Modeling Spatio-Temporal Trends in the Productivity of North Pacific Salmon

Oksana Chkrebtii[1], Jiguo Cao[2,*]
[1] Department of Statistics and Actuarial Science,
Simon Fraser University
[2] Department of Statistical and Actuarial Sciences,
University of Western Ontario
[*] Corresponding author, email: cao@stats.uwo.ca

Abstract

Fisheries management of North Pacific salmon stocks greatly relies on the understanding of changes in spawning and survival over time and across habitats. Underlying the yearly observed number of surviving salmon is a productivity parameter that cannot be directly measured and, moreover, is masked by short-term changes in the observed population. We model the unobserved productivity of chum, sockeye, and two broodlines of pink salmon along the Pacific Coast of North America as a smoothly-varying function of time and spatial location based on the Ricker spawner-recruit model of salmon reproduction. The candidate models belong to the class of Gaussian additive models and require the selection of smoothing parameters that control the trade-off between fit to the data and smoothness of the estimated functions. We select the smoothing parameters by optimizing the pseudo BIC criterion, which incorporates prior knowledge about the degree of smoothness of the estimated functions and is well-suited for detecting low-frequency oscillations in the data, such as those due to long term climate effects. Comparing the candidate models based on fit and model parsimony via the AIC criterion, we find that the productivity components of time and spatial location may be related nonlinearly. We find evidence of an increase in productivity in the mid-1970s for chum and sockeye populations and a North-South inverse relationship in productivity among sockeye and odd-year pink salmon stocks.

KEY WORDS: BIC; Gaussian Additive Models; Salmon stocks.

1 Introduction

The understanding of ecological processes that effect changes in the salmon population is important for fisheries management and conservation initiatives. Our interest lies in modeling spatial and temporal trends in the underlying productivity that drives the observed patterns of salmon spawning off the Pacific Coast of North America. Defined as the ratio of recruits to spawners at low spawner abundance, productivity is an ecologically important variable that describes changes in the stock that are not explained by changes in the abundance of spawners. Observational studies of productivity are restricted by the fact that this quantity cannot directly be measured and that yearly variability in salmon survival rates masks important patterns in this variable (Dorner et al. 2008).

One well-known model describing the productivity of a salmon population, or stock, was developed by Ricker (1975). The Ricker spawner-recruit model relates productivity of a salmon stock to the log ratio of recruits to spawners by the Ricker equation,

$$Y\left(t\right)=a+\delta X\left(t\right)+\varepsilon\left(t\right), \tag{1}$$

where $Y\left(t\right)$ denotes the log ratio of recruits to spawners, and $X\left(t\right)$ denotes the abundance of spawners, or parental stock in brood year t. The term $a \in \mathbb{R}$ represents productivity of the population at low spawner abundance, $\delta \in \mathbb{R}$ represents the density-dependent effect, and ε is a normally distributed error term with mean zero and unknown variance σ^2.

It is now recognized that the productivity parameter, a, varies over time and with the spatial location of the spawning site. For example, Adkison et al. (1996) identify an increase in productivity in the middle of the 1970s for Bristol Bay sockeye stocks that appears to coincide with the large-scale physical change in ocean temperatures which occurred around this time (see, for example, Graham 1994). Mantua et al. (1999) find evidence for five species of Pacific salmon examined of an inverse relationship in the catch of stocks from Alaska and stocks from the US West Coast. They assume that catch numbers are a reflection of productivity and attribute this spatial relationship to climatic changes associated with the Pacific Decadal Oscillation which occurs in cycles and in turn creates conditions that are favorable for stock productivity in the North and less favourable further South, and vice versa.

The basic Ricker equation shown in (1) describes the spawning-recruitment relationship for a particular stock with constant productivity, a, and the density-dependent effect, δ. This has been adapted in the literature to incorporate assumptions about temporal and spatial dependence of the model components. Single-stock models are concerned with changes in the parameters for individual stocks over time and include the Ricker equation itself and variations such as the inclusion of a time-dependent productivity parameter. One example of a single-stock model is proposed in Peterman et al. (2000) and models productivity as an autoregressive process of order 1. Multi-stock versions model several stocks of one species simultaneously. Simultaneous modeling relies on the reasonable assumption that parameters governing the reproductive process at one site contain information about those at a distinct site. Examples include the Ricker equation variant proposed in Su et al. (2004) with productivity and the density-dependent effect that vary with stock location. A list of these models can be found in Dorner et al. (2009).

Simultaneously estimating productivity variables over different stocks has shown some benefits over modeling single stocks (Su et al. 2004). Therefore, we incorporate spatial and temporal trends over multiple spawning sites to study underlying patterns in unobserved stock productivity, under the assumption that productivity, a, is a smooth function of time and spatial location. In this analysis we describe salmon spawning at different sites simultaneously by a multi-stock model. We assume that observations from one spawning site contain information about productivity of nearby spawning sites due to the similarities in ecology and weather conditions between nearby regions. As these factors generally change gradually in space and with time, it is reasonable to assume smoothly-varying productivity effects. Our contribution is to model the productivity parameter a as a smooth function of both time and spatial location. Our interest lies in studying changes in this unobserved productivity parameter.

[Figure 1 about here.]

We base our analysis on observations from 43 pink, 40 chum, and 37 sockeye salmon populations (stocks) off the west coast of Washington State, British Columbia, and Alaska, shown in the map in Figure 1. Each stock is identified by the entry point into the ocean of its juvenile salmon. Also available is the distance (in kilometers) along the shore of each stock's ocean entry point relative to the the southernmost stock. Thus, for each species, all stock locations have an associated along-shore distance relative to the southernmost stock of that species.

We produce separate analyses for each of chum and sockeye salmon over time and across stocks. The data for pink salmon includes both even- and odd-year runs, in which two distinct broodlines spawn (Dorner et al. 2008). Their two-year lifecycle means that one population spawns only in even-numbered years, and the other in odd-numbered years, which allows us to differentiate between the two populations based on the available data, and produce separate analyses for each.

Our paper is organized as follows. Two spatio-temporal models are proposed in Section 2, which model the unobserved productivity of chum, sockeye, and pink salmon as a smooth function of time and spatial location based on the Ricker spawner-recruit model. The functional parameters in the four spatio-temporal models are estimated with the penalized spline smoothing method. Smoothing parameters are selected by minimizing the pseudo BIC criterion (Konishi, Ando, and Imoto 2004). Section 3 compares the two spatio-temporal models, and discusses the spatial and temporal trends in spawning patterns for pink, chum and sockeye salmon. One simulation study is introduced in Section 4 to compare the pseudo BIC and GCV criteria in smoothing parameter selection. Conclusions and discussion are given in Section 5.

2 Method

2.1 Spatio-Temporal Models

Generalized additive models (Hastie and Tibshirani 1986; Wood 2006) are a class of flexible semiparametric models, which represent the mean response by a combination of smooth functions of covariates and parametric terms. They have been used extensively in many areas of applied science, and are especially applicable to problems in ecology, forestry and medicine.

An extension of generalized linear models, generalized additive models allow fitting of both nonparametric and parametric functions of covariates to the data. Under appropriate assumptions, each of the nonparametric functions is represented by a linear combination of basis functions. The model parameters are estimated by optimizing a selected criterion designed to balance smoothness of the estimated functions with fit of the model to the data. Typically, this criterion is taken to be the log likelihood of the data modified to penalize lack of smoothness. The relative importance of these two competing considerations is determined by the choice of a smoothing parameter, which may be selected systematically based on the available data. A detailed treatment of generalized additive models can be found in Wood (2006). In this analysis, we will work with the special case of Gaussian additive models (GAMs).

Let s denote one-dimensional location given by the distance along the shoreline in km from the southernmost stock location, and let t denote the brood year. Two candidate spatio-temporal models are,

$$\begin{aligned}
\text{M1}: \quad & Y(s,t) = a_0 + a_1(s) + a_2(t) + \delta_s X(s,t) + \epsilon(s,t), \\
\text{M2}: \quad & Y(s,t) = a_0 + a(s,t) + \delta_s X(s,t) + \epsilon(s,t),
\end{aligned}$$

where $Y(s,t)$ is the log ratio of the abundance of adult recruits over the abundance of spawners, and $X(s,t)$ is the abundance of spawners at spatial location s and brood year t. In model M1, $a_1(s)$ and $a_2(t)$ are smooth functions of s and t representing spatial and temporal variations in salmon productivity, respectively. In contrast, under model M2, $a(s,t)$ is a smooth function of both s and t representing spatio-temporal variation in productivity simultaneously. Model M1 is more desirable in terms of interpretability by allowing an additive separation of the spatial and temporal processes driving productivity. However, this additivity is a strong assumption and therefore we must consider model M2, where the spatial and temporal processes are linked in ways that are not necessarily additive. The parameter δ_s represents density-dependent effects for each stock, and a_0 is the intercept. We assume normally distributed errors $\epsilon(s,t)$. The existing literature models the density-dependent effect as either constant or varying with location. Arguments for assuming that δ is constant with respect to time include those of Peterman et al. (2000) who argue that time-varying density-dependent effect would have to be associated with uncharacteristically large spawner abundance $X(s,t)$ in order to lead to large-scale temporal variation in survival rates such as those observed in the 1970s, and that it is more likely that a time-varying productivity parameter is responsible. And, in fact, Adkison et al. (1996) find no evidence of a temporal change in these effects during the large shift in productivity identified in the mid-1970s. On the other hand, possible dependence of δ on spatial location is assumed by, for example, Dorner et al. (2009). Therefore, we shall assume that density-dependent effects are not time-varying, but are stock-specific.

2.2 Estimation of Spatio-Temporal Models

The models described above belong to the class of GAMs. We place some smoothness assumptions on the functional parameters, $a_1(s)$, $a_2(t)$, $a(s,t)$, and express them as linear combinations of basis functions,

$$a_1(s) = \sum_{j=1}^{q_1} b_{1j}\,(s)\,c_{1j}; \quad a_2(t) = \sum_{j=1}^{q_2} b_{2j}\,(t)\,c_{2j};$$
$$a(s,t) = \sum_{j=1}^{q_3} b_{3j}\,(s,t)\,c_{3j};$$

Choice of the basis system depends on our assumptions about the properties of the smooth functions that we wish to model. We select a univariate thin-plate regression spline basis for the smooth functions $a_1(s)$ and $a_2(t)$ in model M1. Under model M2, productivity is a bivariate function of the time t and spatial location s covariates, which are measured on different scales. Thus we choose an anisotropic tensor product basis system with thin-plate spline marginal bases, to represent the bivariate productivity function, $a(s,t)$. Optimal knot placement is a feature of the smoothing problem, so that knots are selected automatically for both types of bases. In this analysis we choose $q_1 = q_2 = 9$ and $q_3 = 24$, which are the reasonable defaults in the `mgcv` package in R. Additional details regarding the above basis functions can be found, for example, in Wood (2006).

In each case, we write the linear predictor $\hat{\mathbf{y}} = \boldsymbol{\Phi}\mathbf{c}$ as a linear combination of vectors of covariates and basis functions. The $p \times 1$ coefficient vector $\mathbf{c}$ consists of the unknown parameters in the model, and $\boldsymbol{\Phi}$ is the corresponding design matrix with p columns of covariates and basis functions evaluated at n design points. The model parameters are estimated by minimizing the negative log likelihood modified to penalize lack of smoothness,

$$-\log f\left(\mathbf{y}|\mathbf{c}\right) + \mathbf{J}\left(\hat{\mathbf{y}}|\boldsymbol{\lambda}\right). \tag{2}$$

The penalty term $\mathbf{J}\left(\hat{\mathbf{y}}|\boldsymbol{\lambda}\right)$ is a measure of the roughness of the estimated function $\hat{\mathbf{y}}$ and can be written in general as,

$$\mathbf{J}\left(\hat{\mathbf{y}}|\boldsymbol{\lambda}\right) = \mathbf{c}^{\top}\mathbf{D}\left(\boldsymbol{\lambda}\right)\mathbf{c}.$$

The block-diagonal matrix $\mathbf{D}\left(\boldsymbol{\lambda}\right)$ consists of $q_i \times q_i$ blocks corresponding to the basis coefficients of the i-th smooth function, and rows of zeros corresponding to the parametric coefficients. Details on the form of the penalty for each model are provided in the supplementary materials. The value of the smoothing parameter vector $\boldsymbol{\lambda}$ controls the tradeoff between fit to the data and smoothness of the resulting functions. The selection of the smoothing parameter vector will be discussed in the next subsection.

For our candidate models, the solution to the problem of minimizing the penalized criterion (2) turns out to be,

$$\widehat{\mathbf{c}} = \left[\mathbf{\Phi}^\top \mathbf{W}\mathbf{\Phi} + \mathbf{D}\left(\boldsymbol{\lambda}\right)\right]^{-1} \mathbf{\Phi}^\top \mathbf{W}\mathbf{y},$$

where $\mathbf{W}$ is the inverse covariance matrix of the data, and in this case, $\mathbf{W} = \sigma^{-2}\mathbf{I}$. Given an optimally chosen value of $\boldsymbol{\lambda}$, we fit the candidate models using the `mgcv` package in R. Details are available in Wood (2006).

2.3 Smoothing Parameter Selection

The value of the smoothing parameter $\boldsymbol{\lambda}$ which controls the tradeoff between fit to the data and smoothness of the resulting function must be specified by the analyst. The smoothing parameter may be chosen visually or systematically. In general, systematic methods are based on optimization with respect to $\boldsymbol{\lambda}$ of a criterion measuring features of the fitted model that we deem to be important in a particular context.

Accordingly, a commonly used method for selecting the smoothing parameter is minimization with respect to $\boldsymbol{\lambda}$ of the generalized cross-validation (GCV) criterion,

$$\mathrm{GCV}\left(\boldsymbol{\lambda}\right) = \frac{n||\mathbf{y} - \hat{\mathbf{y}}||^2}{\left[n - \mathrm{edf}\right]^2},$$

where

$$\mathrm{edf} = \mathrm{tr}(\mathbf{\Phi}\left[\mathbf{\Phi}^\top \boldsymbol{W}\mathbf{\Phi} + \mathbf{D}\left(\boldsymbol{\lambda}\right)\right]^{-1} \mathbf{\Phi}^\top \boldsymbol{W}) \tag{3}$$

is the effective degrees of freedom, or the effective number of parameters, of the model. We use the R package `mgcv` to estimate the minimizer $\boldsymbol{\lambda}$ of GCV by numerical optimization.

[Figure 2 about here.]

The left panel of Figure 2 shows the univariate smooth function of time $a_1(t)$ for sockeye salmon stocks estimated under model M1 with smoothing parameter vector chosen by GCV. The resulting estimated model appears to undersmooth the spatial functional parameter, potentially hindering our ability to identify features of interest. Indeed, this is a common problem with GCV-based smoothing parameter selection and in such cases, it is common to upweight the effective model degrees of freedom in the GCV calculation by an adjustment factor (Chambers and Hastie 1992). As this appears somewhat arbitrary, we recommend instead using a Bayes Information Criterion (BIC, Schwarz 1978) to choose $\boldsymbol{\lambda}$ by including some prior knowledge about smoothness of the functional parameters in the automatic selection of the smoothing parameter. The BIC is the posterior probability of a model given the data, and is widely used as a tool for model selection. However, difficulty arises in calculating BIC when choosing among functional models indexed by a smoothing parameter vector under partially improper prior distributions on the functional parameters. In this case, BIC may be estimated up to a constant of proportionality (Kass 1993). This approximation has been called the pseudo Bayes Information Criterion (pBIC, Konishi, Ando, and Imoto 2004) in the literature and is given by,

$$\begin{aligned}\text{pBIC} &= n\log\left(2\pi n\right)+n\log\left(\hat{\sigma}^2\right)+\hat{\sigma}^{-2}\left(\mathbf{y}-\boldsymbol{\Phi}\widehat{\mathbf{c}}\right)^\top\left(\mathbf{y}-\boldsymbol{\Phi}\widehat{\mathbf{c}}\right)\\ &\quad+\left(p-d\right)\log\left(2\pi\right)-\log\left|\mathbf{D}\left(\boldsymbol{\lambda}\right)\right|_{+}+\widehat{\mathbf{c}}^\top\mathbf{D}\left(\boldsymbol{\lambda}\right)\widehat{\mathbf{c}}\\ &\quad-p\log\left(2\pi\right)+p\log n+\log\left|H_{\boldsymbol{\lambda}}\left(\widehat{\mathbf{c}}\right)\right|+O_p\left(n^{-1}\right)\end{aligned}\tag{4}$$

where $|\mathbf{D}(\boldsymbol{\lambda})|_{+}$ is the product of the $p-d$ non-zero eigenvalues of $\mathbf{D}(\boldsymbol{\lambda})$, and $\mathbf{H}_{\boldsymbol{\lambda}}(\widehat{\mathbf{c}})=\frac{1}{n\hat{\sigma}^2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}+\frac{1}{2n}\left[\mathbf{D}^\top(\boldsymbol{\lambda})+\mathbf{D}(\boldsymbol{\lambda})\right]$. We take $\widehat{\mathbf{c}}=\left[\boldsymbol{\Phi}^\top\mathbf{W}\boldsymbol{\Phi}+\mathbf{D}(\boldsymbol{\lambda})\right]^{-1}\boldsymbol{\Phi}^\top\mathbf{W}\mathbf{y}$

by following the argument of Konishi et al. (2004) for models estimated via penalized negative log-likelihood minimization, and the maximum likelihood error term variance estimate $\hat{\sigma}^2 = ||\mathbf{y} - \hat{\mathbf{y}}||^2/n$. We then evaluate and minimize pBIC with respect to $\boldsymbol{\lambda}$ via grid search using the R software.

This criterion is well-suited to analyses of data with low-frequency oscillations, such as those due to climate cycles, that may be obscured by higher-frequency year-to-year variability. For example, a simulation study in Konishi et al. (2004) suggests that smoothing parameters selected by minimizing the GCV score may undersmooth the data compared to those selected by minimizing pBIC. We observe such a pattern in our analysis. The right panel of Figure 2 shows the smooth function $a_1(t)$ estiamted for sockeye salmon stocks under model M1 with $\boldsymbol{\lambda}$ chosen by minimizing pBIC over a grid of smoothing parameter values. This function appears smoother and allows us to observe an increase in productivity in the mid-1970s coinciding with a warming climate in the North Pacific.

Details on the derivation of the approximation in (4) are provided in the supplementary materials for Gaussian additive models with a normally distributed error. This derivation may be more useful for practicioners than the general formula provided by Konishi et al. (2004) for the large class of radial basis function network models under an exponential link function, which is very heavy in notation and difficult to implement for a specific case.

2.4 Estimation of Pseudo BIC

The selection of the smoothing parameter from a grid of possible values can be restated as a problem of model selection. For a given GAM with coefficients $\mathbf{c}$, we denote M $(\boldsymbol{\lambda})$ to be the model indexed by any given smoothing parameter vector $\boldsymbol{\lambda}$ associated with pdf $f(\mathbf{y}|\mathbf{c},\boldsymbol{\lambda})$. The posterior probability of the model given observations $\mathbf{y}$ (see, for example, Raftery 1996) is,

$$P\left(\mathrm{M}\left(\boldsymbol{\lambda}\right)|\mathbf{y}\right)=\frac{P\left(\mathrm{M}\left(\boldsymbol{\lambda}\right)\right)f\left(\mathbf{y}|\boldsymbol{\lambda}\right)}{\sum_{\alpha=1}^{r}P\left(\mathrm{M}\left(\alpha\right)\right)f\left(\mathbf{y}|\boldsymbol{\lambda}_{\alpha}\right)}, \tag{5}$$

where $P\left(\mathrm{M}\left(\cdot\right)\right)$ represents the prior probability of model $M\left(\cdot\right)$. We would like to select among $\mathrm{M}\left(\boldsymbol{\lambda}_1\right),\ldots,\mathrm{M}\left(\boldsymbol{\lambda}_r\right)$ the model with the largest posterior probability by maximizing the numerator of (5) by choice of $\boldsymbol{\lambda}$. If all models are a priori equally probable, this is equivalent to minimizing the pseudo BIC,

$$\mathrm{pBIC} \quad \equiv \quad -2\log f\left(\mathbf{y}|\boldsymbol{\lambda}\right)$$

for model M by choice of smoothing parameter $\boldsymbol{\lambda}$. Assumptions and derivation of an approximate expression for the logarithm of this objective function are provided in the supplementary materials following Konishi et al. 2004. The resulting expression is a function of the maximizer $\widehat{\mathbf{c}}$ of (2) and the maximum likelihood estimate $\hat{\sigma}^2$ of the error term variance, both of which can be obtained directly from the model fitting procedure. Note that, although we may use this result to choose among r smoothing parameters for a particular model, it is not possible to compare non-nested models using the pBIC when priors are improper. This fact is discussed , for example, in Carlin and Louis (2000).

3 Results

3.1 Model Selection

We would like to compare models M1 and M2 in terms of their fit to the data and the effective number of parameters used in their estimation. As discussed previously, model M1 is more easily interpretable than model M2, as it represents spatial and temporal effects additively. But the assumption that these effects are indeed related in an additive way is quite strong. In order to determine whether or not such an assumption is justified, we shall use the Akaike information criterion (AIC) to compare the models. AIC, which was first published in Akaike (1974), evaluates the fit of the model while penalizing models with a large number of effective degrees of freedom. It is defined as

$$\mathrm{AIC} = 2 \cdot \mathrm{edf} - 2 \cdot \log f\left(\mathbf{y}|\hat{\mathbf{c}}\right),$$

where edf is the effective number of parameters in the model, which is calculated in (3), and $f\left(\mathbf{y}|\hat{\mathbf{c}}\right)$ is the maximized value of the likelihood function for the estimated model.

AIC values for each model are provided in the supplementary materials. The comparison reveals that, for each of the observed salmon species, time and spatial productivity effects have an underlying nonlinear relationship. This suggests that complex underlying mechanisms drive the observed productivity, and that additional unmodeled variables that vary over time and location may explain some of the observed variability in the log ratio of recruits to spawners. To explore this further, model M2 residuals are plotted against spatial location and time in the supplementary materials. For all four populations, the residual mean does not appear to change systematically with spatial location, while the nonconstant variance is suggestive of unmodeld changes in spatial variability of the log ratio or recrutis to spaners, Y. Over time, the residual variance for all but the odd-year runs of pink salmon is approximatley constant and the mean does not deviate systematically from zero. For odd-year runs of pink salmon, the residuals show some variability over time which has not been captured by the model. In this case, the unmodeled pattern suggests an increase in log ratio of recruits to spawners starting in the mid-1970s when the North Pacific experienced a climatic warming event.

3.2 Spatial and Temporal Trends in Spawning Patterns of Salmon

[Table 1 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

Table 1 shows results of hypothesis tests for the significance of the functional parameters in Models M1 and M2 for chum, sockeye, and even- and odd-year runs of pink salmon. This corresponds to testing the subset of coefficients $\mathbf{b}_i$ of each smooth function for equality with zero. Under the null hypothesis and with unknown error term variances, the test statistic $\hat{\mathbf{b}}_i \hat{\boldsymbol{V}}_{\boldsymbol{(i)}} \hat{\mathbf{b}}_i / r$ has an approximate $F_{r,n-edf}$ distribution (Wood 2006, p. 194), where $\hat{\mathbf{b}}_i$ is the estimated subset of coefficients, and $\hat{\boldsymbol{V}}_{\boldsymbol{(i)}}$ is the pseudoinverse of $\hat{\gamma}^2 \left[\mathbf{\Phi}_{\boldsymbol{(i)}}{}^\top \mathbf{W} \mathbf{\Phi}_{\boldsymbol{(i)}} + \mathbf{D}_{(\mathbf{i})} \left(\boldsymbol{\lambda} \right) \right]^{-1}$ of rank $r = \text{rank} \left(\left[\mathbf{\Phi}_{\boldsymbol{(i)}}{}^\top \mathbf{W} \mathbf{\Phi}_{\boldsymbol{(i)}} + \mathbf{D}_{(\mathbf{i})} \left(\boldsymbol{\lambda} \right) \right]^{-1} \right)$, and where $\hat{\gamma}^2 = ||\mathbf{y} - \hat{\mathbf{y}}||^2 / (n - \text{edf})$.

Under model M1, the time-dependent function of productivity, $a_2(t)$, is found to be significant for all four groups of salmon. The location-dependent smooth function of productivity, $a_1(s)$, is significant for sockeye stocks. Therefore, we may conclude that along-shore distance explains a significant amount of the variation in observed log ratio of recruits to spawners for this population. Figures 3 and 4 show the estimated functional parameters, $a_1(s)$ and $a_2(t)$ and their confidence intervals under model M1, based on data for each group of salmon. The estimated spatial productivity function for sockeye salmon decreases linearly from South to North. The time-dependent function of productivity, $a_2(t)$, oscillates between positive and negative values. For the two pink salmon species, we note cyclic patterns in this estimated function with amplitude decreasing over time. Among all four groups, the shapes of the estimated smooth functions of time show evidence of an increase in productivity in the mid 1970s consistent with that found, for example, by Adkison et al. (1996) for Alaskan sockeye salmon and attributed to large-scale climatic changes during that period.

[Figure 6 about here.]

[Figure 7 about here.]

Under model M2, we find that the smooth bivariate function of time and spatial location is significant for all four populations studied. Figures 6 and 7 show the estimated bivariate functional parameter, $a(s,t)$ in model M2, estimated from the data for four groups of salmon. Although interpretation of the fitted function is less clear, some interesting patterns are observed. An increase in productivity beginning in the mid-1970s is observed clearly for sockeye and chum stocks, the latter being more pronounced at middle latitudes. For odd-year pink salmon, this increase is more gradual and confined to the mid-to-northern latitudes. Productivity of odd-year pink salmon decreases with time for middle and northern latitude stocks, but remains steady over time for southern latitudes. It is important to note, however, that the very sharp decrease observed in the productivity funtion for even-year pink salmon may be the result of endpoint inaccuracy of the nonparameteric estimator. Only for sockeye salmon stocks do we find evidence of an inverse relationship in productivity between northern and southern stocks, such as that observed in Mantua et al. (1999).

4 Simulations

In the present analysis we select the smoothing parameter by optimization of the pBIC criterion, in order to aid interpretability of the estimated functions. Analysis based on the GCV criterion is also provided in the supplementary materials for comparison. In general, smoothing parameter selection is an unresolved problem (Konishi and Kitagawa 2008), and the relative performance of these two criteria may vary with the smoothness of the underlying function.

We conducted one simulation study to compare the two smoothing parameter selection criteria, pBIC and GCV, for the case when the underlying function exibits low-frequency cyclical variation. Model M2 $Y(s,t) = a_0 + a(s,t) + \delta_s X(s,t) + \epsilon(s,t)$ is selected as the true model. The true values for a_0, $a(s,t)$, and δ_s are set as their corresponding estimate from the real data of odd-year runs of pink salmon stocks. The independent and identically distributed random errors, $\epsilon(s,t)$, are generated from the normal distribution with mean 0 and the same variance as estimated from the read data. The data locations, (s,t), are set as the same with the real data. For each simulated data set, we selected smoothing parameters by minimizing either pBIC or GCV and estimated model M2 in each case. The simulation is implemented with 100 replicates.

The accurancy of estimated parameters for Model M2 is evaluated with two criteria: the mean squared error (MSE) of the fitted values of the response variable to their true values,

$$MSE(\lambda) = \frac{1}{N}\sum_{i=1}^{N}\Big[\big(\hat{a}_{0,\lambda} + \hat{a}_{\lambda}(s_i, t_i) + \hat{\delta}_{s\lambda}X_i\big) - \big(a_0 + a(s_i, t_i) + \delta_s X_i\big)\Big]^2,$$

and the median absolute deviation (MAD) of the fitted values of the response variable to their true values,

$$MAD(\lambda) = \text{median}\Big|\big(\hat{a}_{0,\lambda} + \hat{a}_{\lambda}(s_i, t_i) + \hat{\delta}_{s\lambda}X_i\big) - \big(a_0 + a(s_i, t_i) + \delta_s X_i\big)\Big|,$$

calculated over a grid of equally-spaced 16 points for each of the available 40 spawning sites ($N = 40 \times 16$). Box plots of MSE and MAD values in 100 simulaion replicates are shown in figure (5). As expected, the two criteria perform similarly in terms of these two measures of fit when data is generated from an underlying function with low-frequency signal structure.

The same simulation set-up was used to study model choice via the AIC for data simulated from model M2, as estimated for odd-year runs of pink salmon. Models M1 and M2 were estimated from the 100 simulated data sets using pBIC and their AIC scores were compared. The result of this simulation suggestes that in this setting, AIC selects the correct model M2 over M1 in each case.

5 Conclusions and Discussion

The goal of this work is to estimate the unobserved North Pacific salmon productivity based on the classic framework developed by Ricker (1975). Our contribution is to model spatial and temporal components of salmon productivity as smooth functions of time and spatial location through Gaussian additive modeling. An inherent feature of this class of model is the necessity of selecting appropriate smoothing parameters, which control the trade-off between smoothness of the estimated functions and fit to the data. We suggest using the pseudo BIC objective function, instead of the commonly used GCV criterion, to choose the optimal values of smoothing parameters. In our analysis, the smoothing parameters selected by the pseduo BIC result in more informative estimated functions that do not appear to undersmooth the data.

We study four populations of North Pacific salmon: chum, sockeye and pink salmon from even- and odd-year runs. For each species, we fit two candidate spatio-temporal models to the data, controlling the smoothness of functional parameters by choosing smoothing parameters that minimize the pseudo BIC within each model. We compare the two models based on their relative AIC scores, which measure the fit of models to the data through the negative log likelihood penalized by the effective number of parameters. Model M2 is found to have the lowest AIC scores among the two condidates for each of the four groups of salmon being studied, indicating that productivity is a nonlinear function of time and spatial location.

Inference from sparse data is a limitation of the present analysis. Spatial and temporal distributions of observations for the four groups of salmon are summarized in the design plots provided in the supplementary materials. We chose to include all available data in our analysis in order to be able to estimate density parameters for all stock-stream combinations. It is important to note, however, that systematic considerations are involved in the selection and time of sampling sites in many ecological studies. Since a reliable model of this complex sampling process is beyond the scope of our analysis, we instead make the simplifying assumption that the measured stocks are selected randomly from all existing stocks and that any measurments missing over time are missing at random.

Our analysis reveals that a significant portion of the observed variability in the log ratio of recruits to spawners can indeed be explained by changes in productivity over time and spatial location. A smooth function of productivity varying with time and spatial location is found to be significant for all four populations of salmon studied. This function is generally smooth along the spatial dimension, but shows oscillatory patterns over time, some of which correspond to productivity-related events identified in the literature.

In particular, our results suggest that salmon productivity for the stocks under investigation may be related to long-term climatic changes in the North Pacific. Existing literature on salmon spawning suggests that specific environmental covariates contain information on salmon spawning patterns above what is contained in time and spatial location. For example, Mueter et al. (2002) find evidence of an inverse correlation between sea surface temperatures and survival rates for the species of salmon studied here. For this reason, a natural extension of the models considered here would incorporate additional covariates.

Acknowledgements

The authors would like to thank Prof. Peter Guttorp, the associate editor, and two reviewers for their very constructive suggestions and comments, which were extremely helpful for improving this paper. They would also like to thank Dr. Randall Peterman for providing the data used in this analysis and Dr. David Campbell for his helpful suggestions. This research was supported by a discovery grant (PIN: 328256) from the Natural Science and Engineering Research Council of Canada (NSERC).

References

Adkison, M. D., R. M. Peterman, M. F. Lapointe, D. M. Gillis, and J. Korman (1996). Alternative models of climatic effects on sockeye salmon, oncorhynchus nerka, productivity in bristol bay, alaska, and the fraser river, british columbia. *Fisheries Oceanography 5*(3-4), 137–152.

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

Carlin, B. P. and T. A. Louis (2000). *Bayes and empirical Bayes methods for data analysis*. Boca Raton, Fla: Chapman and Hall/CRC.

Chambers, J. M. and T. J. Hastie (1992). *Statistical models in S*. Boca Raton, Fla: Wadsworth and Brooks/Cole Advanced Books and Software.

Dorner, B., R. M. Peterman, and S. L. Haeseker (2008). Historical trends in productivity of 120 pacific pink, chum, and sockeye salmon stocks reconstructed by using a kalman filter. *Canadian Journal of Fisheries and Aquatic Sciences 65*(9), 1842 – 1866.

Dorner, B., R. M. Peterman, and Z. Su (2009). Evaluation of performance of alternative management models of pacific salmon (oncorhynchus spp.) in the presence of climatic change and outcome uncertainty using monte carlo simulations. *Canadian Journal of Fisheries and Aquatic Sciences 66*(12), 2199–2221.

Graham, N. E. (1994). Decadal-scale climate variability in the tropical and north pacific during the 1970s and 1980s: observations and model results. *Climate Dynamics 10*, 135–162.

Hastie, T. and R. Tibshirani (1986). Generalized additive models. *Statistical Science 1*(3), 297–310.

Kass, R. E. (1993). Bayes factors in practice. *Journal of the Royal Statistical Society, Series D 42*(5), 551–560.

Konishi, S., T. Ando, and S. Imoto (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika 91*(1), 27–43.

Konishi, S. and G. Kitagawa (2008). *Information Criteria and Statistical Modeling.* Springer.

Mantua, N. J., R. C. Francis, and S. R. Hare (1999). Inverse production regimes: Alaska and west coast pacific salmon. *Fisheries 24*(1), 6–14.

Mueter, F. J., R. M. Peterman, and B. J. Pyper (2002). Opposite effects of ocean temperature on survival rates of 120 stocks of pacific salmon (oncorhynchus spp.) in northern and southern areas. *Canadian Journal of Fisheries and Aquatic Sciences 59*(3), 456.

Peterman, R. M., B. J. Pyper, and J. A. Grout (2000). Comparison of parameter estimation methods for detecting climate-induced changes in productivity of pacific salmon (oncorhynchus spp.). *Canadian Journal of Fisheries and Aquatic Sciences 57*(1), 181–191.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika 83*(2), 251–266.

Ricker, W. E. (1975). *Computation and interpretation of biological statistics of fish populations.* Ottawa, ON: Dept. of the Environment, Fisheries and Marine Service.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*(2), 461–464.

Su, Z., R. M. Peterman, and S. L. Haeseker (2004). Spatial hierarchical bayesian models for stock-recruitment analysis of pink salmon (oncorhynchus gorbuscha). *Canadian Journal of Fisheries and Aquatic Sciences 61*(12), 2471 – 2486.

Wood, S. N. (2006). *Generalized additive models : an introduction with R*. Boca Raton, FL: Chapman and Hall.

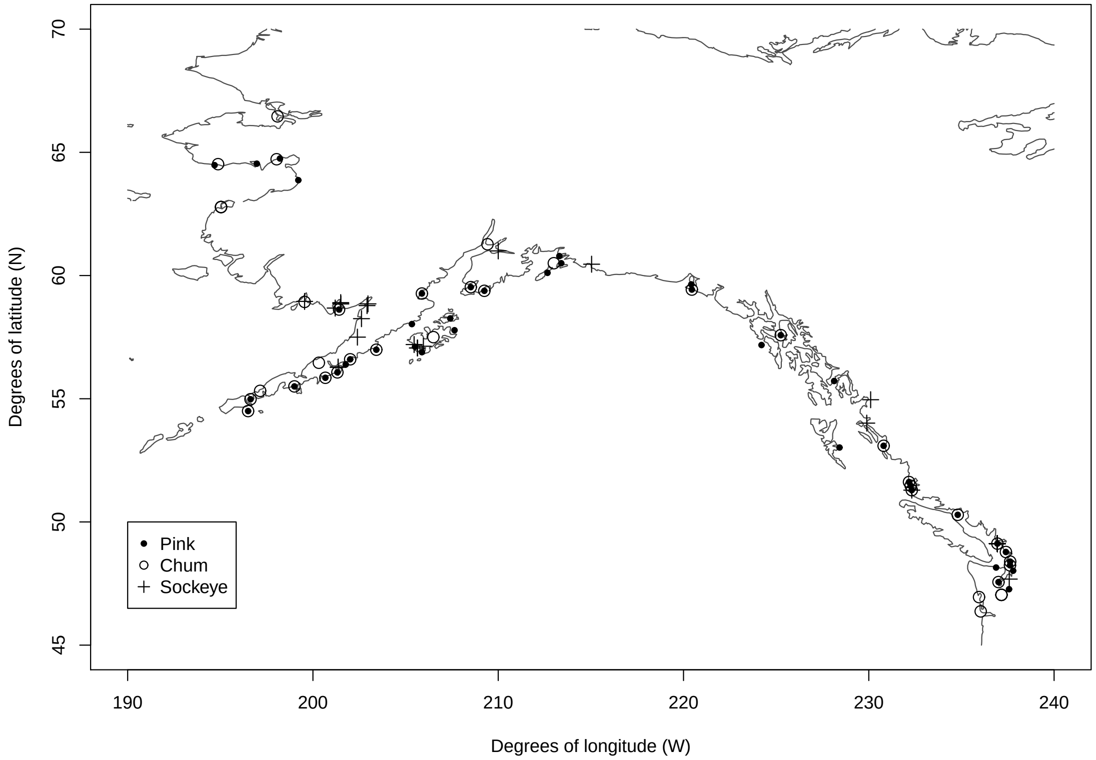

Figure 1: Sampled spawning sites for pink, chum, and sockeye salmon along the North Pacific coast.

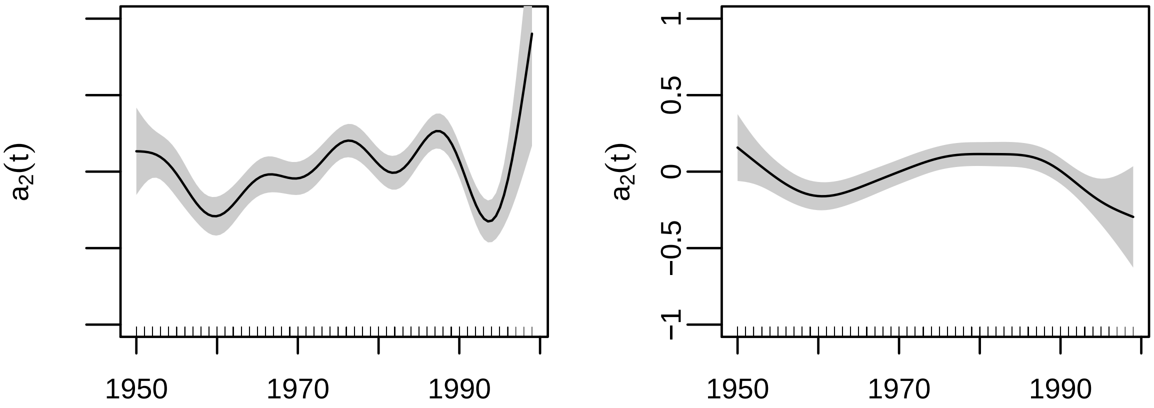

Figure 2: Estimated functional parameter $a_2(s)$ and its confidence intervals for model M1: $Y(s,t) = a_0 + a_1(s) + a_2(t) + \delta_s X(s,t) + \epsilon(s,t)$ from the data of sockeye salmon. The smoothing parameter vector is selected by minimizing GCV (left) and pseudo BIC (right).

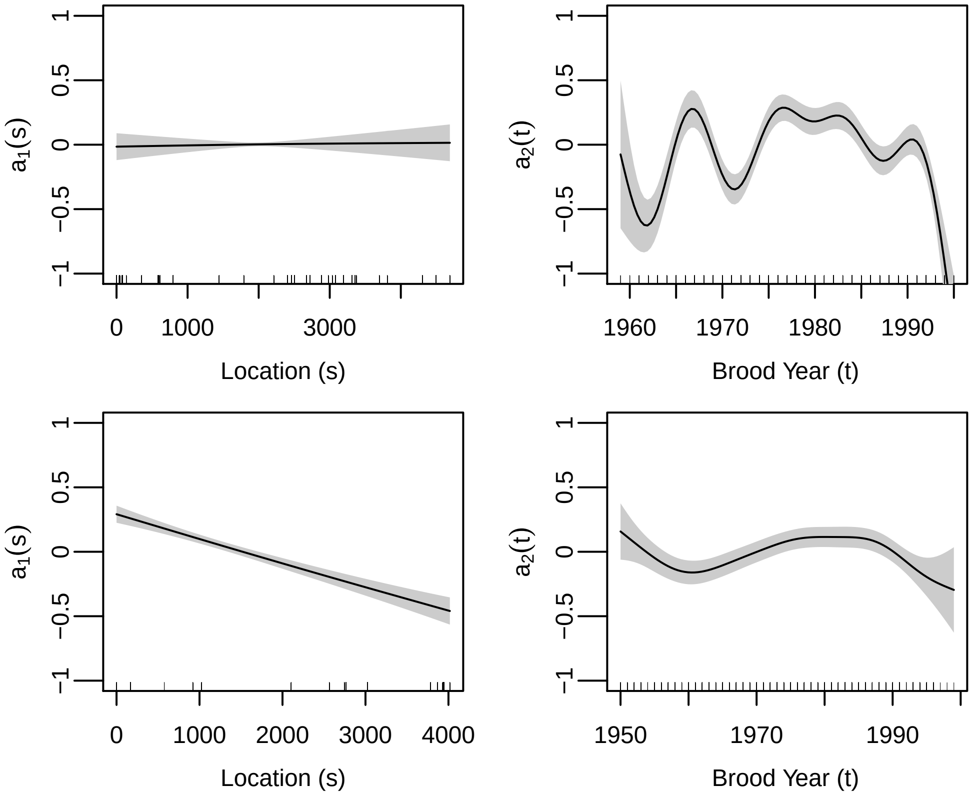

Figure 3: Estimated functional parameters, $a_1(s)$ and $a_2(t)$, and their confidence intervals in model M1: $Y(s,t) = a_0 + a_1(s) + a_2(t) + \delta_s X(s,t) + \epsilon(s,t)$ from the data of chum (top), and sockeye (bottom) salmon. The smoothing parameter vector is selected by minimizing pseudo BIC.

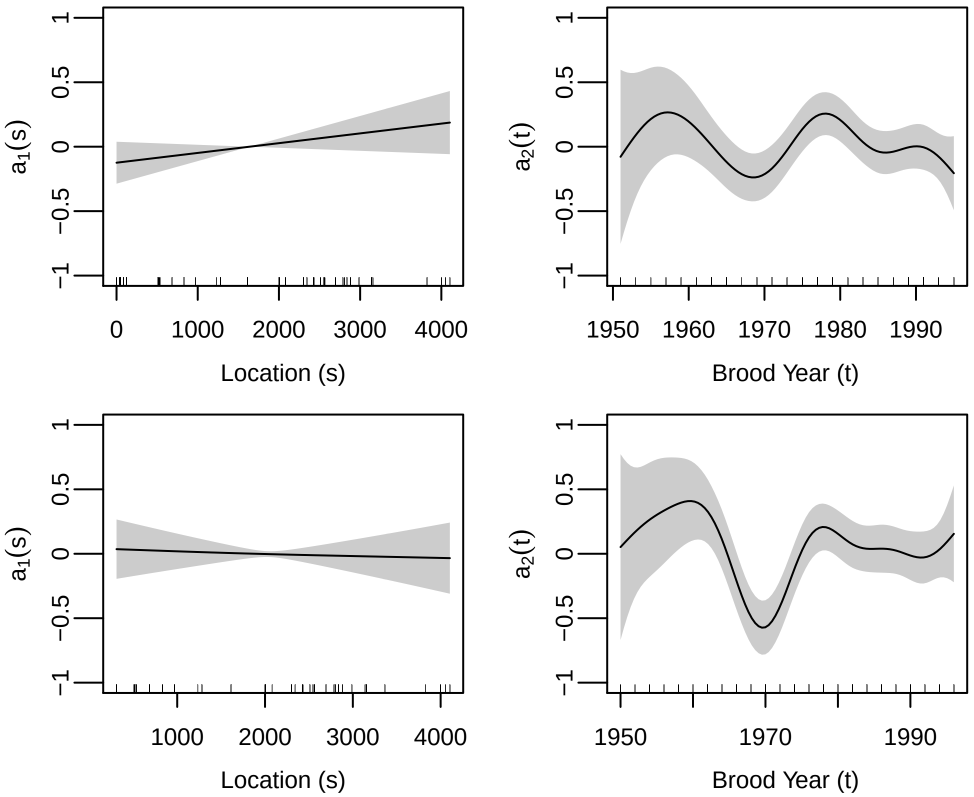

Figure 4: Estimated functional parameters, $a_1(s)$ and $a_2(t)$, and their confidence intervals in model M1: $Y(s,t) = a_0 + a_1(s) + a_2(t) + \delta_s X(s,t) + \epsilon(s,t)$ from the data of odd-year runs (top) and even-year runs (bottom) of pink salmon. The smoothing parameter vector is selected by minimizing pseudo BIC.

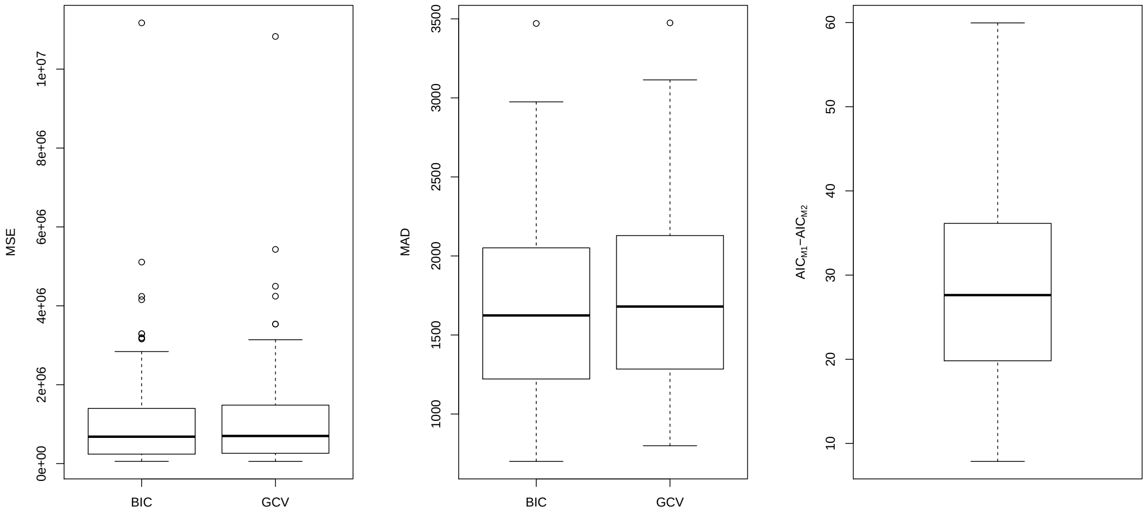

Figure 5: Simulation results for choice of smoothing parameter are shown on the left and centre: the models with smoothing parameters selected via pBIC and GCV perform similarly in terms of MSE and MAD. Simulation results for model choice via the AIC are shown on the right: positive differences in AIC sores between models M1 and M2 determine that M2 is selected over M1 in every case.

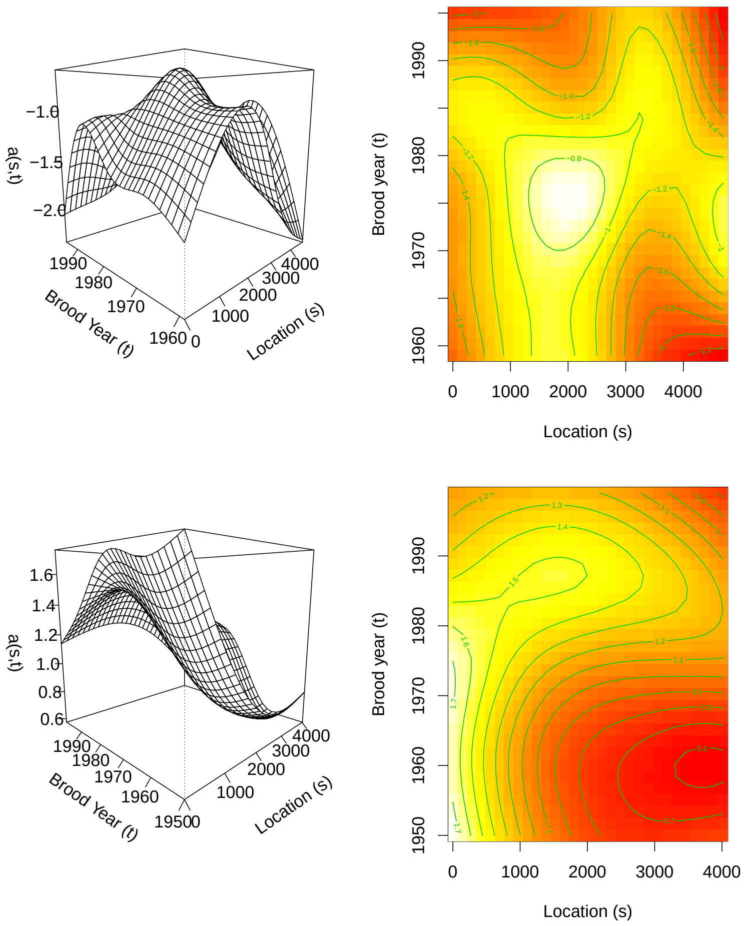

Figure 6: Estimated bivariate functional parameter, $a(s,t)$, in model M2: $Y(s,t) = a_0 + a(s,t) + \delta_s X(s,t) + \epsilon(s,t)$ from the data of chum (top), and sockeye (bottom) salmon. The smoothing parameter vector is selected by minimizing pseudo BIC.

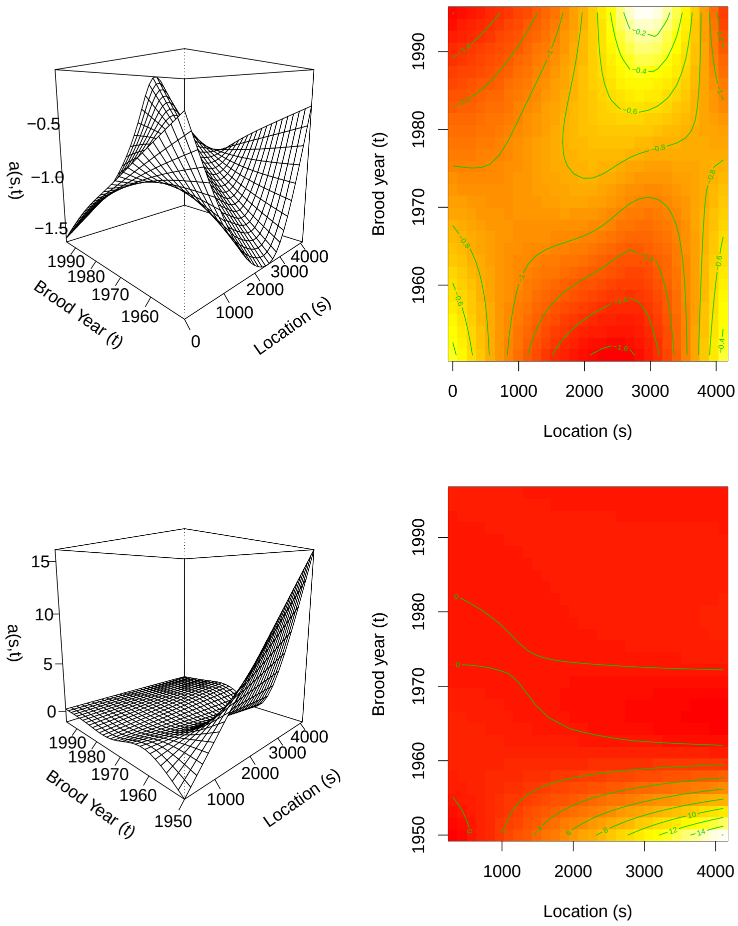

Figure 7: Estimated bivariate functional parameter, $a(s,t)$, in model M2: $Y(s,t) = a_0 + a(s,t) + \delta_s X(s,t) + \epsilon(s,t)$ from the data of odd-year runs (top) and even-year runs (bottom) of pink salmon. The smoothing parameter vector is selected by minimizing pseudo BIC.

Table 1: Statistical tests for the significance of functional parameters in models M1: $Y(s,t) = a_0 + a_1(s) + a_2(t) + \delta_s X(s,t) + \epsilon(s,t)$ and M2: $Y(s,t) = a_0 + a(s,t) + \delta_s X(s,t) + \epsilon(s,t)$ from the data of chum, sockeye, odd-year pink, and even-year pink salmon stocks. "edf" represents the effective degrees of freedom, or the effective number of parameters, of the model, which is defined in (3).

		pink	pink		
Parameters		(odd-year)	(even-year)	chum	sockeye
Model M1					
$a(s)$	edf	1 .0	1.0	1.0	1.0
	F statistic	2.3	0.063	0.049	74
	p-value	1e-1	8e-1	8e-1	3e-18
$a(t)$	edf	5.9	6.9	8.5	4.8
	F statistic	2.4	4.4	16	4.4
	p-value	2e-2	4e-5	5e-24	2e-4
a_0	Estimate	1.3	1.4	1.5	1.7
	Std. Error	0.065	0.075	0.041	0.034
	t statistic	20	19	35	48
	p-value	1e-68	2e-61	<1e-100	<1e-100
Model M2					
$a(s,t)$	edf	8.5	9.4	17	11
	F statistic	3.7	5.2	7.1	12
	p-value	4e-5	4e-7	1e-18	4e-26
a_0	Estimate	1.4	1.4	1.5	1.7
	Std. Error	0.066	0.075	0.041	0.035
	t statistic	21	19	37	48
	p-value	9e-71	1e-61	<1e-100	<1e-100