Received: 23 August 2023 | Revised: 15 May 2024 | Accepted: 24 May 2024

DOI: 10.1002/env.2867

RESEARCH ARTICLE

WILEY

# Categorical data analysis using discretization of continuous variables to investigate associations in marine ecosystems

**Hiroko Kato Solvang**[1] | **Shinpei Imori**[2] | **Martin Biuw**[3] | **Ulf Lindstrøm**[3,4] | **Tore Haug**[3]

[1]Marine Mammals Research Group, Department of Bergen, Institute of Marine Research, Bergen, Norway

[2]Department of Mathematics, Hiroshima University, Higashi-Hiroshima City, Japan

[3]Marine Mammals Research Group, Department of Tromsø, Institute of Marine Research, Tromsø, Norway

[4]Department of Arctic and Marine Biology, UiT The Arctic University of Norway, Tromsø, Norway

**Correspondence**
Hiroko Kato Solvang, Marine Mammals Research Group, Department of Bergen, Institute of Marine Research, PO Box 1870 Nordnes, N-5817 Bergen, Norway.
Email: hiroko.solvang@hi.no

**Abstract**

Understanding and predicting interactions between predators and prey and their environment are fundamental for understanding food web structure, dynamics, and ecosystem function in both terrestrial and marine ecosystems. Thus, estimating the conditional associations between species and their environments is important for exploring connections or cooperative links in the ecosystem, which in turn can help to clarify such directional relationships. For this purpose, a relevant and practical statistical method is required to link presence/absence observations with biomass, abundance, and physical quantities obtained as continuous real values. These data are sometimes sparse in oceanic space and too short as time series data. To meet this challenge, we provide an approach based on applying categorical data analysis to present/absent observations and real-number data. The real-number data used as explanatory variables for the present/absent response variable are discretized based on the optimal detection of thresholds without any prior biological/ecological information. These discretized data express two different levels, such as large/small or high/low, which give experts a simple interpretation for investigating complicated associations in marine ecosystems. This approach is implemented in the previous statistical method called CATDAP developed by Sakamoto and Akaike in 1979. Our proposed approach consists of a two-step procedure for categorical data analysis: (1) finding the appropriate threshold to discretize the real-number data for applying an independent test; and (2) identifying the best conditional probability model to investigate the possible associations among the data based on a statistical information criterion. We perform a simulation study to validate our proposed approach and investigate whether the method's observation includes many zeros (zero-inflated data), which can often occur in practical situations. Furthermore, the approach is applied to two datasets: (1) one collected during an international synoptic krill survey in the Scotia Sea west of the Antarctic Peninsula to investigate associations among krill, fin whale (*Balaenoptera physalus*), surface temperature, depth, slope in depth (flatter or steeper terrain), and temperature gradient (slope in temperature); (2) the other collected by

ecosystem surveys conducted during August–September in 2014–2017 to investigate associations among common minke whales, the predatory fish Atlantic cod, and their main prey groups (zooplankton, 0-group fish) in Arctic Ocean waters to the west and north of Svalbard, Norway. The R code summarizing our proposed numerical procedure is presented in S4S1.

**KEYWORDS**

AIC, climate, fin whale, krill, marine ecosystem assessment, Minke whale

## 1 | INTRODUCTION

Recent climate change and fluctuations in the environmental conditions of oceans have affected the spatial distribution of both zooplankton and fish species known to be the prey of marine mammals. Investigating the associations among biological communities and oceanographic factors is important for exploring connections or cooperative links in the ecosystem, which may in turn further clarify these relationships.

For marine ecosystem assessment, several spatial and temporal data are obtained by oceanographic surveys to investigate the changes occurring due to global climate change or human activity. However, these data can be either too sparsely or too densely distributed, depending on the area in the ocean, for application to a spatial model that assumes the data are distributed uniformly, and as time series data they are much too short as annual observations for application to a conventional series model that considers temporal correlations. For spatially sparse data, several statistical spatial models have been developed (e.g., Smith et al., 2014; Sugasawa et al., 2022; Ver Hoef & Jansen, 2007), and an integrated approach to investigating the trends of short time series data was also proposed (Solvang & Planque, 2020). In addition to such efforts, Solvang et al. (2021) proposed a practical approach to analyzing the data as categorical data for investigating the relationships among common minke whales (*Balaenoptera acutorostrata*) and their prey species using spatial and temporal data. This approach is useful because it is not necessary to consider the statistical properties of spatial sparseness or short time series for the observations in advance.

The original idea of our approach derives from Sakamoto and Akaike (1978), who was the first study to apply a conditional probability model and statistical evaluation by the Akaike Information Criterion (AIC, see Akaike, 1974) to find the best association for the response variable from possible explanatory variables in a contingency table summarizing categorical data. The computational calculation procedure was first developed in Fortran code and called CATDAP (Katsura & Sakamoto, 1980). Now, the code has been edited in R (R Core Team, 2023) and implemented as "*catdap*" in the R package (The Institute of Statistical and Mathematics, 2023). The first computational procedure was available for only categorical data; however, *catdap* in R is able to handle data including continuous variables by applying a histogram model to the continuous variables as explanatory (dependent) variables, and it tries to find the best categorical groups by AIC.

The procedure is automatically conducted without any biological/ecological prior information, and the continuous data are sometimes categorized into many groups. While the outputs may numerically make sense, the interpretation of many categorized groups becomes too difficult for experts, especially those who investigate complicated biological and ecological associations in marine ecosystems. For them, it might be useful if the data were simply categorized as binary to interpret the data as, for example, higher/lower or presence/absence, according to a single threshold. In this study, we propose a useful statistical analysis procedure to find the optimum threshold to simply categorize the continuous data into two groups and present binary data by considering the relationships among the data based on the procedure in CATDAP. This threshold presents the best association between the response variable and the explanatory variable. Based on the proposed procedure, each continuous data item is replaced by two categorical data. Then, we integrate the data as one dataset and finally apply *catdap* to the dataset to find the best causal relationships among several combinations through setting a response variable and explanatory variables for the data.

This article is organized as follows. Section 2 reviews the categorical data analysis based on AIC from Sakamoto and Akaike (1978) and then introduces our proposed method. The simulation study to verify the proposed method is presented in Section 3. Section 4 gives two examples for the real-data analysis, where—one is applied to the two datasets mentioned in the abstract Discussion is provided in Section 5, and conclusion are offered in Section 6.

## 2 | METHOD

### 2.1 | Independent test using AIC evaluation in the approach of Sakamoto and Akaike (1978)

This subsection reviews the categorical data analysis based on AIC provided by Sakamoto and Akaike (1978). We examine the method called CATDAP. For easy understanding, the dataset we take as an example includes four types of data: whale counting number, krill biomass, depth in the sea, and surface temperature of the sea. Let us consider a simple relationship between whale and krill. For the categorical data analysis, a contingency table is set by counting the numbers for two categories, presence and absence, from the data of whale and krill.

Let $W, K \in \{0, 1\}$ be binary random variables. Suppose that we observe $n$ samples $(W_i, K_i)(i = 1, \cdots, n)$ for whale counting number and krill biomass, which are independently and identically distributed with the same distribution as $(W, K)$. For $j, k = 0, 1$, let $n(j, k) = \#\{(W_i, K_i) = (j, k), i = 1, \cdots, n\}$, $n(j, \cdot) = n(j, 0) + n(j, 1)$, and $n(\cdot, k) = n(0, k) + n(1, k)$. Then, we construct a two-way contingency table as below:

| | $W = 1$ | $W = 0$ | Total |
|---|---|---|---|
| $K = 1$ | $n(1, 1)$ | $n(1, 0)$ | $n(1, \cdot)$ |
| $K = 0$ | $n(0, 1)$ | $n(0, 0)$ | $n(0, \cdot)$ |
| Total | $n(\cdot, 1)$ | $n(\cdot, 0)$ | $n$ |

Here, $W = 1$ and $K = 1$ mean presence of whale and presence of krill, and $W = 0$ and $K = 0$ mean absence of whale and absence of krill. Let the cell frequency and joint probability for whale and krill be represented by $n(i_w, i_k)$ $(i_w = 0, 1, \quad i_k = 0, 1)$ and $p(i_w, i_k)$, respectively, where $\sum_{i_w=0}^{1}\sum_{i_k=0}^{1} n(i_w, i_k) = n$ and $\sum_{i_w=0}^{1}\sum_{i_k=0}^{1} p(i_w, i_k) = 1$. Assuming that $n(i_w, i_k)$ has a multinomial distribution with unknown probabilities $p(i_w, i_k)$, the probability is given by

$$P(\{n(i_w i_k)\}|\{p(i_w i_k)\}) = \frac{n!}{\prod_{i_w=0}^{1}\prod_{i_k=0}^{1} n(i_w, i_k)} \prod_{i_w=0}^{1}\prod_{i_k=0}^{1} p(i_w, i_k)^{n(i_w, i_k)}. \tag{1}$$

If the logarithm of the first term $\frac{n!}{\prod_{i_w=0}^{1}\prod_{i_k=0}^{1} n(i_w, i_k)}$ of (1) is replaced by $M$, the log-likelihood of the unknown parameter is given by

$$l(\{p(i_w, i_k)\}) = M + \sum_{i_w=1}^{1}\sum_{i_k=0}^{1} n(i_w, i_k) \log p(i_w, i_k). \tag{2}$$

If it is assumed that whale presence is independent of krill presence, the model is described by

$$\text{Model(i)} : p(i_w, i_k) = \theta(i_w, \cdot)\theta(\cdot, i_k),$$

where $\theta(i_w, \cdot) = \sum_{i_k=0}^{1} p(i_w, i_k)$ and $\theta(\cdot, i_k) = \sum_{i_w=0}^{1} p(i_w, i_k)$. The log-likelihood is represented by

$$l(\{\theta(i_w \cdot), \theta(\cdot i_k)\}) = M + \sum_{i_w=0}^{1}\sum_{i_k=0}^{1} n(i_w, i_k) \log \theta(i_w, \cdot) \ \ \theta(\cdot, i_k), \tag{3}$$

where $\sum_{i_w=0}^{1} \theta(i_w, \cdot) = \sum_{i_k=0}^{1} \theta(\cdot, i_k) = 1$. The maximum likelihood estimators are obtained by maximizing Equation (3) as follows:

$$\widehat{\theta}(i_w, \cdot) = \frac{n(i_w, \cdot)}{n} \text{ and } \widehat{\theta}(\cdot, i_k) = \frac{n(\cdot, i_k)}{n}.$$

On the other hand, if it is assumed that whale presence depends on krill presence, the model is described by

$$\text{Model(d)} : p(i_w, i_k) = \theta(i_w, i_k),$$

where $\sum_{i_w=0}^{1}\sum_{i_k=0}^{1}\theta(i_w, i_k) = 1$. The log-likelihood is given by

$$l(\{\theta(i_w, i_k)\}) = M + \sum_{i_w=0}^{1}\sum_{i_k=0}^{1} n(i_w, i_k)\log\theta(i_w, i_k). \tag{4}$$

The maximum likelihood estimators are given by $\widehat{\theta}(i_w, i_k) = \frac{n(i_w, i_k)}{n}$. The comparison of Model(i) with Model(d) is usually done by Person's Chi-square test of independence (Pearson, 1990). The Chi-square statistics is calculated by the disparity between the expected value and observation and then assessed by a Chi-square distribution with 1 degree of freedom for a null hypothesis supporting Model(i). Sakamoto and Akaike (1978) used a statistical model selection approach from the model candidates Model(i) and Model(d) in CATDAP, without using the Chi-square independence test-based null hypothesis. The discrepancy of a model fitted to a set of observed data by maximum likelihood is evaluated by AIC, given by $-2\times\log$ (maximized likelihood) $+ 2\times$ number of parameters in the model. AICs for Model(i) and Model(d) are given by

$$\mathrm{AIC}_{\mathrm{Model(i)}} = -2\left\{M + \sum_{i_w=0}^{1}\sum_{i_k=0}^{1} n(i_w, i_k)\log\frac{n(i_w, \cdot)n(\cdot, i_k)}{n^2}\right\} + 2\{(2-1)+(2-1)\} \tag{5}$$

and

$$\mathrm{AIC}_{\mathrm{Model(d)}} = -2\left\{M + \sum_{i_w=0}^{1}\sum_{i_k=0}^{1} n(i_w, i_k)\log\frac{n(i_w, i_k)}{n}\right\} + 2(2\times 2-1). \tag{6}$$

To compare the fitting of the two models to the data, $\Delta AIC = \mathrm{AIC}_{\mathrm{Model(i)}} - \mathrm{AIC}_{\mathrm{Model(d)}}$ is calculated. If $\Delta AIC < 0$, Model(i) should be adopted, that is, the presence of whale and krill is independent, otherwise Model(d) should be adopted, which means that the presence of whale is dependent on the presence of krill.

The above example can be expanded to an $m$-way contingency table that consists of a variable to be predicted and the $m-1$ predictors. When considering the $m-1$ explanatory variables for the responsive variable $I_1$ in general, the $2^{m-1}$ models $\left(={}_{m-1}C_{m-1} + \cdots + {}_{m-1}C_0 \text{ where } {}_nC_r = n!/r!(n-r)!\text{for } n \geq r \geq 0\right)$ to compare the combinations of all explanatory variables are given by

$$\begin{aligned}
&\text{Model } (I_1; I_2, \cdots, I_m) : p(i_1|i_2, \cdots, i_m) = \theta(i_1|i_2, \cdots, i_m),\\
&\text{Model } (I_1; I_2, \cdots, I_{m-1}) : p(i_1|i_2, \cdots, i_m) = \theta(i_1|i_2, \cdots, i_{m-1}),\\
&\qquad\qquad\qquad\qquad \cdots\cdots\cdots\\
&\text{Model } (I_1; I_3, \cdots, I_m) : p(i_1|i_2, \cdots, i_m) = \theta(i_1|i_3, \cdots, i_m),\\
&\qquad\vdots\qquad\qquad\qquad\qquad\vdots\\
&\text{Model } (I_1; I_m) : \qquad p(i_1|i_2, \cdots, i_m) = \theta(i_1|i_m),\\
&\qquad\qquad\qquad\qquad \cdots\cdots\cdots\\
&\text{Model } (I_1; I_2) : \qquad p(i_1|i_2, \cdots, i_m) = \theta(i_1|i_2), \text{and}\\
&\text{Model } (I_1; \phi) : \qquad p(i_1|i_2, \cdots, i_m) = \theta(i_1).
\end{aligned} \tag{7}$$

The above models are summarized by

$$\text{Model } (I_1; \mathbf{J}) : \qquad p(i_1|\mathbf{i}) = \theta\left(i_1|\mathbf{j}\right), \tag{8}$$

where $\mathbf{I}$ is defined by the set of the explanatory variables $\{I_2, \cdots, I_m\}$, $\mathbf{J}$ indicates the arbitrary subset, and $\mathbf{i}$ and $\mathbf{j}$ are represented by the realization of $\mathbf{I}$ and $\mathbf{J}$. Therefore, the AIC of Model $(I_1; \mathbf{J})$ is given by

$$\mathrm{AIC}_{\mathrm{Model}\ (I_1;\mathbf{J})} = -2\sum_{i_1 j} n\left(i_1, \mathbf{j}\right)\log\frac{n\cdot n\left(i_1, \mathbf{j}\right)}{n(i_1)n(\mathbf{j})} + 2(c_1 - 1)(c_{\mathbf{J}} - 1), \tag{9}$$

where $c_1$, $n(\mathbf{j})$, and $c_{\mathbf{J}}$ indicate the number of category of $i_1$, the marginal frequency for each combination of explanatory variables, and the number of category for $\mathbf{J}$ (Sakamoto et al., 1986). AIC is used to search for the optimal predictor on which the variable to be predicted has the strongest dependence (Sakamoto & Akaike, 1978). The procedure in CATDAP is carried out to arrange possible combinations for the response variable and calculate AICs for the models (by running "*catdap2*" function with "*additional. output*" by listing all combinations of the considered model in the R package *catdap*). It is useful to search for the optimal combination of variables for the response variable to show a reasonable relationship among variables.

## 2.2 | Categorization using histogram model in CATDAP

The data of whale abundance are recorded in sighting surveys and can be easily divided into presence for observed whales (i.e., data are over zero) and absence for no observation (zero counted data); however, the data of krill abundance are usually represented by a real number. In this case, a method to estimate the threshold for dividing such real numbers into categorical groups is required for categorical data analysis. In CATDAP, a histogram model for the data distribution shape is fitted to the real numbers, and AIC evaluates the best subset obtained by dividing the histogram's bins between the minimum and maximum values of the data. Let the section width of the histogram be indicated by $d$, given by $d = \frac{x_{\max} - x_{\min}}{m-1}$, where $x_{\min}$ and $x_{\max}$ are the minimum and maximum values of the observation and $m$ is the bin used for dividing the histogram. The frequency distribution of the histogram is assumed to follow a multinomial distribution. Let the frequency and the corresponding probability for each section of the histogram be $n(i)$ and $p(i)(i = 1, \cdots, c)$, where $c$ is the number of the section. The probability given a set of frequencies $\{n(i)\}$ is represented by $P(\{n(i)\}|\{p(i)\}) = \frac{n!}{\prod_{i=1}^{c} n(i)!} \prod_{i=1}^{c} p(i)^{n(i)}$.

Therefore, the log-likelihood and AIC for the histogram model fitted to the real numbers are given by $l(\{p(i)\}) = \log \frac{n!}{\prod_{i=1}^{c} n(i)!} + \sum_{i=1}^{c} n(i) \log p(i)$ and $\mathrm{AIC} = -2\left[\log \frac{n!}{\prod_{i=1}^{c} n(i)!} + \sum_{i=1}^{c} n(i) \log p(i)\right] + 2\{(c-1) + n \log d\}$, respectively (Sakamoto et al., 1986). If the explanatory variable in CATDAP is in real numbers, this model fitting is applied to the data for the optimum categorization. Furthermore, CATDAP is applicable for the case of objective variables indicating real numbers.

## 2.3 | Proposed method: Discretization of real-number variable into binary

CATDAP explores the optimum number for categorization of real-number variables, and the categorized number is sometimes identified as more than two by AIC. For example, if temperature is categorized into four variables by minimum AIC, it may be difficult for oceanographic or ecological experts to interpret the meaning of the four categorical groups in a marine ecosystem. Categorization such as high/low temperature groups makes it simpler to interpret the output than using four categorized temperature groups. To achieve this simplified approach, we propose the following procedure to estimate the optimum threshold for two-value discretization of continuous variables. The word "discretization" of real-number variables corresponds to the word "categorization" of continuous variables into two variables.

Note that we observe a continuous random variable $K^c \in \mathbb{R}$ instead of $K$ in subsection 2.1. This means that we observe $n$ samples $\left(W_i, K_i^c\right)(i = 1, \cdots, n)$ that are independently and identically distributed with the same distribution as $(W, K^c)$. When categorizing $K^c$ in a certain way, we can apply the method described in subsection 2.1 to the discretized data. If the discretization were applied to all of the observed data, the distribution of discretized samples would be complicated, that is, the assumption for i.i.d. is not supported and the same sample could not be used for evaluation of the models for conditional dependency or independency using the procedure in subsection 2.1. In consideration of this point, we randomly separate the dataset into two parts, say $G_1 = \left\{\left(W_i K_i^c\right) | 1 \le i \le m\right\}$ and $G_2 = \left\{\left(W_i K_i^c\right) | m+1 \le i \le n\right\}$ with an integer $m$ $(m = n/2)$. Note that $G_1$ and $G_2$ are independent training and testing sets taken at random in a dataset. We consider this the 0-step, and the following steps are: (i) discretize $K^c$ based on $G_1$; (ii) select the best model by using $G_2$ where $K_j^c (j = m+1, \cdots, n)$ is discretized in the manner of step (i); and, (iii) repeat steps (i) and (ii) 1000 times because the selection result depends on data separation. Then, by comparing the averaged criteria for model selection, we select the model with the lower value as the best one. In the following, we explain each step:

Step (i): data discretization.

Suppose that $K^c \in [a, b]$, where $a$ and $b$ are constants satisfying $a < b$. Let $s_1, \cdots, s_L$ be predetermined $L$threshold points between $a$ and $b$, that is, $a < s_1 < \cdots < s_L < b$. Let us fix $l \in \{1, \cdots, L\}$. Define $K_i^c(s) = I\{K_i < s_l\}$ for $i = 1, \cdots, m$, where $I\{\cdot\}$ is an indicator function. Then, $n_{s_l}(1,1)$ is the total number of indices $i$ such that $\left(W_i, K_i^c(s_l)\right) = (1,1)$; $n_{s_l}(1,0)$,

$n_{s_l}(0,1)$ and $n_{s_l}(0,0)$ are defined similarly. Based solely on the elements of $G_1$, the following two-way contingency table is obtained:

| | $W=1$ | $W=0$ |
|---|---|---|
| $K_i^c(s_l)=1$ | $n_{s_l}(1,1)$ | $n_{s_l}(1,0)$ |
| $K_i^c(s_l)=0$ | $n_{s_l}(0,1)$ | $n_{s_l}(0,0)$ |

This table is denoted by $T_l(G_1)$. The values of this table allow us to perform a Pearson's Chi-square independent test. Let $p_l$ be the $p$-value associated with this test statistics and the Chi-square distribution with 1 degree of freedom. This process is applied to each value of $l\,(1,2,\cdots,L)$. Having the set of $L$, $p(G_1)=\min\{p_1,p_2,\cdots,p_l\}$. After determining $\widehat{p}(G_1)$, define $\widehat{s}(G_1)=s_\kappa$ where $\kappa$is such that $\widehat{p}(G_1)=p_\kappa$.

Step (ii): model selection.

Based on $s_\kappa$, consider the counts of the table $T_\kappa(G_2)$. Let $n'_s(j,k)=\#\{(W_i,K_i(s))=(j,k),i=m+1,\cdots,n\}$. Note that given $s=s_\kappa$, the elements of $T_\kappa(G_2)$,$\mathbf{n}'_{s_\kappa}=\left(n'_{s_\kappa}(1,1),n'_{s_\kappa}(1,0),n'_{s_\kappa}(0,1),n'_{s_\kappa}(0,0)\right)$, follows a multinomial probability model with unknown probabilities $p'(i_w,i_k)$ where $i_w\in\{0,1\}$ and $i_k\in\{0,1\}$. Theses probabilities can be modeled under either dependence or independence. This results i model candidates with their corresponding AICs. As usual, the model with the smallest AIC is selected as the best model.

Steps (i) and (ii) are repeated 1000 times, and then the averaged values for the optimum threshold and AIC are calculated as

$$\bar{\mathrm{AIC}}=\frac{\sum_{t=1}^{1000}\mathrm{AIC}^{(t)}}{1000},\quad \bar{s}=\frac{\sum_{t=1}^{1000}\widehat{s}\left(G_1^{(t)}\right)}{1000}$$

where $\bar{\mathrm{AIC}}^{(t)}$ is the AIC of the $t-$th iteration and similarly, $\widehat{s}\left(G_1^{(t)}\right)$ is the optimal threshold from the $t-$th iteration. The averaged AIC is used to evaluate which model is better among the model candidates. The optimum threshold for the best model is used as a final optimum threshold to ensure simple interpretation of the continuous data, for example, large/small or high/low, by biological/ecological experts or relevant stakeholders.

If other continuous data such as temperature or depth should be considered, corresponding to the counted whale number data, the same procedure is applied to two partitions for the continuous data. Here, we observe $n$ samples $\left(W_i,T_i^c\right)$ and $\left(W_i,D_i^c\right)$ $(i=1,\cdots,n)$ for temperature $(T^c)$ and depth $(D^c)$ in addition to krill, which are independently and identically distributed with the same distributions as $(W,T^c)$ and $(W,D^c)$. Similar to the previous example for whale and krill, the two are also randomly separated based on the same integer $m(1<m<n)$ and the same random order as krill like $G_{1,T^c}=\left\{\left(W_iT_i^c\right),i=1,\cdots,m\right\}$ and $G_{2,T^c}=\left\{\left(W_iT_i^c\right),i=m+1,\cdots,n\right\}$ and $G_{1,D^c}=\left\{\left(W_iD_i^c\right),i=1,\cdots,m\right\}$ and $G_{2,D^c}=\left\{\left(W_iD_i^c\right),i=m+1,\cdots,n\right\}$, respectively. The former groups $G_{1,T^c}$and $G_{1,D^c}$ are used to detect the optimum threshold to discretize them into binary form, and the latter groups $G_{2,T^c}$ and $G_{2,D^c}$ are discretized into binary using the threshold, and a new contingency table is made for the four types of data (whale, krill, temperature, and depth). Setting the response variable, an expanded procedure for more explanatory variables than the two variables shown in Section 2.1 is applied to the multi-way contingency table. According to Equation (9), the AICs for all combinations of explanatory variables are given by the conditional probability models. By iterating these procedures 1000 times, the optimum thresholds and AICs can be obtained through averaging their values.

Finally, the best conditional probability model, including the explanatory variables associated with the response variable, is identified. The conceptual outline of the proposed procedure is given by Figure 1. The numerical procedure including the R package *catdap* (2023) is implemented using R code (R Core Team, 2023) summarized in Supplementary S4.

## 3 | SIMULATION STUDY

To validate our proposed method, the following simulation study was conducted. We first generated 1000 data $x$ by the truncated normal distribution given by

$$f(x;\mu,\sigma,a,b)=\frac{1}{\sigma}\frac{\Phi\left(\frac{x-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)},$$

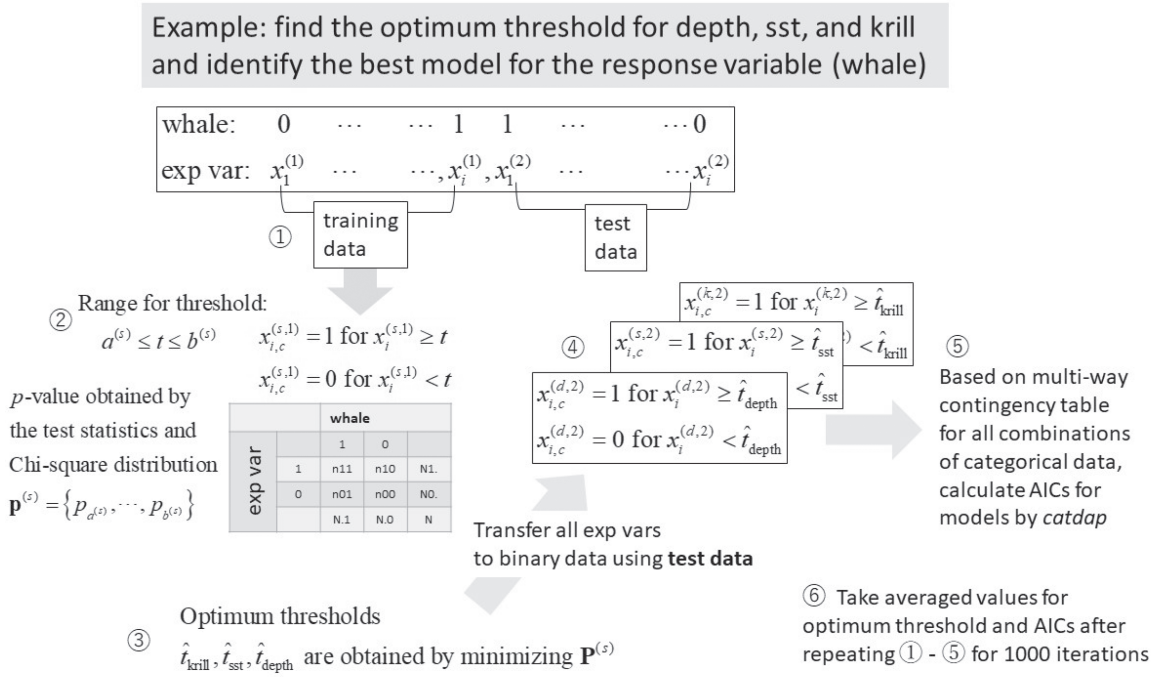


**FIGURE 1** Flow of proposed procedure. ① Set a continuous explanatory variable for a response variable, for example, whale, and split it into training set and test set; ② Categorize the explanatory variable for the training data according to a range for the threshold and make a two-way table using the categorized data and the corresponding whale data. Apply an independent test for the table and calculate a *p*-value of the test statistics for each threshold; ③ Minimize the vector summarizing the obtained *p*-values and obtain the optimum threshold in the case where the *p*-value indicates minimum. Step (i) in the text corresponds to ①, ②, and ③. Apply step (i) to all explanatory variables; ④ Transfer all explanatory variables to binary data for the test data using the optimum thresholds; ⑤ Based on a multi-way contingency table for all combinations of the binary explanatory variables, calculate Akaike Information Criterion (AIC) using *catdap*. Step (ii) in the text corresponds to ④ and ⑤. ⑥ Take the averaged values for the optimum threshold and AIC after repeating steps (i) and (ii) 1000 times. The averaged AIC is used to evaluate which model is better among model candidates. The optimum threshold for the better model is used as a final optimum threshold to divide the continuous values into two-category data, 1 and 0. Step (iii) in the text corresponds to ⑥.

where $\mu$ and $\sigma$ are mean and standard deviation of the distribution for random variable $X$ within $-\infty \le a < b \le \infty$. In this case, the data, called "simb," are mixed by two truncated normal distributions $f(x; 3, 1.75, 1, 10)$ and $f(x; 7, 0.75, 1, 10)$ as Case 1, $f(x; 3, 0.75, 1, 10)$ and $f(x; 7, 1.75, 1, 10)$ as Case 2, and $f(x; 3, 0.75, 1, 10)$ and $f(x; 7, 0.75, 1, 10)$ as Case 3. The histogram for each case presents different two peaks summarized in Figure 2. Then, we generated categorical data, called "csimb," which have the same array size as simb:

$$\text{Case 1: csimb} = \begin{cases} 1 & \text{for } 2.0 \le \text{simb} < 4.0 \text{ or } 6.0 \le \text{simb} < 8.0 \\ 0 & \text{otherwise} \end{cases},$$

$$\text{Case 2: csimb} = \begin{cases} 1 & \text{for } 2.0 \le \text{simb} < 4.0 \text{ or } 6.5 \le \text{simb} < 8.5 \\ 0 & \text{otherwise} \end{cases}, \text{ and}$$

$$\text{Case 3: csimb} = \begin{cases} 1 & \text{for } 2.5 \le \text{simb} < 3.5 \text{ or } 6.5 \le \text{simb} < 7.5 \\ 0 & \text{otherwise} \end{cases}.$$

Furthermore, we prepared another type of data, called "simc," that includes 1000 data $y$ generated by beta distribution, given by

$$f(y; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1},$$

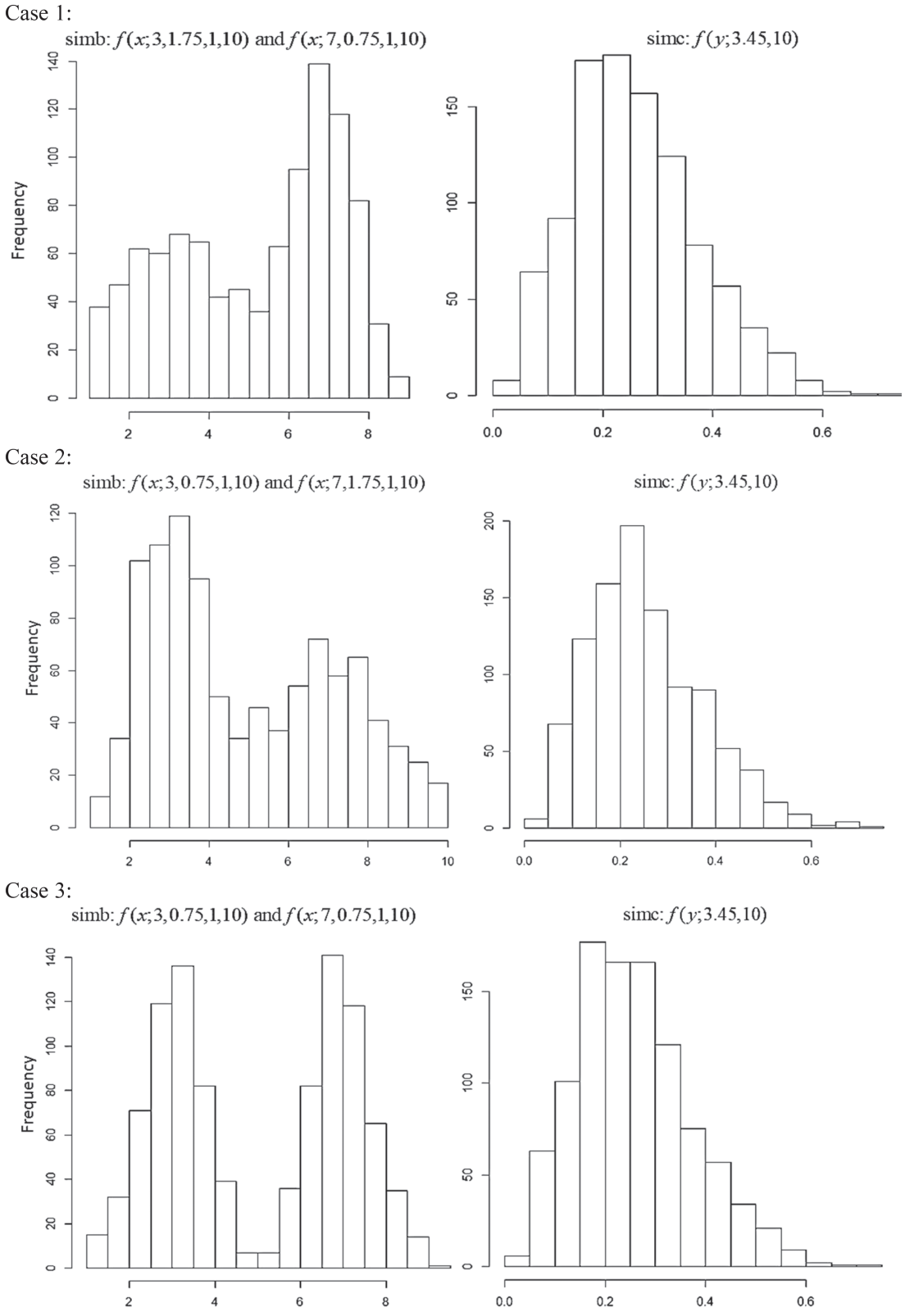


**FIGURE 2** Histograms of simb generated by two truncated normal distributions (left) and simc generated by beta distribution (right).

where Γdenotes a gamma function, and we set $a = 3.45$ and $b = 10$ in this study.

Summarizing these procedures, we now have the following artificial datasets:

simb (Cases 1–3): 1000 real-number data generated by two truncated normal distributions,
csimb: 1000 binary data by categorization of simb based on two thresholds, and.
simc: 1000 real-number data generated by beta distribution.

We illustrated the histograms for simb and simc to consider the ranges to find the optimum thresholds for two variables as seen in Figure 2. We set the range to detect the threshold of simb (Case 1) as 1.5–8.0, simb (Case 2) as 2.0 to 9.0, and simb (Case 3) as 2.0–8.0, and the threshold of simc as 0.1–0.5.

We performed two different simulation studies: Simulation Study 1 to validate the proposed method, and Simulation Study 2 to investigate whether the method's observation includes many zeros (zero-inflated data), which can often occur in practical situations.

## 3.1 | Study 1

The proposed method introduced in Section 2.3 is applied to the two combinations of csimb and simb and csimb and simc using *catdap* (without histogram model). Here, csimb is set as the response variable and simb in Cases 1–3 and simc are set as the explanatory variables. The models are given by

$$\begin{aligned}
&\text{Model 1}: \text{Model}(\text{csimb}; \text{simb}(\text{case}^*), \text{simc})\\
&\text{Model 2}: \text{Model}(\text{csimb}; \text{simb}(\text{case}^*)),\\
&\text{Model 3}: \text{Model}(\text{csimb}; \text{simc}), \text{and}\\
&\text{Null model}: \text{Model}(\text{csimb}; \phi)
\end{aligned}$$

For the outputs, the averaged AIC values for the above three models are summarized in Table 1. The values for Model 2 always become the smallest since the model captures the correct association of simb with csimb. While the averaged AICs of Models 1 and 2 are similar, null and Model 3 having no association with simc are also similar but worse than Models 1 and 2. Supplementary Figure 1 in S1 shows the bar plots for the differences in AICs between Models 2 and 1, and between Models 2 and 3, for each iteration. The upper panel for AIC (Model2) – AIC (Model1) in the three cases mostly indicates negative values, and the lower panel for AIC (Model2) – AIC (Model3) in the three cases indicates all negative values. The table named Simulation Study 1 in Supplementary Table 1 in S3 gives us the counted number in the case of supporting Model 1 as the best model or supporting Model 2 as the best model, which corresponds to summed up iterations showing negative bar plots in Supplementary Figure 1 in S1. It shows that the counted number for Model 2 is larger than the counted number for Model 1, while the counted number for Model 2 in Case 2 indicates a lower level, which corresponds to more positive values in the bar plot for AIC (Model2) – AIC (Model1) in Case 2. Furthermore, the optimum threshold for each iteration in the three cases is shown under bar plots in Supplementary Figure 1 in S1. The averaged optimum thresholds for simb and simc plotted by dashed lines in the lower panels of Supplementary Figure 1 in S1 were:

**TABLE 1** Mean values of Akaike Information Criterion for the models in Simulation Study 1.

| | Case 1 | Case 2 | Case 3 |
|---|---|---|---|
| Model 1 | 507.45 | 605.16 | 661.77 |
| Model 2 | 505.16 | 603.94 | 659.62 |
| Model 3 | 622.58 | 641.97 | 695.37 |
| Null (Independent) | 621.42 | 642.34 | 694.23 |

*Note*: Cases 1, 2, and 3 correspond to the combinations for explanatory variables; simb in Case1 and simc, simb in Case2 and simc, and simb in Case3 and simc. The response variable of the model is csimc. Generated data includes 1000 samples. Model1: Model (csimb; sim(case*), simc), Model2: Model (csimb; simb(case*)), Model3: Model Model (csimb; simc), and Null: Model(csimb; ø), which means an independent model is applied to the data in all three cases.

5.96 and 0.28 for Case 1,

4.00 and 0.20 for Case 2, and

4.73 and 0.26 for Case 3

respectively. These mean values are not related to the threshold of the best model, but they would be useful for expressing a higher/lower level as a simple interpretation of the continuous values. In *catdap*, when AIC summarized in the output indicates zero, the model includes no variable as the explanatory variable (Null model). Therefore, if the AIC of a model indicates a positive value, the explanatory variable assumed in the model is basically not dependent on the response variable and is not necessary as an explanatory variable in the model.

### 3.2 | Study 2

In practical situations, the data obtained by an annual survey are taken from different sample points for each year and often includes the data counted as zero at the data collecting point, for example, zero-inflated fish count data observed in the Barents Sea (Sugasawa et al., 2022). To verify the performance of the proposed method, we considered applying the above four models to csimb, including different numbers of zero data in different sampling data. The prepared data in this study involve 250, 500 and 750 sampling numbers for csimb, simb (Case1), and simc, and then, we intentionally set 30%, 60%, and 90% zero-inflated data in csimb. Table 2 summarizes the averaged AICs according to the above four models. The smallest AIC correctly represents Model 2 for 0%–60% zero-inflated csmib in 250–750 sampling data. On the other hand, the case for 90% zero-inflated csmib in 250–750 sampling data has difficulty supporting the true model. Simulation Study 2 of Supplementary Table 1 in S3 presents outcomes for the counted number of iterations according to the model that was selected as having the smallest AIC in an iteration. In the case of 0 and 30% zero-inflated csimb in 750 and 1000 sampling data, Models 1 and 2 were selected as the best model through the iterations (similar to the outcome in Simulation Study 1). On the other hand, a higher zero-inflated rate and small sampling data induce cause more varied models to be selected as the best model for some iterations. Furthermore, this shows that the smallest AIC with respect to the averaged AIC does not support Model 2 for 90% zero-inflated csimb in 250–750 sampling number data. For 250 sampling number and 90% zero-inflated csmib, the averaged AICs for Model 2 and Null model are not so different, but, the counted number of Null model as seen in the table (250 samples, Simulation Study 2 of Supplementary Table 1 in S3) is considerably larger than the counted number for Model 2.

The above simulation procedure is performed in R (R Core Team, 2023), and simb is generated using the function "*rtruncnorm*" in R (Mersmann et al., 2018) as seen in the Supplementary S4 summarizing R code.

## 4 | REAL-DATA ANALYSIS

### 4.1 | Field observations

We applied our methods to two cases. The first uses visual sightings of fin whales from line transect surveys in the Southern Ocean, while the second case uses visual sightings of minke whales from line transect surveys in the in the Arctic Ocean to the west and north of Svalbard.

#### 4.1.1 | Case 1: Southern Ocean fin whales

Visual observations for cetaceans were carried out onboard three of the six vessels participating in the 2019 Area 48 Survey for Antarctic krill (Krafft et al., 2021): the *R/V Kronprins Haakon* (KPH), the *F/V Cabo de Hornos* (CDH), and the *RRS Discovery* (DIS). Details of the observation methods and protocol can be found in Biuw et al. (2024). Observation transects were split into 1 nm long segments, and the number of fin whale sightings (groups and individuals) were summarized within each segment. All positive sightings (i.e., number of individuals within a segment greater than or equal to 1). We also included surface temperature (°C) and water depth (meter) for a 1-nm segment to investigate the association from

**TABLE 2** Mean values of Akaike Information Criterion for the models in Simulation Study 2.

| **1000 samples** | | | | |
|---|---|---|---|---|
| Applied model | 0% | 30% | 60% | 90% |
| Model 1 | 507.45 | 636.53 | 571.98 | 253.65 |
| Model 2 | 505.16 | 634.34 | 569.20 | 250.63 |
| Model 3 | 622.58 | 695.30 | 591.48 | 253.28 |
| Null model | 621.42 | 694.05 | 590.17 | 252.02 |
| **750 samples** | | | | |
| Applied model | 0% | 30% | 60% | 90% |
| Model 1 | 432.54 | 486.20 | 444.12 | 203.28 |
| Model 2 | 430.25 | 483.74 | 441.90 | 200.97 |
| Model 3 | 460.93 | 522.22 | 454.47 | 201.16 |
| Null model | 459.63 | 521.12 | 453.44 | 200.36 |
| **500 samples** | | | | |
| Applied model | 0% | 30% | 60% | 90% |
| Model 1 | 320.05 | 322.13 | 289.97 | 130.52 |
| Model 2 | 317.30 | 319.71 | 288.14 | 127.45 |
| Model 3 | 332.85 | 348.97 | 296.64 | 127.65 |
| Null model | 331.39 | 347.94 | 295.77 | 126.18 |
| **250 samples** | | | | |
| Applied model | 0% | 30% | 60% | 90% |
| Model 1 | 142.96 | 152.73 | 138.92 | 50.20 |
| Model 2 | 140.02 | 149.06 | 136.53 | 47.54 |
| Model 3 | 174.48 | 167.99 | 140.88 | 48.07 |
| Null model | 173.22 | 166.50 | 140.20 | 46.53 |

*Note*: The data of simb, simc, and csimb include different numbers (750, 500, and 250) sampled from 1000 simulation data (simb is used for Case 1). Furthermore, csimb is reproduced by increasing the zero number 30%, 60%, and 90%. Models are applied to all zero-inflated csimb in 1000, 750, 500, and 250 sampled data.

environmental data for the biological community. Figure 3 gives spatial plots for krill biomass (krill), sighting data of fin whales (w), surface temperature (sst), depth data (depth), slope for the depth (slope), and gradient surface temperature (sstgrd).

Data on water depth came from the ETOPO 1 bathymetric dataset available at https://www.ncei.noaa.gov/products/etopo-global-relief-model. Data were extracted for the middle position of each 1-nm segment throughout the survey tracks. Data on sea surface temperature sst were obtained from the OISST dataset available at https://www.ncei.noaa.gov/products/optimum-interpolation-sst and extracted in the same way as for depth data. The slope for the depth is the maximum rate of change in the depth from that cell to its neighbors calculated by the *Slope* tool of the Surface toolset in ArcGIS https://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/an-overview-of-the-surface-tools.htm. The lower the slope value, the flatter the terrain; the higher the slope value, the steeper the terrain. The gradient surface temperature is also calculated by *Slope* applied to sst.

### 4.1.2 | Case 2: Arctic Ocean minke whales

During the years 2014–2017, ecosystem surveys were conducted in August–September in the Arctic Ocean to the west and north of Svalbard (Solvang et al., 2021). Sampling included all trophic levels from phytoplankton to whales, as well as chemical and physical properties of the water masses in the area. In Solvang et al. (2021), the associations among

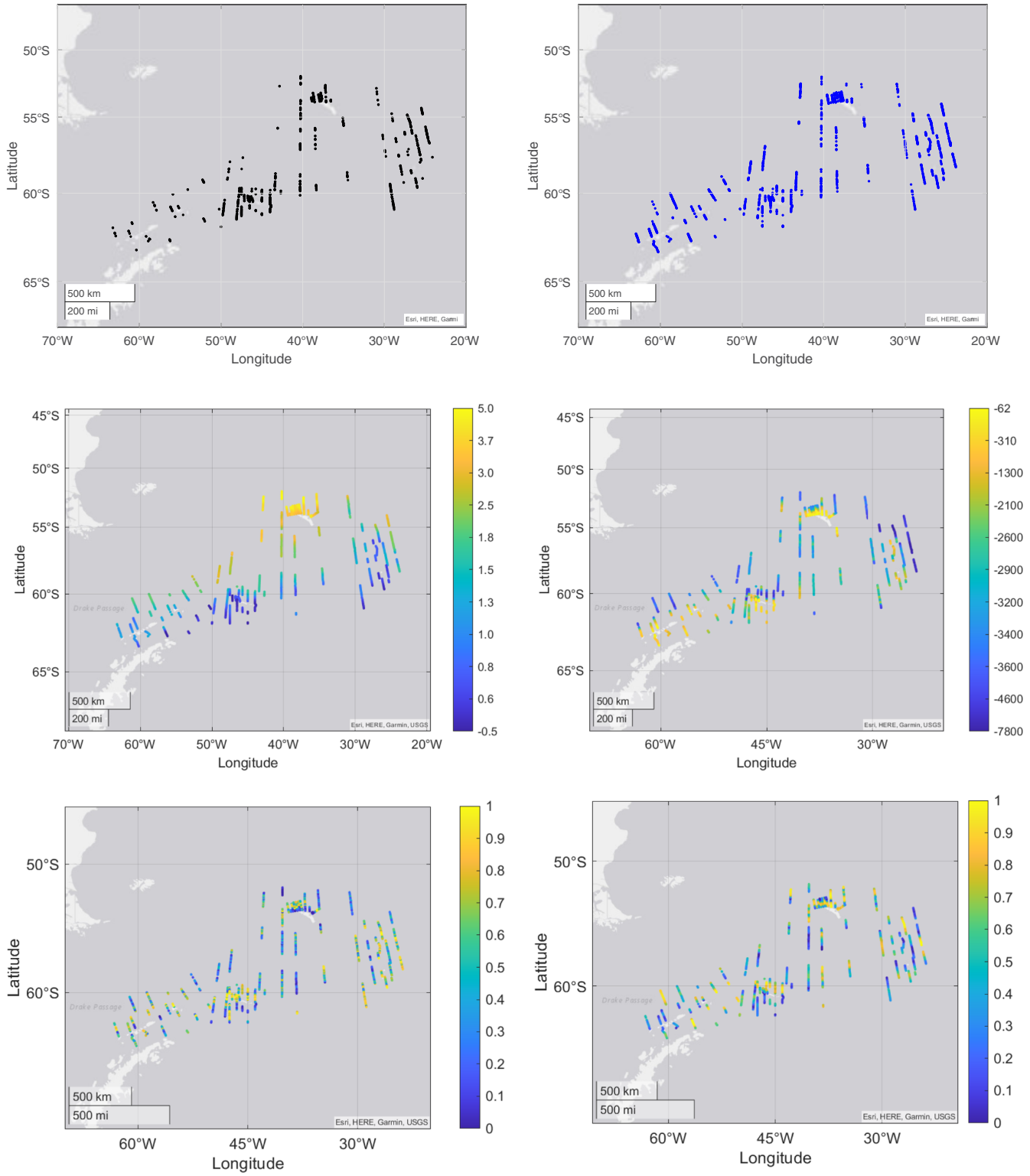


**FIGURE 3** Field observations for Case 1. Upper-left: sighting data of fin whales (counted number), upper-right: krill biomass $(gm^{-2})$, middle-left: surface temperature (°C), middle-right: depth data (m), lower-left: slope for depth (m), and lower-right: gradient temperature (°C). Brighter/darker dots' colors of surface temperature and water depth mean higher/lower temperature and shallower/deeper water depth, respectively.

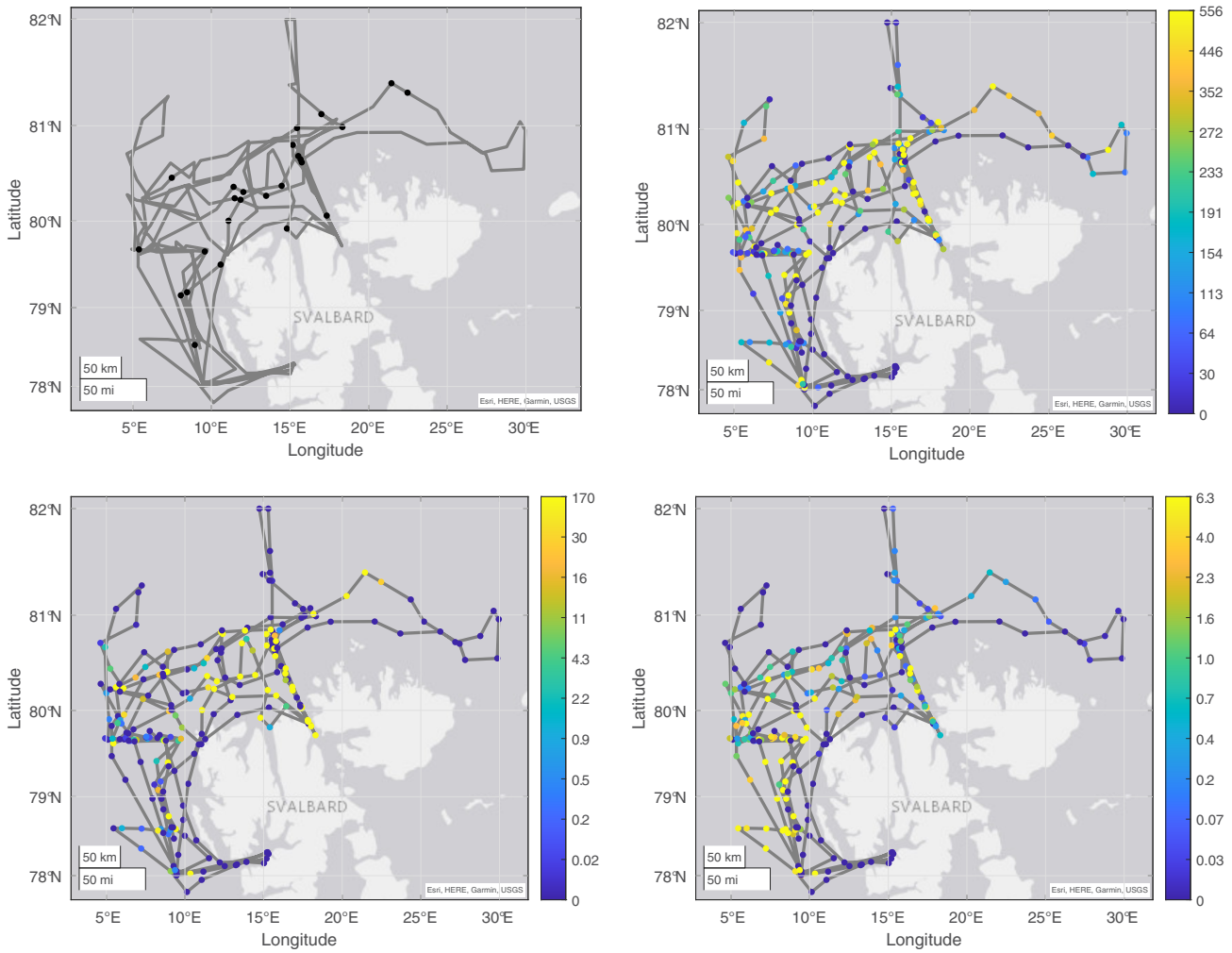


**FIGURE 4** Field observations for Case 2 along cruise track during autumn 2014–2017. Upper-left: sighting data of minke whales (counted number), upper-right: integrated values of $s_A$ $\left(\mathrm{m^2nmi^{-2}}\right)$ for plankton in the upper 200-m depth, lower-left: integrated values of $s_A$ $\left(\mathrm{m^2nmi^{-2}}\right)$ for Cod in the upper 200-m depth, and lower-right: integrated values of $s_A$ $\left(\mathrm{m^2nmi^{-2}}\right)$ for 0-group fish in the upper 200-m depth. Brighter/darker dots' colors of higher/lower integrated values of $s_A$.

minke whale, the predatory fish Atlantic cod, and 0-group fishes and zooplankton as their potential prey species were investigated. The data of minke whale are the counted number and the data of Atlantic cod, 0-group fishes, and zooplankton are acoustic registration as the nautical area scattering strength ($S_A$, dB re 1 $\mathrm{m^2nmi^{-2}}$, $S_A = 10\log 10(s_A)$). Figure 4 presents the data for a grid cell size of 50 km according to the transects of cruise and the aggregated upper 200 m in the depth.

## 4.2 | Preliminary analyses

Before applying the proposed procedure to the data, the relationships among variables are first investigated by the established regression models as a preliminary analysis. The counted number of sighted whales is converted into binary data, 1 for number > 0 and 0 for number = 0. The logistic regression model for whales and the linear regression model for the continuous response variable (James et al., 2013) are considered. The best model is selected by minimum AIC and then, the effect size (Lakens, 2013; Field, 2009) is also investigated. The logistic and linear regression analyses are conducted by *glm* and *lm* in R (R Core Team, 2023).

### 4.2.1 | Case 1

The krill biomass data included a substantial proportion of values at or close to zero, due to the swarming nature of krill, and the survey covering areas of both high and low biomass (Figure 3). Thus, logarithmic transformation of krill observations is required for the data by replacing zero with a small real number. We consider the logistic regression model (model w) for the binary data of whale and the linear regression model (model k) for the logarithmic-transformed krill biomass data. The AICs obtained by all model candidates are summarized in Supplementary Table 3 in S3. The best models selected by minimum AIC are.

$$\text{Model w}: \text{Probability of fin whale presence} = (1 + \exp(-0.76 + 0.014 \times \log(\text{krill}) - 0.21 \times \text{sst} - 3.4e-05 \times \text{Depth} + 0.058 \times \text{slope}))^{-1}, \text{and.}$$

$$\text{Model k}: \log(\text{krill}) = -7.1 + 1.2 \times \text{as factor (presence/absence of fin whale)} + 4.2e-04 \times \text{Depth} - 0.41 \times \text{SST} + 0.26 \times \text{slope.}$$

The *p*-values for the estimated coefficients in model w are 0.00050 for log(krill), 1.06e-09 for SST, 0.12 for Depth, and 3.01e-06 for slope, that is, three explanatory variables (except for Depth) were significantly associated with fin whale presence. On the other hand, the *p*-values for the estimated coefficients in model k are 0.00056 for whale, 4.8e-06 for depth, 0.0020 for sst, and 1.8e-06 for slope, that is, all explanatory variables were significantly associated with log(krill).

In the multinomial logistic regression (model w), the impact of predictor variables in a logistic regression model is explained in terms of the odds ratio, which reflects the effect size measures. The effect sizes for log(krill), SST, Depth, and Slope are given by the odds below:

$$\log(\text{krill}) : \exp(0.014) = 1.01,$$

$$\text{sst} : \exp(-0.21) = 0.81,$$

$$\text{depth} : \exp(-3.4e-05) = 1, \text{and}$$

$$\text{slope} : \exp(0.058) = 1.06.$$

The effect sizes for fin whales presence indicate a low value, 2% and 6% effects for increases in 1.0 logarithmic krill biomass and 1.0 meter for slope, respectively, and no effect for depth, while the effect size of fin whale presence shows 21% negative effects for an increase of 1°C.

For the output by linear model, the eta squared, which is calculated by the effect size of ANOVA (Ben-Shachar et al., 2020) indicates small effect sizes for all explanatory variables as below:

binary data of counted number of fin whales: 4.9e-03,

depth: 6.0e-03,

sst: 4.2e-03, and

slope: 6.0e-03.

### 4.2.2 | Case 2

Solvang et al. (2021) described the logistic regression analysis used to investigate whether certain prey groups were more or less likely to be present or absent when minke whales were present. The modeling was conducted for each prey species. In comparing the models including two and three explanatory variables, we apply the same regression analyses as done in Case 1. For prey species, we apply the linear regression model (model k) for the logarithmic-transformed plankton, cod, and 0-group fish biomass data. The AICs of all model combinations are summarized in Supplementary Table 4 in S3. The minimum AIC selected the models w, p, c, and 0 as below:

$$\text{Model w}: \text{Appearance probability of minke whale} = (1 + \exp(-2.09 + 0.08 \times \log(\text{plk}) + 0.04 \times \log(\text{cod})))^{-1}$$

$$\text{Model p}: \log(\text{plk}) = 0.74 + 0.12 \times \log(\text{cod}) + 0.75 \times \log(0\text{gr})$$

$$\text{Model c}: \log(\text{cod}) = -8.23 + 2.53 \times \text{as.factor(binary data of minke whale)} + 0.30 \times \log(\text{plk}) + 0.21 \times \log(0\text{gr})$$

$$\text{Model 0}: \log(0\text{gr}) = 0.78 + 0.92 \times \log(\text{plk}) + 0.10 \times \log(\text{cod})$$

The *p*-values for the estimated coefficients are 0.074 for log(plk) and 0.16 for log(cod) in Model w, 0.0028 for log(cod) and <2.0e-16 for log(0gr) in Model p, 0.11 for whale, 0.0038 for log(plk), and 0,028 for log(0gr) in Model c, and <2e-16 for log(plk) and 0.025 for log(cod) in Model 0.

The effect sizes for log(plk) and log(cod) in Model w are given by the odds below:

$$\log(\text{plk}) : \exp(0.082) = 1.09, \text{and}$$

$$\log(\text{cod}) : \exp(0.038) = 1.04.$$

The effect sizes for the appearance of minke whales indicate 9% and 4% effects for increases in 1.0 logarithmic plankton and cod biomass.

For output by the linear model, the eta squared, which is calculated by the effect size of ANOVA (Ben-Shachar et al., 2020) indicates small effect sizes for all explanatory variables as follows:

$$\text{Model p}: \log(\text{cod}) : 0.60 \text{ and } \log(0\text{gr}) : 0.69$$

$$\text{Model c}: \text{binary data of counted number of minke whales} : 0.05, \log(\text{plk}) : 0.31 \text{ and } \log(0\text{gr}) : 0.02, \text{and}$$

$$\text{Model 0}: \log(\text{plk}) : 0.78 \text{ and } \log(\text{cod}) : 0.02.$$

### 4.3 | Results by proposed procedure

First, we describe histograms of the observations in Figure 5 to find the region to use in searching for the optimum threshold to obtain two-category data. In Case 1, krill biomass showed a high distribution close to zero because there are many locations indicated by zero. Thus, logarithmic transformation of krill observations is required for the data by replacing zero with a small real number. The histogram of logarithmic-transformed data showed a dense distribution around small values and a long-tailed distribution with a single peak. The range for finding the optimum threshold for making two-category data is set from −3 to 8 (i.e., $a = -3$ and $b = 8$ in $[a, b]$ of subsection 2.3). For the surface temperature (sst), the histogram seems to have two different distributions around 2.0°C. The optimum threshold to make two categorical data is searched for within the range [−2.0, 2.5] (°C). For the water depth (depth), we set the range to [−1,000, 0.0] (m), where negative means under the sea surface. The slope (slope) and temperature gradient (sstgrd) are set to [1,15] (m) and [0.00001, 0.0017] (°C), respectively. In Case 2, we set the range to [2.0, 9.9] for logarithmic zooplankton biomass (plk), [−7.5, 9.0] for logarithmic cod biomass (cd), and [0.1, 9.9] for logarithmic 0-group fishes biomass (0gr).

In applying our proposed method to these data, Supplementary Figure 2 in S1 presents the plots for optimum thresholds in Cases 1 and 2. The threshold detected in each iteration is plotted by the light-gray's points and the averaged optimum thresholds are plotted by black-dashed lines. The averaged optimum thresholds are 2.39 for log(krill), 0.99°C for sst, −386.1 m for depth, 6.52 m for slope, and 0.00087°C for sstgrd in Case 1, and 4.41 for log(plk), 3.94 for log(cd), and 2.79 for log(0gr) in Case 2. Using the averaged optimum thresholds, we could obtain the categorical data for log(krill), sst, depth, slope, and sstgrd in Case 1 and for log(plk), log(cd), and log(x0g) in Case 2. Those data and the categorical data for whale are integrated, and the R function *catdap* is applied to the dataset. The response variable is set as whale or log(krill) in Case 1, and it is set as whale, plk, cd or x0g in Case 2. Against the response variables, we use several combinations of explanatory variables, for example, krill, sst, and depth for whale (in Case 1). Tables 3 and 4 present the calculated AICs for all possible models in Case 1 and Case 2, respectively. Using these categorical data, the smallest AIC selects the model including sst, depth, and slope for fin whale, and the model including depth, slope, and sstgrd for krill in Case 1 (Table 3). In Case 2, the smallest AIC selects the model including plk and cd for minke whale, the model including whale and cd for pl, the model including pl and 0gr for cd, and the model including pl and cd for 0gr (Table 4a). Table 4b summarizes the calculated AICs for all possible models among the prey species of minke whale. Supplementary Table 4 in S3 presents

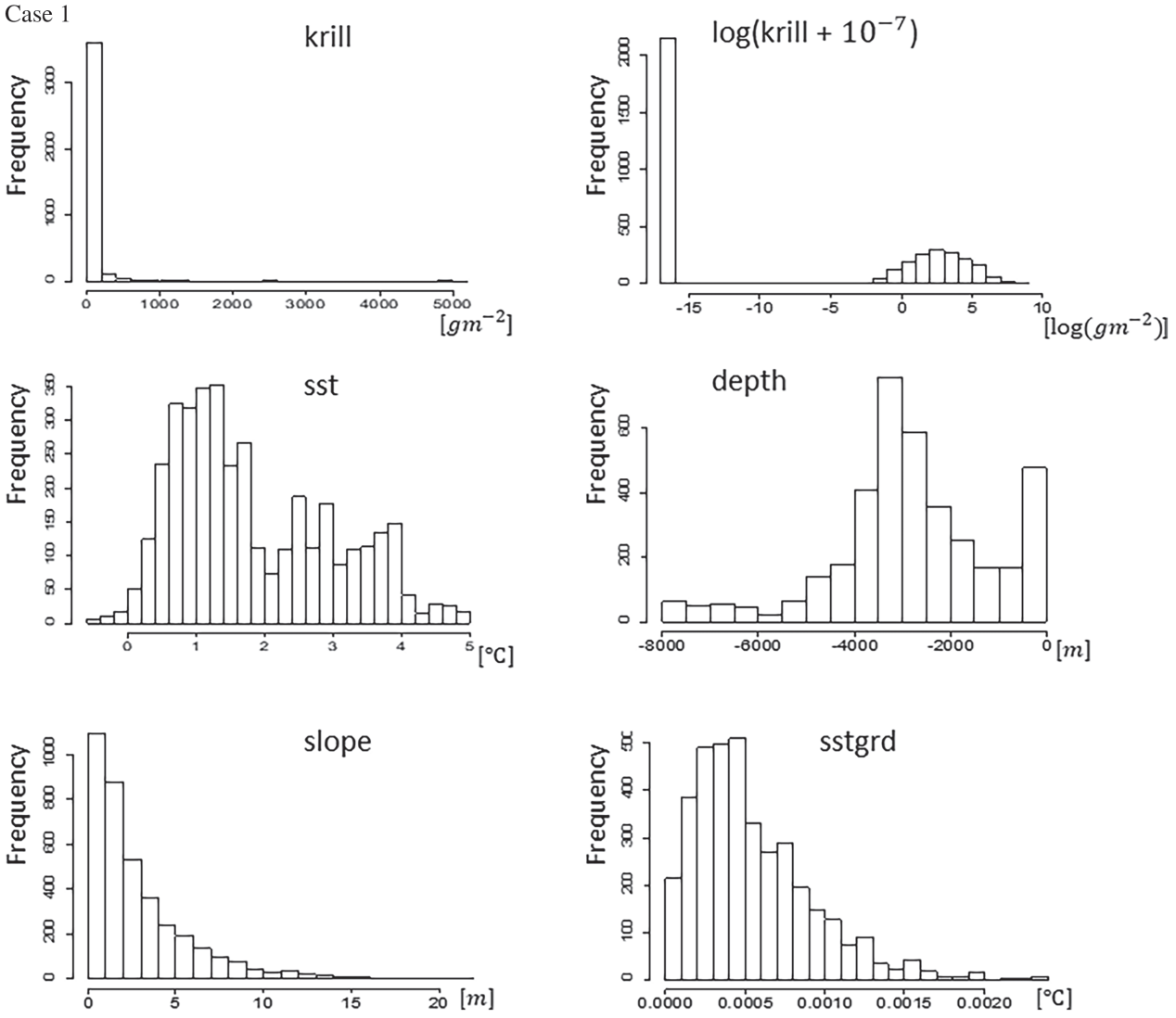


**FIGURE 5** Histograms of observations for Cases 1 and 2. These are used for setting the range [a,b] to detect the optimum threshold in a computational procedure. Case 1: Histograms of observations for krill biomass (upper-left, *x*-axis: biomass $\left(\mathrm{gm}^{-2}\right)$, *y*-axis: frequency), logarithmic transformation of krill biomass (upper-right, *x*-axis: logarithmic-transformed biomass value, *y*-axis: frequency), surface temperature (middle-left, *x*-axis:°C, *y*-axis: frequency), depth (middle-right, *x*-axis: meter, *y*-axis: frequency), slope (lower-left, *x*-axis: meter, *y*-axis: frequency), sst gradient (lower-right, *x*-axis:°C, *y*-axis: frequency). Case 2: Histograms of observations for logarithmic transformation of plankton biomass (top, *x*-axis: logarithmic-transformed biomass ($\mathrm{g\,m}^{-2}$), *y*-axis: frequency), logarithmic transformation of cod biomass (middle, *x*-axis: logarithmic-transformed biomass value, *y*-axis: frequency), and logarithmic transformed 0-group fish (bottom, *x*-axis: meter, *y*-axis: frequency).

the counted number of iterations for the models selected as the smallest AIC models in Cases 1 and 2. For Case 1, the best model for fin whale includes the highest counted number, while the secondary model for krill includes the highest counted number. For Case 2, while the best models for minke whale and plankton include the highest counted number, the best models for cod and 0gr do not necessarily include the highest counted number. Instead, models (cod; W) and (0gr; PL) include the highest counted number. These models are not the best models supported by the smallest AIC. The associations among plankton, cod, and 0gr also indicate a higher number for model (0gr; PL) than model (0gr; PL, CD) while the smallest AIC supports model (0gr; PL, CD). In fact, the field observation in Case 2 is high zero-inflated at 50% zero number in cod, 89% zero in whale, and 30% zero in 0gr.

The smallest AICs indicate that the species are associated with each other. Based on the smallest AIC models for Cases 1 and 2, we diagram the directional relationships among the data based on the best conditional probability models in Figure 6.

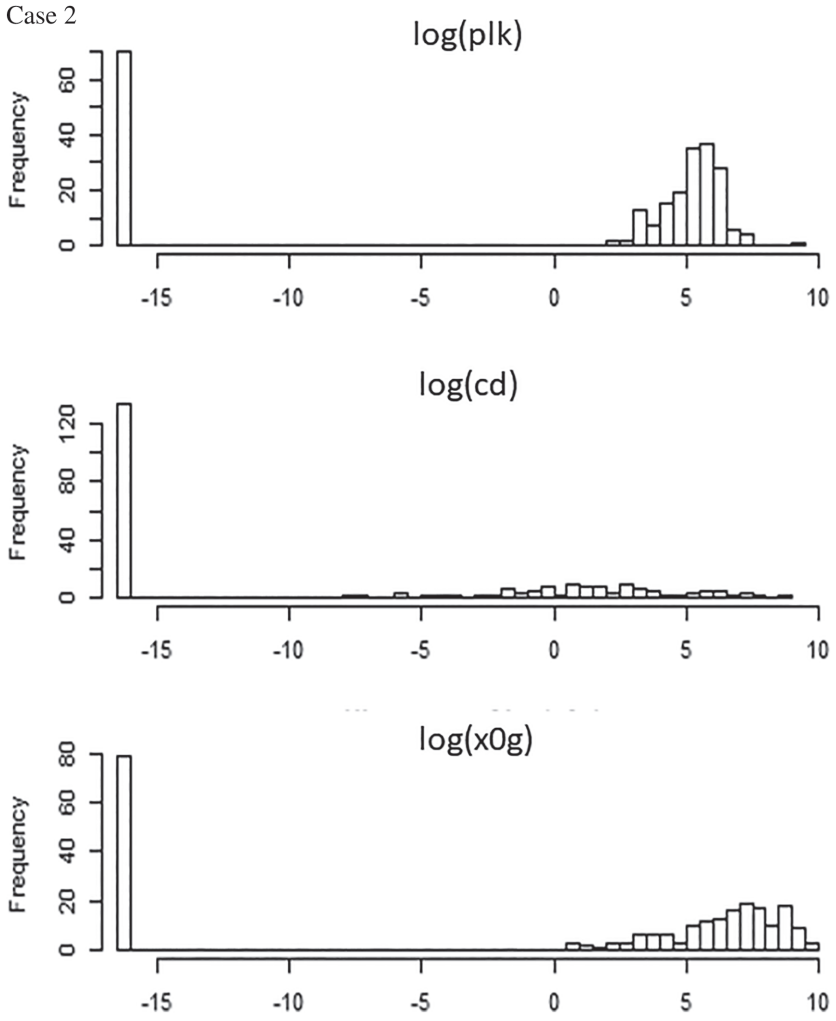


**FIGURE 5** (Continued)

Furthermore, we demonstrate how to apply the histogram model, which has been set for continuous variables in *catdap*. When setting fin whale as the response variable in Case 1, it automatically identifies 14 categories for sst, 26 categories for depth, 5 categories for slope; meanwhile, 3 categories are identified for krill and 1 category for sstgrd. Using these categorical data given by *catdap*, the minimum AIC selects the explanatory model including sst, depth and slope when the response variable is krill. For whale as a response variable, the minimum AIC selects the model including sst and depth. When setting minke whale as the response variable in Case 2, *catdap* automatically identifies three categories for plk, cd, and 0gr. The minimum AIC selects the model including plk and cd, as done in our method.

## 5 | DISCUSSION

As shown for the simulation experiments in Table 1, if the data include enough samples, our procedure identifies likely conditional probability models that explain directional relationships. However, the background of the data usually involves a complicated ecosystem, which is difficult to express by the simple model assumed in the simulation study. Furthermore, if the data include a small sampling number or is more zero-inflated, the model supported by the smallest AIC based on the averaged AIC may not be the consistently best model through computational iterations. To check the reliability of the results, the association from explanatory variables to response variable should be considered from a two-step procedure by the smallest AIC and counted iteration number for the model selected as the best model, especially for the data such as Case 2, which shows 89% zero in whale counted data.

**TABLE 3** Averaged Akaike Information Criterion results obtained by *catdap* applied to the categorical data based on the averaged optimum threshold in Case 1.

| **Response variable** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **To fin whale** | | | | **To krill** | | | |
| From | 1var | krill | 2227.4 | From | 1var | whale | 1728.0 |
| | | sst | 2188.5 | | | sst | 1729.9 |
| | | depth | 2227.6 | | | depth | 1716.1 |
| | | slope | 2217.5 | | | slope | 1730.3 |
| | | sstgrd | 2231.5 | | | sstgrd | 1730.3 |
| | 2var | krill, sst | 2184.6 | | 2var | whale, sst | 1726.0 |
| | | krill, depth | 2226.0 | | | whale, depth | 1714.5 |
| | | krill, slope | 2215.3 | | | whale, slope | 1728.1 |
| | | krill, sstgrd | 2228.1 | | | whale, sstgrd | 1726.9 |
| | | sst, depth | 2187.0 | | | sst, depth | 1716.6 |
| | | sst, slope | 2179.8 | | | sst, slope | 1727.3 |
| | | sst, sstgrd | 2185.1 | | | sst, sstgrd | 1728.1 |
| | | depth, slope | 2213.1 | | | depth, slope | 1714.0 |
| | | depth, sstgrd | 2228.2 | | | depth, sstgrd | 1715.8 |
| | | slope, sstgrd | 2219.0 | | | slope, sstgrd | 1730.2 |
| | 3var | krill, sst, depth | 2186.4 | | 3var | whale, sst, depth | 1715.9 |
| | | krill, sst, slope | 2178.5 | | | whale, sst, slope | 1726.1 |
| | | krill, sst, sstgrd | 2183.2 | | | whale, sst, sstgrd | 1725.4 |
| | | krill, depth, slope | 2212.6 | | | whale, depth, slope | 1713.6 |
| | | krill, depth, sstgrd | 2227.4 | | | whale, depth, sstgrd | 1714.8 |
| | | krill, slope, sstgrd | 2217.4 | | | whale, slope, sstgrd | 1728.8 |
| | | sst, depth, slope | 2177.1 | | | **sst, depth, slope** | **1712.7** |
| | | sst, depth, sstgrd | 2183.3 | | | sst, depth, sstgrd | 1715.3 |
| | | sst, slope, sstgrd | 2177.0 | | | sst, slope, sstgrd | 1726.0 |
| | | depth, slope, sstgrd | 2214.4 | | | depth, slope, sstgrd | 1714.9 |
| | 4var | krill, sst, depth, slope | 2180.2 | | 4var | whale, sst, depth, slope | 1715.8 |
| | | krill, sst, depth, sstgrd | 2186.1 | | | whale, sst, depth, sstgrd | 1717.4 |
| | | krill, sst, slope, sstgrd | 2179.0 | | | whale, sst, slope, sstgrd | 1727.6 |
| | | krill, depth, slope, sstgrd | 2215.6 | | | whale, depth, slope, sstgrd | 1716.1 |
| | | **sst, depth, slope, sstgrd** | **2174.9** | | | sst, depth, slope, sstgrd | 1712.9 |
| | 5var | krill, sst, depth, slope, sstgrd | 2182.3 | | 5var | whale, sst, depth, slope, sstgrd | 1728.4 |

*Note*: For the response variable (whale or krill), several explanatory variables listed in "From" are considered. For whale, the model explained by sst, depth, and slope is the best. On the other hand, for krill, the model explained by depth, slope, and sstgrd is the best. The relationships are illustrated in Figure 5. The numerical values indicate the minimum AICs. The models to response variable are the best models (in bold).

**TABLE 4** Akaike Information Criterion results obtained by *catdap* applied to the categorical data based on the averaged optimum threshold in Case 2.

| a | | Response variable | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | To minke whale | | To plankton | | To cod | | To 0gr | |
| Explanatory variable | 1 | PL | 81.0 | W | 143.1 | W | 66.2 | W | 155.4 |
| | | CD | 78.5 | CD | 138.2 | PL | 66.8 | PL | 114.0 |
| | | 0gr | 78.2 | **0gr** | **98.9** | 0gr | 62.4 | CD | 152.0 |
| | 2 | PL, CD | 80.0 | W, CD | 140.9 | W, PL | 63.6 | W, PL | 115.3 |
| | | PL, 0gr | 81.5 | W, 0gr | 102.1 | **W, 0gr** | **61.5** | W, CD | 151.6 |
| | | **CD, 0gr** | **77.4** | CD, 0gr | 99.4 | PL, 0gr | 63.0 | **PL, CD** | **112.7** |
| | 3 | PL, CD, 0gr | 81.5 | W, CD, 0gr | 105.0 | W, PL, 0gr | 63.9 | W, PL, CD | 116.3 |

| b | | Response variable | | | | | |
|---|---|---|---|---|---|---|---|
| | | To plankton | | To cod | | To 0gr | |
| Explanatory variable | 1 | CD | 138.2 | PL | 63.8 | PL | 114.0 |
| | | **0gr** | **98.9** | **0gr** | **62.4** | CD | 152.0 |
| | 2 | CD, 0gr | 99.4 | PL, 0gr | 63.0 | **PL, CD** | **112.7** |

*Note*: (a) For the response variable (minke whale, plankton, cod, and 0gr), several explanatory variables listed in "From" are considered. For minke whale, the model explained by the presence of plankton and cod is the best. For plankton, the model explained by the presence of cod and 0gr is the best. For cod, the model explained by the presence of plankton and 0gr is the best. Finally, for 0gr, the model explained by the presence of plankton and cod is the best. The relationships are illustrated in Figure 5. (b) For the response variable (plankton, cod, or 0gr) in prey species, several explanatory variables listed in "From" are considered. The best model for each response variable is explained by the presence of the other two species. The numerical values indicate the minimum AICs. The models to response variable are the best models (in bold).

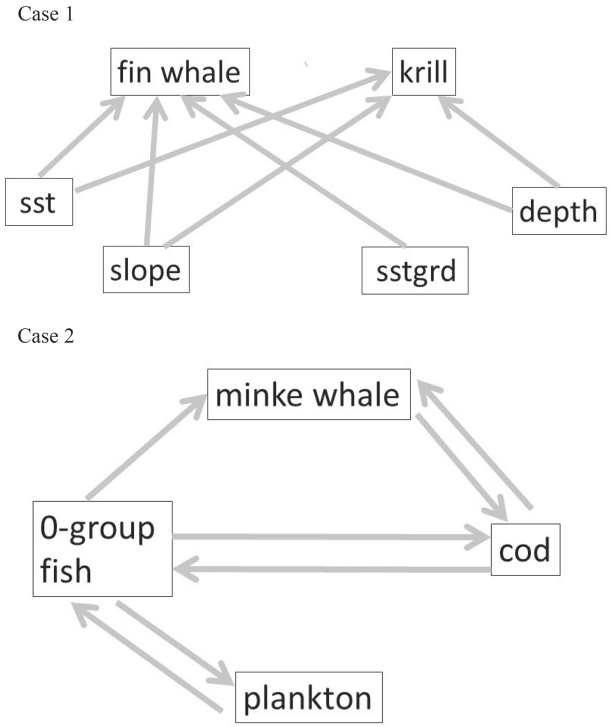


**FIGURE 6** Diagrams of the associations based on the model indicating minimum Akaike Information Criterion by the proposed method. Case 1: among fin whale, krill, surface temperature, depth, slope, and surface temperature gradient; Case 2: among minke whale, plankton, cod, and 0-group fish.

For real data, we also checked the median of AICs instead of the averaged AIC, considering the distribution of AIC. The outputs are summarized in Supplementary Table 5 in S3 for Cases 1 and 2. In Case 1, the identified model for fin whale is the same as the output by the averaged AIC, while the model identified by median selects slightly better model including sstgrd in addition to the sst, depth and slope that the averaged AIC model selected. The models selected by the median of AIC in Case 2 are the same as the model selected by the averaged AIC. By comparing with the median of AIC with the averaged AIC, we confirm that sst, depth and slope are reliable, at least for krill.

In our method, the optimum threshold to divide continuous data into two types of categorical data is found by taking account of the associations with the response variable, for example, whale or krill. The obtained threshold for the logarithmic krill biomass is 2.39, which is among the highest frequencies of the histogram in Figure 5, Case 1. A value less/larger than 2.39 is interpreted as a lower/higher value for the logarithmic biomass. The optimum threshold for surface temperature is 1.01, which is interpreted as a temperature warmer than 1.0°C and a lower temperature of less than 1.0°C. The optimum depth threshold is −386.1, which seems reasonable given the common knowledge of the fin whale's diving depth (Fonseca et al., 2022). The optimum threshold for the logarithmic plankton biomass is 4.41, which is among the highest frequencies of the histogram in Figure 5, Case 2. On the other hand, the optimum thresholds 3.94 for the logarithmic cod biomass and 2.79 for the logarithmic 0-group fish biomass seem to divide the values taking higher frequencies in the histograms. While the averaged optimum threshold is not directly related to employing the threshold of the best model as the optimal one, this could be used as an indicator to distinguish the continuous data into two levels, such as large/small or high/low. This gives a simple interpretation of the model's continuous dependent variables for biological/ecological experts or relevant stakeholders.

The preliminary regression analysis for Case 1 considered in Section 4.3 showed a significant effect from logarithmic krill biomass for fin whale presence and from fin whale presence for the logarithmic krill biomass by the model (Supplementary Table 3 in S3). The relationship from krill to fin whale is reasonable, but the relationship from fin whale to krill is not biologically reasonable. On the other hand, the best model identified by our proposed approach (Table 3) indicates that fin whale and krill may be indirectly related through an association with the slope This suggests that the association of the slope with whale and krill may be caused by the fact that continental slopes often create frontal systems that gather biological material, thus attracting krill and other planktonic organisms. In Case 2, the preliminary analysis and use of our approach supports the idea that plankton and cod are associated with 0-group fish. The 0-group is normally associated with the epipelagic zone, where they consume zooplankton (Eriksen et al., 2020; Solvang et al., 2021) and the 0-group includes cod in addition to redfish, haddock, capelin, and herring (Solvang et al., 2021). For minke whale, the regression model applied in the preliminary analysis selected the model having plankton and cod as the explanatory variables, but the coefficient of plankton is not significant. Our approach selected a model including 0-group fish and cod as the explanatory variables. While a relationship with plankton is not directly shown, the directional relationship from plankton to 0-group fish has already been shown in the outputs. In addition, this was supported by the output by our approach for 0-group fish, and Solvang et al. (2022) suggested that there exists a connection between cod abundance and feeding conditions, such as food competition, for other top predators. For plankton, the association from 0-group fish is supported by our approach and the preliminary analysis; however, it would be difficult to explain the contribution from cod that was seen in the best model by preliminary analysis. For cod, our approach and the preliminary analysis support the association from whale even if the effect size is small. Since the 0-group fish is associated with plankton, the best model by the preliminary analysis may be a redundant model.

Increasing demand for commercial harvesting of krill requires a careful assessment of sustainable harvest levels. As krill-dependent predators, several Southern Hemisphere populations of humpback whales (*Megaptera novaeangliae*) have undergone dramatic population recoveries in recent decades (Baines et al., 2021). Fin whales in the Southern Hemisphere were the most heavily exploited in terms of numbers taken during the period of intense industrial whaling, but the recent abundance of fin whales also suggests that they are undergoing a substantial recovery (Herr et al., 2022). The results of our analysis reflect the recent tendency between fin whales and krill biomass, and they suggest that our proposed method could contribute to the CCAMLR risk assessment and future management systems for Antarctic krill.

## 6 | CONCLUSION

We presented a categorical data analysis by combining the procedure (CATDAP) provided by Sakamoto and Akaike (1978), with a method to replace continuous values by two types of categorized data. The optimum threshold obtained by training data is used to categorize continuous explanatory variables to binary data using test data. This

procedure helps to avoid complicated categorization by a histogram model as conducted for continuous variables in the calculation of *catdap*. The proposed procedure is iterated a large enough number of times, and the averaged threshold is used as the optimum threshold. In practices, such a simple categorization based on the threshold would help to interpret the contribution from environmental factors for the biological community. The averaged AIC is also used for finally determining whether the response variable is dependent on the explanatory variable between the paired data. Using the spatial temporal data obtained by two scientific surveys, our proposed method attempts to objectively investigate the likelihood of several events, such as the number of overlapped locations among data across months or years. The conditional probability model identified by our approach gives us more information on the directional relationships among variables than applying a linear or logistic regression model; however, this is still not enough to infer the causal relationships among them. Knowing the associations among variables estimated by this method would be useful before applying a more specific model to estimate causality such as a directed acyclic graph (Pearl, 1988, 2009), which has been widely applied to Bayesian networks (Koller & Friedman, 2009; Neapolitan, 2003). Furthermore, this method is expected to be used as a practical monitoring tool in integrated ecosystem assessment, especially for following the temporal changes of associations among high-dimensional multivariate data for biological communities and oceanographic data reflecting the impact of human activities, for example, the spatio-temporal data used in several working groups of the International Council for the Exploration of the Sea (e.g., ICES, 2020).

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

## ORCID

*Hiroko Kato Solvang* https://orcid.org/0000-0002-0330-4670
*Shinpei Imori* https://orcid.org/0000-0002-6099-265X

## REFERENCES

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

Baines, M., Kelly, N., Reichelt, M., Lacey, C., Pinder, S., Fielding, S., Murphy, E., Trathan, P., Biuw, M., Lindstrøm, U., Krafft, B., & Jackson, J. (2021). Population abundance of recovering humpback whales *Megaptera novaeangliae* and other baleen whales in the scotia arc, South Atlantic. *Marine Ecology Progress Series*, *676*, 77–94. https://doi.org/10.3354/meps13849

Ben-Shachar, M. S., Lüdecke, D., & Makowski, D. (2020). Effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, *5*(56), 2815. https://doi.org/10.21105/joss.02815

Biuw, M., Lindstrøm, U., Jackson, J. A., Baines, M., Kelly, N., McCallum, G., Skaret, G., & Krafft, B. A. (2024). Estimated summer abundance and krill consumption of fin whales throughout the Scotia Sea during the 2018/2019 summer season. *Scientific Reports*, *14*(1), 7493. https://doi.org/10.1038/s41598-024-57378-3

Eriksen, E., Bagøien, E., Strand, E., Primicerio, R., Prokhorova, T., Trofimov, A., & Prokopchuk, I. (2020). The record-warm Barents Sea and 0-group fish response to abnormal conditions. *Frontiers in Marine Science*, *7*, 338. https://doi.org/10.3389/fmars.2020.00338

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). SAGE Publications Ltd.

Fonseca, C. T., Pérez-Jorge, S., Prieto, R., Oliveria, C., Tobeña, M., Scheffer, A., & Silva, M. A. (2022). Dive behavior and activity patterns of fin whales in a migratory habitat. *Frontiers in Marine Science*, *9*. https://doi.org/10.3389/fmars.2022.875731

Herr, H., Viquerat, S., Devas, F., Lees, A., Wells, L., Gregory, B., Giffords, T., Beecham, D., & Meyer, B. (2022). Return of large fin whale feeding aggregations to historical whaling grounds in the Southern Ocean. *Scientific Reports*, *12*, 9458. https://doi.org/10.1038/s41598-022-13798-7

ICES. (2020). Working group on the integrated assessments of the Norwegian Sea (WGINOR; outputs from 2020 meeting) 3; 35.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. Springer.

Katsura, K., & Sakamoto, Y. (1980). Computer science monograph, No.14, CATDAP, a categorical data analysis program package. The Institute of Statistical Mathematics.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. MIT Press.

Krafft, B. A., Macaulay, G. J., Skaret, G., Knutsen, T., Bergstad, O. A., Lowther, A., Huse, G., Fielding, S., Trathan, P., Murphy, E., Choi, S. G., Chung, S., Han, I., Lee, K., Zhao, X., Wang, X., Ying, Y., Yu, X., Demianenko, K., … Hoem, N. (2021). Standing stock of Antarctic krill (*Euphausia superba* Dana, 1850) (Euphausiacea) in the Southwest Atlantic sector of the Southern Ocean, 2018-19. *Journal of Crustacean Biology*, *41*(3), 1–17. https://doi.org/10.1093/jcbiol/rnab046

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863. https://doi.org/10.3389/fpsyg.2013.00863

Mersmann, O., Trautmann, H., Steuer, D., & Bornkamp, B. (2018). *Package 'truncnorm'*. CRAN.
Neapolitan, R. E. (2003). *Learning Bayesian networks*. Prentice Hall.
Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible Inference*. Morgan Kaufmann.
Pearl, J. (2009). *Causality: Models, reasoning and Inference* (2nd ed.). Cambridge University Press.
Pearson, K. (1990). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine. Series 5*, *50*(302), 157–175.
R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/
Sakamoto, Y., & Akaike, H. (1978). Analysis of cross classified data by AIC. *Annals Institute of Statistical Mathematics, 30*, 185-197. https://doi.org/10.1007/BF02480213.
Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). Akaike information criterion statistics. KTK Scientific Publishers; D. Reidel; Sold and distributed in the U.S.A. and Canada by Kluwer Academic Publishers.
Smith, A. N., Anderson, M. J., Millar, R. B., & Willis, T. J. (2014). Effects of marine reserves in the context of spatial and temporal variation: An analysis using Bayesian zero-inflated mixed models. *Marine Ecology Progress Series*, *499*, 203–216. https://doi.org/10.3554/meps10653
Solvang, H., Haug, T., Gjøsæter, H., Bogstad, B., Hartvedt, S., Øien, N., & Lindstrøm, U. (2021). Distribution of rorquals and Atlantic cod in relation to their prey in the Norwegian high Arctic. *Polar Biology*, *44*, 761–782. https://doi.org/10.1007/s00300-021-02835-2
Solvang, H., Haug, T., & Øien, N. (2022). Recent trends in temporal and geographical variation in blubber thickness of common minke whales (*Balaenoptera acutorostrata* acutorostrata) in the Northeast Atlantic. *NAMMCO Scientific Publications*, *12*. https://doi.org/10.7557/3.6308
Solvang, H. K., & Planque, B. (2020). Estimation and classification of temporal trends to support integrated ecosystem assessment. *ICES Journal of Marine Science*, *77*, 2529–2540. https://doi.org/10.1093/icesjms/fsaa111
Sugasawa, S., Nakagawa, T., Solvang, H. K., Subbey, S., & Alrabeei, S. (2022). Dynamic spatio-temporal zero-inflated Poisson models for predicting capelin distribution in the Barents Sea. *Japanese Journal of Statistics and Data Science.*, *6*, 1–20. https://doi.org/10.1007/s42081-022-00183-x
The Institute of Statistical Mathematics. (2023). Package 'catdap' Categorical Data Analysis Program package, version 1.3.7.
Ver Hoef, J. M., & Jansen, J. K. (2007). Space-time zero-inflated count models of harbor seals. *Environmetrics*, *18*, 697–712. https://doi.org/10.1002/env.873

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Solvang, H. K., Imori, S., Biuw, M., Lindstrøm, U., & Haug, T. (2024). Categorical data analysis using discretization of continuous variables to investigate associations in marine ecosystems. *Environmetrics*, e2867. https://doi.org/10.1002/env.2867