TR-No. 23-05, Hiroshima Statistical Research Group, 1–27

Estimation of Spatial Effects by Generalized Fused Lasso for Nonuniformly Sampled Spatial Data: An Analysis of the Body Condition of Common Minke Whales (*Balaenoptera acutorostrata acutorostrata*) in the Northeast Atlantic

Mariko Yamamura[1], Hirokazu Yanagihara[2*], Mineaki Ohishi[3], Keisuke Fukui[4], Hiroko Solvang[5], Nils Øien[5] and Tore Haug[6]

[1]Department of Statistics, Radiation Effects Research Foundation
5-2 Hijiyama Park, Minami-ku, Hiroshima 732-0815, Japan
[2]Mathematics Program, Graduate School of Advanced Science and Engineering, Hiroshima University
1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8526, Japan
[3]Center for Data-driven Science and Artificial Intelligence, Tohoku University
41 Kawauchi, Aoba-ku, Sendai, Miyagi 980-8576, Japan
[4]Faculty of Social Safety Sciences, Kansai University
7-1 Hakubaicho, Takatsuki, Osaka 569-1116 Japan
[5]Marine Mammals Research Group, Institute of Marine Research
P.O Box 1870 Nordnes NO-5817 Bergen, Norway
[6]Marine Mammals Research Group, Institute of Marine Research
P.O Box 6606 Stakkevollan NO-9296 Tromsø, Norway

Abstract

When using spatial data to estimate the effects of various factors on a given response variable, it is important to identify the influence of spatial effects. In this paper, we focus on a method for estimating spatial effects in a dataset detailing the body condition of common minke whales (*Balaenoptera acutorostrata acutorostrata*) in the Northeast Atlantic. The central idea of our approach is to estimate the spatial effects discretely rather than continuously, i.e., as coefficients of dummy variables indicating small spaces divided from the space to be analyzed. Using adjacency information for each segmented small space, our method joins neighboring spaces by using generalized fused lasso. This procedure is expected to be robust to nonuniformity in the density of spatial data, as adjacency information for the data is used rather than the coordinates of the spatial data. For model flexibility, effects other than spatial effects on the response variable are estimated in a penalized spline regression. The penalized spline regression is optimized via generalized ridge regression in order to reduce the time required for optimizing smoothness. To illustrate its applicability, we apply our estimation method to body condition data for common minke whales.

(Last Modified: July 30, 2023)

Key words: Backfitting algorithm, Generalized fused lasso, Graph trend filtering, Penalized spline regression, Spatio-temporal analysis.

*Corresponding author
E-mail address: yanagi-hiro@hiroshima-u.ac.jp (Hirokazu Yanagihara)

1. Introduction

In statistical spatial analysis, it is practically difficult to assume a uniform collection of spatial data from the target area, since the data can be both densely and sparsely distributed within that area. This can be seen, for example, in the locations at which body condition data for the common minke whale are collected in Norwegian waters. The body condition data include blubber thickness, girth, and length measured individually, and are typically recorded by commercial whalers in the Northeast Atlantic. The region for commercial whaling corresponds to the management areas defined by the International Whaling Commission. As shown in Figure 1, the region includes the Eastern Barents Sea (EB), the Svalbard-Bear Island area (ES), the Norwegian Sea and coastal zones off North Norway including the Lofton area (EW), the North Sea (EN), and the Western Norwegian Sea and Jan Mayen area (CM) subregions (see Solvang *et al.*, 2021). The yellow dots in the figure correspond to the positions of whaling during the period from 1993 to 2017. As can be seen, these positions are not uniformly sampled in the management subareas but rather are densely and sparsely sampled in specific areas within them.

The common minke whales migrate to higher latitudes for intensive feeding and, consequently, seasonal fattening. The deposited fat is stored as energy reserves for overwintering at lower latitudes. Accordingly, it is expected that the body condition of the whales in the summer grounds will reflect food availability during their most intensive feeding period and thus indicate how well the high-latitude ecosystems can support the population (Solvang *et al.*, 2017). Body condition data are recorded, along with the year, month, day, latitude and longitude, and sex of each individual whale. Using these data, temporal and spatial effects on body condition have heretofore been analyzed using ordinal linear, random effects and varying coefficient regression models (Solvang *et al.*, 2017), as well as a varying coefficient model in canonical correlation analysis (Yamamura *et al.*, 2016). In the varying coefficient model in Solvang *et al.* (2017) and Yamamura *et al.* (2016), where a polynomial was set as the coefficient, the estimated varying coefficients indicated significant temporal and spatial effects on blubber thickness according to whale migration in summer during the observed period.

While using polynomials in the varying coefficient model can provide a rough indication of the status of the spatial effect in the commercial whaling region, it is difficult to establish more

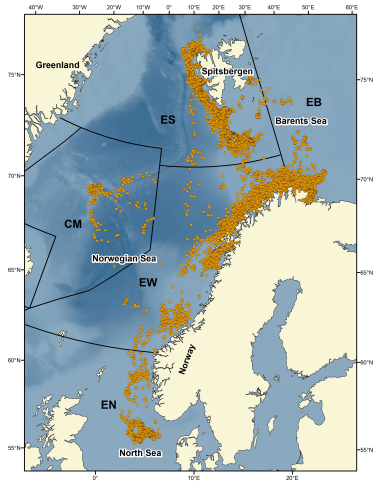

Figure 1. Capture positions in period from 1993 to 2017

precise changes in the corresponding spaces. In such cases, a nonparametric regression model using basis function expansion (see e.g., Härdle, 1990; Green & Silverman, 1994; Perperoglou *et al.*, 2019) may be more appropriate than using a polynomial function. Importantly, however; the model estimation may be sensitive to the density of the data (for scatter plot smoothing, see Yanagihara & Ohtaki, 2003). In particular, as mentioned above, the density of the spatial data that we are dealing with here is nonuniform, i.e., there is both densely and sparsely sampled spatial data. To address this challenge, we propose a method that alternately fits nonparametric regression and graph trend filtering via generalized fused lasso (see Wang *et al.*, 2016) using a backfitting algorithm (see e.g., Hastie & Tibshirani, 1990). Generalized fused lasso (GFL) is a generalization of the original fused lasso proposed by Tibshirani *et al.* (2005) with respect to the adjacencies that can be handled. Figure 2 shows the nature of the GFL estimation. In the GFL, we first segment the original space into small spaces. The GFL fuses the segmented small spaces with similar spatial effects. This procedure is expected to avoid overfitting influenced by the special state where the data are densely or sparsely sampled since only the adjacency information between the data is used rather than the coordinates of the data.

Our focus is specifically on the reduction of the time required to optimize the nonparametric regression by applying the noniterative optimization method for smoothness in penalized spline regression proposed by Yanagihara (2012). The method was applied in the formulation of generalized ridge regression (GRR; Hoerl & Kennard, 1970) to penalized spline regression. Using the results of Yanagihara (2018), we are able to explicitly obtain the smoothing parame-

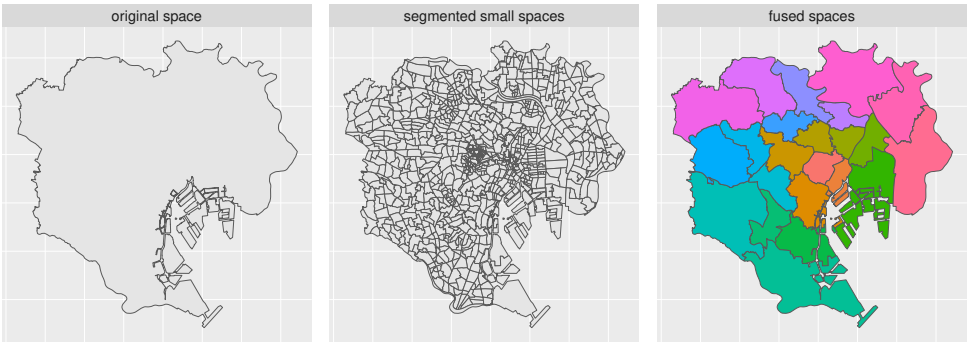

Figure 2. GFL estimation method

ters that minimize the generalized cross-validation (GCV) criterion (Craven & Wahba, 1979), which, unlike the C_p criterion (Mallows, 1973; 1995), does not require an asymptotically unbiased estimator of variance. Fukui *et al.* (2020) applied this penalized spline regression to the additive model and reported that the method not only dramatically reduces computation time, but also improves prediction accuracy. Unfortunately, Fukui *et al.* (2020) did not optimize the placement and number of basis functions. Of course, while it is desirable to optimize these hyper parameters, doing so effectively is quite troublesome. Therefore, in the present study, by using the same concept as the pseudospline in Hastie (1996), we optimize the rank of the smoother matrix as an alternative to those optimizations.

Solvang *et al.* (2022) reported the results of an analysis of minke whale blubber thickness similar to that described in this paper. Their estimation method is a predecessor of our proposed estimation method.

The remainder of the paper is organized as follows: Section 2 introduces the proposed model and estimation procedure; Section 3 describes the simulation study used to assess the predictive performance of our proposed method by applying it to nonuniformly sampled spatial dataset; Section 4 gives the numerical results produced by applying our method to body condition data for common minke whales; Section 5 provides a brief summary of the proposed procedures. Technical details are given in the Appendix.

2. Estimation Method

2.1. Vectorial Form of the Model

We first need to segment the space to be analyzed into m small spaces. If there is already a small space segmentation, such as by town or street address, this would suffice; otherwise, a space segmentation is established using, for example, a Voronoi diagram, so that the density

of the included data is approximately the same. Let y_{ij} be the ith response variable at the jth segmented small space ($i = 1, \ldots, n_j$; $j = 1, \ldots, m$), where n_j and m are the number of samples and the number of small spaces, respectively. Here, n is the combined sample size for all spaces, i.e., $n = \sum_{j=1}^{m} n_j$. We can express the additive model with a spatial effect as

$$y_{ij} = \sum_{\ell=1}^{p} f_\ell(x_{\ell,ij}) + \mu_j + \varepsilon_{ij}, \quad (i = 1, \ldots, n_j;\ j = 1, \ldots, m), \tag{2.1}$$

where f_ℓ is a trend indicating the influence of the ℓth explanatory variable, and $f_\ell(x_{\ell,ij}) = \beta_\ell x_{\ell,ij}$ not including the constant term when it is a multiple regression analysis, μ_j is the spatial effect of the ith sample at the jth small space, and ε_{ij} is the error term for the ith sample at the jth small space. It is assumed that ε_{ij}'s ($i = 1, \ldots, n_j$; $j = 1, \ldots, m$) are identically and independently distributed according to the distribution with $E[\varepsilon_{ij}] = 0$ and $Var[\varepsilon_{ij}] = \sigma^2$. In this paper, the function f_ℓ satisfies $\sum_{j=1}^{m} \sum_{i=1}^{n_j} f_\ell(x_{\ell,ij}) = 0$, which is a common constriction in additive models for a stable estimation.

In this paper, $f_\ell(x_{\ell,ij})$ is estimated nonparametrically using basis function expansions; however, for some $x_{\ell,ij}$, it may be desirable to estimate its value linearly. Therefore, we express $f_\ell(x_{\ell,ij})$ semiparametrically as follows:

$$\sum_{\ell=1}^{p} f_\ell(x_{\ell,ij}) = \sum_{\ell=1}^{p_1} \beta_\ell \left(x_{\ell,ij} - \bar{x}_\ell\right) + \sum_{\ell=p_1+1}^{p} \left\{s_\ell(x_{\ell,ij}) - \bar{s}_\ell\right\}, \tag{2.2}$$

where $\bar{x}_\ell$ and $\bar{s}_\ell$ are sample means of $x_{\ell,ij}$ and $s_\ell(x_{\ell,ij})$, respectively, i.e., $\bar{x}_\ell = n^{-1} \sum_{j=1}^{m} \sum_{i=1}^{n_j} x_{\ell,ij}$ and $\bar{s}_\ell = n^{-1} \sum_{j=1}^{m} \sum_{i=1}^{n_j} s_\ell(x_{\ell,ij})$. In (2.2), the first through p_1th explanatory variables are fitted using a linear model with regression coefficients $\beta_1, \ldots, \beta_{p_1}$. On the other hand, the p_1th through pth explanatory variables are fitted with an additive model based on spline regression. The following cubic spline without a constant term is considered as s_ℓ:

$$s_\ell(x_{\ell,ij}) = \beta_{\ell,1} x_{\ell,ij} + \beta_{\ell,2} x_{\ell,ij}^2 + \beta_{\ell,3} x_{\ell,ij}^3 + \sum_{g=1}^{b_0} \alpha_{\ell,g} (x_{\ell,ij} - \tau_{\ell,g})_+^3, \tag{2.3}$$

where b_0 is the number of basis functions, knot $\tau_{\ell,g}$ is the $100 \times g/(b_0 + 1)$ percentile of the ℓth explanatory variable, and $(x - \tau)_+^3 = I(x > \tau)(x - \tau)^3$ is the cubic power basis function. Here, $I(A)$ is an indicator function, i.e., $I(A) = 1$ if A is true and $I(A) = 0$ if A is not true. Let $\bar{c}_{\ell,g}$ ($g = 1, 2, 3$) and $\bar{b}_{\ell,g}$ ($g = 1, \ldots, b_0$) be sample means defined by

$$\bar{c}_{\ell,g} = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} x_{\ell,ij}^g, \quad \bar{b}_{\ell,g} = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} (x_{\ell,ij} - \tau_{\ell,g})_+^3.$$

Then, $\bar{s}_\ell$ is given as

$$\bar{s}_\ell = \beta_{\ell,1}\bar{c}_{\ell,1} + \beta_{\ell,2}\bar{c}_{\ell,2} + \beta_{\ell,3}\bar{c}_{\ell,3} + \sum_{g=1}^{b_0} \alpha_{\ell,g}\bar{b}_{\ell,g}.$$

Let $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ be k- and b-dimensional vectors, respectively, where $k = p_1 + 3(p - p_1)$ and $b = b_0(p - p_1)$, which are given by

$$\boldsymbol{\beta} = \left(\beta_1, \ldots, \beta_{p_1}, \boldsymbol{\beta}'_{p_1+1}, \ldots, \boldsymbol{\beta}'_p\right)', \quad \boldsymbol{\alpha} = \left(\boldsymbol{\alpha}'_{p_1+1}, \ldots, \boldsymbol{\alpha}'_p\right)', \tag{2.4}$$

and let $\boldsymbol{x}_{ij}$ and $\boldsymbol{b}_{ij}$ be k- and b-dimensional vectors defined as

$$\begin{aligned}
\boldsymbol{x}_{ij} &= \left(x_{1,ij} - \bar{x}_1, \ldots, x_{p_1,ij} - \bar{x}_{p_1}, \boldsymbol{c}_{p_1+1}(x_{p_1+1,ij})', \ldots, \boldsymbol{c}_p(x_{p,ij})'\right)', \\
\boldsymbol{b}_{ij} &= \left(\boldsymbol{b}_{p_1+1}(x_{p_1+1,ij})', \ldots, \boldsymbol{b}_p(x_{p,ij})'\right)',
\end{aligned}$$

where each subvector is given by

$$\begin{aligned}
\boldsymbol{\beta}_\ell &= (\beta_{\ell,1}, \beta_{\ell,2}, \beta_{\ell,3})', & \boldsymbol{c}_\ell(x) &= \left(x - \bar{c}_{\ell,1}, x^2 - \bar{c}_{\ell,2}, x^3 - \bar{c}_{\ell,3}\right)', \\
\boldsymbol{\alpha}_\ell &= (\alpha_{\ell,1}, \ldots, \alpha_{\ell,b_0})', & \boldsymbol{b}_\ell(x) &= \left((x - \tau_{\ell,1})_+^3 - \bar{b}_{\ell,1}, \ldots, (x - \tau_{\ell,b_0})_+^3 - \bar{b}_{\ell,b_0}\right)'.
\end{aligned}$$

Here, the notation "$'$" denotes the transpose of a vector or matrix. By using the above four vectors, (2.2) can be expressed as

$$\sum_{\ell=1}^{p} f_\ell(x_{\ell,ij}) = \boldsymbol{\beta}'\boldsymbol{x}_{ij} + \boldsymbol{\alpha}'\boldsymbol{b}_{ij}. \tag{2.5}$$

Let $\boldsymbol{y}_j$ and $\boldsymbol{\varepsilon}_j$ be n_j-dimensional vectors, $\boldsymbol{X}_j$ be an $n_j \times k$ matrix and $\boldsymbol{B}_j$ be an $n_j \times b$ matrix, which are given by

$$\begin{aligned}
\boldsymbol{y}_j &= (y_{1j}, \ldots, y_{n_j j})', & \boldsymbol{\varepsilon}_j &= (\varepsilon_{1j}, \ldots, \varepsilon_{n_j j})', \\
\boldsymbol{X}_j &= (\boldsymbol{x}_{1j}, \ldots, \boldsymbol{x}_{n_j j})', & \boldsymbol{B}_j &= (\boldsymbol{b}_{1j}, \ldots, \boldsymbol{b}_{n_j j})'.
\end{aligned}$$

Then, the vectorial form of the model for the jth space ($j = 1, \ldots, m$) is given by

$$\boldsymbol{y}_j = \boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{B}_j\boldsymbol{\alpha} + \mu_j\mathbf{1}_{n_j} + \boldsymbol{\varepsilon}_j, \quad (j = 1, \ldots, m), \tag{2.6}$$

where $\mathbf{1}_{n_j}$ is an n_j-dimensional vector of 1s. Using the vectorial forms of the first through mth spaces in (2.6), the vectorial form for all samples is given by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B}\boldsymbol{\alpha} + \boldsymbol{R}\boldsymbol{\mu} + \boldsymbol{\varepsilon}, \tag{2.7}$$

where $\boldsymbol{y}$ and $\boldsymbol{\varepsilon}$ are n-dimensional vectors, $\boldsymbol{X}$ is an $n \times k$ matrix, $\boldsymbol{B}$ is an $n \times b$ matrix, $\boldsymbol{\mu}$ is an m-dimensional vector, and $\boldsymbol{R}$ is an $n \times m$ matrix, which are given by

$$\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_m')', \qquad \boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1', \ldots, \boldsymbol{\varepsilon}_m')', \quad \boldsymbol{X} = (\boldsymbol{X}_1', \ldots, \boldsymbol{X}_m')',$$
$$\boldsymbol{B} = (\boldsymbol{B}_1', \ldots, \boldsymbol{B}_m')', \quad \boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)', \quad \boldsymbol{R} = \begin{pmatrix} \mathbf{1}_{n_1} \otimes \boldsymbol{e}_1' \\ \vdots \\ \mathbf{1}_{n_m} \otimes \boldsymbol{e}_m' \end{pmatrix}.$$

Here $\boldsymbol{e}_j$ $(j = 1, \ldots, m)$ is an m-dimensional vector of which the jth element is 1 and the others are 0, and $\otimes$ denotes the Kronecker product (see e.g., Harville, 1997).

In this paper, unknown parameters $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\mu}$ in (2.7) are estimated by the least squares (LS), penalized least squares (PLS), and GFL methods, respectively. The estimation of $\{\boldsymbol{\beta}, \boldsymbol{\alpha}\}$ and $\boldsymbol{\mu}$ is alternated by the backfitting algorithm (for more detail about the backfitting algorithm, see, e.g., Hastie & Tibshirani, 1990). This process is as follows:

Step 1. *Initialize*: Set $\hat{\boldsymbol{\mu}} = \bar{y}\mathbf{1}_m$, where $\bar{y}$ is the sample mean of $\boldsymbol{y}$, i.e., $\bar{y} = n^{-1}\mathbf{1}_n'\boldsymbol{y}$.

Step 2. *Cycle*: Update $\{\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}\}$ using the noniterative penalized spline regression proposed by Yanagihara (2012) with $\boldsymbol{y} - \boldsymbol{R}\hat{\mu}$ as the vector for response variables, and then update $\hat{\boldsymbol{\mu}}$ using the GFL with $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{B}\hat{\boldsymbol{\alpha}}$ as the vector of response variables.

Step 3. *Iterate*: Repeat step 2 until $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\mu}}$ change by less than a prespecified threshold.

After estimating the parameters, the estimate of f_ℓ $(\ell = p_1 + 1, \ldots, p)$ is given by

$$\hat{f}_\ell(x) = \hat{\boldsymbol{\beta}}_\ell'\boldsymbol{c}_\ell(x) + \hat{\boldsymbol{\alpha}}_\ell'\boldsymbol{b}_\ell(x) \quad (\ell = p_1 + 1, \ldots, p), \tag{2.8}$$

where $\hat{\boldsymbol{\beta}}_\ell$ and $\hat{\boldsymbol{\alpha}}_\ell$ are subvectors of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ corresponding to the division in (2.4), respectively. In our method, the smoothness of the penalized spline regression can be optimized without iterations, dramatically reducing the time required for optimization compared to existing methods (see Fukui *et al.*, 2020). In addition to the reduction in computation time, Fukui *et al.* (2020) also report improved prediction accuracy compared to existing methods.

2.2. Estimations of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ via the GRR

In this subsection, the estimation of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in (2.7) is considered under the given $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}$. Let $\boldsymbol{Q}$ be a $b \times b$ orthogonal matrix which diagonalizes $\boldsymbol{B}'(\boldsymbol{I}_n - \boldsymbol{G})\boldsymbol{B}$ as

$$\boldsymbol{Q}'\boldsymbol{B}'(\boldsymbol{I}_n - \boldsymbol{G})\boldsymbol{B}\boldsymbol{Q} = \boldsymbol{D} = \mathrm{diag}(d_1, \ldots, d_b), \quad (d_1 \geq \cdots \geq d_b), \tag{2.9}$$

where $\boldsymbol{G}$ is the projection matrix to the subspace spanned by the columns of $\boldsymbol{X}$, i.e., $\boldsymbol{G} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, $\boldsymbol{I}_n$ is an $n \times n$ unit matrix, and $d_1, \ldots, d_b$ are positive eigenvalues of $\boldsymbol{B}'(\boldsymbol{I}_n - \boldsymbol{G})\boldsymbol{B}$.

According to Fukui *et al.* (2020), $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are estimated by minimizing the penalized residual sum of squares (PRSS) in (A.1) in Appendix A.1. Here, $\boldsymbol{\Theta}$ is a $b \times b$ diagonal matrix given

by $\boldsymbol{\Theta} = \mathrm{diag}(\theta_1, \ldots, \theta_b)$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_b)'$ are non-negative multiple smoothing parameters. The minimizers of the PRSS are given by (A.2) in Appendix A.1; the GCV criterion for optimizing $\boldsymbol{\theta}$ is given by (A.4) in Appendix A.1. The $\boldsymbol{\theta}$ that minimizes GCV in (A.4) can be obtained in explicit form from the results of Fukui *et al.* (2020).

As noted in Section 1, we optimize the rank of the smoother matrix as an alternative to optimizing the placement and number of basis functions. Let $\boldsymbol{P}$ be an $n \times b$ matrix defined by

$$\boldsymbol{P} = (\boldsymbol{I}_n - \boldsymbol{G})\boldsymbol{B}\boldsymbol{Q}\boldsymbol{D}^{-1/2}. \tag{2.10}$$

It follows from (A.3) that the underlying smoother matrix is $\boldsymbol{P}\boldsymbol{D}(\boldsymbol{D} + \boldsymbol{\Theta})^{-1}\boldsymbol{P}'$. Here, we consider reducing the rank of the underlying smoother matrix from b to γ, i.e., using the rank-reduced smoothing matrix $\boldsymbol{P}_{[\gamma]}\boldsymbol{D}_{[\gamma]}(\boldsymbol{D}_{[\gamma]} + \boldsymbol{\Theta}_{[\gamma]})^{-1}\boldsymbol{P}'_{[\gamma]}$ rather than the underlying smoother matrix, where $\boldsymbol{P}_{[\gamma]}$ denotes the matrix consisting of the first through the γth columns of $\boldsymbol{P}$, and $\boldsymbol{D}_{[\gamma]}$ and $\boldsymbol{\Theta}_{[\gamma]}$ denote the $\gamma \times \gamma$ matrices consisting of the elements of the first through the γth columns and the first through the γth rows of $\boldsymbol{D}$ and $\boldsymbol{\Theta}$, respectively. Let $\boldsymbol{Q}_{[\gamma]}$ denote the $b \times \gamma$ matrix consisting of the first through the γth columns of $\boldsymbol{Q}$. Then, using $\boldsymbol{P}_{[\gamma]}\boldsymbol{D}_{[\gamma]}(\boldsymbol{D}_{[\gamma]} + \boldsymbol{\Theta}_{[\gamma]})^{-1}\boldsymbol{P}'_{[\gamma]}$ as the smoothing matrix is equivalent to using the minimizers of the following PRSS as the estimates:

$$\mathrm{PRSS}_{\boldsymbol{\theta}}(\boldsymbol{\alpha}, \boldsymbol{\beta} \,|\, \hat{\boldsymbol{\mu}}, \gamma) = \left\| \boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{\mu}} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{B}\boldsymbol{Q}_{[\gamma]}\boldsymbol{Q}'_{[\gamma]}\boldsymbol{\alpha} \right\|^2 + \boldsymbol{\alpha}'\boldsymbol{Q}\boldsymbol{\Theta}\boldsymbol{Q}'\boldsymbol{\alpha}, \tag{2.11}$$

where $\hat{\boldsymbol{\mu}}$ is the estimate of μ obtained in the previous iteration consisting of (2.18) and (2.19) given by $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}(\hat{\lambda})$. It is clear that $\mathrm{PRSS}_{\boldsymbol{\theta}}(\boldsymbol{\alpha}, \boldsymbol{\beta} \,|\, \hat{\boldsymbol{\mu}}, b)$ coincides with PRSS in (A.1). From elementary linear algebra, we know that the minimizers of (2.11) are invariant for any $\theta_{\gamma+1}, \ldots, \theta_b$. Hence, it is sufficient to optimize $\boldsymbol{\theta}_{[\gamma]} = (\theta_1, \ldots, \theta_\gamma)'$. A simple calculation implies that the GCV criterion for optimizing $\boldsymbol{\theta}_{[\gamma]}$ is defined by replacing $\boldsymbol{P}$, $\boldsymbol{D}$, and $\boldsymbol{\Theta}$ in (A.4) with $\boldsymbol{P}_{[\gamma]}$, $\boldsymbol{D}_{[\gamma]}$, and $\boldsymbol{\Theta}_{[\gamma]}$. Hence, the optimum $\boldsymbol{\theta}_{[\gamma]}$ is given by

$$\hat{\boldsymbol{\theta}}_{[\gamma]} = \arg\min_{\boldsymbol{\theta}_{[\gamma]} \in \mathbb{R}^\gamma} \frac{(\boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{\mu}})'\{\boldsymbol{I}_n - \boldsymbol{G} - \boldsymbol{P}_{[\gamma]}\boldsymbol{D}_{[\gamma]}(\boldsymbol{D}_{[\gamma]} + \boldsymbol{\Theta}_{[\gamma]})^{-1}\boldsymbol{P}'_{[\gamma]}\}^2(\boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{\mu}})}{[1 - \{k + \mathrm{tr}(\boldsymbol{D}_{[\gamma]}(\boldsymbol{D}_{[\gamma]} + \boldsymbol{\Theta}_{[\gamma]})^{-1})\}/n]^2}.$$

Let $\boldsymbol{z}_{[\gamma]}$ be a γ-dimensional vector consisting of the first through the γth elements of $\boldsymbol{z}$ defined by

$$\boldsymbol{z} = (z_1, \ldots, z_b)' = \boldsymbol{P}'(\boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{\mu}}), \tag{2.12}$$

and let $t(1 \,|\, \gamma) \leq \cdots \leq t(\gamma \,|\, \gamma)$ ($\gamma \in \{1, \ldots, b\}$) be the order statistics of $z_1^2, \ldots, z_\gamma^2$. Using these vectors and matrix, we define the following statistic:

$$s^2(a \,|\, \gamma) = \begin{cases} \dfrac{(\boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{\mu}})'(\boldsymbol{I}_n - \boldsymbol{G} - \boldsymbol{P}_{[\gamma]}\boldsymbol{P}'_{[\gamma]})(\boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{\mu}})}{n - k - \gamma} & (a = 0) \\ \dfrac{(n - k - \gamma)s^2(0 \,|\, \gamma) + \sum_{j=1}^{a} t(j \,|\, \gamma)}{n - k - \gamma + a} & (a = 1, \ldots, \gamma) \end{cases}. \tag{2.13}$$

Let $a^\star(\gamma)$ be the integer satisfying

$$a^\star(\gamma) \in \mathcal{A}(\gamma) = \left\{ a = \{0, 1, \ldots, \gamma\} \;\middle|\; s^2(a \,|\, \gamma) \in \pi(a \,|\, \gamma) \right\},$$
$$\pi(a \,|\, \gamma) = \begin{cases} (0, t(1 \,|\, \gamma)] & (a = 0) \\ (t(a \,|\, \gamma), t(a+1 \,|\, \gamma)] & (a = 1, \ldots, \gamma - 1) \\ (t(\gamma \,|\, \gamma), \infty) & (a = \gamma) \end{cases}. \tag{2.14}$$

From Yanagihara (2018), there exists only one $a \in \{0, 1, \ldots, \gamma\}$ satisfying $s^2(a \,|\, \gamma) \in \pi(a \,|\, \gamma)$, which means that $a^\star(\gamma)$ can be one of the values from 0 to γ, i.e., $\#(\mathcal{A}(\gamma)) = 1$. By using $s^2(a^\star(\gamma) \,|\, \gamma)$ and $\boldsymbol{z}$ in (2.12), we can prepare the $\gamma \times \gamma$ diagonal matrix $\boldsymbol{V}(\gamma) = \mathrm{diag}(v(1 \,|\, \gamma), \ldots, v(\gamma \,|\, \gamma))$ of which the jth diagonal element is given by

$$v(j \,|\, \gamma) = I\left(s^2(a^\star(\gamma) \,|\, \gamma) \leq z_j^2\right)\left(1 - \frac{s^2(a^\star(\gamma) \,|\, \gamma)}{z_j^2}\right).$$

By a similar calculation to Yanagihara (2018) and Fukui *et al.* (2020), the minimizers of (2.11) after optimizing $\boldsymbol{\theta}_{[\gamma]}$ can be obtained in explicit form as follows:

$$\hat{\boldsymbol{\alpha}}(\gamma) = \boldsymbol{Q}_{[\gamma]} \boldsymbol{V}(\gamma) \boldsymbol{D}_{[\gamma]}^{-1/2} \boldsymbol{z}_{[\gamma]}, \quad \hat{\boldsymbol{\beta}}(\gamma) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\left(\boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{\mu}} - \boldsymbol{B}\hat{\boldsymbol{\alpha}}(\gamma)\right). \tag{2.15}$$

The rank γ is optimized by minimizing an extended GCV (EGCV; Ohisihi *et al.*, 2020) with $\log n$ as

$$\hat{\gamma} = \arg\min_{\gamma \in \{1, \ldots, b\}} \frac{(\boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{\mu}})'(\boldsymbol{I}_n - \boldsymbol{G} - \boldsymbol{P}_{[\gamma]}\boldsymbol{V}(\gamma)\boldsymbol{P}_{[\gamma]}')^2(\boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{\mu}})}{\{1 - (k + \gamma)/n\}^{\log n}}. \tag{2.16}$$

The EGCV with $\log n$ is asymptotically equivalent to the Bayesian information criterion (BIC) proposed by Schwarz (1978) under the normality assumption. Since an optimization of the rank γ is in a sense the same as optimizing the number of basis functions to be used, we use a consistent model selection criterion for selecting γ. Hence, in this study, since the GCV is used to optimize the smoothing parameters $\boldsymbol{\theta}$, EGCV with $\log n$ is correspondingly used to optimize γ. Using the optimized γ, estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ after optimizing γ are given by $\hat{\boldsymbol{\alpha}}(\hat{\gamma})$ and $\hat{\boldsymbol{\beta}}(\hat{\gamma})$, respectively.

2.3. Estimation of $\boldsymbol{\mu}$ via the GFL

In this subsection, the estimation of $\boldsymbol{\mu}$ in (2.7) is considered under the given $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$. The penalized residual sum of squares (PRSS_λ) of the GFL is as follows:

$$\mathrm{PRSS}_\gamma\left(\boldsymbol{\mu} \mid \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}\right) = \left\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{B}\hat{\boldsymbol{\alpha}} - \boldsymbol{R}\boldsymbol{\mu}\right\|^2 + \lambda \sum_{j=1}^{m} \sum_{\ell \in \mathcal{D}_j} \omega_{j\ell} |\mu_j - \mu_\ell|, \tag{2.17}$$

where $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ are the estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ obtained in the previous iteration consisting of (2.15) and (2.16) given by $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}(\hat{\gamma})$ and $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\hat{\gamma})$, λ is a non-negative tuning parameter, and $\mathcal{D}_j$ is the neighboring set of the jth space ($j = 1, \ldots, m$). The neighboring set $\mathcal{D}_j$ is the set for which i is one of elements if the ith space is adjacent to the jth space ($j = 1, \ldots, m; i \neq j$) . For example, if the second, third and fifth spaces are adjacent to the first space, then $\mathcal{D}_1 = \{2, 3, 5\}$. Here, m_j is the number of elements in $\mathcal{D}_j$, i.e., $m_j = \#(\mathcal{D}_j)$. Furthermore, let $\omega_{j\ell}$ be the weight for adaptive lasso (Zou, 2006), $\omega_{j\ell} = |\bar{\mu}_j - \bar{\mu}_\ell|^{-1}$, where $\bar{\mu}_j$ is the jth element of $\bar{\boldsymbol{\mu}}$ defined by

$$\bar{\boldsymbol{\mu}} = (\boldsymbol{O}_{m,k+b}, \boldsymbol{I}_m)\{(\boldsymbol{X}, \boldsymbol{B}, \boldsymbol{R})'(\boldsymbol{X}, \boldsymbol{B}, \boldsymbol{R})\}^{-1}(\boldsymbol{X}, \boldsymbol{B}, \boldsymbol{R})'\boldsymbol{y},$$

where $\boldsymbol{O}_{m,k+b}$ is an $m \times (k + b)$ matrix of zeros.

Let $\hat{\boldsymbol{\mu}}(\lambda)$ be the minimizer of (2.17), i.e.,

$$\hat{\boldsymbol{\mu}}(\lambda) = \arg\min_{\boldsymbol{\mu} \in \mathbb{R}^m} \mathrm{PRSS}_\lambda\left(\boldsymbol{\mu} \mid \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}\right). \tag{2.18}$$

We optimize λ by minimizing the GCV as

$$\hat{\lambda} = \arg\min_{\lambda \in \Lambda} \frac{\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{B}\hat{\boldsymbol{\alpha}} - \boldsymbol{R}\hat{\boldsymbol{\mu}}(\lambda)\|^2}{(1 - \mathrm{df}(\lambda)/n)^2} \tag{2.19}$$

where $\mathrm{df}(\lambda)$ is the number of distinct elements in $\hat{\boldsymbol{\mu}}(\lambda)$. The $\hat{\boldsymbol{\mu}}(\lambda)$ and $\hat{\lambda}$ are recursively calculated. Based on the domain for λ, $\Lambda = \{\lambda_1, \ldots, \lambda_{100}\}$ is set, where $\lambda_a = \lambda_{100}(0.75)^{100-a}$ ($a = 1, \ldots, 100$). Here, λ_{100} is given by

$$\lambda_{100} = \max_{j=\{1,\ldots,m\}} \frac{n_j|\tilde{\mu}_j - \hat{\mu}_\infty|}{\sum_{\ell \in \mathcal{D}_j} \omega_{j\ell}},$$

and then,

$$\tilde{\mu}_j = \frac{\mathbf{1}_{n_j}'(\boldsymbol{y}_j - \boldsymbol{X}_j\hat{\boldsymbol{\beta}} - \boldsymbol{B}_j\hat{\boldsymbol{\alpha}})}{n_j}, \quad \hat{\mu}_\infty = \frac{\mathbf{1}_m'\boldsymbol{R}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{B}\hat{\boldsymbol{\alpha}})}{n}.$$

Setting λ from λ_1 to λ_{100}, the calculation is recursively conducted to obtain $\hat{\boldsymbol{\mu}}(\lambda_1), \ldots, \hat{\boldsymbol{\mu}}(\lambda_{100})$, respectively. Finally, the optimum estimation for $\boldsymbol{\mu}$ is derived. Since $\hat{\boldsymbol{\mu}}(\lambda_a)$ ($a = 1, \ldots, 100$) cannot be obtained by closed forms, unlike $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$, the coordinate descent algorithm for the GFL (see e.g., Ohishi *et al.*, 2021) should be applied to determine $\hat{\boldsymbol{\mu}}(\lambda_a)$.

3. Simulation Study

To validate our proposed method for nonuniformly sampled spatial data, we conducted a simulation study. In order to produce nonuniformly sampled spatial data for the simulation, we generated n-dimensional vectors $\boldsymbol{u}_1 = (u_{1,1}, \ldots, u_{1,n})'$ and $\boldsymbol{u}_2 = (u_{2,1}, \ldots, u_{2,n})'$ as the

longitude and latitude. First ϵ $(= n/100)$ of $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are generated uniformly as follows:

$$u_{1,1}, \ldots, u_{1,\epsilon} \sim i.i.d.\ U(1, 10),$$
$$u_{2,1}, \ldots, u_{2,\epsilon} \text{ are mutually independent and } u_{2,i} \sim \begin{cases} U(g_1(u_{1,i}), g_2(u_{1,i})) & (u_{1,i} \in [2, 6]) \\ U(g_1(u_{1,i}), 11) & (u_{1,i} \notin [2, 6]) \end{cases},$$

where g_1 and g_2 are functions defined by

$$g_1(x) = \log x, \quad g_2(x) = (x - 4)^2 + 7. \tag{3.1}$$

The remaining n_ϵ $(= n - \epsilon)$ of $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are generated nonuniformly as follows:

(1) The n_ϵ-dimensional random vector $\boldsymbol{a}_0 = (a_{0,1}, \ldots, a_{0,n_\epsilon})'$ is generated from the mixture normal distribution, as $a_{0,1}, \ldots, a_{0,n_\epsilon} \sim i.i.d.\ 0.2 \cdot N(0, 1) + 0.3 \cdot N(15, 5)$.

(2) The $n_\epsilon/2$-dimensional random vectors $\boldsymbol{a}_1 = (a_{1,1}, \ldots, a_{1,n_\epsilon/2})'$ and $\boldsymbol{a}_2 = (a_{2,1}, \ldots, a_{2,n_\epsilon/2})'$ are set by transferring the first $n_\epsilon/2$ of $\boldsymbol{a}_0$ to $[1, 10]$ and the remainder of $\boldsymbol{a}_0$ to $[2, 6]$, respectively, i.e., given as

$$\begin{aligned} a_{1,i} &= 1 + 9\frac{a_{0,i} - \min_{j=1,\ldots,n_\epsilon/2} a_{0,j}}{\max_{j=1,\ldots,n_\epsilon/2} a_{0,j} - \min_{j=1,\ldots,n_\epsilon/2} a_{0,j}} \\ a_{2,i} &= 2 + 4\frac{a_{0,n_\epsilon/2+i} - \min_{j=n_\epsilon/2+1,\ldots,n} a_{0,j}}{\max_{j=n_\epsilon/2+1,\ldots,n} a_{0,j} - \min_{j=n_\epsilon/2+1,\ldots,n} a_{0,j}} \end{aligned} \quad (i = 1, \ldots, n_\epsilon/2).$$

(3) The random variables $\xi_{1,i}$ and $\xi_{2,i}$ $(i = 1, \ldots, n_\epsilon/2)$ are generated by exponential distributions with the parameters 2.5 and 3.0, respectively.

(4) The $(n - \epsilon)/2$-dimensional random vectors $\boldsymbol{b}_1 = (b_{1,1}, \ldots, b_{1,n_\epsilon/2})'$, and $\boldsymbol{b}_2 = (b_{2,1}, \ldots, b_{2,n_\epsilon/2})'$ are defined by

$$b_{1,i} = g_1(a_{1,i}) + \xi_{1,i}, \quad b_{2,i} = g_2(a_{2,i}) - \xi_{2,i} \quad (i = 1, \ldots, n/2).$$

(5) The n_ϵ longitude and latitude are given by $(u_{1,\epsilon+1}, \ldots, u_{1,n})' = (\boldsymbol{a}_1', \boldsymbol{a}_2')'$, and $(u_{2,\epsilon+1}, \ldots, u_{2,n})' = (\boldsymbol{b}_1', \boldsymbol{b}_2')'$.

We also create the following four divided subregions:

$$\begin{aligned} &\text{Subregion A } (u_{1,i} \le 2.7, u_{2,i} > 4.7); \quad \text{Subregion B } (u_{1,i} > 2.7, u_{2,i} > 4.7); \\ &\text{Subregion C } (u_{1,i} \le 3.1, u_{2,i} \le 4.7); \quad \text{Subregion D } (u_{1,i} > 3.1, u_{2,i} \le 4.7). \end{aligned} \tag{3.2}$$

When actually dividing the space for use of the GFL, the space is divided along these subregions, as are the data for the common minke whales in Norwegian waters shown in Figure 1.

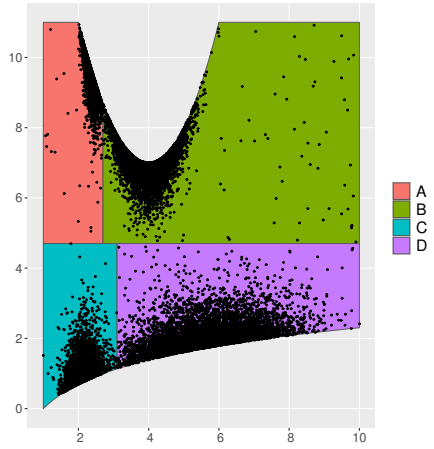

Figure 3. Sample points

Figure 3 shows the four subregions and the sample points for data generated by the simulation procedure. The domain for obtaining the sample is defined as $\mathcal{R} \subset [1, 10] \times [0, 11]$, where $\mathcal{R}$ is given by

$$(x, y) \in \mathcal{R} \Longleftrightarrow x \in [1, 10] \text{ and } \begin{cases} g_1(x) < y \leq 11 & (11 < g_2(x)) \\ g_1(x) < y < g_2(x) & (11 \geq g_2(x)) \end{cases},$$

where functions g_1 and g_2 are given by (3.1). In a practical situation, the area within the domain corresponds to ocean, while the area outside the domain corresponds to land. The simulation data are generated by the following model:

$$\begin{aligned} y_i &\sim N\left(\eta(x_{1,i}, \ldots, x_{4,i}, u_{1,i}, u_{2,i}), 1\right) \quad (i = 1, \ldots, n), \\ \eta(x_{1,i}, \ldots, x_{4,i}, u_{1,i}, u_{2,i}) &= f_1(x_{1,i}) + \cdots + f_4(x_{4,i}) + g(u_{1,i}, u_{2,i}), \end{aligned}$$

where $x_{\ell,i}$ $(\ell = 1, \ldots, 4; i = 1, \ldots, n)$ is generated independently by $U(0, 1)$, and $f_\ell(x)$ $(\ell = 1, \ldots, 4)$ is the true trend given by

- f_1 (Hastie *et al.*, 2001): $f_1(x) = \dfrac{\sin(12x + 0.2)}{x + 0.2}$.
- f_2 (partial linear trend): $f_2(x) = \begin{cases} -60(x - 17/60)^2 + 16/15 & (x < 1/4) \\ 4x & (1/4 \leq x < 3/4) \\ 80(x - 29/40)^2 + 59/20 & (3/4 \leq x) \end{cases}$.
- f_3 (linear trend): $f_3(x) = 6x$.

- f_4 (Wand, 2000): $f_4(x) = 8\{1.5\phi((x-0.35)/0.15) - \phi((x-0.8)/0.04)\}$, where $\phi(x)$ is the distribution function for the standard normal distribution.

The shape of each trend is shown in Figure 4. Here, $g(x, y)$ indicates a true spatial effect, which is defined using the following function, g_0:

$$g_0(x, y) = 15\left|(x-5)\exp\left\{-\frac{1}{5}(x-5)^2 - \frac{1}{5}x - \frac{1}{5}(y-5)^2\right\}\right|.$$

For the true spatial effects, three cases are considered. Cases 1 and 2 involve a spatial effect that changes discretely. Let $E_1^*, \ldots, E_\delta^*$ be polygons denoting the true subareas for the simulation domain, where δ is the number of subareas, and let $\boldsymbol{\zeta}_\ell = (\zeta_{\ell,1}, \zeta_{\ell,2})'$ be the center of gravity of the polygon vertices in the true subarea E_ℓ^* ($\ell = 1, \ldots, \delta$). Now, for $(x, y) \in E_\ell^*$, the discrete spatial effect for the response variable, g, is defined by

$$g(x, y) = g_0(\zeta_{\ell,1}, \zeta_{\ell,2})$$

In Case 1, $\delta = 9$; in Case 2, $\delta = 25$. The values $E_1^*, \ldots, E_\delta^*$ are determined using Voronoi decomposition, the details of which will be described later. Case 3 includes continuous variation of the spatial effect, given by

$$g(x, y) = g_0(x, y).$$

Figure 5 shows the spatial effect $g(x, y)$ for each case. The right, middle and left panels indicate the $g(x, y)$ for Cases 1, 2 and 3, respectively.

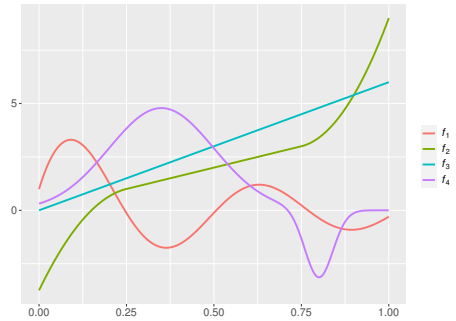

Figure 4. True trends in the simulation study

Prediction accuracy is evaluated using two types of MSE, which are given for data points and for points other than data points gridded in the space, defined by

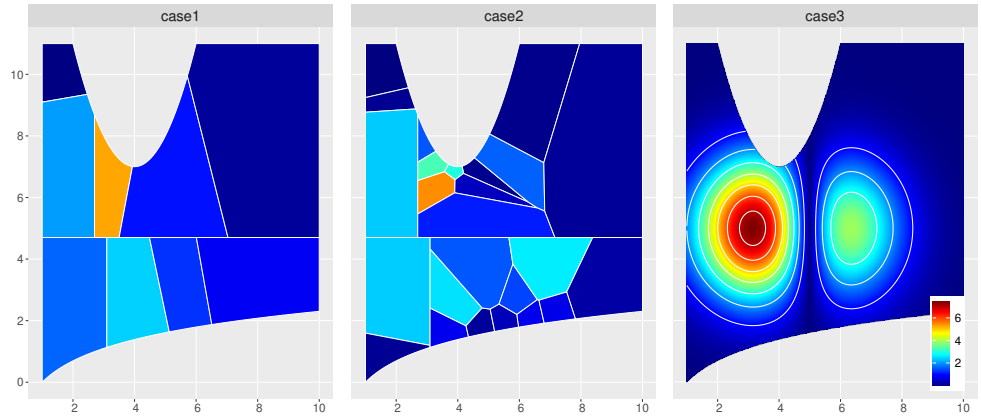

Figure 5. True spatial effects in simulation study

$$\mathrm{MSE} = \begin{cases} E\left[\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\{\eta(x_{1,i},\ldots,x_{4,i},u_{1,i},u_{2,i}) - \hat{\eta}(x_{1,i},\ldots,x_{4,i},u_{1,i},u_{2,i})\}^2\right] & \text{(sample points)} \\ E\left[\dfrac{1}{100}\displaystyle\sum_{i=1}^{100}\{\eta(\tau_i,\ldots,\tau_i,\kappa_i,\nu_i) - \hat{\eta}(\tau_i,\ldots,\tau_i,\kappa_i,\nu_i)\}^2\right] & \text{(grid points)} \end{cases},$$

where $\tau_i = (i-1)/99$ $(i = 1,\ldots,100)$ is obtained by dividing $[0,1]$ into 99 equal parts, and (κ_i, ν_i) $(i = 1,\ldots,100)$ is obtained by dividing the domain $[1,10]\times[0,11]$ into 9 vertical equal parts and 100 points in the 13 horizontal grids contained in $\mathcal{R}$ (see Figure6). MSE is determined

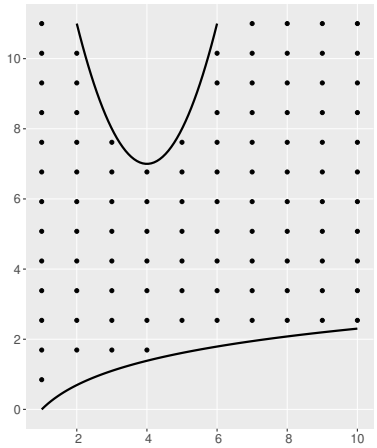

Figure 6. Grid points (κ_i, ν_i) $(i = 1,\ldots,100)$

using 1,000 Monte Carlo iterations. The number of basis functions b_0 is 15. The segmented space for applying the GFL is divided for each area in (3.2). The division in the subregion is conducted so as to be around the n_0 sample number included in each segmented small space, where n_0 = 50, 100, 300, 500, 1,000 for n = 5,000, 10,000, 20,000. The segmented small spaces are determined by Voronoi decomposition as follows:

(i) Let n_{G} be the number of samples of the subregion G, and h_{G} be a value of the ceiling function n_{G}/n_0, i.e., $\lceil n_{\mathrm{G}}/n_0 \rceil$ (G = A, B, C and D). Then, for each subregion, the sample is divided into h_{G} clusters using the k-mean method.

(ii) Voronoi decomposition is carried out using the centers of the clusters obtained in (i). The generated Voronoi partitions are denoted as segmented small spaces.

We compared our proposed method with the following two methods:

- GAM (generalized additive model): Here, f_1, f_2, f_3, f_4, and $g(x, y)$ are estimated with the 'mgcv' package in R (see Wood, 2023). In this package, $g(x, y)$ is estimated using the thin-plate spline technique (see e.g., Green & Silverman, 1994, chap. 7).

- PR (polynomial regression): Here, the basis functions are polynomial basis functions as used in Yamamura *et al.* (2016) and Solvang *et al.* (2017). With this method, we select the best degrees of the polynomials in f_1, f_2, f_3, f_4, and $g(x, y)$ from all the possible combinations, from the first to the sixth, using the Bayesian information criterion (BIC).

Table 1. MSE for sample points

Case	n	Proposed (n_0)						GAM	PR
		50	100	300	500	700	1,000		
1	5000	**0.218**	0.334	0.788	0.660	0.637	1.488	0.397	0.787
	10000	**0.143**	0.197	0.454	0.641	0.741	0.599	0.385	0.786
	20000	**0.098**	0.148	0.274	0.342	0.471	0.507	0.418	0.797
2	5000	**0.267**	0.385	0.810	0.913	0.924	1.216	0.405	0.863
	10000	**0.168**	0.242	0.321	0.606	0.815	0.956	0.415	0.856
	20000	**0.129**	0.173	0.318	0.338	0.521	0.567	0.437	0.870
3	5000	**0.085**	0.112	0.273	0.470	0.480	0.477	0.109	0.331
	10000	**0.055**	0.070	0.142	0.218	0.273	0.435	0.112	0.320
	20000	**0.038**	0.041	0.080	0.115	0.151	0.209	0.102	0.321

Tables 1 and 2 provide a summary of the MSE for the data points and the grid points. For the data points, in virtually all cases, our proposed method showed higher prediction accuracy

Table 2. MSE for grid points

Case	n	Proposed (n_0)						GAM	PR
		50	100	300	500	700	1,000		
1	5000	0.345	**0.334**	0.411	0.362	0.356	0.901	0.766	29.531
	10000	**0.205**	0.283	0.360	0.488	0.356	0.378	0.794	22.464
	20000	**0.208**	0.214	0.363	0.299	0.418	0.474	0.667	8.712
2	5000	**0.289**	0.296	0.569	0.868	0.901	1.532	0.605	26.583
	10000	0.227	0.216	**0.207**	0.484	0.556	0.996	0.725	17.894
	20000	**0.187**	0.207	0.256	0.163	0.313	0.499	1.146	6.699
3	5000	1.208	1.358	1.829	2.659	2.719	3.027	**1.156**	7.916
	10000	**0.887**	1.093	1.573	1.753	2.235	2.528	1.080	5.014
	20000	**0.661**	0.760	1.280	1.582	1.601	1.855	1.072	2.917

Table 3. Running time for sample points

Case	n	Proposed (n_0)						GAM	PR
		50	100	300	500	700	1,000		
1	5000	6.290	3.107	1.375	0.915	0.876	0.656	1.680	0.710
	10000	16.194	5.783	2.195	1.614	1.496	1.318	2.473	1.206
	20000	32.854	13.924	4.782	3.422	2.926	2.715	4.224	2.397
2	5000	6.671	3.415	1.419	0.992	0.972	0.646	1.683	0.710
	10000	13.238	6.637	2.305	1.723	1.534	1.346	2.481	1.205
	20000	35.904	13.868	4.676	3.522	2.914	2.732	4.211	2.393
3	5000	6.903	2.944	1.354	0.903	0.877	0.629	1.683	0.709
	10000	13.829	6.378	2.173	1.587	1.373	1.380	2.468	1.205
	20000	35.676	15.517	4.886	3.459	2.886	2.732	4.206	2.391

as well as better accuracy for finer division. Except for Case 3 with $n = 5,000$, the proposed method had a better prediction accuracy than either the GAM or PR. Among the three methods, the prediction accuracy of PR proved to be the worst. It should be noted that PR is not appropriate for sparse data points in the area since it produces a large curve with a sixth order polynomial even if the prediction accuracy for the data points is better. Table 3 summarizes the running time for the three methods when applied to the sample points. Since iterative calculation is not necessary for the estimation, applying polynomials is fastest. When $n_0 = 50$ and 100, the running time for the GAM appeared to be shorter than that for the proposed method. This is because when n_0 is small, the fusing of the elements of $\boldsymbol{\mu}$ by the GFL must be repeated more frequently. Although the proposed method with $n_0 = 300$ did not have the best prediction accuracy among the proposed methods, the running time was nearly the same as that for the GAM, and the MSEs for the sample and grid points in the proposed method with $n_0 = 300$ were tended to be slightly better than those of the GAM. Table 4 shows the following

Table 4. Running time for proposed method as percentage of 'mgcv' package running time

Case	n	n_0					
		50	100	300	500	700	1,000
1	5000	52.20	37.02	21.36	15.94	14.29	10.70
	10000	85.35	49.26	26.85	20.44	21.24	19.07
	20000	61.94	52.97	36.85	28.35	27.44	24.53
2	5000	49.30	40.79	21.88	14.80	16.50	11.51
	10000	69.24	51.67	26.19	20.11	21.76	17.23
	20000	67.21	50.88	34.86	31.09	26.67	23.89
3	5000	61.65	34.90	21.26	15.36	14.66	11.66
	10000	72.02	50.98	25.65	20.65	17.74	19.16
	20000	68.07	61.19	37.29	29.57	25.85	24.23

percentage for the sample points:

$$100 \times \frac{\text{running time with the proposed method}}{\text{running time when only } f_1, f_2, f_3 \text{ and } f_4 \text{ are estimated by the 'mgcv' package}} \ (\%).$$

This percentage indicates how much faster the smoothing based on the GRR is, relative to the time for smoothing using the 'mgcv' package. According to our results, the GRR-based smoothing reduces the running time dramatically, to as little as $1/10$ of the time for smoothing with the 'mgcv' package.

4. Real Data Example

By applying the proposed approach to the body condition data for the common minke whales obtained in Norwegian scientific and commercial whaling operations in the Northeast Atlantic during the months from April to September, we test its practicality and assess its real-world performance. Previous analyses of such body condition data collected from 1993 through 2013 have been conducted using the varying coefficients model (Solvang *et al.*, 2017) and the canonical correlation procedure for varying coefficient estimation (Yamamura *et al.*, 2016). For our study, we updated the data to include the years 2014-2017. The body condition metric typically takes into account the structural size, e.g., body length of the whale. Immediately after death, the whales are taken onboard and hauled across the foredeck of the boat. Total body length is measured in a straight line from the tip of the upper jaw to the apex of the tail fluke notch. Blubber thickness is measured at three sites (Figure 7): dorsally behind the blowhole (BT1) and behind the dorsal fin (BT2), and laterally just above the center of the flipper (BT3). Blubber measurements are made perpendicular from the skin surface to the muscle-connective tissue interface. Length measurements are recorded to the nearest centimeter, while blubber

measurements are taken to the nearest millimeter. For all captured whales, the year, month, day, and location (latitude and longitude) of capture are also recorded.

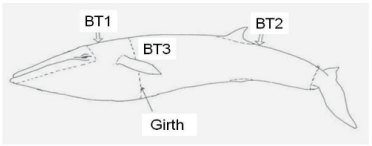

Figure 7. Measurement sites on whale body

Solvang *et al.* (2017) recognized that BT2 is difficult to measure consistently and potentially produces more measurement errors. The particular challenge is the large local variation in blubber thickness between the actual spot and close neighboring areas on the whale's body. Because of this, we excluded BT2 from our analysis. Furthermore, we excluded data from 1993-1996 which presented a different tendency over the entire period. After these exclusions, data for 11,509 whales were used in our analyses. BT1 served as the response variable y_{ij}. Table 5 shows summary statistics for length, BT1, latitude and longitude of the capture position, and calendar day. Calendar day is sequentially counted from the first day of April to the day that the measurement of caught whales is completed in September. For example, calendar days 74, 10, and 173 indicate that the measurements were completed on June 13th, April 10th, and September 20th, respectively. The explanatory variables $x_{1,ij}$, $x_{2,ij}$, $x_{3,ij}$ and $x_{4,ij}$ correspond to

Table 5. Data description

	Length (cm)	BT1 (mm)	Latitude	Longitude	Calendar day (April 1=1)
Mean	744.71	37.25	71.30	17.31	73.46
(S.D.)	97.00	9.91	5.54	8.98	24.03
Min.	350.00	2.00	56.50	-9.13	9.00
Max.	990.00	100.00	81.30	35.03	172.00

sex, year, calendar day, and length, i.e., $p = 4$ in (2.1). Since sex is a dummy variable (0 for male, 1 for female), we estimated it linearly. This means that $p_1 = 1$ in equation (2.2). As in the simulation study, the number of basis functions b_0 was set to 15. Based on the 5 subregions shown in Figure 1, segmented small spaces that included at least 300 individuals were defined. Finally, the number of segmented small spaces m was set to 37. Figure 8 shows the 5 subregions and the m-spaces that were segmented. In Figure 8, the red frames represent the 5 subregions, the blue frames represent the 37 segmented spaces, and the dots correspond to the

position of whaling. The $\hat{\boldsymbol{\mu}}$ for each space is estimated. If $\hat{\boldsymbol{\mu}}$ is the same as in the neighboring spaces, these spaces are fused together as one space.

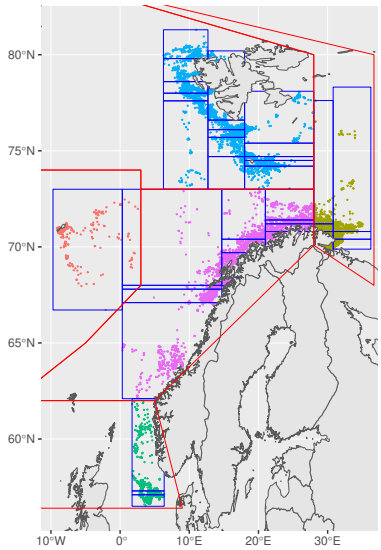

Figure 8. Segmented spaces

Table 6. Running time, estimated coefficient, and optimal values of hyperparameters

running time (sec)	iteration (times)	$\hat{\beta}_1$	$\hat{\lambda}$	df($\hat{\lambda}$)	$\hat{\gamma}$
2.39	8	3.55	4.72×10^{-3}	23	8

The estimation results are shown in Table 6 and Figures 9 and 10. Table 6 shows the running time to obtain the estimation result, the number of iterations until the backfitting algorithm converges, the estimated regression coefficient for sex, the optimal λ derived from (2.19), the number of final spaces after being fused (df($\hat{\lambda}$) in (2.19)), and the optimal γ derived from (2.16). Figures 9 and 10 show the estimation results for $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_m)'$, and for $f_\ell(x)$ (year, calendar day, and length) derived from (2.8), respectively. In Figure 10, the marker "|" on the horizontal axis of each figure represents the sample points. The estimates in the figure were centered by the weighted mean of the respective spatial effects, i.e., $\bar{\mu} = n^{-1} \sum_{j=1}^{m} n_j \hat{\mu}_j$. Both figures also include 95% simultaneous confidence intervals constructed by the model-based bootstrap method (see Appendix A.2 for details on how to compute simultaneous confidence intervals using this method). In Figure 9, the middle panel is the estimated spatial effect, the

left and right panels are the lower and upper bounds of the 95% simultaneous confidence interval for the spatial effect, and the dots correspond to the whaling position. In Figure 10, the right, middle, and left panels are the estimated $f_\ell(x)$ for year, calendar day, and length, respectively. The gray zones indicate the 95% simultaneous confidence intervals for $f_\ell(x)$. From the GFL estimation results, the number of subdivided areas has been reduced to 23 (from 37). The northern area around Svalbard showed the highest value (2.34) among the spaces, while the coastal area of mid-Norway had the lowest value (−3.24). Based on a previous study of length distributions in catches from the Northeastern Atlantic stock of minke whale (Øien, 1988), and focusing on positive estimates for the coastal areas around Svalbard and mid-Norway, we investigated the proportion of immature individuals whose length was less than 700 centimeters. Spatial effect estimates, counts of immature and total individuals, and the percentage of immature whales in the coastal areas of Svalbard and mid-Norway are summarized in Table 7. The areas showing the highest percentages are in mid-Norway, in the Vestfjorden area, as compared to the coastal areas of Svalbard, a trend that supports the outcome reported by Øien in 1988. Vestfjorden is known as an area where calves spend most of their summer (Jonsgård, 1951). The coastal areas from Trondheimsfjorden up to and including Vestfjorden are more suitable for immature whales, likely because the shallow waters provide a measure of protection and offer reasonable feeding conditions, as demonstrated by the higher (positive) spatial effects on BT1.

Table 7. Proportion of individuals with a length of less than 700 centimeters in coastal areas

Area	Positive estimates seen in coastal area	Number of individuals less than 700 *cm* in length / total number of individuals	Percentage of individuals less than 700 *cm* in length
Svalbard	2.13	42/268	16
	0.51	70/289	24
	0.76	58/232	25
	1.84	129/643	20
	2.34	60/303	19
Mid-Norway	1.28	82/312	26
	0.94	268/319	84
	0.61	189/338	56

In addition, the coastal areas of Svalbard are known to be important feeding grounds for females, where the minke whales are nourished and accumulate fat reserves not only in their blubber but also in their visceral fat during the summer (Solvang *et al.*, 2022). The effect at around 65 degrees North and the Jan Mayen area suggests that feeding on summering herring

may contribute to this (Solvang *et al.*, 2022).

The result for the effect of sex was 3.55, which is reasonable, as females are larger than males in general. For the year changes, BT1 decreased until 2015; however, the more recent tendency is an increase in BT1. The effect for length showed a sharp increasing trend up to 500 cm, followed by a linear increasing trend. This result is reasonable insofar as whales with greater body length tend to have greater blubber thickness. As for the confidence intervals, the width of the intervals for calendar day and length widened towards both ends of the data. The same tendency was observed for year, although it did not widen as rapidly as in the case of calendar day and length. This is likely due to the fact that calendar day and length were observed sparsely at the ends of the data, while year was observed relatively uniformly.

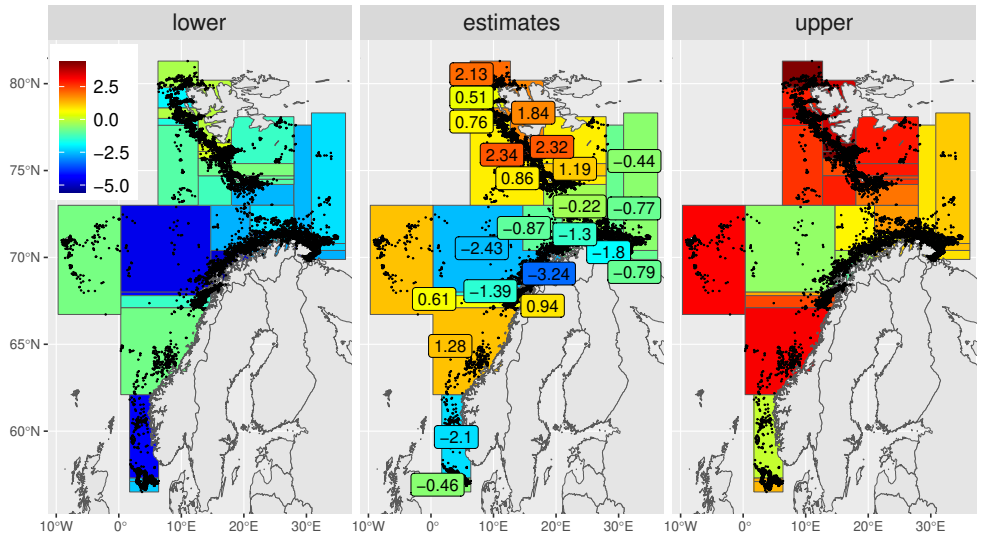

Figure 9. Estimated spatial effect on BT1

5. Conclusion and Discussion

To estimate the spatial effects in densely and sparsely sampled spatial data, we have proposed a method that alternately applies nonparametric regression and graph trend filtering via the GFL using a backfitting algorithm. In the GFL, the spatial data are segmented in advance into small spaces of approximately the same density. The segmented small spaces are fused by the GFL in order to reduce arbitrariness in the segmentation and to improve prediction accuracy. In addition, by dividing the spatial data into small spaces of similar density, a wide space

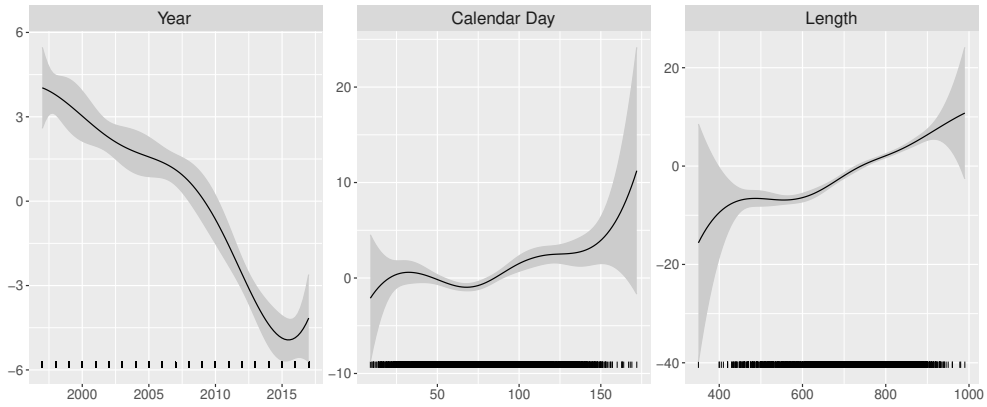

Figure 10. Estimated effects on BT1 by year, calendar day and length

is created where the data are sparsely scattered and a narrow space is created where the data is densely scattered. This method is suitable for trend estimation for nonuniformly sampled spatial data, since the wide space created by sparsely sampled data will lead to non-sensitive trend estimation. In our nonparametric regression, we applied penalized spline regression based on the GRR as proposed by Yanagihara (2012). In this method, the smoothing parameters that minimize the GCV are obtained in explicit form, which drastically reduces the time required for optimization. In addition, in an ordinary additive model, it is necessary to optimize the placement and number of basis functions for each explanatory variable; however, since the task of optimizing these values is quite tedious, it is common either to ignore it or to muddle through with the appropriate optimizations. In the proposed method, we avoid such tedious optimizations by optimizing the rank of the smoother matrix, which allows us to achieve high prediction accuracy by strictly performing the task, albeit a simplified version, rather ignoring the optimizations altogether or muddling through the process.

In a series of simulations, we compared the proposed method to two others: a method that uses the 'mgcv' package (GAM) and the method proposed in Yamamura *et al.* (2016) and Solvang *et al.* (2017), which estimates both f and the spatial effects using polynomials (PR). The comparisons focused on their respective MSEs for the sample points and grid points, as well as their relative running times. The PR approach proved to be very fast, as it can be applied to multiple regression with polynomial equations; however, the MSE at the grid points, which is the extrapolation prediction, was very poor because the spatial effects and the f values are estimated with polynomial equations. The proposed method was superior to the GAM in terms of the MSE for both the sample and grid points, but takes considerably longer than the GAM when the size of the segmented space is made too small. Even though it requires

substantial running time, it is clear that the estimation of effects other than the spatial effect by the penalized spline regression based on GRR reduces running time drastically. Since the key to improving prediction accuracy is how small the size of the segmented spaces can be made, further reduction of running time will be an important issue to be addressed in the future. The fact that the GAM running time is not longer can be attributed to the use of a thin-plate spline to estimate the spatial effects. The thin-plate spline places basis functions without bandwidths at all sample points, thus eliminating the need to optimize the placement, number, or bandwidth of the basis functions.

To show the performance of the proposed method when applied to real world data, we applied our method to data collected for the common minke whale in order to estimate the effects on blubber thickness (BT1) of sex, year of observation, month and day of observation (calendar day), total length of the whale, and capture location. The sex effect was a natural result, with females having a thicker BT1 than males. The effect of year showed a decreasing trend until 2015, but subsequently followed an increasing trend. The effect of calendar day could roughly be described as showing an increasing trend; however, the trend was downward in May but turned upward in June. The effect of length also showed an increasing trend in a rough sense, while a more detailed view showed a leveling off in the range of 500 to 600 *cm*. The effect of location was greater on the north side and lower on the south side. Similar tendencies were reported by Solvang *et al.* (2022). Although our confidence intervals might seem too wide compared to those obtained with the 'mgcv' package, it should be noted that the confidence intervals derived from the 'mgcv' package are point-wise values (for example, they correspond to a particular year), whereas our confidence intervals are simultaneous values, i.e., they include the estimated curve. It is thus natural that our confidence intervals are wider than those derived with the 'mgcv' package. Nevertheless, for calendar day and length, the intervals at the endpoints were too wide, no matter how small the number of data points. Finding a way to avoid fluctuations in the estimated trend at the endpoints represents another future challenge. Finally, the results derived from the proposed method (in particular, the predictors at non-sampled points) depended on which initial space segmentation is used. To produce a better predictor, it will be necessary to study more closely the influence of the initial segmentation.

The R code for the proposed method is available at `https://github.com/ohishim/amgfl`.

Acknowledgment

This research was supported by JSPS Bilateral Program Grant Number JPJSBP 120219927. The first, second, third and fourth authors' research was supported in part by JSPS KAKENHI Grant Number JP20H04151. The second author's research was supported in part by JSPS KAKENHI Grant Number JP23H00809, and the third author's research was supported in part

by JSPS KAKENHI Grant Number JP21K13834. The Radiation Effects Research Foundation (RERF), Hiroshima and Nagasaki, Japan is a public interest foundation funded by the Japanese Ministry of Health, Labor and Welfare (MHLW) and the US Department of Energy (DOE). This publication was supported by RERF. The views of the authors do not necessarily reflect those of the two governments.

References

Craven, P., & Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, **31**, 377–403.

Fukui, K., Ohishi, M., Yamamura, M., & Yanagihara, H. (2020). A fast optimization method for additive model via partial generalized ridge regression, *Intelligent Decision Technologies 2020, Proceedings of the 12th KES International Conference on Intelligent Decision Technologies* (*KES-IDT-20*) (eds. I. Czarnowski, R. J. Howlett & L. C. Jain), 279–290, Springer.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press.

Harville, D. A. (1997). *Matrix Algebra from a Statistician's Perspective*. Springer-Verlag, New York.

Hastie, T. (1996). Pseudosplines. *J. Roy. Stat. Soc. Ser*. **B**, **58**, 379–396.

Hastie, T., & Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall, London.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning*. Springer, New York.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Jonsgård, Å. (1951). Studies on the little piked whale or minke whale (*Balaenoptera acutorostrata* Lacépède). *Norsk Hvalfangstiid*, **40**, 209–232.

Green, P. J., & Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall/CRC.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, **15**, 661–675.

Mallows, C. L. (1995). More comments on C_p. *Technometrics*, **37**, 362–372.

Ohishi, M., Yanagihara, H., & Fujikoshi, Y. (2020). A fast algorithm for optimizing ridge parameters in a generalized ridge regression by minimizing a model selection criterion. *J. Stat. Plan. Infer.*, **204**, 187–205.

Ohishi, M., Fukui, K., Okamura, K., Itoh, Y., & Yanagihara, H. (2021). Coordinate optimization for generalized fused Lasso. *Comm. Statist. Theory Methods*, **50**, 5955–5973.

Øien, N. (1988). Length distributions in catches from the northeastern Atlantic stock of minke whales. *Rep. Int. Whal. Commn*., **38**, 289–295.

Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., & Schmid, M. (2019). A review of spline function procedures in R. *BMC Med. Res. Methodol*., **19**: 46, 1–16.

Politis, D. (2014). Bootstrap confidence intervals in nonparametric regression without an additive model. In *Topics in Nonparametric Statistics*: *Proceedings of the First Conference of the International Society for Nonparametric Statistics* (eds. M. G. Akritas, S. Lahiri & D. N. Politis), 271–282, Springer.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464.

Solvang, H. K., Skaug, H. J., & Øien, N. I. (2021). Abundance of common minke whales in the Northeast Atlantic based on survey data collected over the period 2014-2019. *IWC Scientific Committee 68C*, SC/68C/ASI/04.

Solvang, H. K., 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 Sci. Publ.*, **12**. https://doi.org/10.7557/3.6308

Solvang, H. K., Yanagihara, H., Øien, N., & Haug, T. (2017). Temporal and geographical variation in body condition of common minke whales (*Balaenoptera acutorostrata acutorostrata*) in the Northeast Atlantic. *Polar Bio.*, **40**, 667–683.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused Lasso.

J. Roy. Stat. Soc. Ser. **B**, **67**, 91–108.

Wand, M. P. (2000). A comparison of regression spline smoothing procedures. *Stat. Compt.*, **15**, 443–462.

Wang, Y. X., Sharpnack, J., Smola, A. J., & Tibshirani, R. J. (2016). Trend filtering on graphs. *J. Mach. Learn. Res*., **17**, 3651–3691.

Wood, S. (2023). Mgcv: mixed GAM computation vehicle with automatic smoothness estimation. R Package version 1.8-42. `https://cran.r-project.org/web/packages/mgcv/index.html`

Yamamura, M., Yanagihara, H., Solvang, H. K., Øien, N., & Haug, T. (2016). Canonical correlation analysis for geographical and chronological responses. *Procedia Comput. Sci.*, **96**, *Knowledge-Based and Intelligent Information & Engineering Systems*: *Proceedings of the 20th International Conference KES-2016* (eds. R. J. Howlett, L. C. Jain, B. Gabrys, C. Toro & C. P. Lim), 1096–1105.

Yanagihara, H. (2012). A non-iterative optimization method for smoothness in penalized spline regression. *Stat. Comput.*, **22**, 527–544.

Yanagihara, H. (2018). Explicit solution to the minimization problem of generalized cross-validation criterion for selecting ridge parameters in generalized ridge regression. *Hiroshima Math. J.*, **48**, 203–222.

Yanagihara, H., & Ohtaki, M. (2003). Knot-placement to avoid over fitting in *B*-spline scedastic smoothing. *Comm. Statist. Simulation Comput*., **32**, 771–785.

Yanagihara, H., Kamo, K., Imori, S., & Yamamura, M. (2017). A study on the bias-correction effect of the AIC for selecting variables in normal multivariate linear regression models under model misspecification. *REVSTAT-Stat. J*., **15**, 299–332.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Am. Stat. Assoc*., **101**, 1418-1429.

Appendix

A. Mathematical Details

A.1. GCV Criterion for Optimizing θ

According to Fukui *et al.* (2020), $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are estimated by minimizing the following penalized residual sum of squares (PRSS):

$$\mathrm{PRSS}_{\boldsymbol{\theta}}(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \hat{\boldsymbol{\mu}}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{B}\boldsymbol{\alpha} - \boldsymbol{R}\hat{\boldsymbol{\mu}}\|^2 + \boldsymbol{\alpha}'\boldsymbol{Q}\boldsymbol{\Theta}\boldsymbol{Q}'\boldsymbol{\alpha}, \tag{A.1}$$

where $\boldsymbol{\Theta}$ is a $b \times b$ diagonal matrix given by $\boldsymbol{\Theta} = \mathrm{diag}(\theta_1, \ldots, \theta_b)$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_b)'$ are non-negative multiple smoothing parameters. Here, $\hat{\boldsymbol{\mu}}$ is the estimate of $\boldsymbol{\mu}$ obtained in the one previous iteration consisting of (2.18) and (2.19) given by $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}(\hat{\lambda})$. Using the result from Fukui *et al.* (2020) after treating $\boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{\mu}}$ as a vector of new response variables in (A.1) yields the minimizers of (A.1) as

$$\bar{\boldsymbol{\alpha}}_{\boldsymbol{\theta}} = \boldsymbol{Q}(\boldsymbol{D} + \boldsymbol{\Theta})^{-1}\boldsymbol{D}^{1/2}\boldsymbol{z}, \quad \bar{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{\mu}} - \boldsymbol{B}\bar{\boldsymbol{\alpha}}_{\boldsymbol{\theta}}), \tag{A.2}$$

where $\boldsymbol{z}$ is given by (2.12). From elementary linear algebra, we have

$$\boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{\mu}} - \boldsymbol{B}\bar{\boldsymbol{\alpha}}_{\boldsymbol{\theta}} - \boldsymbol{X}\bar{\boldsymbol{\beta}} = \left\{\boldsymbol{I}_n - \boldsymbol{G} - \boldsymbol{P}\boldsymbol{D}(\boldsymbol{D} + \boldsymbol{\Theta})^{-1}\boldsymbol{P}'\right\}(\boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{\mu}}), \tag{A.3}$$

where $\boldsymbol{G} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ and $\boldsymbol{P}$ is given by (2.10). Therefore, the GCV criterion for optimizing $\boldsymbol{\theta}$ can be given by

$$\frac{(\boldsymbol{y}-\boldsymbol{R}\hat{\boldsymbol{\mu}})'\{\boldsymbol{I}_n-\boldsymbol{G}-\boldsymbol{P}\boldsymbol{D}(\boldsymbol{D}+\boldsymbol{\Theta})^{-1}\boldsymbol{P}'\}^2(\boldsymbol{y}-\boldsymbol{R}\hat{\boldsymbol{\mu}})}{[1-\{k+\mathrm{tr}(\boldsymbol{D}(\boldsymbol{D}+\boldsymbol{\Theta})^{-1})\}/n]^2}. \tag{A.4}$$

A.2. Simultaneous Confidence Intervals via Bootstrap

We here describe the computation of simultaneous confidence intervals for $f_\ell(x)$ ($\ell = p_1+1,\ldots,p$) and $\boldsymbol{\mu}$ using the model-based bootstrap method (see e.g., Politis, 2014). (These are calculated in lieu of confidence intervals for each point.) First, to express the ath bootstrap resample of $\boldsymbol{y}$ ($a = 1,\ldots,N_\mathrm{B}$), the following $n\times n$ random matrix is prepared, as in Yanagihara *et al.* (2017):

$$\boldsymbol{D}_a = (\boldsymbol{d}_{a,1},\ldots,\boldsymbol{d}_{a,n})', \quad \boldsymbol{d}_{a,1},\ldots,\boldsymbol{d}_{a,n} \sim i.i.d.\ MN_n(1, n^{-1}\boldsymbol{1}_n),$$

where $MN_n(1,n^{-1}\boldsymbol{1}_n)$ denotes the n-variate one-trial multinomial distribution with the same cell probabilities n^{-1}. Additionally, we assume that $\boldsymbol{D}_1,\ldots,\boldsymbol{D}_{N_\mathrm{B}}$ are mutually independent. The ath bootstrap resample of $\boldsymbol{y}$ can then be expressed as

$$\boldsymbol{y}_a^\star = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{B}\hat{\boldsymbol{\alpha}} + \boldsymbol{R}\hat{\boldsymbol{\mu}} + (\bar{y}-\bar{\mu})\boldsymbol{1}_n + \boldsymbol{D}_a\hat{\boldsymbol{e}}, \quad \hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{B}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\alpha}} - \boldsymbol{R}\hat{\boldsymbol{\mu}} - (\bar{y}-\bar{\mu})\boldsymbol{1}_n,$$

where $\bar{\mu}$ is a weighed mean of $\hat{\boldsymbol{\mu}}$, i.e., $\bar{\mu} = n^{-1}\boldsymbol{1}_n'\boldsymbol{R}\hat{\boldsymbol{\mu}} = n^{-1}\sum_{j=1}^m n_j\hat{\mu}_j$. Let $\hat{\boldsymbol{\beta}}_{a,\ell}^\star$, $\hat{\boldsymbol{\alpha}}_{a,\ell}^\star$ and $\hat{\mu}_{a,j}^\star$ be estimates of $\boldsymbol{\beta}_\ell$, $\boldsymbol{\alpha}_\ell$ and μ_j derived from the ath bootstrap resample $\boldsymbol{y}_a^\star$, and let $\hat{\boldsymbol{\xi}}_{a,\ell}^\star$ and $\boldsymbol{h}_\ell(x)$ be $(3+b_0)$-dimensional vectors defined as

$$\hat{\boldsymbol{\xi}}_{a,\ell}^\star = \left(\hat{\boldsymbol{\beta}}_{a,\ell}^{\star\prime}, \hat{\alpha}_{a,\ell}^{\star\prime}\right)', \quad \boldsymbol{h}_\ell(x) = (\boldsymbol{c}_\ell(x)', \boldsymbol{b}_\ell(x)')'.$$

We can now prepare means and variances

$$\bar{\boldsymbol{\xi}}_\ell^\star = \frac{1}{N_\mathrm{B}}\sum_{a=1}^{N_B}\hat{\boldsymbol{\xi}}_{a,\ell}^\star, \quad \boldsymbol{S}_\ell^\star = \frac{1}{N_\mathrm{B}}\sum_{a=1}^{N_B}\left(\hat{\boldsymbol{\xi}}_{a,\ell}^\star - \bar{\boldsymbol{\xi}}_\ell^\star\right)\left(\hat{\boldsymbol{\xi}}_{a,\ell}^\star - \bar{\boldsymbol{\xi}}_\ell^\star\right)',$$
$$\bar{\mu}_j^\star = \frac{1}{N_\mathrm{B}}\sum_{a=1}^{N_B}\hat{\mu}_{a,j}^\star, \quad s_j^\star = \frac{1}{N_B}\sum_{a=1}^{N_\mathrm{B}}\left(\hat{\mu}_{a,j}^\star - \bar{\mu}_j^\star\right)^2,$$

along with the following set

$$C_\ell = \left[\min_{i=1,\ldots,n_j;j=1,\ldots,m} x_{\ell,ij}, \max_{i=1,\ldots,n_j;j=1,\ldots,m} x_{\ell,ij},\right].$$

The maximum values can be calculated as

$$t_{a,\ell}^\star = \max_{x\in C_\ell}\frac{|\boldsymbol{h}_\ell(x)'(\boldsymbol{\xi}_{b,\ell}^\star - \bar{\boldsymbol{\xi}}_\ell^\star)|}{\sqrt{\boldsymbol{h}_\ell(x)'\boldsymbol{S}_\ell^\star\boldsymbol{h}_\ell(x)}}, \quad t_{a,0}^\star = \max_{j=1,\ldots,m}\frac{|\hat{\mu}_{a,j}^\star - \bar{\mu}_j^\star|}{\sqrt{s_j^\star}}.$$

The maximum value $t_{a,\ell}$ is obtained at, for example, the 100 points into which the C_ℓ interval is divided. Now let $z_\ell^\star$ and $z_0^\star$ be $(1-\alpha/2)$-quantiles of $t_{1,\ell}^\star, \ldots, t_{N_\mathrm{B},\ell}^\star$ and $t_{1,0}^\star, \ldots, t_{N_\mathrm{B},0}^\star$, respectively. Then, the $(1-\alpha)$-simultaneous confidence intervals for $f_\ell(x)$ and μ_j are given by

$$f_\ell(x) \in \left[\hat{f}_\ell(x) - z_\ell^\star \sqrt{\boldsymbol{h}_\ell(x)' \boldsymbol{S}_\ell^\star \boldsymbol{h}_\ell(x)},\ \hat{f}_\ell(x) + z_\ell^\star \sqrt{\boldsymbol{h}_\ell(x)' \boldsymbol{S}_\ell^\star \boldsymbol{h}_\ell(x)} \right]$$
$$(x \in C_\ell; \ell = p_1 + 1, \ldots, p),$$
$$\mu_j \in \left[\hat{\mu}_j - z_0^\star \sqrt{s_j^\star},\ \hat{\mu}_j + z_0^\star \sqrt{s_j^\star} \right] \ (j = 1, \ldots, m).$$