Biometrics 67, 1142–1152
September 2011
DOI: 10.1111/j.1541-0420.2010.01534.x

Exploring Spatial and Temporal Variations of Cadmium Concentrations in Pacific Oysters from British Columbia

Cindy Xin Feng,[1] Jiguo Cao,[1]*, and Leah Bendell[2]

[1]Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada

[2]Department of Biological Sciences, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada

email: jca76@sfu.ca

Summary. Oysters from the Pacific Northwest coast of British Columbia, Canada, contain high levels of cadmium, in some cases exceeding some international food safety guidelines. A primary goal of this article is the investigation of the spatial and temporal variation in cadmium concentrations for oysters sampled from coastal British Columbia. Such information is important so that recommendations can be made as to where and when oysters can be cultured such that accumulation of cadmium within these oysters is minimized. Some modern statistical methods are applied to achieve this goal, including monotone spline smoothing, functional principal component analysis, and semi-parametric additive modeling. Oyster growth rates are estimated as the first derivatives of the monotone smoothing growth curves. Some important patterns in cadmium accumulation by oysters are observed. For example, most inland regions tend to have a higher level of cadmium concentration than most coastal regions, so more caution needs to be taken for shellfish aquaculture practices occurring in the inland regions. The semi-parametric additive modeling shows that oyster cadmium concentration decreases with oyster length, and oysters sampled at 7 m have higher average cadmium concentration than those sampled at 1 m.

Key words: Cadmium concentration; Functional principal component analysis; Monotone smoothing; Semi-parametric additive model.

1. Introduction

Pacific oysters (*Crassostrea gigas*) are cultivated along the northwest coast of North America from Washington to Alaska and accumulate levels of cadmium that exceed some international tolerances. Health guidelines for the European Community set the tolerance of cadmium concentration at 6.3 μg Cd/g dry weight basis, and Asian export markets set the tolerance at 13.5 μg Cd/g dry weight basis (Kruzynski, 2004). In 1999, several shipments of oysters cultured in the province of British Columbia (BC), Canada, were rejected by the Hong Kong Food and Environmental Hygiene Department for exceeding the 13.5 μg Cd/g dry weight basis import limit. A subsequent shellfish survey by the Canadian Food Inspection Agency (CFIA) confirmed these shipments were not unusual and reported a mean cadmium value of 17.7 μg Cd/g dry weight basis for BC oysters cultured over the broad geographic area (Schallie, 2001). In 2000, Fisheries and Oceans Canada provided possible sources where cadmium might be originating. They concluded that the cadmium in BC oysters is mainly due to the geology of the area (Kruzynski, 2000), but the source of cadmium for these oysters is still uncertain. Consequently, this issue has recently been studied extensively.

A primary interest of our analysis is to study how oyster cadmium concentrations vary over space and time. A second objective is to investigate how cadmium concentrations depend on oyster growth over time. We illustrate how spline-smoothing techniques can be employed to address both of these concerns.

1.1 *The Motivating Data Sets*

In July 2000, in collaboration with the British Columbia Ministry of Agriculture Food and Fisheries (BCMAFF), Simon Fraser University initiated a grow-out study whereby Pacific oysters from the same seed source (Coast Seafoods, Washington State, USA) of the same age were deployed to existing oyster culture locations along the western coast of BC. The deployment dates of all oysters were the same within each site. Representative locations in both the east (mainland) and outer west (oceanic) were included. Deployment occurred along lines that were approximately 8 m long with seeded shells inserted at 30 cm intervals from the surface. Oysters were sampled approximately bimonthly between December 2002 (D2) to February 2004 (F4), from shallow (about 1 m depth) to deep (about 7 m depth) positions along the long-line. Oyster shell length (at the maximum length), was recorded in the field at time of sampling. Then, the sampled oysters were transported on ice to the laboratory, where they were killed and frozen whole until cadmium concentration analysis.

In this article, we consider thirteen sites identified in Figure 1. In the remainder of the article, the location of each site is denoted as the two-letter abbreviation of its name with the number in brackets indicating which region the site is from. The five southernmost sites located in the region of Barkley Sound (BS) on the westmost side of Vancouver Island are (1)Poett Nook (PN), (1)Useless Inlet-3 (BM), (1)Useless Inlet-4 (JF), (1)Useless Inlet-5 (PC), and (1)Webster Island/

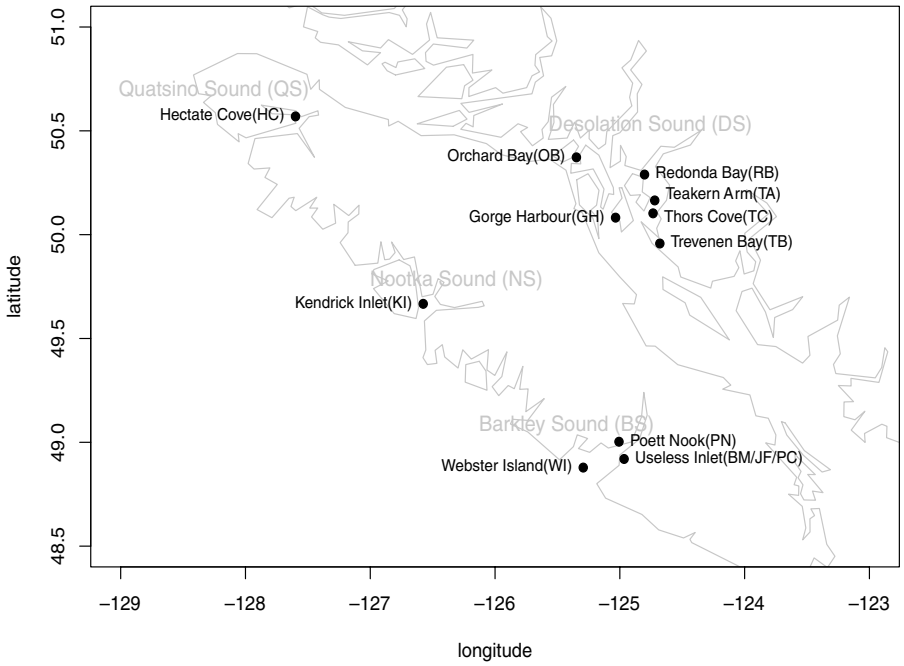

Figure 1. The geographical range of samples of cultured Pacific oysters along the west coast of British Columbia analyzed for oyster cadmium concentrations.

Effingham Inlet (WI). Moving northward, the site (2)Kendrick (KI) is in the region of Nootka Sound (NS) located to the north of Barkley Sound. Six sites from the region of Desolation Sound (DS), located on the west coast of the British Columbia mainland and on the east side of Vancouver Island include (3)Orchard Bay (OB), (3)Redonda Bay (RB), (3)Teakerne Arm (TA), (3)Gorge Harbour (GH), (3)Thors Cove (TC), and (3)Trevenen Bay (TB). The northernmost site considered is (4)Hecate Cove (HC) located in the region of Quatsino Sound (QS). Note that the region of Nootka Sound (NS) or the region of Quatsino Sound (QS) include only one site due to sampling difficulty in the specific geographical location.

1.2 *Statistical Methods*

In this study, we seek to explain the variation in observed cadmium concentrations and note that since measurements are taken in this study at unequally spaced time points, traditional techniques from time series analysis are difficult to implement. Bendell and Feng (2009) demonstrated the presence of certain temporal variation and clustering patterns for oyster cadmium concentrations contained in BC oysters via preliminary visual and simple statistical approaches. The average oyster cadmium concentrations and oyster length at the same depth at each site are measured at some discrete time points. Smoothing splines are used to estimate the average oyster cadmium concentration as a smooth nonparametric function at each site. The average oyster length will be nondecreasing over time, so a monotone smoothing technique (Ramsay, 1988) is used to estimate the average oyster length as a monotone function of time.

In addition to identifying the temporal trends in cadmium accumulation by oysters, we also sought to assess spatial influences on measured cadmium concentrations. Specifically, we aim to detect those regions where the oyster cadmium concentration over the entire deployment time is the highest, and hence provide advice to shellfish farmers to avoid these areas as possible farming sites. Classical multivariate principal components analysis (PCA) has often been adopted to identify these features. Here we employ functional alternatives, namely functional principal component analysis (FPCA), which incorporate the smoothing techniques into PCA (Deville, 1974; Besse and Ramsay, 1986; Hall, Muller, and Wang, 2006). Ramsay and Silverman (2005) give a nice introduction about FPCA.

Oyster cadmium concentration may also depend on the explanatory variables over time (i.e., oyster length and growth rate). Bendell and Feng (2009) showed that cadmium concentrations in oysters were linked to a number of factors, such as region, depth, growth rate, and oyster length by using a standard multivariate linear regression model. All of these factors played important roles in determining final tissue concentrations and ultimately the amounts of cadmium transferred to higher tropic levels. However, the linear relationship between the cadmium concentration and these covariates may not be a valid assumption. We relax this strict linear relationship constraint by using a semi-parametric additive model, which allows for flexible dependence structures (Ferraty and Vieu, 2000; Malfait and Ramsay, 2003; Chiou, Muller, and Wang, 2004; Antoniadis and Sapatinas, 2007; Crambes, Kneip, and Sarda, 2009).

Oysters' growth rate appears to be associated with temperature and food availability, which in turn might be linked to the amount of cadmium contained in oysters. Bendell and Feng (2009) attempted to look at the role of growth in influencing oyster cadmium concentrations by calculating the

growth rate as the change in oyster length divided by the total Julian days (the time interval between time 0 when the oysters were first deployed to the sampling date) to adjust for different deployment times across the sites. The main drawback of their method is that they considered global growth rates by dividing the change in length by the length of the time-period between deployment and sampling. Here, we present an alternative way to consider local (instantaneous) growth rates, which are calculated as the first derivatives of the monotone smoothing curves for the oyster lengths, alleviating the limitation of the previous approach.

The rest of the article is organized as follows. In Section 2, we provide an overview of some smoothing techniques used in this analysis, including penalized smoothing, monotone smoothing, FPCA, and semi-parametric additive modeling. These methods are then illustrated in Section 3 using our motivating oyster data where the capacity of the smoothing approaches to handle several features of this data set is demonstrated. Important results on spatial and temporal variation of cadmium concentrations are discussed. Concluding remarks are provided in Section 4.

2. Methodology

This section reviews the methods used in this application. The average measurements of oysters sampled at the same time points were modeled as a function of time at each site. The growth rates of oysters at each site are estimated as the first derivatives of the monotone smoothing spline estimator for oyster lengths. We also use FPCA for detecting the spatial variation of the average oyster cadmium concentrations. The effects of a number of factors on the oyster cadmium concentration functions are examined by semi-parametric additive modeling.

2.1 *Spline Smoothing*

Let y_{sgi} represent the measurement of cadmium concentration on the ith oyster sampled at the gth time point from site s, $s = 1, \ldots, N$, $g = 1, \ldots, G$, $i = 1, \ldots, n_{sg}$, where n_{sg} denotes the number of oysters sampled from site s at time g. Let $x_s(t)$ denote the mean cadmium concentration curve for each site, which is represented as a linear combination of basis functions

$$x_s(t) = \sum_{k=1}^{b+d+1} c_{sk}\,\phi_k(t) = \boldsymbol{\phi}(t)^T \mathbf{c}_s, \tag{1}$$

where $\mathbf{c}_s$ is the vector of B-spline coefficients c_{sk}, $k = 1, \ldots, K$, corresponding to the kth spline effect at site s, $\boldsymbol{\phi}(t)$ is the vector of cubic B-spline basis function $\phi_k(t)$, b is the number of break-points or knots, and d is the degree of the polynomial within each segment—cubic splines ($d = 3$) are often used. There are many equivalent bases for the spline space but the most popular is the so-called B-spline basis due to its numerical stability and computational efficiency (de Boor, 1978). To implement spline smoothing, the basis coefficient vector $\mathbf{c}_s$ needs to be estimated, for example, using least squares. The fitted curve is determined by the number and location of the knots. Ramsay (1988) and Zhou and Shen (2001) discuss how to choose the number and location of the knots. The drawback of the dependence of splines on suitable knot placement has been discussed in the literature (Hastie, Tibshirani, and Friedman, 2001; Durban et al., 2005). Our study uses the penalized fitting strategy to alleviate the importance of knot locations (Wood, 2000) by putting one knot at each distinct time point with measurements, and a roughness penalty term is used to control the smoothness of the fitted function. This eliminates the need to choose knot locations and makes estimated curves more stable at the cost of some increase in bias. The basis coefficient vector $\mathbf{c}_s$ is estimated by minimizing the penalized sum of squared error (PENSSE) loss function,

$$\begin{aligned}\text{PENSSE}(\mathbf{c}_s) = &\sum_{g=1}^{G}\sum_{i=1}^{n_{sg}}\{y_{sgi} - x_s(t_g)\}^2 \\ &+ \lambda_s \int_{t_1}^{t_G} \left\{\frac{d^2}{dt^2} x_s(t)\right\}^2 dt,\end{aligned} \tag{2}$$

where t_g represent the actual time at the gth time point. The second term penalizes the roughness of the fitted function. The smoothing parameter λ_s for site s determines the trade-off between the fit of the data and the smoothness of the fitted function. Ramsay and Dalzell (1991) suggest λ_s can often be chosen by inspection of the curve smoothness or through an automated procedure such as generalized cross-validation (GCV; Craven and Wahba, 1979).

Taking the derivative of (2) with respect to the parameter vector $\mathbf{c}_s$ and solving for $\mathbf{c}_s$ yields

$$\begin{aligned}\widehat{\mathbf{c}}_s = &\left[\sum_{g=1}^{G} n_{sg}\left\{\boldsymbol{\phi}(t_g)\boldsymbol{\phi}(t_g)^T\right\} + \lambda_s \int_{t_1}^{t_G} \frac{d^2}{dt^2}\boldsymbol{\phi}(t)\frac{d^2}{dt^2}\boldsymbol{\phi}(t)^T dt\right]^{-1} \\ &\times \left\{\sum_{g=1}^{G}\sum_{i=1}^{n_{sg}} y_{sgi}\boldsymbol{\phi}(t_g)\right\}.\end{aligned}$$

The estimate for the smooth function is then

$$\hat{x}_s(t) = \boldsymbol{\phi}(t)^T \widehat{\mathbf{c}}_s\,. \tag{3}$$

2.2 *Monotone Spline Smoothing*

In principle, the average oyster length should not decrease over time, even when the noise inherent to any data set may suggest otherwise. To account for this, we employ a monotone smoothing technique (Ramsay, 1988) to model oyster length over time. A strictly monotone smooth function has a strictly positive first derivative. Let $l_s(t)$ represent the average oyster length function at site s. The growth rate $dl_s(t)/dt$ must be positive, so it is expressed as the exponential of an unconstrained function $W_s(t)$: $dl_s(t)/dt = \exp[W_s(t)]$. By integrating both sides of this equation, $l_s(t)$ can be written as

$$l_s(t) = \beta_{s0} + \int_{t_1}^{t} e^{W_s(u)} du\,.$$

By using this representation, $W_s(u)$ can be flexibly estimated as a linear combination of basis functions, $W_s(u) = \sum_k c_{sk}\phi_k(u)$, defined similarly to (1). Here, we need to estimate β_{s0} and c_{s1}, …, c_{sK}. We estimate these parameters by minimizing,

$$\begin{aligned}&\text{PENSSE}(\beta_{s0}, c_{s1}, \ldots, c_{sK})\\ &= \sum_{g=1}^{G}\sum_{i=1}^{n_{sg}}\left\{\ell_{sgi} - \beta_{s0} - \int_{t_1}^{t_i} e^{W_s(t)}dt\right\}^2\\ &\quad + \lambda_s \int_{t_1}^{t_G}\left\{\frac{d^2W_s(t)}{dt^2}\right\}^2 dt\,,\end{aligned}$$

where ℓ_{sgi} represent the length for the ith oyster sampled from site s at the gth time point. In this situation, we cannot obtain closed forms for the estimates of β_{s0} and the spline coefficients c_{s1}, …, c_{sK}. The Newton–Raphson iteration method is used to obtain the coefficient estimates. It is easily implemented and converges quickly. To avoid converging to local minima, one can try different starting values for basis coefficients.

2.3 *Functional Principal Component Analysis*

Here we outline the statistical methodology of FPCA, which we use in the following section to examine the oyster data set. FPCA is a multivariate technique that can partition variability among the measurements into components of decreasing "importance." In this application, we treat the distribution of mean curves for the cadmium concentration, defined in (3) as the "response." We subtract the mean curve $\bar{x}(t) = \sum_{s=1}^{N}\hat{x}_s(t)/N$ from each curve and use $\hat{z}_s(t) = \hat{x}_s(t) - \bar{x}(t)$, to implement FPCA, as our interest is primarily in characterizing the deviations of the $\hat{x}_s(t)$ from the average curve. The first functional principal component weight function $\xi_1(t)$ is estimated by maximizing sum of squared functional principal component (FPC) scores $\sum_s f_{s1}^2$, where

$$f_{s1} = \int_{t_1}^{t_G}\xi_1(t)\hat{z}_s(t)\,dt, \quad s = 1, \ldots, N\,,$$

subject to

$$\|\xi_1\|^2 = \int_{t_1}^{t_G}\xi_1^2(t)\,dt = 1. \tag{4}$$

The second functional principal component weight function $\xi_2(t)$ is estimated by maximizing sum squared FPC scores, subject to the constraint $\|\xi_2\|^2 = 1$ and the additional constraint

$$\int_{t_1}^{t_G}\xi_1(t)\xi_2(t)\,dt = 0. \tag{5}$$

Other functional principal component weight functions can be estimated in the same way.

Searching for the mutually orthonormal and normalized weight functions is equivalent to the problem of eigenanalysis of the variance–covariance function or operator, defined by

$$v(t,t') = N^{-1}\sum_{s=1}^{N}\hat{z}_s(t)\hat{z}_s(t'),$$

then any eigenfunction $\xi_p(t)$, $p = 1$, …, P, satisfies the functional eigenequation

$$\int_{t_1}^{t_G} v(t,t')\xi_p(t')dt' = \rho_p\xi_p(t),$$

for an appropriate eigenvalue ρ_p. The proportion of each eigenfunction $\xi_p(t)$ taking account of total variation among N curves is calculated as $\rho_p/\sum_{p=1}^{P}\rho_p$. In practice, the first P_L eigenfunctions are chosen such that $\sum_{p=1}^{P_L}\rho_p/\sum_{p=1}^{P}\rho_p$ is greater than some threshold, because they account for most of the total variation. The first two components in the oyster data set accounted for much of the variation, providing enough information regarding the principal sources of variation between mean concentration curves.

To control the smoothness of eigenfunctions, Ramsay and Silverman (2005) introduce a smoothed PCA approach. The first eigenfunction $\xi_1(t)$ is estimated by maximizing the penalized sample variance

$$\text{PCAPSV}(\xi(t)) = \frac{\text{var}\int_{t_1}^{t_G}\xi(t)\hat{z}_s(t)\,dt}{\|\xi_1\|^2 + \lambda\int_{t_1}^{t_G}\left\{\frac{d^2\xi(t)}{dt^2}\right\}^2}\,dt,$$

subject to $\|\xi_1\|^2 = 1$. The smoothing parameter λ controls the trade-off between the maximization of the sample variance and the roughness of the first eigenfunction. Each subsequent eigenfunction, $\xi_j(t)$, $j = 2, 3, \ldots$, is estimated by maximizing the penalized variance PCAPSV (ξ(t)) subject to two constraints $\|\xi_j\|^2 = 1$ and the modified form of orthogonality

$$\int_{t_1}^{t_G}\xi_j(t)\xi_k(t)\,dt + \int_{t_1}^{t_G}\left\{\frac{d^2\xi_j(t)}{dt^2}\right\}\left\{\frac{d^2\xi_k(t)}{dt^2}\right\}dt = 0 \quad \text{for } k = 1, \ldots, j-1.$$

Ramsay and Silverman (2005) explain in detail how to find these eigenfunctions by solving a single eigenvalue problem in Section 9.4 of their book. Silverman (1996) shows the theoretical advantages of this approach.

2.4 *Semi-Parametric Additive Model*

Oyster cadmium concentration trends for the sampling sites can be explained by variables such as the depth, region, oyster length, and oyster growth rate. A semi-parametric additive model is proposed to investigate which regions tend to have greater cadmium concentrations, and which sizes of oysters tend to have high concentrations within any region, after adjusting for depth and growth rate effects.

Let y_{sgi} denote the measurement of cadmium concentration on the ith oyster sampled at the gth time point from site s, $s = 1, \ldots, N$, $g = 1, \ldots, G$, $i = 1, \ldots, n_{sg}$, where n_{sg} denotes the number of oysters sampled from site s at time g. We investigate the semi-parametric additive model:

$$\begin{aligned}M_1 : \log(y_{sgi}) &= \alpha + s_0(t_g) + s_1(l_{si}(t_g)) + s_2(r_{si}(t_g))\\ &\quad + \beta_d I(\text{depth}_{si} = 7m)\\ &\quad + \sum_{h=1}^{H}\beta_h I(\text{region}_{si} = h) + \epsilon_{si}(t_g),\end{aligned}$$

where α represents the overall mean, t_g is the actual time at the gth time point, the nonparametric smooth function $s_0(\cdot)$ represents the overall mean trend. The growth rate $r_{si}(t)$ is estimated as the first derivative of the monotone smoothing spline estimator of oyster lengths $l_{si}(t)$, $s_1(\cdot)$ and $s_2(\cdot)$ are nonparametric smooth functions of observed oyster length

and estimated oyster growth rate, respectively. Here $s_0(\cdot)$, $s_1(\cdot)$ and $s_2(\cdot)$ are not constrained to be of any pre-specified parametric form. Instead, we model these terms as linear combinations of cubic B-splines: $s_k(\cdot) = \sum_{j=1}^{p_k} c_{kj}\phi_{kj}(\cdot)$, $k = 0, 1$ and 2, where c_{kj} are coefficients of the smooth. The linear coefficients, β_d and β_h, are discrete effects for the depth and region in our study where the effects of being at the depth 1 m and region BS are set as baseline as these are set to be the reference levels for these factor variables. We use $\epsilon_{si}(t)$ to denote independent errors with mean zero and common variance.

To avoid overfitting, M_1 is estimated by penalized maximum likelihood estimation (Hastie and Tibshirani, 1990). The semi-parametric additive model is estimated by minimizing the PENSSE loss function:

$$\begin{aligned}\text{PENSSE} = \sum_{s,i,g}\Bigg[\log(y_{sgi}) - \Bigg\{\alpha + s_0(t_g) + s_1(l_{si}(t_g)) \\ + s_2(r_{si}(t_g)) + \beta_d I(\text{depth}_{si} = 7m) \\ + \sum_{h=1}^{H}\beta_h I(\text{region}_{si} = h)\Bigg\}\Bigg]^2 \\ + \lambda_0\int\left\{\frac{d^2 s_0(t)}{dt^2}\right\}^2 dt + \lambda_1\int\left\{\frac{d^2 s_1(l)}{dl^2}\right\}^2 dl \\ + \lambda_2\int\left\{\frac{d^2 s_2(r)}{dr^2}\right\}^2 dr\,,\end{aligned}$$

where the smoothing parameters λ_0, λ_1, λ_2, determine the amount of smoothing for each of the smooth terms. The smoothing parameters are estimated with a computationally efficient method by applying GCV in generalized ridge regression problems (Wood, 2004). The above model is fitted using an `R` package `mgcv` (Wood, 2004).

Note that although oyster growth rate is estimated as the first derivative of the monotone smoothing spline estimator of the oyster length, it has little correlation with the oyster length as correlation coefficient equals to -0.17 (scatter plot is shown in Web Figure S4). It is not rare for the first derivative of a variable to be independent from that variable itself. For example, although the velocity of a moving car is the first derivative of the position function, the velocity is independent from the position of the car.

Bendell and Feng (2009) estimate linear effects of oyster growth rate and oyster length and effects of depth and region using a standard multiple linear regression model

$$\begin{aligned}M_2 : \log(y_{sgi}) = \alpha + \beta_l l_{si}(t_g) + \beta_r r_{si}(t_g) + \beta_d I(\text{depth}_{si} = 7m) \\ + \sum_{h=1}^{H}\beta_h I(\text{region}_{si} = h) + \epsilon_{si}(t_g)\end{aligned}$$

where β_l and β_r are linear coefficients for oyster length and oyster growth rate, respectively.

To compare the two models, we employ the Akaike information criterion (AIC; Akaike, 1974), which penalizes the complexity of the model for using a large number of parameters. The standard multiple linear regression model M_2 is less complex than the semi-parametric additive model M_1 and easier to interpret. However, M_1 is appealing in terms of flexibility in the trends for the covariate effects and avoids restricting the trend to a linear form.

3. Results

3.1 *Data Representation*

In this subsection, all plots are provided in the Web Appendix to save space. Web Figure S1 displays smoothed functions of average oyster cadmium concentrations over time by penalized spline smoothing at each of the experimental sites. During the initial sampling time in winter 2002 and 2003, the oysters appear to exhibit high cadmium concentrations, which decrease over the summer of 2003, and subsequently increase towards winter 2003, though the patterns vary from site to site. Also, the oyster cadmium concentrations tend to be higher for oysters sampled from 7 m depth than those sampled at 1 m.

The curves for the average oyster length are estimated using monotone spline smoothing, as the average oyster length should not decrease over time. As an illustration, Figure 2 displays the fitted curves of the lengths and growth rates for oysters sampled from site WI at 1 m and 7 m, respectively. The fitted curves for all the experimental sites are displayed in Web Figures S2 and S3. Web Figure S2 shows that the oysters sampled at 7 m tend to be smaller than those sampled at 1 m possibly due to food availability at different depths. Web Figure S3 shows that the trends and variations of the growth rate are not aligned across the sites. Note that the estimated growth rate at the first measurement time from site PN at 1 m is beyond 15, and it is much higher than the other estimated growth rates, which are approximately less than 8. This may be caused by the boundary effect of spline smoothing, which often yields unreliable estimates at the boundaries. Therefore, this extreme estimate is removed when exploring the effect of growth rate on cadmium concentration in the semi-parametric additive model M1 and the multiple linear regression model M2.

3.2 *Spatial Variability*

To visualize FPCA results, we examine plots of the overall mean function and the functions obtained by adding and subtracting a suitable multiple of the eigenfunctions (Silverman, 1995). The multiple is $\kappa\sqrt{\rho_p}$, where κ represents a correcting factor to adjust the magnitude of the effect of $\xi_p(t)$ with respect to the square root of the eigenvalue ρ_p. The correcting factor κ can be set to be any value subjectively to adjust the magnitude of the effect of the eigenfunction $\xi_p(t)$ with respect to the square root of the eigenvalue ρ_p. Here, we choose $\kappa = 0.2$ subjectively to produce a clear visual impression of the effect of principal components on the overall mean function.

Figure 3 displays the overall mean curve and the effect of adding $(+)$ and subtracting $(-)$ a multiple of each of the first two weighting functions for 1 m and 7 m, respectively. The first FPC displayed in the upper left panel of Figure 3 accounts for about 89% of the variation of the average cadmium concentration for the oysters sampled at 1 m among the thirteen experimental sites. The effect of the first eigenfunction is approximately to add or subtract a constant to cadmium concentration throughout the time period. This indicates that about 89% of variability between sites is accounted for by

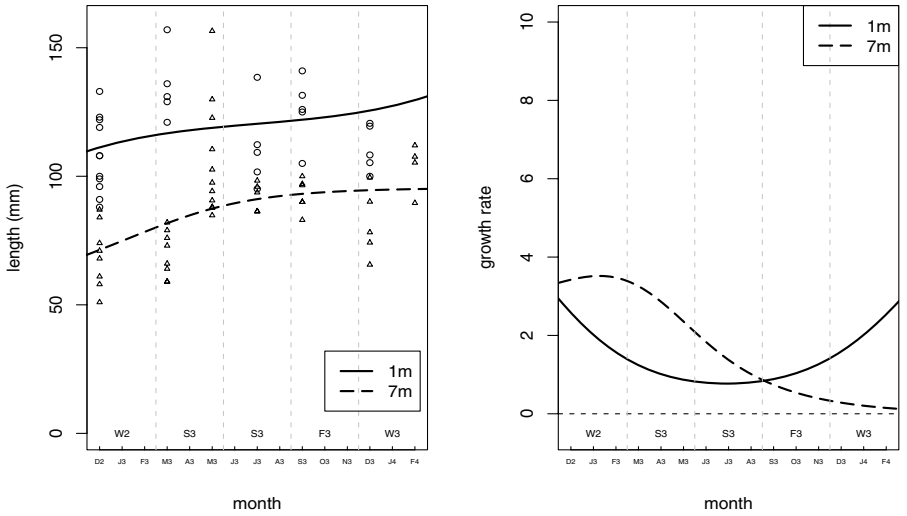

Figure 2. The left panel shows the monotone spline smoothing curves of oyster lengths for oysters sampled from site WI at 1 m (solid curves) and 7 m (dashed curves), respectively. The circles and triangles are the measured oyster lengths sampled at 1 m and 7 m depth, respectively. The right panel shows the estimated growth rates for the oysters sampled from site WI at 1 m (solid curves) and 7 m (dashed curves), respectively. The growth rates are estimated as the first derivatives of the monotone smoothing functions of the oyster lengths. The labels above the x-axis represent the seasons, ranging from winter 2002 (W2) to winter 2003 (W3). The labels below the x-axis represent the months, ranging from December 2002 (D2) to February 2004 (F4).

the average cadmium concentration differences. The second eigenfunction explains about 10% of the variation after accounting for the variability of the first eigenfunction, indicating that about 10% of the variation among sites is the change of cadmium concentration from winter 2002 and winter 2003. Similar patterns are observed for the variation of the cadmium concentration for the oysters sampled at 7 m, except that the second eigenfunction represents the change of cadmium concentration after August 2003.

One of the important features of FPCA is the ability to examine the scores of each curve on each eigenfunction (Ramsay and Dalzell, 1996; Ramsay and Silverman, 2002). The bottom two panels in Figure 4 displays the first FPC score against the second FPC score for 1 m and 7 m, respectively. There appears to be some regional groupings, although there is some overlap between Barkley Sound, Nootka Sound, and Quatsino Sound. One referee suggests to use a hierarchical clustering tree for group classification by considering the $n \times 2$ matrix with row entries taken as the principal component scores associated with eigenvalues ρ_1 and ρ_2. The top two panels in Figure 4 show the clustering tree for the samples at 1 m and 7 m, respectively. For samples at 1 m, three groups are obtained by cutting the tree at height 25: group 1=TB, OB, GH and TC, which score highly on the first PC; group 2= RB, TA, JF, BM and WI, which score moderately on the first PC; group 3=PC, HC, PN and KI, which score low on the first PC. For samples at 7 m, three groups are obtained by cutting the tree at height 25: group 1=OB, WI, TA and TC, which score highly on the first PC; group 2= GH, TB, BM and JF, which score moderately on the first PC; group 3=PN, KI, PC, RB and HC, which score low on the first PC. Therefore, cadmium concentrations for the oysters sampled from those inland sites may have higher cadmium concentration on average than the coastal sites at 1 m depth. In fact, the form of the groups is completely guided by the first principal component coordinates because this first axis accounts for about 90% of the entire variability. This axis can then serve as a pollution index because it sorts the observations by mean cadmium concentration. Similar grouping patterns are also found for the oysters sampled at 7 m depth, except for site WI, exhibiting high cadmium concentration on average. It is interesting to note that the oysters sampled from WI are more influenced by oceanic processes rather than direct anthropogenic influences. Possible sources at this one site could also be related to forestry practices within this region, e.g., forest canopy removal with resulting erosion of soils naturally high in cadmium. Cadmium contaminated fertilizer applied during reforestation could also contribute to observed oyster cadmium concentrations at this site.

3.3 *The Semi-Parametric Additive Model*

The semi-parametric additive model M_1 and the standard multiple linear regression model M_2 are compared in terms of AIC. The AIC is 856.68 for M_1 and 936.41 for M_2. A model with a lower AIC score is preferred as it achieves a more optimal combination of fit and parsimony. As a result, AIC favors M_1 over M_2. In comparison to M_2, M_1 used nonparametric functions of length and growth rate to explain the variation left after accounting for the effects of depth and region.

Figure 5 displays the estimated model terms with 95% confidence intervals. The top left panel shows that the oyster cadmium concentration averaged over thirteen sites has the lowest value in summer 2003 and relatively higher values in winter 2002 and 2003. A longer series of data would be needed to test if this pattern is consistent over years. The other two top panels show that the oyster length and growth rate tend to

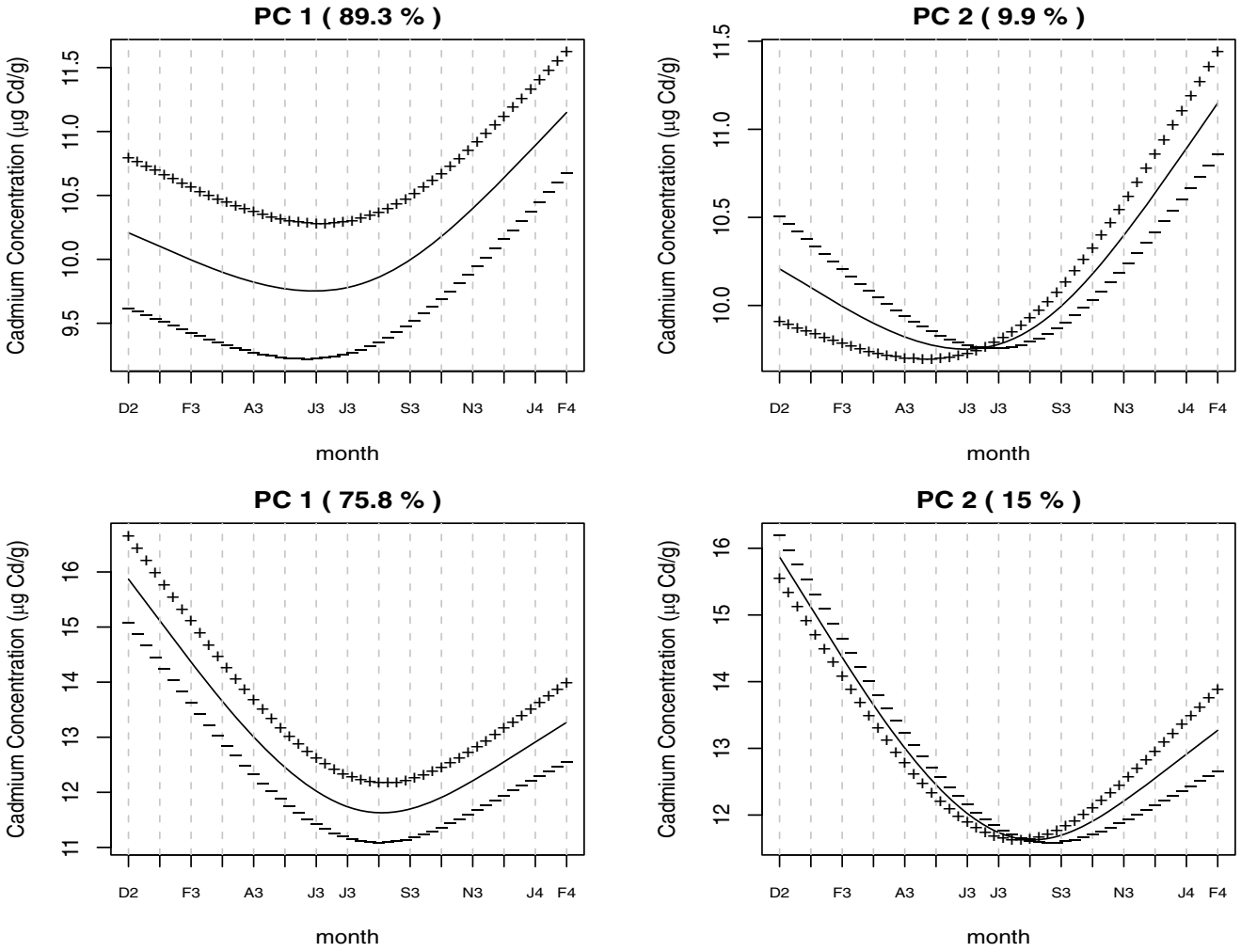

Figure 3. The top two panels and the bottom two panels display the mean oyster cadmium concentration curve and the effects of adding (+) and subtracting (−) a small multiple of each eigenfunctions for oysters sampled at 1 m and 7 m depth, respectively. The percentages indicate the amount of total spatial variation accounted by functional principal components. The labels of the x-axis represent the months, ranging from December 2002 (D2) to February 2004 (F4).

have nonlinear relationships with oyster cadmium concentration. The average cadmium concentration decreases with the oyster length, indicating that smaller oysters have higher cadmium concentration than their longer counterparts. The third panel shows that the partial growth rate effect appears to decrease with the growth rate up to about 2.5 mm per month and then becomes statistically nonsignificant. Note that the confidence interval for the partial effect of growth rate gets wider as growth rate gets larger, as there are fewer oysters with higher values of growth rate.

Figure 5 also displays the comparison for oyster cadmium concentrations between 1 m and 7 m and among all the regions after adjusting for the smooth terms of length and growth rate effects in model M_1. On average, oyster cadmium concentrations are higher for the oysters sampled at 7 m than those from 1 m and higher for those from region DS than those from the other regions with regions BS and QS having significantly lower oyster cadmium concentrations than the other locations, confirming the results of functional PCA.

The results in Table 1 for model M_1 indicate that the functional effects of oyster length and growth rate on the cadmium concentration are significant and the average oyster cadmium concentration at a lower depth of 7 m is significantly higher than that at depth of 1 m by about 0.23 μg Cd/g. Also, the average oyster cadmium concentration in oysters from region DS is significantly higher than the other regions by about 0.47 μg Cd/g. It is worthwhile noting that the model terms that are common to both models have remarkably similar coefficients, since depth and region are independent from oyster length and oyster growth rate. To be more precise, the nonparametric functions of length and growth rate explain the remaining variation of cadmium concentration conditional on the depth and region variables.

4. Concluding Remarks

In this article, we investigate the nature of the spatial and temporal distributions of oyster cadmium concentration within the regions of our study. We illustrate some

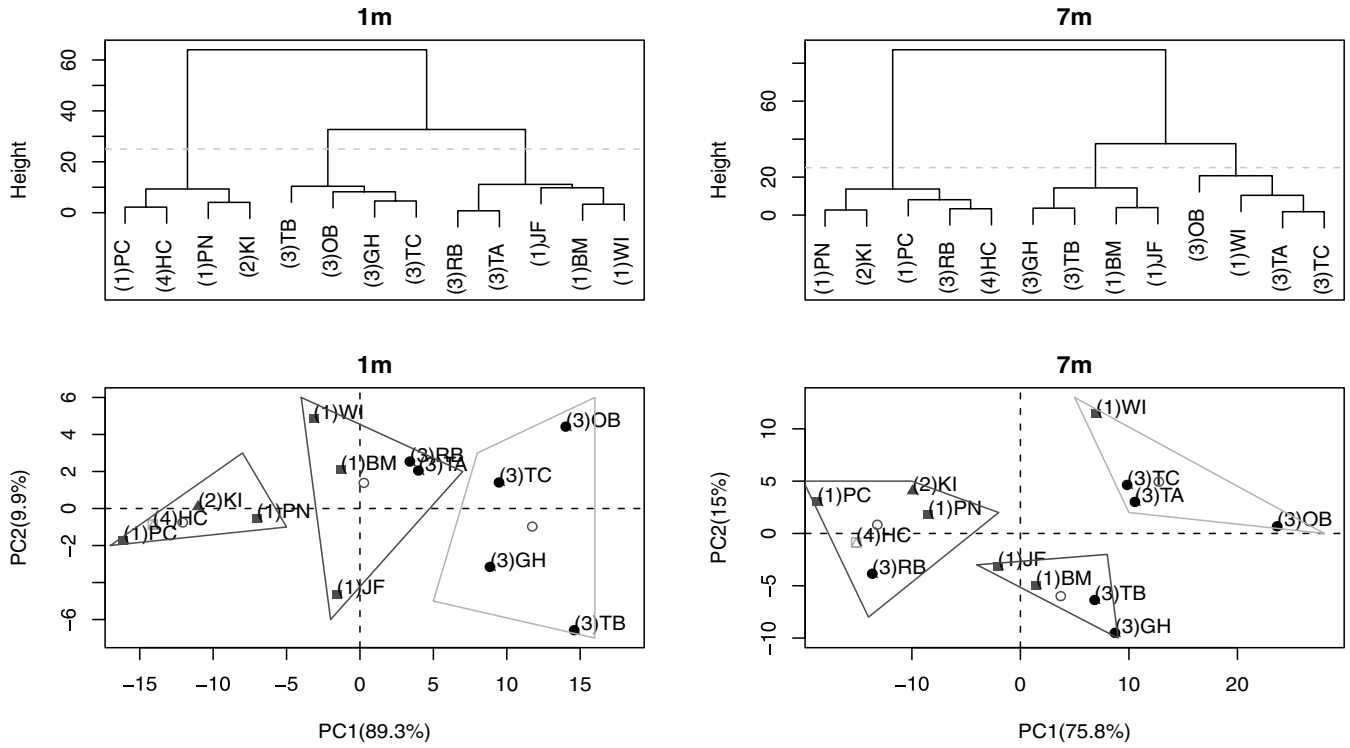

Figure 4. The upper two panels show the hierarchical clustering trees for the $n \times 2$ matrix with row entries taken as the principal component scores associated to eigenvalues ρ_1 and ρ_2 for 1 m and 7 m depth, respectively. The bottom two panels show the first two functional principal component scores at 1 m and 7 m depth, respectively. The location of each site is shown by the two-letter abbreviation of its name with the number in the bracket indicating which region the site is from. The sites from Barkley Sound are symbolized as the solid squares; the site from Nootka Sound is symbolized as the solid triangle; the sites from Desolation Sound are symbolized as the solid circles; and the site from Quatsino Sound is symbolized as the square with triangle inside. The convex hulls are added to the clusters with the cluster centroid symbolized as the open circle for each of the clusters, provided the trees are cut at height 25.

statistical methodologies to provide a route for statistical analysis directed at enhancing biological insight. Those methodologies can readily be applied to a wide variety of marine ecological data characterized by being irregularly spaced and noisy, while allowing spatial clustering and potentially interesting and important factors to be identified.

To handle missing and irregularly spaced temporal measurements, we have investigated the use of penalized spline smoothing technique to estimate the mean curve of oyster cadmium concentration. We also adopt the monotone spline smoothing method to impose nondecreasing constraints on oyster length curve estimation. Oyster growth rate is characterized as the first derivative of the estimated curve for oyster length. To the best of our knowledge, few attempts have been made so far in the marine ecological literature to impose shape restrictions on the growth curve. The prime advantage of using these spline techniques is to relax the parametric assumption on the curve shapes commonly seen in ecological and biological literature.

The functional PCA technique is investigated to identify the spatial variation of the average oyster cadmium concentration over the experimental sites, which provides a good indication of which sites are similar and might assist future allocation of sampling efforts. There appears to be some regional grouping, although there is some overlap between Barkley Sound, Nootka Sound, and Quatsino Sound. Possible cadmium sources (Kruzynski, 2000, 2004) from different regions are quite different, though. For the oysters sampled from sites located in Desolation Sound, given their close proximity to terrestrial influences, possible cadmium sources could include cadmium contaminated phosphate fertilizers and local septic tanks. Therefore, the spatial clustering pattern suggests an upland continental source versus a marine source in the coastal area. The oysters sampled from Webster Island, however, are more influenced by oceanic processes rather than direct anthropogenic influences. Possible cadmium sources at this site may also be related to forestry practices within this region, e.g., forest canopy removal with resulting erosion of soils naturally high in cadmium. Cadmium contaminated fertilizer applied during reforestation may also contribute to observed oyster cadmium concentrations at this site. Further investigation with more sampling sites and longer time duration of experiments are needed to test our hypothesis.

In this study, we have seen that the semi-parametric additive model ensures a better fit than the standard multiple

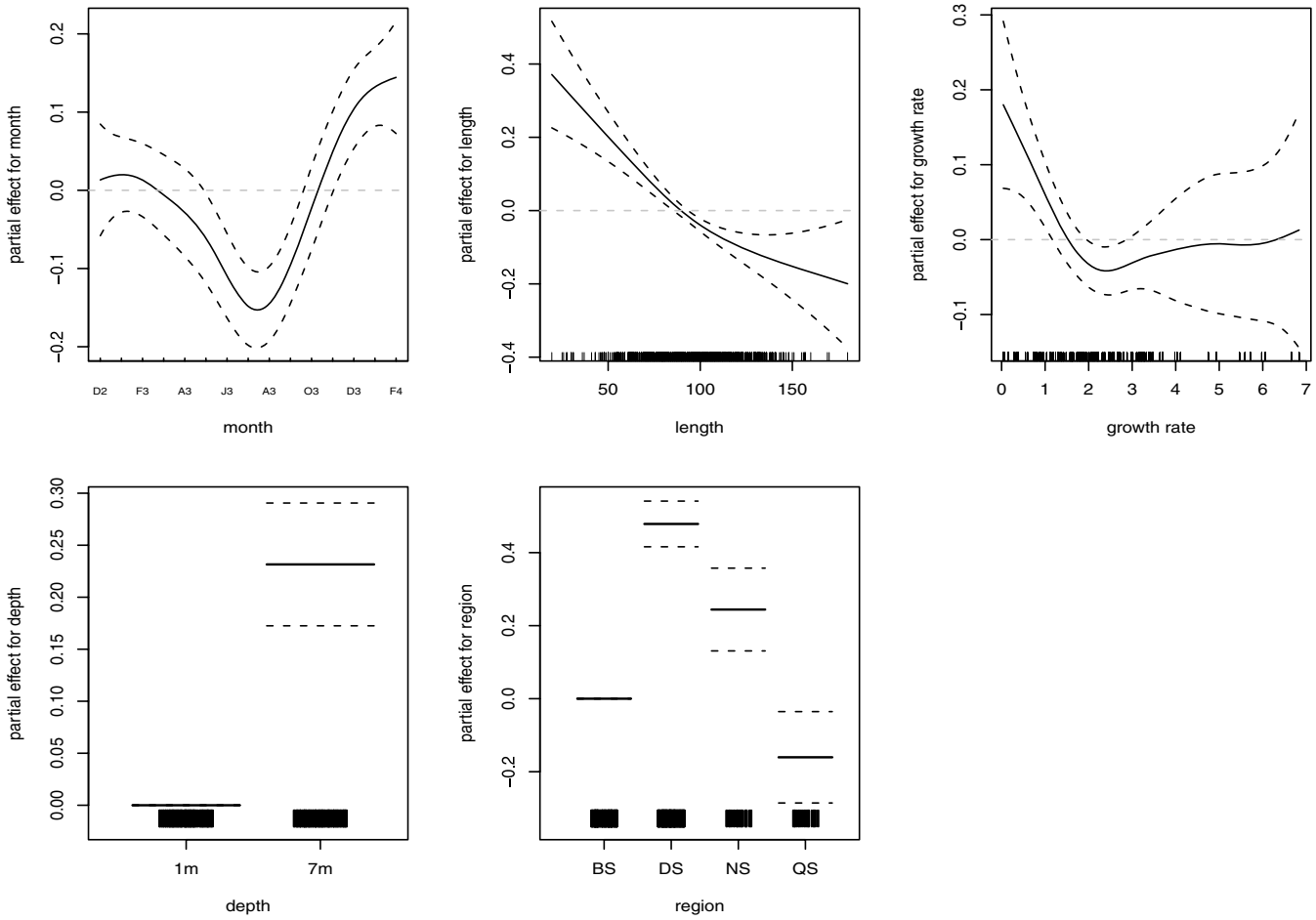

Figure 5. The estimated partial effects of covariates on cadmium concentration in oysters in model M_1. The top left panel displays the estimated partial effect over months $s_0(t)$; the top middle panel displays the estimated partial effect of oyster length $s_1(l_{si}(t))$; and the top right panel displays the estimated partial effect of oyster growth rate $s_2(r_{si}(t))$. The x-axis tick labels in the top left panel represent the months ranging from December 2002 (D2) to February 2004 (F4). The growth rate is calculated as the first derivative of the monotone smoothing curve for the oyster length. The bottom two panels show the estimated partial effects for each level of depth and region, respectively. The effects of being at the depth 1 m and region BS are set as baseline due to the default contrasts having been used. In all the panels, the dashed lines indicate the 95% confidence intervals for the partial effects.

linear regression model. More importantly, it has the ability to examine the nonlinear relationships between the cadmium concentration and a set of covariates when there is no prior knowledge that these relationships should be linear. The nonparametric term of the overall mean trend for oyster cadmium concentrations in the model implies that oysters may have greater cadmium concentration during the colder winter months than the warmer summer months. This may be due to phytoplankton blooms in early spring. However, a longer time series of data is needed to verify this hypothesis properly.

Our model also reveals that oyster cadmium concentrations are significantly different at two depths on average, being notably higher at a depth of 7 m. This is possibly due to the dilution with oysters at 1 meter being heavier than those at 7 m, therefore, the amount of metal to tissue is greater at 7 m than at 1 m, even though the amount of metal is the same. Therefore, it may be advisable for shellfish farmers to avoid harvesting oysters at lower depths. The model also shows that oyster cadmium concentration decreases as oyster length increases, which may be attributed to the fact that oyster has grown more tissue relative to the amount of metal accumulated.

If the interest is to investigate the influence of environmental factors (i.e., temperature, salinity, turbidity, chlorophyll) to rates at which organisms assimilate and utilize energy for maintenance, growth, and reproduction, we may consider models that describe processes involved in the oyster growth such as those constructed with the dynamic budget energy theory (Bourles, Alunno-Bruscia, and Pouvreau, 2009). Such models are based on ecophysiological modeling that details the physiological processes and energetics of an organism in response to environmental fluctuations.

Table 1

Results for the semi-parametric additive model (M_1) and standard linear regression model (M_2). The effects of being at the depth 1 m and region BS are set as baseline due to the default contrasts having been used. Note "edf" represents the effective degrees of freedom of the functional parameters.

Semi-parametric additive model (M_1)

	Estimate	SE	p-value
Parametric coefficients:			
(Intercept)	2.01	0.03	<0.001
Depth 7 m	0.23	0.03	<0.001
Region DS	0.48	0.03	<0.001
Region NS	0.24	0.06	<0.001
Region QS	-0.16	0.06	0.01
Approximate significance of smooth terms:			
	edf	p-value	
$s_0(t)$	3.59	<0.001	
s_1(Length)	1.56	<0.001	
s_2(Growth rate)	3.36	0.008	

Standard linear regression model (M_2)

	Estimate	SE	p-value
(Intercept)	2.37	0.07	<0.001
Length	-0.003	0.01	<0.001
Growth rate	-0.017	0.001	0.10
Depth 7 m	0.25	0.03	<0.001
Region DS	0.47	0.03	<0.001
Region NS	0.23	0.06	<0.001
Region QS	-0.17	0.06	0.01

5. Supplementary Materials

Web Appendices and Figures referenced in Sections 2.4 and 3.1 are available under the Paper Information link at the Biometrics website `http://www.biometrics.tibs.org`. The statistical methods used in this article are implemented using the *`fda`* package and `mgcv` package in `R` (`R` Development Core Team, 2010). The `R` code is provided in the supplementary file.

Acknowledgements

The authors thank Dr Charmaine Dean and Dr Douglas Woolford of the Department of Statistics and Actuarial Science at Simon Fraser University for their valuable statistical advice. The authors would also like to thank Environment Canada for the support and permission to publish this paper and BC Ministry of Agriculture, Food and Fisheries and Department of Fisheries and Oceans for their assistance in carrying out this study. We also thank Dr Thomas A. Louis, an associate editor, and two anonymous referees for their constructive comments, which greatly improve the article.

References

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

Antoniadis, A. and Sapatinas, T. (2007). Estimation and inference in functional mixed-effects models. *Computational Statistics and Data Analysis* **51**, 4793–4813.

Bendell, L. I. and Feng, C. X. (2009). Spatial and temporal variations in cadmium concentrations and burdens in the pacific oyster (Crassostrea gigas) sampled from the Pacific North-west. *Marine Pollution Bulletin* **58**(8), 1137–1143.

Besse, P. and Ramsay, J. O. (1986). Principal components analysis of sampled functions. *Psychometrika* **51**, 285–311.

Bourles, Y., Alunno-Bruscia, M., and Pouvreau, S. (2009). Modelling growth and reproduction of the pacific oyster Crassostrea gigas: Advances in the oyster-DEB model through application to a coastal pond. *Journal of Sea Research* **62**, 62–71.

Chiou, J. M., Muller, H. G., and Wang, J. L. (2004). Functional response models. *Statistica Sinica* **14**, 675–693.

Crambes, C., Kneip, A., and Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *Annals of Statistics* **37**, 35–72.

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

de Boor, C. (1978). *A Practical Guide to Splines*. New York: Springer-Verlag.

Deville, J. C. (1974). Méthodes statistiques et numériques de l'analyse harmonique. *Annales de l'INSEE* **15**, 7–97.

Durban, M., Harezlak, J., Wand, M., and Carroll, R. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine* **24**, 1153–1167.

Ferraty, F. and Vieu, P. (2000). Dimension fractale et estimation de la régression dans des espaces vectoriels semi-normés. *Comptes Rendus de l'Académie des Sciences Paris–Series I Mathematics* **330**, 403–406.

Hall, P., Muller, H. G., and Wang, J. L. (2006). Properties of principal components methods for functional and longitudinal data analysis. *Annals of Statistics* **34**, 1493–1517.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. New York: Chapman and Hall.

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

Kruzynski, G. M. (2000). Cadmium in BC farmed oysters: A review of available data, potential sources, research needs and possible mitigation strategies. Canadian stock assessment secretariat research document 2000, 1437 pp. *Fisheries and Oceans Science*.

Kruzynski, G. M. (2004). Cadmium in oysters and scallops: The BC experience. *Toxicology Letters* **148**, 159–169.

Malfait, N. and Ramsay, J. O. (2003). The historical functional linear model. *The Canadian Journal of Statistics* **31**(2), 115–128.

R Development Core Team. (2010). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science* **3**, 425–461.

Ramsay, J. O. and Dalzell, C. J. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society, Series B* **53**, 539–572.

Ramsay, J. O. and Dalzell, C. J. (1996). Functional data analyses of lip motion. *Journal of the Acoustical Society of America* **99**, 3718–3727.

Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis: Methods and Case Studies*. New York: Springer.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*, 2nd edition. New York: Springer.

Schallie, K. (2001). Results of the 2000 survey of cadmium in B.C. oysters. *Proceedings of a Workshop on Possible Pathways of*

Cadmium into The Pacific Oyster Crassostrea Gigas as Cultured on The Coast of British Columbia **65**, 31–32.

Silverman, B. W. (1995). Incorporating parametric effects into functional principal components analysis. *Journal of the Royal Statistical Society, Series B* **57**, 673–689.

Silverman, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *Annals of Statistics* **24**, 1–24.

Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society, Series B* **62**, 413–428.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association* **99**, 673–686.

Zhou, S. and Shen, X. (2001). Spatially adaptive regression splines and accurate knot selection schemes. *Journal of the American Statistical Association* **96**, 247–259.

Received November 2009. *Revised September* 2010.
Accepted October 2010.