**Functional Data Analysis in Ecosystem Research: the Decline of Oweekeno Lake Sockeye Salmon and Wannock River Flow**

L.M. Ainsworth*, R. Routledge, and J. Cao

Department of Statistics & Actuarial Science

Simon Fraser University, Burnaby, B.C. Canada V5A 1S6

*Corresponding author: Department of Statistics & Actuarial Science

Simon Fraser University, 8888 University Drive, Burnaby, B.C. Canada V5A 1S6. Tel: 01-778-782-3803; fax: 01-778-782-4368

*E-mail addresses*: lmainswo@stat.sfu.ca (L.M. Ainsworth), routledg@stat.sfu.ca (Rick Routledge), jiguo_cao@sfu.ca (J. Cao)

Research was supported by the Natural Sciences and Engineering Research Council of Canada, the Research and National Program on Complex Data Structures and the Tula Foundation.

## SUMMARY

Functional regression is a natural tool for exploring the potential impact of the physical environment (continuously monitored) on biological processes (often only assessed annually). This paper explores the potential use of functional regression analysis and the closely related functional principal component analysis for studying the relationship between river flow (continuously monitored) and salmon abundance (measured annually). The specific example involves a depressed sockeye salmon population in Rivers Inlet, BC. Particular attention is given to (i) the role of subject matter expertise and cross-validation techniques in guiding decisions on basis functions and smoothing parameters, and (ii) the importance of restricting the time domain for the continuously monitored variable to a scientifically meaningful period of time. In addition, we derive a joint confidence region for the functional regression coefficient function, and discuss its use relative to the more commonly used pointwise confidence intervals. The analysis points to a substantial negative correlation between early spring river flow and marine survival of the sockeye salmon that subsequently migrate down the inlet.



KEYWORDS: Functional Linear Regression; Joint Confidence Region; Functional Principal Components; Salmon Abundance; Rivers Inlet, British Columbia

# 1. INTRODUCTION

Recent crises associated with stratospheric ozone depletion and greenhouse gas emissions, as well as more locally concentrated pollution events have sparked global interest in assessing biological impacts of changes to the physical environment. In this context, it is natural to consider regressing a measure of biological performance on physical parameters. Accurate, automated and frequent physical measurements are often readily available; more labor intensive biological estimates are typically less accurate and generated less frequently. This paper focuses on the use and development of functional data analysis (FDA) as a tool for identifying key features of a physical variable and how these relate to biological processes.

We consider a specific example: the impact of river flow on marine survival of an anadromous fish population. While river flow has been monitored continuously and relatively accurately, each annual cohort of out-migrating juvenile fish generates a single, relatively inaccurate marine survival rate. We consider regressing annual marine survival on the river flow measurements for the year in which the fish cohort migrates to sea. Hence, the explanatory variable is a function. Functional regression analysis is designed to handle such a complication and is therefore a natural tool to use in this context.

Functional regression analysis differs from classical regression in that the regression coefficient is a function, $\boldsymbol{\beta}$. This function can be thought of as a weighting of the continuously monitored, autocorrelated covariate. It is the shape of this regression coefficient function that is of primary interest here. In contrast, we typically visualize the line re-

lating the outcome, $y$, to the covariate, $x$. The functional analogy is the line relating $y$ to the weighted sum of the time varying covariate, $\int \beta(t)x_i(t)dt$. Here we use a linear regression model but the regression coefficient function may also be non-linear.

The regression coefficient function indicates the direction and magnitude of the covariate effect throughout the time domain. Our analyses highlight a need for more rigorous methodology for delineating the time windows over which the physical variable, river flow, is significantly related to the biological variable, marine survival. To address this need we develop a joint confidence region for $\boldsymbol{\beta}$ which allows simultaneous inference across the time domain.

Functional regression methodology exists for inference at a given time point as well as for prespecified linear combinations such as time contrasts and linear probes (Ramsay and Silverman, 2006). A linear probe could be used, for example, to assess the impact of March river flow on sockeye salmon marine survival rates, by using a box function to weight all March river flow values equally and giving zero weight to all remaining values.

There is a great deal of work on developing methods for simultaneous inference in the classical linear regression context as well as for nonparametric regression. Liu, Lin, and Piegorsch (2008) review theory and methodology associated with simultaneous confidence bands for linear regression. Scheffé's (1953, 1959) familiar method for comparing all contrasts readily extends to the multiple regression context. Since it generates confidence bands that are simultaneously valid over unbounded regions, these tend to be conservative for practical applications where the domain is finite.

For simple linear regression, tighter confidence regions can be derived by exploiting the underlying geometry. However, extensions, even merely to multiple linear regression

problems, are notably challenging. Liu, Jamshidian, Zhang, and Donnelly (2005) propose simulation-based methodology for obtaining more exact coverage probabilities for the special case when the independent variables are constrained to lie in a hyper-rectangle.

Other extensions include Hardle and Marron's (1991) bootstrap-based simultaneous error bars for nonparametric kernel regression estimates. Loader and Sun (1997) and Sun and Loader (1994) obtain approximate simultaneous confidence bands for a function of a covariate using the 'tube formula'. This method can be used for multidimensional $x$ and a wide class of linear estimates.

Improved confidence regions have also been proposed for cases in which the covariates are related to each other. For instance, Liu, Wynn, and Hayter (2008) account for functional relationships between the covariates in order to provide improved simultaneous confidence bands for polynomial regression. Wu, Chiang, and Hoover (1998) considered the varying-coefficient model $Y(t) = X(t)\gamma(t) + \epsilon(t)$ where repeated measurements within a subject are correlated over time. They established an asymptotic distribution for a kernel estimate of $\gamma(t)$ and constructed a class of approximate pointwise and simultaneous confidence bands for $\gamma(t)$.

To our knowledge, the only methods available for simultaneous inference on the functional regression coefficient function are simulation-based. For example, Hardle and Marron (1991) use a bootstrap approach for a kernel regression estimator, and Herwartz and Xu (2009) use a bootstrap approach to inference for functional coefficient models. Bootstrap re-sampling is computationally intensive. Hence, an easily programmed alternative is useful.

This paper develops a joint confidence region for the regression coefficient function, $\boldsymbol{\beta}$,

to accommodate simultaneous inference across the time domain. Using numerical quadrature and the properties of basis functions, we reduce functional regression estimation to standard multiple linear regression and use Scheffé's (1953) methodology to generate a joint confidence region for $\boldsymbol{\beta}$ which allows simultaneous inference at the quadrature points. This is straightforward to compute and easy to implement. We explore the adequacy of this approach in the current context, paying special attention to the inclusion of a penalty term, the use of cross-validation for guiding decisions regarding choice of smoothing parameter, and the impact of smoothing on the coverage probabilities of the joint confidence regions. This provides valuable knowledge for assessing the potential usefulness of FDA in ecosystem research.

The paper is structured as follows. In Section 2, we introduce the application, the impact of Wannock River flow on marine survival of sockeye salmon in Rivers Inlet. In Section 3, we discuss functional data analysis techniques and show how Scheffé's method can be adapted to generate a joint confidence region for a functional regression coefficient function. In Section 4, we present the analysis of the example data. In Section 5, we explore the properties of the proposed confidence region via a simulation study. Section 6 wraps up with a discussion of the issues and considerations of particular relevance to applications of functional data analysis to environmental data.

## 2. SALMON ABUNDANCE AND RIVER FLOW IN RIVERS INLET

Sockeye salmon returning to Rivers Inlet on the British Columbia Central Coast once supported the third largest sockeye salmon fishery in the province. Returns that formerly often exceeded a million adults dwindled to a few thousand by 1999 (Rutherford and Wood, 2000). Since then, the population has shown only sporadic signs of partial recovery. The cause of this conservation disaster, and even the date of onset for the decline, remains uncertain.

Adult sockeye salmon escaping the fishery spawn in freshwater gravel beds in the late summer and fall. The young salmon emerge from the gravel, spend a year rearing in Oweekeno Lake, at the head of the inlet, and then migrate down the Wannock River to the inlet and make their way to sea, late the following spring. Most of the surviving adults return at age 4 or 5.

The migrating sockeye salmon smolts in this population are remarkably small (Burgner, 1991). Smaller smolts typically experience lower marine survival rates (Koenings, Geiger, and Hasbrouk, 1993), and the small size of these particular smolts has long been a concern (Foskett, 1958). Abundant food in Rivers Inlet could provide a vital opportunity for the juvenile sockeye salmon to grow rapidly and thereby increase their chance of survival.

Limited evidence available in the aftermath of 1999 was primarily from the freshwater environment. Fry abundance in the lake remained relatively high while adult returns were decreasing. This and other evidence suggests that the cause of the decline was likely to be found in the early-marine phase of the sockeye salmon life cycle (McKinnell, Wood,

Rutherford, Hyatt, and Welch, 2001). This largely circumstantial evidence prompted research on the juvenile migration down the inlet. Evidence to date (Buchanan, 2006) has shown that the juvenile sockeye salmon migration peaks at the first new moon in late May or early June.

A major spring event in many temperate ecosystems is the first spring phytoplankton bloom. The timing and size of this event depend on many factors such as inlet conditions and local weather (Reynolds, 2006; Tommasi, 2008). In Rivers Inlet, our limited observations to date suggest that the spring bloom typically occurs sometime in March or April (Tommasi, 2008).

Cleorn (1991) and Mallin et al. (1993) reported that, for systems influenced by runoff from nutrient-rich watersheds (San Francisco Bay and the Neuse River estuary in North Carolina respectively), plankton blooms were associated with periods of high river discharge. By contrast, the Wannock River drains the nutrient-poor Oweekeno Lake (McKinnell et al., 2001). Thus, high river flows generate the formation of a nutrient-poor, brackish surface layer. This layer can not only decrease the intensity of solar radiation penetrating to the more nutrient-rich marine water, but can also disrupt phytoplankton development through entrainment and horizontal advection as the lighter, brackish layer flows down the inlet. High river flows in spring are typically generated by intense Pacific storms, whose winds and low light levels will directly inhibit phytoplankton development. It is generally accepted that a period of water column mixing, followed by a period of water column stability, is necessary for the development of a spring phytoplankton bloom (Mann and Lazier, 2006). It is therefore reasonable to anticipate that high early spring river flows would disrupt the development of the plankton food chain, create a food short-

age for the juvenile salmon, and thereby decrease their marine survival rate. A primitive retrospective analysis of the correlation between early spring river discharge and marine sockeye salmon survival has supported this hypothesis.

Here, we exploit functional data analysis to solidify and extend the preliminary analysis of the relationship between river flow and marine sockeye salmon survival. We do not attempt to tease out the differential effects of high river flow itself versus the storms and associated winds, decreased solar radiation, etc., that typically generate the higher flows in late winter and early spring. Related, ongoing interdisciplinary field research is generating insight on the plankton dynamics that may in turn help to separate the impacts of these confounding factors.

## 3. FUNCTIONAL DATA ANALYSIS

A key concept in FDA is the representation of functional data as a series of basis functions which project the original data onto a smaller space (Ramsay and Silverman, 2006). A basis function system is a set of known and mathematically independent functions. Any function can be approximated arbitrarily well by taking a weighted sum or 'linear combination' of a sufficiently large number, $K$, of these basis functions. That is, we can summarize, $\beta(t) = \sum_{k=1}^{K} c_k \phi_k(t)$ where the $\phi_k, k = 1, 2, \ldots, K$, are a set of basis functions. For instance, one may use a Fourier Series to summarize cyclic, seasonal trends in data. On the other hand, B-splines are not restricted to be periodic and often provide

flexibility for modelling deviations from seasonal trends. B-splines are also computationally efficient, as they have 'compact support'. That is, any B-spline basis function is only non-zero over a small interval (Ramsay and Silverman, 2006).

3.1 Functional Regression Analysis

Functional regression analysis handles both functional and scalar variables. In the present context, we consider functional regression analysis of a scalar variable, marine survival, on a univariate functional variable, river flow. We model the log marine survival index, $Y_i$, as

$$Y_i = \int \beta(t)x_i(t)dt + \epsilon_i, \qquad i = 1, 2, \ldots, n \tag{1}$$

where the $\epsilon_i$ are independent with mean zero and variance, $\sigma^2$, and $x_i(t)$ is the covariate for year $i$ at time $t$. Here, we assume $Y_i$ to be normally distributed. Alternatively, one can consider a generalized functional linear model such as that used by Muller and Stadtmuller (2005).

Using a numerical quadrature method, Simpson's Rule (Atkinson, 1989), we can write the above model as

$$Y_i \approxeq \sum_{j=1}^{J} w_j\beta(t_j)x_i(t_j) + \epsilon_i,$$

where $t_j, j = 1, 2, \ldots, J$, is a sufficiently fine grid of quadrature points, and $w_j$ are the quadrature weights.

We write, $\beta(t)$ as a linear combination of basis functions,

$$\beta(t) = \sum_{k=1}^{K} c_k\boldsymbol{\phi}_k(t),$$

Let $\boldsymbol{\beta} = (\beta(t_1), \beta(t_2), \beta(t_3), \ldots, \beta(t_J))'$, $\boldsymbol{c} = (c_1, \ldots, c_K)$ and $\boldsymbol{\Phi}$, a matrix of order $J$ by $K$, with elements $\phi_k(t_j)$ in the $j^{th}$ row and $k^{th}$ column. Then, in matrix notation, we have

$\boldsymbol{\beta} = \boldsymbol{\Phi c}$, a simple linear transformation of $\boldsymbol{c}$.

Hence,

$$Y_i \approxeq \sum_{k=1}^{K} c_k \sum_{j=1}^{J} w_j \boldsymbol{\phi}_k(t_j) x_i(t_j) + \epsilon_i.$$

Letting

$$z_{ik} = \sum_{j=1}^{J} w_j \boldsymbol{\phi}_k(t_j) x_i(t_j),$$

we get

$$Y_i \approxeq \sum_{k=1}^{K} c_k z_{ik} + \epsilon_i.$$

Let $\boldsymbol{Z}$, be an order $n$ by $K$ matrix with elements $z_{ik}$ in the $i^{th}$ row and $k^{th}$ column and let $\boldsymbol{\epsilon}$ be the $n$ by 1 vector with elements $\epsilon_i$. Then in matrix notation,

$$\boldsymbol{Y} = \boldsymbol{Zc} + \boldsymbol{\epsilon}$$

Since we assume that $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$, we can use standard linear regression to solve for $\boldsymbol{c}$,

$$\hat{\boldsymbol{c}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y},$$

and the covariance matrix for $\hat{\boldsymbol{c}}$ is,

$$Cov(\hat{\boldsymbol{c}}) = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}' Var(\boldsymbol{\epsilon}) \boldsymbol{Z}[(\boldsymbol{Z}'\boldsymbol{Z})^{-1}]' = \sigma^2 (\boldsymbol{Z}'\boldsymbol{Z})^{-1}.$$

Since $\boldsymbol{\beta} = \boldsymbol{\Phi c}$, is a linear transformation of $\boldsymbol{c}$, we have,

$$Cov(\hat{\boldsymbol{\beta}}) = \boldsymbol{\Phi} Cov(\hat{\boldsymbol{c}}) \boldsymbol{\Phi}' = \boldsymbol{\Phi}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}' Var(\boldsymbol{\epsilon}) \boldsymbol{Z}[(\boldsymbol{Z}'\boldsymbol{Z})^{-1}]' \boldsymbol{\Phi}' = \sigma^2 \boldsymbol{\Phi}(\boldsymbol{Z}'\boldsymbol{Z})^{-1} \boldsymbol{\Phi}'$$

3.2 Joint Confidence Region for $\hat{\boldsymbol{\beta}}$

Theorem 1

Let $Y_i = \int \beta(t)x_i(t)dt + \epsilon_i, \boldsymbol{\epsilon} \sim N(0, \sigma^2\boldsymbol{I}), i = 1, 2, \ldots, n$, denote the functional regression model with $\hat{\boldsymbol{\beta}} = \boldsymbol{\Phi}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y}$ and $\boldsymbol{Z}$ a design matrix relating $y$ to $\boldsymbol{\beta}$.

A $100(1-\alpha)$ % joint confidence region for $\hat{\boldsymbol{\beta}}$ at quadrature points $t_j$ is

$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'(\boldsymbol{\Phi}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{\Phi}')^{-1}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \leq rs^2F_{r,n-r}(\alpha) \tag{2}$$

where $s^2 = \sum(y_i - \hat{y}_i)^2/(n-r)$ and $r$ is the number of basis functions. For an outline of the proof see Appendix A.

Theorem 1 provides us with a joint confidence region for $\hat{\boldsymbol{\beta}}$ at the quadrature points, $t_j$. We could make the quadrature points infinitesimally close together. However, there is no advantage to interpolating beyond the resolution of the river flow data.

3.3 Roughness Penalty

A basic identity in statistics tells us that the mean square error, a commonly used 'loss function', consists of bias and variance: $MSE[\hat{x}(t)] = Bias^2[\hat{x}(t)] + Var[\hat{x}(t)]$. If allowing for some small amount of bias results in a substantial decrease in variability, then the mean square error may be reduced. A completely unbiased estimate can be obtained by a curve which fits the observed data exactly. However, such an estimate is subject to a large degree of random variation which manifests as rapid local fluctuations in the curve. This is the key reason for imposing smoothness on the estimated curve and thereby introducing bias.

By requiring the estimate to vary smoothly, we 'borrow strength' from data values at neighbouring time points. This is an idea that has been used extensively in the spatial

literature (see for example Mollié, 1996). If the underlying function is indeed smooth, then the estimate will be more stable. The trick is to find the 'best' balance between bias and variance. If we choose a large number of basis functions, we can fit the data very well but we risk fitting noise. To counteract this tendency we can either choose fewer basis functions, smoother basis functions, or impose a roughness penalty on the least squares criterion.

Thus, we add a roughness penalty on $\beta(t)$. The penalty term is often chosen to be $PEN = \int \{\frac{d^2\beta(t)}{dt^2}\}^2 dt$. Then the criterion for estimating the basis coefficients is $PENSSE = (\boldsymbol{Y} - \boldsymbol{Zc})'(\boldsymbol{Y} - \boldsymbol{Zc}) + \lambda \int \{\frac{d^2\boldsymbol{\beta}(t)}{dt^2}\}^2 dt$. In this case, curvature at $t$ increases the penalty and introduces bias. The smoothing parameter, $\lambda$, controls the amount of smoothing and measures the trade-off between fit to the data and smoothness of the function.

The vector of penalized least squares estimates of the basis function coefficients is then,

$$\hat{\boldsymbol{c}}_\lambda = (\boldsymbol{Z}'\boldsymbol{Z} + \lambda \boldsymbol{R})^{-1}\boldsymbol{Z}'\boldsymbol{y},$$

where $\boldsymbol{R} = \int (\frac{(d^2\boldsymbol{\phi}(t)}{dt^2})(\frac{d^2\boldsymbol{\phi}(t)}{dt^2})' dt$. Let $\boldsymbol{B} = (\boldsymbol{Z}'\boldsymbol{Z} + \lambda \boldsymbol{R})^{-1}$ and we have $cov(\hat{\boldsymbol{c}}_\lambda) = \sigma^2 \boldsymbol{B}\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{B}'$. Again, we use Simpson's Rule to approximate the integral in $\boldsymbol{R}$. As before, we assume multivariate normal $\boldsymbol{\epsilon}$ and consider a joint confidence region based on Theorem 1. This confidence region is strictly valid only for unbiased estimates of $\boldsymbol{c}$. For an appropriate amount of smoothing we expect to have a small amount of bias and thus expect the region to be a reasonable approximation.

As in Section 3.1, we transform $\hat{\boldsymbol{c}}_\lambda$ to get the penalized regression function estimate,

$\hat{\boldsymbol{\beta}}_\lambda = \boldsymbol{\Phi}\hat{\boldsymbol{c}}_\lambda$. Based on Theorem 1, we derive an approximate joint confidence region for $\hat{\boldsymbol{\beta}}_\lambda$,

$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda)'(\boldsymbol{\Phi}\boldsymbol{B}\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{B}'\boldsymbol{)}\boldsymbol{\Phi}')^{-1}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\lambda) \leq rs^2 F_{r,n-r}(\alpha) \tag{3}$$

where, $r$, is taken to be the effective degrees of freedom. One suggestion for the effective degrees of freedom made by Buja, Hastie and Tibshirani (1989) is $r = trace(A)$, $\boldsymbol{A} = \boldsymbol{Z}\boldsymbol{B}\boldsymbol{Z}'$, the hat matrix; another is $r = trace(AA')$.

The appropriateness of the joint confidence region for the penalized case depends on the amount of bias introduced. In Section 5, we run a simulation study to investigate coverage probabilities and choice of smoothing parameter.

A practical consideration when using a roughness penalty is how to choose the smoothing parameter. The optimal amount of smoothing may depend on the context of the study. However, the generalized cross-validation measure,

$$GCV(\lambda) = \frac{nSSE(\lambda)}{(n-r)^2} \tag{4}$$

provides an objective criteria which is often used to choose $\lambda$ (Craven and Wahba, 1979; Wahba, 1985). A plot of $GCV(\lambda)$ along with plots of $\hat{\boldsymbol{\beta}}$ for variety of $\lambda$ values, shows the influence of the smoothing parameter. These plots are particularly useful for discussions with biologists and other applied researchers, who can offer valuable insight into which curves appear most reasonable in light of background knowledge in the application area. In our experience, choosing the point at which the GCV($\lambda$) value levels off tends to provide a good choice from a practical perspective. This is not a novel approach, an analogous rule of thumb is used for choosing the number of components to retain in a principal components analysis.

# 4. SAMPLE APPLICATION: THE RELATIONSHIPS BETWEEN RIVER FLOW AND SALMON ABUNDANCE

## 4.1 Data

An index of annual sockeye salmon marine survival was determined based on data extracted from McKinnell et al. (2001). The marine survival index is the estimated number of adult returns divided by an index of premigration, juvenile abundance. The index values are available for 29 of the years from 1964 to 1998, but are unavoidably inaccurate. To promote the aptness of the standard regression model, we base our analysis on the logarithm of the marine survival index.

Daily river flow measurements for the Wannock River are available from Environment Canada ($www.wsc.ec.gc.ca/products/hydat/main\_e.cfm?cname = hydat\_e.cfm$). The Wannock River is located on the central coast of British Columbia approximately 100 km north of Port Hardy. It flows from the outlet of the Oweekeno Lake west approximately 6.5 km to the head of Rivers Inlet, has a mean channel width of approximately 100m, and drains roughly half the entire inlet watershed. Water conditions tend to be highly turbid from April to December due to glacial run-off from the Monarch Icefield and Silverthrone glacier.

In functional data analysis, the resolution of the raw data is an important consideration. In order to capture the peaks and valleys in functional data, the data need to be collected on a sufficiently dense time scale. Daily river flow measurements are felt to have sufficiently high resolution to capture important fluctuations which might relate to

marine survival of salmon.

It is also important to choose a set of basis functions which mimic the characteristics of the process generating the data. Thus, we decomposed river flow data into a yearly, seasonal mean trend modelled using a Fourier series, and smaller scale deviations from these seasonal trends modelled using B-splines. That is, residual river flow is calculated as log river flow minus the seasonal Fourier fit to log river flow. The 'detrended' residual function is then approximated using B-splines. Let $\boldsymbol{x}(t)$ denote the residual river flow. A knot was placed every 2 days in order to produce a function which closely mimics the discretely collected data points. Here, we do not address the negligible bias introduced in this initial data preparation step.

The Fourier fit to yearly log river flow indicates an early-spring low. This is followed by a steep rise generated initially by snow melt and then sustained through the summer by glacial melt-waters. A declining trend accelerates as the summer progresses. The decline is then slowed by fall rains, but subsequently resumes as winter causes more of this precipitation to fall as snow, and then as the stormy season begins to ease toward winter's end.

4.2 Smoothing, Basis Functions and Functional Linear Regression

Recall that we regress marine survival on log river flow. We use B-splines with a knot every 14 days to model the regression coefficient function. A penalty term is introduced and the generalized cross validation measure, GCV($\lambda$), is used to study the balance between bias and variance and its impact on regression coefficient function estimation.

Figure 1 shows the estimated regression coefficient function, the pointwise confidence intervals and the joint confidence region for a variety of smoothing parameters in three

time windows. The final graph in each panel displays the GCV($\lambda$) values. The minimum value of GCV($\lambda$) suggests that a smoothing parameter of $\lambda = 10^{-1}$, $\lambda = 10^{-2.8}$ and $\lambda = 10^{-2.8}$ are appropriate for the time frames July 1 to June 30, December 1 to June 30, and February 1 to May 31, respectively.

As done by Ramsay and Silverman (2006) we use the ratio of sums of squares to create an approximate $F$-test for the regression. Using $\lambda$ chosen based on the minimum GCV criteria suggests the regressions are significant for all time intervals: July to June, $F_{2.0,27.0} = 4.6, p = 0.015, R^2 = 0.25$; December to June, $F_{2.5,26.5} = 4.13, p = 0.018, R^2 = 0.28$; February to May, $F_{2.5,26.5} = 4.15, p = 0.017, R^2 = 0.29$.

When one chooses the full year as the time domain, the GCV criterion leads to a regression coefficient function with no curvature and a broad significant time window from the beginning of January to mid-May. The line produced by such smoothing is not biologically consistent; it does not make sense for the function to be continuously decreasing to its largest (negative) value on June 30 when the juvenile migration is virtually over. If one chooses to ignore the GCV criterion and uses less smoothing ($\lambda = 10^{-5}$ or less), then the curve suggests an initial dip in the $\boldsymbol{\beta}$ function early in the calendar year. This is more consistent with the background knowledge on the early spring phytoplankton bloom outlined in Section 2, but the joint confidence region does not confirm the significance. It appears that the information in the data is insufficient to support any detailed inference over this broad time domain. To draw further inference, we restrict the time domain.

The second row of Figure 1 shows results for the restricted time window, December 1 to June 30. Since this extends from early winter through to the time when the juvenile sockeye salmon migration is typically winding down, such a restriction seems justifiable.

The estimated regression function corresponding to the minimum GCV($\lambda$) is curved and the joint confidence region indicates that at least the portion of this window from the beginning of February through to the beginning of April is significant. Also note that GCV($\lambda$) reaches a minimum around $\lambda = 10^{-3}$ and then begins to increase. This is biologically consistent as it implies that a straight line fit is not sufficient to describe the regression coefficient function.

Nonetheless, we are not entirely comfortable with this analysis. The GCV($\lambda$) value changes very little between $\lambda = 10^{-4}$ and $\lambda = 10^{-2.8}$. However, the smaller smoothing parameter produces a regression coefficient function with a more distinct dip in March and early April that returns to 0 at the end of June as the fish migration winds down. Backing off even further to $\lambda = 10^{-5}$ produces a function with even more emphasis on the importance of early spring flow during the key time when the first spring phytoplankton bloom appears likely to occur. Hence, it appears that the GCV criterion promotes oversmoothing in this context as well.

If we use further background information to restrict the time domain to February 1 through May 31 (Figure 1 row 3), then the results are even more encouraging. The GCV criterion leads to choosing $\lambda = 10^{-2.8}$ and the corresponding $\boldsymbol{\beta}$ function has a biologically reasonable form. Graphs with less smoothing also appear reasonably shaped. These are in accord with field observations of early signs of phytoplankton increases in February, the first bloom in early April, and repeated periods of high abundance through the later spring. Thus, judicious use of background knowledge allows us to zoom in on the key time window and thereby generate a more definitive assessment.

In all cases, the analysis suggests the same general conclusion, a negative relationship

between marine survival of salmon and early spring river flow. However, in the absence of any background knowledge to focus the time domain, functional regression analysis was unable to detect the curve in the regression coefficient function. Circumstantial evidence suggests a tendency for the GCV criterion to promote oversmoothing in this context. This is explored further in the simulation study discussed in Section 5.

### 4.3 Functional Principal Components Analysis

Functional principal components analysis (FPCA) provides another approach to exploring this data. Whereas functional regression analysis highlights time points which are significantly related to marine survival, FPCA highlights the time intervals accounting for the greatest annual variability in river flow. See Ramsay and Silverman (2006) for a full discussion of FPCA.

Marine survival can be regressed on the functional principal components of river flow. We present results for the time interval December 1 to May 31. Figure 2 presents the effect of adding and subracting a suitable multiple of the principal components to the overall mean curve. The functional principal components highlight variability in the following time frames: Component 1 (Figure 2a), during January and February with a prolonged tail extending over March and April; Component 2 (Figure 2b), the contrast between December and February-March flows; and Component 3, (Figure 2c), March and early April flows. The first three functional principal components account for approximately 36%, 24%, and 18% of the variability in the residual river flow curves respectively.

The first and third components are correlated with marine survival (r = -0.38 and r = -0.45) while the second is not (r=0.08). Standard multiple linear regression indicates that the $1^{st}$ and $3^{rd}$ functional principal component scores are significant predictors of marine

survival ($F_{2,26} = 6.81, p = 0.004, R^2 = 0.34, AdjR^2 = 0.29$). On the other hand, the $3^{rd}$ rotated principal component (Figure 2d) is even more highly correlated with marine survival (r=-0.54). This component captures both March and April flows as well as flow in January.

Both functional regression analysis and regression using functional principal component scores indicate a strong negative association between marine survival and river flow in March and early April. The FPCA analysis also suggests a potential relationship in the early part of the year. In order to validate the results found here we are pursuing parallel analysis in similar inlets along the B.C. coast.

## 5. SIMULATION STUDY

For the purpose of the simulation study we focused on the time domain February 1 to May 31. The regression coefficient function (Section 4.2) estimated using smoothing parameter $\lambda = 10^{-2.8}$, was taken as the true regression coefficient function, $\boldsymbol{\beta}$, and $\sigma^2$ was set to the estimated error variance, $s^2 = 1.5$. We simulated 10,000 datasets such that $y_i^* = \int \beta(t)x_i(t)dt + \epsilon_i^*$, with $\epsilon_i^* \sim N(0, \sigma^2 I)$.

The regression coefficient function, pointwise confidence intervals, and joint confidence regions were estimated for the simulated data sets using smoothing parameter values of, $\lambda = 10^{-4}, 10^{-2.8}, 10^{-1}$. The pointwise confidence interval coverage, the coverage corresponding to simultaneous inference and the joint confidence region coverage probabilities

are presented in Table 1. The top half of the table uses degrees of freedom, $trace(A)$ whereas the bottom half uses $trace(AA')$. The pointwise coverage is close to nominal except when the curve is smoothed to essentially a line ($\lambda = 10^{-1}$). However, the issue of multiple testing arises when one uses pointwise intervals for simultaneous inference. As expected, using pointwise intervals to make inference across the entire time domain leads to unacceptably low coverage probabilities (presented in brackets in Table 1).

The joint confidence regions are slightly liberal when the amount of smoothing is the same as that used to generate the data ($\lambda = 10^{-2.8}$) but conservative for less smoothing ($\lambda = 10^{-4}$). When the curve is restricted to be very smooth ($\lambda = 10^{-1}$), coverage is unsatisfactorily low. As noted above, this is not surprising as fitting a line to a 'U' shaped curve introduces a great deal of bias and renders such regions meaningless.

Using the alternative definition of effective degrees of freedom, $trace(AA')$, has little effect on the coverage probabilities for the pointwise intervals or the joint confidence regions for a large amount of smoothing. On the other hand, when a moderate amount of smoothing is used, $\lambda = 10^{-4}$, the coverage probabilities match the nominal values.

In practice, one might consider choosing the amount of smoothing, $\lambda_{GCV}$, via the minimum GCV criterion. Here we consider smoothing values in the range, $\lambda = 10^{-6}$ to $\lambda = 10^{-1}$ at intervals of $10^{0.5}$ and get $\lambda_{GCV}$ using effective degrees of freedom defined as $trace(A)$. For the simulation study we know the true curve and can determine the 'gold standard' for smoothing, $\lambda_{MSE}$, based on the mean square error (MSE). Coverage probabilities corresponding to this choice of smoothing parameter are presented in Table 1. Both pointwise intervals and joint regions using $\lambda_{GCV}$, give coverage probabilities well below nominal levels. The coverage probabilities obtained for $\lambda_{MSE}$, a case not possible

in practice, are very close to nominal values. Here, the minimum GCV criterion leads to oversmoothing; the $1^{st}, 2^{nd}$, and $3^{rd}$ quartiles for $\lambda_{GCV}$ and $\lambda_{MSE}$ are $10^{(-3.5,-2,-1)}$ and $10^{(-3,-3,-2)}$ respectively. Thus, use of this criteria led to excess bias and it is not surprising that the confidence regions are unreliable.

Here, slightly less smoothing than that suggested by the minimum GCV criterion leads to better inference. Figure 3 shows the coverage probabilities and average GCV($\lambda$) values for a variety of smoothing parameters used in the simulation study. We see that the point at which the GCV($\lambda$) curve begins to level off corresponds to the end of the range of $\lambda$ for which coverage probabilities attain approximately nominal values.

Again, the evidence suggests choosing the smoothing parameter based on the point at which the GCV($\lambda$) plot begins to level off. It also seems prudent to look at the GCV($\lambda$) plots along side plots of the estimated regression coefficient function. For instance, for the December 1 to May 31 time domain (Figure 1, row 2), the GCV($\lambda$) value changes very little between $\lambda = 10^{-4}$ and $\lambda = 10^{-2.8}$ yet, the corresponding curves differ in an important way: the rougher curve reveals a distinct dip in March and early April.

The appropriateness of our joint confidence regions for penalized least squares depends on the amount of bias introduced. Figure 4 shows histograms of $F_{n,n-r}$ and of the empirical values $\hat{F}$ when no smoothing is used. Our empirical results confirm the appropriateness of the $F-$distribution when no smoothing is considered. For the case considered here, using the minimum GCV criterion leads to oversmoothing and bias so that the $F_{n,n-r}-$distribution is a poor approximation to the empirical distribution. However, the approximation does seem to be valid for a moderate amount of smoothing such as that recommended here: the point at which the GCV($\lambda$) curve begins to level off

($\lambda = 10^{-4}$).

## 6. DISCUSSION

We view functional data analysis as an important tool for empirical model building in ecosystem research. A joint confidence region for the regression coefficient function and an approximate $F$-test for the significance of the relationship between river flow and marine survival are valuable even if one can obtain only vague estimates of the overall shape of the regression coefficient function. Our sample application highlights the value of background knowledge in refining the time window for the analysis and in selecting the appropriate amount of smoothing. Through such refinements, we are able to uncover evidence of a key significant time period during March and early April.

Scheffé's approach is used to create a joint confidence region which is valid for unbiased estimates. For this small sample size the use of the minimum GCV criteria tends to overestimate the optimal amount of smoothing. This oversmoothing introduces substantial bias so that the coverage probabilities of the joint confidence region are below target. On the other hand, for small $\lambda$, the coverage probabilities are large due to increased variability, $Var(\hat{\boldsymbol{\beta}})$, related to overfitting the data. For optimal smoothing, the proposed confidence region has good coverage.

In our case, the optimal amount of smoothing is well approximated by choosing the leveling off point of the GCV($\lambda$) curve. We recommend the user look not for the value of

the smoothing parameter that minimizes the GCV($\lambda$) value, but where this function starts to level off and where small increases in GCV($\lambda$) lead to substantial increased flexibility in the shape of the curve. Such a subjective decision requires ongoing collaboration with subject-matter specialists. An objective criterion for such a choice would be valuable.

Could a researcher generate comparable inferences with simpler techniques? One could, e.g., run a very large number of simple regressions of marine survival vs. average river discharge over a wide variety of time windows and seek the one with the highest correlation. This was in fact the basis for an earlier analysis. However, such an approach is not only tedious but also generates difficulties associated with assessing the significance of an extreme from many tests on the same data set. The joint confidence region allows the researcher to address this key issue. It is primarily useful as a guide in addressing the question, "What portion of the time domain is unequivocally significant?". It is not perfect: inferences depend on subjective judgments on the level of smoothing. But, this is also the case in standard inferences in spectral analysis as well as in Bayesian approaches to confidence intervals for regression with a penalty (Wahba, 1983, Silverman, 1985). In any case, it is a considerable improvement over naive use of standard pointwise intervals and does not require computationally intensive bootstrap resampling.

The active collaboration between scientists and statisticians in this project generates iterative refinements. Exploratory field research generated the simplistic statistical exploration using regression on simple time averages. This analysis highlighted the potential importance of the early spring time period, which in turn instigated more focused field research. Insight from the field research has helped to refine and focus the exploratory functional data analyses. Our collaborative team is currently constructing parsimonious,

mechanistic models of both the inlet hydrodynamics and the first spring phytoplankton bloom - the key event establishing the development of the food chain base. We plan to be integrally involved in the assessment and refinement of these models.

In summary, it is important to bear in mind that there are many subtle aspects to functional data analysis which require subjective judgment calls. Of particular importance are choices of time windows and the degree of smoothing. These require judicious use of scientific background knowledge. Since these techniques also require advanced statistical expertise - especially in calculating and interpreting confidence regions - we encourage their use in collaborative research projects where statistical experts are engaged as full collaborators. We recommend their use in the early phases of iterative, collaborative research projects.

**ACKNOWLEDGEMENTS**

We would like to thank the TULA Foundation, NSERC and the NPCDS for financial assistance and the IRMACS Centre for providing facilities for this research.

## APPENDIX A

Proof of Theorem 1

A $100(1-\alpha)\%$ confidence region for $\hat{\boldsymbol{\beta}}$ at quadrature points $t_j$ is,

$$(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'(\boldsymbol{\Phi}(Z'Z)^{-1}\boldsymbol{\Phi}')^{-1}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) \leq rs^2F_{r,n-r}(\alpha) \tag{5}$$

where $s^2=\sum(y-\hat{y})^2/(n-r)$ and $r$ is the number of basis functions.

Using the standard multiple linear regression result that a $100(1-\alpha)$ % simultaneous confidence band for $\hat{\boldsymbol{c}}$ is,

$$(\boldsymbol{c}-\hat{\boldsymbol{c}})'(Z'Z)(\boldsymbol{c}-\hat{\boldsymbol{c}}) \leq rs^2F_{r,n-r}(\alpha) \tag{6}$$

We have $\hat{\boldsymbol{\beta}}=\boldsymbol{\Phi}\hat{\boldsymbol{c}}$ a linear transformation of $\hat{\boldsymbol{c}}$. Note that $\hat{\boldsymbol{c}} \sim N_r(\boldsymbol{c},\sigma^2(\boldsymbol{Z}'\boldsymbol{Z})^{-1})$. Therefore, using standard multivariate normal theory we have $\hat{\boldsymbol{\beta}} \sim N_J(\boldsymbol{\Phi}\hat{\boldsymbol{c}},\sigma^2\boldsymbol{\Phi}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{\Phi}')$ where $J$ is the number of quadrature points. Note that $\sigma^2\boldsymbol{\Phi}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{\Phi}'$ does not have full rank; it has rank $r$.

We proceed by showing that the confidence region for $\hat{\boldsymbol{c}}$ holds and then transform to $\hat{\boldsymbol{\beta}}$. Consider the symmetric square root matrix $(\boldsymbol{Z}'\boldsymbol{Z})^{1/2}$. Let $V=(\boldsymbol{Z}'\boldsymbol{Z})^{1/2}(\boldsymbol{c}-\hat{\boldsymbol{c}})$. Then

$$\begin{aligned} E(\boldsymbol{V}) &= 0 \\ Cov(\boldsymbol{V}) &= (\boldsymbol{Z}'\boldsymbol{Z})^{1/2}Cov(\hat{\boldsymbol{c}})(\boldsymbol{Z}'\boldsymbol{Z})^{1/2} \\ &= \sigma^2(\boldsymbol{Z}'\boldsymbol{Z})^{1/2}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}(\boldsymbol{Z}'\boldsymbol{Z})^{1/2} \\ &= \sigma^2 I, \end{aligned}$$

where I is an identity matrix of dimension $r$ by $r$ and $\boldsymbol{V}$ is normally distributed since it

consists of linear combinations of the $\hat{c}_i$. Thus,

$$\begin{aligned} \boldsymbol{V}'\boldsymbol{V} &= (\boldsymbol{c}-\hat{\boldsymbol{c}})'(\boldsymbol{Z}'\boldsymbol{Z})^{1/2}(\boldsymbol{Z}'\boldsymbol{Z})^{1/2}(\boldsymbol{c}-\hat{\boldsymbol{c}}) \\ &= (\boldsymbol{c}-\hat{\boldsymbol{c}})'(\boldsymbol{Z}'\boldsymbol{Z})(\boldsymbol{c}-\hat{\boldsymbol{c}}) \end{aligned}$$

is distributed $\sigma^2\chi^2_r$.

Again, by standard linear regression theory, we have that $\boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (n-r)s^2$ is distributed as $\sigma^2\chi^2_r$ independent of $\hat{\boldsymbol{c}}$ and hence independent of $V$. Thus, $\frac{\chi^2_r/r}{\chi^2_{n-r}/(n-r)} = \frac{\boldsymbol{V}'\boldsymbol{V}/r}{s^2}$ has an $F_{r,n-r}-$distribution and the confidence ellipsoid $(\boldsymbol{c}-\hat{\boldsymbol{c}})'(Z'Z)(\boldsymbol{c}-\hat{\boldsymbol{c}}) \leq rs^2F_{r,n-r}(\alpha)$.

We use the linear transformation, $\hat{\boldsymbol{\beta}} = \boldsymbol{\Phi}\hat{\boldsymbol{c}}$ and take $\hat{\boldsymbol{c}} = \boldsymbol{\Phi}^{-1}\hat{\boldsymbol{\beta}}$ where $\boldsymbol{\Phi}^{-1}$ is the generalized inverse of $\boldsymbol{\Phi}$. Thus,

$$\begin{aligned} (\boldsymbol{c}-\hat{\boldsymbol{c}})'(Z'Z)(\boldsymbol{c}-\hat{\boldsymbol{c}}) &= (\boldsymbol{\Phi}^{-1}\boldsymbol{\beta}-\boldsymbol{\Phi}^{-1}\hat{\boldsymbol{\beta}})'(\boldsymbol{Z}'\boldsymbol{Z})(\boldsymbol{\Phi}^{-1}\boldsymbol{\beta}-\boldsymbol{\Phi}^{-1}\hat{\boldsymbol{\beta}}) \\ &= [\boldsymbol{\Phi}^{-1}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})]'(\boldsymbol{Z}'\boldsymbol{Z})[\boldsymbol{\Phi}^{-1}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})] \\ &= (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'(\boldsymbol{\Phi}^{-1})'(\boldsymbol{Z}'\boldsymbol{Z})\boldsymbol{\Phi}^{-1}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})] \\ &= (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'[\Phi(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{\Phi}]^{-1}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}). \end{aligned}$$

Finally, we have $(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'(\boldsymbol{\Phi}^{-1})'(\boldsymbol{Z}'\boldsymbol{Z})\boldsymbol{\Phi}^{-1}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) \leq rs^2F_{r,n-r}(\alpha)$ and obtain a joint confidence region for $\hat{\boldsymbol{\beta}}$ at the quadrature points, $t_j$.

# References


[1] Atkinson, K.E. (1989), *An Introduction to Numerical Analysis*, John Wiley and Sons, New York. ISBN 0-471-5002302.

[2] Buchanan, S. (2006), "Factors Influencing the Early Marine Ecology of Juvenile Sockeye Salmon in Rivers Inlet," British Columbia. M.Sc. thesis, Simon Fraser University, B.C.

[3] Burgner, R. L. (1991), "Sockeye Salmon," in *Pacific Salmon Life Histories*, eds. Groot, C., and Margolis, L. UBC Press, Vancouver, BC, Canada.

[4] Cloern, J.E. (1991), "Tidal Stirring and Phytoplankton Bloom Dynamics in an Estuary," *Journal of Marine Research*, 49, 203-221.

[5] Craven, P. & 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, 337-402.

[6] Foskett, D.R. (1958), "The Rivers Inlet Sockeye Salmon," *Journal of the Fisheries Research Board of Canada*, 15, 867-889.

[7] Hardle, W. & Marron, J.S. (1991), "Bootstrap Simultaneous Error Bars for Nonparametric Regression," *Annals of Statistics*, 19(2), 778-796.

[8] Herwartz, H. & Xu, F. (2009), "A New Approach to Bootstrap Inference in Functional Coefficient Models," *Computational Statistics & Data Analysis*, 53(6), 2155-2167.

[9] Koenings, J.P., Geiger, H.J., & Hasbrouk, J.J. (1993), "Smolt-to-AduIt Survival Patterns of Sockeye Salmon (Oncorhynchus nerka): Effects of Smolt Length and Geographic Latitude When Entering the Sea," *Canadian Journal of Fisheries and Aquatic Science*, 50, 600-611.

[10] Liu, W., Jamshidian, M., Zhang, Y., & Donnelly, J. (2005), "Simulation-based Simultaneous Confidence Bands in Multiple Linear Regression With Predictor Variables Constrained in Intervals," *Journal of Computational and Graphical Statistics*, 14(2), 459-484.

[11] Liu, W., Lin, S. & Piegorsch, W.W. (2008), "Construction of Exact Simultaneous Confidence Bands for a Simple Linear Regression Model," *International Statistical Review*, 76(1), 39-57.

[12] Liu, W., Wynn, H.P. & Hayter, A.J. (2008), "Statistical Inferences for Linear Regression Models When the Covariates Have Functional Relationships: Polynomial Regression," *Journal of Statistical Computation and Simulation*, 78(4), 315-324.

[13] Loader, C. & Sun, J. (1997), "Robustness of Tube Formula Based Confidence Bands," *Journal of Computational and Graphical Statistics*, 6(2), 242-250.

[14] Mallin, M.A., Paerl, H.W., Rudek, J., & Bates, P.W. (1993), "Regulation of Estuarine Primary Production by Watershed Rainfall and River Flow," *Marine Ecology - Progress Series*, 93, 199-203.

[15] Mann, K. & Lazier, J. (2006), *Dynamics of Marine Ecosystems: Biological-physical Interactions in the Oceans*, Blackwell Publishing, Malden, MA.

[16] McKinnell, S.M., Wood, C.C., Rutherford, D.T., Hyatt, K.D. & Welch D.W. (2001), "The Demise of Oweekeno Lake Sockeye Salmon," *North American Journal of Fisheries Management*, 21, 774-791.

[17] Mollié, A. (1996), "Bayesian Mapping of Disease," in *Markov Chain Monte Carlo in Practice*, eds. W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, New York, Chapman & Hall.

[18] Muller, H. & Stadtmuller, U. (2005), "Generalized Functional Linear Models," *Annals of Statistics*, 33(2), 774-805,.

[19] Ramsay, J. & Silverman, (2006), *Functional Data Analysis* Springer, New York.

[20] Reynolds, C. (2006), *Ecology of Phytoplankton*, Cambridge University Press, Cambridge, UK.

[21] Rutherford, D. & Wood, C. (2000), "Assessment of Rivers and Smith Inlet Sockeye Salmon, With Commentary on Small Sockeye Salmon Stocks in Statistical Area 8," *Fisheries and Oceans Science*, Research Document.

[22] Scheffé, H. (1953), "A Method for Judging all Contrasts in Analysis of Variance," *Biometrika*, 40, 87-104.

[23] Scheffé, H. (1959), *The Analysis of Variance*, New York: Wiley.

[24] Silverman, B.W. (1985), "Some Aspects of the Spline Smoothing Approach to Non-parametric Regression Curve Fitting (with discussion)," *Journal of the Royal Statistical Society B*, 47, 1-52.

[25] Sun, J. & Loader, C. (1994), "Simultaneous Confidence Bands for Linear Regression and Smoothing," *Annals of Statistics*, 22(3), 1328-1345.

[26] Tommasi, D. (2008), "Seasonal and Inter Annual Variability of Primary and Secondary Productivity in a Coastal Fjord," M.Sc. Thesis, Simon Fraser University.

[27] Wahba, G. (1985), "A Comparison of GCV and GML for Choosing the Smoothing Parameter in the Generalized Spline Smoothing Problem," *The Annals of Statistics*, 13(4), 1378-1402.

[28] Wu, C.O., Chiang, C. & Hoover, D.R. (1998), "Asymptotic Confidence Regions for Kernel Smoothing of a Varying-Coefficient Model With Longitudinal Data," *Journal of the American Statistical Association*, 93(444), 1388-1402.

## List of Tables

1. Pointwise and joint confidence region coverage probabilities based on 10,000 simulated datasets and a smoothing parameter between $10^{-6}$ and $10^{-1}$.

## List of Figures

1. The regression coefficient function for each time frame and for several values of the smoothing parameter along with confidence regions and a plot of GCV($\lambda$). Pointwise confidence intervals are indicated with a dotted line, joint confidence regions with a solid line. Vertical lines are located at the first of each month.

2. The mean residual river flow and the effects of adding and subtracting a multiple of each principal component.

3. Coverage probabilities and GCV($\lambda$) values based on the simulation study

4. Theoretical and empirical F-ratios for no smoothing parameter. The solid line indicates the theoretical F-distribution, the dashed line is the theoretical cut point and the dotted line is the empirical cut point for a 95% confidence region.

**Table 1**: Pointwise and joint confidence region coverage probabilities based on 10,000 simulated datasets and a smoothing parameter between $10^{-6}$ and $10^{-1}$

| Smoothing Parameter | Pointwise Confidence Interval (Simultaneous Inference) | | | Joint Confidence Region | | |
|---|---|---|---|---|---|---|
| | 0.90 | 0.95 | 0.99 | 0.90 | 0.95 | 0.99 |
| degrees of freedom = $trace(A)$ | | | | | | |
| $\lambda = 10^{-4}$ | 0.89 (0.53) | 0.94 (0.71) | 0.99 (0.92) | 0.93 | 0.97 | 0.99 |
| $\lambda = 10^{-2.8}$ | 0.88 (0.60) | 0.93 (0.76) | 0.98 (0.93) | 0.87 | 0.93 | 0.98 |
| $\lambda = 10^{-1}$ | 0.85 (0.41) | 0.91 (0.61) | 0.98 (0.87) | 0.70 | 0.83 | 0.95 |
| $\lambda_{GCV}$ | 0.81 (0.35) | 0.88 (0.50) | 0.96 (0.73) | 0.77 | 0.87 | 0.97 |
| $\lambda_{MSE}$ | 0.89 (0.63) | 0.94 (0.77) | 0.98 (0.92) | 0.88 | 0.94 | 0.98 |
| degrees of freedom = trace(AA') | | | | | | |
| $\lambda = 10^{-4}$ | 0.89 (0.52) | 0.94 (0.71) | 0.99 (0.91) | 0.90 | 0.95 | 0.99 |
| $\lambda = 10^{-2.8}$ | 0.88 (0.61) | 0.93 (0.76) | 0.98 (0.93) | 0.84 | 0.92 | 0.98 |
| $\lambda = 10^{-1}$ | 0.85 (0.42) | 0.92 (0.62) | 0.98 (0.88) | 0.71 | 0.83 | 0.95 |
| $\lambda_{GCV}$ | 0.80 (0.32) | 0.87 (0.46) | 0.95 (0.71) | 0.78 | 0.88 | 0.97 |
| $\lambda_{MSE}$ | 0.89 (0.63) | 0.94 (0.76) | 0.98 (0.92) | 0.87 | 0.93 | 0.98 |

**Figure 1**: The regression coefficient function for each time frame and for several values of the smoothing parameter along with confidence regions and a plot of GCV($\lambda$). Point-wise confidence intervals are indicated with a dotted line, joint confidence regions with a solid line. Vertical lines are located at the first of each month.

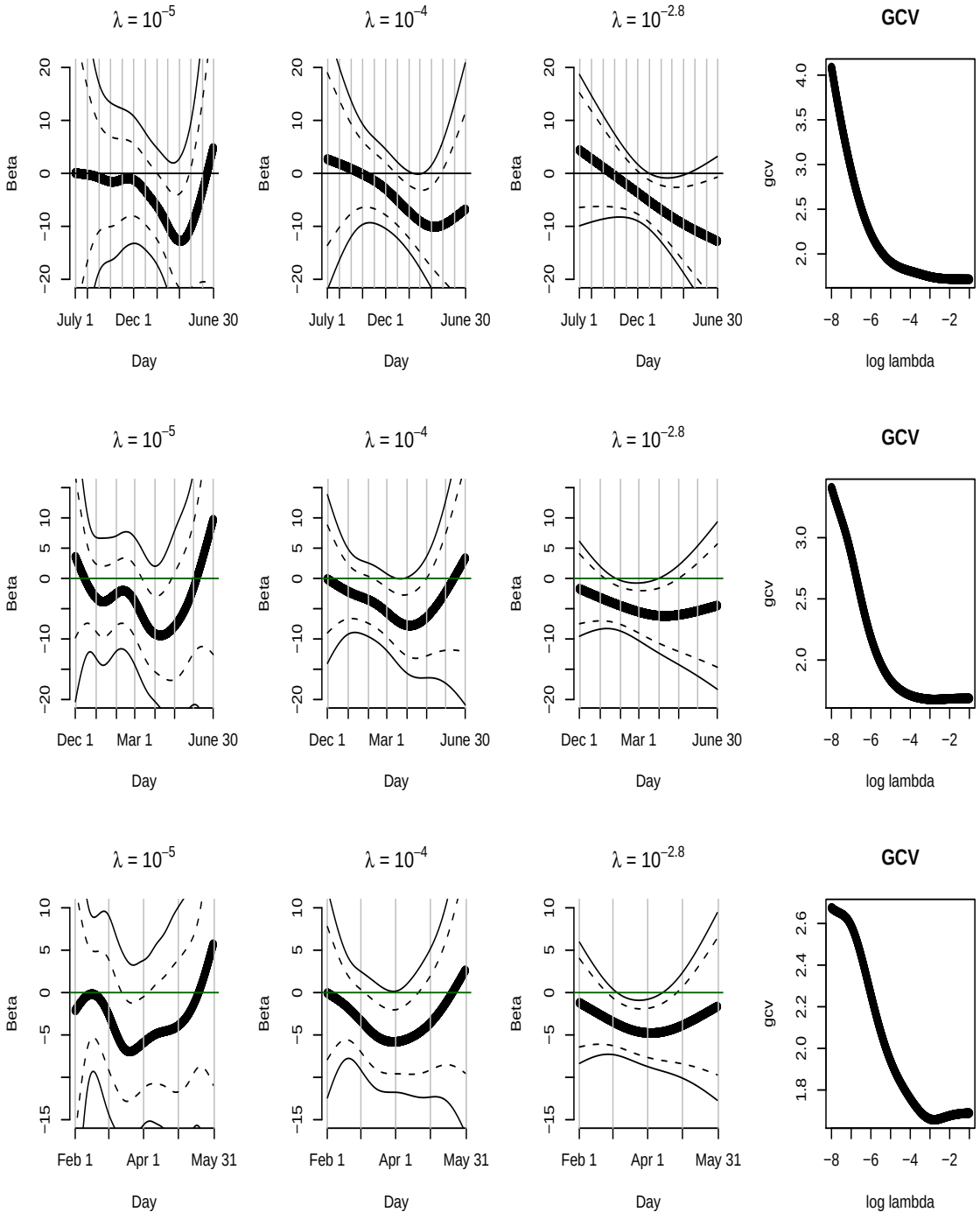

**Figure 2**: The mean residual river flow and the effects of adding and subtracting a multiple of each principal component.

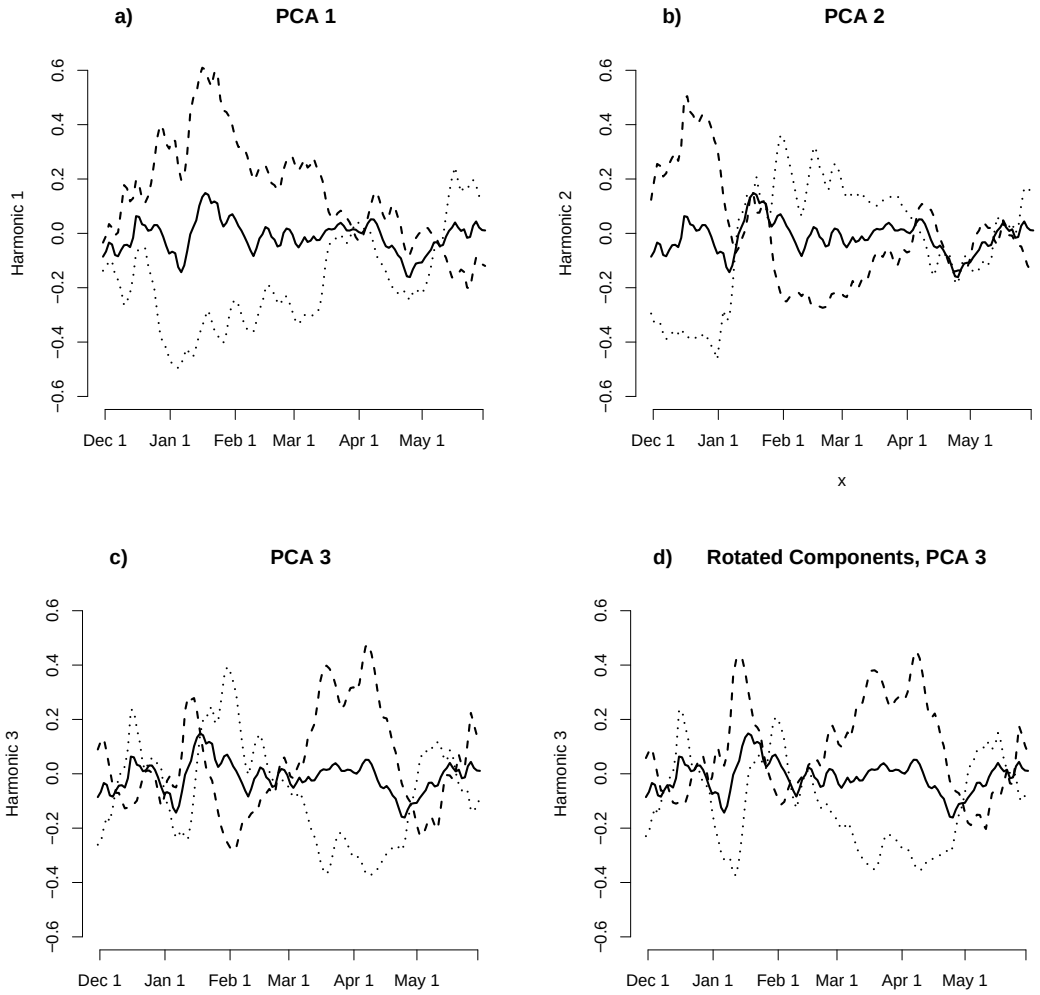

**Figure 3**: Coverage probabilities and GCV($\lambda$) values based on the simulation study

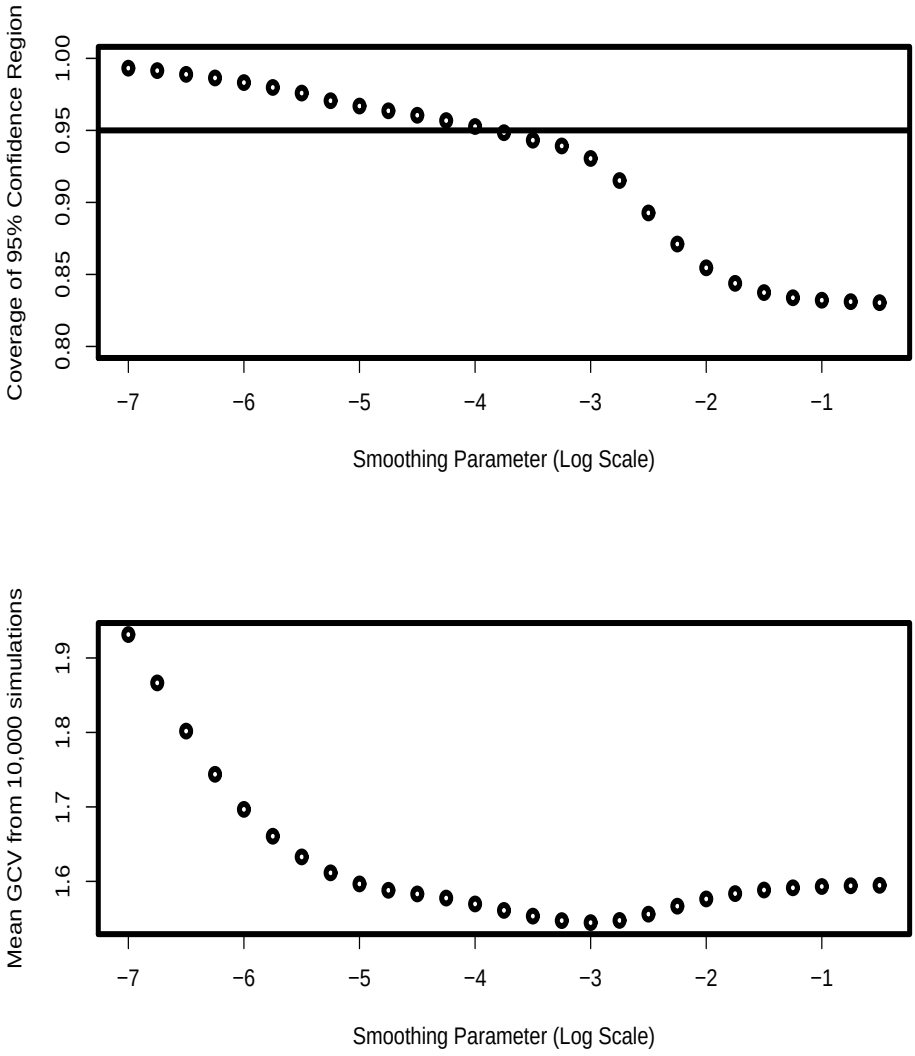

**Figure 4**: Theoretical and empirical F-ratios for no smoothing parameter. The solid line indicates the theoretical F-distribution, the dashed line is the theoretical cut point and the dotted line is the empirical cut point for a 95% confidence region.