs explored (top-row panel Figure 3.8). When PS changed over-time, expressed in the OM with temporal shifts in $\beta^{pref}(t)$ (Figure 3.4), models that accounted for PS ($\mathcal{M}_2$ and $\mathcal{M}_3$) reduced

bias in $\tilde{I}_t$, but could not remove it. This bias was expected because none of the EMs explored contained time-varying parametrisations with respect to β^{pref}. Models that contained time-varying GF ($\mathcal{M}_2$ and $\mathcal{M}_3$) estimated more precise abundance indices compared with EMs that had time-invariant GFs ($\mathcal{M}_0$ and $\mathcal{M}_1$).

Figure 3.8 (right hand column) also showed PS model $\mathcal{M}_1$ had a different trend in relative error to the PS model $\mathcal{M}_3$. The PS model with time-invariant GF ($\mathcal{M}_1$) had similar trends in bias compared to conventional geostatistical EMs $\mathcal{M}_0$ and $\mathcal{M}_2$. These models over-estimated relative abundance in years with high PS ($\beta^{pref} = 3$) and underestimated relative abundance in years with lower PS ($\beta^{pref} = 1.5$).

The EM that had the least biased estimates of relative abundance was $\mathcal{M}_3$. This EM under-estimated abundance when data was sampled under high PS and over-estimated abundance when PS was lower. The estimated values for β^{pref} (Table 3.5) were higher in EM $\mathcal{M}_3$ compared to $\mathcal{M}_1$. This coupled with the flexible time-varying GF assumptions is believed to be why $\mathcal{M}_3$ was the least biased estimator of relative abundance.

Table 3.5: Median and 95% quantiles of $\widehat{\beta}^{pref}$ across 100 simulations.

EM	Constant	Increase	Decrease	Random Draw
$\mathcal{M}_1$	1.11 (0.97 - 1.26)	1.85 (1.63 - 2.1)	1.84 (1.6 - 2.1)	1.23 (0.99 - 1.43)
$\mathcal{M}_3$	1.51 (1.47 - 1.54)	2.13 (2.01 - 2.27)	2.09 (1.98 - 2.25)	1.52 (1.42 - 1.64)

An additional EM was explored that included an estimable time-varying parameter denoted by β_t^{pref}. Initial model runs encountered model convergence issues (estimated parameters constrained at bounds). The only way to obtain a stable model was to assume restrictive constraints through the use of highly restrictive penalties on $\beta^{pref}(t)$ and use penalized likelihood estimation. For this reason, the extended model wasn't explored any further.

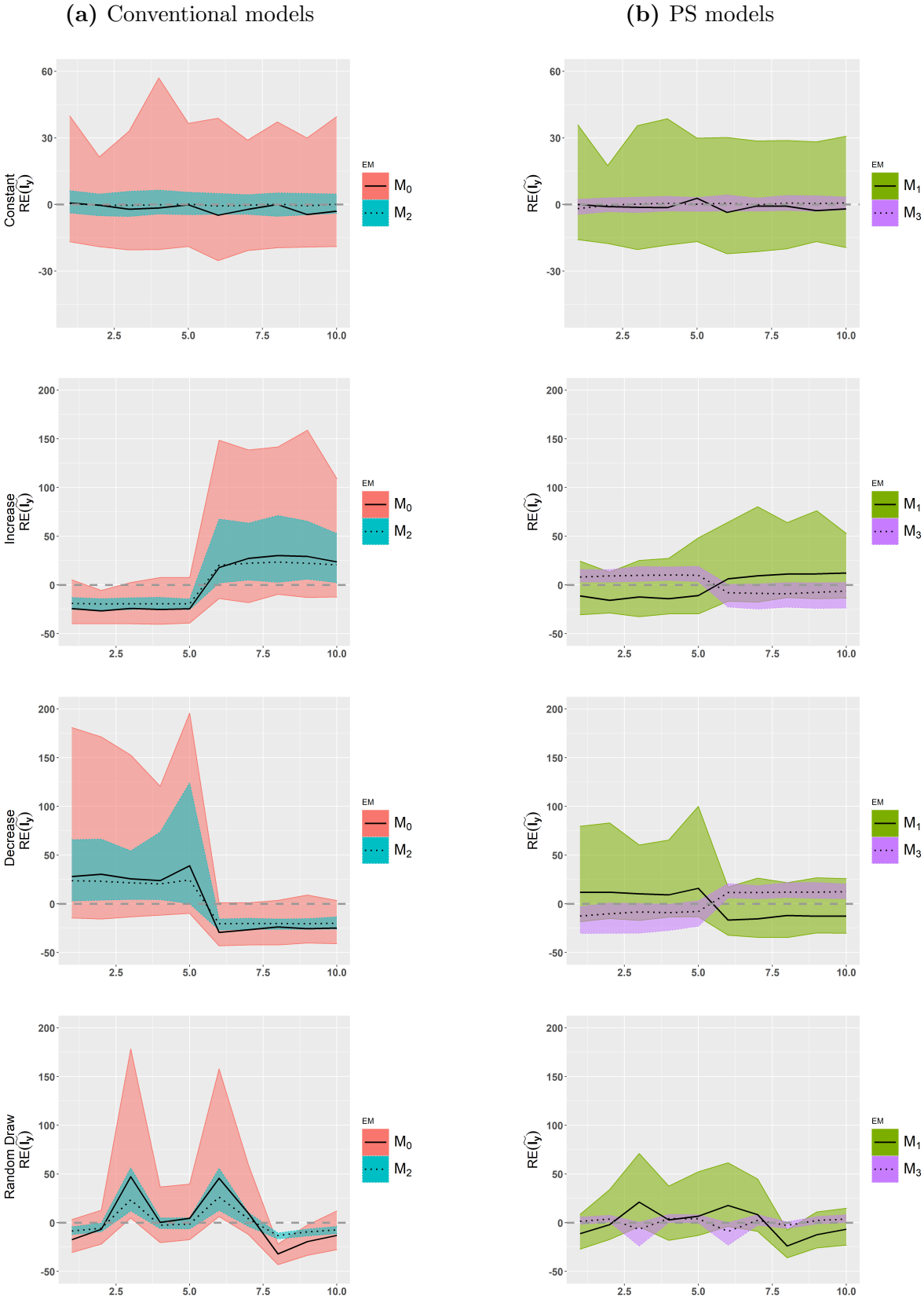


Figure 3.8: Relative error as a percentage from 100 simulations in estimated scaled index of abundance ($\widehat{\tilde{I}_y}$). Shaded area represent 95% quantiles and black lines are median relative error values among the 100 simulations.

3.6 Preferential sampling correlation metric

The previous simulations highlighted the importance of preferential sampling and its influence on geostatistical model inference in a CPUE standardisation. These simulations were presented in the context of known sampling mechanisms but in reality, fishing decisions will be more complex. This section explores a correlation metric derived from conventional geostatistical models with the purpose of identifying preferential sampled data and indicate if a PS model is appropriate. The aim of the metric was to identify correlations between predicted catch rates and sampling frequency for discrete cells that were sampled.

The correlation metric used the Pearson correlation coefficient. The correlation is calculated using predicted values from a conventional geostatistical model (Section 2.3) and the number of observations in each cell (sample size for each cell). The spatial domain is partitioned consistent with the extrapolation grid used for deriving indices of abundance (Equation 3.5) and the correlation is calculated for the subset of cells that contain at least one observation denoted by ζ_t at time t

$$r_t = \frac{\tilde{n}_t \left(\sum\limits_{j \in \zeta_t} n(j,t) \widehat{y}(s_j,t) - \sum\limits_{j \in \zeta_t} n(j,t) \sum\limits_{j \in \zeta_t} \widehat{y}(s_j,t) \right)}{\sqrt{\tilde{n}_t \left(\sum\limits_{j \in \zeta_t} n(j,t)^2 (\sum\limits_{j \in \zeta_t} - \sum\limits_{jin\zeta_t} n(j,t))^2 \right) \tilde{n}_t \left(\sum_j \widehat{y}(s_j,t)^2 (\sum\limits_{j \in \zeta_t} - \sum\limits_{j \in \zeta_t} \widehat{y}(s_j,t))^2 \right)}}, \tag{3.10}$$

where, $n(j,t)$ is the number of observations in cell j at time t, $\widehat{y}(s_j,t)$ is the model predicted value in cell j with midpoint s_j, and $\tilde{n}_t$ is the number of cells that are sampled at time t.

This correlation metric is conditional on model fits that generate $\widehat{y}(s_j,t)$. If this model does not fit the data well this metric is expected to be of limited value. Using simulations in the following section, properties of this metric were explored.

Methods

Two applications of the metric were used to explore its properties. The first application was a simulation study that followed from the simple simulation (Section 3.5.1),

but misspecified the intensity function. The second retrospectively applied the metric to the previous spatio-temporal simulations (Section 3.5.2).

For the first simulation, the response variable was identical to Equation 3.6,

$$\begin{aligned} y_i &\sim \mathcal{N}\left(\eta_i, \sigma_y^2\right) \\ \eta_i &= \beta_0 + \omega(s_i) \\ \omega(\boldsymbol{s})|\tau, \kappa &\sim \mathcal{GF}\left(\mathbf{0}, \boldsymbol{\Sigma}_\omega\right) \end{aligned} \tag{3.11}$$

The divergence from the simple simulation is with respect to $\Lambda(\boldsymbol{s})$. The intensity function was extended to include an auxiliary covariate which is expressed as a GF denoted by $\chi(\boldsymbol{s})$,

$$\begin{aligned} \lambda(s_i) &= \exp\left(\beta^{pref}\left(\beta_0 + \omega(s_i)(1-p) + \chi(s_i)p\right)\right) \\ \chi(\boldsymbol{s})|\sigma_\chi, \kappa_\chi &\sim \mathcal{GF}\left(\mathbf{0}, \boldsymbol{\Sigma}_\chi\right) \ . \end{aligned}$$

The parameter $p \in [0, 1]$ was varied among OM scenarios. This weighted the contributions of the auxiliary covariate $\chi(s_i)$ relative to $\omega(s_i)$ when defining the intensity function $\Lambda(\boldsymbol{s})$. As $p \to 0$ the resulting sampling point pattern tended towards preferentially sampled data. In contrast, as $p \to 1$ the point pattern was only a function of the auxiliary covariate $\chi(s_i)$.

Five OM scenarios were developed, each exploring different values of p with parameters defined in Table 3.7. An illustration of the resulting point patterns is given in Figure 3.9.

All five OM scenarios used the same 100 realisations of $\zeta(\boldsymbol{s})$, $\omega(\boldsymbol{s})$ and y_i. They each generated different sampling locations which corresponded to different simulated data sets due to their different assumptions regarding $\lambda(s_i)$. Each OM scenario generated 100 simulated data sets where two EMs described in Table 3.6 were fitted to the simulated data sets. Relative error in $\widehat{N}$ was used for EM comparison.

Table 3.6: EM model assumptions.

Model label	Systematic component	Intensity function (PS)
$\mathcal{M}_0$	$\eta_i = \beta_0 + \omega(s_i)$	-
$\mathcal{M}_1$	$\eta_i = \beta_0 + \omega(s_i)$	$\lambda(s_i) = \exp\{\beta^{pref}\left(\beta_0 + \omega(s_i)\right)\}$

Table 3.7: Model parameters assumed for the operating model

parameter	values
β_0	4
σ^2_y	0.1
σ^2_M	1.5
κ (ρ)	18.86 (0.15)
σ^2_χ	1.5
κ_χ (ρ_χ)	18.86 (0.15)
p	0, 0.25, 0.5, 0.75, 1
β^{pref}	1.5

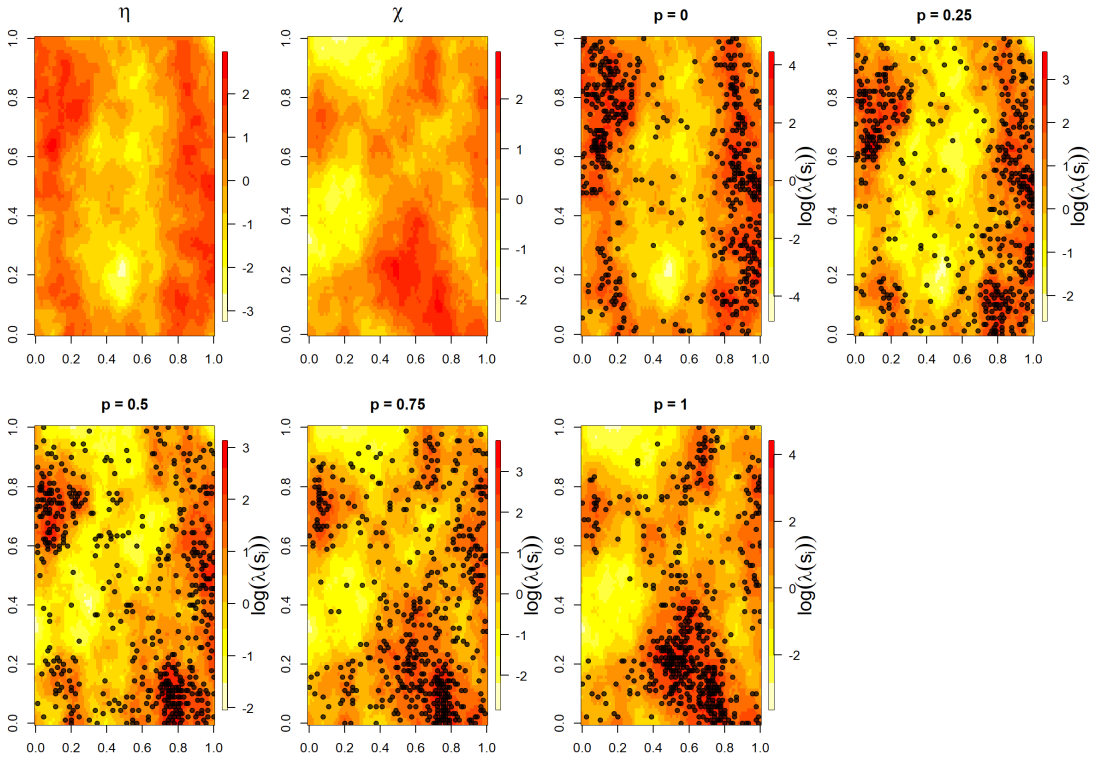


Figure 3.9: A single realisation with n = 500 of varying sampling point patterns across simulation scenarios.

Results

This simulation showed that when $p \geq 0.5$ the PS model ($\mathcal{M}_1$) looked to be biased and a less precise estimator of N compared to conventional EM (Figure 3.10). This

suggested when auxiliary factors had as much influence in fishing location as catch rates, the PS model might result in more biased inference compared with conventional geostatistical models. The resulting median value for r_t from the OM scenario with $p = 0.5$ was 0.33 (Figure 3.11).

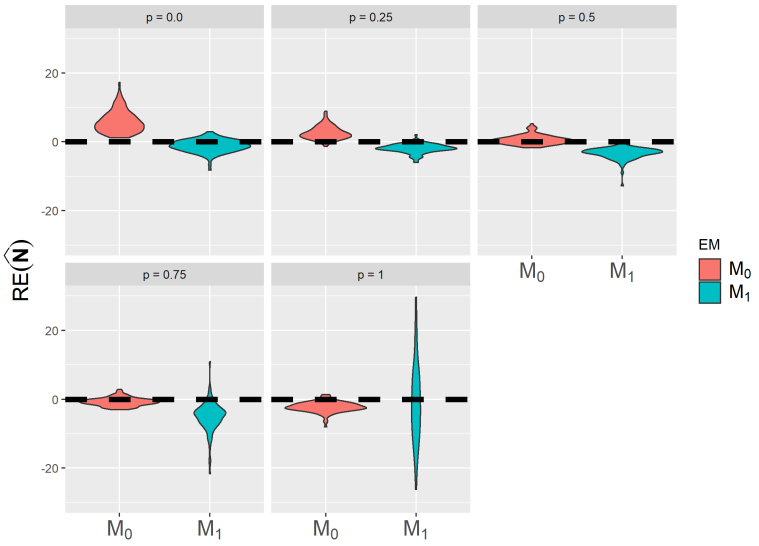


Figure 3.10: Relative error of $\widehat{N}$ between the two EMs from 100 simulations.

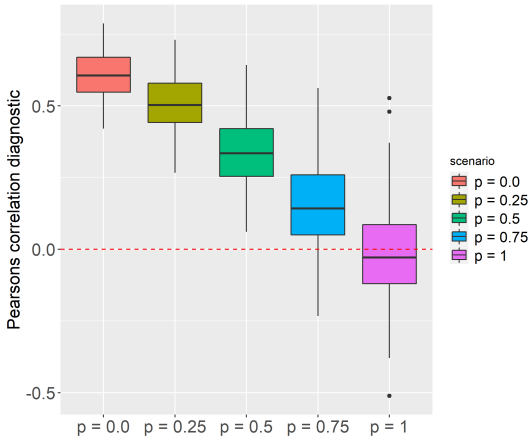


Figure 3.11: Summary of the diagnostic for all OM scenarios.

When the correlation metric was retrospectively applied to the spatio-temporal simulations from the previous Section 3.5, there were no obvious trends in the metric

over time, shown in Figure 3.12. This indicated the metric may have limited use in identifying temporal trends in the sampling mechanism.

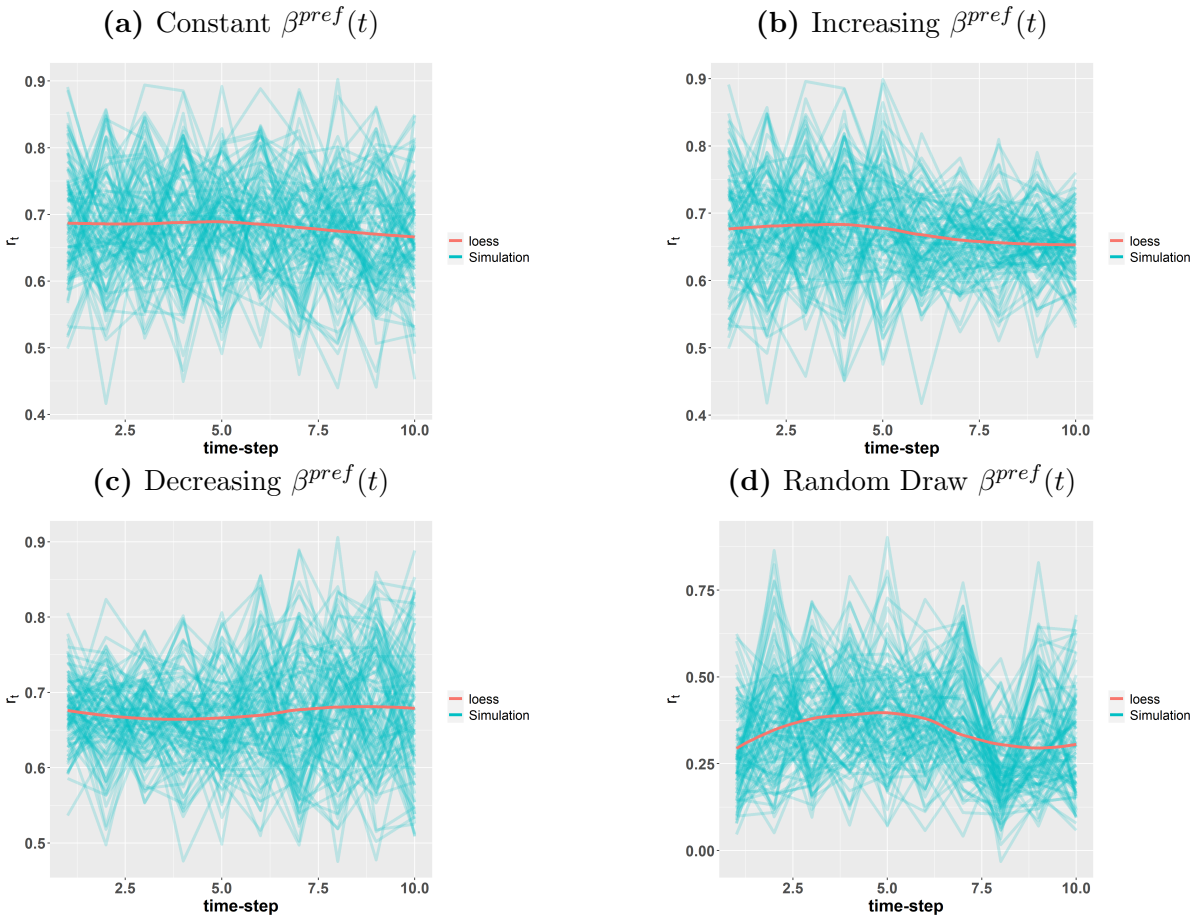


Figure 3.12: Correlation metric (r_t from Equation 3.10) applied to EM $\mathcal{M}_1$ from simulations in section 3.5

3.7 Discussion

This chapter developed a geostatistical model that accounted for preferentially sampled data for CPUE standardisation (the PS model). Simulations were conducted to illustrate bias from conventional geostatistical models with preferentially sampled data, and how the PS model can alleviate that bias (Section 3.5.1). Additional simulations showed that when the PS model was correctly specified, it could estimate unbiased estimates of β^{pref} for models with spatio-temporal GFs.

Section 3.5.2 highlighted that even if an EM correctly specified covariates affecting catch rates, if the sampling mechanism changed over time, substantial biases in indices of relative abundance could occur. This reaffirmed previous recommendations from Branch et al. (2006) and Hilborn (1985) that CPUE standardisation studies should put more resources and effort into understanding fishing behaviour and market conditions in addition to factors that affect catch rates.

The proposed correlation metric explored in Section 3.6 could be a useful metric for identifying preferentially sampled data in general but was limited in identifying trends in the PS process over time. This highlighted an area of future research. Methods that can identify trends on the PS process would be valuable when interpreting and using estimated abundance indices in stock assessments. The correlation metric explored in this chapter ignored unsampled regions due to the breakdown in linearity between the two variables. A metric based on the Poisson distribution or the spatial point pattern test and its extensions (Andresen 2016, Wheeler et al. 2018) should be considered in future research.

This chapter extended the recent simulation study by Ducharme-Barth et al. (2022) and highlighted the need for a caveat in their concluding statement (Ducharme-Barth et al. 2022, pg 10) "unsurprisingly, the random and preferential spatial sampling patterns produced abundance indices that were approximately unbiased". A useful caveat would be "assuming the PS sampling process does not change over-time". Because fishers operate in dynamic systems with changing catch limits and market conditions, this caveat should not be over looked.

All the geostatistical and PS models described in this chapter were compiled into a publicly available R package called `CPUEspatial`, found at the following url https://github.com/Craig44/CPUEspatial. The aim of this chapter was to demonstrate if PS models were applicable to CPUE standardisation analysis and whether they could improve upon existing methods. Having an R package offers the stock assessment community a tool that is easy to use for applying models demonstrated in this thesis.

This section used simulations to validate the proposed PS model had identifiable parameters. Results focused on biases in derived quantities that pertain to a CPUE standardisation. An important consideration not mentioned here is model comparisons and goodness of fit, this is discussed in detail in the next chapter where

both conventional models and the PS model are applied to data from a major New Zealand fishery.

Results from simulation were derived under ideal conditions, where the EM and OM had consistent assumptions and model formulations. Results from such studies are not easily extrapolated to real data sets (Francis 2012). Chapter 6 conducts a more comprehensive simulation using a more realistic operating model to address this limitation.

Chapter 4

The Chatham Rise hoki fishery

4.1 Overview

This chapter describes characteristics of a major New Zealand hoki (*Macruronus novaezelandiae*) trawl fishery and applies the PS model with conventional models to estimate an index of relative abundance. Prediction error and estimated abundance indices are compared between the PS model and conventional models to further investigate the utility of the PS model.

4.2 Introduction

Hoki is New Zealand's largest commercially fished species by volume, peaking at around 250 000 metric tonnes in the 2000s (Fisheries New Zealand 2018). Hoki has a wide geographic distribution around New Zealand which supports multiple fisheries. The Chatham Rise is located to the east of New Zealand's South Island (Figure 4.1) and sustains the second largest New Zealand hoki fishery

Currently, hoki is managed as one management unit in New Zealand but the stock assessment model (McKenzie 2017) assumes two separate biological stocks, an Eastern and Western stock. It also assumes both stocks have a common nursery ground on the Chatham Rise, but have separate spawning and resident grounds. The Eastern stock also resides on the Chatham Rise outside of the spawning period (O'Driscoll et al. 2011). The Western stock migrates south to their resident ground in the Sub-Antarctic once they reach maturity (Figure 4.2).

The Chatham Rise hoki trawl fishery was chosen for this analysis because it had characteristics described in the list below. These were assumed desirable for applying and evaluating the PS model.

- It has high spatial resolution data dating back to 1991 (Figure 4.1).
- It is a highly targeted fishery. Ninety six percent of all hoki caught (by weight) was by hoki targeted fishing events (Figure 4.3).
- It is fished by long range vessels with extended fishing trip durations. Eighty percent of hoki biomass caught was from trips at sea for 20 days or longer. Because they are not constrained to locations near ports, this provides fishers with greater opportunity to explore a wider range of fishing locations and cover more ground. Thus, they were more likely to exploit regions of high density
- There is a long-term scientific trawl survey (Ballara et al. 2017, Stevens et al. 2021) that provided additional data on hoki abundance, spatial distribution and biological information for the region.

The presence of a fishery-independent survey means management of this fishery does not depend on abundance indices from fishery-dependent catch and effort data. Nonetheless, it was thought to be a useful case study fishery for exploring properties of the PS model.

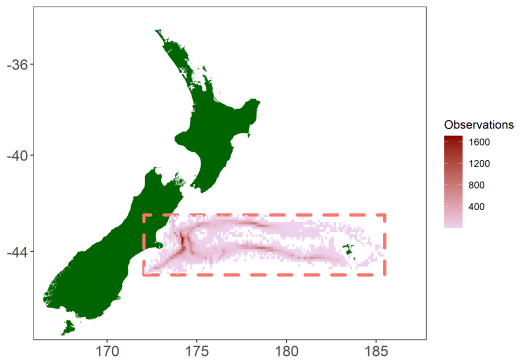


Figure 4.1: Spatial distribution of fishing events (observations) that caught hoki from the Chatham Rise trawl fishery.

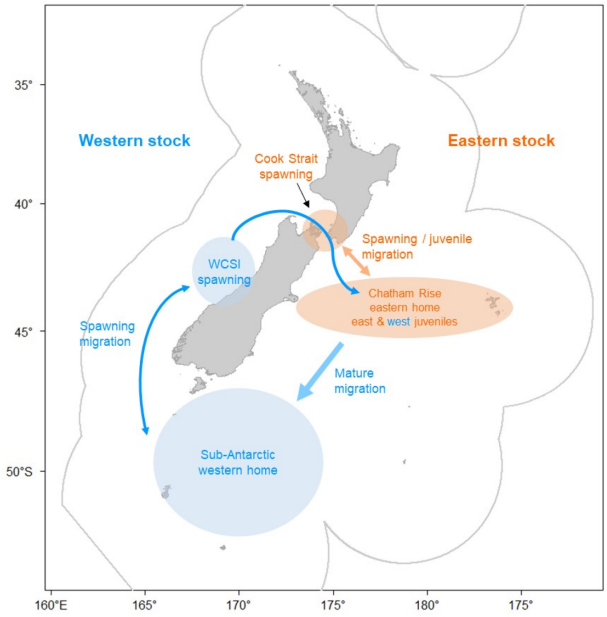


Figure 4.2: Hoki stock assumptions from the 2016 stock assessment. Source Fisheries New Zealand (2018)

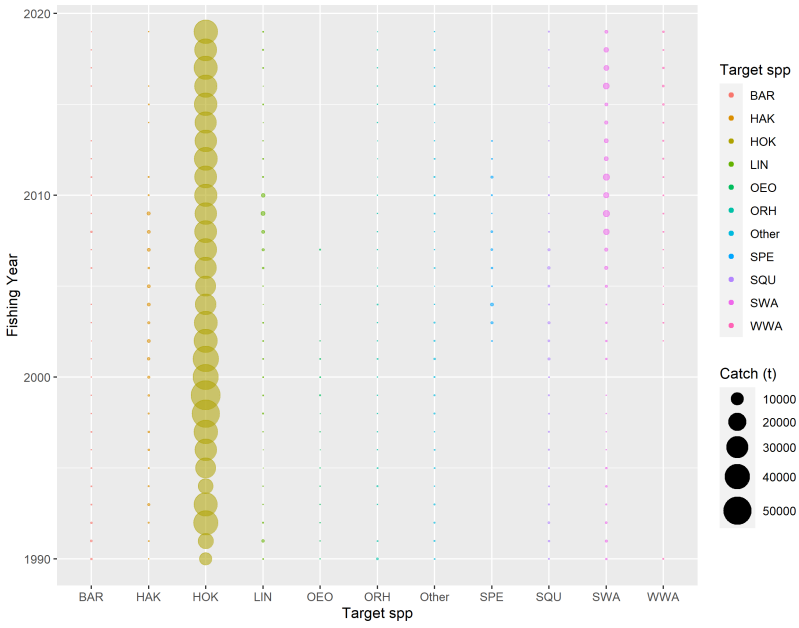


Figure 4.3: Total catch by target species (HOK = hoki, see Table C.1 in Appendix C for all species code names).

4.3 Fishery characteristics and data

The Chatham Rise hoki fishery ("the fishery") is predominately a bottom trawl fishery that operates on the slopes between 250-900 m depths (Figure 4.4). Data from the fishery sampling programme (Ballara & O'Driscoll 2020) and the scientific survey (Ballara et al. 2017), show hoki occupy different depth ranges at different sizes. Smaller fish are found in shallower waters (200-400 m) and large fish occupy deeper regions (400-950 m). This stratification of fish by size and depth coupled with vessels targeting larger fish (McKenzie 2017), is reflected in the spatial distribution of fishing shown in Figure 4.4.

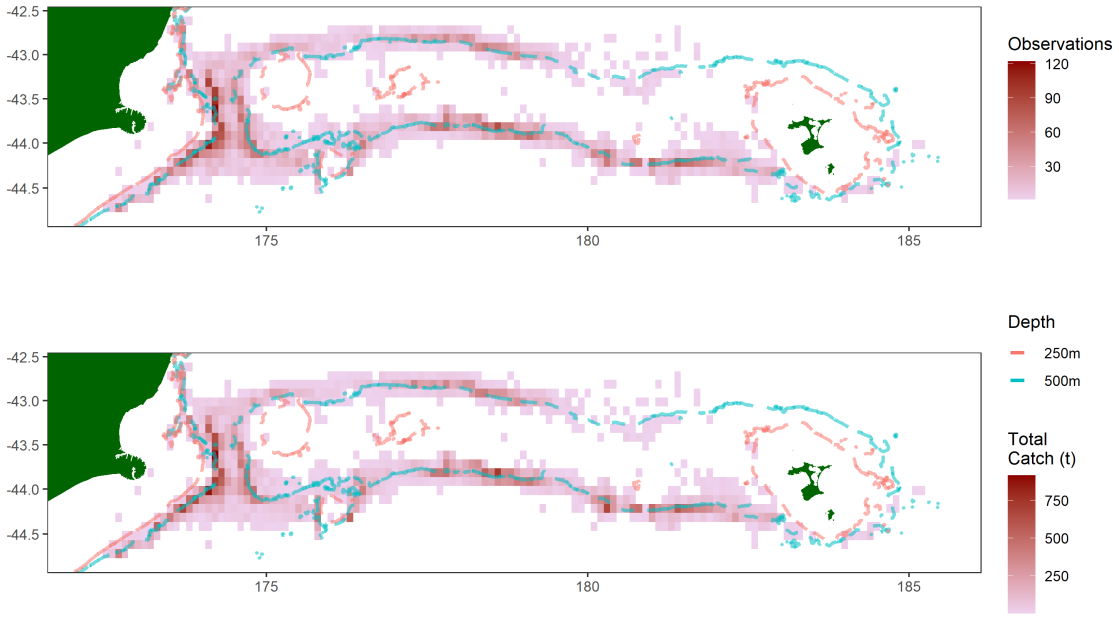


Figure 4.4: Spatial distribution of catch (t) and number of fishing events from the fishery, for the month of January between the fishing years of 2001 and 2018.

The fishery-dependent catch and effort data set used in this analysis was extracted from the Fisheries New Zealand Enterprise Data Warehouse as extract 12649B on 19 December 2019 and consisted of all fishing events associated with all fishing trips that reported a positive catch of hoki, hake (*Merluccius australis*), or ling (*Genypterus blacodes*) between fishing years 1989–90 to 2018–19 (Ballara & O'Driscoll 2020). A total of 2 397 127 fishing events were extracted of which 486 876 were on the Chatham Rise and 250 000 of these reported hoki in the catch.

Recorded fishery-dependent catch and effort data generally contain records that are missing values or have reporting and typographic errors such as fishing events on land or unrealistically large catches (Catch > 50 tonnes). These fishing events were identified in the data set and removed to avoid influence in the analysis. This is a process coined "grooming" in the CPUE standardisation literature (Bentley et al. 2012). Another consideration of fishery-dependent catch and effort data is they often contain a number of events which occur outside the stock habitat i.e., outside known depth range. In addition, some vessels may only operate in the fishery for short periods of time or sporadically over time. Information from these fishing events are unlikely to inform changes in relative abundance over the long term. A set of rules were applied (Table 4.1) to develop a "core" data set which represented consistent fishing behaviour over time, space and with respect to fishing method.

To avoid including fishing events with poor gear performance in the core data set i.e., fishing event abandoned due to gear failure. Fishing events had to be longer than an hour and catch a minimum of 500 kilos. Core fishing events also had to be less than 10 hours in duration. There has been concern in the past regarding long duration fishing events. These have been attributed to nets being left in the water after hauling (Ballara & O'Driscoll 2020), leading to non representative duration records.

In 2001 the fishing industry introduced a Code of Practice (COP) for hoki-targeted fishing events on the Chatham Rise with the intention of protecting regions with high density of small fish (less than 60 cm) (Ballara & O'Driscoll 2020), this was later replaced with the Deepwater Trawl hoki Operational Procedures (Group 2018). Both of these industry plans outlined spatial and temporal restrictions termed Hoki Management Areas (HMAs) (Figure 4.5 bottom panel). Due to these restrictions the core data set only considered fishing events from 2001.

In addition to sub setting by fishing year, the core data set only considered fishing events for the month of January. This was informed by timing of the scientific survey which also occurs in January (Stevens et al. 2021). January is assumed to be the month least effected by movement of fish from spawning migrations and from the Western fish migrating south to resident grounds (Livingston et al. 2004). This was to reduce variability between years in abundance signals associated with movement of fish in and out of the fishing ground.

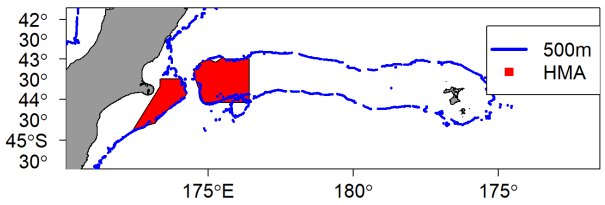


Figure 4.5: Hoki management areas (HMA).

Table 4.1: Rules that were applied to define a core data set and grooming. The effect on catch and number of fishing events removed during each rule are visualised in Figure C.3 in Appendix C.

- vessel needed to record at least 25 recorded fishing events
- fishing depth $\leq$ 950m
- recorded start latitude and longitude of fishing event
- gear method was Bottom Trawl (BT) and Precision Seafood Harvest (PSH)
- no NA's for variables of interest (Table 4.2)
- fishing years 2001-2018
- fishing events in the Month of January
- fishing event duration $>$ 1 hour and $<$ 10 hours
- 0.5 $<$ Catch $<$ 50

The data set used for model inference only contained fishing events that reported targeting hoki within the core data set (hoki-target dataset). The focus on hoki-target fishing events was based on the assumption that targeted fishing events were more likely to exhibit preferential sampling.

Table 4.2: Variables available for the hoki Chatham Rise fishery catch and effort core data set.

Label	Description	Variable type [1]	levels (range)
fyear	fishing year (Sep - Oct)	factor (x)	2001 - 2018
month	calender month	factor (x)	January
fweek	fishing week	factor (x)	14-18
start.time	start time of tow	continuous (x)	0 - 24
longitude	start longitude	continuous (x)	E 172.4° - E 185.1°
latitude	start latitude	continuous (x)	S44.5° - S 42.5°
observer.present	observer on board	bool (x)	0 or 1
method	gear method	factor (x)	BT, PSH
stat area	statistical area	factor (x)	16 levels
HOK.catch	hoki catch	continuous (y)	
area_swept	trawl footprint	continuous (offset)	0.1-4.1km^2
distance	distance	continuous (x)	4 -43 nm
duration	duration of fishing event	continuous (x)	1 -10 hours
effort.depth	gear depth	continuous (x)	157 - 950 m
spread	Trawl door spread	continuous (x)	11- 68 m
vessel.key	unique vessel identifier	factor (x)	27 levels

[1] x = covariate, y = response variable

An exploration of covariates and the response variable (Table 4.2) was conducted to identify any systematic changes over time, in addition to identifying multicollinearity. The exploration confirmed the covariate area was highly correlated with distance and duration (Figure 4.6). This was expected because area is a function of those two covariates. There was a change in fishing method in recent years with more hoki caught by the new modular precision seafood gear method (Fisheries New Zealand 2019) (Figure C.2) compared with conventional bottom trawl method. Plots and summary tables from the exploration are considered supplementary and supplied in Appendix C.

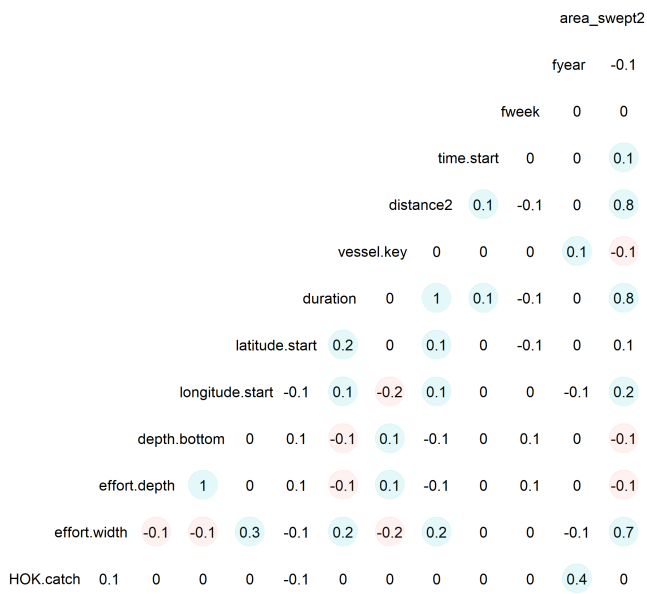


Figure 4.6: Pearson correlation coefficients for covariates and response variable. Blue circles indicate positive correlations, and red circles indicate negative correlations.

4.4 Variable selection using conventional models

The response variable was hoki weight per area fished (t/km^2). The gamma distribution with a log link was assumed due to the right skewed nature of the response variable (Figure 4.7) and its ease of implementation in TMB. The previous catch rate analysis by Ballara & O'Driscoll (2020) fitted to the natural log of catch and assumed normal distribution with identity link function. This approach was not applicable for geostatistical models because the linear predictor and thus abundance indices and standard errors would be in log-space due to the identity link. TMB does not have any base lognormal distribution functions (i.e., `dlnorm`, `rlnorm`, etc) and was not implemented due to time constraints.

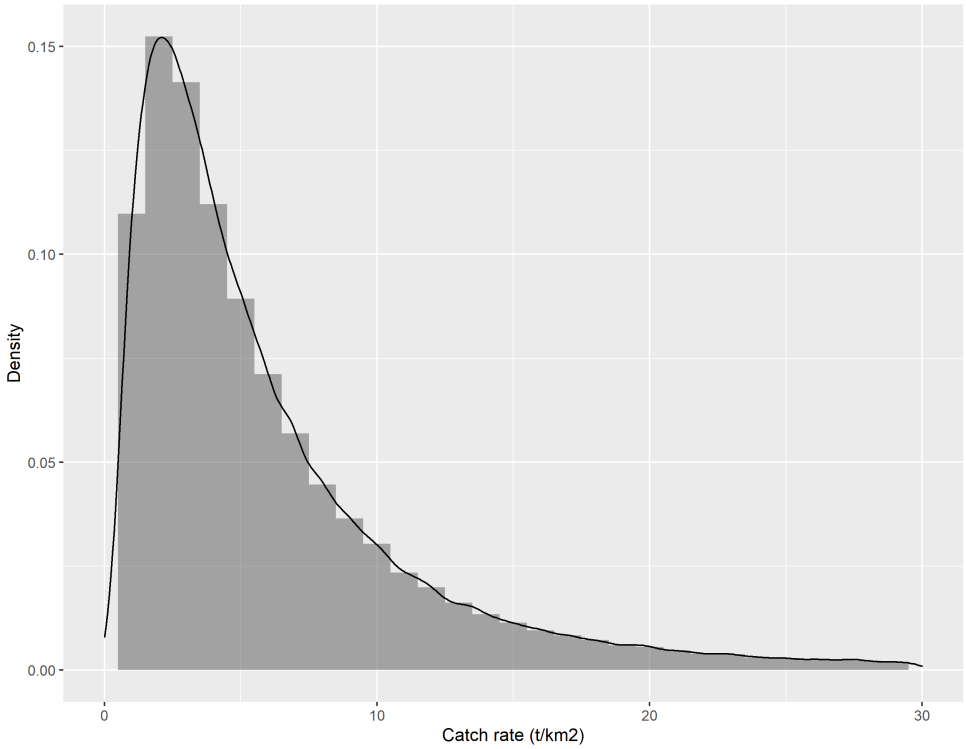


Figure 4.7: Density of catch rate (weight per area fished) from the hoki-target dataset.

Due to the number of covariates (Table 4.2) and large sample size, variable selection was conducted without GF or preferential sampling assumptions. It was too impractical to consider covariate selection in addition to different GF mesh assumptions.

A forward stepwise procedure using BIC for variable selection and percent deviance explained as a stopping criterion (Equation 4.1) was used to identify practically significant covariates. All covariates from Table 4.2 were offered during variable selection, where continuous variables were offered twice, firstly as a linear term and secondly as a thin-plate spline-based smoother (see Section 2.2 for details on spline-based smoother). The starting model for the stepwise procedure contained fishing year as the only covariate in the model. The algorithm began by iterating over all covariates testing them additively for practical significance. A practically significant covariate had the lowest BIC and had to explain at least 1% of deviance (Equation 4.1). If multiple covariates explained more than 1% of deviance, the covariate with the lowest BIC was selected. The algorithm repeated the previous step with the

remaining covariates until no additional covariate could explain at least 1% in $D_{\%}$ where,

$$D_{\%} = \left(1 - \frac{D_{residual}}{D_{null}}\right) 100 \ , \tag{4.1}$$

and D_{null} is an intercept only model. The stepwise procedure resulted in the model formula defined in Code block 4.1 with summary statistics supplied in Table 4.3.

```
y ~ fyear + s(distance, bs = "ts")+ s(spread, bs = "ts")
      + s(time.start, bs = "ts") + vessel.key,
      family = Gamma(link = "log"), offset = log(area))
```

Code Block 4.1: R code for the model selected from the stepwise procedure.

Table 4.3: Percent deviance explained, BIC and effective degrees of freedom for each additional term added from the stepwise procedure.

Term	$D_{\%}$	BIC	Effective degrees of freedom
$fyear$	21.3	67192.7	18
$+ s(distance)$	38.3	64142.4	26.1
$+ s(spread)$	44.6	62842.8	34.9
$+ s(time.start)$	47.2	62324.9	42.5
$+ vessel.key$	49.1	62091.3	67.6

The first two selected terms were distance and door spread. The estimated effect for both covariates (Figure 4.8) was similar, with decreasing catch rates for longer and wider fishing events. These estimated effects could be attributed to many factors for example, the spread effect could have a herding effect (Winger et al. 2004). This is where the doors resuspend sediment which can increase the "effective" spread of narrow trawls. The distance effect could also be the consequence of different fishing tactics, where short tows may be targeting clear aggregations. In contrast, longer tows may not be fishing obvious aggregations (often distinguished by acoustic back scatter) and so tend to have lower catch rates. Unfortunately, information required to disentangle these possible factors are not available in the data set.

The estimated effect for time of day was consistent with survey observations (Stevens et al. 2021, Livingston et al. 2004) where catch rates increased during daylight hours as mid-water prey move deeper in the water column. This is assumed to lead hoki closer to the bottom floor which increases their availability to bottom contact fishing methods.

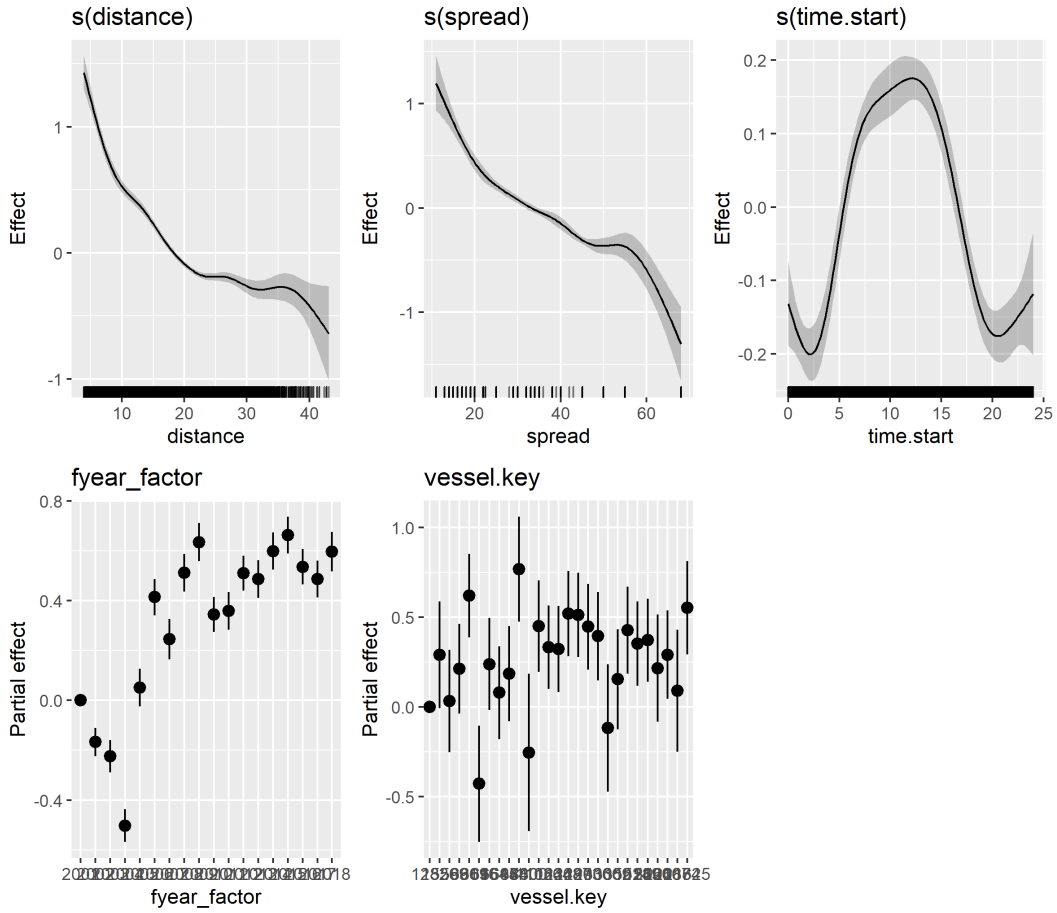


Figure 4.8: Estimated covariate effects from variable selection with conventional model.

4.5 Geostatistical and PS models

The model defined in Code Block 4.1 was extended to include GF and preferential sampling assumptions outlined in Table 4.4. This required defining a mesh (see Section 2.3) using the universal transverse mercator projection coordinate system. Due to New Zealand's location in the southern latitudes, working with latitude and longitude are an inconvenience for distance based analyses. In addition to defining a mesh, the spatial domain was partitioned into equal area grid for spatial extrapolation. A range of mesh assumptions were explored with the final mesh presented in Figure 4.9. This assumed linear basis functions between vertices and was the finest

spatial resolution achieved whilst maintaining a reasonable model run time (under an hour to converge).

GF assumptions followed the model

$$\boldsymbol{\omega}|\boldsymbol{\theta_\omega} \sim GF(\mathbf{0}, \Sigma_\omega) \ ,$$
$$\boldsymbol{\epsilon_t}|\boldsymbol{\theta_\epsilon} \sim GF(\mathbf{0}, \Sigma_\epsilon) \ ,$$

where, Σ_ϵ and Σ_ω were assumed isotropic Matérn covariances (Section 2.3).

Conventional geostatistical models presented in earlier sections (Section 2.3) assumed sample locations were ancillary when estimating spatial abundance and related parameters. In contrast, the PS model assumes sample locations contain information on spatial abundance. For model comparisons, this chapter assumed conventional geostatistical models had sample locations drawn from a homogenous point process. This is equivalent to assuming the PS model with $\beta^{pref} = 0$ (Equation 3.4). Assuming a homogenous point process for sample locations is analogous to treating sample locations as ancillary because both assumptions state there is no information in sample locations when estimating spatial abundance and related parameters. Regarding a conventional geostatistical model as one that generates locations according to a homogenous point process allows it to be nested within the PS model. This makes the log likelihoods comparable.

Table 4.4: Geostatistical and preferential models applied to the hoki Chatham Rise trawl fishery.

Model label	Systematic component	Intensity function (PS)
$\mathcal{M}_0$	$\boldsymbol{X_i\beta}$ = Code Block 4.1	-
$\mathcal{M}_1$	$\boldsymbol{X_i\beta} + \omega(s_i)$	$\lambda(s_i) =$ constant
$\mathcal{M}_2$	$\boldsymbol{X_i\beta} + \epsilon(s_i, t_i)$	$\lambda(s_i) =$ constant
$\mathcal{M}_3$	$\boldsymbol{X_i\beta} + \omega(s_i)$	$\lambda(s_i) = \exp\{\beta^{pref}\,(\omega(s_i))\}$
$\mathcal{M}_4$	$\boldsymbol{X_i\beta} + \epsilon(s_i, t_i)$	$\lambda(s_i, t_i) = \exp\{\beta^{pref}\,(\epsilon(s_i, t_i))\}$

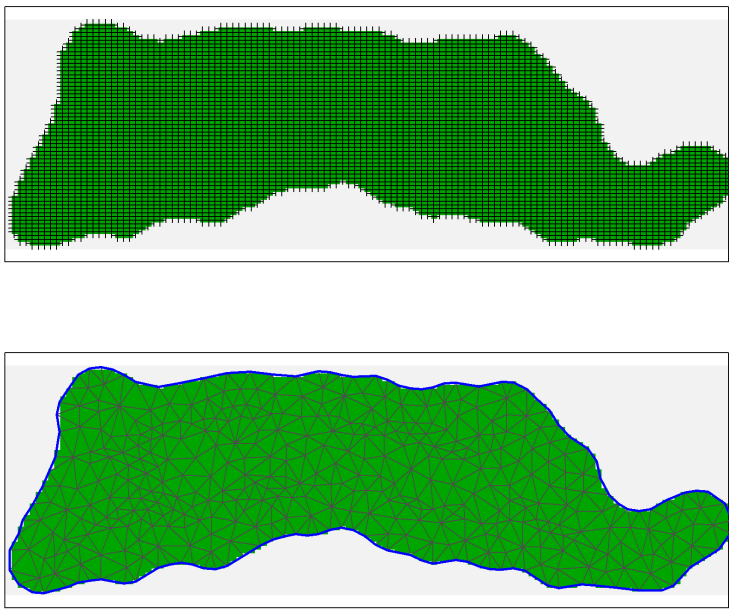


Figure 4.9: Spatial domain (green) with extrapolation grid (top panel) compared with mesh resolution (bottom panel).

4.6 Goodness of fit and model comparison methods

The metrics used to evaluate goodness of fit and conduct model comparison were randomised quantile residuals (Section 2.4), residual deviance, K-folds cross validation and prediction error using a bycatch data set.

Randomised quantile residuals and residual deviance were calculated for the catch rate observations. The K-folds cross validation tests a model's out-of-sample predictions by withholding a portion of the training data during inference for testing. These approaches are all based on the core data set to evaluate a models goodness of fit and prediction error. Two additional data sets that were considered for evaluating prediction error were the research survey data, and fishing events that caught hoki from the core data set that were not hoki-target events.

The ideal testing data set would have been the research survey data due to its randomised spatial design over the domain and standardised sampling methods, i.e.,

uses the same vessel and gear over the entire thirty year time-series. The randomised spatial design of the research survey was thought to be advantageous for evaluating prediction error of models in regions that are not fished by the fishery. This data set was not explored due to time-constraints and the need to account for factors such as, the research vessel using a different mesh compared to the fishery (contact selectivity) and the survey sampling areas that contain high densities of young fish (spatial availability of exploitable population). These areas with high densities of young fish do not contain hoki-target fishing events (Group 2018). Catch rates from the survey can be quite large in these regions but are not considered part of the target population of the fishery, which the hoki-target models are predicting.

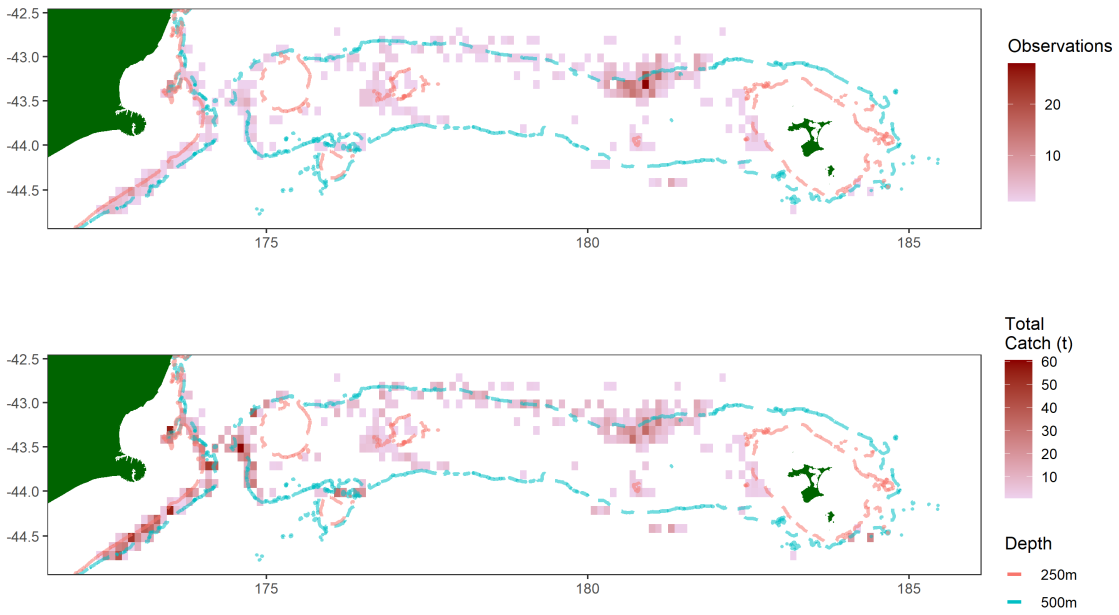


Figure 4.10: Spatial distribution of the bycatch data set.

The other testing data set considered, consisted of fishing events from the core data set that caught hoki but were not hoki-target events termed the bycatch data set. This data set contained the same vessels and fishing methods that are in the hoki-target data set. The downside of this testing dataset is it will likely have operational and behaviour differences to that of the hoki-target data i.e., skippers may configure or deploy trawl gear to optimise the non-hoki target species. Although there are pitfalls in this data it was used to compare model prediction error because

it contained sample locations at the edges of the hoki-target data set but were still in exploited hoki habitat (Figure 4.10). An expected benefit of using the PS model is its ability to predict into unsampled regions assuming preferential sampling is occurring. This data set was thought to be useful for exploring this model behaviour.

The K-folds cross validation split the hoki-target data set into ten folds, where each fold consisted of 10% of the data randomly removed within each year. It is best practice to remove spatial blocks of data for cross validating spatio-temporal models (Roberts et al. 2017). This was not appropriate here due to the PS model interpreting unsampled space as low values leading to poor predictions. For each fold two metrics were calculated, the root mean square error (RMSE, Equation 4.2) for the hold out catch rates and the joint likelihood which was evaluated for both catch rate and location likelihood components (Equation 4.3). RMSE calculated for catch rates was conditional on locations being sampled. Final model comparisons (Table 4.5) were done using median values for RMSE and the predicted joint likelihood across all ten folds. RMSE for the k^{th} fold was calculated as

$$RMSE_k = \sqrt{\frac{\sum_{j=1}^{n_k} (y_j - \widehat{y}_j)^2}{n_k}} \ , \tag{4.2}$$

where, n_k is the number of samples in the test data set for fold k. The predicted joint likelihood for the k^{th} fold denoted by ll_k was

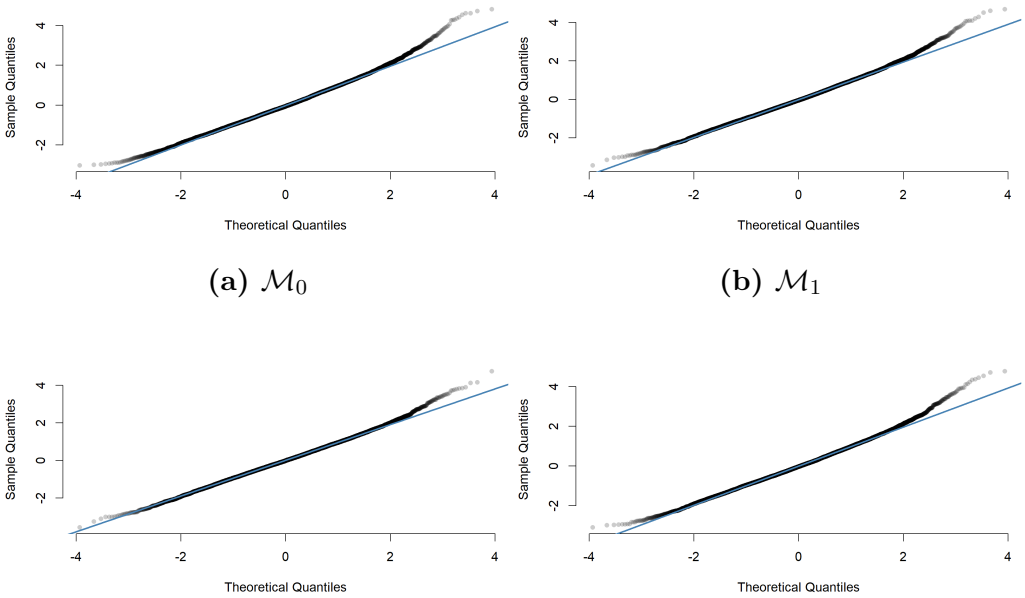
$$ll_k = \sum_{j=1}^{n_k} \log \ f\left(y_j|\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{u}}\right) \log \ g\left(s_j|\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{u}}\right) \ , \tag{4.3}$$

where, $f(.)$ is the catch rate density, $g(.)$ is the spatial point process density (Equation 3.2), $\widehat{\boldsymbol{\theta}}$ are estimated parameters and $\widehat{\boldsymbol{u}}$ are empirical Bayes estimates (Equation 2.15).

4.7 Results

Model $\mathcal{M}_3$ encountered convergence issues. It produced large parameter gradients and a non-positive definite hessian matrix (Section 2.4). To investigate this further additional log-likelihood profiles were conducted for the β^{pref} parameter for both $\mathcal{M}_3$ and $\mathcal{M}_4$. Due to the convergence issue $\mathcal{M}_3$ was not considered in model comparisons.

Residual patterns for all models based on QQ plots of randomised quantile residuals exhibited heavy tails (Figure 4.11). This suggested all models struggled to fit large catch rates and to a lesser extent small catch rates. Apart from a slight deviation in the tails of the QQ plots it was concluded that all the models were providing a reasonable fit to the catch rate observations.



(a) $\mathcal{M}_0$ **(b)** $\mathcal{M}_1$

(c) $\mathcal{M}_2$ **(d)** $\mathcal{M}_4$

Figure 4.11: QQ plots of randomised quantile residuals.

Table 4.5 outlines summaries of the model comparisons and shows that residual deviance based on catch rates and RMSE of catch rates from the K-folds cross validation both favoured model $\mathcal{M}_2$. This was also supported by a slightly better pattern in residuals (Figure 4.11). Prediction error based on the bycatch dataset favoured $\mathcal{M}_1$. Although this testing data set may have pitfalls due to potential operational differences, this result did show the time invariant GF assumption (model $\mathcal{M}_1$) may have better out-of-sample prediction skill than the time-varying GF model ($\mathcal{M}_2$) and could hint at over-fitting in model $\mathcal{M}_2$. The predictive joint log-likelihood which measured the fit to both fishing location and catch rates favoured the PS model $\mathcal{M}_4$. This was due to the conventional geostatistical models inability to account for spatial heterogeneity in sample locations, due to the homogenous point process assumption.

Table 4.5: Summary statistics from model comparisons.

Model label	Residual deviance	Median RMSE (K-folds)	Median ll_k (K-folds)	RMSE (bycatch dataset)
$\mathcal{M}_0$	4467.8	4.15	-	4.49
$\mathcal{M}_1$	4158.3	4.08	-18242.82	3.76
$\mathcal{M}_2$	3550.2	4.01	-18216.92	4.23
$\mathcal{M}_3$[1]	–	–	–	–
$\mathcal{M}_4$	4209.8	4.10	-15691.99	4.02

[1] Model was deemed not to converge based on likelihood profiles of β^{pref} in Figure 4.12

Table 4.6: Estimated parameters from GF.

Parameter	$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$
Marginal variance σ_M	-	0.29	0.26	–	0.4
Range	–	2.77	2.74	–	2.76
β^{pref}	–	–	–	–	8.2

The PS model $\mathcal{M}_4$ estimated a value of 8.2 for β^{pref} (Table 4.6). This was a much larger value than was explored in previous simulations (Section 3.5.2). Such a large value indicated extreme preferential sampling and raised concerns regarding identifiability. In particular, whether the estimated value was based on a local minimum rather than a global one. To address this concern, a log-likelihood profile was conducted (Figure 4.12). This indicated a global minimum was found for model $\mathcal{M}_4$ but $\mathcal{M}_3$ illustrated no clear minimum with the profile sloping towards the upper bound.

A comparison of estimated terms among the EMs including estimated spline-based smoothers and fixed effect were similar among are supplied in Appendix C. In general, the estimated effects were similar to those estimated and presented during variable selection in Figure 4.8.

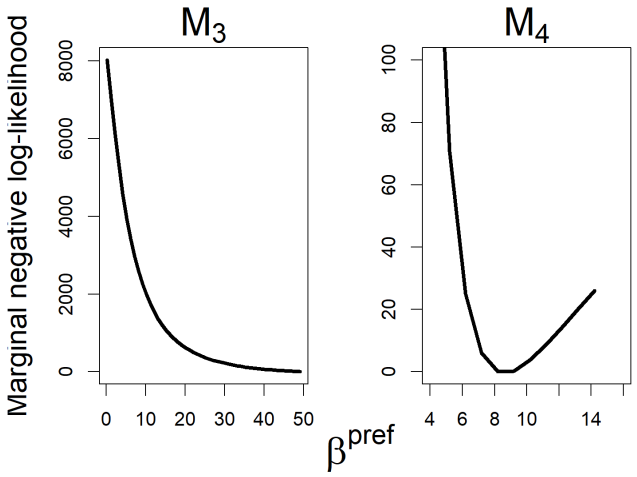


Figure 4.12: Marginal negative log-likelihood profile (scaled to have minimum at zero) for β^{pref} for both PS models.

The large estimated value of β^{pref} from model $\mathcal{M}_4$ resulted in the point process being very influential on the estimated spatial GF. This is illustrated in Figure 4.13 and Figure 4.14 (bottom right panel), where the PS model $\mathcal{M}_4$ estimated high spatial abundance in regions that were sampled and very low values else where. The influence of the point process also resulted in variable estimates of spatial abundance between years that may be the result of slightly different sampling patterns between years. In 2012 the North West and South West regions were not sampled and the PS model estimated very low spatial abundance in these regions. Compare that with 2001 which had fishing events in these locations and the PS model estimated levels of high abundance.

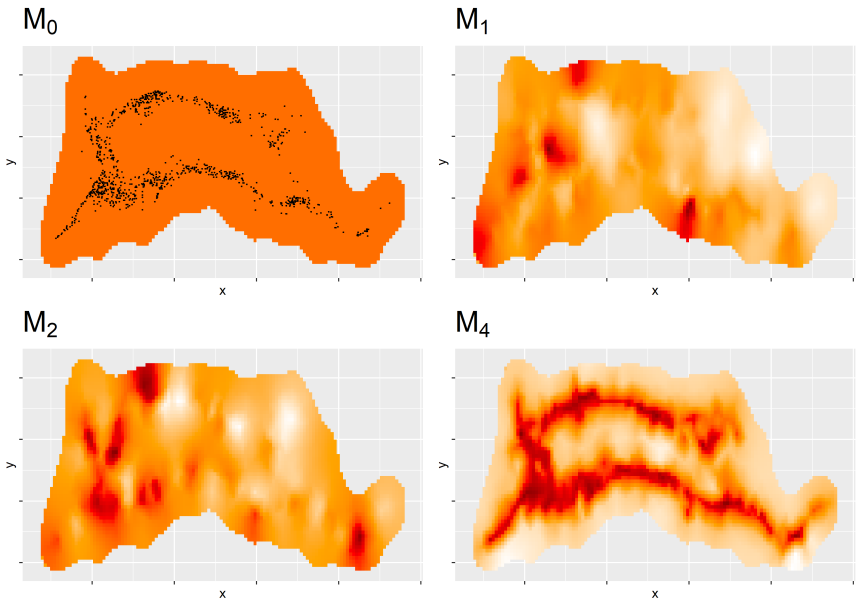


Figure 4.13: Spatial estimates of abundance for the year 2001. Black dots in the top left panel indicate fishing locations from the hoki-target data set for this year. Cells with dark red indicate large values with pale colours indicating low values.

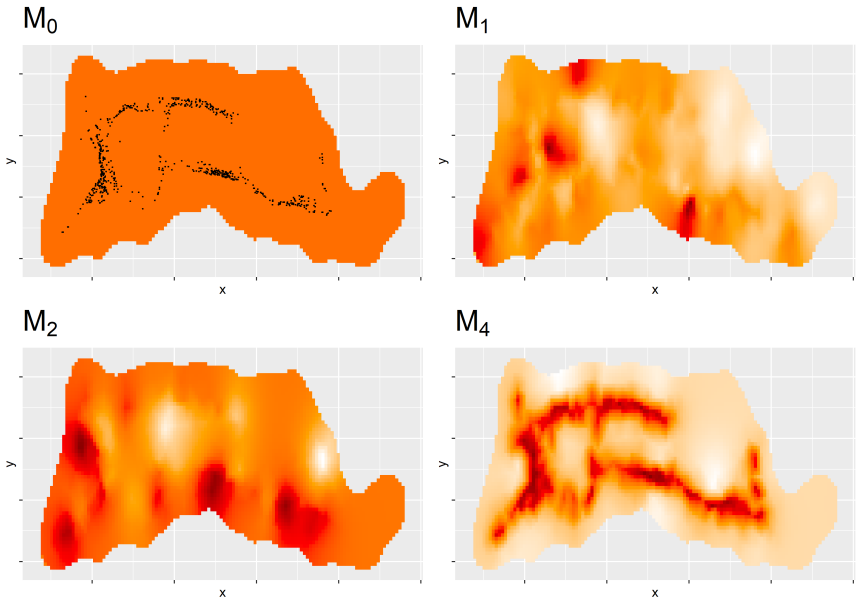


Figure 4.14: Estimates of spatial abundance for the year 2012. Black dots in the top left panel indicate fishing locations from the hoki-target data set for this year. Cells with dark red indicate large values with pale colours indicating low values.

Figure 4.13 and Figure 4.14 illustrate how different the estimated spatial distributions was between models. Estimated spatial distributions for all years and models are supplied in Appendix C but were similar to the patterns shown in 2001 and 2012 years.

Although the models estimated quite different spatial abundance patterns, they estimated similar indices of biomass over time (Figure 4.15). Given the unit of the response variable was in weight and not numbers, the estimated index tracked changes in biomass over time rather than abundance. An index of biomass estimated from the scientific survey (Stevens et al. 2021) was added and showed very similar trends to those estimated from the models explored in this chapter. Although the fishery grounds are within the survey area, there are regions of the survey that are not fished due to the COP (Group 2018). This meant the survey index and fishery-dependent indices are based on different age compositions of hoki. However, because larger fish contribute more to biomass (i.e., the relationship between fish length and weight is cubic) the relative biomass monitored by the two data sets was in agreement.

Figure 4.15: Estimated indices of relative biomass between fishery-dependent models explored here and the independent scientific survey.

Estimated spatial intensity for all years from model $\mathcal{M}_4$ are shown in Figure 4.16. This figure highlights how informative the location dataset was on estimated spatial distributions. Areas that had many fishing events had high estimated intensity values and unsampled areas estimated low intensity values which was expected due to the PS model assumptions.

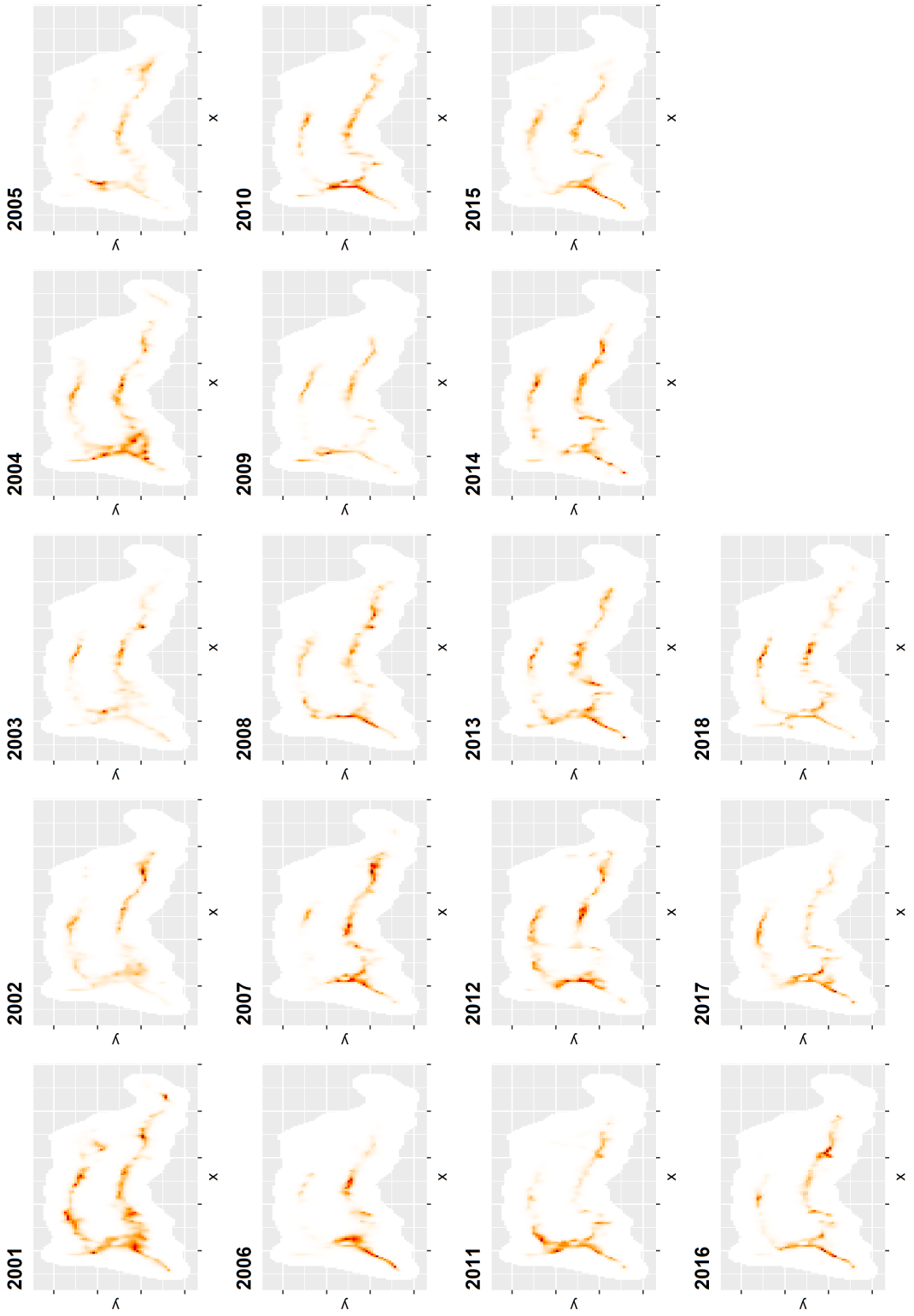


Figure 4.16: Estimated intensity spatial distribution ($\widehat{\lambda}(s_i)$) from model $\mathcal{M}_4$.

The correlation metric described in Section 3.6 was applied to models $\mathcal{M}_1$ and $\mathcal{M}_2$ (Figure 4.17). This exhibited large inter-annual variability with values ranging from 0.0 to 0.4. Based on simulations from the previous section (Section 3.6) this implied there could be other factors contributing to fishing sampling locations and may be a reason for identifiability issues with β^{pref} in model $\mathcal{M}_3$.

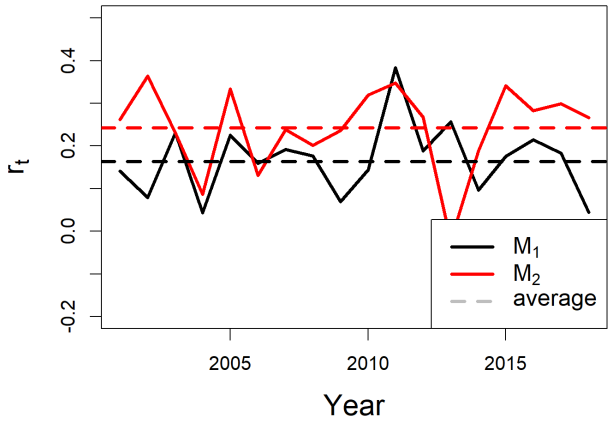


Figure 4.17: Correlation metric (r_t) (Section 3.6) applied to $\mathcal{M}_1$ and $\mathcal{M}_2$.

4.8 Discussion

This chapter applied a geostatistical model that accounted for preferentially sampled observations (the PS model) to data from a major New Zealand trawl fishery. The purpose of this analysis was to identify if the PS model could be fit to real fishery-dependent catch and effort data. Of the two PS models applied, the one that contained a spatio-temporal GF ($\mathcal{M}_4$) successfully converged, whereas the PS model with a time-invariant GF ($\mathcal{M}_3$) did not. The convergence issue for $\mathcal{M}_3$ was due to β^{pref} being estimated at boundary constraints. This indicated the complexity and variability of real data required a flexible PS model.

Comparisons between estimated indices of relative biomass, residuals, and out-of-sample predictions were made between the PS model (that did converge) and conventional geostatistical models. The conventional geostatistical model with spatio-temporal GFs ($\mathcal{M}_2$) had the lowest residual deviance and prediction error from the

K-folds cross validation. The conventional geostatistical model with time-invariant GF ($\mathcal{M}_1$) had the lowest prediction error based on the bycatch dataset. However, the PS model was favoured by the predictive joint log-likelihood metric. This highlighted the difficulty in conducting model comparisons among conventional geostatistical models and the PS model.

Spatial estimates of abundance from the PS model that converged, illustrated that the PS model did behaviour as expected. That is, regions with a high frequency of fishing events was estimated with high abundance and areas with no or a low frequency of fishing events, were estimated as regions of low abundance (Figure 4.13). Comparisons among estimated spatial distributions of abundance show there was considerable difference among the EMs, but ultimately this did not result in different trends in relative biomass over time (Figure 4.15). The estimated indices of relative abundance from all converged models were compared to an index from the research survey, which corroborated the estimated trends from the CPUE standardisation.

Although the study by Rufener et al. (2021) focused on species distribution maps. They also did not identify significant differences in estimated parameters when they incorporating a preferential sampling model for the fishery-dependent data. This result was in an integrated analysis using both fishery-independent data and fishery-dependent data for species distribution maps. Their approach of integrating fishery-independent and fishery-dependent data should be considered to further validate the PS model.

When comparing the PS model with conventional geostatistical models using catch rate residuals and residual deviance (Table 4.5 and Figure 4.11). The PS model did not perform as well we had initially anticipated, which in hindsight, was not surprising. The PS model assumed the same spatial covariates and GFs when describing spatial distributions of both catch rates and fishing locations. For complex data generating systems such as commercial fishing, it is more likely this assumption will be violated for some fishing events due to fishers considering additional factors when choosing locations to fish such as; weather conditions, other vessels fishing in the area, market conditions, previous catch rates etc. These factors will introduce a conflict between the location data set and catch rate data set resulting in a compromised fit. In order for a fishery to satisfy the PS assumption exactly, skippers would need perfect knowledge of the spatial distribution of exploitable fish and fish it relative to the density. Future extensions of the PS model should consider spatial covariates

or GFs that are not shared in the linear predictor for catch rates and fishing locations, as explored in Diggle et al. (2013), Conn et al. (2017). This would relax this assumption of the PS model and perhaps lead to better fits.

The predictive joint log-likelihood was explored as an alternative metric for model comparison during K-folds cross validation. This measured a models fit to both catch rate and sample location data sets. The issue with our exploration of this metric was in order to apply it, we had to assume a point process for the conventional geostatistical models. We made the assumption that sample locations were drawn from a homogenous point process as this is an analogous assumption to treating them as ancillary. However, the sample locations are clearly not generated from a homogenous point process (Figure 4.4). This led to the predictive joint log-likelihood favouring the PS model which had the ability to describe variability in observed locations. This comparison of the predictive joint log-likelihood would be more meaningful if, as advocated in the previous paragraph, additional spatial covariates or GFs were included in both the catch rate and point process models that were not shared.

This application along with other PS model studies (Dinsdale 2018, Conn et al. 2017) explored prediction error in catch rates as a method to evaluate PS model performance to conventional geostatistical models. This application highlighted a possible limitation in using prediction error for this purpose. When preferentially sampled data is randomly split into training and testing data sets (i.e., K-folds cross validation). Testing data sets will naturally contain a high frequency of observations in areas that are preferentially sampled (spatially imbalanced), and thus they will contribute more to summary statistics such as the root mean squared error than less frequent observations in poorly sampled regions. These less frequent observations in poorly sampled regions are where the PS model is expected to outperform conventional models (assuming the PS assumption is satisfied). This strength of the PS model will not be "valued" due to the abundance of observations in highly fished locations, which the conventional geostatistical models are expected to be as good if not better because they do not have to trade off fits to catch rates due to the location data set. An alternative is to split data in a more structured way between training and testing datasets i.e., spatial blocking (Roberts et al. 2017). However, as mentioned earlier, the PS model assumptions can make this impractical. Another consideration is to explore weighted prediction calculations that can account for the spatial imbalance of testing data sets.

The application of the PS model to this fishery highlighted how influential the point process was on the estimated spatial distribution of abundance relative to the catch rate data. We did not consider adjusting the influence or "weight" the likelihood of the point process in the PS model explored. Future research should explore alternative distributional assumptions regarding the point process and how much influence or "weight" it should have relative to the catch rate data. This is an analogous problem to the data weighting topic with respect to integrated stock assessment models (Francis 2011, Maunder & Punt 2013).

In addition to looking at different likelihoods for the point process, the lognormal should also be explored as an alternative to the catch rate response variable. This may address the heavy tails exhibited in the residuals, but is unlikely to change the general conclusions. Another addition for the catch rate component would be to explore cyclic splines for cyclic covariates. Cyclic covariates have minimum and maximum values that join, i.e time of day.

Even though this analyses was applied to a subset of the original data extract, it still was close to reaching computational limits. The inclusion of the point process on top of the Gaussian Field calculations required a lot of computer memory (RAM) and CPU usage. We can envision limitations in applying the PS model to much larger data sets, unless there is a reduction in spatial resolution of the model, which may be a limitation. Future work should consider computational efficient alternative point processes such as in Simpson et al. (2016).

Log-likelihood profiles indicated identifiability issues for the β^{pref} in model $\mathcal{M}_3$ which was estimated at the upper bound. This has been observed in other applications of PS models (Conn et al. 2017). To understand assumptions and data properties that led to poor estimation of β^{pref} in the model that failed to converge, addition simulations were conducted in Chapter 6. This motivated the creation of an innovative agent-based model program described in the following chapter.

Chapter 5

A generalised agent-based operating model

5.1 Overview

This chapter describes the innovative generalised agent-based model (ABM), CABM (C++ Agent Based Model). CABM is a C++ program that was created for the simulation study based on the Chatham Rise hoki fishery in Chapter 6. Section 5.2 describes the building blocks of ABMs, benefits of ABM operating models and previous usage of ABMs in the fisheries literature. Section 5.3 combines these building blocks to create CABM, a flexible and powerful general purpose ABM program.

5.2 Introduction

An ABM is a model that represents a fish stock as a collection of agents. An agent is defined as one or more fish with homogenous characteristics, i.e. length, weight and sex. When an agent represents a single individual, the ABM becomes an individual-based model (IBM) (Grimm & Railsback 2013). Given fish stocks consist of millions if not billions of individuals, ABMs are often more practical than IBMs due to computational limitations, i.e. it requires large amounts of memory to record and modify millions of agents. ABMs use functions to grow, move, create and kill agents over time, termed agent dynamics. When summaries are made over all agents, stock level quantities are observed. Simulating stocks with this high level of detail allows

heterogeneity in key dynamics such as growth and mortality. This is an advantage of ABMs as this heterogeneity is often approximated in other operating models (OM).

A common type of OM used in stock assessment literature is a "cohort-based" OM (Punt 2003, Lee et al. 2011, Deroba et al. 2014, Sippel et al. 2017). These partition fish stocks into cohorts with shared characteristics, i.e. age, stock, region, length and combinations of these. Cohorts are managed inside the "cohort-based" OM as elements of an array, termed the partition. Productivity parameters relate to elements of the partition. As fish move between elements of the partition, they take on productivity parameters of the destination element without regard to their previous productivity parameters, i.e. if a fish moves from a low growth area (low mean length at age) to a high growth area (high mean length at age) it will immediately take on the larger length at age even if it was a small sized fish previously.

In contrast, ABMs link productivity assumptions to individual agents. As agents change attributes such as age, length or region, they retain their history of productivity parameters, i.e. if an agent moves from a low growth area to a high growth area, it will retain its length from its time in the low growth area and take on high growth rates in the future. This example demonstrates one of the key benefits of ABMs compared with "cohort-based" OMs, that is they can emulate more realistic dynamics. ABMs are also naturally suited to describe a range of stock assessment processes and observations that are inherently individual-based, such as tag release and recapture events.

It is common for simulation studies to have estimation models as operating models i.e., Lee et al. (2011, 2012), Stewart & Monnahan (2017) and in Chapter 3. Conclusions from these simulations are tied to strict caveats such as "assuming all other model assumptions are correctly specified . . . ". Due to complexities in the real world, conclusions based on such simulations have been brought into question (Francis 2012). This criticism was a motivator for the development of CABM. Simulations using OM that emulate realistic biology and complexities, will yield results that are more general. This is a strength of using CABM as an OM.

ABMs have been used to investigate stock assessment methods in the past (Webber 2015, Cao et al. 2016, Phillips et al. 2018) but were almost always bespoke due to the difficulties in generalising software code for a wide range of problems. These ABMs are difficult to apply to problems outside their intended bespoke design. An objective of this thesis was to build a generalised ABM that could easily be applied

to a range of applications, including its use for research questions that fall outside the scope of this chapter. This objective is currently being achieved with its use as an OM to explore key uncertainties in stock assessments for snapper (*Pagrus auratus*) and kahawhai (*Arripis trutta*) in New Zealand (McKenzie et al. 2021).

CABM was created to emulate a stock over a fine spatial resolution domain, apply realistic movement (Section 5.3.1.4) and replicate complex fishing processes (Section 5.3.1.5). These were all attributes of an OM that were required to simulate fishery-dependent catch and effort data to further explore properties of the PS model (Chapter 6).

CABM was coded in C++ for execution speed and encompasses best coding practices, including modular code design and unit testing. The codebase is publicly available on GitHub (`https://github.com/Craig44/CABM`). CABM has generalised model structures including spatial resolution and flexible timing of agent dynamics within a year. These two attributes make CABM unique to other ABMs reviewed (Webber 2015, Cao et al. 2016).

The following sections describe CABM's core dynamics that were used in the Chatham Rise simulation study (Chapter 6). However, CABM is a generalised program that can be used to emulate a range of life-histories and spatial structures that are not described in the following sections. For a more comprehensive description of CABM and all its functionality, we recommend users see the user manual (Marsh 2022).

5.3 The CABM model

CABM expresses a fish stock as a collection of spatially referenced agents within a discrete spatial domain with R cells. Agents in spatial cell indexed by r ($r = (1, \ldots, R)$) are denoted by $\boldsymbol{X_r} = \{X_{r,1}, \ldots, X_{r,n_r}\}$, where n_r is the number of agents in cell r. Agent $X_{r,i}$ is also a set where each element in $X_{r,i}$ is an agent attribute, as outlined in Table 5.1.

CABM applies an annual cycle each year which is then repeated over a specified number of years. The annual cycle is made up of discrete time blocks hereinafter referred to as time-steps. Each time-step applies a sequence of agent dynamics (Section 5.3.1), outputs summaries of agent information and simulates synthetic data. Agent dynamics use agent specific attributes (Table 5.1) along with other assumed

parameter values to define interactions and outcomes for agents over time. CABM assumes a closed spatial domain, meaning, no agents can leave the spatial extent (i.e. no immigration or emigration).

Table 5.1: Attributes recorded for all agents, where $X_{r,i} = \{a_i, l_i, \ldots, x_i, y_i\}$.

Index	Description (variable type)
a_i	age (integer)
l_i	length (continuous)
lb_i	length bin (integer)
mat_i	maturity (boolean)
s_i	sex (boolean, 0 = male, 1 = female)
w_i	weight (continuous)
$n_i = \tilde{n}$	multiplier that represents how many individual fish per agent (continuous). See Section 5.3.2 for the calculation of $\tilde{n}$
M_i	natural mortality rate (continuous)
L_i^{inf}	growth parameters (continuous)
r_i^b	natal cell (birth cell)
r_i	current cell
x_i, y_i	coordinates over the domain (continuous)

5.3.1 Agent Dynamics

Agent dynamics are functions responsible for modifying agents (growth), moving agents, creating agents (recruitment) and deleting agents (mortality). The inputs for these functions use both agent specific attributes (Table 5.1) and assumed parameter values. Agent dynamics are predominately stochastic, where agent actions are based on a randomly generated realisation from a random variable. For example, a dynamic that results in a binary outcome for an agent (i.e. an agent getting caught by a fishing interaction or an agent maturing) can be expressed by the Bernoulli random variable $I \sim \mathcal{B}ern(p)$, where p is the probability of an event occurring

$$I \begin{cases} 1, & \text{event occurs} \\ 0, & \text{event does not occur} \end{cases}.$$

Agent dynamics generally iterate over a collection of agents and for each agent, they apply an outcome based on a realisation from a random variable. The following example demonstrates the notation used when describing how an agent dynamic

applies an event to all agents in cell r,

$$\begin{aligned} p_i &= f(X_{r,i}, \boldsymbol{\theta}) \ , \ \forall \ X_{r,i} \ \in \boldsymbol{X_r} \ , \\ I_i &\sim \mathcal{B}ern(p_i) \\ I_i &\begin{cases} 1, & \text{event occurs for the } i^{th} \text{ agent} \\ 0, & \text{event does not occur for the } i^{th} \text{ agent} \end{cases} , \end{aligned}$$

where p_i denotes the i^{th} agents ($X_{r,i}$) specific probability of the event occurring. This is the result of the function $f(.)$ which uses agent attributes and parameters $\boldsymbol{\theta}$.

The agent dynamics described in the following sections define p_i as a function of specific agent attributes (Table 5.1). In all descriptions $X_{r,i}$ in $f(X_{r,i}, \boldsymbol{\theta})$ is replaced with the specific agent attribute. For example, if p_i was a function of an agent's length (l_i), it would be defined as,

$$p_i \ = f(l_i, \boldsymbol{\theta}) \ , \ \forall \ X_{r,i} \ \in \boldsymbol{X_r} \ .$$

5.3.1.1 Growth

Growth is the agent dynamic responsible for changing an agent's length and weight over time. Length and weight are commonly used as inputs to other agent dynamics or used when calculating stock level quantities, i.e. spawning stock biomass. When an agent is created via the recruitment dynamic (Section 5.3.1.3), it is assigned a growth parameter from a population level distribution. There are two growth models available in CABM, the Von Bertalanffy and generalised Schnute growth model (Schnute 1981). The Von Bertalanffy is described below due to its use in the Chatham Rise simulation study.

When the Von Bertalanffy growth model is assumed, each agent is assigned an asymptotic length parameter denoted by $L_{i,\infty}$ from the following lognormal distribution

$$L_{i,\infty} \sim \mathcal{LN}\left(\mu = \log\left(\bar{L}_{\infty_r}\right) - 0.5\sigma^2_{L_\infty}, \sigma^2 = \sigma^2_{L_\infty}\right) \ ,$$

where $\bar{L}_{\infty_r}$ is the mean asymptotic length in cell r with variance $\sigma^2_{L_\infty}$. For CABM models that have spatially varying values of $\bar{L}_{\infty_r}$ and movement, CABM has the ability to update an agent's specific growth parameter ($L_{i,\infty}$) when it moves into a different cells. CABM will draw a new agent specific parameter from the lognormal distribution above with $\bar{L}_{\infty_r}$ of the destination cell allowing for spatially varying growth.

When growth is specified in the annual cycle for time step t, CABM will iterate over all cells and agents and increment each agent's length following

$$l_{t+\Delta,i} = l_{t,i} + p^t_\Delta((L_{i,\infty} - l_{t,i})(1 - \exp\{-k\})) \ . \tag{5.1}$$

Where k is the global growth coefficient, $l_{t,i}$ is the i^{th} agents length in time-step t and p^t_Δ denotes the proportion of annual increment to be added in time-step t.

The growth dynamic changes an agents weight after changing it's length using the following allometric length-weight relationship,

$$w_i = \alpha l_i^\beta \ ,$$

where α and β are length weight coefficients which are equal for all agents in the system.

A feature that was initially explored but not implemented due to time constraints was extending growth models to include local environmental covariates as discussed in Shelton & Mangel (2012).

5.3.1.2 Natural Mortality

Natural mortality is the agent dynamic responsible for removing (killing) agents due to factors other than fishing. These include predation, starvation and disease. It is applied to all agents with an agent specific survivorship probability denoted by S_i. This survivorship probability can be a function of age or length and can be applied multiple times during a year. As with the growth dynamic (Section 5.3.1.1), agent specific mortality rates are assigned to agents based on the following population level distribution

$$M_i \sim \mathcal{LN}\left(\mu = \log\left(\bar{M}\right) - 0.5\sigma^2_M, \sigma^2 = \sigma^2_M\right) \ .$$

Natural mortality is then applied to agents in cell r at time-step t following

$$S_i = e^{-M_i p_\Delta^t S(a_i)}, \quad \forall\ X_{r,i} \in \boldsymbol{X_r}\ ,$$
$$I_i \sim \mathcal{B}ern(1 - S_i)\ ,$$
$$I_i \begin{cases} 1, & \text{agent dies} \\ 0, & \text{agent lives} \end{cases},$$

where p_Δ^t is the proportion of annual mortality assumed in time-step t and $S(a_i)$ is an age-based function (but could also be length-based). Age varying mortality is often explored in stock assessment models (Horn & Francis 2010, Lee et al. 2011, McKenzie 2017), where young or smaller fish experience higher rates of natural mortality relative to older or larger fish.

5.3.1.3 Recruitment and Spawning

Recruitment is the agent dynamic responsible for creating new agents in the model each year. This dynamic uses a global measure of spawning stock biomass (SSB), denoted by SSB_t as a proxy for a fish stock's fecundity. The Beverton-Holt stock-recruitment function denoted by $SR(SSB_t; h, SSB_0)$, is used to describe the proportion of R_0, where R_0 is the long-term average recruitment expected with no fishing. The Beverton-Holt stock-recruitment function is defined as

$$SR(SSB_t; h, SSB_0) = \frac{4hS_t - 1}{SSB_0(1-h) + SSB_0(5h-1)}\ ,$$

where, SSB_0 is the long-term average spawning stock biomass expected with no fishing and h is the steepness parameter that defines the proportion of R_0 expected when $SSB_t = 0.2SSB_0$ (20% average un-fished spawning stock biomass) (Mace & Doonan 1988).

Recruitment is spatially distributed based on a spatial recruitment layer denoted by P_r ($\sum_r P_r = 1$). The number of agents created by the recruitment dynamic in cell r follows

$$R_{t,r} = \left\lfloor P_r \frac{R_0}{\tilde{n}} SR(SSB_{t-a_{min}}; h, SSB_0) e^{\epsilon_t} \right\rceil\ ,$$
$$\epsilon_t \sim \mathcal{LN}\left(-0.5\sigma_R^2, \sigma_R^2\right)\ ,$$

where, ϵ_t denotes the annual recruitment deviation around the stock recruitment function due to factors such as environmental conditions, $\lfloor . \rceil$ is the rounding operator, $\tilde{n}$ is the number of individual fish an agent represents (defined in Section 5.3.2) and a_{min} is the first age represented in CABM.

SSB (SSB_t) can be calculated in every time-step and is the result of a linear interpolation based on SSB before fishing and natural mortality denoted by SSB_t^{pre} and after fishing and natural mortality denoted by SSB_t^{post}. CABM assumes SSB occurs halfway through the mortality dynamic by

$$SSB_t = 0.5SSB_t^{pre} + 0.5SSB_t^{post} \ . \tag{5.2}$$

The method for calculating SSB for SSB_t^{pre} and SSB_t^{post} was the same and is shown below for SSB_t^{pre}

$$SSB_t^{pre} = \sum_r \sum_{\forall X_{r,i} \in \boldsymbol{X}_r} w_i n_i I_i \ ,$$
$$I_i \sim \mathcal{B}ern(P^M(a_i)) \ ,$$

where, $P^M(a_i)$ is the probability of an agent with age a_i contributing to the spawning stock biomass, w_i is the weight of an agent and n_i number of individuals in an agent.

5.3.1.4 Movement

Agents move around the spatial domain based on environmental and physical covariates specified for each cell, termed preference covariates. This movement is similar to that used in the individual based model by Phillips et al. (2018) and population model by Dunn et al. (2012). CABM assumes agents have knowledge of preference covariates in adjacent cells and will move on average towards cells that have preferable covariate values and move away from cells that contain unpreferable covariate values. How far an agent moves also depends on the preference covariates of the cell an agent is in. If an agent resides in a preferable cell, it will be encouraged to stay (constrained movement), whereas if it resides in an unpreferable cell it may move larger distances.

CABM calculates the overall preference for all cells using preference covariate values and functions that describe preference of agents to values of the covariates (Equation 5.3). CABM then calculates the gradient of the overall preference for all

cells which is used to move agents around the domain. The preference gradient is analogous to hills and valleys on a topographic landscape. An agent can be likened to a ball that is dropped onto a topographic landscape and is allowed to make discrete movements. The ball will move towards the valleys, just as the agents will move towards regions with preferable covariates.

The preference value in cell r for preference covariate x is denoted by P_r^x and calculated by,

$$P_r^x = f(x_r; \boldsymbol{\theta}^x), \quad P_r^x \in [0, 1] \ ,$$

where x_r is the covariate value in cell r, $f(x_r; \boldsymbol{\theta}^x)$ denotes the preference function with parameters $\boldsymbol{\theta}^x$. A value of zero for P_r^x indicates unpreferable covariate values and a value of one indicates highly preferable values. Functional forms of preference functions available in CABM are described in the following list and illustrated in Figure 5.1.

- Double normal

$$f(x_r; \mu, \sigma_L, \sigma_R) = \begin{cases} 2^{-[(x_r-\mu)/\sigma_L]^2}, & \text{if } x_r \leq \mu \\ 2^{-[(x_r-\mu)/\sigma_R]^2}, & \text{if } x_r \geq \mu \end{cases} \ ,$$

- normal

$$f(x_r; \mu, \sigma) = 2^{-[(x-\mu)/\sigma]^2} \ ,$$

- logistic

$$f(x_r; a_{50}, a_{to95}) = 1/[1 + 19^{(a_{50}-x)/a_{to95}}] \ ,$$

- inverse logistic

$$f(x_r; a_{50}, a_{to95}) = 1.0 - 1/[1 + 19^{(a_{50}-x)/a_{to95}}] \ .$$

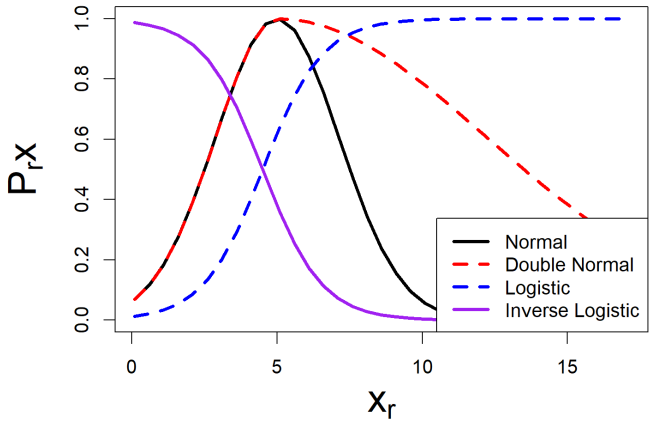


Figure 5.1: Examples of preference functions available in CABM.

The overall preference in cell r is the geometric mean over all preference covariates

$$P_r = \left(\prod_{k=1}^{n_k} P_r^k \right)^{1/n_k} . \tag{5.3}$$

Gradients of P_r in the north-south direction denoted by v_r and east-west direction denoted by u_r are calculated using the finite difference method. The spatial domain can be represented as a two-dimensional array, where cell r is expressed by row index j and column index k. Gradients for central cells are calculated as

$$\nabla u_r = \frac{P_{j,k+2} - P_{j,k-2}}{2} ,$$
$$\nabla v_r = \frac{P_{j+2,k} - P_{j-2,k}}{2} .$$

Gradients for boundary cells are calculated based on the difference between the two edge cells. To prevent gradients sloping towards the boundaries and encouraging agents towards the edge of the domain, low preference values are imputed in edge cells of the spatial domain.

Agents move in the x and y directions with a biased random walk using the preference gradients. For the i^{th} agent who currently resides in cell r_i, with coordinates x_i and y_i. Its new coordinates after movement will be

$$x_i' = x_i + \mathcal{N}\left(c\nabla u_{r_i}, \sigma_{r_i}\right) ,$$
$$y_i' = y_i + \mathcal{N}\left(c\nabla v_{r_i}, \sigma_{r_i}\right) ,$$

where x_i' and y_i' are new coordinates after movement. If the new coordinates are with in a different spatial cell, then the cell reference r_i is updated for the agent.

Depending on the spatial temporal resolution of the model, ∇v and ∇v can be scaled by the multiplier c. This is to align average movement distances with the expected distance travelled by the species under investigation.

This dynamic assumes an agent has perfect local knowledge of environmental covariates in adjacent cells, and will move on average based on the gradient from their current cell. The standard deviation of the biased walk σ_r is specified as a function of the preference for the current cell, as

$$\sigma_r = \sqrt{2 * D_r * t} \ . \tag{5.4}$$

Here t is the time interval (if the process is applied n times in a year then $t = 1/n$), D_r inversely relates to the preference of cell r and parameter D_{max} by

$$D_r = D_{max} \left(1 - \frac{P_r}{\zeta + P_r} \right) \ . \tag{5.5}$$

Here ζ is an arbitrary constant controlling the curvature of the function, following Bertignac et al. (1998) (illustrated in Figure 5.2). This function controls the variability of movement, as $P_r \to 0$ then $D_r \to D_{max}$ and vice versa. This assumes agents will remain in areas with preferable habitat (constrained movement) and will move farther away when they are in areas of unpreferable habitat.

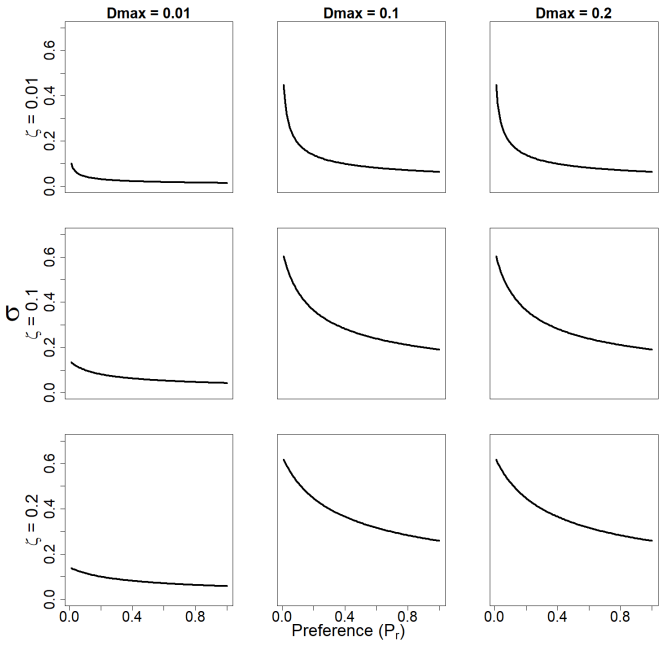


Figure 5.2: Values of σ_r as a function of ζ (rows) D_{max} (cols) and preference of a cell P_r.

If the biased random walk attempts to move an agent outside the spatial domain the movement is rejected and the agent remains at the same location

$$\begin{aligned} x_i' &= x_i \qquad \text{if , } x_i' < x_{min} \text{ or } x_i' > x_{max} \text{ ,} \\ y_i' &= y_i \qquad \text{if , } y_i' < y_{min} \text{ or } y_i' > y_{max} \text{ .} \end{aligned}$$

An agent's cell attribute r_i (see Table 5.1) is updated after the movement based on x_i' and y_i'.

CABM can apply multiple preference movement dynamics in a time-step, where each preference movement dynamic can be associated to a specific demographic of the stock, i.e., age or length. This allows CABM to link certain movement dynamics with specific demographics. It is common for fish at different life stages to occupy different niches (i.e. hoki (McKenzie 2017) and sablefish (Goethel et al. 2020)). CABM does this by assigning a probability function denoted by $S^k(.)$ to the k^{th} preference movement dynamic. Agents in cell r are moved by the k^{th} preference movement

dynamic based on the following action,

$$I_i \sim \mathcal{B}ern(S^k(l_i)) \ , \forall \ X_{r,i} \in \boldsymbol{X_r} \ ,$$

$$I_i \begin{cases} 1, & \text{agent } i \text{ will move based on movement dynamic } k \\ 0, & \text{agent } i \text{ remains in current location} \end{cases} ,$$

where, l_i is an agents length (can also be age).

5.3.1.5 Fishing Mortality

The fishing dynamic used in for the Chatham Rise CABM model (there are a range of different fishing dynamics available in CABM) is based on the approach used in Truesdell et al. (2017). They applied this fishing dynamic to apply spatially heterogeneous fishing mortality to sessile scallop populations. CABM has extended upon this study by applying this fishing dynamic to mobile species. The core assumption is that the spatial distribution of fully selected fishing mortality is a function of vulnerable (exploitable) biomass to the fishery and auxiliary fishing covariates such as distance from shore, trawlable ground and closed areas.

If auxiliary fishing covariates are ignored, then this dynamic assumes fishing mortality is spatially distributed according to the Ideal Free Distribution (IFD) (Gillis 2003), with the highest levels of fishing mortality being applied in cells with the most vulnerable biomass.

Auxiliary fishing covariates represent spatial factors that reflect a "cost" to fishing within the spatial domain. When they are assumed in this dynamic, fully selected fishing mortality is spatially distributed based on vulnerable biomass and auxiliary fishing covariates, known as the Weighted Ideal Free Distribution (WIFD) (Truesdell et al. 2017). Auxiliary fishing covariates can be used to exclude fishing completely in cells (i.e. closed or unfishable regions) or dissuade fishing in some cells relative to others (i.e. constrain a fishery to fish closer to port).

To apply this dynamic users need to specify a vector of relative fishing mortalities (an element for each spatial cell in the model) denoted by $\boldsymbol{E} = (E_1, \ldots, E_R)$. CABM converts these relative fishing mortality values to actual fishing mortality values via,

$$F_r = E_r \lambda \ ,$$

where λ is an estimable parameter and F_r is the fully selected fishing mortality applied to agents in cell r. CABM will estimate λ which is explained in more detail below.

The order of elements in $\boldsymbol{E}$ is not important because each element is assigned to a spatial cell by CABM using the ranking variable denoted by $\tilde{\boldsymbol{E}} = (\tilde{E}_1, \ldots, \tilde{E}_R)$. This ranking variable is calculated by CABM and is a function of vulnerable biomass denoted by V_r and fishing covariates (Equation 5.6). The highest value of $\boldsymbol{E}$ is assigned to the cell with the highest value of the ranking variable value and the second highest value of $\boldsymbol{E}$ gets assigned to the cell with the second highest value of the ranking variable, and so on, until all $\boldsymbol{E}$ are assigned to a spatial cell.

Although the order of $\boldsymbol{E}$ specified by users is not important, the distribution of $\boldsymbol{E}$ is. Variability in $\boldsymbol{E}$ will result in CABM applying unequal fishing mortality rates over the spatial domain (i.e. some cells will have higher fishing mortality rates than others). The simplest assumption is to assume no variability in $\boldsymbol{E}$ (all values are the same), which assumes equal fishing mortality in all cells. For more realistic assumptions, variability in $\boldsymbol{E}$ can be assumed which will result in spatially heterogeneous fishing mortalities.

Fishing mortality is often assumed to be a function of fishing effort, i.e. Schaefer (1943). Using this link, spatial distributions of effort can be used to inform values of $\boldsymbol{E}$. The Chatham Rise fishery exhibited a right skewed distribution for the number of fishing events in a spatial cell (Figure 5.3), which lead to the use of the lognormal distribution for $\boldsymbol{E}$ in the Chatham Rise CABM model (Section 6.3.5).

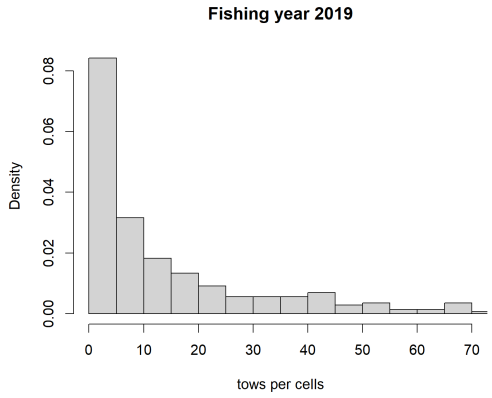


Figure 5.3: Distribution of fishing events (tows) per cell in 2019 from the Chatham Rise hoki fishery catch effort data set.

CABM calculates $\tilde{\boldsymbol{E}}$ as a function of vulnerable biomass and auxiliary fishing covariates for all cells. For cell r the vulnerable biomass is calculated as

$$V_r = \sum_{\forall X_{r,i} \in \boldsymbol{X_r}} I_i w_i n_i \ ,$$
$$I_i \sim \mathcal{B}ern(S(a_i)) \ ,$$

where, $S(a_i)$ is the age based selectivity interacting with agent i with corresponding age a_i, weight denoted by w_i and represents n_i individual fish.

The ranking variable is then calculated as,

$$\tilde{E}_r = \frac{1}{n_k + 1} \left(\sum_{k=1}^{n_k} M_r^k w_k + \frac{V_r}{\max\limits_j \{V_j\}} \right) \ , \tag{5.6}$$

where,

- V_r is the vulnerable biomass and $\max\limits_j \{V_j\}$ is the maximum vulnerable biomass over all cells,
- n_k is the number of axillary covariates (other than vulnerable biomass),
- z_r^k is the value of the k$^{\text{th}}$ covariate in cell r
- $M_r^k = g^k(z_r^k, \boldsymbol{\theta_{zk}}), \in [0, 1]$, with $g^k()$ being a "cost" function for covariate k, with parameters $\boldsymbol{\theta_{zk}}$. The function $g^k()$ can take the same functional forms as preference functions (Section 5.3.1.4) within CABM. Values returned from this function represent a "cost" to fishing in each cell. Values close to one encourage fishing and values close to zero dissuade fishing,
- w_k is the weighting factor for the k$^{\text{th}}$ auxiliary covariate. This influences the contribution of each covariate relative to the vulnerable biomass, i.e. a value of one would give equal weight to the covariate relative to vulnerable biomass, assuming $M_r^k = 1$.

The function $g^k()$ allows for non-linear relationships between axillary fishing covariates and fishing mortality. A hypothetical example is used to demonstrate this non-linearity. It is plausible to assume inshore fishing fleets may be constrained to fish close to port due to factors such as fuel constraints, storage space for fish etc. Within

a certain distance from port, fishers are likely to choose locations that will yield the best catch rates because there will be little constraints. However, the likelihood of fishing at locations far away will decrease due to the factors that constrain the fleet. This could be expressed using the inverse logistic function for M_r^k ($M_r^k = g^k()$) shown in Figure 5.4.

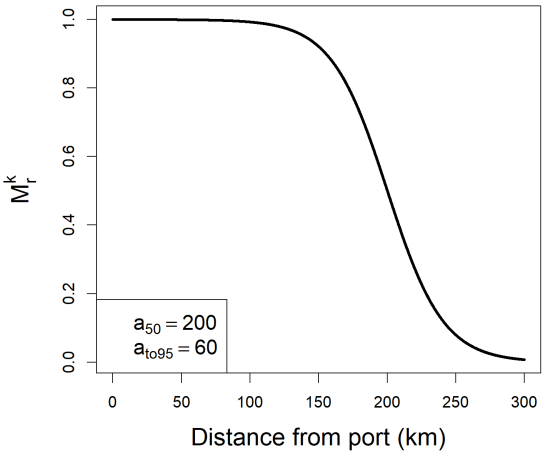


Figure 5.4: A hypothetical cost function for an inshore fleet being constrained to fish closer to port.

Once CABM calculates $\tilde{\boldsymbol{E}}$ for all spatial cells (Equation 5.6), it assigns $\boldsymbol{E}$ to spatial cells based on $\tilde{\boldsymbol{E}}$, with the largest value of $\boldsymbol{E}$ being assigned to the cell with largest value of $\tilde{\boldsymbol{E}}$ and so on. Once $\boldsymbol{E}$ has been assigned to a spatial cell, CABM performs a sum of squares minimisation routine to estimate the estimable parameter λ.

CABM requires users to supply a catch over the entire spatial domain denoted by C. The sum of squares minimisation is done each year the fishing dynamic is applied.

$$\begin{aligned}
\widehat{\lambda} &= \arg\min_{\lambda} \left(C - \widehat{C}\right)^2 \ , \\
\widehat{C} &= \sum_{r} \sum_{\forall X_{r,i} \in \boldsymbol{X_r}} I_i w_i n_i \ , \\
I_i &\sim \mathcal{B}ern(p_i) \ , \\
p_i &= 1 - e^{-E_r S(a_i) \lambda} \ .
\end{aligned} \tag{5.7}$$

In order to have a unique solution for $\widehat{\lambda}$ during the above minimisation, CABM fixes the stochastic realisation of the Bernoulli random variable. CABM implements the Bernoulli random variable using a uniform random variable. Given $x \sim \mathcal{B}ern(p_i)$, CABM generates a realisation of x following

$$\begin{aligned} y &\sim \mathcal{U}niform\,(0,1) \ , \\ x &= \begin{cases} 1, & \text{if } y < p_i \\ 0, & \text{if } y \geq p_i \end{cases} \ . \end{aligned} \tag{5.8}$$

This implementation allows a single draw of y for each agent at the beginning of minimisation. As different values of λ are trialed during minimisation, p_i will change based on Equation 5.7 ($p_i = 1 - e^{-E_r S(a_i)\lambda}$) possibly resulting in a different outcome of I_i due to the uniform implementation of the Bernoulli random variable (Equation 5.8).

Cell specific catch (C_r) is a derived quantity and reported by CABM. It is calculated during Equation 5.7 as

$$\begin{aligned} C_r &= \sum_{\forall X_{r,i} \in \boldsymbol{X_r}} I_i w_i n_i \ , \\ I_i &\sim \mathcal{B}ern(p_i) \ , \\ p_i &= 1 - e^{-E_r S(a_i)\widehat{\lambda}} \ . \end{aligned}$$

5.3.2 Initialising CABM

CABM calculates the number of individuals that an agent represents during initialisation. It is initially derived by calculating the total number of individuals by extending the maximum age a_{max} by four times to approximate the plus group,

$$\begin{aligned} \widetilde{N} &= \sum_{a=a_{min}}^{4\times a_{max}} R_0 e^{-M^* a} \ , \\ \tilde{n} &= \frac{\widetilde{N}}{n_{agents}} \ , \end{aligned}$$

where, a is the age, a_{min} is the minimum age, M^* is the initial natural mortality rate (it should be the same as $\bar{M}$ from the morality process, see Section 5.3.1.2), R_0 is the average number of recruited individuals expected in the absence of fishing

(see Section 5.3.1.3) and n_{agents} is the number of agents assumed from the users to model the initial stock. The choice of n_{agents} is a tradeoff between model run time and agent resolution of the stock. As n_{agents} increases CABM moves towards an IBM ($\tilde{n} \to 1$) but this comes at computational cost and larger model run times.

Once CABM calculates $\tilde{n}$, it creates the number of agents for the first age (a_{min}) in each cell denoted by $R_{r,a_{min}}$. This is calculated as

$$R_{r,a_{min}} = \lfloor R_0 / \tilde{n} P_r \rceil \ ,$$

where P_r is the proportion of initial population allocated to cell r and $\lfloor . \rceil$ denotes the rounding operator (the number of agents created needs to be an integer). CABM then creates the remaining number of agents for remaining ages in cell r following

$$R_{r,a} = \lfloor R_{r,a-1} \exp\{-M^*\} \rceil \ , \ a > a_{min} \ .$$

When agents are created, they are also assigned agent attributes based on their age and agent specific attributes. The above actions from CABM assume an equilibrium age-structure of agents in each cell, but ignore movement and other dynamics that may effect starting conditions. To account for these dynamics, CABM then iterates over the annual cycle without fishing dynamics for a user defined number of cycles denoted by n_{init}. This populates the agents around the spatial domain according to the annual cycle assumptions.

5.3.3 Input files for CABM

CABM is a command line program that reads text files as input. The text files have a specific structure which define parameter values, agent dynamics and spatial resolution for a model. Syntax style of inputs files were based on the generalised stock assessment packages Casal2 (Doonan et al. 2016), SPM (Dunn et al. 2012) and CASAL (Bull et al. 2012). A core design philosophy of CABM was having input files that are interpretable and easy to read. An excerpt of an example input file is provided in Figure 5.5. This described the `@model` component which defines model structures and illustrates readability of the input files and flexibility in specifying spatial structures and agent dynamics.

```
## Model structure
@model
start_year 1972 # first year in CABM
final_year 2016 # last year in CABM
min_age 1       # minimum age of agents
time_steps Summer Autumn Winter Spring # Time-steps in annual cycle
length_bins  1:110 # length bins that agents record as they grow (i.e. 1 cm increments)
length_plus true   # is the last length bin an accumulation length bin
sexed false        # are there male and female agents in the state
nrows 20           # number of rows in the spatial domain
ncols 50           # number of columns in the spatial domain

@time_step Summer
processes recruitment growth fishing
```

Figure 5.5: An example input configuration file for CABM. Blue and red text indicate commands expected by the ABM, black text is specified by users and green text refers to comments that are ignored by CABM.

5.4 Summary

This chapter has described the generalised ABM program, CABM. The aim was to provide descriptions on dynamics and core assumptions used in operating models during the simulation study in the subsequent chapter. Simulations using CABM in the subsequent chapter, allowed us to explore properties of geostatistical models with a realistic operating model.

Although CABM was initially created to be used as an OM for this thesis. We designed CABM to be a flexible operating model that can emulate a range of life histories and spatial structures. These flexibilities coupled with CABM's easy to read input syntax files make CABM a unique operating model.

Chapter 6

Chatham Rise hoki ABM

6.1 Overview

This chapter conducted a simulation study using CABM (Chapter 5) as the operating model to further explore the PS model. The objective of the simulation study was to identify factors that cause β^{pref} in the PS model to be estimated at boundary constraints. This was observed for one of the PS models in the Chatham Rise hoki fishery case study (see Section 4.6), in addition to be observed in other PS model investigations (Cochran 2007).

6.2 Introduction

CABM was developed to emulate hoki spatial distributions and productivity assumptions on the Chatham Rise. Multiple Chatham Rise CABM operating models (CH-CABM OMs) were developed using multiple sources of information including; the case study (Chapter 4), age-structured stock assessment model (McKenzie 2017), long term survey (Ballara et al. 2017) and catch at age sampling programme (Ballara & O'Driscoll 2020). Although the CH-CABM OMs were conditioned on information from the Chatham Rise hoki fishery, they were not intended to exactly replicate dynamics of hoki on the Chatham Rise. The stock assessment for hoki is one of New Zealand's most complex age-structured stock assessments (McKenzie 2017). It assumes four areas, two biological stocks, sexual dimorphism and six separate fisheries. CH-CABM OMs only considered one of the four areas (the Chatham Rise)

and assumed a single closed population, averaged growth parameters over both sex's and included only a single fishery.

The aim of the simulation study was to identify factors that cause β^{pref} in the PS model to be estimated at boundary constraints. These factors were; the proportion of spatial domain fished, auxiliary covariates that affect fishing spatial distributions and spatial closures. These factors were implemented in the CH-CABM OMs via different fishing mortality dynamic assumptions (Section 6.3.5).

Eight CH-CABM OMs were developed which all had identical assumptions regarding the following agent dynamics; growth, recruitment, natural mortality and movement. The CH-CABM OMs only differed in their fishing dynamic assumptions. These factors were explored because they were identified in the Chatham Rise hoki fishery case study, or hypothesised to affect the estimation of β^{pref}. Further details regarding why these factors were chosen and how they were implemented are given in Section 6.3.5.

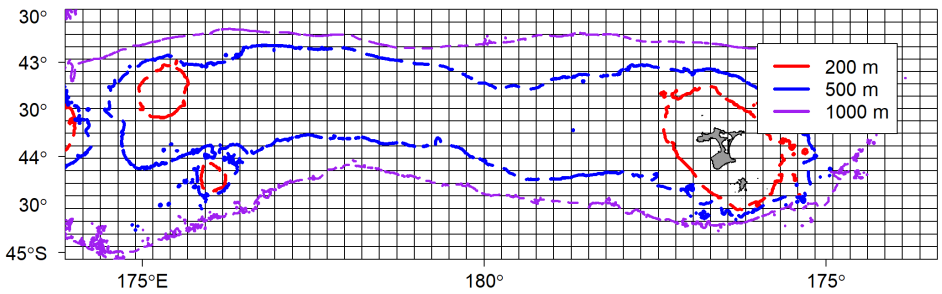


Figure 6.1: Spatial resolution of the CH-CABM OMs.

The CH-CABM OMs partitioned the Chatham Rise into 20 rows and 50 columns (1000 cells Figure 6.1). The CH-CABM OMs assumed a single annual time-step which applied the following agent dynamics;

1. ageing (all agents become one year older),
2. movement (Section 5.3.1.4),
3. recruitment (Section 5.3.1.3),

4. growth (Section 5.3.1.1),
5. calculate spawning stock biomass before mortality S_t^{pre} (Section 5.3.1.3),
6. mortality; half natural mortality (Section 5.3.1.2), fishing mortality (Section 5.3.1.5), remaining natural mortality (Section 5.3.1.2),
7. calculate S_t^{post} (Section 5.3.1.3).

The following sections outline parameter values and assumptions for the agent dynamics in the eight CH-CABM OMs.

6.3 Agent dynamics

6.3.1 Growth

The Von Bertalanffy growth parameters assumed for CH-CABM (see Section 5.3.1.1) are specified in Table 6.1. Von Bertalanffy growth parameters were averaged over both sexes from the stock assessment model and assumed $\bar{L}_{\infty_r}$ was constant over space.

To validate length distributions for a collection of agents, a CH-CABM OM was run and reported lengths of 10 000 agents at each age (Figure 6.2). These were compared with the population level growth assumptions.

Table 6.1: Von Bertalanffy parameters assumed for the CH-CABM OMs.

Parameter	Value
$\bar{L}_{\infty_r}$	96.97
$\sigma^2_{L_\infty}$	0.1
k	0.217
α	$4.79e^{-09}$
β	2.89
p^t_Δ	1.0

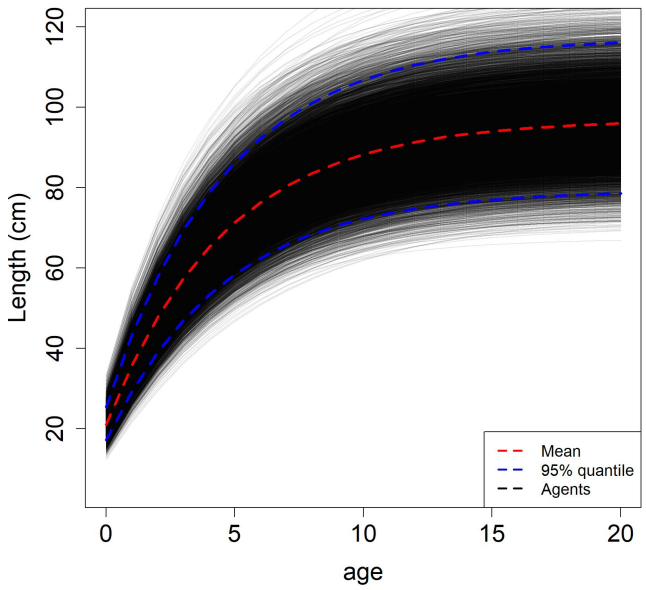


Figure 6.2: Length from 10 000 hoki agents at each age (black lines). The red line indicates mean growth and blue lines indicate 95% quantile of the population level lognormal length at age.

6.3.2 Natural mortality

Natural mortality is applied twice in the annual cycle, before and after fishing mortality. In each application it only applies half of the annual natural mortality rate. It assumes natural mortality is age-invariant and does not vary among agents, i.e. $M_i = \bar{M}$ (Equation 5.3.1.2) which is the same assumption in the stock assessment model.

Table 6.2: Natural mortality parameters assumed for the CH-CABM OMs.

Parameter	Value
$\bar{M}$	0.296
$S(a_i)$	1 (constant)

6.3.3 Movement

There were two movement dynamics applied in the annual cycle. Both assumed depth was the only preference covariate (Figure 6.3). Depth was chosen because

repeated research surveys have found a relationship with fish size and depth (Ballara & O'Driscoll 2020). Specifically, larger fish occupy deeper regions, and smaller fish occupy shallower regions.

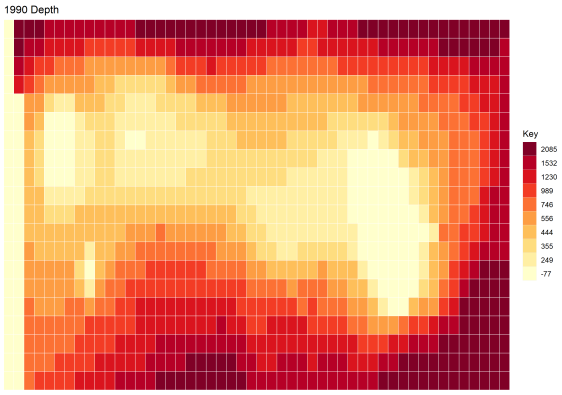


Figure 6.3: Depth by cell in the CH-CABM OMs.

To mimic this behaviour, two preference movement dynamics were assumed. A "shallow" preference movement which moved smaller agents to areas of shallow depths and a "deep" preference movement, which moved larger fish to deeper areas. The probability of an agent moving by either of these movement dynamics is shown in Figure 6.4 (left panel). The preference to depth for each of these movement dynamics is also shown in Figure 6.4 (right panel).

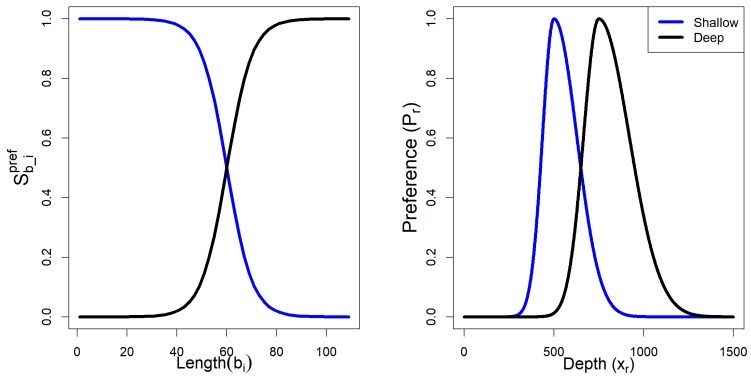


Figure 6.4: The left panel shows the probability of an agent with length l_i moving according to both movement dynamics. The right panel shows the preference to depth for each of the movement dynamics.

Figure 6.4 shows that smaller length fish (approximately less than 60 cm) will prefer to be in depths between 250-650 m depths and larger fish (approximately greater than 60 cm) prefer depths between 600 - 1100 m. These lengths and depth preferences were informed by size discrepancy and depths from research trawls and catch at age sampling (Ballara et al. 2017, Ballara & O'Driscoll 2020). The spatial distribution of overall preference for both movement dynamics is shown in Figure 6.5.

The remaining parameters that describe movement were c, D_{max} and ζ. These were the same for both preference movements. They describe the average movement distance and relationship between variance of the bias random walk and preference covariates. Values chosen are outlined in Table 6.3. These were based on iteratively running the model with different values and observing the resulting spatial distributions. Large standard deviation values resulted in a near homogeneous spatial distribution, and low values resulted in almost no movement, i.e. static spatial distribution. The values chosen were somewhat arbitrary but landed in the middle of these extremes.

Table 6.3: Parameters assumed for both shallow and deep movement dynamics.

Preference dynamic	parameter	values
both	ζ	0.02
both	D_{max}	0.1
both	c	1

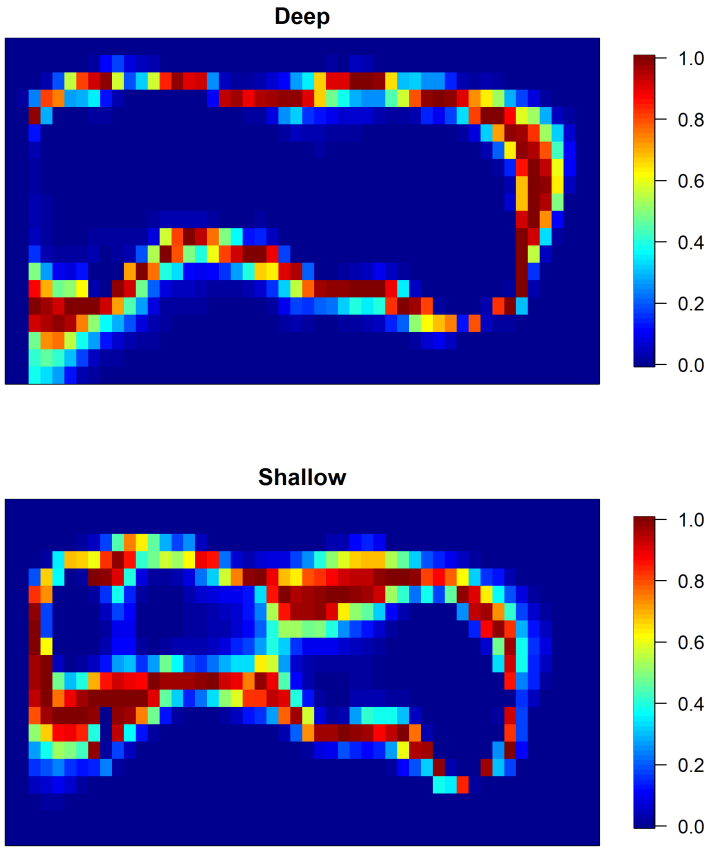


Figure 6.5: Overall preference (Equation 5.3 from Section 5.3.1.4) by cell for both "shallow" and "deep" movement dynamics.

6.3.4 Recruitment

Parameters assumed for the CH-CABM OMs are supplied in Table 6.4. The year class parameters and R_0 were derived by averaging the estimated values for the Eastern and Western stocks from the 2016 estimated stock assessment values (McKenzie 2017).

Table 6.4: Biological parameters assumed in CH-CABM.

parameter	values
R_0	3.0159 10^8
σ_R	0.6
h	0.75
P_r	Figure 6.6
maturity $P^M(a_i)$	Figure 6.7
$YCS_t = e^{\epsilon_t}$	Figure 6.8

Spatial allotment of recruits (P_r) is shown in Figure 6.6. This was derived from shallow movement preference dynamic (Section 5.3.1.4). Juvenile fish are found in shallower waters (O'Driscoll et al. 2011) and the OM emulated this by seeding recruits between the depth range 250-650 m. Values for the remaining parameters h, ϵ_t, σ_R were taken from the 2016 stock assessment.

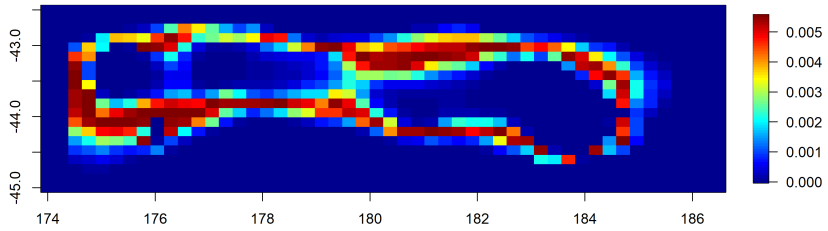


Figure 6.6: Proportion of spatial allocation of recruitment events P_r.

.

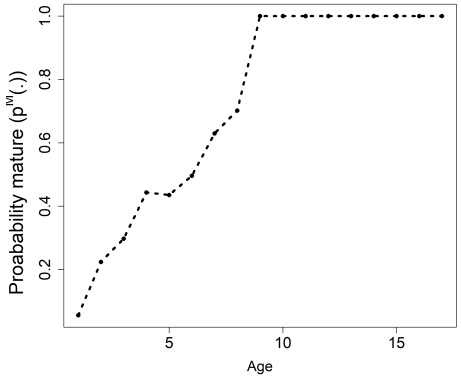


Figure 6.7: Probability of an agent being mature by age. This was assumed in the 2016 assessment model.

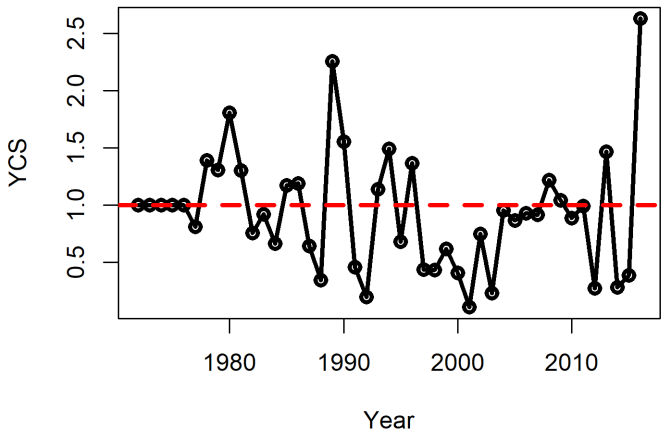


Figure 6.8: Year class strength parameters ($YCS_t = e^{\epsilon_t}$) assumed for each year in CH-CABM.

6.3.5 Fishing mortality

Eight CH-CABM OMs were developed (Table 6.5 and illustrated in Figure 6.14) with varying assumptions regarding the fishing mortality dynamic. Each OM corresponded to a different fishing scenario, where a fishing scenario included one or more of the following factors; the proportion of spatial domain available to fishing, presence of an auxiliary fishing covariate (WIFD) and spatial closures.

Spatial sampling coverage was proposed as a factor that could cause estimation issues for the PS model because the Chatham Rise case study exhibited a consistent sampling pattern between 300-950 m depth with little sampling outside this depth range (Figure 4.16). This high concentration of sampling in core fishing areas followed by very low sampling at the edges, was thought to be a contributing factor for β^{pref} being estimated at boundary constraints. This factor was also identified in Ducharme-Barth et al. (2022), as a factor that effects geostatistical model inference.

Spatial sampling coverage was implemented in the OMs by assuming a percentage of cells had no fishing in them (i.e., a percentage of $\boldsymbol{E}$ equal to zero). Four levels of spatial coverage were explored. To this end, 100%, 50%, 20% and 10% of the cells were available for fishing. The labels for these fishing scenarios were IFD constant, IFD 50%, IFD 20% and IFD 10% respectively (Table 6.5). The lower the proportion

of cells available the higher the fishing mortality was expected for cells available to fishing.

An auxiliary fishing covariate was explored as a factor based on economic reasoning, in addition to its exploration in other studies (Truesdell et al. 2017, Smith 2005). The auxiliary covariate assumed in the CH-CABM OMs was distance from land, denoted by D_r (Figure 6.10a). The purpose was to include an economic "cost" for fishing further away from ports. Spatial distribution plots of catch and effort from the Chatham Rise hoki case study did exhibit a higher concentration of fishing effort closer to land compared with further offshore (Figure 6.9). Although we do acknowledge that this observed spatial distribution may be the result of factors other than distance from land i.e., underlying hoki abundance.

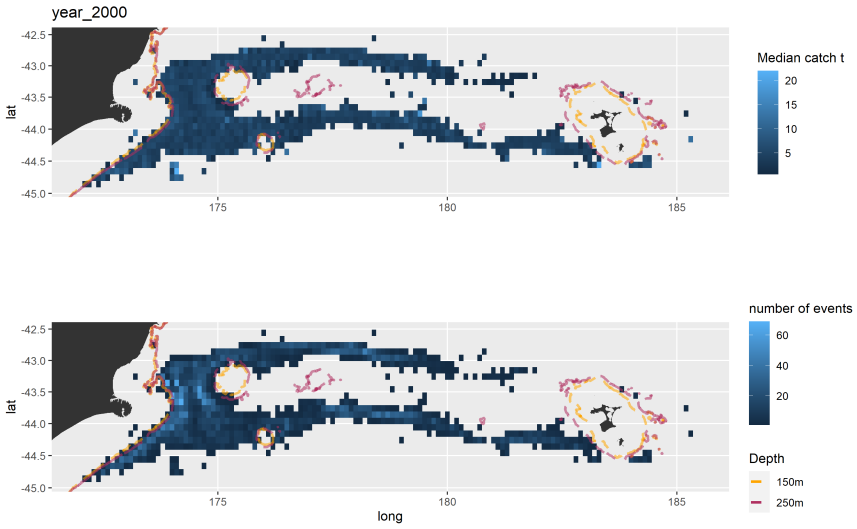
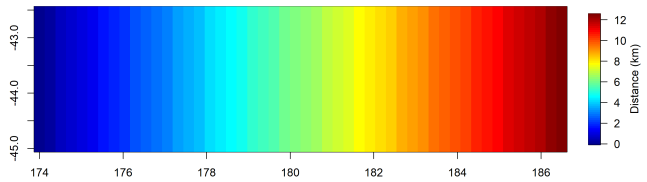


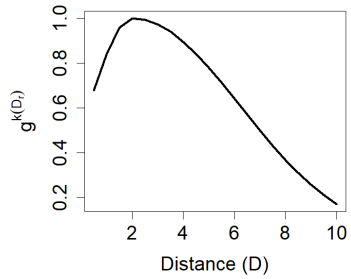
Figure 6.9: Median catch and number of fishing events per cell for the year 2000 from the Chatham Rise hoki fishery.

The Chatham Rise hoki fishery (Chapter 4) is fished by predominately offshore trawlers. This long distance capability of the fleet was thought to be a useful attribute when exploring the PS model, because the fleet would not be constrained to fish close to ports. However, this factor was still included in this simulation to see if the PS model was robust to departures in this assumption. Two fishing scenarios were assumed with different weighting multipliers ($w_k = (0.5, 1)$ from Equation 5.6), which were labelled WIFD 50% 1 and WIFD 50% 2 respectively.

The shape of the "cost" function for distance from land was double normal (Figure 6.10b). This was to dissuade fishing from shallow coastal waters, and then decrease as they remove far from land.



(a) Auxiliary fishing covariate distance from land (D_r) in each cell of the CH-CABM OMs.



(b) Fishing preference function (M_r^k from Equation 5.6) for the auxiliary fishing covariate distance from land.

Figure 6.10: Auxiliary covariate assumptions for WIFD fishing scenarios.

Spatial closures were added due to the Chatham Rise hoki fishery introducing voluntary spatial closures (Group 2018). The analysis of the Chatham Rise hoki fishery (Section 4.3) selected catch and effort data to be after the introduction of these voluntary management actions. The intention of selecting data after this event was to remove the effect of shifting fishing effort and behaviour due to these introduced measures (Section 4.3). However, the closed area was still present in the Chatham Rise hoki fishery analysis and hence was included in this simulation. Two scenarios were explored with respect to spatial closures. The "Full closure" scenario assumed the spatial closure for the entire time-series, whereas the "Partial closure" scenario assumed the spatial closure was introduced in the year 2010. As discussed later in Section 6.4, catch and effort data sets were simulated for the years 2000 to 2013. The choice of introducing the spatial closure in the year 2010 was arbitrary but intended to represent a scenario where a closure was introduced part-way through the time-

series. Cells that were closed to fishing for the spatial closure scenario are shown in Figure 6.11.

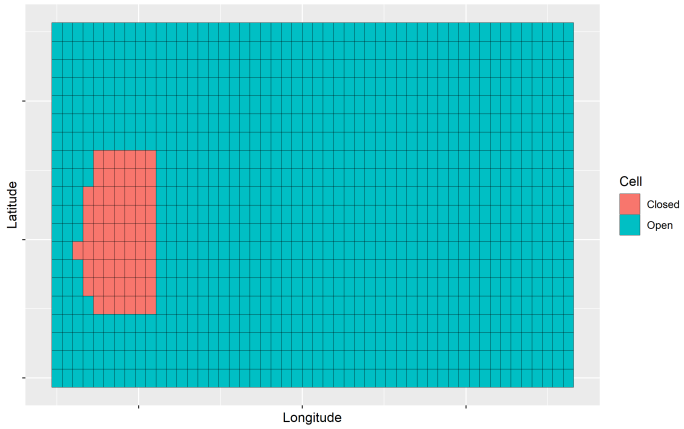


Figure 6.11: Closed area assumed for the two fishing scenarios, loosely based on closed areas from Group (2018).

Table 6.5: Summary of the eight CH-CABM OMs (fishing scenarios). These are visually illustrated in Figure 6.14 for the year 2010.

Fishing scenario	Variables for $\tilde{\boldsymbol{E}}$ [1]	$\boldsymbol{E}$ distribution	Closed area
IFD Constant	V_r	constant $\boldsymbol{E} = 0.2$	no
IFD 50%	V_r	$\boldsymbol{E} \sim \mathcal{LN}(0.3, 0.6)$, with $50\% = 0$	no
IFD 20%	V_r	$\boldsymbol{E} \sim \mathcal{LN}(0.3, 0.6)$, with $80\% = 0$	no
IFD 10%	V_r	$\boldsymbol{E} \sim \mathcal{LN}(0.3, 0.6)$, with $90\% = 0$	no
WIFD 50% 1	V_r and D_r, $\omega_D = 0.5$	$\boldsymbol{E} \sim \mathcal{LN}(0.3, 0.6)$, with $50\% = 0$	no
WIFD 50% 2	V_r and D_r, $\omega_D = 1$	$\boldsymbol{E} \sim \mathcal{LN}(0.3, 0.6)$, with $50\% = 0$	no
Full closure	V_r and D_r, $\omega_D = 0.35$	$\boldsymbol{E} \sim \mathcal{LN}(0.3, 0.6)$, with $50\% = 0$	yes
Partial closure [2]	V_r and D_r, $\omega_D = 0.35$	$\boldsymbol{E} \sim \mathcal{LN}(0.3, 0.6)$, with $50\% = 0$	yes

[1] V_r = Vulnerable biomass, D_r = Distance from land with fishing preference function illustrated in Figure 6.10b
[2] Closed area Figure 6.11 from 2010:2013

The lognormal distribution was used for generating relative fishing mortality values ($\boldsymbol{E}$). This was chosen to mimic observed right skewed effort (number of fishing events per cell) distributions observed in the Chatham Rise hoki fishery (Figure 6.12). Annual catch (C_t in Equation 5.7) was based on catches from the stock assessment model (McKenzie 2017) and visualised in Figure 6.13.

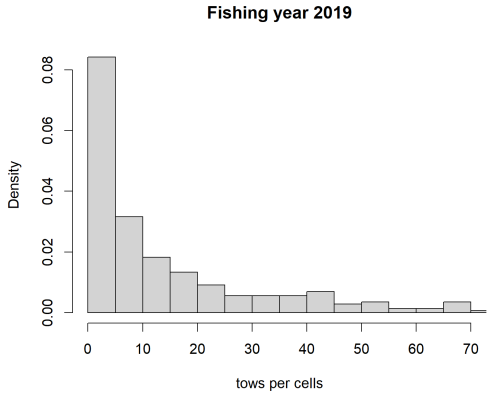


Figure 6.12: Distribution of tows (fishing events) per cell in 2019 from the Chatham Rise hoki fishery catch effort data set.

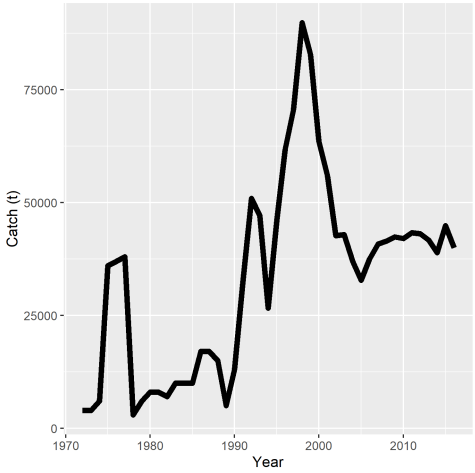


Figure 6.13: Total catch assumed in al years for all CH-CABM OMs. C from Equation 5.7 which is used to estimate λ and convert relative fishing mortalities into fully selected fishing mortalities.

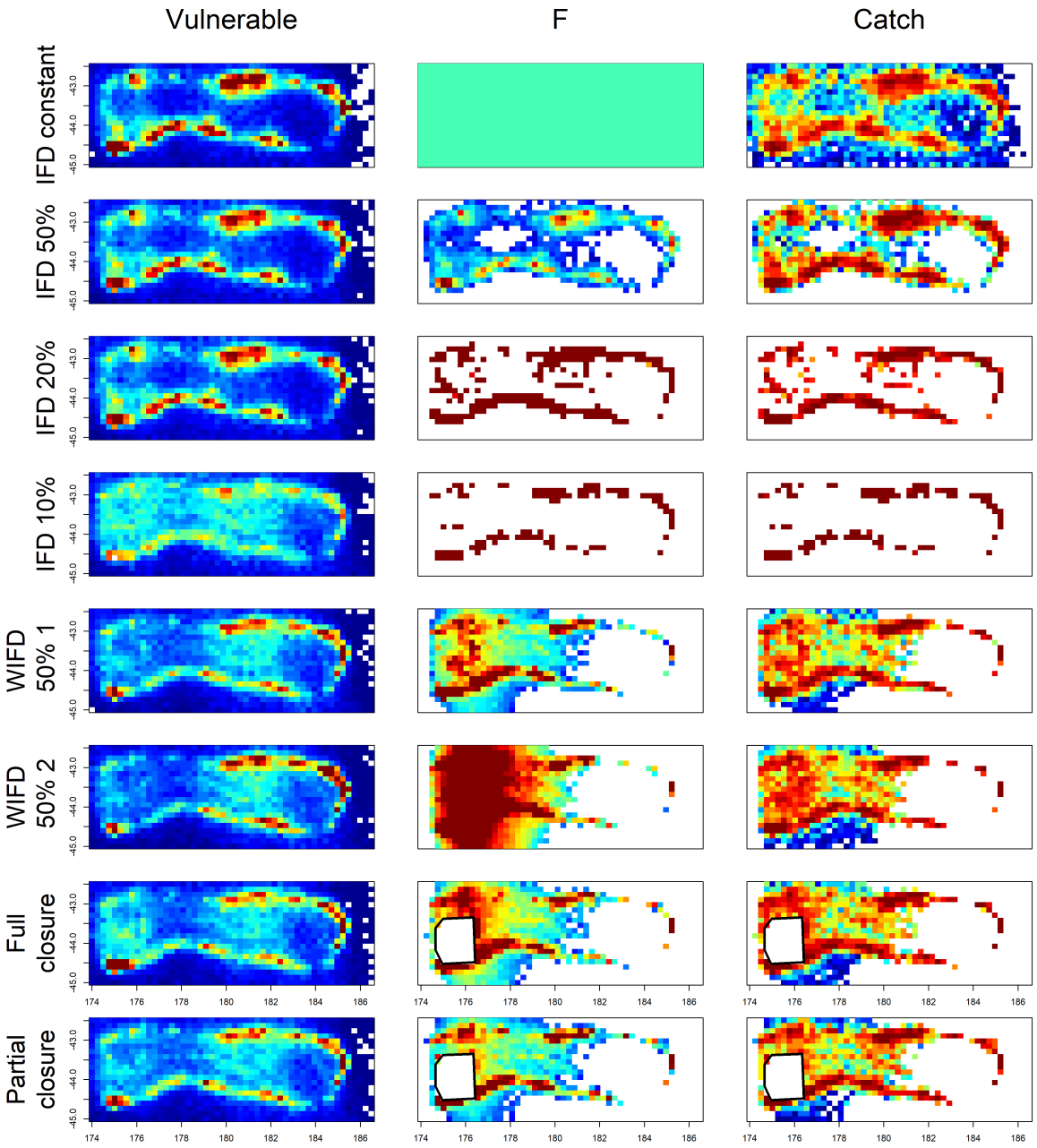


Figure 6.14: Spatial distributions from one realisation for the year 2010 of vulnerable biomass (vulnerable), fishing mortality (F) and catch for all eight CH-CABM OMs. Blue coloured cells indicates low values and red\maroon indicates high values.

The objective of this simulation study was to investigate the PS model. This section outlined how eight CH-CABM OMs applied spatially varying fishing mortalities and what the resulting spatial distributions of catch were (Figure 6.14). Additional

steps were required to convert CH-CABM OM output into simulated fishing event-level catch and effort data sets (see Section 6.4). The core assumption during these additional steps was that the expected number of fishing events simulated for a cell, was a function of catch in that cell. More specifically, the larger the catch in a cell, the more fishing events were simulated in that cell to remove that catch and vice versa.

The PS model assumes sampling intensity (number of fishing events) is a function of the underlying vulnerable abundance. This assumption can be visually inspected for a given OM by comparing the spatial distribution of catch with the spatial distribution of vulnerable biomass in Figure 6.14. When the spatial distributions between catch and vulnerable biomass are similar (i.e., have high values in the same cells and low values in the same cells), then this suggests agreement with the PS model assumptions. Figure 6.14 shows the IFD scenarios look to be consistent with the PS model assumptions, although it becomes difficult when there is less fishable area. This is in contrast to the WIFD and spatial closure scenarios, which look to deviate in some regions.

6.3.6 Initialisation

Model years and initial parameters used in the CH-CABM OMs are outlined in Table 6.6. The CH-CABM OMs all used 500 000 agents to represent the stock, this assumption led to each agent representing 2 300 individual fish. This was chosen for computational convenience in order to conduct simulations in a reasonable time frame.

Table 6.6: Initial parameter values.

Parameter	Value
n_{agents}	500 000
model years	1972-2013
ages	1-17 ($a_{min} - a_{max}$)
$\tilde{n}$	2300
n_{init}	40
R_0	Section 6.3.4
M^*	0.296 (Section 6.3.2)

6.4 Simulating catch and effort data from the CH-CABM OMs

The previous sections described eight CH-CABM OMs that assumed different fishing scenarios loosely based on the Chatham Rise hoki fishery. Each of the eight CH-CABM OMs generated 100 realisations of vulnerable biomass, fishing mortality and spatial catch distributions (Figure 6.14) over time. Due to the stochastic nature of the agent dynamics, each realisation resulted in a different random number sequence and in turn, resulted in different outcomes for agents and thus different spatial distributions of abundance and fishing.

Geostatistical models and the PS model require fishing event-level data whereas the CH-CABM OMs produced catch and vulnerable biomass at the spatial resolution of the OMs (20 x 50 cells). To convert the CH-CABM OM output to fishing event-level data an additional algorithm was applied. This algorithm used spatial catch and vulnerable biomass from the CH-CABM OMs to simulate high spatial resolution catch and effort data sets. The algorithm was only applied for a subset of OM years (2000 to 2013). This was done for computational convenience, but does mimic scenarios where high spatial resolution fishery dependent catch and effort data are only available for later years of the fishery time-series, i.e. Langley & Bentley (2014).

The algorithm for generating fishing event-level data sets for a given year used spatially explicit vulnerable biomass denoted by V_r, density $d_r = \frac{V_r}{A_r}$ (where A_r is the area of cell r) and catch removed C_r. The core assumption of the algorithm (described in detail below) was that cells that had large catches, had more fishing events and cells with smaller catches had less fishing events.

This was implemented by generating an area swept for each fishing event using an average expected catch. This resulted in cells of low densities having fishing events with large area swept. The fishing event algorithm followed:

- Iterate over all cells in the domain,
  - for cell r set $\tilde{C}_r = C_r$, $\tilde{d}_r = d_r$,

    While $\tilde{C}_r > 0$

    1. Generate an area swept, denoted by a_k for fishing event k based on an average catch $\bar{c} = 5.5$ if $\bar{c} > \tilde{C}_r$ then $\bar{c} = \tilde{C}_r$

    $$a_k = \frac{\bar{c}}{\tilde{d}_r}$$

    2. Simulated catch for fishing event k was generated using density in the cell, area swept and a dispersion parameter from the gamma distribution,

    $$\theta_k = \frac{a_k \tilde{d}_r}{\phi}$$
    $$c_k \sim \Gamma(\theta_k, \phi)$$

    d_r is the density using the updated vulnerable biomass and $\phi = 2$ is the shape parameter (roughly equates to a CV$\approx$ 0.7).

    3. Record the following attributes for each fishing event; catch, area, latitude and longitude. Latitude and longitude were assigned values based on a uniform draw between cell boundaries.

    4. Remove catch from cell $\tilde{C}_r = \tilde{C}_r - c_k$ and $\tilde{V}_r = \tilde{V}_r - c_k$ and $\tilde{d}_r = \tilde{V}_r / A_r$

Step 1 and 2 introduced two sources of variation. The variation from the Gamma distribution, and secondly the depletion in density within each cell as catch is removed during the algorithm. An illustration of spatial distributions of fishing events generated for each CH-CABM OM is given in Figure 6.15.

Alternative fishing event algorithms were initially explored that applied a constant area swept approach. The constant area approach resulted in having moderate to high levels of fishing events in cells with low densities and moderate catches. This reflected an avoidance behaviour which was not thought to be realistic.

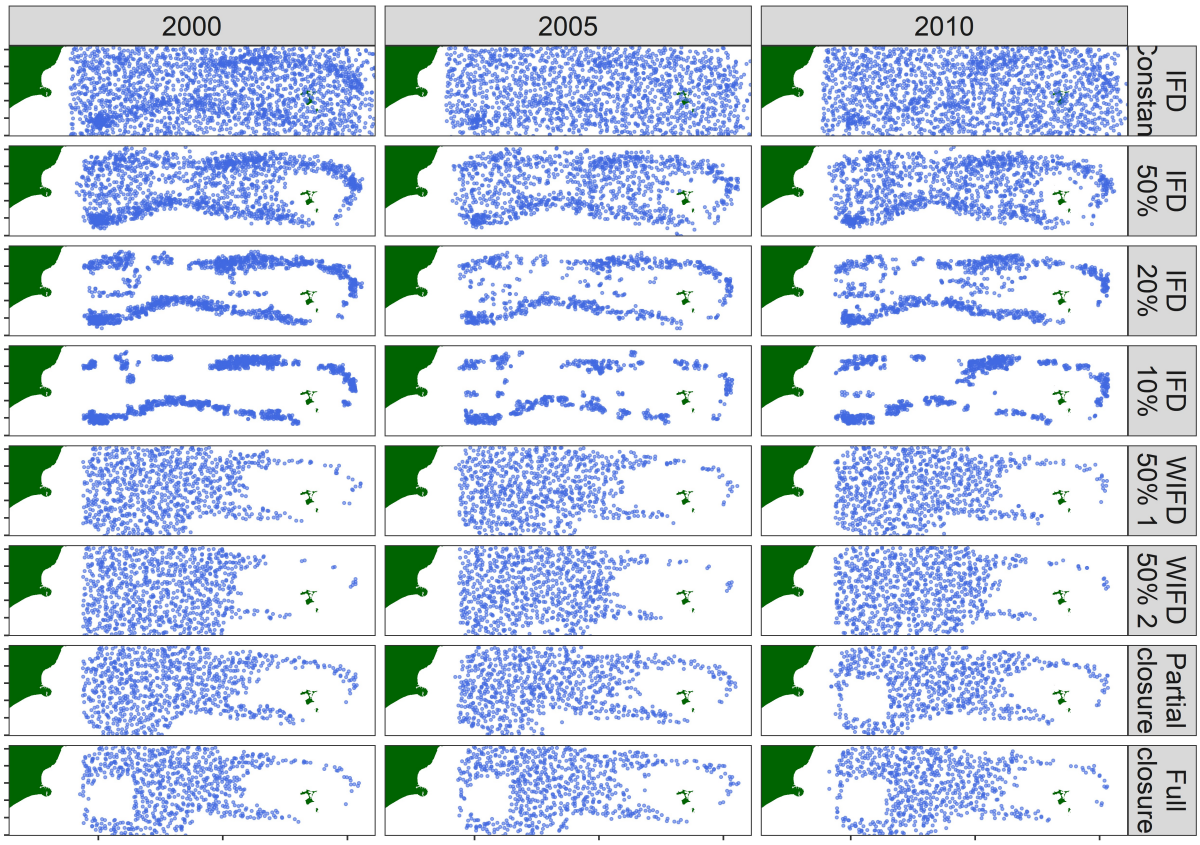


Figure 6.15: Simulated fishing event locations for one realisation of the eight OMs for the years 2000, 2005 and 2010 across the eight fishing scenarios.

6.5 Estimation models

The previous sections described how 800 simulated catch and effort data sets were simulated (eight CH-CABM models, where each was run 100 times). Each simulated catch and effort data set was split into a training and testing data set. Ten percent of fishing events within each year were randomly withheld for a testing data set. The remaining ninety percent of simulated fishing events were used as the training data set for parameter estimation.

Five EMs were fitted to each simulated training data set (Table 6.7). These models varied in GF and PS assumptions. All EMs assumed the response variable (catch) was drawn from the gamma distribution,

$$y_i \sim \Gamma(\theta_i, k) \ , \tag{6.1}$$
$$\theta_i = \frac{a_i exp(\eta_i)}{k} \ ,$$

where, k and θ_i are the shape and scale parameters of the Gamma distribution respectively, a_i is area for the i^{th} observation and η_i is the systematic component described in Table 6.7. The only covariate in the systematic covariate was year $X_i^t \boldsymbol{\beta^t}$ in addition to GF assumptions.

Table 6.7: Estimation models applied to each set of simulated data.

Model label	Systematic component	Intensity function (PS)
$\mathcal{M}_0$	$\eta_i = X_i^t \boldsymbol{\beta^t}$	-
$\mathcal{M}_1$	$\eta_i = X_i^t \boldsymbol{\beta^t} + \omega(s_i)$	-
$\mathcal{M}_2$	$\eta_i = X_i^t \boldsymbol{\beta^t} + \epsilon(s_i, t_i)$	-
$\mathcal{M}_3$	$\eta_i = X_i^t \boldsymbol{\beta^t} + \omega(s_i)$	$\lambda(s_i) = exp\{\beta^{pref}(\omega(s_i))\}$
$\mathcal{M}_4$	$\eta_i = X_i^t \boldsymbol{\beta^t} + \epsilon(s_i, t_i)$	$\lambda(s_i, t_i) = exp\{\beta^{pref}(\epsilon(s_i, t_i))\}$

GF assumptions followed the model

$$\boldsymbol{\omega} | \boldsymbol{\theta_\omega} \sim GF(\mathbf{0}, \Sigma_\omega) \ ,$$
$$\boldsymbol{\epsilon_t} | \boldsymbol{\theta_\epsilon} \sim GF(\mathbf{0}, \Sigma_\epsilon) \ ,$$

where both Σ_ϵ and Σ_ω were assumed to have an isotropic Matérn covariance.

Each of the fishing scenarios resulted in a different spatial distributions of fishing (Figure 6.15) over the domain. This led to different mesh assumptions being applied over the spatial domain for each fishing scenario, shown in Figure 6.16. Estimated indices of relative abundance were derived by summing extrapolated spatial abundance within the mesh boundary. The mesh boundary is indicated by the blue line in Figure 6.16. This meant for some fishing scenarios parts of the CABM spatial domain were not extrapolated to. This was also done when calculating prediction error for testing data sets.

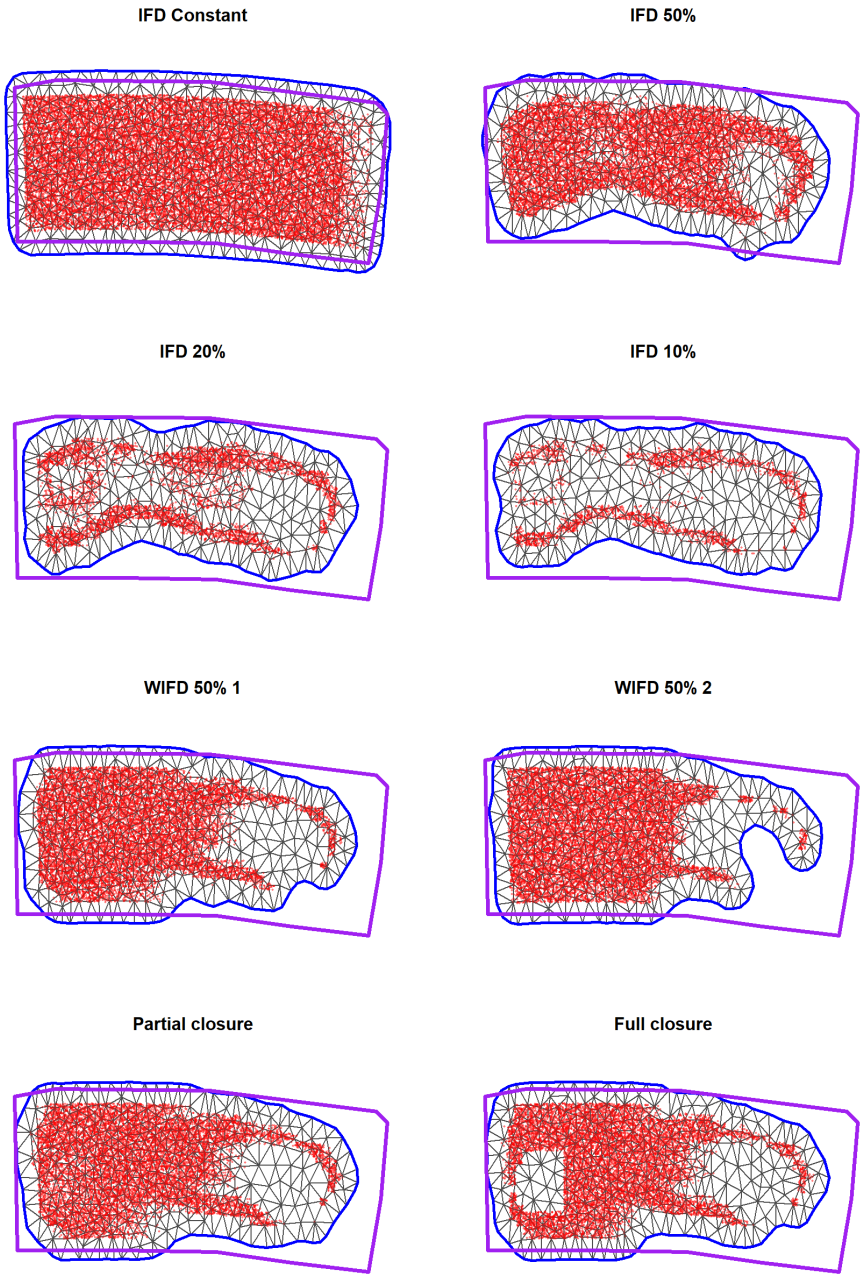


Figure 6.16: Mesh assumptions for the eight fishing scenarios. Purple polygon indicates CH-CABM OM boundary. Blue polygon indicates boundary for mesh. Red dots are fishing locations for one realisation.

An additional testing data set was derived using true density values from the OM for each cell. For all years in the simulation, densities of all cells were extracted from

each CH-CABM OM. The density in cell r, d_r, was calculated by

$$d_r = \frac{V_r}{A_r} \ ,$$

where A_r is the area of the cell.

The root mean squared error (RMSE) was used to evaluate prediction error for both test data sets (10% of the simulated data and the OM derived densities) in catch rates among EMs and OMs.

6.6 Results

The most important factor for PS model convergence (of the factors explored in the simulation) was the proportion of area available to fishing. Fishing scenarios IFD 10% was an extreme case of this factor which resulted in low convergence rates for both PS models (Table 6.8). Almost all the other fishing scenarios resulted in a high percentage of successfully converged models across all EMs.

Table 6.8: Percentage of EMs that successfully converged[1] for each fishing scenario.

	$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$
IFD Constant	100	100	100	99	100
IFD 50%	100	100	100	96	100
IFD 20%	100	100	99	100	100
IFD 10%	94	94	95	7	6
WIFD 50% 1	100	100	100	100	100
WIFD 50% 2	100	100	100	97	98
Partial closure	98	98	98	100	100
Full closure	100	100	100	100	100

[1] Model convergence based on Section 2.4.

For each OM, the first simulated catch and effort data set was used to conduct negative log-likelihood profiles. The profiles were with respect to the β^{pref} parameter for EM $\mathcal{M}_3$ and $\mathcal{M}_4$. These profiles are shown in Figure 6.17 for $\mathcal{M}_3$ and Figure 6.18 for $\mathcal{M}_4$. These show that for a given fishing scenario there was little difference in the profile shape between PS model $\mathcal{M}_3$ and $\mathcal{M}_4$. They also highlight that fishing scenario IFD 20% and IFD 10% had profile shapes that closely resembled the log-likelihood profile observed in the Chatham Rise hoki fishery case study (Section 4.7, Figure 4.12).

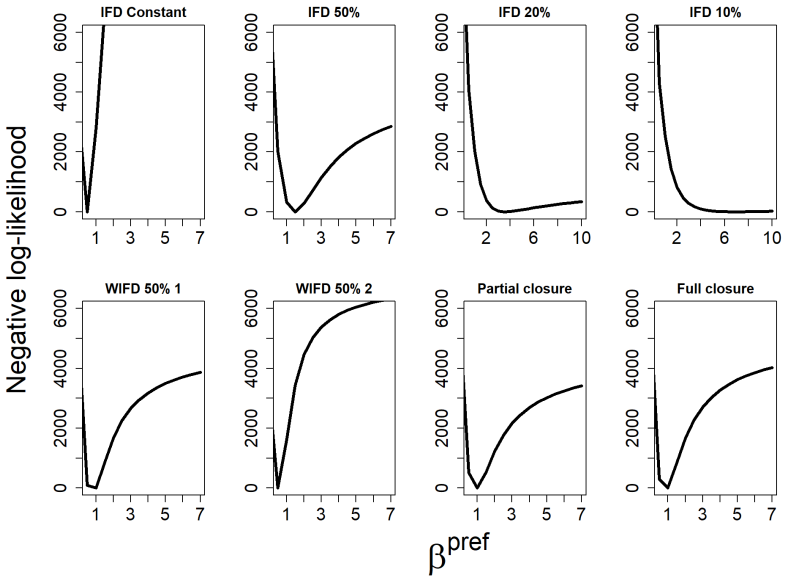


Figure 6.17: Negative log-likelihood profiles for β^{pref} from EM $\mathcal{M}_3$.

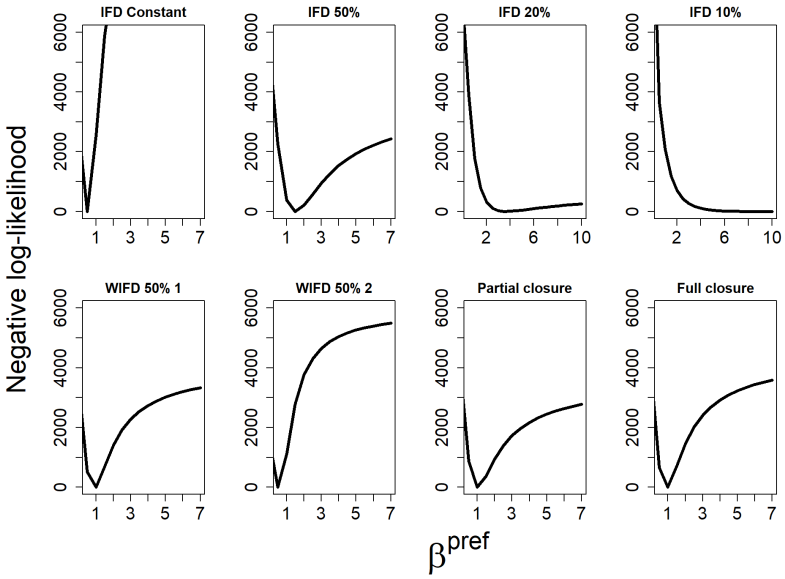


Figure 6.18: Negative log-likelihood profiles for β^{pref} from EM $\mathcal{M}_4$.

RMSE of catch rates using the two testing data sets (Section 6.4) are summarised in Table 6.9 and Table 6.10. The testing data set that contained 10% of the simulated

data sets that was withheld during estimation showed that on average, EM $\mathcal{M}_1$ had the lowest prediction error followed by EM $\mathcal{M}_3$ (Table 6.9).

Results from prediction error in catch rates using the testing data set based on densities at cell midpoints from the CH-CABM OMs (Table 6.10), showed, that on average EM $\mathcal{M}_3$ had the lowest RMSE followed by EM $\mathcal{M}_1$.

Table 6.9: Percentage of EMs that had the lowest RMSE by OM and EM. The testing data set was of 10% of simulated data withheld.

	$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$
IFD Constant	0	43	1	55	1
IFD 50%	0	98	0	2	0
IFD 20%	0	94	2	4	0
IFD 10%	9	86	5	0	0
WIFD 50% 1	0	53	3	44	0
WIFD 50% 2	0	50	4	43	3
Partial closure	0	58	0	42	0
Full closure	0	59	3	38	0

Table 6.10: Percentage of EMs that had the lowest RMSE by OM and EM. The testing dataset used densities from each CH-CABM OM cell midpoint (whether sampled or not).

	$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$
IFD Constant	0	100	0	0	0
IFD 50%	0	0	0	100	0
IFD 20%	0	0	0	97	3
IFD 10%	6	89	3	1	0
WIFD 50% 1	0	0	0	100	0
WIFD 50% 2	11	1	0	86	2
Partial closure	0	0	0	100	0
Full closure	0	0	0	100	0

Estimated spatial distributions for the full spatial closure fishing scenarios are shown in Figure 6.19 for the year 2010. This Figure demonstrated that the PS model had the expected and undesirable effect of estimating little to no abundance in the closed area due to the lack of sampling induced by the closure.

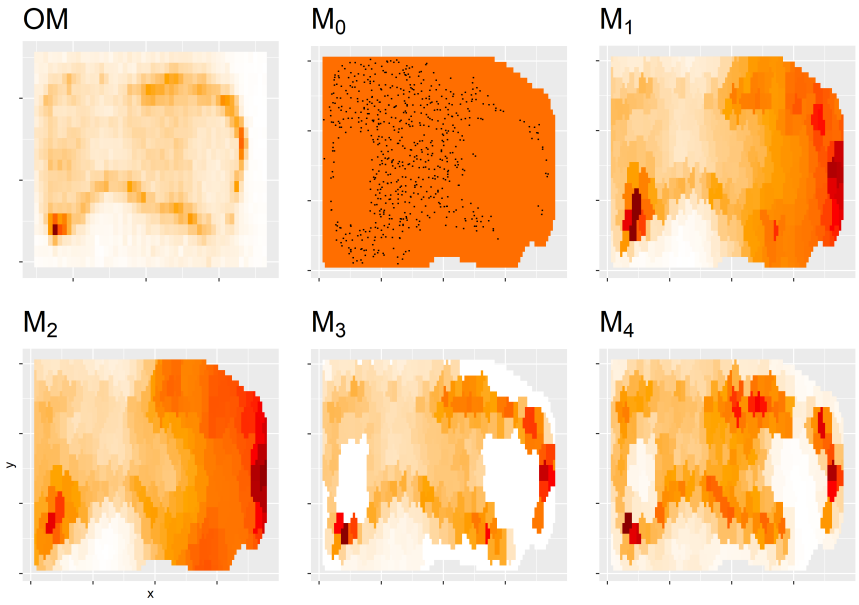


Figure 6.19: Spatial distributions of abundance for the year 2010 from the fishing scenario Full Closure. Black dots in panel labelled $\mathcal{M}_0$ indicate fishing locations for that year.

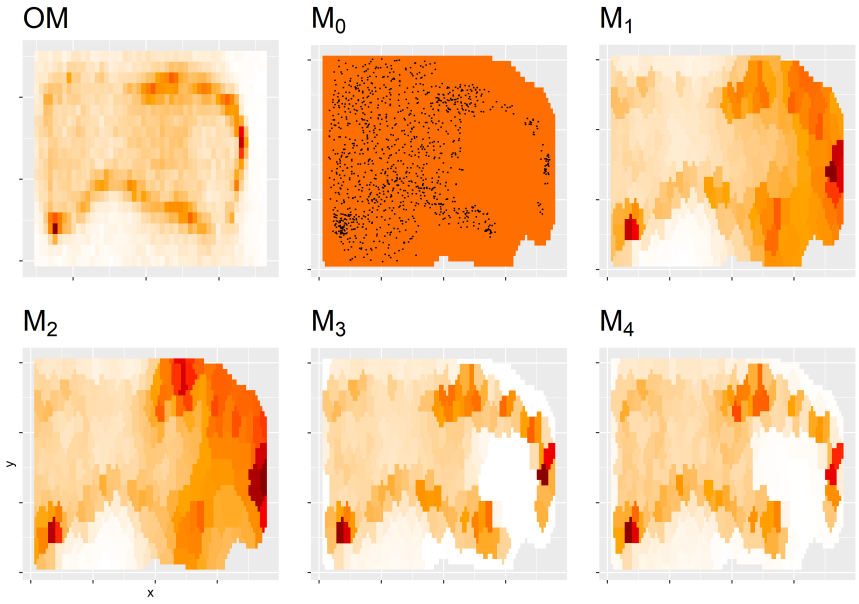


Figure 6.20: Spatial distributions of abundance for the year 2010 from the fishing scenario WIFD 50% 1. Black dots in panel labelled $\mathcal{M}_0$ indicate sample locations for that year.

Estimated spatial distributions are also shown for fishing scenario WIFD 50% 1 in Figure 6.20 for the year 2010. This scenario constrained the fishery to be close to land with fishing events that occurred far to the right of the spatial domain occurring in areas of high abundance. EM $\mathcal{M}_2$ was sensitive to these fishing events to the far right estimating high abundance to the right of the domain. Whereas the PS models predicted unsampled space with almost no abundance, which was not reflective of the abundance in the OM (Top left panel of Figure 6.20).

6.7 Discussion

We applied CABM to emulate hoki spatial distributions and productivity assumptions on the Chatham Rise. The simulation study identified that the proportion of area fished was the most important factor (of the factors explored) for obtaining high convergence rates for the PS model. Fishing scenarios that had highly concentrated fishing with respect to space (i.e. fishing scenario IFD 10%) resulted in the PS model having low convergence rates and log-likelihood profiles for β^{pref} that sloped towards the upper boundary constraint. This result was consistent with the recent simulation study by Ducharme-Barth et al. (2022), who also identified spatial sampling coverage to be a factor that degraded geostatistical model inference.

The hoki Chatham Rise fishery case study and CH-CABM OM IFD 10% both exhibited highly concentrated fishing around the main depth contours, with very little fishing and hence sampling of catch rates in surrounding regions. This lack of sampling in surrounding regions resulted in little information to estimate β^{pref}, indicated by flat log-likelihood profiles and β^{pref} being estimated at boundary constraints. This highlighted the need for a level of sampling in areas of low to moderate abundance when applying the PS model.

The Chatham Rise hoki fishery has a wide spatial extent but spatially concentrated fishing. This was assumed to be a desirable attribute when applying the PS model. It turns out that when fishing is this spatially concentrated it becomes problematic for the PS model to estimate β^{pref}. One future extension of the PS model is to extend the catch rate observation model to include zero observations (i.e. delta-model (Thorson et al. 2015) or Poisson-log link model (Thorson 2018)) or consider adding a small constant to zero observations (e.g. as noted by Maunder &

Punt (2004)). This would allow more observations from fishing in regions of low to moderate abundance and possibly resolve the above limitation for the PS model.

The simulation study showed that geostatistical EMs with time-invariant GF assumptions ($\mathcal{M}_1$ and $\mathcal{M}_3$) had lower prediction error in catch rates compared with geostatistical EMs with time-varying GF assumptions ($\mathcal{M}_2$ and $\mathcal{M}_4$). This result is likely driven by the assumed movement dynamic in the OM. Agents moved around the domain based on depth which is a static covariate over time. This may mean results from this simulation are only relevant to non-transient species. Future research should consider extending this simulation to transient species whose movements are linked with more seasonal or time-varying environmental variables, i.e. billfish and temperature fronts (Bigelow et al. 1999).

Two results from this simulation were consistent with the Chatham Rise hoki application (Chapter 4). The first being, when testing data sets used to evaluate prediction error in catch rates were a random sample of the simulated data. Conventional geostatistical models were concluded to perform better than the PS model using the RMSE statistic. However, when the testing data set was derived from midpoints of the OM, the PS model was generally favoured over conventional geostatistical models. The later testing data set was more uniformly distributed through space and demonstrated the expected benefits in out-of-sample predictions from the PS model.

The second result was how influential the point process was in predicted spatial distributions. Any OM scenario that contained patches of space that was not sampled (i.e. closed areas and WIFD fishing scenarios), the PS model interpreted as very low abundance. Caution should be made when using the PS model when there are shifts in sampling spatial distributions that are not due to expected catch rates. As mentioned in the case study this could be resolved by exploring alternative point process likelihoods or exploring the data weighting concept. Another possible solution is to exclude areas that are not fished due to factors such as closures i.e. using a physical barrier spatial mode i.e. Bakka et al. (2019). This latter suggestion will only be suitable if the areas are closed to fishing for the entire time-series.

One limitation of this simulation study and simulation studies from chapter 3, is they did not explore misspecification in the linear predictor of catch rates. There have been many documented pitfalls in analysing fishery-dependent catch and effort data when estimating abundance indices, these include; systematic shifts in fishing

effort (Fonteneau et al. 1999), increase in fishing efficiencies over time (effort-creep) (Maunder et al. 2006, Ducharme-Barth et al. 2022), local depletion causing non-linear relationships between CPUE and abundance (Harley et al. 2001*a*, Abraham & Neubauer 2015) and more. We would expect both geostatistical models and the PS model to be sensitive to departures in this assumption and should be considered in future research.

Chapter 7

Estimators for two-dimensional systematic surveys

7.1 Overview

Previous chapters explored methods for estimating abundance from non random data due to preferential sampling. Two-dimensional systematic surveys are another sampling method employed to estimate the abundance of spatial populations that also departs from random sampling.

This chapter investigates estimators for total abundance and its associated variance using data collected from a systematic survey with only a single primary sampling unit (PSU). This survey design randomly selects the location of the first sampling unit, from which, the location of all other sampling units are deterministically specified. This lack of random sampling means there is no exact design-based variance estimator for the population total.

Section 7.3 reviews a range of approximated design-based variance estimators in addition to the under utilized semi-parametric variance estimator proposed by Fewster (2011). These estimators rely on the design-based point estimator for the population total when calculating confidence intervals. We also explore the utility of a geostatistical model-based estimator for the population total and variance in Section 7.3.5.

Simulations are conduced in Section 7.4 to compare confidence interval coverage among estimators, in addition to exploring properties of the geostatistical model-based estimator.

7.2 Introduction

Two-dimensional systematic surveys are a commonly applied survey design for estimating the abundance of spatial populations (Millar & Olsen 1995, Fewster 2011). Systematic surveys enforce a uniform sampling pattern over the entire spatial domain. Unlike random sampling and preferential sampling (Chapter 3), systematic surveys avoid over and under sampling space.

Systematic surveys can have the advantage over random sampling strategies of being more practical to implement, and in simulation studies, the estimated population density has been shown to provide considerably lower variance than random designs under many conditions (e.g., McGarvey et al. 2016). However, the most common systematic survey design consists of just a single primary sampling unit (PSU) and consequently there is no applicable design-based estimator of this variance. This has led to the development and application of several approximations based on variations of design-based variance estimators designed for random sample designs (Wolter 1984, D'Orazio 2003).

This chapter reviewed a range of approximated design-based variance estimators that have been applied to two-dimensional systematic surveys. These variance estimators all use a design-based point estimator for the population mean or total when deriving confidence intervals. In addition to reviewing existing estimators, we explore the utility of a geostatistical model-based estimator for the total population and its variance. The geostatistical model proposed (Section 7.3.5) applied the SPDE approach with Matérn covariance (Lindgren et al. 2011) for spatial Gaussian Fields. This approach is a point of difference from other geostatistical model-based estimators applied to systematic surveys, i.e. (Simmonds & Fryer 1996, Walline 2007, Aune-Lundberg & Strand 2014). In these geostatistical model applications, kriging equations were used for spatial model inference.

Previous chapters have shown the value of geostatistical model-based estimators for spatial population inference. There is also increasing literature demonstrating how geostatistical model-based estimators are efficient estimators for population totals when applied to random survey designs (Shelton et al. 2014, Thorson et al. 2015). These recent studies coupled with advancements in spatial Gaussian Field (GF) methods (Lindgren et al. 2011, Shelton et al. 2014) and statistical software for efficient inference of complex spatial hierarchical models (Kristensen et al. 2016,

Krainski et al. 2018), are why we believe geostatistical model-based estimators may outperform approximated design-based estimators.

Recent comparison studies of variance estimators for systematic surveys with a single PSU have lacked model-based estimators (McGarvey et al. 2016, Strand 2017, D'Orazio 2003). Exceptions include Aune-Lundberg & Strand (2014), Simmonds & Fryer (1996), who, as mentioned earlier explored an estimator based on the concept of semivariance (or kriging) and Bartolucci & Montanari (2006), who also explored a model-based estimator that assumes the spatial population are generated from a model with homoscedastic and uncorrelated errors. One reason provided for the lack of model-based estimators is they "are more difficult to code and implement than the simpler design-based systematic survey variance estimators" (McGarvey et al. 2016, pg 244).

A variance estimator under utilized in recent comparison studies is the boxlot estimator (also called the striplet estimator for one-dimensional systematic surveys) proposed by Fewster (2011). This is a hybrid estimator that describes a semiparametric variance estimator coupled with the design-based point estimator for the population total when deriving confidence intervals. We believe this method has been under utilized due to its computational complexity and lack of convenient software. This is an estimator that we believe may improve upon approximated design-based estimators. The original paper explored the boxlet estimator in simulations that assumed high spatial sampling coverage (4% and 25% spatial coverage). Simulations in later sections (Section 7.4) explore its utility using simulations assuming much lower spatial sampling coverage.

This chapter extended upon recent comparison studies by including the boxlet estimator and exploring a geostatistical model-based estimator using the SPDE approach. The objective was to compare confidence interval coverage of the boxlet and geostatistical model-based estimator with existing methods and continue the search for an optimal estimator for this survey design.

7.3 Estimation methods

Consider a spatial survey region $\mathcal{D} \subset \mathbb{R}^2$ of area A. For simplicity i indexes a sampling unit (quadrat) that is spatially referenced by coordinates in $\mathcal{D}$. A two-dimensional systematic survey that is based on a single PSU requires the specification of the

location of the first sampling unit, from which the location of all other sampling units are deterministically specified. The response variable in this setting is either numbers (abundance) or weight (biomass) in each of the n sampling units denoted by $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$.

The population characteristic of focus is the total population within the survey region, denoted by N, or equivalently the mean density over $\mathcal{D}$, $d_{\mathcal{D}} = N/A$. If α is the area of each sampling unit then $y_{\mathcal{D}} = \alpha d_{\mathcal{D}}$ is the expected mean value. Then, $N = \kappa y_{\mathcal{D}}$ where $\kappa = A/\alpha$ is the inverse of the sampling fraction. Design-based methods applied to systematic surveys all use the sample mean estimator of $y_{\mathcal{D}}$, $\bar{y} = \sum_{i=1}^{n} y_i/n$, and hence the estimator $\widehat{N} = \kappa\bar{y}$ of N, with variance estimator

$$\widehat{\text{var}}\left(\widehat{N}\right) = \kappa^2 \widehat{\text{var}}(\bar{y}) \ .$$

In systematic surveys with multiple PSUs the systematic design-based variance estimator ($\widehat{\text{var}}_{sys}(\bar{y})$) is known (Madow & Madow 1944) and follows,

$$\widehat{\text{var}}_{sys}(\bar{y}) = \frac{1}{B}\sum_{b=1}^{B}(\bar{y}_b - \bar{y})^2$$

where, $b = 1, \ldots, B$ are the PSUs, each assumed to contain the same number of sampling units, and y_b is the sample mean within PSU b. However, B is typically a relatively small number, so this estimated variance generally has few degrees of freedom.

For the systematic design of interest in this study there is just a single PSU (the position of the first quadrat) and hence no design-based variance estimator exists. For this reason, as shown below, some authors have explored the use of the simple random sampling (SRS) variance estimator, ad-hoc estimators utilizing stratified random survey variance estimators (Millar & Olsen 1995, Strand 2017) and adjusted SRS variance estimators (Ambrosio Flores et al. 2003, Strand 2017, Brus & Saby 2016, McGarvey et al. 2016).

This study focused on survey designs that have low sampling spatial coverage. Due in part to our focus on fisheries surveys which often sample a small proportion of the spatial domain, i.e. Strømme & Iilende (2001), Millar & Olsen (1995), Simmonds & Fryer (1996). This results in the finite population correction being negligible and is excluded in subsequent estimator descriptions.

7.3.1 Simple Random Sample (design-based)

The simplest approach is to treat the data as if all points were collected from a SRS design, and to use the usual SRS variance estimator,

$$\widehat{\text{var}}_{srs}(\bar{y}) = \frac{s_y^2}{n} = \frac{1}{n(n-1)} \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad . \tag{7.1}$$

Several simulation studies have demonstrated that this estimator tends to over-estimate the true variance, and the over-estimation can be substantial when positive autocorrelation exists (Fewster 2011, Dunn & Harrison 1993).

7.3.2 Post-stratification (design-based)

This class of estimators was proposed by Wolter (1984) and has been extended to systematic spatial surveys (e.g., Dunn & Harrison 1993, Fewster 2011, Millar & Olsen 1995). Each stratum contains a fixed number of neighbouring sampled units from the systematic design, and the stratified random sample variance estimator of $\bar{y}$ is used (Fewster et al. 2009) (this is ad-hoc because stratified designs assume simple random sampling within strata). Millar & Olsen (1995) provided a variation on this approach, by showing that it could be applied using strata that overlapped.

Consider H overlapping strata where $h = (1, 2, \ldots, H)$ index the strata, and let stratum h contain n_h sampling units (in this case $n_h = 4$ in Figure 7.1). Let the set of sampling units in strata h be denoted by the set S_h. Then the post-stratified estimate of variance is

$$\widehat{\text{var}}_{str}(\bar{y}) = \sum_{h=1}^{H} w_h^2 \frac{s_h^2}{n_h} \ , \tag{7.2}$$

where

$$s_h^2 = \frac{1}{n_h - 1} \sum_{i \in S_h} (y_i - \bar{y_h})^2, \text{ and } w_h = \frac{n_h}{\sum_h n_h} \ .$$

This formulation follows from Millar & Olsen (1995) and assumes overlapping strata. However, there are alternative overlapping and non-overlapping post-stratified variance estimators (see D'Orazio (2003)). These alternative estimators are due in part to their original application for one dimensional systematically sampled data (Wolter 1984). As discussed by D'Orazio (2003), extending these one-dimensional estimators to two-dimensions is not straightforward.

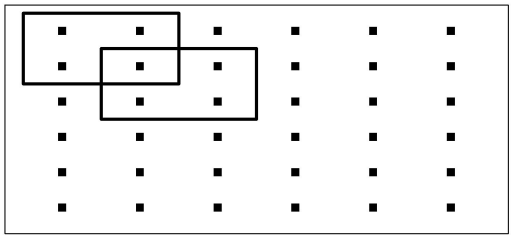


Figure 7.1: Visualising post-stratification of systematically sampled data. Points indicate sampling locations and black polygons represent overlapping strata.

7.3.3 Correlation adjustment (hybrid)

Defined in Wolter (1984) and investigated for spatial systematic samples (Ambrosio Flores et al. 2003, Strand 2017, Brus & Saby 2016, McGarvey et al. 2016), this estimator makes an adjustment to $Var_{srs}(\bar{y})$ (Equation 7.1) based on the amount of spatial auto-correlation that exists in the systematic sample.

$$\begin{aligned} \widehat{\text{var}}_{adj}(\bar{y}) =& \widehat{\text{var}}_{srs}(\bar{y})\left[1 + \frac{2}{ln(\rho)} + \frac{2}{(\rho^{-1} - 1)}\right] && \text{if } \rho > 0 \\ =& \widehat{\text{var}}_{srs}(\bar{y}) && \text{if } \rho \leq 0 \ . \end{aligned} \tag{7.3}$$

Here, ρ is often substituted for by an estimate of global spatial-autocorrelation such as the commonly used Morans I (Anselin 1995).

7.3.4 Boxlet estimator (hybrid)

The boxlet method proposed by Fewster (2011) fits a density surface to the spatial domain $\mathcal{D}$ using a two-dimensional generalised additive model (GAM) (Wood 2017). For any possible location, b, of the first quadrat, this enables calculation of the expected total count within the sampled quadrats, η_b and within $\mathcal{D}$, $\eta_{\mathcal{D}}$. This calculation is done using finite approximation over a fine lattice within $\mathcal{D}$ (these lattice points correspond to the centroids of the so-called boxlets). Let $p_b = \eta_b/\eta_{\mathcal{D}}$

denote the proportion of the expected total count that is expected within the sampled quadrats.

Let $y^+ = \sum_{i=1}^{n} y_i$ be the number of individuals counted in the systematic sample. The boxlet method partitions the variance of y^+ using the law of total variance (Equation 7.4). The first term ($\text{var}_b(E[y^+|b])$) is the variance in the expected value of y^+ due to the possible starting position, b, of the first quadrat. The second component ($E_b[\text{var}(y^+|b)]$) quantifies the expected (over b) magnitude of the variability in y^+ for a given b,

$$\text{var}(y^+) = \text{var}_b(E[y^+|b]) + E_b[\text{var}(y^+|b)] \ . \tag{7.4}$$

Under the assumption that the counts in each quadrat are Poisson distributed, then given b it follows that y^+ is binomially distributed according to a $\text{Bin}(N, p_b)$ distribution. This gives

$$\begin{aligned} \text{var}(y^+) &= \text{var}_b(Np_b) + E_b[Np_b(1-p_b)] \\ &= E_b[N^2 p_b^2] + E_b[Np_b]^2 + E_b[Np_b(1-p_b)] \ . \end{aligned} \tag{7.5}$$

The boxlet method estimates (7.5) by substituting N with $\widehat{N}$, and numerically evaluating the E_b and var_b terms (that is, with respect to variability in p_b) over all boxlet centroids within the scope of placement of the first quadrat. This leads to the boxlet variance estimator y^+ and hence for $\bar{y}$,

$$\widehat{\text{var}}_{box}(\bar{y}) = \frac{\widehat{\text{var}}_{box}(y^+)}{n^2} \tag{7.6}$$

7.3.5 Geostatistical model-based estimator (model-based)

The geostatistical model-based approach partitions the spatial domain $\mathcal{D}$ into $m >> n$ quadrats, with abundances (at the time of the survey) of $\boldsymbol{y} = \{y_i, i = 1, \ldots, m\}$, and hence $N = \sum_i^m y_i$. The abundances are assumed to be a realisation from a model,

$$\boldsymbol{y} \sim f\left(\boldsymbol{X}, \boldsymbol{\theta}\right) \ ,$$

with auxiliary covariates denoted by $\boldsymbol{X}$ and model parameters denoted by $\boldsymbol{\theta}$. This natural incorporation of auxiliary covariates is often viewed as an advantage for the model-based approach (Johnson et al. 2010, Chambers & Clark 2012, Ståhl et al. 2016, Shelton et al. 2014).

Let Ω be the index of the set of n sampling units selected by the systematic survey. The estimate of total population can be decomposed as

$$\widehat{N} = \sum_{i \in \Omega} y_i + \sum_{j \notin \Omega} E[y_j | y_i, i \in \Omega] \ . \tag{7.7}$$

The $y_i, i \in \Omega$ are known and so inference on un-sampled units $y_j, j \notin \Omega$ is the goal.

The proposed geostatistical model follows

$$y_i | \boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{u}, \phi \sim f(y_i | \mu_i, \phi) \qquad i = 1, \ldots, m \tag{7.8}$$

$$\boldsymbol{u} | \boldsymbol{\theta} \sim \mathcal{GF}(\mathbf{0}, \boldsymbol{\Sigma_u}) \tag{7.9}$$

where $\mu_i = g^{-1}(\boldsymbol{X_i^T}\boldsymbol{\beta} + u_i)$, where $\boldsymbol{X_i^T}$ is the covariate terms for observation i, $\boldsymbol{u} = (u_i, i = 1, \ldots, m)$ is a vector of random effects assumed to follow a Gaussian Field with isotropic Matérn correlation structure (Equation 2.8), and $g\,()$ is the link function mapping the linear predictor component to $E[y_i] = \mu_i$. Parameter ϕ denotes any required dispersion parameters for $f(\ )$, and $\boldsymbol{\theta}$ are hyper-parameters for the random effects.

Geostatistical models have been used with systematic surveys primarily in the acoustic literature (Walline 2007, Simmonds & Fryer 1996) but also in land cover use (Aune-Lundberg & Strand 2014). These studies used kriging equations for model inference, which differs from the SPDE approach (Lindgren et al. 2011) that is used for geostatistical models presented here (see (Gómez-Rubio 2020, Chatper 7) and Chang et al. (2015) for more detail on the difference between these two approaches). The SPDE approach was chosen for its flexibility in response variable distributions, flexibility in incorporating spatio-temporal GFs, ease of including covariates and availability of convenient software for its implementation, i.e. INLA (Lindgren & Rue 2015).

The software chosen to implement the spatial model (Equations 7.8 and 7.9), and estimator for the total population (Equation 7.7) and corresponding variance was Template Model Builder (TMB) (Kristensen et al. 2016). TMB was chosen for its ease of implementing bespoke models (Equation 7.7 only involves using a subset of fitted values) and computational speed (Muff et al. 2020) required for the purpose of performing 1 000's of simulations. TMB also has many inbuilt features that enhance its applicability. This includes automatic calculation of standard errors for

functions of parameters using Taylor series expansion (Section 2.4), which is commonly known as the generalised delta-method (Millar 2011, Fournier et al. 2012). TMB also includes the "epsilon" method of bias correction (Tierney et al. 1989, Thorson & Kristensen 2016), which approximates the integral that equates to the expectation of non-linear functions of random-effect variables. Thorson & Kristensen (2016) showed the "epsilon" method to be superior to other bias correction methods commonly used within fisheries models for estimating abundance. Using these estimators of variance, TMB will automatically report estimates of $SE\left[\widehat{N}\right]$.

For spatially variable populations it is often better to estimate and provide standard errors for $\log \widehat{N}$, which has generally been found to be more quadratic (Bolker et al. 2013). When $\log \widehat{N}$ is of focus, confidence intervals are

$$CI_{95\%} = exp\left(\log \widehat{N} \pm 1.96 \ \widehat{SE}\left[\log \widehat{N}\right]\right) \ . \tag{7.10}$$

Considerations when using model-based estimators include; assessing sensitivity to starting values, model convergence, model selection, goodness of fit, etc. As in all regression analysis, these are expected for any valid inferences.

7.4 Simulations

7.4.1 Description

The repeated simulation framework of Fewster et al. (2009) and Fewster (2011) was used to compare variance estimators, and investigate confidence interval coverage of variance estimators described in the previous sections. A repeated survey maintained the same number of individuals (focusing on abundance) denoted by N within the survey domain $\mathcal{D}$. The survey domain was assumed to be a square with x-axis and y-axis ranging from 0 to 100 (i.e. for location $s_i = \{x_i, y_i\}$, $x_i \in [0, 100]$ and $y_i \in [0, 100]$). Each repeated survey had two sources of variability, the first was due to the simulated population spatial distribution. Each repeated survey randomly generated a spatial population within the survey domain according to one of four spatial distributions (Figure 7.2). The second source of variation was due to sampling variability under survey replication. Each repeated survey generated a different PSU, resulting in different sampling unit locations.

None of the four spatial distributions explored (Figure 7.2) corresponded to the geostatistical model. Rather, we wanted to assess its performance under a variety of previously considered scenarios, i.e. Fewster et al. (2009), Johnson et al. (2010), Fewster (2011), McGarvey et al. (2016). Exploring multiple spatial population distributions will provide insight into robustness to departures from model assumptions, which is an important consideration for model-based estimators (Smith 1976).

The first spatial distribution was labelled "uniform" and assumed the population was distributed in the x and y directions based on independent marginal uniform distribution (also known as complete spatial randomness). The population labelled "beta" assumed the population was distributed with the following independent marginal beta distributions $x \sim Beta\,(4,1)\times 100$ and $y \sim Beta\,(1,4)\times 100$ (they are multiplied by 100 so the coordinates are consistent with survey domain). This distribution mimics a diffusing population from the lower-right corner of the survey domain. The remaining two spatial population were assumed to be drawn from an inhomogeneous Poisson point process (IPP) with intensity function based on a spatially varying habitat covariate. The R package `DSpat` (Johnson et al. 2014) was used to simulate these two different IPP spatial populations. They differed in how patchy the spatial distribution of the habitat covariate was (shown in Figure 7.2) but had the same intensity function for each level of the habitat covariate. The population labelled "IPP (diffuse)" contained large patches of habitat over the domain and "IPP (patchy)" consisted of smaller patches of habitat.

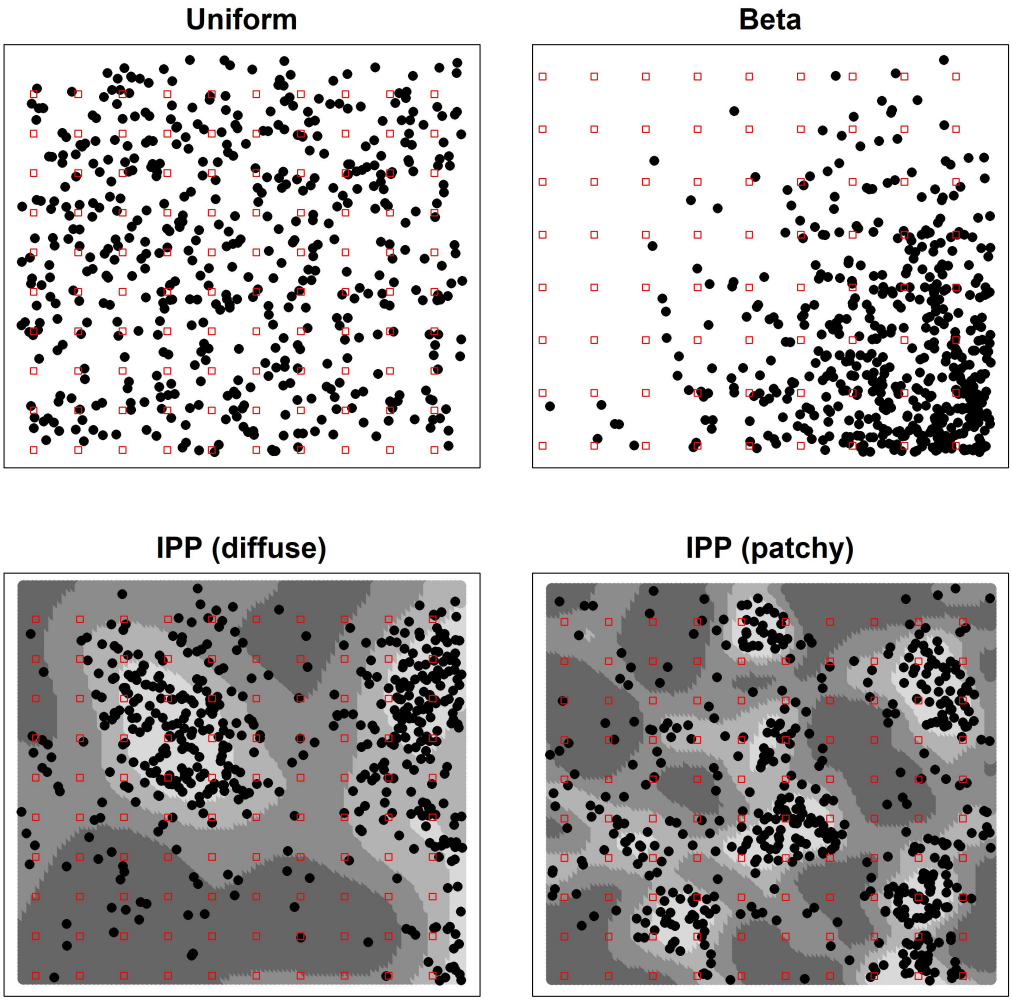


Figure 7.2: Examples of spatial distributions explored with $N = 500$ individuals. Black dots indicate individual locations, red boxes indicated sampling units and grey shades in IPP plots indicate different habitat.

For each spatial distribution two systematic sampling designs were explored. The designs differed in the area of individual sampling units and thus spatial coverage of the survey. The first sampling design (sample type 1) assumed sampling units were square quadrats with dimensions 1 x 1 (1 sampling unit covered 0.01% of the spatial domain). The second sampling design (sample type 2) assumed sampling units had dimensions 0.5 x 0.5 (1 sampling unit covered 0.0025% of the spatial domain). Both sampling designs assumed a 10 x 10 sampling units (100 sampling units) of equal

distance over the domain (illustrated in Figure 7.2). This resulted in sample type 1 having 1% spatial coverage and sample type 2 having 0.25% spatial coverage.

All spatial populations assumed $N = 1\ 000\ 000$ individuals in the survey domain. A second set of simulations was conducted with 500 000 individuals but the results were very similar. Due to the lack of contrast between these population sizes, we chose to present results for the $N = 1\ 000\ 000$ simulations.

This simulation resulted in eight different simulation scenarios (two sampling designs and four spatial distributions), where each simulation scenario was repeated 1 000 times.

The geostatistical model assumed the response variable was a Poisson random variable with the following assumptions,

$$\begin{aligned} y_i &\sim Poisson(\lambda_i) \ , \\ \lambda_i &= a_i exp\{\mu + u(s_i)\} \ , \\ \boldsymbol{u}|\kappa, \tau &\sim \mathcal{GF}\left(\mathbf{0}, \boldsymbol{\Sigma_u}\right) \ , \end{aligned}$$

where $\boldsymbol{\Sigma_u}$ is assumed to have an isotropic Matérn correlation structure. The mesh assumed during simulations was panel B in Figure 7.3. This assumed Neumann boundaries (Lindgren 2012), which increases the variance of the GF near the boundary.

The negative binomial distribution was explored as an alternative to the Poisson in early work but resulted in low convergence rates, this is discussed further in Section 7.5. Additional information is given in Appendix B regarding starting values and convergence criteria for the geostatistical model-based estimator. Models that did not converge were dropped from the results.

An additional sensitivity analysis was conducted for the geostatistical model for the IPP (diffuse) and IPP (patchy) spatial distributions with sample type 1. For these simulation scenarios, three geostatistical models were applied. The models differed in mesh assumptions, which are illustrated in Figure 7.3. This was to explore the sensitivity to this model assumption.

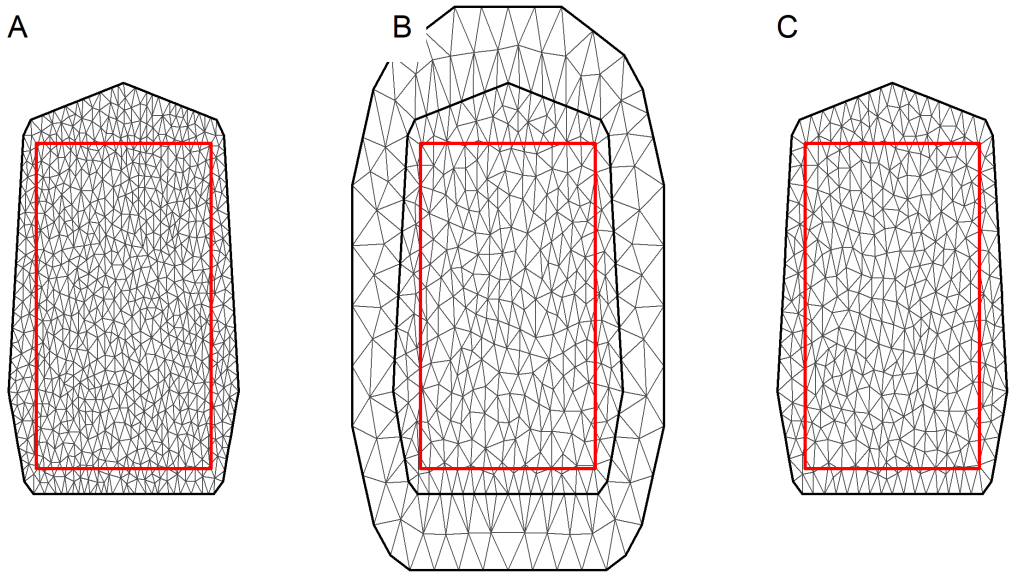


Figure 7.3: Three different mesh assumptions explored for the geostatistical model-based estimator. The red border represents the survey domain.

7.4.2 Results

Table 7.1 gives the 95% confidence interval coverage for all estimators and population scenarios. This table indicated there was no estimator that consistently obtained the desired confidence coverage across all scenarios.

In general, design-based approximations (SRS and STR) over-estimated the variance for spatially aggregating populations (IPP distributions). The Boxlet and geostatistical model-based estimators were more consistent across all spatial distributions. However, they did result in lower confidence interval coverage for spatially clumping populations than was desired (IPP distributions).

Table 7.1: 95% Confidence interval coverage expressed as a percentage.

Distribution	N	Sampling design[1]	SRS	STR	Adj	Boxlet	Model-based
Uniform	1 000 000	1	95	92	90	96	95
	1 000 000	2	95	92	90	95	95
Beta	1 000 000	1	98	42	46	94	99
	1 000 000	2	99	42	44	97	98
IPP (patchy)	1 000 000	1	100	100	98	78	89
	1 000 000	2	100	100	97	83	89
IPP (diffuse)	1 000 000	1	100	98	85	85	79
	1 000 000	2	100	99	86	88	84

SRS = Simple random sample (Section 7.3.1)
STR = Post stratified (Section 7.3.2)
Adj = Correlation adjusted (Section 7.3.3)
[1] 1 = sample type 1, 2 = sample type 2

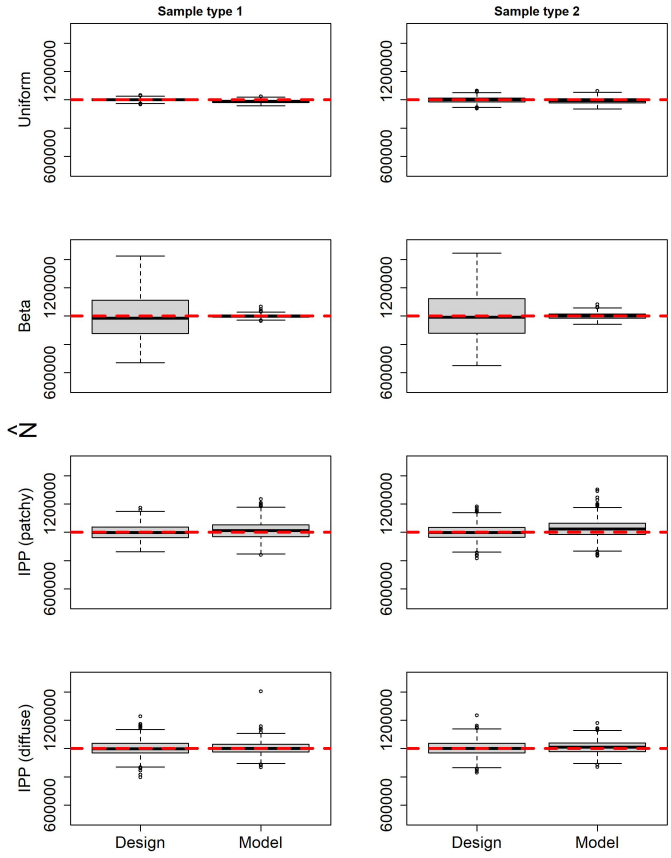


Figure 7.4: Estimates of total population abundance between the design and geostatistical model-based point estimators.

Table 7.2: Root mean squared error for point estimators of the population total for design-based and model-based.

Distribution	Sampling design	Design	Model
Uniform	1	20057.7	20086.9
	2	9997.5	9976.2
Beta	1	166487.7	61291.3
	2	166495.9	11399.5
IPP (patchy)	1	54276.1	72587.8
	2	52955.5	68275.7
IPP (diffuse)	1	54235.1	47958.9
	2	54877.3	45085.1

There were two results that stood out when the population was spatially distributed according to the beta assumptions. Firstly, the geostatistical model point estimator for total population ($\widehat{N}$) was more precise than the design-based estimator (Figure 7.4). Secondly, the correlation adjusted and post-stratified variance estimators both performed poorly for this spatial distribution (Table 7.1).

For both IPP spatial distribution scenarios, the geostatistical model-based estimator for N looked to have a small positive bias (Figure 7.4). However, only the IPP (patchy) spatial distribution resulted in the model-based point estimator having a larger root mean squared error than the design-based point estimator (Table 7.2).

The sensitivity analysis indicated the model-based point estimator for the IPP (patchy) spatial distribution was sensitive to alternative mesh assumptions (Figure 7.5). However, for the IPP (diffuse) spatial distribution, different mesh assumptions did not affect the model-based point estimator (Figure 7.6).

Table 7.3 gives the 95% confidence interval coverage for the mesh sensitivity analysis, which showed the coverage did differ among mesh assumptions for both spatial distributions explored. Mech C resulted in the best confidence interval coverage, which was surprising given Mesh A had a finer spatial resolution and mesh B had a barrier which was anticipated to be more appropriate given there was a physical boundary around the domain.

The estimated spatial distributions from the mesh sensitivity analysis for IPP (patchy) is shown for a single survey replication in Figure 7.7. This highlighted how similar the spatial predictions were among mesh assumptions.

Table 7.3: Confidence interval coverage of mesh sensitivities.

Distribution	Sampling design	Mesh A	Mesh B	Mesh C
IPP (patchy)	1	96	89	94
IPP (diffuse)	1	84	79	87

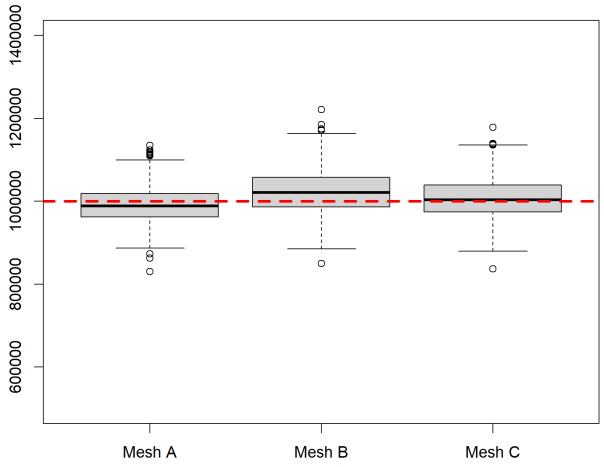


Figure 7.5: Point estimates of population total for mesh sensitivity analysis for IPP (patchy).

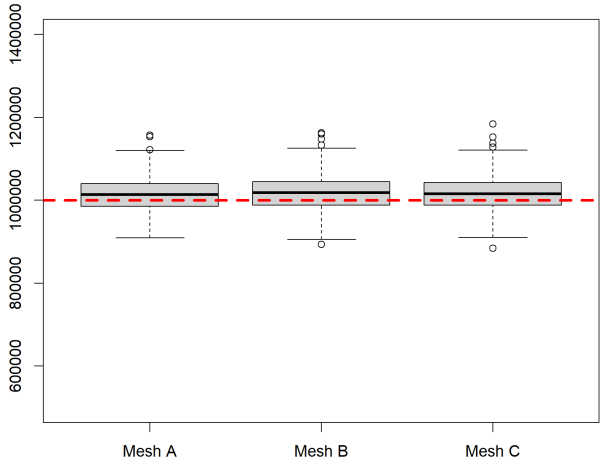


Figure 7.6: Point estimates of population total for mesh sensitivity analysis for IPP (diffuse).

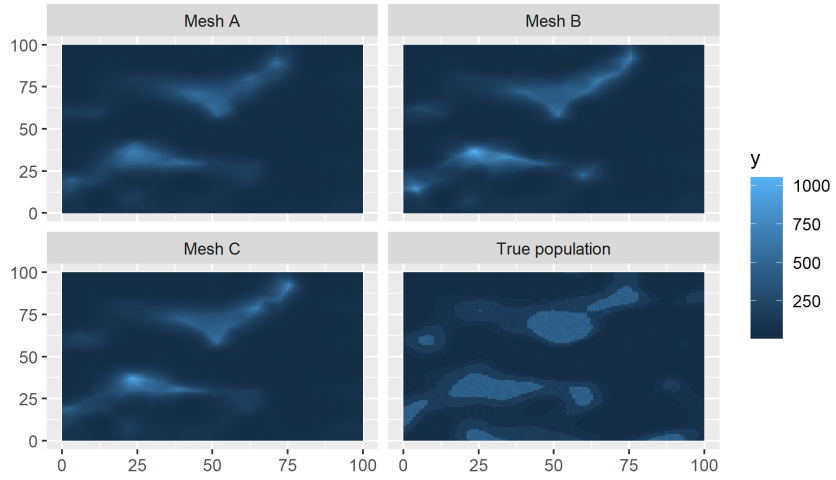


Figure 7.7: Spatial abundance compared with the mesh sensitivity analysis for IPP (patchy) spatial distribution.

Table 7.4 outlines the convergence rates of the geostatistical model-based approach. The uniform distribution resulted in the lowest convergence rate. This was due to the low spatial auto-correlation and the model having difficulty estimating hyper parameters of the Matérn covariance.

Table 7.4: Model convergence rates for the geostatistical model-based estimator. $C_{\%}$ is the percentage of models that successfully converged, and ST = sample type.

Dist	Uniform				Beta				IPP (patchy)				IPP (diffuse)			
ST	2		1		2		1		2		1		2		1	
N	5e5	1e6	5e5	1e6	5e5	1e6	5e5	1e6	5e5	1e6	5e5	1e6	5e5	1e6	5e5	1e6
$C_{\%}$	90.6	88.7	87.5	88.6	100	100	99.9	98.6	97.5	98.7	99.6	99.4	100	98.3	99.9	99.7

7.5 Discussion

This chapter conducted a review and comparison study of variance estimators for two-dimensional systematic surveys. By including the semiparametric boxlet method proposed by Fewster (2011) and exploring a geostatistical model-based estimator using the SPDE approach, this research extended previous comparison studies, which have predominately focused on design-based approximations (D'Orazio 2003, McGarvey et al. 2016).

Results from this comparison study did not identify any single best variance estimator for all spatial distributions explored. This result follows from previous

comparison studies who found an optimal variance estimator will depend on the population attributes such as periodicity and spatial correlation etc. (Aune-Lundberg & Strand 2014, McGarvey et al. 2016). However, there were valuable insights for some spatial distributions and estimators.

Almost all estimators produced the desired coverage for the uniform spatial distribution. This was considered a validation simulation scenario due to this spatial distribution being unlikely to occur in spatial populations. This was the only simulation scenario that the SRS estimator gave appropriate coverage. When populations were spatially varying, the SRS estimator consistently over estimated the variance which resulted in a conservative estimator.

In cases similar to the beta spatial distribution where the population is spatially distributed as a single concentration on or near the survey boundary, the design-based point estimator was less efficient than the model-based point estimator. This had consequences in the confidence interval coverage. In particular, for the post-stratified and correlation adjusted estimators which produced small estimates of variance. This, coupled with a variable point estimator, led them to have poor confidence interval coverage and should be avoided when the sampled population is spatially distributed in this manner. The boxlet estimator had the best confidence interval coverage for this spatial distribution, and is recommended even though it was using the design-based point estimator.

The IPP spatial distributions produced mixed results. The root mean squared error of the design-based point estimator was better for the patchy spatial distribution but the model-based point estimator was better for the diffuse spatial distribution. Although the model-based point estimator looked to be slightly biased, it was more efficient leading to a lower root mean squared error.

The post-stratified and correlation adjusted estimators produced similar results because they are similar estimators. D'Orazio (2003) showed how the post-stratified estimator with overlapping strata can be expressed as an adjusted SRS estimator based on serial correlation, which is analogous to the correlation adjusted estimators.

An innovation from this research was the development of the `systematicsurvey` R package found at `https://github.com/Craig44/systematicsurvey`. One of the reasons given for the lack of model-based variance estimators applied to systematic survey data is their difficulty to code and implement (McGarvey et al. 2016). This R package was an attempt to overcome this barrier and allow future users to easily

compare and contrast a range of variance estimators and encourage future research and exploration.

This study focused on surveys that only collected information on the variable of interest, in this case it was population abundance. Scientific surveys often collect data on physical and environment covariates in conjunction with the variable of interest. Model-based estimators have the benefit of incorporating these covariates for improved spatial predictions, in addition to testing for relationships between covariates and the variable of interest (Johnson et al. 2010). Additional benefits of the geostatistical-model estimator also include that, it naturally outputs high spatial resolution distribution maps, can be applied to irregular spatial domains and, if surveys are repeated over time, the model-based estimator can be extended to include spatio-temporal correlation. Given these additional benefits, we recommend the use and exploration of the geostatistical model-based estimator.

A limitation identified for the boxlet method is its restriction to integer measures of abundance. Ecological surveys can often measure biomass as opposed to abundance which may limit its usage. That being said, it is a method whose performance was consistent and often had the best confidence interval coverage, which highlights that it should be considered in future surveys.

The confidence coverage of the geostatistical model-based estimator was found to be sensitive to assumptions regarding the mesh. We found that the confidence interval coverage varied by between 1-8% among the meshes explored. Neither the high resolution mesh (Mesh A) or the mesh with a boundary (Mesh B) were the best performing mesh assumptions which indicates an area of further research. The mesh with a boundary was meant to inflate the variance of the GF to account for a physical boundaries (Lindgren 2012), but in this simulation resulted in lower estimated standard errors. When presenting the results for the geostatistical model-based estimator relative to all other estimators we used the mesh with the boundary (Mesh B) because there was an implied boundary around the domain. This sensitivity analysis suggested that the results presented in Table 7.1 for confidence interval coverage of the geostatistical estimator may be pessimistic when different mesh assumptions are considered.

An unexpected result from both the boxlet and geostatistical model-based estimator was that the confidence interval coverage for the IPP spatial distributions was often worst for sampling type 1 (1% spatial coverage) compared to sampling

type 2 (0.25% spatial coverage). It was anticipated that the opposite would happen, that is, the higher the spatial sampling coverage, the better the estimators would perform. This could be the result of how the domain is partitioned for model extrapolations. Both the boxlet and geostatistical estimator partitioned the spatial domain into discrete cells based on the size of the sampling unit. This will require further simulations to identify if partitioning will effect the performance of these estimators.

Finally, there were estimators that have been used in the literature but not included here due to time-constraints. These include the geostatistical model-based estimator using kriging (Walline 2007, Aune-Lundberg & Strand 2014) and covered grid estimators (McGarvey et al. 2016). These should be considered in future comparison and simulation studies.

Chapter 8

Conclusions and future research

This thesis set out to develop and investigate spatial methods for improved estimates of abundance and associated variance estimates from preferentially and systematically sampled data. Geostatistical models using the SPDE approach were the main modelling framework for inference of spatial populations in this thesis. Results from this thesis are considered timely and complimentary to the current and active field of geostatistical modeling in the fisheries domain (Thorson et al. 2015, Grüss & Thorson 2019, Grüss et al. 2019, Pennino et al. 2019, Thorson 2019, Ducharme-Barth et al. 2022).

8.1 Preferentially sampled data

Due to the economic cost of fishery-independent research, many fisheries globally will continue to rely on fishery-dependent catch and effort data for estimating indices of abundance. Fishing is, by nature, a targeted activity which results in fishery-dependent catch and effort being sampled preferentially. Chapter 3 described a geostatistical model that accounted for preferentially sampled data (the PS model) for CPUE standardisation. Simulations in that chapter illustrated how the PS model could alleviate biases in abundance estimates and spatial predictions when the data was preferentially sampled.

Subsequent simulations in Chapter 3 explored the consequence of systematic changes in the preferential sampling process over time. All estimation models estimated biased indices of relative abundance. However, the PS model had the least biased estimates. This simulation study reinforced the importance of understanding

fishing behaviour in a CPUE standardisation, something that is often advocated for, as in Hilborn (1985), Branch et al. (2006).

A Pearson correlation metric was explored to identify if data was systematically sampled. However, this metric did not qualitatively identify (by visual inspection) temporal trends when the preferential sampling process was known to change over time. Diagnostics that could identify systematic shifts in preferential sampling would be valuable for geostatistical model-based inference using preferentially sampled data. These diagnostics could be used to split a time-series into periods with consistent sampling processes, possibly resulting in less biased indices. Alternative metrics such as the spatial point pattern test and its extensions (Andresen 2016, Wheeler et al. 2018) should be considered in future research.

Applying the PS model to the Chatham Rise hoki fishery (the case study) in Chapter 4 demonstrated that while the PS model with spatio-temporal GF could fit to real data, the PS model with a time-invariant GF did not successfully converge. In addition to demonstrating the utility of PS models, the case study motivated a range of innovative developments and indicated a range of areas for future research.

The case study prompted an exploration of model comparison methods and identified the predictive joint log-likelihood as a promising method for evaluating the usefulness of the PS model. When we applied this metric to the Chatham Rise case study, we had to assume the fishing locations for the conventional geostatistical model were from a homogenous point process. This lack of flexibility in describing spatial variation in the fishing locations led to the diagnostic strongly favouring the PS model. If this assumption was relaxed and conventional geostatistical models contained covariates or GF for the point process, it would explain variability in fishing locations and provide evidence on the appropriateness of the PS model. The downside to this approach is it would increase complexity due to additional variable selection procedures for the point process in addition to adding computational overheads.

The other model comparison approach explored throughout this thesis and used in other PS model investigations (Conn et al. 2017, Dinsdale 2018) was prediction error in catch rates. We found that if preferentially sampled data was randomly split into testing and training data sets for evaluating prediction error (i.e. in K-folds cross validation), it would likely favour the conventional geostatistical model. Testing data sets that were randomly sub sampled from preferentially sampled data

would naturally contain a high frequency of observations in areas that were preferentially sampled, and thus contribute more to summary statistics (i.e. root mean square error) than less frequent observations in poorly sampled regions. The strength of the PS model is predicting in low and unsampled regions (assuming the PS assumption is satisfied) that would not be "valued" using this approach. Simulations using the CABM model (Chapter 6) reinforced this conclusion. Testing data derived by randomly splitting testing and training data sets often favoured conventional geostatistical models. Whereas, when testing data sets were derived using cell midpoints over the spatial domain (i.e. more uniformly distributed), the PS model was often chosen. It is unlikely that in practice such a uniform testing data set would be available. Therefore future research should consider how testing data sets are selected during procedures such as K-folds cross validation in order to use prediction error in catch rates when comparing PS models to conventional geostatistical models.

One of the major innovations motivated by the case study was creating the general purpose ABM program, CABM (Chapter 5). CABM was originally created to emulate complex movement, productivity and fishing dynamics for hoki on the Chatham Rise. However, CABM was intentionally developed to be flexible with respect to life histories and spatial assumptions. This allowed CABM to be tuned to simulate a range of important fishery measures such as correlated age frequencies (Pennington & Volstad 1994) using clustered agents. CABM allowed us to conduct simulations (Chapter 6) in situations in which the operating model was not the estimation model, leading to improved generality in our results.

Results from the CABM Chatham Rise simulation identified that the most important factor for estimating β^{pref} (e.g., the strength of the PS) was the area fished relative to the spatial extent. We anticipated that target fisheries would be optimal for applying the PS model and was a factor in choosing the Chatham Rise hoki fishery as a case study. Results from this simulation highlighted the need to investigate the spatial distribution of fishing locations more closely for future case studies, with a particular emphasis on contrast - that is, some regions having high sampling frequency and some regions having low sampling frequency. The complexity of the CABM Chatham Rise simulation led to a range of additional results that were outside the scope of this study. These raised a range of interesting questions that should be addressed in future research.

To summarise, our exploration of models that estimate indices of abundance from preferentially sampled data resulted in developing an R package for fitting the PS model and conventional geostatistical models `CPUEspatial` (https://github.com/Craig44/CPUEspatial), conducting simulations to illustrate the benefits of the PS model, applying the PS model to the Chatham Rise hoki fishery and creating the general purpose ABM program CABM. However, there were elements of the PS model that were not explored in detail that would benefit from further research. The most important being:

- extend the catch rate likelihood to include zero observations. This would provide more samples in regions of low to moderate abundance that would help estimate β^{pref};
- allow for additional spatial covariates and GF assumptions (systematic component) for both the catch rate and location expected values that are not shared. This would relax the PS assumption and better allow for model comparison using the joint predictive log-likelihood metric;
- investigate alternative point process models for sample location data sets. Due to time constraints only one point process likelihood was explored. However, there are alternatives which have additional computational benefits i.e. Simpson et al. (2016);
- allow the PS model to exclude closed areas, i.e. physical boundary mesh Bakka et al. (2019).
- explore alternative goodness of fit methods for catch rates. The geostatistical models employed randomised quantile residuals (Dunn & Smyth 1996, Scudilio & Pereira 2020) (Equation 2.16). A possible concern with these residuals is their lack of power for identifying misspecified models. The analogy is to the Pearson's residuals, which have been shown to be a flawed diagnostic for misspecified hierarchical models (Thygesen et al. 2017). Alternatives include the one-step prediction residuals proposed by (Thygesen et al. 2017) and simulated quantile residuals (Hartig 2020). Future simulations should explore the utility of these approaches.

8.2 Systematically sampled data

Our study of estimators for data sampled from a two-dimensional systematic survey with a single PSU was an extension of previous comparison studies (McGarvey et al. 2016, D'Orazio 2003). By including the semiparametric boxlet estimator of Fewster (2011) and exploring a geostatistical model-based estimator using the SPDE approach, we extended the search for a preferred estimator for this survey design.

We anticipated the boxlet estimator and geostatistical model-based estimator would improve upon approximated design-based estimators, given their ability to model spatial heterogeneous populations. Although the geostatistical model-based point estimator was more efficient for the total population for some of the spatial distributions, neither the boxlet nor geostatistical model-based estimator achieved the desired confidence interval coverage across all spatial distributions. This result suggests that an optimal estimator would likely depend on population attributes such as periodicity and spatial correlation etc.

This study motivated the development of the R package `systematicsurvey` found at https://github.com/Craig44/systematicsurvey. This R package was created to overcome some of the barriers given for the under-utilised boxlet and geostatistical model-based estimators. This will give future researchers the ability to apply a range of estimators. If estimators give conflicting results, the literature should be referred to i.e., Strand (2017), McGarvey et al. (2016). This will help identify which estimator may be most appropriate for a particular spatial distribution of the sampled population.

Results from the simulation study identified two areas of future research for the geostatistical model-based estimator; firstly, how to diagnose and fix non-converged models. Convergence issues were encountered for the uniform spatial distribution but could occasionally occur in other situations. The convergence issue was due to the GF variance parameter being estimated close to zero $\sigma^2_M \approx 0$. Exploring methods such as starting parameter values were not considered during the simulation due to time constraints but may have resolved this. Secondly, mesh assumptions were found to influence the confidence interval coverage. Our results suggested that a mesh with no boundary and spatial resolution close to the spatial resolution of the sampling units would achieve better confidence interval coverage but more research is needed to investigate optimal mesh assumptions in detail.

There were two aspects of the geostatistical estimator that were not explored in detail during this thesis and should be in future work. These include:

- explore alternative spatial covariance structures. The geostatistical model-based estimator assumed isotropic Matérn covariance. An alternative that should be explored is the anisotropic Matérn covariance (Deng 2008). Anisotropy Matérn covariances allow spatial autocorrelation to be a function of both direction and distance. This is expected to be more appropriate in environments with hard geographic boundaries on one side of the domain such as beaches and sub or inter-tidal environments,

- exploring goodness of fit metrics, as mentioned earlier in the preferential sampling discussion. Goodness of fit metrics would be valuable for the geostatistical model-based estimator. Such diagnostics may provide insight into suitable mesh assumptions.

In summary, the geostatistical model-based estimator did not completely resolve the estimation issue for systematic surveys and needs additional exploration. However, its use is recommended to systematic surveys for its ability to produce spatial distribution maps. Moreover, it can include covariates and be applied to irregular spatial domains.

Appendix A

Notation

Table A.1: Distributions definitions, and parameter meaning

Notation	Distribution
$\mathcal{N}(\mu, \sigma^2)$	Normal
$\mathcal{LN}(\mu, \sigma^2)$	Log normal distribution
$Poisson(\lambda)$	Poisson
$\mathcal{B}in(n, p)$	Binomial
$\mathcal{B}ern(p)$	Bernoulli

Appendix B

Geostatistical model-based estimator settings

In general simulations can be difficult to get 100% convergence as optimisation performance can be dependent on bounds and starting values (Millar 2011). A phased estimation approach was implemented which followed

- Phase 1: estimate μ whilst $\boldsymbol{\omega}, \kappa, \tau$ were fixed at zero,
- Phase 2: μ set to MLE estimate from phase 1,$\kappa = \log(\sqrt(8)/\Delta_{0.1})$ $\tau = 1/var(y)$ and $\boldsymbol{\omega}$ equal to zero.

$\Delta_{0.1}$ is the *a priori* distance that the spatial auto-correlation equals 0.1, in our case we set this to 10% x-axis.

Models were implemented using **R** inbuilt optimiser `nlminb`, convergence was based on the maximum absolute gradient of fixed effect parameters being less than 0.001, as well as the model producing a positive definite hessian matrix.

Appendix C

Characterisation of the hoki Chatham Rise fishery

Table C.1: Species code with common name

Code	Common name	Scientific name
SWA	Silver warehau	*Seriolella brama*
HOK	Hoki	*Macruronus novaezelandiae*
BAR	Barracouta	*Thyrsites atun*
HAK	Hake	*Merluccius australis*
BYX	Alfonsino	*Beryx splendens*
FLA	Flatfish	species of *Rhombosolea*
SQU	Squid	*Nototodarus sloanii*
TAR	Tarakihi	*Nemadactylus macropterus*
RCO	Red cod	*Pseudophycis bachus*
ORH	Orange roughy	*Hoplostethus atlanticus*

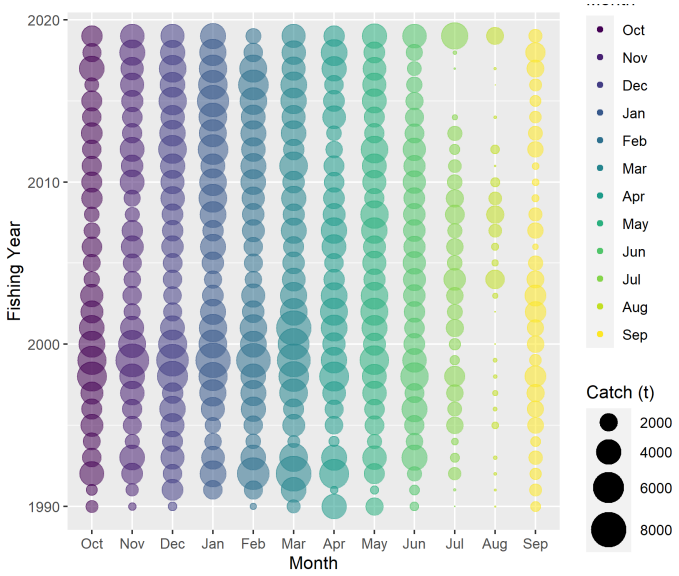


Figure C.1: Hoki catch by year and month

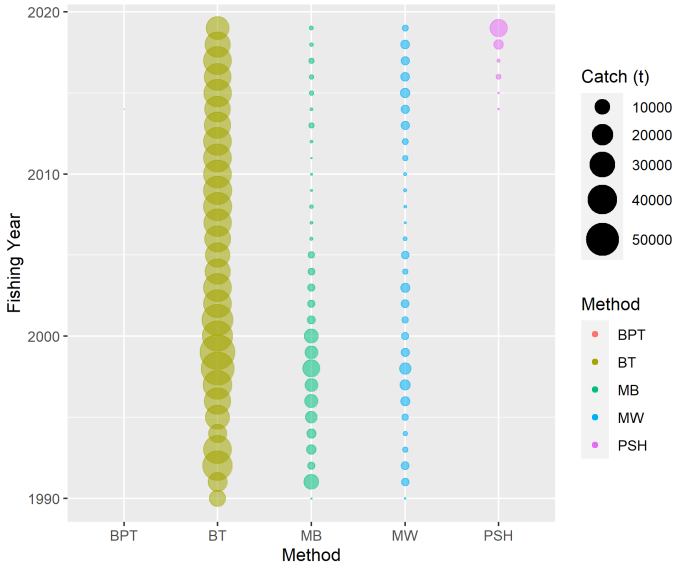


Figure C.2: Hoki catch by year and fishing method. BT = bottom trawl, MB = mid bottom trawl, MW = midwater trawl, BPT = bottom paired trawl, PSH = precision seafood harvest net.

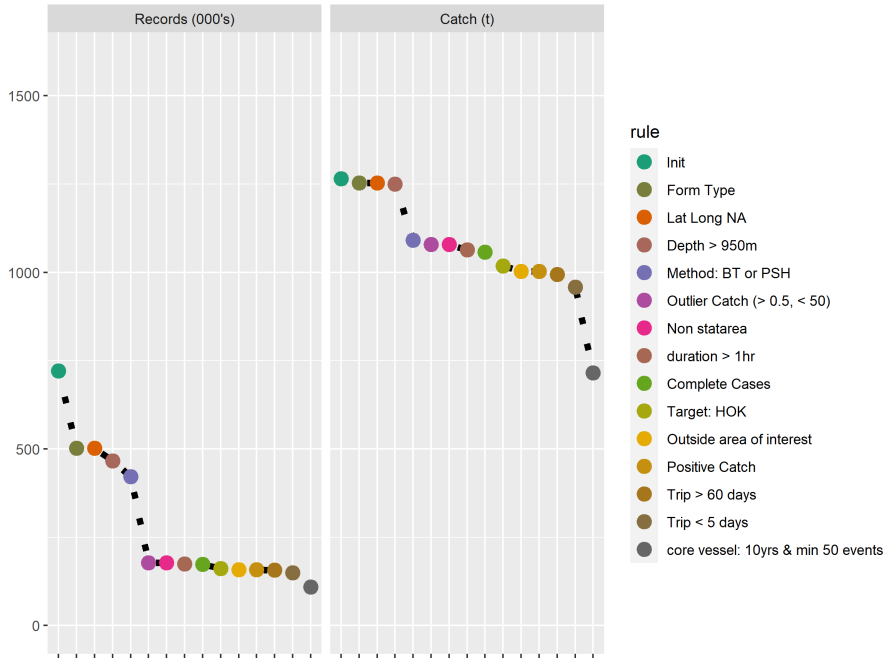


Figure C.3: The effect of grooming rules (Table 4.1) on catch and number of observations

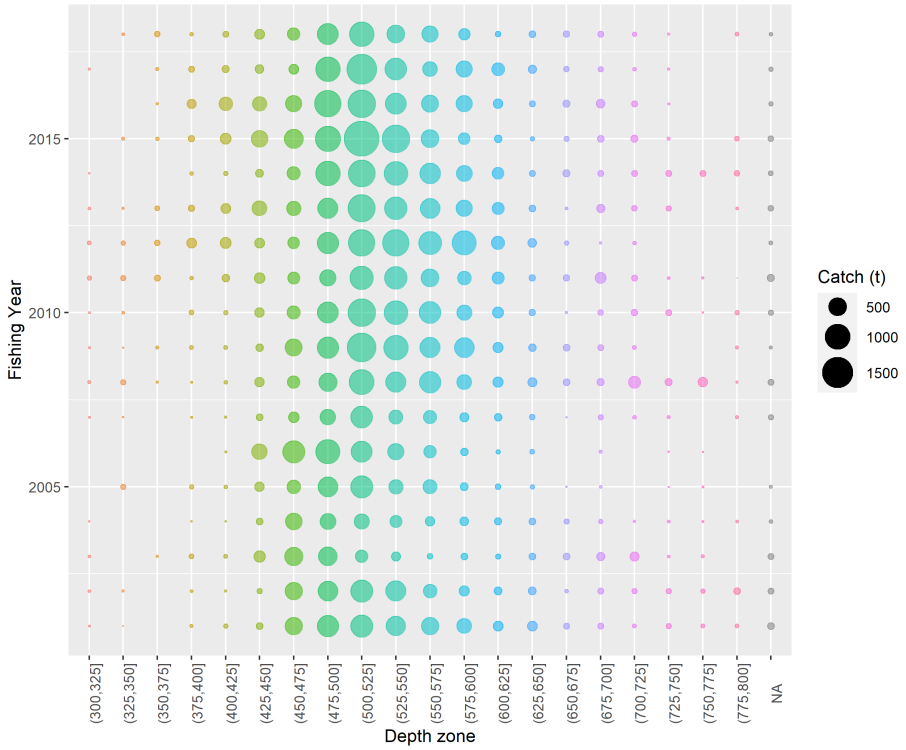


Figure C.4: Hoki catch by depth and year.

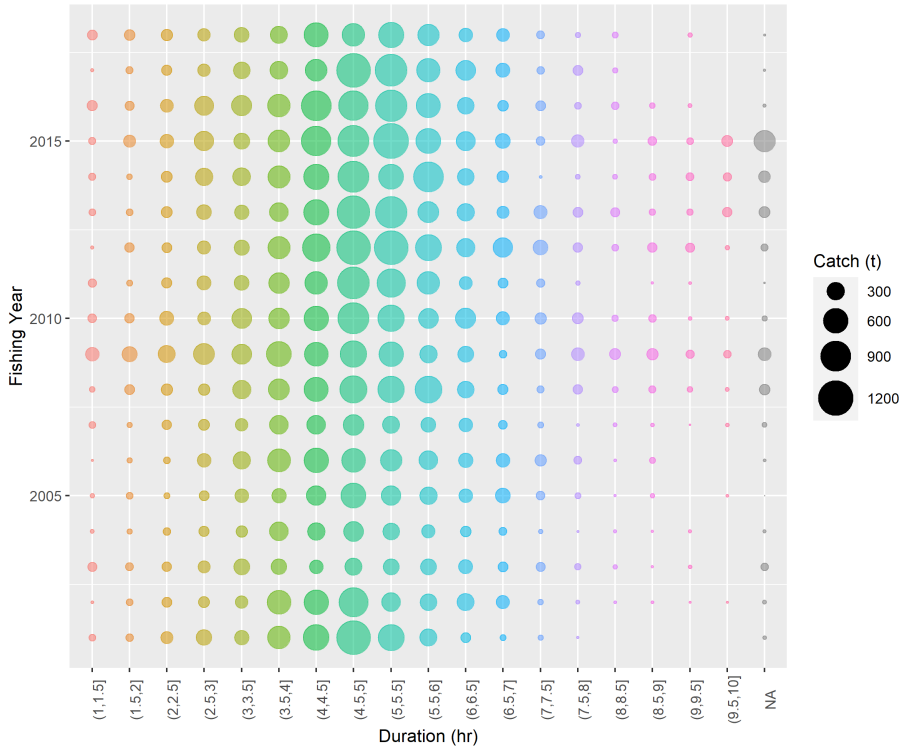


Figure C.5: Hoki catch by duration and year.

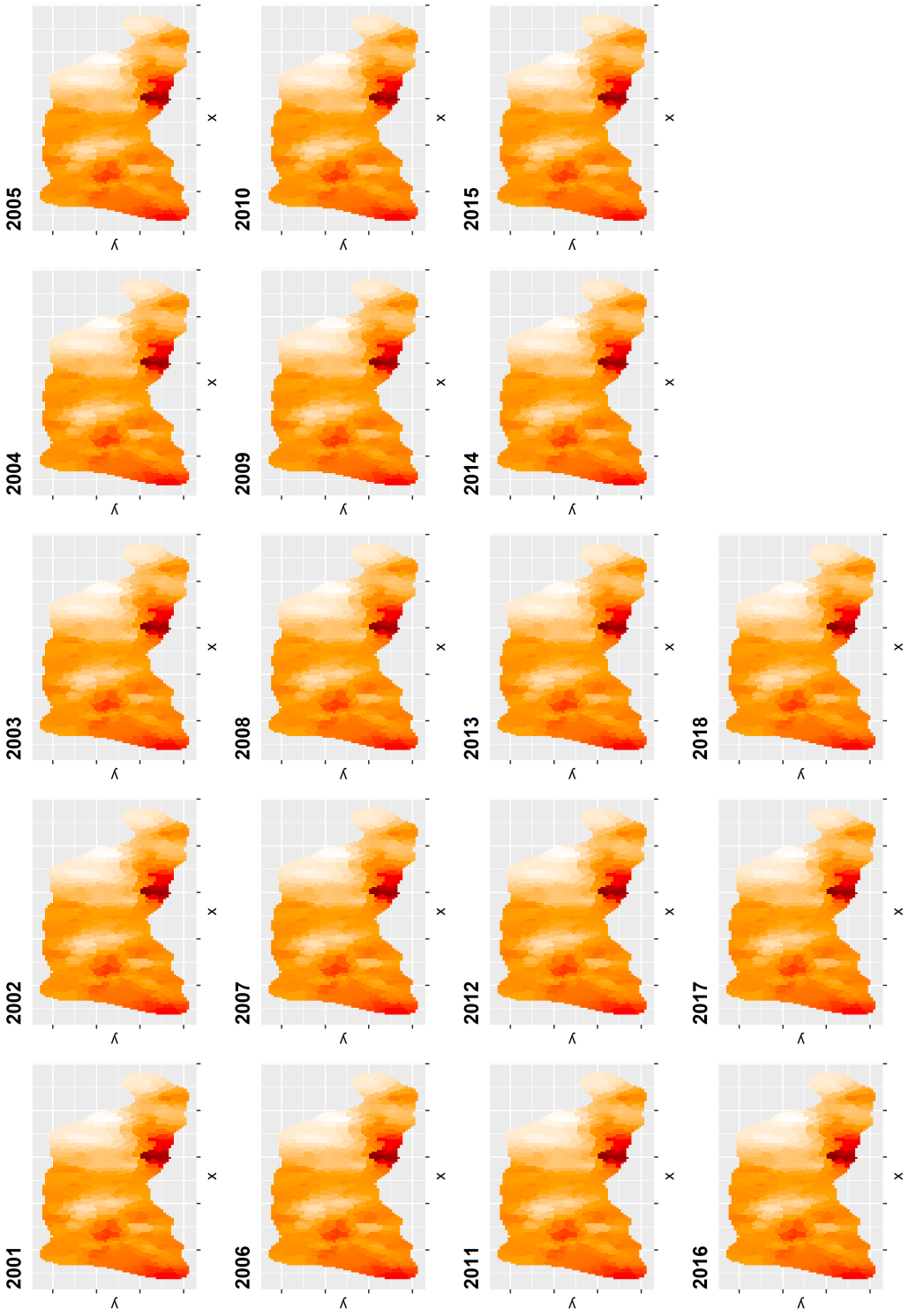


Figure C.6: Estimated spatial distribution from model $\mathcal{M}_1$

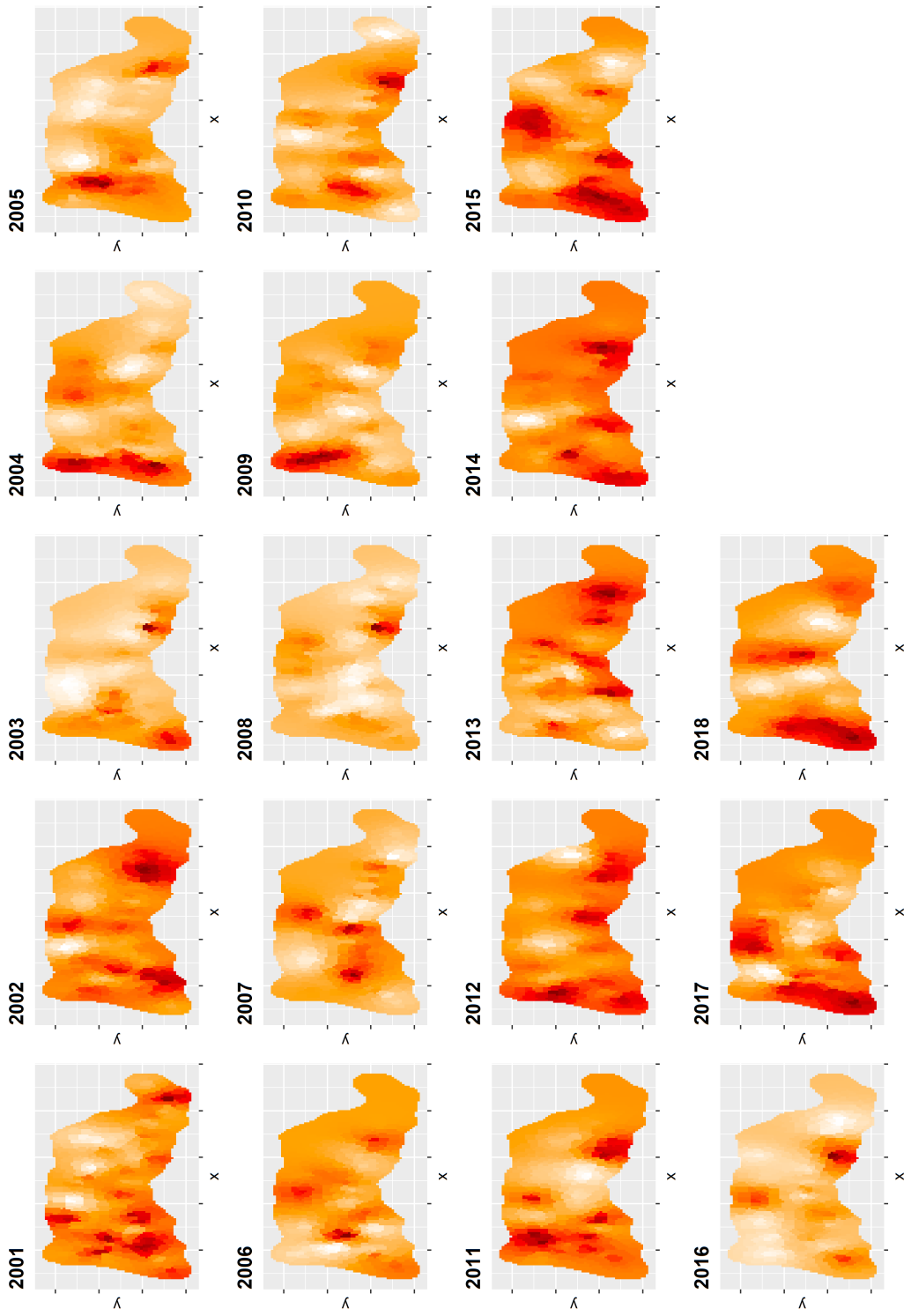


Figure C.7: Estimated spatial distribution from model $\mathcal{M}_2$

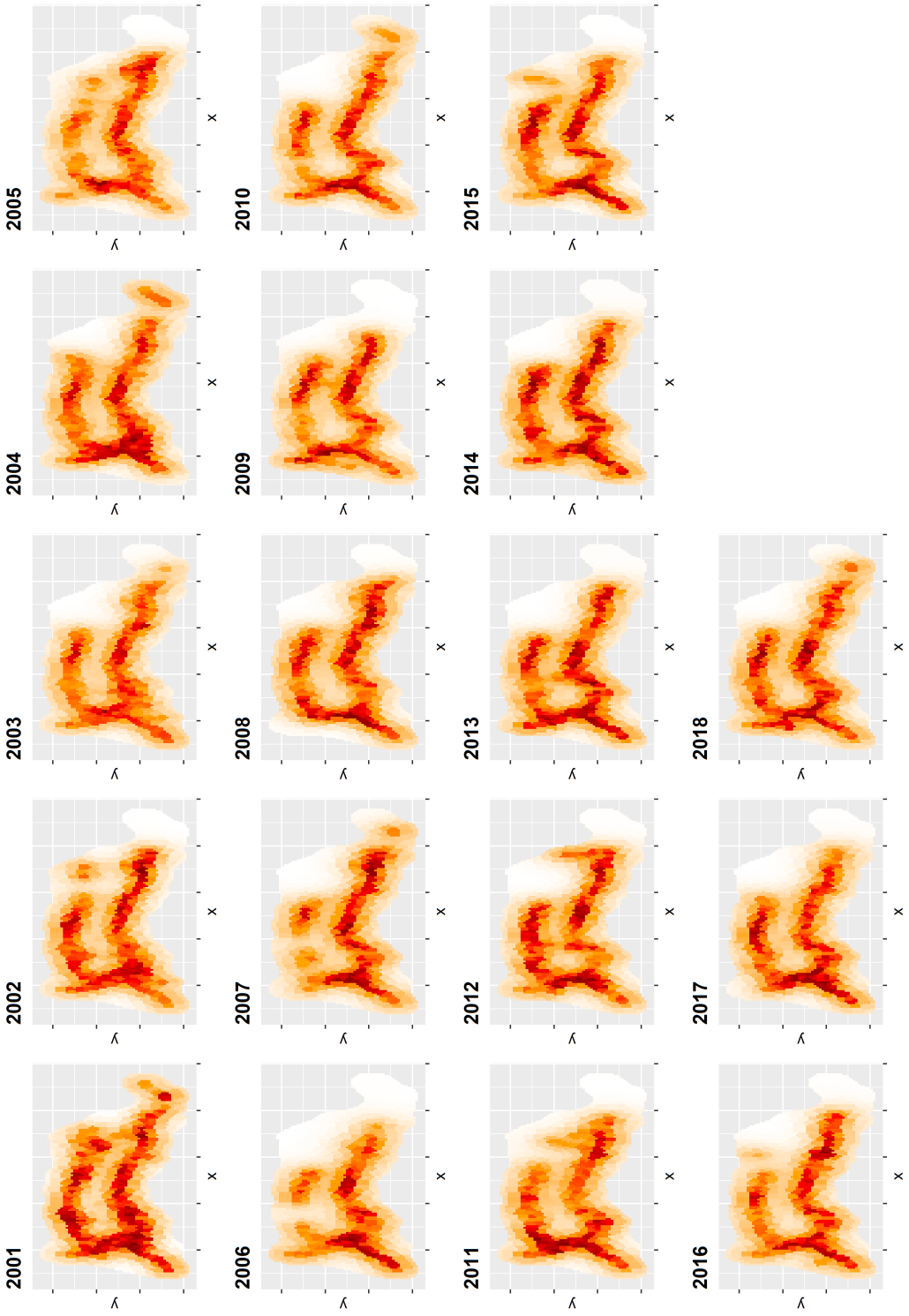


Figure C.8: Estimated spatial distribution from model $\mathcal{M}_4$

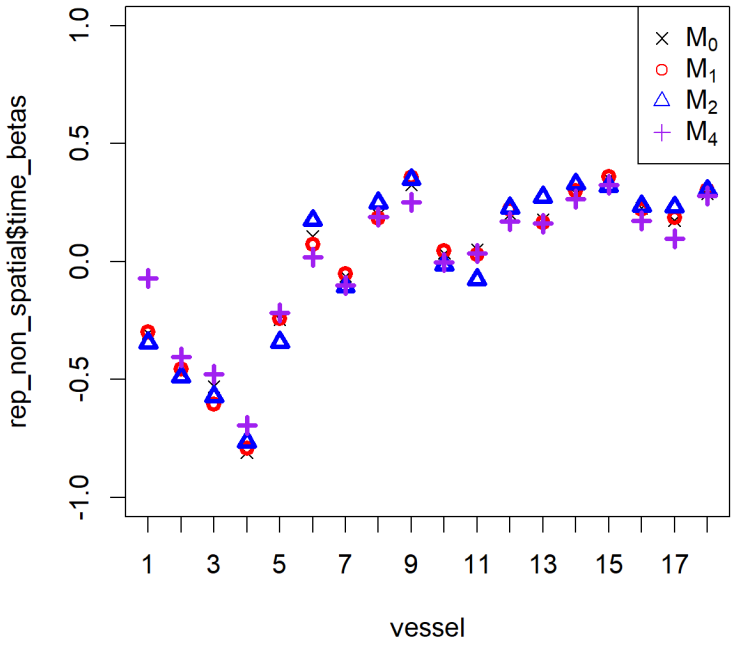


Figure C.9: Estimated year effect among EMs.

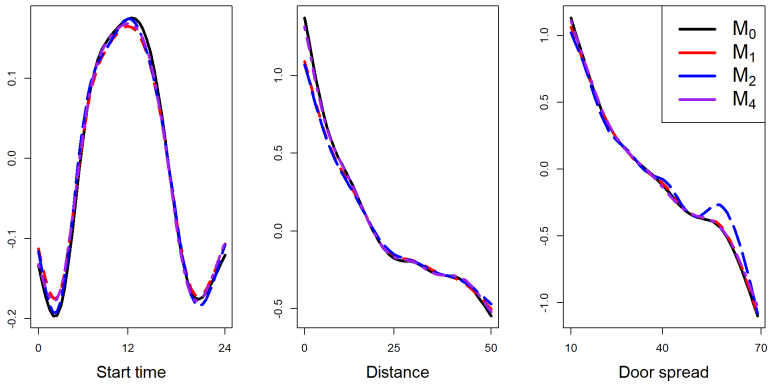


Figure C.10: Estimated splines among EMs.

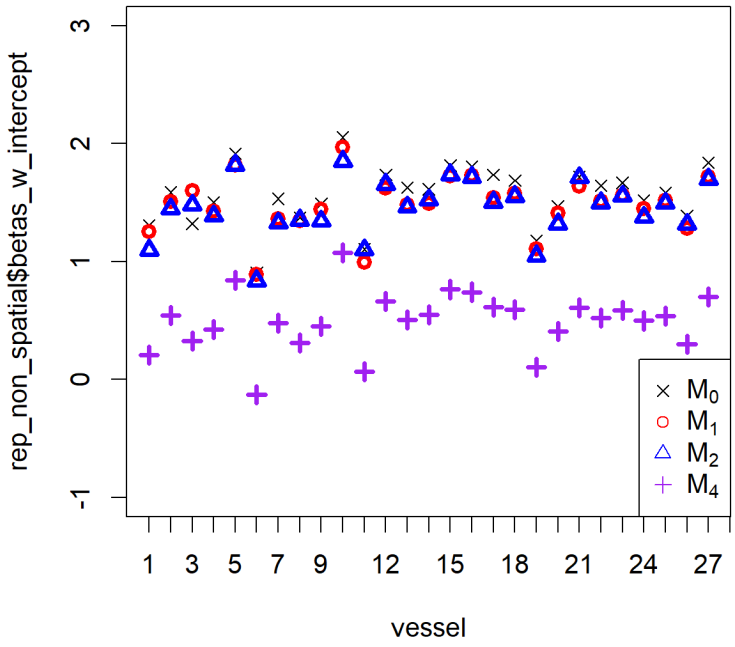


Figure C.11: Estimated vessel effect among EMs.

Bibliography

Abraham, E. R. & Neubauer, P. (2015), Relationship between small-scale catch-per-unit-effort and abundance in new zealand abalone (pāua, haliotis iris) fisheries, Technical report, PeerJ PrePrints.
https://peerj.com/preprints/1388/

Ambrosio Flores, L., Iglesias Martínez, L. & Marín Ferrer, C. (2003), 'Systematic sample design for the estimation of spatial means', *Environmetrics: The official journal of the International Environmetrics Society* **14**(1), 45–61.

Andresen, M. A. (2016), 'An area-based nonparametric spatial point pattern test: The test, its applications, and the future', *Methodological Innovations* **9**, 11.

Anselin, L. (1995), 'Local indicators of spatial association LISA', *Geographical Analysis* **27**(2), 93–115.

Aune-Lundberg, L. & Strand, G.-H. (2014), 'Comparison of variance estimation methods for use with two-dimensional systematic sampling of land use/land cover data', *Environmental Modelling & Software* **61**, 87–97.

Bakka, H. (2018), 'How to solve the stochastic partial differential equation that gives a matérn random field using the finite element method', *arXiv preprint arXiv:1803.03765* .

Bakka, H., Vanhatalo, J., Illian, J. B., Simpson, D. & Rue, H. (2019), 'Non-stationary Gaussian models with physical barriers', *Spatial Statistics* **29**, 268–288.

Ballara, S. L., Ladroit, Y., O'Driscoll, R. L. & Stevens, D. W. (2017), 'Trawl survey of hoki and middle depth species on the Chatham Rise, January 2016 (TAN1601)', *New Zealand Fisheries Assessment Research Document* (08).

Ballara, S. L. & O'Driscoll, R. L. (2020), 'Catches and size and age structure of the 2018–19 hoki fishery', *New Zealand Fisheries Assessment Research Document* (22).

Bartolucci, F. & Montanari, G. E. (2006), 'A new class of unbiased estimators of the variance of the systematic sample mean', *Journal of Statistical Planning and Inference* **136**(4), 1512–1525.

Begg, G. A. & Waldman, J. R. (1999), 'An holistic approach to fish stock identification', *Fisheries Research* **43**(1-3), 35–44.

Bentley, N., Kendrick, T. H., Starr, P. J. & Breen, P. A. (2012), 'Influence plots and metrics: tools for better understanding fisheries catch-per-unit-effort standardizations', *ICES Journal of Marine Science* **69**(1), 84–88.

Bertignac, M., Lehodey, P. & Hampton, J. (1998), 'A spatial population dynamics simulation model of tropical tunas using a habitat index based on environmental parameters', *Fisheries Oceanography* **7**(3-4), 326–334.

Bertsekas, D. P. (1997), 'Nonlinear programming', *Journal of the Operational Research Society* **48**(3), 334–334.

Bigelow, K. A., Boggs, C. H. & He, X. (1999), 'Environmental effects on swordfish and blue shark catch rates in the US North Pacific longline fishery', *Fisheries Oceanography* **8**(3), 178–198.

Bischof, C. & Griewank, A. (1992), ADIFOR-A FORTRAN system for portable automatic differentiation, *in* '4th Symposium on Multidisciplinary Analysis and Optimization', p. 4744.

Blangiardo, M. & Cameletti, M. (2015), *Spatial and spatio-temporal Bayesian models with R-INLA*, John Wiley & Sons.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H. & White, J.-S. S. (2009), 'Generalized linear mixed models: A practical guide for ecology and evolution', *Trends in Ecology & Evolution* **24**(3), 127–135.

Bolker, B. M., Gardner, B., Maunder, M., Berg, C. W., Brooks, M., Comita, L., Crone, E., Cubaynes, S., Davies, T., de Valpine, P. et al. (2013), 'Strategies for fitting nonlinear ecological models in R, AD Model Builder, and BUGS', *Methods in Ecology and Evolution* **4**(6), 501–512.

Box, G. E. (1979), Robustness in the strategy of scientific model building, *in* 'Robustness in Statistics', Elsevier, pp. 201–236.

Branch, T. A., Hilborn, R., Haynie, A. C., Fay, G., Flynn, L., Griffiths, J., Marshall, K. N., Randall, J. K., Scheuerell, J. M., Ward, E. J. et al. (2006), 'Fleet dynamics and fishermen behavior: lessons for fisheries managers', *Canadian Journal of Fisheries and Aquatic Sciences* **63**(7), 1647–1668.

Bravington, M. V., Skaug, H. J. & Anderson, E. C. (2016), 'Close-kin mark-recapture', *Statistical Science* **31**(2), 259–274.

Brown, K. S. & Sethna, J. P. (2003), 'Statistical mechanical approaches to models with many poorly known parameters', *Physical review E* **68**(2), 021904.

Brus, D. J. (2021), 'Statistical approaches for spatial sample survey: Persistent misconceptions and new developments', *European Journal of Soil Science* **72**(2), 686–703.

Brus, D. & Saby, N. (2016), 'Approximating the variance of estimated means for systematic random sampling, illustrated with data of the french soil monitoring network', *Geoderma* **279**, 77–86.

Bull, B., Francis, R. I. C. C., Dunn, A., McKenzie, A., Gilbert, D., Smith, M., Bian, R. & Fu, D. (2012), 'CASAL (C++ algorithmic stock assessment laboratory): CASAL User Manual v2'.

Campbell, R. A. (2015), 'Constructing stock abundance indices from catch and effort data: Some nuts and bolts', *Fisheries Research* **161**, 109–130.

Cao, J., Guan, W., Truesdell, S., Chen, Y. & Tian, S. (2016), 'An individual-based probabilistic model for simulating fisheries population dynamics', *Aquaculture and Fisheries* **1**, 34–40.

Carruthers, T. R., Ahrens, R. N., McAllister, M. K. & Walters, C. J. (2011), 'Integrating imputation and standardization of catch rate data in the calculation of relative abundance indices', *Fisheries Research* **109**(1), 157–167.

Chambers, R. & Clark, R. (2012), *An Introduction to Model-Based Survey Sampling with Applications*, Vol. 37, Oxford University Press.

Chang, K.-L., Guillas, S. & Fioletov, V. (2015), 'Spatial mapping of ground-based observations of total ozone', *Atmospheric Measurement Techniques* **8**(10), 4487–4505.

Chatfield, C. (1995), *Problem Solving: A Statistician's Guide*, CRC Press.

Cochran, W. G. (2007), *Sampling Techniques*, John Wiley & Sons.

Conn, P. B., Thorson, J. T. & Johnson, D. S. (2017), 'Confronting preferential sampling when analysing population distributions: diagnosis and model-based triage', *Methods in Ecology and Evolution* **8**(11), 1535–1546.

Cragg, J. G. (1971), 'Some statistical models for limited dependent variables with application to the demand for durable goods', *Econometrica: Journal of the Econometric Society* pp. 829–844.

Cressie, N. (2015), *Statistics for spatial data*, John Wiley & Sons.

Deng, M. (2008), 'An anisotropic model for spatial processes', *Geographical Analysis* **40**(1), 26–51.

Deroba, J., Butterworth, D. S., Methot Jr, R., De Oliveira, J., Fernandez, C., Nielsen, A., Cadrin, S., Dickey-Collas, M., Legault, C., Ianelli, J. et al. (2014), 'Simulation testing the robustness of stock assessment models to error: some results from the ICES strategic initiative on stock assessment methods', *ICES Journal of Marine Science* **72**(1), 19–30.

Diggle, P. J., Menezes, R. & Su, T.-l. (2010), 'Geostatistical inference under preferential sampling', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **59**(2), 191–232.

Diggle, P. J., Moraga, P., Rowlingson, B., Taylor, B. M. et al. (2013), 'Spatial and spatio-temporal log-gaussian cox processes: extending the geostatistical paradigm', *Statistical Science* **28**(4), 542–563.

Dinsdale, D. (2018), Methods for preferential sampling in geostatistics, PhD thesis, University of British Columbia.

Doonan, I., Large, K., Dunn, A., Rasmussen, S., Marsh, C. & Mormede, S. (2016), 'Casal2: New Zealand's integrated population modelling tool', *Fisheries Research* **183**, 498–505.

D'Orazio, M. (2003), 'Estimating the variance of the sample mean in two-dimensional systematic sampling', *Journal of Agricultural, Biological, and Environmental Statistics* **8**(3), 280.

Ducharme-Barth, N. D., Grüss, A., Vincent, M. T., Kiyofuji, H., Aoki, Y., Pilling, G., Hampton, J. & Thorson, J. T. (2022), 'Impacts of fisheries-dependent spatial sampling patterns on catch-per-unit-effort standardization: A simulation study and fishery application', *Fisheries Research* **246**.

Dunn, A., Rasmussen, S. & Mormede, S. (2012), 'Spatial Population Model User Manual, SPM v1.1-2012-09-06 (rev 4806)', *CCAMLR; WG-FSA-12/46* .

Dunn, D. C., Boustany, A. M., Roberts, J. J., Brazer, E., Sanderson, M., Gardner, B. & Halpin, P. N. (2014), 'Empirical move-on rules to inform fishing strategies: a New England case study', *Fish and Fisheries* **15**(3), 359–375.

Dunn, P. K. & Smyth, G. K. (1996), 'Randomized quantile residuals', *Journal of Computational and Graphical Statistics* **5**(3), 236–244.

Dunn, R. & Harrison, A. R. (1993), 'Two-dimensional systematic sampling of land use', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **42**(4), 585–601.

FAO (2020), 'The State of World Fisheries and Aquaculture 2020', *Sustainability in action* .
https://www.fao.org/documents/card/en/c/ca9229en/

Fewster, R. M. (2011), 'Variance estimation for systematic designs in spatial surveys', *Biometrics* **67**(4), 1518–1531.

Fewster, R. M., Buckland, S. T., Burnham, K. P., Borchers, D. L., Jupp, P. E., Laake, J. L. & Thomas, L. (2009), 'Estimating the encounter rate variance in distance sampling', *Biometrics* **65**(1), 225–236.

Fisheries New Zealand (2018), Fisheries Assessment Plenary, May 2018: stock assessments and stock status, Technical Report May, Compiled by the Fisheries Science and Information Group, Fisheries New Zealand, Wellington, New Zealand.

Fisheries New Zealand (2019), Approval for the inshore use of a precision seafood harvesting modular harvest system trawl net under regulation 71a of the fisheries (commercial fishing) regulations 2001, Technical report, Fisheries New Zealand.

Fonteneau, A., Gaertner, D. & Nordström, V. (1999), 'An overview of problems in the CPUE-abundance relationship for the tropical purse seine fisheries'.

Fournier, D. A., Skaug, H. J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M. N., Nielsen, A. & Sibert, J. (2012), 'AD model builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models', *Optimization Methods and Software* **27**(2), 233–249.

Fournier, D. & Archibald, C. P. (1982), 'A general theory for analyzing catch at age data', *Canadian Journal of Fisheries and Aquatic Sciences* **39**(8), 1195–1207.

Francis, R. I. C. C. (1999), 'The impact of correlation in standardised CPUE indices. New Zealand Fisheries Assessment', *New Zealand Fisheries Assessment Research Document* **42**.

Francis, R. I. C. C. (2011), 'Data weighting in statistical fisheries stock assessment models', *Canadian Journal of Fisheries and Aquatic Sciences* **68**(6), 1124–1138.

Francis, R. I. C. C. (2012), 'The reliability of estimates of natural mortality from stock assessment models', *Fisheries Research* **119-120**, 133–134.

Francis, R. I. C. C. (2017), 'Revisiting data weighting in fisheries stock assessment models', *Fisheries Research* **192**, 5–15.

Francis, R. I. C. C., Hurst, R. J. & Renwick, J. A. (2001), 'An evaluation of catchability assumptions in New Zealand stock assessments', *New Zealand Fisheries Assessment Report* **1**.

Fuglstad, G.-A., Simpson, D., Lindgren, F. & Rue, H. (2019), 'Constructing priors that penalize the complexity of Gaussian random fields', *Journal of the American Statistical Association* **114**(525), 445–452.

Gábor, A. & Banga, J. R. (2015), 'Robust and efficient parameter estimation in dynamic models of biological systems', *BMC systems biology* **9**(1), 74.

Gabriel, W. L. & Mace, P. M. (1999), A review of biological reference points in the context of the precautionary approach, *in* V. R. Restrepo, ed., 'Proceedings of the Fifth National NMFS Stock Assessment Workshop Providing Scientific Advice to Implement the Precautionary Approach Under the Magnuson-Stevens Fishery Conservation and Management Act', United States, National Marine Fisheries Service, pp. 34–45.

Gavaris, S. (1980), 'Use of a multiplicative model to estimate catch rate and effort from commercial data', *Canadian Journal of Fisheries and Aquatic Sciences* **37**(12), 2272–2275.

Gay, D. M. (1990), 'Usage summary for selected optimization routines', *Computing science technical report* **153**, 1–21.

Geange, S. W., Rowden, A. A., Nicol, S., Bock, T. & Cryer, M. (2020), 'A Data-Informed Approach for Identifying Move-on Encounter Thresholds for Vulnerable Marine Ecosystem Indicator Taxa', *Frontiers in Marine Science* **7**.
https://www.frontiersin.org/article/10.3389/fmars.2020.00155

Gelman, A. & Hill, J. (2006), *Data analysis using regression and multilevel/hierarchical models*, Cambridge university press.

Gillis, D. M. (2003), 'Ideal free distributions in fleet dynamics: a behavioral perspective on vessel movement in fisheries analysis', *Canadian Journal of Zoology* **81**(2), 177–187.

Gillis, D. & Peterman, R. (1998), 'Implications of interference among fishing vessels and the ideal free distribution to the interpretation of CPUE', *Canadian Journal of Fisheries and Aquatic Sciences* **55**(1), 37–46.

Girardin, R., Hamon, K. G., Pinnegar, J., Poos, J. J., Thébaud, O., Tidd, A., Vermard, Y. & Marchal, P. (2017), 'Thirty years of fleet dynamics modelling using discrete-choice models: What have we learned?', *Fish and Fisheries* **18**(4), 638–655.

Goethel, D., Hanselman, D., Rodgveller, C., Fenske, K., Shotwell, K., Echave, Katy Malecha, P., Siwicke, K. & Lunsford, C. (2020), 'Assessment of the sablefish stock in alaska', *NPFMC Bering Sea, Aleutian Islands and Gulf of Alaska SAFE* .

Goethel, D. R., Quinn, T. J. & Cadrin, S. X. (2011), 'Incorporating spatial structure in stock assessment: movement modeling in marine fish population dynamics', *Reviews in Fisheries Science* **19**(2), 119–136.

Gómez-Rubio, V. (2020), *Bayesian inference with INLA*, CRC Press.

Grimm, V. & Railsback, S. F. (2013), *Individual-based Modeling and Ecology*, Princeton University Press.

Group, D. (2018), Deepwater Trawl hoki Operational Procedures, Technical report, Deepwater Group (Version 18).
https://deepwatergroup.org/newsresources/op-manual/

Grüss, A. & Thorson, J. T. (2019), 'Developing spatio-temporal models using multiple data types for evaluating population trends and habitat usage', *ICES Journal of Marine Science* **76**(6), 1748–1761.

Grüss, A., Walter III, J. F., Babcock, E. A., Forrestal, F. C., Thorson, J. T., Lauretta, M. V. & Schirripa, M. J. (2019), 'Evaluation of the impacts of different treatments of spatio-temporal variation in catch-per-unit-effort standardization models', *Fisheries Research* **213**, 75–93.

Gulland, J, A. (1974), 'Catch per unit effort as a measure of abundance', *Collective Volume of Scientific Papers ICCAT* **3**, 1–11.

Haddon, M. (2010), *Modelling and Quantitative Methods in Fisheries*, CRC press.

Hannah, R. W. (2003), 'Spatial changes in trawl fishing effort in response to footrope diameter restrictions in the us west coast bottom trawl fishery', *North American Journal of Fisheries Management* **23**(3), 693–702.

Harley, S. J., Myers, R. A. & Dunn, A. (2001*a*), 'Is catch-per-unit-effort proportional to abundance?', *Canadian Journal of Fisheries and Aquatic Sciences* **58**(9), 1760–1772.

Harley, S., Myers, R. & Dunn, A. (2001*b*), 'A meta-analysis of the relationship between catch-per-unit-effort and abundance', *Canadian Journal of Fisheries and Aquatic Sciences* **58**, 1705–1772.

Hartig, F. (2020), *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. R package version 0.2.7.
https://CRAN.R-project.org/package=DHARMa

Hastie, T. J. & Tibshirani, R. J. (1990), *Generalized additive models*, Vol. 43, CRC press.

Higham, D. J. & Higham, N. J. (2016), *MATLAB guide*, SIAM.

Hilborn, R. (1985), 'Fleet dynamics and individual variation: why some people catch more fish than others', *Canadian Journal of Fisheries and Aquatic Sciences* **42**, 2–13.

Hilborn, R. & Walters, C. J. (1992), *Quantitative Fisheries Stock Assessment: Choice, Dynamics and Uncertainty*, Chapman and Hall.

Hinton, M. G. & Maunder, M. N. (2003), Methods for standardizing CPUE and how to select among them, Technical report, The sixteenth meeting of the Standing Committee on Tuna and Billfish, SCRS/2003/034.

Horn, P. & Francis, R. I. C. C. (2010), 'Stock assessment of hake (*Merluccius australis*) on the Chatham Rise for the 2009–10 fishing year', *New Zealand Fisheries Assessment Report* **14**, 65.

Hoyle, S. D. & Langley, A. D. (2020), 'Scaling factors for multi-region stock assessments, with an application to Indian Ocean tropical tunas', *Fisheries Research* **228**, 105586.

Hutton, T., Mardle, S., Pascoe, S. & Clark, R. A. (2004), 'Modelling fishing location choice within mixed fisheries: English North Sea beam trawlers in 2000 and 2001', *ICES Journal of Marine Science* **61**(8), 1443–1452.

Hyun, S.-Y. & Seo, Y. I. (2018), 'The systematic sampling for inferring the survey indices of Korean groundfish stocks', *Fisheries and Aquatic Sciences* **21**(1), 1–9.

Johnson, D., Laake, J. & VerHoef, J. (2014), *DSpat: Spatial Modelling for Distance Sampling Data*. R package version 0.1.6.
https://CRAN.R-project.org/package=DSpat

Johnson, D. S., Laake, J. L. & Ver Hoef, J. M. (2010), 'A model-based approach for making ecological inference from distance sampling data', *Biometrics* **66**(1), 310–318.

Krainski, E. T., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F. & Rue, H. (2018), *Advanced spatial modeling with stochastic partial differential equations using R and INLA*, CRC Press.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H. & Bell, B. M. (2016), 'TMB: Automatic differentiation and Laplace Approximation', *Journal of Statistical Software* **70**(5), 1–21.

Langley, A. (2020), 'Stock assessment of snapper in SNA 8 for 2020', *New Zealand Fisheries Assessment Report* p. 20.

Langley, A. & Bentley, N. (2014), 'Fishery characterisation and Catch-Per-Unit-Effort indices for giant stargazer in STA 5', *New Zealand Fisheries Assessment Report* **64**, 49.

Lee, H.-H., Maunder, M. N., Piner, K. R. & Methot, R. D. (2011), 'Estimating natural mortality within a fisheries stock assessment model: an evaluation using simulation analysis based on twelve stock assessments', *Fisheries Research* **109**(1), 89–94.

Lee, H.-H., Maunder, M. N., Piner, K. R. & Methot, R. D. (2012), 'Can steepness of the stock–recruitment relationship be estimated in fishery stock assessment models?', *Fisheries Research* **125**, 254–261.

Lindgren, F. (2012), 'Continuous domain spatial models in R-INLA', *The ISBA Bulletin* **19**(4), 14–20.

Lindgren, F. & Rue, H. (2015), 'Bayesian Spatial Modelling with R-INLA', *Journal of Statistical Software* **63**(19), 1–25.
http://www.jstatsoft.org/v63/i19/

Lindgren, F., Rue, H. & Lindström, J. (2011), 'An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(4), 423–498.

Little, L. R., Wayte, S. E., Tuck, G. N., Smith, A. D., Klaer, N., Haddon, M., Punt, A. E., Thomson, R., Day, J. & Fuller, M. (2011), 'Development and evaluation of a cpue-based harvest control rule for the southern and eastern scalefish and shark fishery of Australia', *ICES Journal of Marine Science* **68**(8), 1699–1705.

Livingston, M, E., Bull, B. & Stevens, D, W. (2004), Migration patterns during the life-cycle of hoki (*Macruronus novaezelandiae*): an analysis of trawl survey data in New Zealand waters 1991—2002, Technical report, Final Research Report for Ministry of Fisheries Research Project HOK2000/01 Objective 6. (Unpublished report held by Ministry of Fisheries, Wellington.

Lohr, S. L. (2009), *Sampling: design and analysis*, Nelson Education.

Mace, P. M. & Doonan, I. (1988), A generalised bioeconomic simulation model for fish population dynamics, Technical report, MAFFish, NZ Ministry of Agriculture and Fisheries.

MacLean, J. & Evans, D. (1981), 'The stock concept, discreteness of fish stocks, and fisheries management', *Canadian Journal of Fisheries and Aquatic Sciences* **38**(12), 1889–1898.

Madow, W. G. & Madow, L. H. (1944), 'On the theory of systematic sampling, i', *The Annals of Mathematical Statistics* **15**(1), 1–24.

Marsh, C. (2022), 'C++ Agent Based Model (CABM) User Manual', https://github.com/Craig44/CABM/blob/master/Documentation/UserManual/CABM.pdf.

Maunder, M. N. & Piner, K. R. (2014), 'Contemporary fisheries stock assessment: many issues still remain', *ICES Journal of Marine Science* **72**(1), 7–18.

Maunder, M. N. & Punt, A. E. (2004), 'Standardizing catch and effort data: a review of recent approaches', *Fisheries Research* **70**(2-3), 141–159.

Maunder, M. N. & Punt, A. E. (2013), 'A review of integrated analysis in fisheries stock assessment', *Fisheries Research* **142**, 61–74.

Maunder, M. N., Sibert, J. R., Fonteneau, A., Hampton, J., Kleiber, P. & Harley, S. J. (2006), 'Interpreting catch per unit effort data to assess the status of individual stocks and communities', *ICES Journal of Marine Science* **63**(8), 1373–1385.

Maunder, M. N., Thorson, J. T., Xu, H., Oliveros-Ramos, R., Hoyle, S. D., Tremblay-Boyer, L., Lee, H. H., Kai, M., Chang, S.-K., Kitakado, T. et al. (2020), 'The need for spatio-temporal modeling to determine catch-per-unit effort based indices of abundance and associated composition data for inclusion in stock assessment models', *Fisheries Research* **229**, 105594.

McCullagh, P. & Nelder, J. A. (2019), *Generalized linear models*, Routledge.

McGarvey, R., Burch, P. & Matthews, J. M. (2016), 'Precision of systematic and random sampling in clustered populations: habitat patches and aggregating organisms', *Ecological Applications* **26**(1), 233–248.

McKenzie, A. (2017), 'Assessment of hoki (*Macruronus novaezelandiae*) in 2016', *New Zealand Fisheries Assessment Report* **11**.

McKenzie, J., Marsh, C. & Langley, A. (2021), 'Assessment implications of time-varying R0 for SNA 8 parameterised stock', Requst from info@mpi.govt.nz. Working group report: FNZ Project SNA201903A.

Methot Jr, R. D. & Wetzel, C. R. (2013), 'Stock synthesis: a biological and statistical framework for fish stock assessment and fishery management', *Fisheries Research* **142**, 86–99.

Millar, R. B. (2011), *Maximum likelihood estimation and inference: with examples in R, SAS and ADMB*, Vol. 111, John Wiley & Sons.

Millar, R. B. & Olsen, D. (1995), 'Abundance of large toheroa (*Paphies ventricosa Gray*) at Oreti Beach, 1971–90, estimated from two-dimensional systematic samples', *New Zealand Journal of Marine and Freshwater Research* **29**(1), 93–99.

Montero, J.-M., Fernández-Avilés, G. & Mateu, J. (2015), *Spatial and spatio-temporal geostatistical modeling and kriging*, John Wiley & Sons.

Muff, S., Signer, J. & Fieberg, J. (2020), 'Accounting for individual-specific variation in habitat-selection studies: Efficient estimation of mixed-effects models using bayesian or frequentist computation', *Journal of Animal Ecology* **89**(1), 80–92.

Nottingham, C. D. (2021), Geostatistical tools supporting improved management practices for the New Zealand surfclam fishery, PhD thesis, The University of Auckland.

O'Driscoll, R., MacGibbon, D., Fu, D., Lyon, W. & 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**.

Pascoe, S., Bustamante, R., Wilcox, C. & Gibbs, M. (2009), 'Spatial fisheries management: a framework for multi-objective qualitative assessment', *Ocean & Coastal Management* **52**(2), 130–138.

Pati, D., Reich, B. J. & Dunson, D. B. (2011), 'Bayesian geostatistical modelling with informative sampling locations', *Biometrika* **98**(1), 35–48.

Pennington, M. & Volstad, J. H. (1994), 'Assessing the effect of intra-haul correlation and variable density on estimates of population characteristics from marine surveys', *Biometrics* pp. 725–732.

Pennino, M. G., Paradinas, I., Illian, J. B., Muñoz, F., Bellido, J. M., López-Quílez, A. & Conesa, D. (2019), 'Accounting for preferential sampling in species distribution models', *Ecology and Evolution* **9**(1), 653–663.

Phillips, J. S., Gupta, A. S., Senina, I., van Sebille, E., Lange, M., Lehodey, P., Hampton, J. & Nicol, S. (2018), 'An individual-based model of skipjack tuna (Katsuwonus pelamis) movement in the tropical Pacific ocean', *Progress in Oceanography* **164**, 63–74.

Poos, J., Bogaards, J., Quirijns, F., Gillis, D. & Rijnsdorp, A. (2010), 'Individual quotas, fishing effort allocation, and over-quota discarding in mixed fisheries', *ICES Journal of Marine Science* **67**(2), 323–333.

Punt, A. E. (2003), 'The performance of a size-structured stock assessment method in the face of spatial heterogeneity in growth', *Fisheries Research* **65**(1-3), 391–409.

R Core Team (2020), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
https://www.R-project.org/

Reynolds, R. W., Smith, T. M., Liu, C., Chelton, D. B., Casey, K. S. & Schlax, M. G. (2007), 'Daily high-resolution-blended analyses for sea surface temperature', *Journal of Climate* **20**(22), 5473–5496.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W. et al. (2017), 'Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure', *Ecography* **40**(8), 913–929.

Rufener, M.-C., Kristensen, K., Nielsen, J. R. & Bastardie, F. (2021), 'Bridging the gap between commercial fisheries and survey data to model the spatiotemporal dynamics of marine species', *Ecological Applications* pp. 1–15.

Salas, S. & Gaertner, D. (2004), 'The behavioural dynamics of fishers: management implications', *Fish and Fisheries* **5**(2), 153–167.

Sampson, D. B. (1991), 'Fishing tactics and fish abundance and their influence on catch rates', *ICES Journal of Marine Science* **48**, 291–201.

Sampson, D. B. (1992), 'Fishing technology and fleet dynamics: predictions from a bioeconomic model', *Marine Resource Economics* **7**(1), 37–58.

Schaefer, M. B. (1943), 'The theoretical relationship between fishing effort and mortality', *Copeia* **1943**(2), 79–82.

Schnute, J. (1981), 'A versatile growth model with statistically stable parameters', *Canadian Journal of Fisheries and Aquatic Sciences* **38**(9), 1128–1140.

Scudilio, J. & Pereira, G. H. (2020), 'Adjusted quantile residual for generalized linear models', *Computational Statistics* **35**(1), 399–421.

Shelton, A. O. & Mangel, M. (2012), 'Estimating von Bertalanffy parameters with individual and environmental variations in growth', *Journal of Biological Dynamics* **6**, 3–30.

Shelton, A. O., Thorson, J. T., Ward, E. J. & Feist, B. E. (2014), 'Spatial semiparametric models improve estimates of species abundance and distribution', *Canadian Journal of Fisheries and Aquatic Sciences* **71**(11), 1655–1666.

Simmonds, E. J. & Fryer, R. J. (1996), 'Which are better, random or systematic acoustic surveys? a simulation using north sea herring as an example', *ICES Journal of Marine Science* **53**(1), 39–50.

Simpson, D., Illian, J. B., Lindgren, F., Sørbye, S. H. & Rue, H. (2016), 'Going off grid: Computationally efficient inference for log-Gaussian Cox processes', *Biometrika* **103**(1), 49–70.

Sippel, T., Lee, H. H., Piner, K. & Teo, S. L. (2017), 'Searching for M: Is there more information about natural mortality in stock assessments than we realize?', *Fisheries Research* **192**, 135–140.

Smith, M. D. (2005), 'State dependence and heterogeneity in fishing location choice', *Journal of Environmental Economics and Management* **50**(2), 319–340.

Smith, S. J. (1990), 'Use of statistical models for the estimation of abundance from groundfish trawl survey data', *Canadian Journal of Fisheries and Aquatic Sciences* **47**(5), 894–903.

Smith, T. M. F. (1976), 'The foundations of survey sampling: a review', *Journal of the Royal Statistical Society: Series A (General)* **139**(2), 183–195.

Ståhl, G., Saarela, S., Schnell, S., Holm, S., Breidenbach, J., Healey, S. P., Patterson, P. L., Magnussen, S., Næsset, E., McRoberts, R. E. et al. (2016), 'Use of models in large-area forest surveys: comparing model-assisted, model-based and hybrid estimation', *Forest Ecosystems* **3**(1), 1–11.

Sterba, S. K. (2009), 'Alternative model-based and design-based frameworks for inference from samples to populations: From polarization to integration', *Multivariate Behavioral Research* **44**(6), 711–740.

Stevens, D., O'Driscoll, R., Ballara, S. & Schimel, A. (2021), 'Trawl survey of hoki and middle depth species on the Chatham Rise, January 2020 (TAN2001)', *New Zealand Fisheries Assessment Report* **33**.

Stewart, I. J. & Monnahan, C. C. (2017), 'Implications of process error in selectivity for approaches to weighting compositional data in fisheries stock assessments', *Fisheries Research* **192**, 126–134.

Strand, G.-H. (2017), 'A study of variance estimation methods for systematic spatial sampling', *Spatial Statistics* **21**, 226–240.

Strømme, T. & Iilende, T. (2001), 'Precision in systematic trawl surveys as assessed from replicate sampling by parallel trawling off namibia', *South African Journal of Marine Science* **23**(1), 385–396.

Tenningen, M., Slotte, A. & Skagen, D. (2011), 'Abundance estimation of Northeast Atlantic mackerel based on tag recapture data—A useful tool for stock assessment?', *Fisheries Research* **107**(1-3), 68–74.

Thorson, J. T. (2018), 'Three problems with the conventional delta-model for biomass sampling data, and a computationally efficient alternative', *Canadian Journal of Fisheries and Aquatic Sciences* **75**(9), 1369–1382.

Thorson, J. T. (2019), 'Guidance for decisions using the Vector Autoregressive Spatio-Temporal (VAST) package in stock, ecosystem, habitat and climate assessments', *Fisheries Research* **210**, 143–161.

Thorson, J. T. & Kristensen, K. (2016), 'Implementing a generic method for bias correction in statistical models using random effects, with spatial and population dynamics examples', *Fisheries Research* **175**, 66–74.

Thorson, J. T., Maunder, M. N. & Punt, E. (2020), 'The development of spatio-temporal models of fishery catch-per-unit-effort data to derive indices of relative abundance', *Fisheries Research* **230**.

Thorson, J. T., Shelton, A. O., Ward, E. J. & Skaug, H. J. (2015), 'Geostatistical delta-generalized linear mixed models improve precision for estimated abundance indices for West Coast groundfishes', *ICES Journal of Marine Science* **72**(5), 1297–1310.

Thygesen, U. H., Albertsen, C. M., Berg, C. W., Kristensen, K. & Nielsen, A. (2017), 'Validation of ecological state space models using the Laplace approximation', *Environmental and Ecological Statistics* **24**(2), 317–339.

Tierney, L., Kass, R. E. & Kadane, J. B. (1989), 'Fully exponential laplace approximations to expectations and variances of nonpositive functions', *Journal of the American Statistical Association* **84**(407), 710–716.

Tobler, W. R. (1970), 'A computer movie simulating urban growth in the detroit region', *Economic Geography* **46**, 234–240.

Truesdell, S. B., Hart, D. R. & Chen, Y. (2017), 'Effects of unequal capture probability on stock assessment abundance and mortality estimates: an example using the US Atlantic sea scallop fishery', *Canadian Journal of Fisheries and Aquatic Sciences* **74**(11), 1904–1917.

Van Putten, I. E., Kulmala, S., Thébaud, O., Dowling, N., Hamon, K. G., Hutton, T. & Pascoe, S. (2012), 'Theories and behavioural drivers underlying fleet dynamics models', *Fish and Fisheries* **13**(2), 216–235.

Vignaux, M. (1996), 'Analysis of vessel movements and strategies using commercial catch and effort data from the New Zealand hoki fishery', *Canadian Journal of Fisheries and Aquatic Sciences* **53**, 2126—-2136.

Walline, P. D. (2007), 'Geostatistical simulations of eastern Bering Sea walleye pollock spatial distributions, to estimate sampling precision', *ICES Journal of Marine Science* **64**(3), 559–569.
https://doi.org/10.1093/icesjms/fsl045

Webber, D. (2015), Modelling Complexity and Uncertainty in Fisheries Stock Assessment, PhD thesis, Victoria University of Wellington.

Wheeler, A. P., Steenbeek, W. & Andresen, M. A. (2018), 'Testing for similarity in area-based spatial patterns: Alternative methods to Andresen's spatial point pattern test', *Transactions in GIS* **22**(3), 760–774.

Williams, P. J., Hooten, M. B., Womble, J. N., Esslinger, G. G. & Bower, M. R. (2018), 'Monitoring dynamic spatio-temporal ecological processes optimally', *Ecology* **99**(3), 524–535.

Winger, P. D., Walsh, S. J., He, P. & Brown, J. A. (2004), 'Simulating trawl herding in flatfish: the role of fish length in behaviour and swimming characteristics', *ICES Journal of Marine Science* **61**(7), 1179–1185.

Wolter, K. M. (1984), 'An investigation of some estimators of variance for systematic sampling', *Journal of the American Statistical Association* **79**(388), 781–790.

Wolter, K. M. (2007), *Introduction to variance estimation*, Statistics for Social and Behavioral Sciences, 2nd edn, Springer, New York.

Wood, S. N. (2017), *Generalized Additive Models: An Introduction with R*, CRC press.
https://doi.org/10.1201/9781315370279

Xu, H., Lennert-Cody, C. E., Maunder, M. N. & Minte-Vera, C. V. (2019), 'Spatiotemporal dynamics of the dolphin-associated purse-seine fishery for yellowfin tuna (*Thunnus albacares*) in the eastern Pacific Ocean', *Fisheries Research* **213**, 121–131.

Index