Effects of stratification of survey design on cetacean abundance estimation using species distribution models

層別化調査が種の分布モデルによる鯨類個体数推定値に及ぼす影響

Department of Risk management and Environmental Science

Graduate school of Environment and Information Science.

Yokohama National University

Yasutoki Shibata

Index

List of parameters

$\hat{N}$: Estimated abundance

N_{true} : The true abundance that was either 2000 or 6000

n : Number of detections

$\hat{S}$: Estimated mean group size

L : The total length of the transect lines

$\hat{w}$: The estimated effective strip half-width

A : The size of the study area

l_i : The length of segment i

θ : The parameter to be estimated

f : The function to be estimated

v : Grid cell

P_v : The probability of generating an individual school within grid cell v when the cell is selected

λ_v : The number of schools determined by the generating model at grid cell v

e : The ID of generating equations

ϕ : The ID of estimating equations

α, m, σ : The parameters to be estimated in the estimation procedure

ω : Partial regression coefficient

s : One dimensional smoothing function with smoothing parameters selected by generalized cross validation

SST : Sea surface temperature

MTEM : Mean water temperature from 0 to 200m

MSAL : Mean salinity from 0 to 200m

LAT : Latitude

d : The ID of stratum

RB : Relative bias

RSD : Relative standard error

RRMSE : Relative root mean square error

I. Introduction

History and development of species distribution model

An interesting question of how living organisms such as plants, birds, fish and mammals are distributed in space has a long history which has inspired many ecologists to seek explanations. Most modelling approaches developed for estimating living organism species distributions have their roots by quantifying relationships between species distribution and environment covariates. Three phases seem to have marked the history of species distribution model (SDM) [1]: (i) non-spatial statistical quantification of species environment relationship based on empirical data, (ii) expert-based (non-statistical, non-empirical) spatial modelling of species distribution, and (iii) spatially explicit statistical and empirical modelling of species distribution. An example of modelling approach of an early date which using correlations between species distributions and climate seems to be those of Johnston (1924), estimating the invasive spread of a cactus species in Australia [2]. Earliest developments in computer based modelling of species distribution seem to be started in the mid-1970s, stimulated by numerous environment covariates available at that time [3]. To the best of my knowledge, the earliest species distribution modelling attempt seems to be the niche

based estimation of spatial distribution of crop species by Henry Nix and collaborators in Australia [4].

In the early 1980s, there were parallel developments in computer and statistical sciences and strong theoretical support to ecology [5]. As a result, the number of related publications increased very much. A synthesis review of SDM can be found in Guisan & Zimmermann 2000) [6]. In recent years, modelling of species distribution has become an increasingly important tool to address various issues in ecology, biogeography, evolution and, more recently, in conservation biology and climate change research ([7] [8] [9] [10] [11]).

SDMs are models relating field observations to covariates based on statistically or theoretically derived response surfaces [6]. Species data could be presence/absence or abundance observations based on random sampling scheme [12]. Covariates can exert direct or indirect effects on species distribution, and are optimally chosen to reflect three main types of influences on the species distribution [1]: (i) limiting factors (or regulators), defined as factors controlling species eco-physiology (e.g. temperature, water, soil composition); (ii) disturbances, defined as all types of perturbations affecting environmental systems (natural or human-induced) and (iii) resources, defined as all compounds that can be assimilated by organisms (e.g. energy and water). Relationships

between species distribution and covariates could cause different types of spatial patterns to be observed at different scales, often in a hierarchical manner [13]. For more details on SDM building, refer to Guisan & Zimmermann (2000) [6].

Abundance estimation and species distribution model

Effective management and conservation strategies of vulnerable species and exploited bioresources require reliable information on abundance [14]. Information on abundance permits evaluation of the threats of the target bioresource, which is a necessary step in the establishment of successful conservation measures [15]. Conventionally, line-transect (LT) estimator have been widely used to estimate abundance of marine and terrestrial mammals, birds and plants [16] [17] [18] [19]. The estimator is assumed that the tracklines are set at random in a study area where any point has an equal probability of being surveyed, so it has been called a design-based method [20] [21]. However, the estimator has a difficulty in quantifying the relationships between the distribution of animals and environmental variables such as topography and temperature. This is because the spatial resolution of estimated abundance is very low (e.g. in thousands of km^2), compared to spatial resolution of environmental factors [22]. This is true at least for cetaceans [22].

In contrast, an analysis method exists in which inference is made by using a relationship between the spatial distribution of animals and that of covariates [22] [20]. This is known as a model-based method [20] [21] or SDM where the response variable may be count data [1]. SDMs have also been applied to data from a random sampling scheme [1], such as a LT survey for abundance estimation of cetaceans [14] [22] [20]. In LT surveys, the estimates arising from survey stratification are intended to produce estimates of abundance with smaller variance than non-stratified surveys [16]. Therefore abundance estimation from SDMs for cetaceans is often based on stratified LT survey as well.

Some simulation studies have considered effects of the stratification on estimation performance of the probability of presence in grid cells with logistic regression [23], [24]. However, these studies considered effect of the stratification on the probability but abundance estimation and in cases where mis-specification of covariates is occurred. Here, mis-specification of covariates is defined as a situation that the true covariates defining the spatial distribution of living organisms are not used for estimation with SDMs. I cannot always use the true covariates that define spatial distribution as explanatory variables. Mis-specification of covariates can adversely affect hypothesis tests of association between a response variable and covariates [25].

Abundance estimation is unbiased with a small variance is desired. Therefore, evaluations of abundance estimation performance should be based on an index combining the bias and variance [26] such as root mean square error (RMSE). In our study, RMSE is defined as the root of the sum of the square of the bias (the expected difference between estimated parameter and the true parameter) and variance of the estimated parameter (below equation (18)). RMSE shows “risk” of estimated parameter and smaller risk implies better estimation [27]. The bias and RMSE can be identified only when the true abundance is known. Therefore simulation study is indispensable to evaluate bias and RMSE in abundance estimation

Our aims is to examine effects of survey design stratification in terms of relative bias [28], relative standard deviation [29] and relative root mean squared error [30] of SDMs with mis-specified covariates when estimating school abundance thorough a LT survey. In the second chapter, an abundance estimation method by a fundamental definition of LT estimator and by using SDM was defined by equations, and a situation and its problem about mis-specification of covariates was explained. The third chapter considered the influence of when a survey design is changed in the situation where mis-specification of covariates has occurred. Comprehensive consideration was performed in Chapter 4.

II. Line transect estimator and species distribution model

Here I briefly outline the principles of a LT survey. In a LT survey, observers travel at constant speed on designed tracklines and can use binoculars to sight animal (or object) groups. Group size, detection distances and angles to sighted animal groups are recorded. Detection distances and angles are converted into perpendicular distances to the transect lines. An effective strip half-width [16] can be estimated from a detection function of the perpendicular distances. The conventional LT estimator of abundance $\hat{N}^{(LT)}$ is described as;

$$\hat{N}^{(LT)} = \frac{n\hat{S}}{2\hat{w}L}A, \qquad (1)$$

where n is the number of detections, $\hat{S}$ is the estimated mean group size, L is the total length of the transect lines, $\hat{w}$ is the estimated effective strip half-width and A indicates the size of the study area. Here I assumed detection probability is 1.

In the case of abundance estimation by SDM, researcher estimates a relationship between covariates as explanatory variables and observed number of animals in unit grid cell as response variables from sample data (e.g. on surveyed tracklines). Then researcher can estimate number of animals at Non-surveyed grid cells in study area if value of covariates were obtained there from such as a satellite data. Abundance can be estimated by summed up the number of animals in each grid cells.

Here I show an equation as an example of SDM as below. Suppose that the total length of tracklines was divided into *I* small contiguous segments. Let the length of segment *i* be l_i. I denote the expected observed number of animals within segment *i*, by $E(n_i)$, which can be modelled with covariates by using SDM with Poisson distribution.

$$E(n_i) = \exp\left(\ln(2l_i w) + \theta + \sum_k f_k(x_{k,i}) \right), \quad (2)$$

where the offset variable $2l_i w$ is the rectangle of segment *i*. θ and f_k are parameters and functions to be estimated. Each $x_{k,i}$ is the covariate in segment *i* of the *k* th covariates. If the study area could be divided into small grid cell *v*, then, I can estimate the expected number of schools $\hat{N}_v^{(SDM)}$ by using $\hat{\theta}$, $\hat{f}_k$ estimated from equation 9-10 and the covariates $x_{k,v}$.The estimated abundance in the whole study area $\hat{N}^{(SDM)}$ for SDMs was obtained by summation of $\hat{N}_v^{(SDM)}$.

III. Effects of stratification and mis-specification of covariates in SDM for abundance estimation from virtual line transect survey data

Data for generating simulation scenarios

A data set obtained through the following survey was used as a reference data set to generate simulation data and examine the relative performance of the LT estimator and the SDMs. The cetacean sighting survey was conducted in the sector of the Southern Ocean south of the Indian Ocean in the austral summer of 1999/2000. The survey was conducted as a part of the Japanese Whale Research Program under Special Permit in the Antarctic (JARPA). The annual JARPA surveys were carried out over an austral summer, usually from December through to March. The survey area was divided into northern and southern strata. Boundary between southern and northern strata was set at about 83 km from the sea ice edge. It was reported that Antarctic minke whales were densely aggregated along sea ice edge while they are sporadic in the north [31]. Sparsely and densely zigzag tracklines were set in the northern and southern stratum, respectively, according to the distribution pattern of Antarctic minke whales. The starting points of tracklines were selected randomly. Sighting survey vessels (SVs)

conducted the cetacean sightings, oceanographic observations and a hydroacoustic survey. Distance and angle from the vessels to sighted Antarctic minke whale were recorded. The nominal steaming speed of SVs on the tracklines was 10.5 knots. The details of the JARPA survey were written by e.g. [32], [33]. The area bounded by 60° S and the ice edge (approximately 65° S) from 110° E to 130° E was used as a target of area for our simulations since the shape of this area was almost rectangle that could be avoided considering of scraggly ice edge lines. I used data of sightings within 2.8 km from the tracklines and then there are a total of 201 sighted schools in 3,435 km of sighting effort. The 2.8 km distance was used as a right truncation distance of Antarctic minke whale in JARPA [34].

I chose five continuous variables as candidates for using simulation; sea surface temperature (*SST*), mean water temperature from 0 to 200 m (*MTEM*), mean salinity from 0 to 200 m (*MSAL*), depth (*DEPTH*) and latitude (*LAT*). I removed *DEPTH* and *MSAL* from our analysis because *DEPTH* had relatively high correlation with other factors especially with *SST* and *MSAL* had a weak correlation to the number of school.

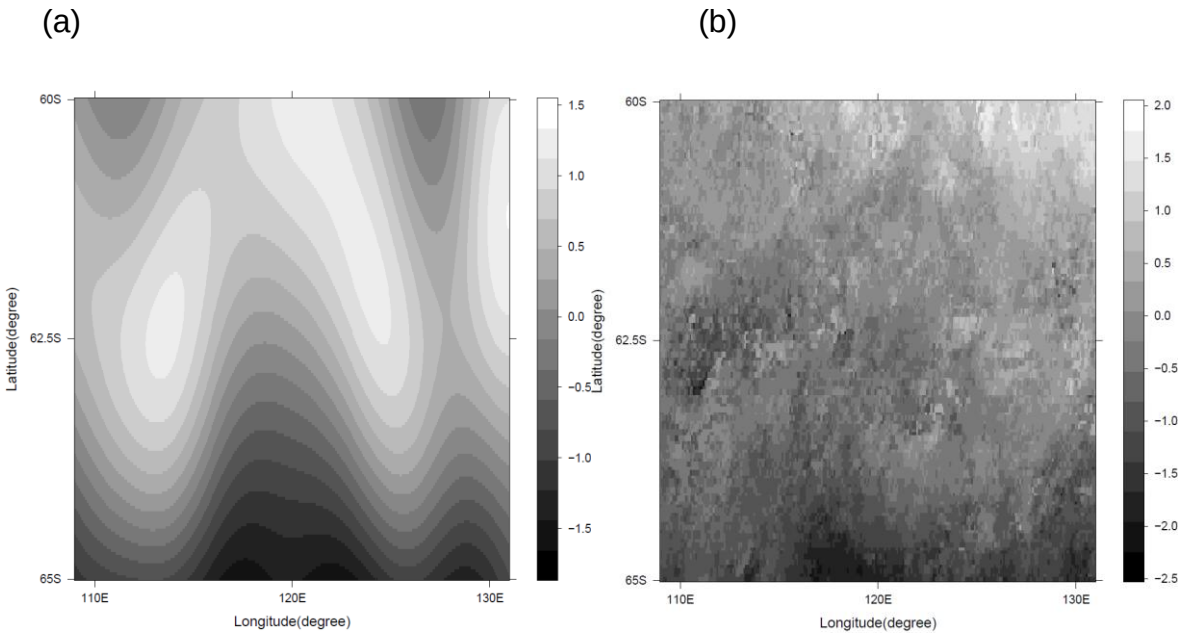


Fig. 1. The distribution of covariates value in virtual space, (a) mean water temperature from 0 to 200 m estimated by ordinary Kriging and (b) mean sea surface temperature from December to March in 1999/2000.

SST data were obtained from NOAA Pathfinder AVHRR version 5 of PO.DAAC Ocean ESIP Tool (POET, http://poet.jpl.nasa.gov/) on a 4 by 4 km grid resolution. The mean *SST* at each grid cell was represented by the mean *SST* from December to March from daytime monthly data in 1999/2000. *MTEM* was calculated using data obtained from a conductivity, temperature and depth profiler (CTD) and expendable conductivity, temperature and depth profiler (XCTD). *MTEM* has been used as an indicator of distribution of Antarctic krill [35] [36] whose spatial distribution of

aggregation is correlated with that of the Antarctic minke whale [32]. *MTEM* of the survey area was estimated using ordinary Kriging using 3 years moving average of CTD and XCTD data from the JARPA survey (1995/1996, 1999/2000, 2003/2004) by the Geostatistical Analyst extension of ArcGIS v9.3 {ESRI, 2009 #425} (Fig. 1).

Generation of dummy data

Three covariates were used for generating spatial distributions of schools; *MTEM*, *SST* and *LAT*. I divided the virtual survey area into 77,284 small grid cells and the area of each grid cell *v* was 4 by 4 km that was matched by resolution of *SST*. A resolution of *MTEM* and *LAT* were also given as 4 by 4 km here. The area of grid cell *v* at the north and east boundaries of a virtual survey area (see below "Assumed survey area and designs" section) was corrected because I could not divided the lengths of 556 km nor 2222 km by 4 km. The probability of generating an individual school within grid cell *v* when that cell is selected, denoted by $P_v^{(e)}$, was given as

$$P_v^{(e)} = \frac{\exp[\log(\lambda_v^{(e)})(1+0.1\varepsilon_v^{(e)})]}{\max(\exp[\log(\lambda_v^{(e)})(1+0.1\varepsilon_v^{(e)})])}, \qquad (3)$$

where $\lambda_v^{(e)}$ indicates the logarithm of a value determined by the generating model (see below equations (4-9)) at the grid cell *v* given by the generating equation. (*e*=4, 5,…, 9);

$\varepsilon_v^{(e)}$ represents an independent standardized normal variable whose mean and S.D. were set as 0 and 1, respectively. I assumed process uncertainty in generating schools because the school density may be affected other environmental factors that are not considered here. I iterated 120 times below the estimation procedure by adding random values in equation (3). $\lambda_v^{(e)}$ was described respectively as

$$\log\left(\lambda_v^{(4)}\right) = \theta_4 + \alpha_{41} MTEM_v + \alpha_{42} SST_v, \quad (4)$$

$$\log\left(\lambda_v^{(5)}\right) = \theta_5 + \alpha_5 SST_v, \quad (5)$$

$$\log\left(\lambda_v^{(6)}\right) = \theta_6 + \alpha_6 LAT_v, \quad (6)$$

$$\log\left(\lambda_v^{(7)}\right) = \theta_7 + \frac{\alpha_{71}}{\sqrt{2\pi}\sigma_{71}} \exp\left(-\frac{\left(MTEM_v - m_{71}\right)^2}{2{\sigma_{71}}^2}\right), + \frac{\alpha_{72}}{\sqrt{2\pi}\sigma_{72}} \exp\left(-\frac{\left(SST_v - m_{72}\right)^2}{2{\sigma_{72}}^2}\right) \quad (7)$$

$$\log\left(\lambda_v^{(8)}\right) = \theta_8 + \frac{\alpha_8}{\sqrt{2\pi}\sigma_8} \exp\left(-\frac{\left(SST_v - m_8\right)^2}{2{\sigma_8}^2}\right), \quad (8)$$

$$\log\left(\lambda_v^{(9)}\right) = \theta_9 + \frac{\alpha_9}{\sqrt{2\pi}\sigma_9} \exp\left(-\frac{\left(LAT_v - m_9\right)^2}{2{\sigma_9}^2}\right), \quad (9)$$

where $MTEM_v$, SST_v and LAT_v were covariates obtained at the centre of grid cell *v*; $\theta_4, \theta_5, \theta_6, \theta_7, \theta_8, \theta_9$ were the terms of intercept and $\alpha_{41}, \alpha_{42}, \alpha_5, \alpha_6$ were partial regression coefficients, $\alpha_{71}, \alpha_{72}, \alpha_8, \alpha_9$ were scale parameters, m_{71}, m_{72}, m_8, m_9 and $\sigma_{71}, \sigma_{72}, \sigma_8, \sigma_9$ were the expected values and standard deviation of the normal random

variables, respectively. All other distributions from combinations among *MTEM*, *SST* and *LAT* were similar to the distributions from equations 4-9. I chose combinations of *MTEM + SST*, *SST* and *LAT* because these three were enough to express the difference distribution each other. Here I had fitted the equations 3-8 to the above satellite and JARPA data by using Poisson regression with log-link function where response variable was the number of school on the *i*th area on the surveyed tracklines. We showed the theoretical distributions of $\lambda_v^{(e)}$ (Fig. 2). Effective strip width was calculated approximately 0.6 and 1.4 km in northern and southern stratum, respectively. The tracklines were divided into 20 km equidistance segments and 201 sighted schools in 3,435 km of sighting effort. Therefore *i*th area on the trackline was constituted by 20×2×0.6 km in the northern stratum, 20×2×1.4 km in the southern stratum, respectively. Then estimated parameters with equations (4-9) were used into equation 3 to generate simulated spatial distributions of schools. The parameter values are shown in Table 1 and those were used for generating spatial distributions of schools.

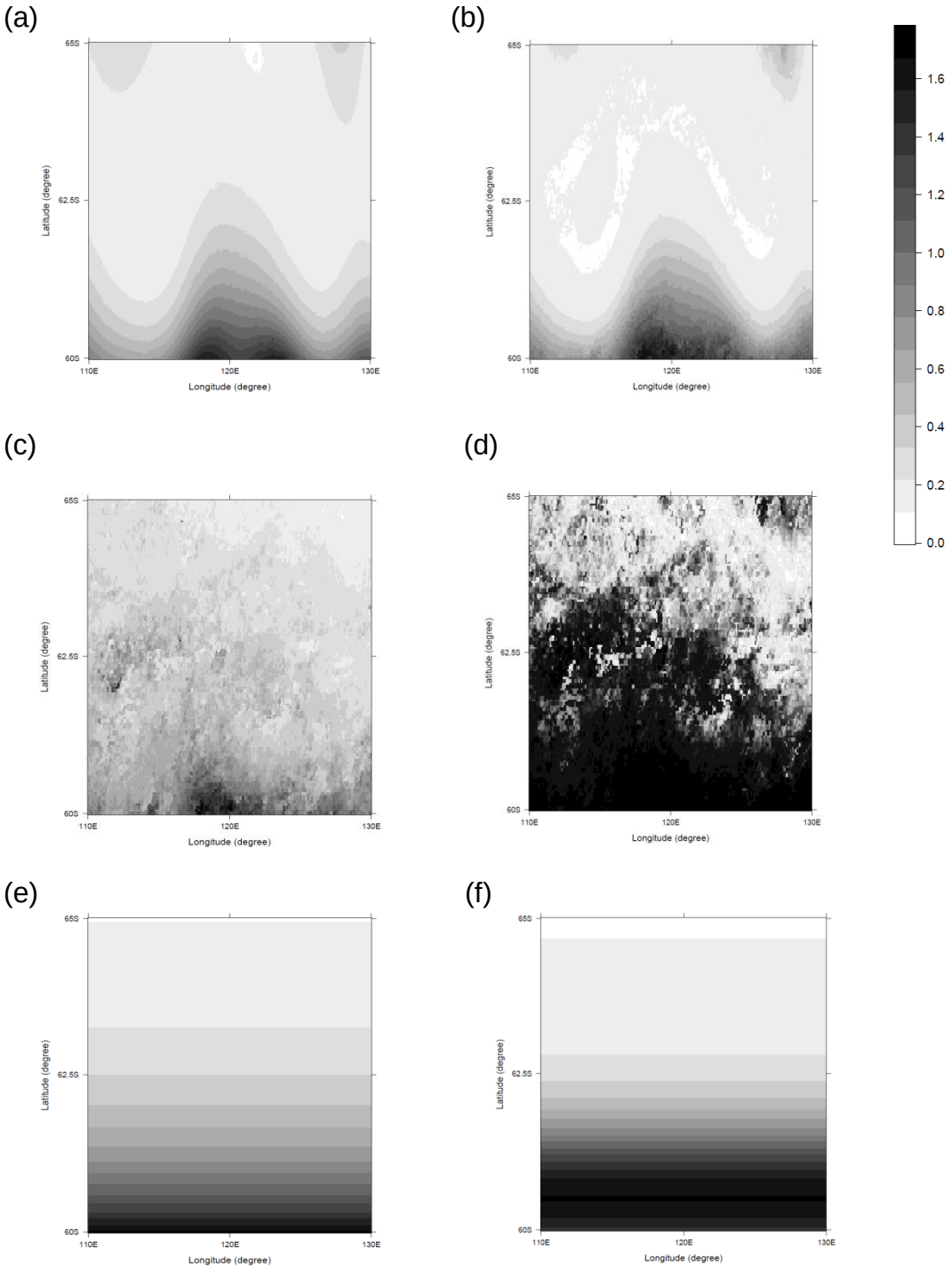


Fig. 2. The theoretical distributions of $\lambda_v^{(e)}$. The alphabets from (a) to (f) are corresponded to $\lambda_v^{(3)}$ to $\lambda_v^{(8)}$, respectively.

Table 1. The values of estimated parameters from JARPA. All the parameters have converged.

θ	Value	α	Value	m	Value	σ	Value
θ_4	-10.81	α_{41}	-1.17				
θ_5	-10.19	α_{42}	-0.06				
θ_6	-8.59	α_5	-0.96				
θ_7	2.00	α_6	-0.01				
θ_8	0.25	α_{71}	-7.63	m_{71}	0.87	σ_{71}	0.93
θ_9	-2.28	α_{72}	-6.30	m_{72}	-0.41	σ_{72}	1.70
		α_8	-2.78	m_8	0.51	σ_8	0.36
		α_9	594.86	m_9	30.79	σ_9	85.00

Responses of the spatial distribution of schools to the environmental covariates were allowed to take one of two shapes: linear or Gaussian [37]. A linear response (equations 4-6) was characterized by a steady increase or decrease in the probability of occurrence, while a Gaussian (equations 7-9), as a Non-linear response, responds to the particular environmental factor with a symmetrical and decreasing probability of occurrence around a mean value [37].

The virtual spatial distributions of schools were generated by using the

rejection sampling algorithm [38]: First, I selected grid cell *v* in the virtual space. I calculated $P_v^{(e)}$ at the grid cell *v* and if a random number which is drawn from the uniform distribution $\mathrm{U}(0,1)$, was smaller than the $P_v^{(e)}$ a school occurred. The detail position of generated school is randomly and homogeneously selected within the grid cell. This procedure was iterated until the number of generated schools became the given abundance (either 2000 or 6000, see below).

Assumed survey area and designs

The study area was fixed as an about 556 km × 2222 km rectangle, whose area approximately corresponded to the area between 60° S and 65° S, and between 110° E and 130° E. Two patterns of simulated school abundances (2000 and 6000 schools) were set for evaluating the difference of abundance on stratification that were based on an approximate minimum and maximum number of schools through 1991/92-2003/04 [34] within the scaled area (from 110° E to 130° E). The period was corresponding to the start and end year of JARPA survey period in the area by same stratified sampling scheme. Therefore I could evaluate the effect of school abundance under plausible condition. Effort allocation was modelled using three different survey designs: single

stratum, two and three strata as shown in Fig. 3. Here, the boundaries in the case of two and three strata were divided by 93 and 278 km from the south. Those divisions and effort allocation within strata were based on the past JARPA surveys (e.g. [34], [39]). It was reported that Antarctic minke whales were densely aggregated along sea ice edge while they are sporadic in the north [31]. In JARPA, therefore, sparsely and densely zigzag tracklines were set in the northern and southern strata, respectively, according to the distribution pattern of Antarctic minke whales. Parallel sets of vertical tracklines towards longitude with equally-spaced intervals were prepared for each stratum and each set of tracklines has 2,780 km total length that was based on the 1999/00 JARPA survey. The place of tracklines was randomly changed (systematic random sampling) each iteration where its total effort and intervals were fixed. I omitted the process of estimating a detection function to focus on the effects of survey design stratification and mis-specification of covariates and assumed a fixed distance that all schools could be counted. Here all schools within 2.8 km perpendicular distance from tracklines on both sides were counted. The single stratum survey design was considered as a control experiment.

The tracklines were divided into 5 km segments which were used a previous study of spatial modelling of Antarctic minke whale abundance in the Antarctic [22].

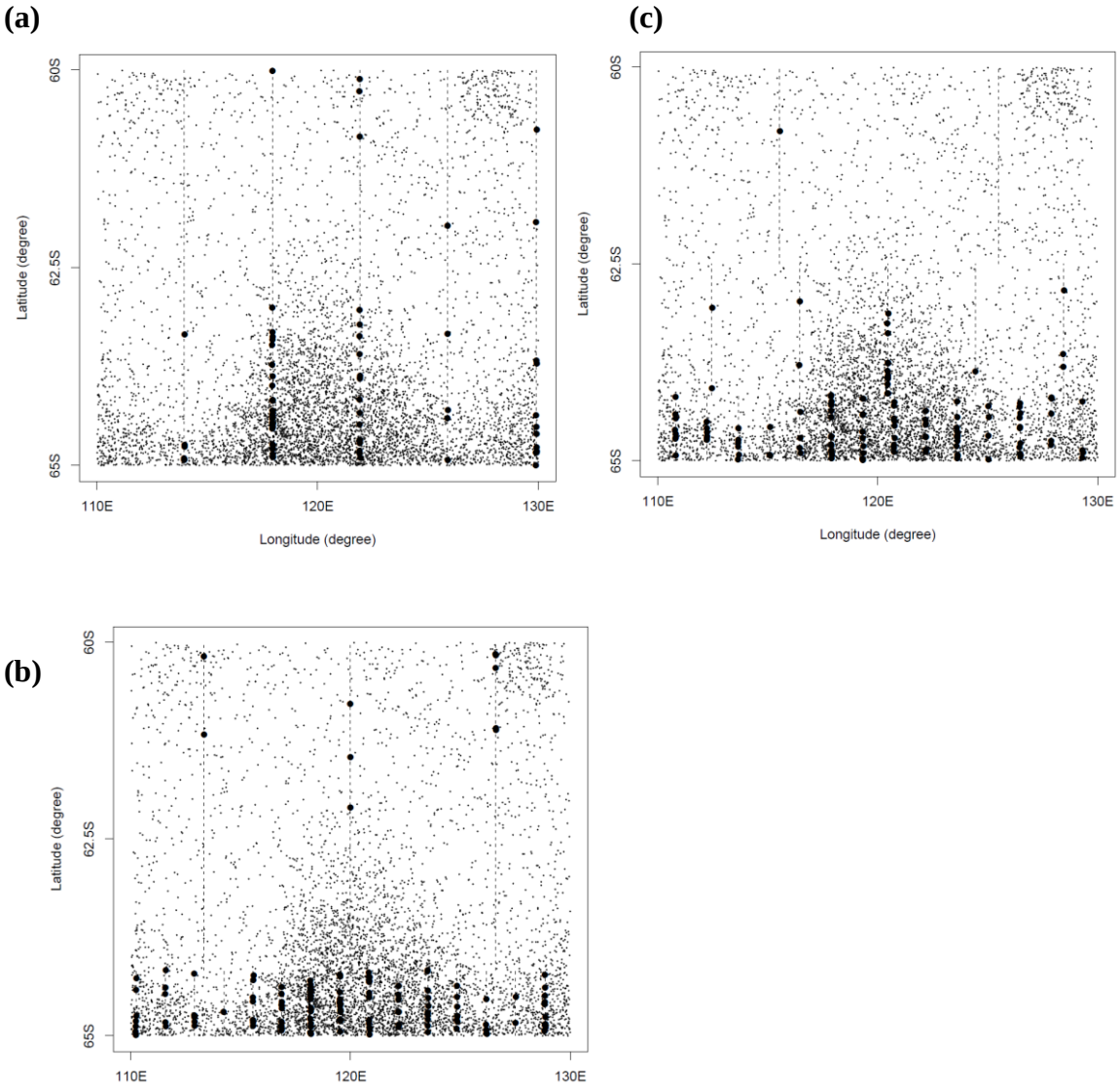


Fig. 3. The different stratification types, (a) single stratum, (b) two strata and (c) three strata. The places of tracklines were set as systematic random sampling scheme. The small dots represent the individual virtual schools. The large dots on the tracklines show simulated detections. In this figure, the true abundance was set to 6000 and the distribution was generated from Eq. 3.

The number of school within each 5 km × 2.8 km × 2 (for right and left side)

rectangular segment *i* is denoted by n_i. Covariates, $MTEM_i$, SST_i and LAT_i were

obtained from the nearest grid cell *v* to the midpoint of segment *i* on the tracklines. Also, the three covariates were obtained at each grid cell *v* ($MTEM_v$, SST_v and LAT_v) that was the centre of each 4 km grid cell in the virtual survey area.

Models and estimation of abundance

Two statistical models were used for SDMs; generalized linear model (GLM) [40] and generalized additive model (GAM) [41]. Both models have been used as SDM frequently [1]. Especially GAM has been commonly used for abundance estimation of cetaceans [14] [22] [20] [42]. A benefit of using GAM is its flexibility in capturing Non-linear cetacean habitat relationships [43]. R version 2.12.2 [44] and package *mgcv* version 1.7-2 [45] were used to fit the GAM.

Suppose that the total length of tracklines was divided into *I* small contiguous segments. Let the length of each segment be l, w is the distance that all schools can be detected (here 2.8 km) and $2lw$ is the area of rectangular segment *i*. I denote the expected number of schools within area of rectangular segment *i*, by $E(n_i)$, which can be modelled with covariates by using GLM or GAM with Poisson distribution. For GLMs,

$$\log(E(n_i)) = \log(2lw) + \theta + \omega_{11} MTEM_i + \omega_{12} SST_i, \quad (10)$$

$$\log(E(n_i)) = \log(2lw) + \theta + \omega_{21} SST_i, \quad (11)$$

$$\log(E(n_i)) = \log(2lw) + \theta + \omega_{31} LAT_i, \quad (12)$$

where intercept (θ) and partial regression coefficient (ω) are parameters to be estimated in each scenario (see the next section) and iteration. I assumed that explanatory variables were not affected by iteration. For GAMs,

$$\log(E(n_i)) = \log(2lw) + \theta + s(MTEM_i) + s(SST_i), \quad (13)$$

$$\log(E(n_i)) = \log(2lw) + \theta + s(SST_i), \quad (14)$$

$$\log(E(n_i)) = \log(2lw) + \theta + s(LAT_i), \quad (15)$$

where the *s* is a 1 dimensional smoothing function with smoothing parameters selected by generalized cross validation. Then, I can estimate the expected number of schools $\hat{N}_v^{(\phi)}$ ($\phi = 10,...,15$) from equations 10-15 because I can obtain the explanatory variables at each grid cell *v* ($MTEM_v$, SST_v and LAT_v). The estimated abundance $\hat{N}$ for SDMs was obtained by summation of $\hat{N}_v^{(\phi)}$.

In the LT estimator, the abundance in stratum *d*, denoted by $\hat{N}_d^{(LT)}$, was estimated by

$$\hat{N}_d^{(LT)} = \frac{\sum_{i=1}^{I_d} n_{i,d}}{2wL_d} A_d, \quad (16)$$

where $n_{i,d}$ was the number of schools within rectangular segment *i* in stratum *d*

($i = 1, \cdots, I_d$), L_d was the total length of the tracklines in each stratum *d* and A_d was the area of stratum *d*. The LT estimate of abundance $\hat{N}$ was obtained by summation of $\hat{N}_d^{(LT)}$.

Scenarios

I prepared 4 factors to make simulation scenarios that were composed of three types of stratification (St), two types of models for generation of virtual school distribution (NL; linear or Non-linear), three sets of covariates for generation (Gen; $MTEM + SST, SST, LAT$) and three sets of covariates for estimation (Est; same as Gen). Here equations (4-9) could show combination of *NL* and *Gen*, and equations (10-15) were distinguished by *Est*. Here the LT estimator did not require covariates Est to estimate the abundance $\hat{N}$. Thus, the number of scenarios was 54 for SDMs and 18 for LT. Here the two cases were examined as for SDMs. One was that the true covariates are used for estimation but the form of the relationship (equations (4-9)) was mis-specified. The other was mis-specification of covariates. As I described above, I iterated the estimation procedure 120 times for each scenario by adding random values in equation 2. No interaction terms between variables were included and no model

selection was carried out. The names of variables are shown in Table 2.

Table 2. Variable names used to generate the scenarios.

Variable	N. of Levels	Explanations
St	3	Single stratum, two strata, three strata with equal total effort among those three types of stratification
NL	2	Linear or Non-linear formulae in the model for generating schools
Gen	3	3 types of variables for generation
Est	3	3 types of variables for prediction

Evaluation of uncertainties

Three measures of model performance were used; relative bias (RB), relative standard deviation (RSD) and relative root mean square error (RRMSE). All three indices were standardized by the true abundance and calculated as ratio by multiplying 100. These definitions are given by the following formulae;

$$RB\left(\hat{N}\right)=100\,E\left(\hat{N}-N_{true}\right)\Big/N_{true}, \quad (16)$$

$$RSD\left(\hat{N}\right)=100\,SD\left(\hat{N}\right)\Big/N_{true}, \quad (17)$$

$$RRMSE\left(\hat{N}\right)=100\sqrt{E\left(\left(\hat{N}-N_{true}\right)^2\right)}\Big/N_{true}, \quad (18)$$

where

$$SD\left(\hat{N}\right)=\sqrt{\left(1\Big/120\right)\sum_{iter}^{120}\left(\hat{N}-E\left(\hat{N}\right)\right)^2}, \quad (19)$$

N_{true} was the true abundance that was either 2000 or 6000. Outliers that were more than 4 times or less than 1/4 of the true abundance from the results of the GAM were removed from the result.

Results

I show the estimated abundance by SDM with mis-specified covariates can have bias under stratified survey designs as shown below. The median value of RB for the SDMs *Gen*Est* (asterisk "*" shows summarizing calculated all the RB, RSD and RMSE as a function of all combinations of the two factors before and after the asterisk each SDM and LT estimator, respectively) was more than 10% when the true covariates were *MTEM + SST* and *LAT* with mis-specification of covariates (Fig. 4a). On the other hand, the median value of RB of LT estimator was within $\pm$5% regardless of types of **(a)**

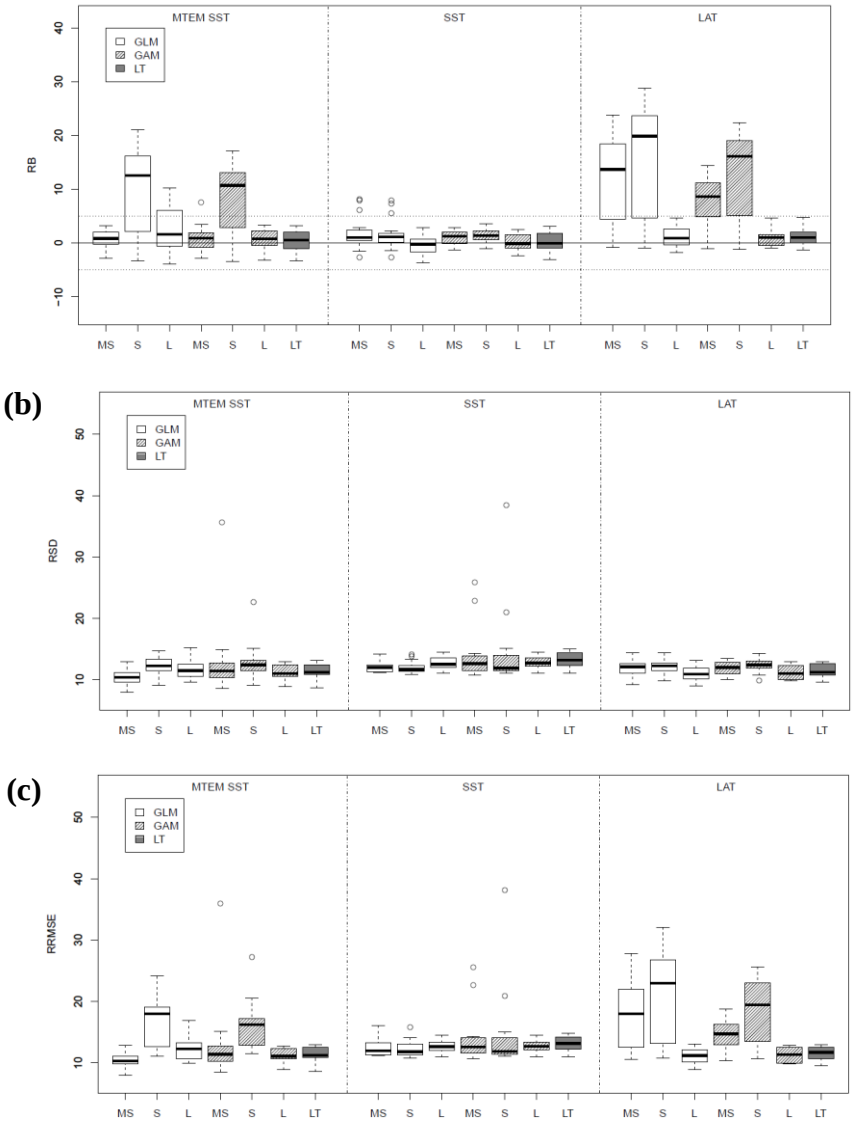


Fig. 4. (a) RB, (b) RSD and (c) RRMSE as function of types of covariates for generation (equations 4-9) and estimation (equations 10-15). Here the abundance was set to 6000. White, hatched and black denote GLM, GAM and LT respectively. The text in each panel shows that *MTEM SST*, *SST* and *LAT* indicates covariates for generation in equations 4-9 (*Gen*) The horizontal axis shows the type of covariates for estimation in equations 10-15. MS, S and L indicate (*MTEM+SST*, *SST* and *LAT* respectively). LT shows the line transect estimator. Horizontal dotted lines in (a) indicate ±5RB lines. Bold line, box limits, error bars and circles indicate the median, 25th-75th percentiles, minimum and maximum values within the length of the range times 1.5 and values outside of this range, respectively.

true covariates. The medians of RSD were not large difference regardless of mis-specification but estimated abundance from GAM had made some extremely large RSD (Fig 4b). RRMSE (Fig. 4c) was consistent with RB (Fig. 4a) rather than RSD (Fig. 4b). Therefore the variation of RRMSE would be caused by that of RB.

I examined the effect of stratification on the values of RRMSE for the SDMs with mis-specification of covariates. I prepared summarized RB with respect to *St* with other factors to examine the effect of stratification on RB (Fig. 5) and RRMSE (Fig. 6) with mis-specification of covariates. Regardless of the types of covariates used for estimation, most of the RBs were within ±5% in single stratum scenarios (Fig. 5a). On the other hand, most of RBs were affected to a greater extent for two and three strata scenarios (Fig. 5a). Also, regardless of the covariates for generation (Fig. 5b) and model linearity (Fig. 5c), most of the RBs were included within ±5% for single stratum scenarios. These results suggest that estimated abundance by SDM with mis-specified covariates can have bias under stratified survey. The RRMSEs for the SDMs with mis-specification of covariates in stratified scenarios were larger than the single stratum same as RB (Fig. 6). Compared to SDMs, RBs of the LT estimator were almost always within ±5% and had a small RRMSE (Figs. 4-6). The RRMSE for *Gen*Est, St*Est*, *St*Gen*, *St*TL* and *St*NL* when N_{true}=2000 (Fig. 7) were similar to N_{true}=6000.

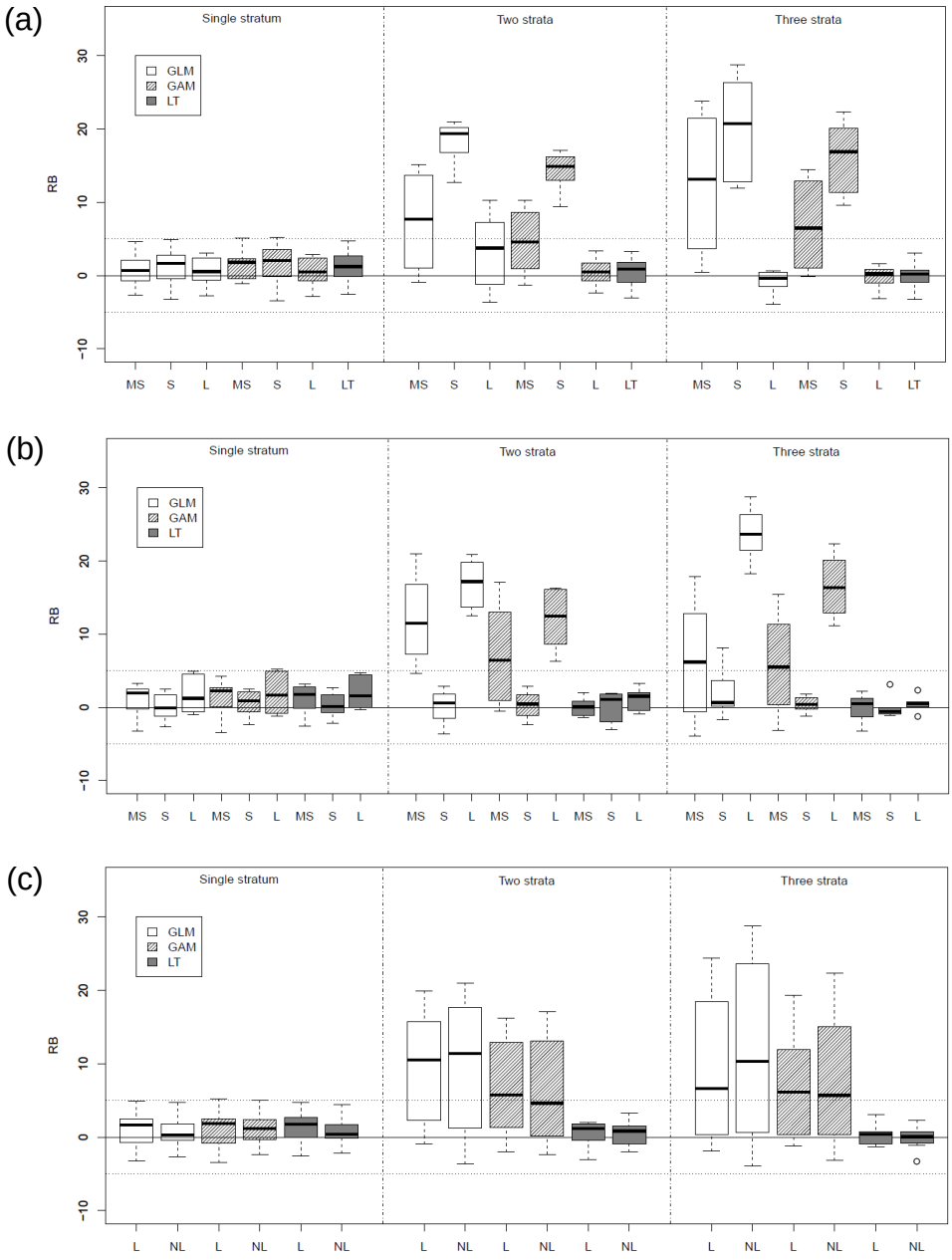


Fig. 5. Box plots of the calculated RB using (a) *St*Est* (b) *St*Gen* and (c) *St*NL* across other factors with mis-specified covariates. Here the abundance was set to 6000. White, hatched and black denote GLM, GAM and LT respectively. (a) The horizontal axis shows the type of covariates for estimation in equations (10-15). MS, S and L indicate *MTEM+SST*, *SST* and *LAT* respectively. (b) The horizontal axis shows the type of covariates for generation in equations (4-9). MS, S and L indicate *MTEM+SST*, *SST* and *LAT* respectively. (c) The horizontal axis shows the type of generation formulae, either linear (L) or Non-linear (NL).

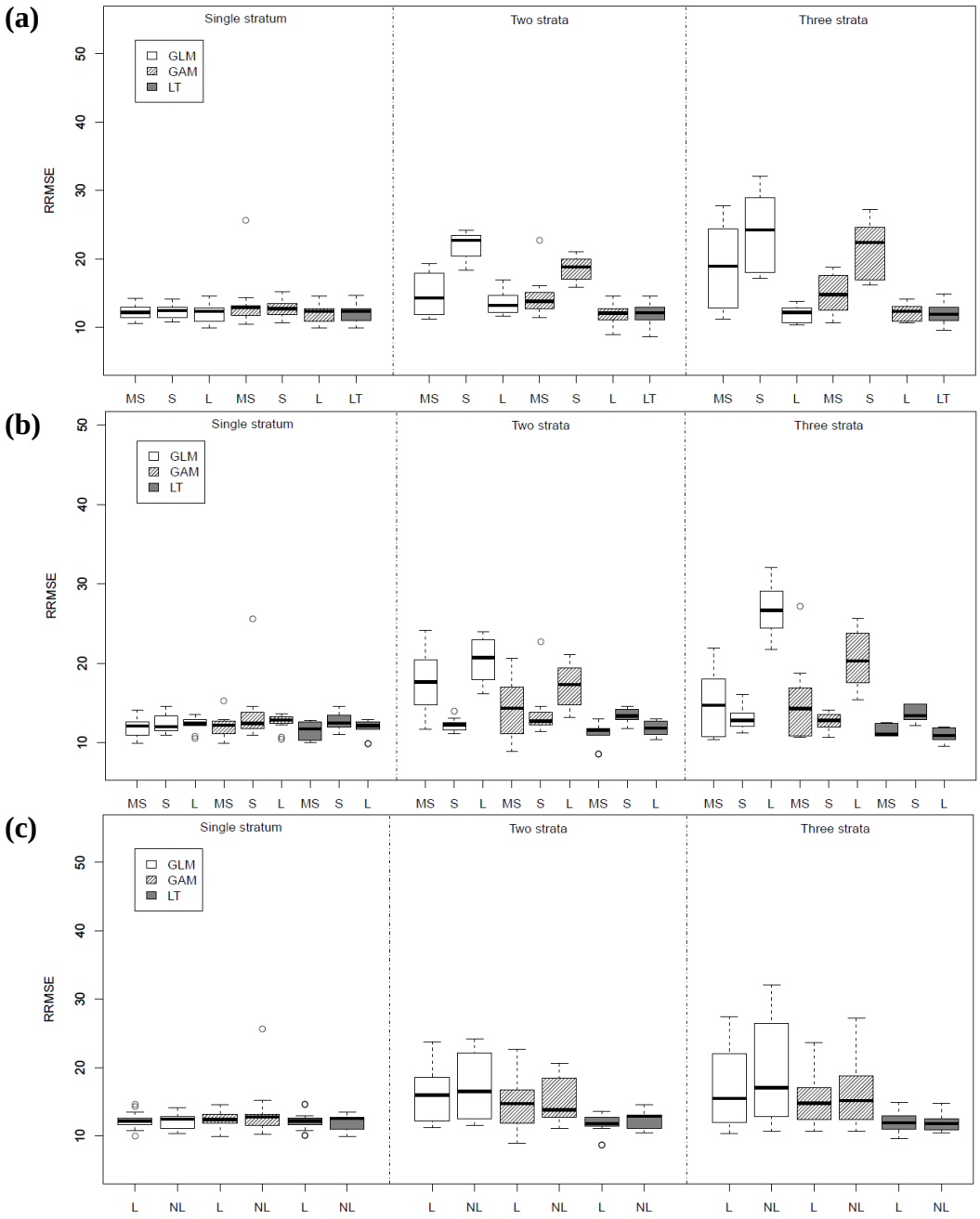


Fig. 6. Box plots of the calculated RRMSE using (a) *St*Est* (b) *St*Gen* and (c) *St*NL* across other factors with mis-specified covariates. Here the abundance was set to 6000. (a) The horizontal axis shows the type of covariates for estimation. MS, S and L indicate *MTEM+SST*, *SST* and *LAT* respectively. (b) The horizontal axis shows the type of covariates for generation S and L indicate *MTEM+SST*, *SST* and *LAT* respectively. (c) The horizontal axis shows the type of generation formulae, either linear (L) or Non-linear (NL).

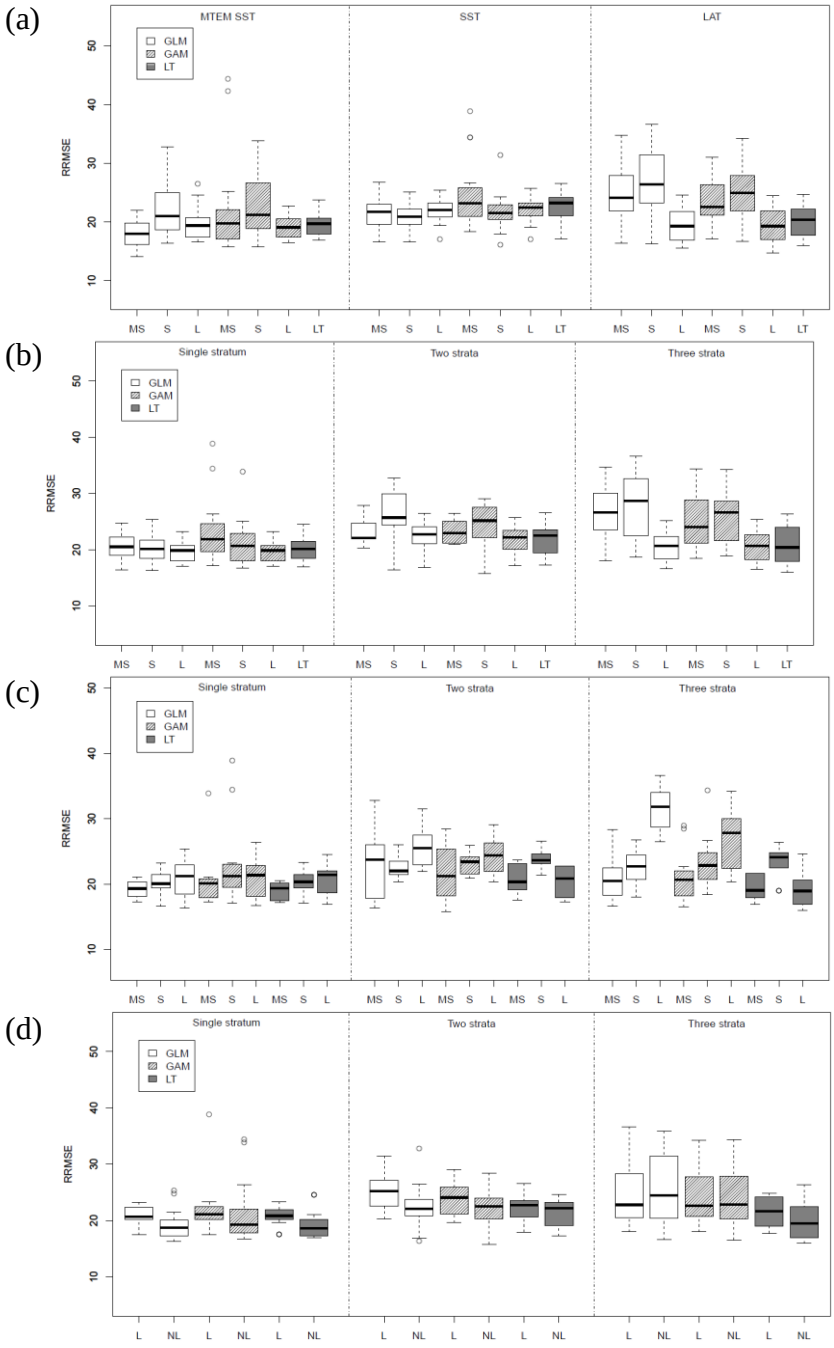


Fig. 7 Box plots of the calculated RRMSE using (a) *Gen*Est*, (b) *St*Est*, (c) *St*Gen* and (d) *St*NL* across other factors with mis-specified covariates. Here the abundance was set to 2000.

When $N_{true} = 2000$, the mean number of detected schools was 24.9 (±5.0), 40.4 (±9.0) and 41.0 (±8.3) in single stratum, two strata and three strata, respectively. The

numbers in parentheses are standard deviations. When $N_{true} = 6000$, there were 75.9 (±9.0), 122.6 (±20.8) and 123.9 (±19.6), respectively. The single stratum scenario always had the lowest mean detection numbers because stratification allocated more effort to the southern portion of the study area where whale schools concentrated. Therefore it suggested that bias in estimated abundance under stratified sampling was not caused by sample size because the sample size in single stratum survey was the smallest.

Discussion

Usually, the classification of samples into groups so that samples within groups have same feature (e.g. biological or economic) before a survey is called as stratification. However, in the field survey, it is difficult to stratify the samples before survey. Therefore geographical stratification is usually used in the field survey because animals at the same geographical region will be thought as having same feature among the animals.

I have shown RRMSE was large because the bias was large, when mis-specified covariates with stratified surveys were used for estimation. This indicates that the true covariates for SDM will be needed to obtain unbiased estimation under stratified survey or non-random sampling. On the other hand, to obtain any true

covariates will be difficult for researchers *a priori*, because no one knows what the truth is. Carrying out the single stratum survey may be more feasible for abundance estimation rather than expending effort to estimate the truth. In this result, the estimated abundance by SDM with LAT did not have large bias. However, there is the situation that had a bias in estimates (Appendix 1). Therefore LAT is not always good covariates for abundance estimation. On the other hand, I did not find conditions when abundance estimation by SDM under multiple strata survey is small bias. Therefore it should be found theoretically in future works.

I showed SDMs were vulnerable to stratifications when mis-specified model was used (Figs. 4-6). Some previous studies reported that GAMs had a smaller coefficient of variation (CV) than the LT estimator [1, 7, 9]. The smaller CV will indicate better estimation if the estimation was unbiased. However, I showed that in this case the abundance estimates by SDMs were biased (Fig. 5a). Smaller CV with biased estimation means a consistent bias in estimation that is far from the true value. It would be dangerous for reliable abundance estimation to make CV lower based on mis-specification of covariates for SDMs especially in case of stratification. Using a dummy data is desirable when I assess between two estimating methods because I can evaluate not only CV but also bias. Unlike the previous studies, I could not conclude

that SDM is better than LT estimator for abundance estimation. If I used smaller sample, I do not know which method is better. In this study, however, the result was same even $N_{true} = 2000$. It is needed to prove the situation that SDM would be the better than LT estimator. On the other hand, if we used the horizontal tracklines against latitude not the vertical tracklines, the result might change. Because the value of $\lambda_v^{(e)}$ was large at south not north (Fig. 3) therefore the detected number of schools on the horizontal tracklines would have large variance among tracklines than that of vertical tracklines. It might effect on the calculated indices. Although we usually set the vertical tracklines against the gradients of spatial distribution of schools for keeping small variance among tracklines [16], it is important when using a data from merchant vessels which tracklines are not statistically designed.

From the 2005/06 austral summer season, the Second Phase of the Japanese Whale Research Program under the Special Permit in the Antarctic (JARPA II) was started. Same as the JARPA survey, JARPA II was keeping stratified random sampling scheme. Under the stratified survey design, the estimated abundance by using SDM may have bias and the LT estimator will be better to estimate abundance unless the mis-specification of covariates was occurred in SDM. On the other hand, if the Antarctic minke whale mainly belonged to the distribution of SST, the estimated

abundance by SDM with the covariates might be useful (see Fig. 4a) because not only abundance but also and an additional spatial distribution of Antarctic minke whale can be gained.

On the other hand, the effect of model selection was not examined because I had focused on the effect of stratification of survey design for SDMs with mis-specification of covariates. Usually, however, a researcher will carry out model selection based on an information criterion such as AIC [46] to improve estimation. Even if the independent covariates that have been sampled did not include the true covariates, the best model will be chosen based on the information criterion. Though such a "best model" may mis-specify covariates, it may have better performance than other mis-specified models. Therefore model selection should be considered to examine the effects of mis-specification of covariates and stratification of survey designs in future works.

Stratification had a great effect on the bias of estimated abundance of SDM if covariates were mis-specified. However this does not imply that stratification always produces undesirable effects. If true model was known, stratification would give a good estimator with small bias and variance. Single stratum surveys may be a good survey design for model-based methods for abundance estimation. Although well-designed

survey stratification is intended to produce the conventional LT estimator with smaller variance than non-stratified surveys, the estimates by SDM with mis-specified covariates may have bias.

IV. Conclusions

Survey design for SDM with mis-specified covariates

I showed that the case when an estimated abundance had bias by SDM with mis-specified covariates can be occurred when survey stratification was carried out. If it assumes that the mis-specification of covariates has usually occurred, survey stratification may be unsuitable for abundance estimation. On the other hand, if the true models assumed in this study (equations 4-9) had approximated the real whale school distribution in the Antarctic, SDM with LAT as covariates would be a good model because bias in estimated abundance would be small. Therefore if the aim of using SDM in the Antarctic was only focused on abundance estimation, SDM with LAT as covariates would be better than other SDMs considered in this study. However, as I showed, SDM with LAT as covariates is not always good (Supplementary 1).

To my knowledge, all preceding articles that use SDMs have not considered possibility of mis-specification of covariates. Actually, we do not know whether mis-specification of covariates occurred or not in each study, and occurring mis-specification of covariates with stratified survey does not directly mean a large bias in estimated abundance which is clearly obtained from this study (e.g. Chapter III Fig. 4). At least, we should pay attention to carry out stratified survey for abundance

estimation with SDM.

Future works

In this study, totally, mis-specified covariates with single stratum and stratified survey had small and large bias in estimated abundance, respectively. On the other hand, it is not clear a theoretical reason why stratified survey arise bias in estimates. In the future works, under the situation where mis-specification of covariates is occurred, development of the optimal survey design for SDM is desired.

V. Acknowledgements

I gratefully acknowledge a large number of people who collected the cetacean sighting and oceanographic data. I would like to thank the member of Centre for Research into Ecological and Environmental Modelling and School of Mathematics and Statistics in University of St. Andrews, our lab members and Centre for Ocean Studies and Integrated Education of Yokohama National University and anonymous reviewers. This work was supported by Grant-in-Aid for JSPS Fellows 201103934 to Y.S. and JSPS grant (10002451) and the Global COE (E-03) by MEXT to H.M. Special thanks to James Lawrence for English proof-reading.

VI. Appendix A case where bias in estimated abundance arise

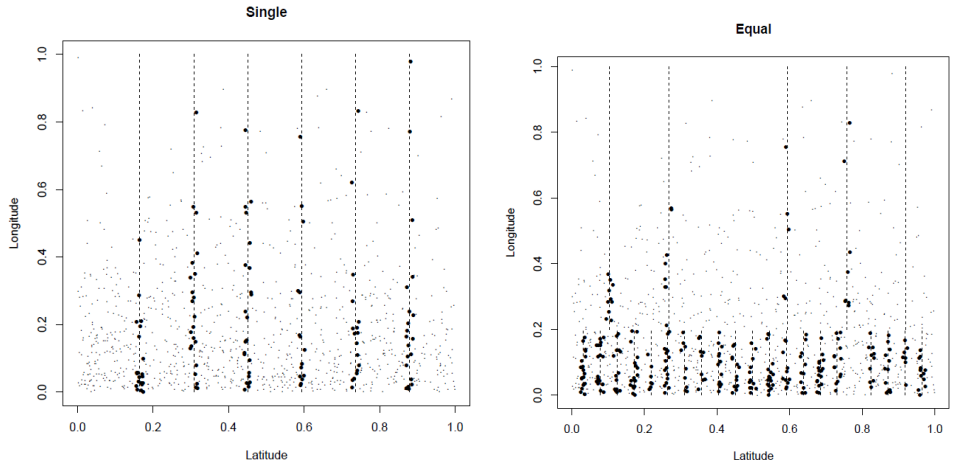
I concluded the estimated abundance by SDM with LAT as covariates were small bias because the models with LAT could have similar estimated abundance $\hat{N}_{lat} = \sum_{Long} \hat{N}_v$ to true $N_{lat} = \sum_{Long} N_v$. However the LAT will not always good covariate corresponding to the true school distribution and survey stratification. I show the numerical experiment that the estimated abundance from SDM with covariates as LAT has bias.

Define survey designs

The study area was fixed as a [0,1]×[0,1] rectangle. Effort allocation was modelled using four different survey designs: single stratum (*Single*), two strata with equal effort allocation between north and south strata (*Equal*), the case south strata has more effort than north (*M_South*) and the case north strata has more effort than south (*M_North*) as shown in Fig.1. Here, the boundaries in the case of two strata were divided by 0.2. I set a total effort of every survey design fixed as 8. The design *Equal* has the total effort allocation of south area was set as 4 and 4 in north area. The design *M_South* has the total effort allocation of south area was set as 5.6 and 2.4 in the north area. The design *S_South* has the total effort allocation of south area was set as 1.6 and 6.4 in the north area. Parallel sets of vertical tracklines towards longitude with equally-spaced intervals were prepared for each stratum and each set of tracklines. The place of tracklines was

randomly changed (systematic random sampling) each iteration where its total effort and intervals were fixed. I omitted the process of estimating a detection function to focus on the effects of survey designs and mis-specification of covariates. Therefore all animals within 0.01 perpendicular distances from tracklines on both sides were counted. The tracklines are divided into 0.05 segments. The numbers of schools within each 0.05 $\times 0.01 \times 2$ (for right and left side) rectangular segment i is denoted by n_i ,where coordinates of point i are midpoint of each segment. Examples of realization of set tracklines, generated and detected animals each survey design was shown in Fig. **A**1.

(a) (b)



(c) (d)

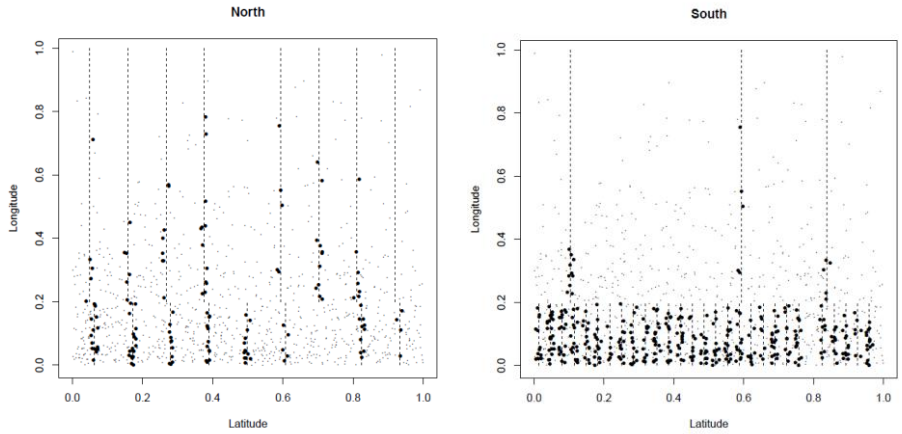


Fig. A.1 The types of survey stratification. Each figure shows (a) *Single*, (b) *Equal*, (c) *North* and (d) *South*, respectively. Break lines show tracklines which places were set as systematic random sampling scheme. The small points represent the individual virtual schools. The large points on the tracklines show simulated detections. In this figure, the true abundance was set to 1000 and the distribution was generated from equation (20).

Define the true models

I prepared the two true model described as;

$$f(LAT,LONG) = \exp(-0.5LAT - 5LONG - 0.1LAT \times LONG), \quad (20)$$

$$f(LAT,LONG) = \left(\frac{1}{\sqrt{2\pi}0.2}\right)^2 \exp\left(\frac{(LAT-0.5)^2}{0.2^2}\right)\exp\left(\frac{(LONG-0.5)^2}{0.2^2}\right), \quad (21)$$

where the value of *LAT* and *LONG* was set between 0 to 1.The probability of occurrence of an individual school at any *LAT* and *LONG*, denoted by $P_{LAT,LONG}$, is given as:

$$P_{LAT,LONG} = \frac{\exp[f(LAT,LONG)(1+0.1\varepsilon_{LAT,LONG})]}{\max(\exp[f(LAT,LONG)(1+0.1\varepsilon_{LAT,LONG})])}, \quad (22)$$

where $\varepsilon_{LAT,LONG}$ represents an independent standardized normal variable whose mean and S.D. are 0 and 1, respectively. The value of $P_{LAT,LONG}$ is divided by max ($P_{LAT,LONG}$) to ensure its maximum value reaches at 1 because $P_{LAT,LONG}$ represents a probability. The virtual distributions of individual animals were generated by using the rejection sampling algorithm: If a random number which is drawn from the uniform distribution $\mathrm{U}(0,1)$, was smaller than $P_{LAT,LONG}$ in a randomly selected coordinates *c* in the virtual space, a school occurred there. This procedure has been continued until generated school had been reached at true abundance. Here I set the true number of

animals as 1000 within the study area.

Models for estimation of abundance

I used a Generalized Linear Model (GLM). Suppose that the total length of tracklines is divided into *I* small contiguous segments. Let the length of each segment be κ. I denote the expected observed number of individual animals within segment *i*, by $E(n_i)$, which can be modelled with covariates by using GLM with Poisson distribution.

$$\ln(E(n_i)) = \ln(2\kappa w) + \alpha + \beta LAT_i, \qquad (23)$$

where the offset variable $2\kappa w$ is the area of segment *i*, α and β are the parameters for estimation and *w* is the effective strip half-width (here 0.01). The estimated abundance $\hat{N}$ for SDMs is obtained by integration of $\hat{N}_{LAT,LONG}$ in the whole study area.

Evaluation of uncertainties

I used standardized relative difference between true and estimated abundance (SRD) which was used to measure uncertainty. The definition is given by the following

formulae;

$$SRD\left(\hat{N}\right)=100\left(\hat{N}-N_{true}\right)\Big/N_{true}, \qquad (24)$$

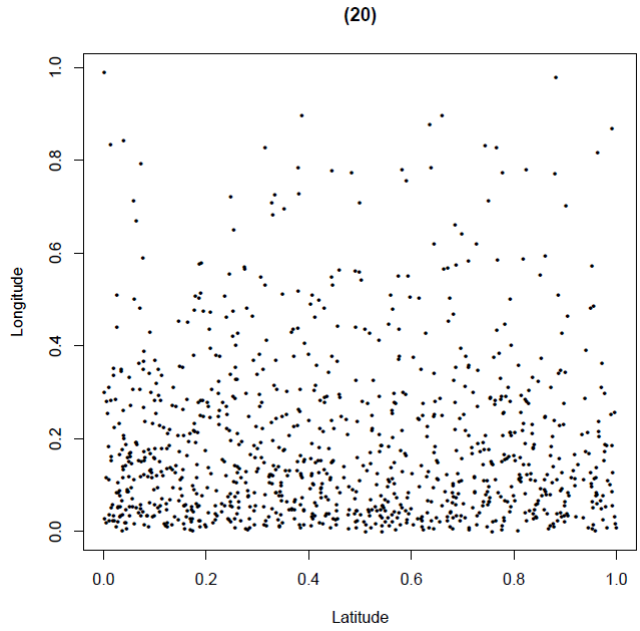
where N_{true} is the true abundance (here 1000). By changing $\varepsilon_{x,z}$, I simulated 1000 times estimation for obtaining the index.

Result

The mean numbers of detected animals in *Single* corresponding to the true models (Eq. 20 - 21) were 160 (12) and 162 (18). In the same way, those were 288 (14) and 121 (17) for the *Equal*, 373 (15) and 92 (16) for the *South*, and 160 (12) and 163 (18) for the *North*. The values in parenthesis are standard deviation.

Although means of estimated abundance based on *Equal* and *South* had biased in case of the true model was equation 21 (-14% and -19%, respectively), as for all the other cases had not bias and mean of estimated abundance were included in ±5% (Fig. **A1.** 2).

(a)



(b)

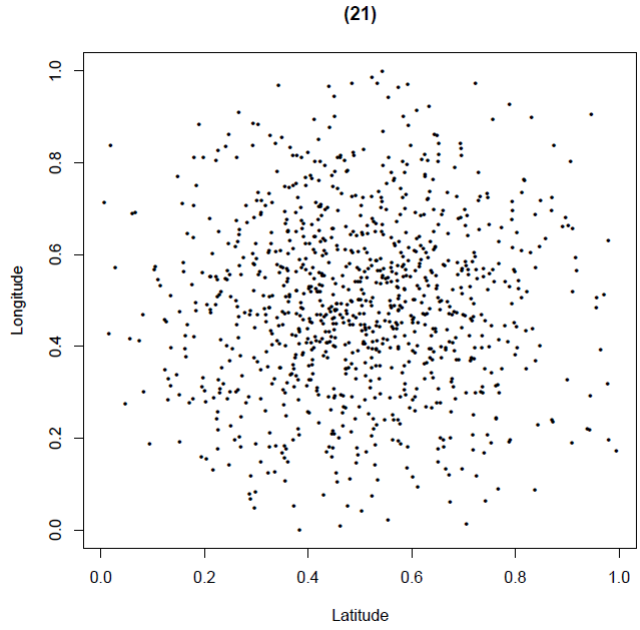


Fig. A.2 The distribution of virtual school generated from (a) equation 20 and (b) 21, respectively. The small points represent the individual virtual schools.

In order to see the approximation situation by LAT in each survey designs, a mean of estimated abundance and the true at the time of 1000 simulations were computed by each grid by integration of longitude (Fig. **S1.** 3). The break line showed 95% prediction interval. While the situation that estimated abundance had bias (Fig **S1.** 3), there was a density slope which originally cannot be appeared (Fig **S1.** 5, **S1.** 7). On the other hand, mean of estimated abundance included in ±5% when the survey designs were *Single* and *North*, even though it was not able to approximate well (Fig **S1.** 4, **S1.** 6). These results indicate that the bias of an estimated abundance by SDM with LAT would be affected according to true school distribution and survey designs.

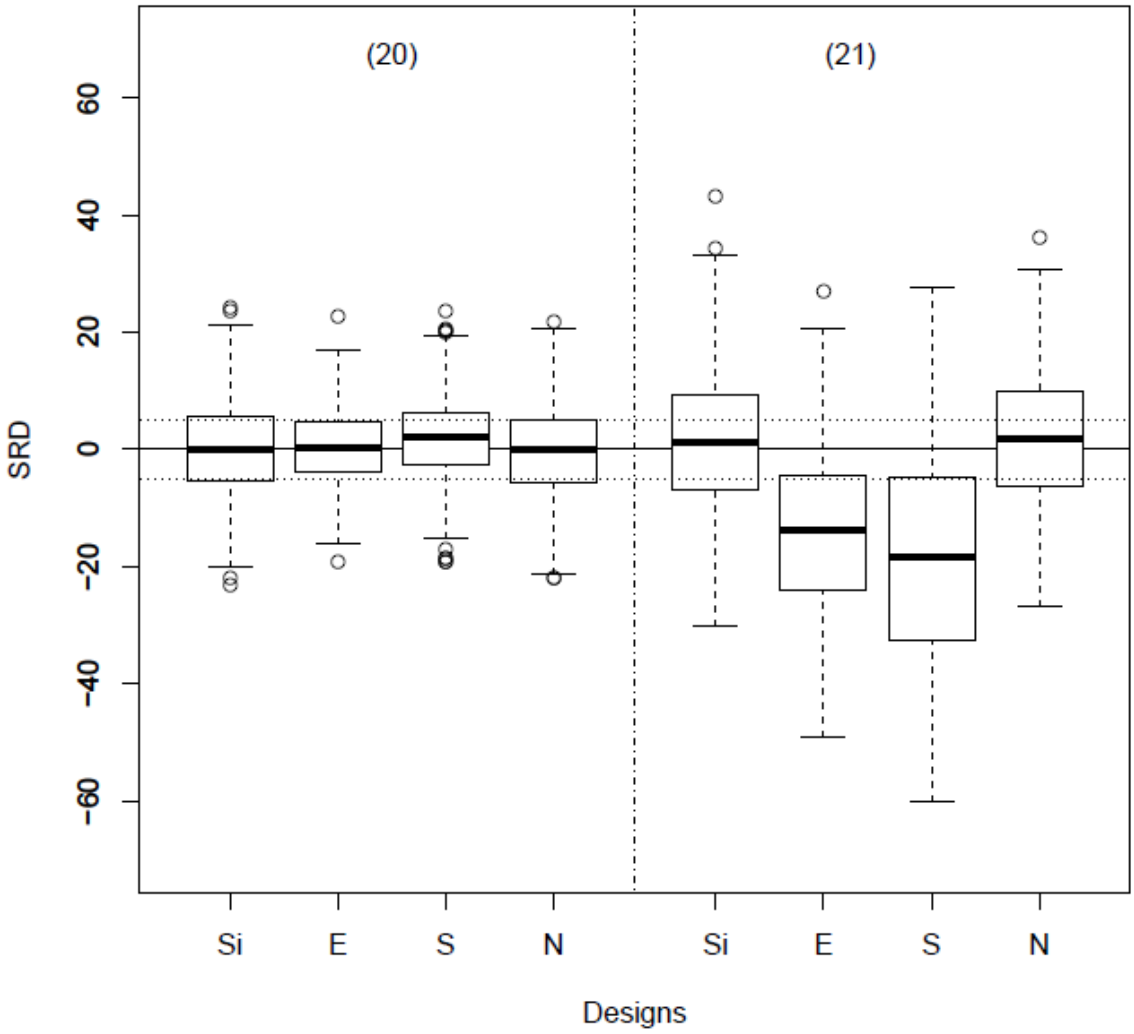
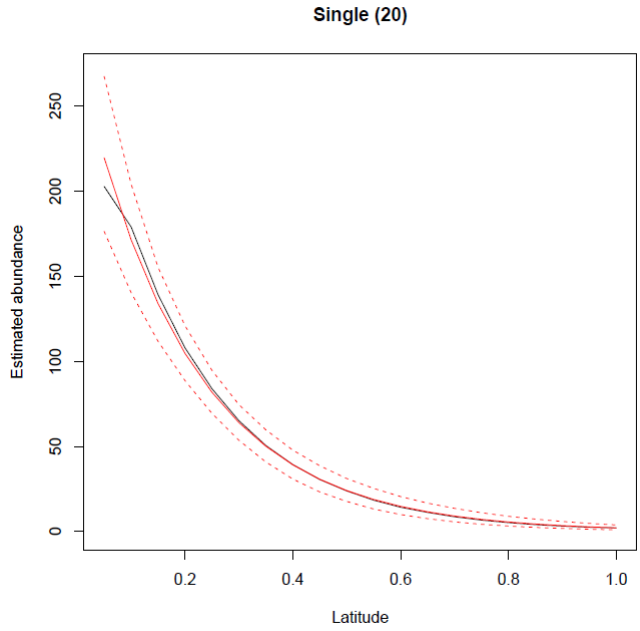


Fig. A.3 Box plots of the calculated SRD. Here the abundance was set to 1000. The horizontal axis shows the type of survey designs; Si, E, S and N indicate Single, Equal, South and North, respectively. Vertical axis shows standardized relative difference between estimated abundance and true abundance. The number in parenthesis at the upper of each four panels shows equations from 20-21.

(a)



(b)

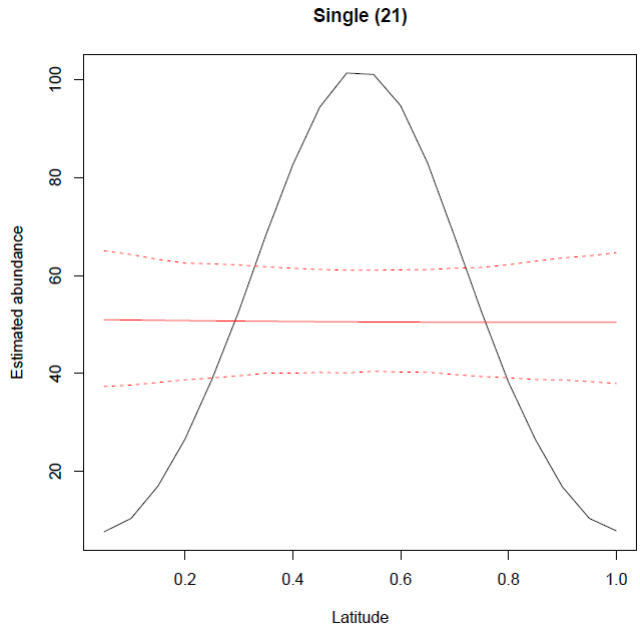
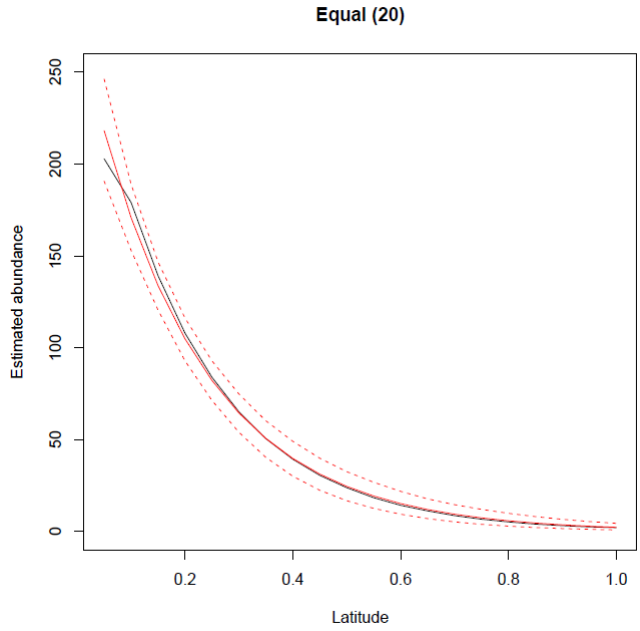


Fig. A. 4 The estimated abundance integrated by longitude. The survey design was fixed as *Single* and true abundance was generated from (a) equation 20 and (b) 21. The black and red line show mean true abundance and estimated abundance integrated by longitude after 1000 iteration. Break red line shows 95% prediction intervals.

(a)



(b)

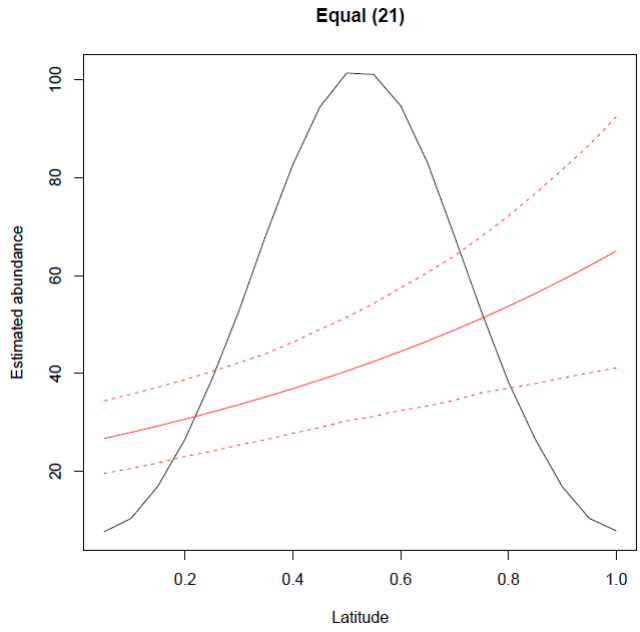
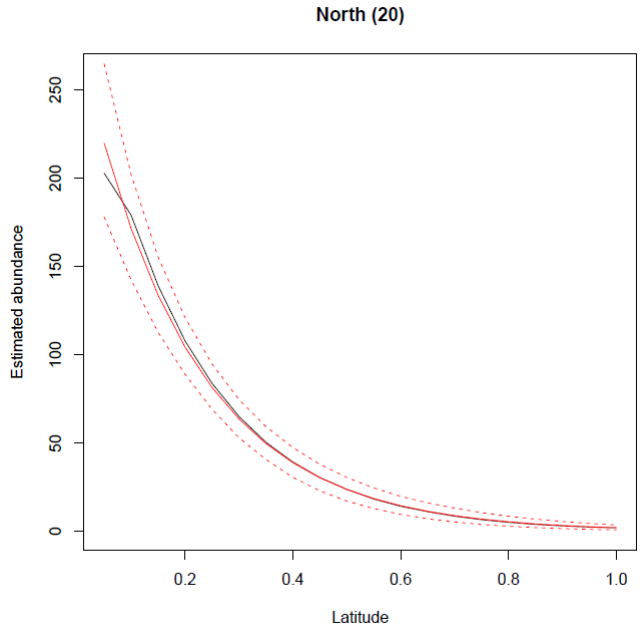


Fig. A. 5 The estimated abundance integrated by longitude. The survey design was fixed as *Equal* and true abundance was generated from (a) equation 20 and (b) 21. The black and red line show mean true abundance and estimated abundance integrated by longitude after 1000 iteration. Break red line shows 95% prediction intervals.

(a)



(b)

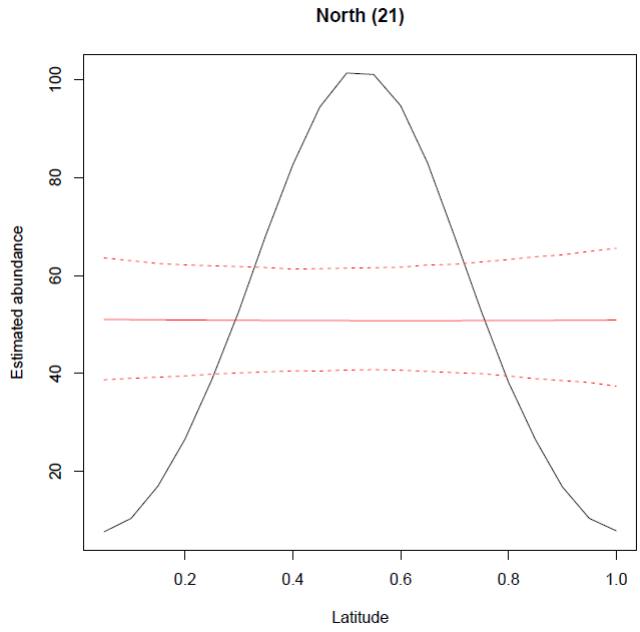
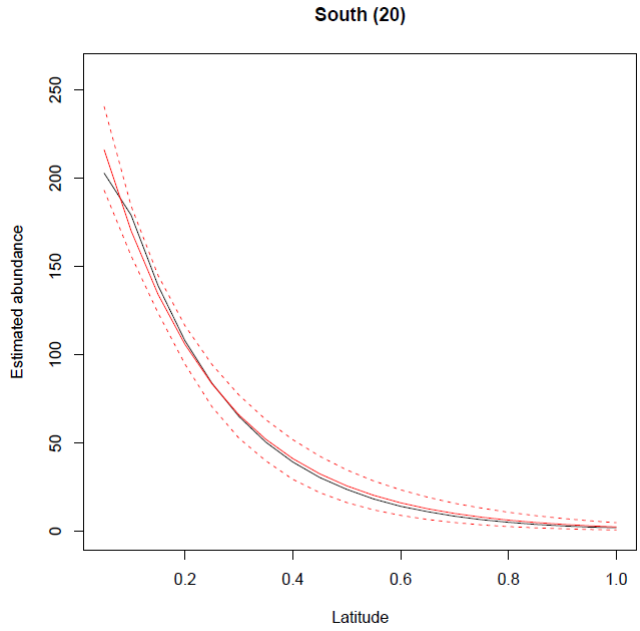


Fig. A. 6 The estimated abundance integrated by longitude. The survey design was fixed as *North* and true abundance was generated from (a) equation 20 and (b) 21. The black and red line show mean true abundance and estimated abundance integrated by longitude after 1000 iteration. Break red line shows 95% prediction intervals.

(a)



(b)

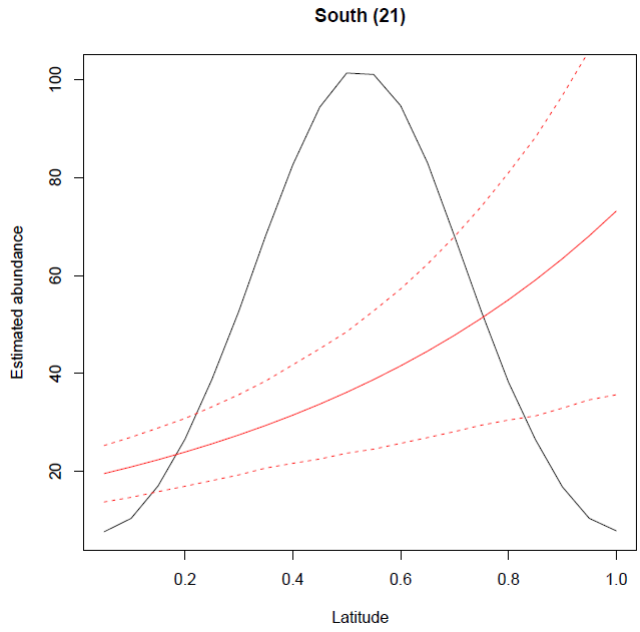


Fig. A. 7 The estimated abundance integrated by longitude. The survey design was fixed as *South* and true abundance was generated from (a) equation 20 and (b) 21. The black and red line show mean true abundance and estimated abundance integrated by longitude after 1000 iteration. Break red line shows 95% prediction intervals.

Reference

1. Guisan, A. and W. Thuiller, *Predicting species distribution: offering more than simple habitat models.* Ecology Letters, 2005. **8**(9): p. 993-1009.
2. Johnston, T.H., *The relation of climate to the spread of prickly pear.* Transactions of the Royal Society of South Australia, 1924. **48**: p. 269-295.
3. Austin, M., *Role of regression analysis in plant ecology.* Proc. Ecol. Soc. Aust., 1971. **6**: p. 63-75.
4. Nix, H., McMahon, J. & Mackenzie, D., *Potential areas of production and the future of pigeon pea and other grain legumes in Australia.* The potential for pigeon pea in Australia. Proceedings of Pigeon Pea (Cajanus cajan (L.) Millsp.) Field Day eds Wallis, E.S. & Whiteman, P.C.), 1977. University of Queensland, Queensland,Australia**,**: p. 5/1–5/12.
5. Peters, R.H., *A critique for ecology*. 1991: Cambridge University Press.
6. Guisan, A. and N.E. Zimmermann, *Predictive habitat distribution models in ecology.* Ecological Modelling, 2000. **135**(2): p. 147-186.
7. Austin, M.P., *An ecological perspective on biodiversity investigations: examples*

from Australian eucalypt forests. Annals of the Missouri Botanical Garden, 1998: p. 2-17.

8. Austin, M., A. Nicholls, and C. Margules, *Measurement of the realized qualitative niche: environmental niches of five Eucalyptus species.* Ecological Monographs, 1990. **60**(2): p. 161-177.

9. Vetaas, O.R., *Realized and potential climate niches: a comparison of four Rhododendron tree species.* Journal of Biogeography, 2002. **29**(4): p. 545-554.

10. Leathwick, J., D. Whitehead, and M. McLeod, *Predicting changes in the composition of New Zealand's indigenous forests in response to global warming: a modelling approach.* Environmental Software, 1996. **11**(1): p. 81-90.

11. Andelman, S.J. and M.R. Willig, *Alternative configurations of conservation reserves for Paraguayan bats: considerations of spatial scale.* Conservation Biology, 2002. **16**(5): p. 1352-1363.

12. Graham, C.H., et al., *New developments in museum-based informatics and applications in biodiversity analysis.* Trends in Ecology & Evolution, 2004. **19**(9): p. 497-503.

13. Pearson, R.G., T.P. Dawson, and C. Liu, *Modelling species distributions in Britain: a hierarchical integration of climate and land ‐ cover data.* Ecography,

2004. **27**(3): p. 285-298.

14. De Segura, A.G., et al., *Comparing cetacean abundance estimates derived from spatial models and design-based line transect methods.* Marine Ecology Progress Series, 2007. **329**: p. 289-299.

15. Reeves, R.R., et al., *Dolphins, whales, and porpoises: 2002-2010 conservation action plan for the world's cetaceans*. 2003, Gland, Switzerland: World Conservation Union.

16. Buckland, S., et al., *Introduction to distance sampling estimating abundance of biological populations.* 2001.

17. Marques, F.F.C., et al., *Estimating deer abundance from line transect surveys of dung: sika deer in southern Scotland.* Journal of Applied Ecology, 2001. **38**(2): p. 349-363.

18. Roberts, J.P. and G.D. Schnell, *Comparison of survey methods for wintering grassland birds.* Journal of Field Ornithology, 2006. **77**(1): p. 46-60.

19. Buckland, S., et al., *Line transect methods for plant surveys.* Biometrics, 2007. **63**(4): p. 989-998.

20. Hedley, S.L. and S.T. Buckland, *Spatial models for line transect sampling.* Journal of Agricultural, Biological, and Environmental Statistics, 2004. **9**(2): p.

181-199.

21. Thomas, L., R. Williams, and D. Sandilands, *Designing line transect surveys for complex survey regions.* Journal of Cetacean Research and Management, 2007. **9**(1): p. 1.

22. Hedley, S.L., S.T. Buckland, and D.L. Borchers, *Spatial modelling from line transect data.* Journal of Cetacean Research and Management, 1999. **1**(3): p. 255-264.

23. Hirzel, A. and A. Guisan, *Which is the optimal sampling strategy for habitat suitability modelling.* Ecological Modelling, 2002. **157**(2-3): p. 331-341.

24. Reese, G.C., et al., *Factors affecting species distribution predictions: A simulation modeling experiment.* Ecological Applications, 2005. **15**(2): p. 554-564.

25. Begg, M.D. and S. Lagakos, *Effects of mismodeling on tests of association based on logistic regression models.* The Annals of Statistics, 1992. **20**(4): p. 1929-1952.

26. Brose, U., N.D. Martinez, and R.J. Williams, *Estimating species richness: sensitivity to sample coverage and insensitivity to spatial patterns.* Ecology, 2003. **84**(9): p. 2364-2377.

27. Magnus, J.R., *The traditional pretest estimator.* Theory of Probability and Its Applications, 1999. **44**(2): p. 293-308.

28. Dominici, F., et al., *On the use of generalized additive models in time-series studies of air pollution and health.* American journal of epidemiology, 2002. **156**(3): p. 193-203.

29. Hoenig, J.M., et al., *Models for tagging data that allow for incomplete mixing of newly tagged animals.* Canadian Journal of Fisheries and Aquatic Sciences, 1998. **55**(6): p. 1477-1483.

30. Jones, F.A. and H.C. Muller-Landau, *Measuring long distance seed dispersal in complex natural environments: an evaluation and integration of classical and genetic methods.* Journal of Ecology, 2008. **96**(4): p. 642-652.

31. Nishiwaki, S., H. Ishikawa, and Y. Fujise, *Review of general methodology and survey procedure under the JARPA.* Paper SC/D06/J2, International Whaling Commission Scientific Committee, 2006.

32. Murase, H., et al., *Relationship between the distribution of euphausiids and baleen whales in the Antarctic (35 E-145 W).* Polar Biology, 2002. **25**(2): p. 135-145.

33. Ishikawa, H., et al., *Cruise report of the Japanese Whale Research Program*

under Special Permit in the Antarctic (JARPA) area IV and eastern part of area III in 1999/2000. 2000, Paper SC/52/O20.

34. Hakamada, T., K. Matsuoka, and S. Nishiwaki, *An update of Antarctic minke whales abundance estimate based on JARPA data including comparison to IDCR/SOWER estimates*. 2005, Paper JA/J05.

35. Hosie, G.W., Schultz, M. B., Kitchener, J. A., Cochran, T. G. and Richards, K, *Macrozooplankton community structure off East Antarctica (80-150ºE) during the Austral summer of 1995/1996.* Deep Sea Res. II, 2000. **47**(2437-2463).

36. Naganobu, M.a.H., T., *Environmental factors for geographical distribution of Euphausia superba Dana.* National institute of polar research special issue, 1986. **40**(191-193).

37. Meynard, C.N. and J.F. Quinn, *Predicting species distributions: a critical comparison of the most common statistical models using artificial species.* Journal of Biogeography, 2007. **34**(8): p. 1455-1469.

38. Gilks, W.R. and P. Wild, *Adaptive rejection sampling for Gibbs sampling.* Applied Statistics, 1992: p. 337-348.

39. Matsuoka, K., et al., *Overview of minke whale sightings surveys conducted on IWC/IDCR SOWER Antarctic cruises from 1978/79 to 2000/01*. Journal of

Cetacean Research and Management, 2003. **5**: p. 173-201

40. McCullagh, P. and J.A. Nelder, *Generalized linear models*. 1989: Chapman & Hall/CRC.

41. Hastie, T.J. and R.J. Tibshirani, *Generalized additive models*. 1990: Chapman & Hall/CRC.

42. Cañadas, A. and P. Hammond, *Model-based abundance estimates for bottlenose dolphins off southern Spain: implications for conservation and management.* Journal of Cetacean Research and Management, 2006. **8**(1): p. 13.

43. Redfern, J., et al., *Techniques for cetacean–habitat modeling.* 2006.

44. team, R.d.c., *R: a language and environment for statistical computing.* R foundation for Statistical Computing, 2011.

45. Wood, S.N., *mgcv: GAMs and generalized ridge regression for R.* Future, 2001. **1**: p. 20.

46. Akaike, H., *A new look at the statistical model identification.* Automatic Control, IEEE Transactions on, 1974. **19**(6): p. 716-723.