LOCAL PARAMETRIC POISSON MODELS FOR FISHERIES DATA

by

IRENE MEI LING YEE
B.Sc., THE UNIVERSITY OF BRITISH COLUMBIA, 1986

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

Department of Statistics

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

September, 1988

Department of Statistics

The University of British Columbia
Vancouver, Canada

Date Sept. 29, 1988

ABSTRACT

Poisson process is a common model for count data. However, a global Poisson model is inadequate for sparse data such as the marked salmon recovery data that have huge extraneous variations and noise. An empirical Bayes model, which enables information to be aggregated to overcome the lack of information from data in individual cells, is thus developed to handle these data. The method fits a local parametric Poisson model to describe the variation at each sampling period and incorporates this approach with a conventional local smoothing technique to remove noise. Finally, the overdispersion relative to the Poisson model is modelled by mixing these locally smoothed, Poisson models in an appropriate way. This method is then applied to the marked salmon data to obtain the overall patterns and the corresponding credibility intervals for the underlying trend in the data.

TABLE OF CONTENTS

LIST OF TABLES

TABLE

LIST OF FIGURES

FIGURE

FIGURE

FIGURE

FIGURE

ACKNOWLEDGEMENTS

I would like to thank Dr. James Zidek for his guidance, assistance and encouragement in producing this thesis. I would like to express my gratitude to Dr. Mohan Delampady for his helpful comments and careful reading on this work. I am indebted to Dr. Jon Schnute from the Pacific Biological Station for providing the data sets and offering advice related to these data. Also I am grateful to technical staff members at the Station, in particular, Brian Kuhn for producing ancillary data sets when their importance emerged in the course of the analysis. I am very grateful for the comments and suggestions from Dr. Keith Knight and Dr. Jian Liu.

I would also like to thank my parents and my brothers for their support and encouragement throughout the years.

The financial support of the Natural Science and Engineering Research Council of Canada is gratefully acknowledged.

1. INTRODUCTION

This thesis develops an empirical Bayes model for marked salmon data collected over time. The method, which employs a hierarchical prior distribution, is used because the data are sparse and the empirical Bayes approach enables information to be aggregated to overcome the lack of information from data in individual cells. The novelty of our approach lies in our use of locally parametric Poisson models and smoothing techniques to obtain estimates of underlying trend in the tagged salmon data.

Over years, data on the return of tagged salmon are collected and prepared for the Mark Recovery Program(MRP) database. This database, which is described in detail in the Appendix, consists of the release data on tagged and untagged salmon, the data on individual marked salmon observed when returning from the ocean for spawning, and data on the sampling periods for each of the fishing regions. Since the database contains a vast amount of information, only selected sample data sets, such as the benchmark data sets, are analyzed here. Other data sets are also formatted like the benchmark data because this benchmark is well documented. With these data, various questions can be posed and investigated.

One topic of interest is the relationship between the size of smolts at release and their return rate as measured by observed marked salmon counts. Another is the comparison of marked salmon counts from different brood years and fishing regions. In this study, only the marked recoveries are examined. In addition, only two species of salmon, chinook and coho, are considered.

In tackling the two problems of interest described above, we first develop a model for the observed fish counts. The Poisson model is a conventional choice for count data. However, it will not be adequate for 'noisy' data with large sampling variation. Our solution to this problem adopts a local Poisson model to describe the variation at each sampling period. Noise is removed by local smoothing. Finally, the overdispersion relative to the Poisson model is modelled by mixing these locally smoothed, Poisson models in an appropriate way.

In Chapter 2, a brief description of the benchmark data sets is given. In addition, some relevant recent studies are summarized for completeness and later comparison or use. We discuss modelling Poisson processes with overdispersion, time series techniques for evaluating long-term trend effects, models for handling contagious or self-inhibiting processes, a local smoothing procedure for obtaining nonlinear regression estimates, and a Bayesian nonparametric smoothing method for modelling locally regular processes. Finally, the empirical Bayes method with hierarchical priors, the basis of this thesis, is reviewed.

To obtain insight for further investigation, the data are carefully examined in Chapter 3. The results indicate that some data pooling might be desirable to partially integrate the separate models for the marked recoveries observed in each catch region. The data from commercial fisheries appear to be more reliable and consistent than those from sport fisheries and escapement; thus, only the commercial data are used in the modelling stage of our analysis.

The local parametric Poisson models are developed in Chapter 4. Smoothing techniques are also developed there for removing noise, and estimating long-term trends in data. The main inferences are estimates of the Poisson intensity functions and the calculation of their corresponding credibility intervals.

Finally, in Chapter 5, the proposed models are fitted to selected coho and chinook data sets. A summary of the estimates and the corresponding credibility intervals is given. The problems of missing values and edge effects are also addressed there.

2. BACKGROUND REVIEW

2.1 The Benchmark Data Set

The benchmark data set is established in the Pacific Biological Station(PBS), which is a research branch of the Canadian Department of Fisheries and Oceans(DFO) in Nanaimo. The tag codes in this benchmark, which are obtained from the MRP database, form a sample data set for statistical analyses and exchanging data with other agencies. Complete documentation, including the selected formats and information related to the tag codes, is available in the 1986 report 'A Canadian MRP Data Benchmark'.

The benchmark consists of release data of tagged and associated juvenile salmon, and recovery data of adult marked salmon for selected tag codes. Data related to the sampling periods for recovering marked salmon are obtained from the MRP database directly. The following is a brief description of these data.

The benchmark release data contain a code for each release group of juvenile salmon. In particular, the origin, age, and average size of fish, the number of tagged and associated fish, as well as the site and date of the release are included for each group.

When adult fish are recovered, not every fish is inspected for mark. Different recovery methods have different sampling and reporting procedures associated with them. These recovery methods are mainly of three types:

i. commercial fisheries,

ii. sport fisheries, and

iii. escapement -- fish that are not captured by any fishery.

Each benchmark recovery data record includes the code found on each recovered and tagged salmon, the time, region and method of recovery. Times are usually recorded as year, month and statistical week (about 5 per calendar month). The recovery regions are geographic catch regions divided according to each of the fishing methods: troll, net, and sport. For each marked recovery, there is one record, except for escapement data. Thus, redundant sample information may appear on numerous records of individual tagged fish from the same sample. Fortunately, the fields of each data record are organized in such a way that data for an entire fish sample can easily be obtained.

The period for observing marked salmon is different in each catch region; thus, the data on sampling periods (described later in section 3.2) are important for determining whether a record is missing because of no sampling, or simply because there is no recovery during that period. Therefore, together with the recovery data, the distribution of marked recoveries in each catch region and the abundance of each group of tagged salmon can be observed over time. With all these data, many related questions can be tackled.

2.2 Statistical Techniques

2.2.1 Negative binomial and mixed Poisson regression

In this subsection, we describe for completeness and comparison, a model which bears some resemblance to that adopted in this thesis. However, it seems less flexible than ours and so has been set aside during the current investigation.

Suppose the response variable Y, a count, and a vector $\mathbf{x}$ of explanatory variables are specified. In general, let U | V denote the conditional distribution of U given V, where U and V are any two random variables with a joint distribution. Then a Poisson model for the response is as follows:

$Y \mid \mathbf{x}$ is Poisson distributed with mean $\mu(\mathbf{x})$,

where $\mu(\mathbf{x})$ is to be estimated.

Very often data exhibit extra-variation or overdispersion relative to the proposed Poisson model. For the count data with no covariates, the negative-binomial distribution is a popular choice for handling the extra-Poisson variation. To handle covariates, this result can be generalized to

$$P(Y = y \mid \mathbf{x}) = \frac{\Gamma(y + \alpha^{-1})}{y!\,\Gamma(\alpha^{-1})} \left[\frac{\alpha\,\mu(\mathbf{x})}{1 + \alpha\,\mu(\mathbf{x})}\right]^{y} \left[\frac{1}{1 + \alpha\,\mu(\mathbf{x})}\right]^{\alpha^{-1}}$$

$$y = 0, 1, \ldots, \qquad (2.1)$$

where $\alpha \geq 0$ is called the index or dispersion parameter.
The mean and variance of Y given $\mathbf{x}$ are

$$E(Y \mid \mathbf{x}) = \mu(\mathbf{x}) \quad \text{and} \quad Var(Y \mid \mathbf{x}) = \mu(\mathbf{x}) + \alpha\,\mu(\mathbf{x})^2.$$

Note that (2.1) yields the Poisson model if a → 0.

Lawless(1987) studies these negative-binomial models and examines their properties in detail. He reviews the maximum likelihood and moment estimation procedures for estimating the dispersion parameter and regression parameters. In addition, he compares the asymptotic covariance structures, efficiency and robustness of the parameters estimated by these two methods.

Since Poisson regression models are very useful, a test of the Poisson hypothesis is often of interest. One method is to test $\alpha = 0$ within the negative-binomial model. Lawless suggests some useful statistics such as the likelihood-ratio and the standardized dispersion, for testing this hypothesis. He also gives a note of caution that the result of any test depends on the size of the sample and $\mu(\mathbf{x})$.

2.2.2 Using SABL to decompose time series data

A method which is extensively used in this thesis will now be described. Suppose observations of a time series are taken at equally spaced time-points and the problem of interest is that of determining the long term trend in the deseasonalized series. Nicholls, Heathcote and Cunningham(1987) suggests a method, implemented in a software called SABL, that deseasonalizes the data, possibly after a transformation, without actually modelling the seasonal components. This method decomposes the series into three additive components by means of a minimization criterion and robust data smoothing techniques. The results at time t are the 'trend'(T_t), 'seasonal'(S_t) and 'irregular'(I_t) components. Let Y_t^T denote the transformed response at time t. Then

$$Y_t^T = T_t + S_t + I_t.$$

Nicholls, et.al.(*ibid*) explain that to construct the additive model, the original data must be transformed so as to minimize the interaction between the trend and the seasonal components. This criterion is reasonable since if their interaction were not at its minimum, then for example, if the trend were increasing, the seasonal component might also increase. With robust smoothing techniques based on moving medians, the trend and seasonal components can be determined iteratively. These robust estimates will not be affected by outliers because these outliers will be incorporated in the irregular component of the series. (To give more flexibility to users, SABL allows them to choose a particular transformation, and to select the widths of smoothing windows for the trend and seasonal components.)

After the decomposition, the seasonally adjusted series is obtained by simply subtracting off the seasonal component to give

$$Y_t^T = T_t + I_t.$$

This series can be converted back to the original response scale by applying the inverse transformation to Y_t^T. Once the trend and irregular components are computed, they may be plotted for visual inspection so that one can model the trend. The model can then be validated by other time series procedures, such as the Box-Jenkins autoregressive moving average (ARMA) technique.

2.2.3 Time series analysis of a contagious process

An alternative model to ours is described in this subsection and one of its deficiencies is noted. However, it promises to have some value and will be investigated further in future work.

Holden(1987) developed a model for rare events like the daily aircraft hijackings in US between 1968 and 1972, for example. The proposed model is for stationary processes. It incorporates the assumption that the contagiousness of an event eventually declines to zero, and that the rate of occurrences levels off over a long period with occasional, temporary peaks when an occurrence excites the process. With modification, the model can also incorporate the effects of exogenous time series.

The data is potentially applicable to the commercial marked salmon data of a given tag code observed in a catch region since there is a long period of no recovery during the winter season. However, the leveling off phase of epidemics is not fully reflected in our data because of the definition of the yearly sampling 'periods' (See Table 3.4). During the sampling season, there are only occasional recoveries which can be thought of as rare events. But an important similarity is that the observed recoveries are serially correlated. Thus, we conclude that Holden's model might be adapted for modelling the salmon data in spite of its deficiencies with respect to our data.

Holden assumes that the observed sequence of daily counts, $\{ N_t \}$, is a sequence of Poisson variates with means given by some sequence, $\{ \lambda_t \}$, which incorporates the stimulating effects of previous incidents. The linear contagion model for rare events is given by

$$\lambda_t = E\left[N_t \mid N_u, u < t \right] = \nu + \delta_t, \qquad (2.2)$$

where

$$\delta_t = \sum_{i=1}^{\infty} W_i N_{t-i}, \qquad (2.3)$$

$W_i \geq 0$ (i = 1,2,...) and t is an integer. For a discrete-time process (2.2), N_t conditioned on the history $\{ N_u, u < t \}$ has a Poisson distribution with mean λ_t. $\sum_{i=1}^{\infty} W_i$ is required to be less than one to ensure that $\mu \equiv E(\lambda_t) > 0$. The quantity ν is the rate at which events are generated by factors other than contagion (assumed constant).

The lag structure of W_i ($i = 1,2,\ldots$) in (2.3) describes the contagiousness of an event i periods after its occurrence. To get a finite number of parameters, simply set W_i to zero after some maximum lag, or assume that W_i has a specified functional form, such as the lag weights associated with a given ARMA process. Then the time-series techniques suggested by Box-Jenkins may be used to obtain the parameter estimates.

2.2.4 Smoothing techniques

i. Estimating smooth functions by the local scoring algorithm

To provide additional perspective on the approach taken in this thesis, a very recently proposed method, similar to our own in spirit, will now be described.

For likelihood-based regression models with response variable Y, such as normal linear regression, one usually assumes a linear form in the covariates $X_1,X_2,\ldots,X_p$. A set of n independent realizations of these random variables will be denoted by $(y_1,x_{11},\ldots,x_{1p})$, $\ldots,(y_n,x_{n1},\ldots,x_{np})$. Hastie and Tibshirani(1986) propose the class of generalized additive models which replace the linear form $\sum_1^p \beta_j X_j$ by a sum of smooth functions $\sum_1^p s_j(X_j)$. The $s_j(\cdot)$ are unspecified functions that are estimated using a scatterplot smoother in an iterative procedure called the *local scoring* algorithm.

Any regression procedure can be viewed as a method for estimating $E(Y \mid X_1, X_2, \ldots, X_p)$. The additive model assumes the following form for this conditional expectation:

$$E(Y \mid X_1, X_2, \ldots, X_p) = s_0 + \sum_{i=1}^{p} s_j(X_j), \qquad (2.4)$$

where the smooth $s_j(\cdot)$'s are standardized so that $E(s_j(X_j)) = 0$. These functions are estimated one at a time using a scatterplot smoother.

A simple class of scatterplot smoother estimates are the local average estimates,

$$\hat{s}(x_i) = \underset{j \in N_j}{\mathbf{Ave}} \{y_j\}$$

where **Ave** represents some averaging operator like the mean and N_j is a neighborhood of x_i (a set of indices of points whose x values are close to x_i). The type of neighborhoods considered in Hastie and Tibshirani's paper are *symmetric nearest neighborhoods*. Associated with this is the span or window size w, which is the proportion of points contained in each neighborhood. Other more complicated estimates of $E(Y \mid X)$ can be used, such as kernel or spline smoothers.

The span w is selected to tradeoff between the bias and variability of the estimate. A data-based criterion is derived for selection. Let $\hat{s}_w^{-i}(x_i)$ be the smoother of span w at x_i, having removed (x_i, y_i) from the sample. Then the *cross-validation* sum of squares (CVSS) is defined by

$$\text{CVSS}(w) = (1/n) \sum_{i=1}^{n} \left[y_i - \hat{s}_w^{-i}(x_i) \right]^2 \qquad (2.5)$$

The optimal span w is that which gives the smallest value of CVSS(w). This criterion effectively weighs bias and variance based on the sample. Note that the E(CVSS(w)) can be shown to be approximately equal to the integrated prediction squared error (PSE)

$$\text{PSE} = E(Y - s(X))^2,$$

and that CVSS is approximately unbiased for the expected prediction squared error. In addition to these desirable properties, CVSS is recommended because it is computationally efficient for obtaining the optimal value of w.

For the additive model in (2.4), the *local scoring* algorithm estimates the $s(\cdot)$'s by iteratively smoothing the adjusted dependent variable on X, and it requires a choice of span which can be estimated using the CVSS(w) in (2.5). Theoretically, this technique can be viewed as an empirical method of maximizing the expected log-likelihood, or equivalently, of minimizing the Kullback-Leibler distance to the true model. It is called *local scoring* because the Fisher scoring update is computed using a local estimate of the score.

ii. Bayesian nonparametric smoothing method for local regular process

A potential refinement of the approach adopted in this thesis is described by Ma(1986) who improves on the Bayesian nonparametric approach proposed by Weerahandi and Zidek(1988) for smoothing stochastic processes. The processes of concern are of the form R = S + N, where S is a smooth function and N is an independent noise process. R is assumed to be observed at a sequence of time-values, t_i, i = 1,..,n, and S is assumed to be locally regular, that is, expandable in a Taylor series to the ρth term about $t = t_{n+1}$. Then an a priori structural model for the data is

$$R = X\beta + \varepsilon, \tag{2.6}$$

where

$R = (R_1,\ldots,R_n)'$ is a vector of n observations,

$X = (1,X_1,\ldots,X_p)$ is an n by $(\rho+1)$ matrix,

where **1** is a vector of ones and

$X_j' = ([t_1 - t_{n+1}]^j/j!,\ldots,[t_n - t_{n+1}]^j/j!)$,

$\beta = (\beta_o,\beta_1,\ldots,\beta_p)'$ is vector of coefficients,

where $\beta_o = S(t_{n+1})$ and $\beta_i = D^i S(t_{n+1})$ with D as the operator of differentiation, and

$\varepsilon = \eta + N$ is the error term,

where both η and N are vectors; specifically η is the remainder of the Taylor expansion of $S(t_i)$ and N is the noise with variance σ^2.

One further assumption underlying this approach is that the expansion errors and all other a priori uncertainty about R, β and the smoothing parameter c, related to the variance of the noise ε, have a joint multivariate normal distribution.

The "smoothing parameter" c controls the degree of smoothness of the estimated R. The main objective of Ma's study(*ibid)* is a simple method to compute an estimate of c and to obtain $\hat{R}$, a smooth estimate of R. His method estimates c, and computes a measure of accuracy for any given $\hat{R}$, called the *predictive squared error* PSE (to be defined later) for each fixed order p which reflects the degree of local regularity of S. The value of p that has the minimum PSE is chosen to be the optimal value.

For each fixed p, the parameter c can be estimated by *cross-validation* which chooses the value of c that minimizes the cross-validated sum of squared (a similar method is described in section 2.2.4.i). Ma(*ibid*) develops a simpler alternative called the *backfitting* method and compares it to *cross-validation*. His new method is recommended for obtaining c because it is easier to implement and computationally more efficient than the *cross-validation* approach.

Ma's method may be described as follows. Suppose in equation (2.6), S has p + 1 derivatives. Then the a priori model is of the form

$$R_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_p x_i^p + \beta_{p+1} x_i^{p+1} + \varepsilon_i, \qquad (2.7)$$

where

R_i is the *i*th component of the vector R,

ε_i is the *i*th component of the error vector ε, and

$x_i = (t_1 - t_{n+1})^i / i!$, $i = 1,\ldots,p+1$.

The *backfitting* method uses the fact that $c = \delta^2 / \sigma^2$, where δ^2 is the prior variance of $D^{p+1}S/(p+1)!$, and σ^2 is the variance of the noise N. Then, for the order p, c_p is given by

$$c_p \approx \frac{\delta^2}{\sigma^2}$$

$$\cong \frac{\text{sample variance of } (\hat{\beta}_{p+1})}{\hat{\sigma}^2 \, ((p+1)!)} \qquad (2.8)$$

Thus, if equation (2.7) holds, by the same argument, (2.8) can be used to estimate the values of c_j for any $j = 0,\ldots,p$. However, to use equation (2.8) c_{p+1} is needed. This parameter can be estimated by *cross-validation*, or it can simply be set to zero assuming p is large. Then, the c_j with $j \le p$ is estimated by *backfitting* using (2.8).

The value of p is optimal in the sense that the $\hat{R}$, estimated by the pth order locally regular fit, minimizes the *predictive squared error* (PSE)

$$\text{PSE}(j) \equiv 1/(n-m) \sum_{i=1}^{n-m} \left[R_{m+i} - \hat{R}_{m+i}(\hat{c}_j) \right]^2, \quad j = 0,\ldots,p,$$

where m is the fixed span of observations (m < n), $p + 1$ is a fixed integer, and R_i is the observed value at time t_i. For each chosen value of ρ, the *backfitting* method is used to compute the c_j for $j \leq \rho$, and the related R_i's are estimated using generalized least squares procedure. The PSE(j) is computed for each $j \leq \rho$ and these values are compared to obtain the optimal $j \leq \rho$ which minimizes the PSE.

2.2.5 Empirical Bayes(EB) and Hierarchical Bayes(HB) analyses

i. Introduction

Let $\mathbf{X} = (X_1,...,X_n)'$ be a vector of n independent random variables which come from a common distribution f with parameter θ. Given a sample of n observations, $\mathbf{x} = (x_1,...,x_n)'$, all relevant information about θ is contained in the observed likelihood function $f(\mathbf{x} \mid \theta)$. Thus, a θ with large $f(\mathbf{x} \mid \theta)$ is more plausibly the true θ than a θ with small $f(\mathbf{x} \mid \theta)$. Likewise, the occurrence of $\mathbf{x}$ would be more plausible if $f(\mathbf{x} \mid \theta)$ were large. Therefore, as a corollary of the Likelihood Principle, only the observed $\mathbf{x}$ should be relevant to conclusions about θ. (More details on the Bayesian analysis can be found in Berger(1985).)

Suppose prior knowledge about θ is given by the distribution $\Pi(\theta)$. Bayesian analysis combines this prior information and the sample information using Bayes rule into what is called the posterior distribution of θ given $\mathbf{x}$, from which all decisions and inferences may be made. This posterior distribution $\Pi(\theta \mid \mathbf{x})$, which reflects the updated beliefs about θ after observing the sample is defined as follows. Let the joint density of **X** and θ be

$$h(\mathbf{x},\theta) = \pi(\theta)\ f(\mathbf{x}\ |\theta), \tag{2.9}$$

and the marginal density of **X** is

$$m(\mathbf{x}) = \int_{\Theta} f(\mathbf{x}\ |\theta)\ dF(\theta). \tag{2.10}$$

Then, providing $m(\mathbf{x}) \neq 0$,

$$\pi(\theta\ |\mathbf{x}) = \frac{h(\mathbf{x},\theta)}{m(\mathbf{x})}. \tag{2.11}$$

When no prior information about θ is available, what is needed in such situations is a noninformative prior, by which is meant a prior that contains no information about θ. A reasonable choice of such a prior is to give equal weights to all possible values of θ. A typical noninformative prior density is $\pi(\theta) = 1$, the uniform density on $\mathbb{R}$. Given the prior, the analysis can proceed in a conventional Bayesian fashion.

ii. Empirical Bayes(EB) analysis

Assume $\pi(\theta)$ has a given functional form, and choose the density of this given form which closely matches the prior beliefs. We assume $\pi \in \Gamma$ with

$$\Gamma = \{\ \pi\ :\ \pi(\theta) = g(\theta\ |\lambda)\ \text{where}\ \lambda \in \Lambda\ \}. \tag{2.12}$$

Here g is a specified function. Then the choice of prior reduces to the choice of $\lambda \in \Lambda$ which is usually called a hyperparameter of the prior. The Type II maximum likelihood estimate (ML-II estimate) of π is such that

$$m(x \mid \hat{\pi}) = \sup_{\pi \in \Gamma} m(x \mid \pi),$$

where

$$m(x \mid \pi) = \int f(x \mid \theta)\, \pi(\theta)\, d\theta, \text{ and}$$

Γ is the set described in (2.12). The marginal density of x given π, $m(x \mid \pi)$, reflects the plausibility of π with the data in hand. This function is clearly maximized by choosing π to be concentrated where $f(x \mid \theta)$ is maximized (as a function of θ). Thus, it is reasonable to consider $m(x \mid \pi)$ as a likelihood function for π. Then

$$\sup_{x \in \Gamma} m(x \mid \pi) = \sup_{\lambda \in \Lambda} m(x \mid g(\theta \mid \lambda))$$

so that the selection is just a maximization over the hyperparameter λ (ML-II hyperparameter).

iii. Hierarchical Bayes(HB) analysis

It is often convenient to elicit subjective prior information in stages. For two stage priors, for example, the initial prior is $\pi_1(\theta \mid \lambda)$, where λ is a hyperparameter in Λ. Instead of estimating λ, as in the empirical Bayes analysis, λ is given a second stage prior distribution $\pi_2(\lambda)$. This could be a proper prior, but more often it is an appropriate noninformative prior. Sometimes λ is written in the form of $\lambda = (\lambda_1, \lambda_2)$ for ease of computation. Then

$$\pi_2(\lambda) = \pi_{2,1}(\lambda^1|\lambda^2)\ \pi_{2,2}(\lambda^2). \tag{2.13}$$

The posterior distribution of θ is then expressed in terms of the posterior distribution at various stages of the hierarchical structure. The procedure is as follows.

If all densities below exist and are non-zero, then

$$\Pi(\theta\ |x) = \int_\Lambda \pi_1(\theta\ |x,\lambda)\ \pi_{2,1}(\lambda^1|x,\lambda^2)\ \pi_{2,2}(\lambda^2|x)\ d\lambda. \tag{2.14}$$

Here

$$\pi_1(\theta\ |x,\lambda) = \frac{f(x\ |\theta)\ \pi_1(\theta\ |\lambda)}{m_1(x\ |\lambda)}, \tag{2.15}$$

where

$$m_1(x\ |\lambda) = \int f(x\ |\theta)\,\Pi_1(\theta\ |\lambda)\,d\theta,$$

$$\pi_{2,1}(\lambda^1|x,\lambda^2) = \frac{m_1(x\ |\lambda)\ \pi_{2,1}(\lambda^1|\lambda^2)}{m_2(x\ |\lambda^2)}, \tag{2.16}$$

$$m_2(x\ |\lambda^2) = \int m_1(x\ |\lambda)\ \pi_{2,1}(\lambda^1|\lambda^2)\ d\lambda^2,$$

$$\pi_{2,2}(\lambda^2|x) = \frac{m_2(x\ |\lambda^2)\ \pi_{2,2}(\lambda^2)}{m(x)},$$

and

$$m(x) = \int m_2(x\ |\lambda^2)\ \pi_{2,2}(\lambda^2)\ d\lambda^2. \tag{2.17}$$

iv. Comparison of EB and HB

First, the advantages of HB over EB are considered. The EB estimates of the hyperparameters obtained from the ML-II approach and then using the first stage prior in a standard Bayesian way albeit with the hyperparameter replaced by its ML-II estimate ignores the inherent uncertainty about the hyperparameter. It leads to unduly optimistic estimates. The HB approach incorporates such uncertainty automatically. Furthermore, with only slight theoretical difficulty, HB can incorporate actual subjective prior information at the second stage.

Even though HB has many advantages over the EB approach, it is more difficult to apply because of its greater computational complexity among other things. As well, Savage's Principle of Precise Measurement asserts that when the likelihood is "peaked" relative to the prior, the EB method is justifiable as a good approximation to HB. So, in particular, there is a large amount of data available in the marginal likelihood, the ML-II estimate produces a reliable estimate of the hyperparameter without an additional stage of prior modelling and conventional Bayes estimation.

3. DATA EXPLORATION

3.1 Introduction

The structure of each of four sets of data is examined in this section as a preliminary step in model development. These sets include the benchmark release data, the benchmark rollup recovery data, the rollup recovery data for replicated tag codes, and the sampling period data. These sample data sets involve two species of marked salmon: chinook(124) and coho(115). (The codes in parentheses are species Hart codes used by DFO for species identification.) Note that the rollup recovery data set for replicated tag codes and the sampling period data sets are not part of the benchmark data. These data are obtained directly from the MRP database, and they are formatted like the benchmark so that the documentation for the benchmark can be used for reference.

The release, recovery and sampling period data have, respectively, 31, 33 and 15 fields in their records. The fields are of varying lengths depending on the type of information contained. A blank field in a release or recovery data record represents a missing value. Zero in one or both of the catch and sample fields of a sampling period data record indicates no sample is taken for that time period. With this background, we now begin a careful examination of these data sets.

3.2 Benchmark Release and Recovery Data Sets

The tag codes in the benchmark release and recovery data subsets are chosen so as to include a wide range of total recoveries and to have several release years represented. In particular, some codes associated with true scientific replicates are selected for coho. A preliminary examination of these two data sets will indicate the problems and the sort of information available. The results are discussed below.

3.2.1 Missing values in the two benchmark data subsets

i. Release data

Each record in the release data set contains the data for one tag code. Since there are only 9 tag codes for chinook and 27 for coho in this benchmark subset, all records are examined together. Table 3.1 gives the 36 tag codes and their corresponding codes in the analysis. Table 3.2 shows that none of these records have any missing values. Also, note that fields 11(number tagged), 12(adipose only) and 13(unclipped) are all nonzero which indicates that not all released fish are tagged. Thus, for each release group, information such as brood year, production area, size at release and time of release, is important in associating unmarked and marked fish. The size of each release considered here is in thousands, but Figures 3.1a and 3.1b show that the chinook release, in general, is much larger in size than the coho.

ii. Recovery data

For the benchmark rollup recovery data subset, there are more than 2000 records for the combined chinook and coho data. For illustrative purposes, only the chinook records are considered here. Some fields in a record contain information only relevant to one of the three recovery methods: commercial fisheries, sport fisheries and escapement. Thus, we have to know the number of records corresponding to each recovery method before computing the percentage of missing values for each field.

The results, as shown in Table 3.3, show that one third of these 33 fields in the recovery data subset are missing more than 75 percent of its values. Most of them are about the physical characteristics of salmon, such as average fork length and percentage of mature females in the sample. Two other important fields with a high percentage of missing values are concerned with recoveries from sport fisheries. As a result, the reliability of the sport data is questionable.

3.2.2 The structure of observed recoveries

The observed counts over time are of particular interest because the results may reflect the survival rate of tagged salmon and the trend of observed recoveries. These counts may also indicate some relationship between the return rate and factors that affect the survival of salmon. Plots of tag codes 021827(chinook) and

081842(coho) are given as illustrations.

The results indicate that chinook start to return a year after their release while the coho return the year they are released. The recorded recovery time has three components: year, month and statistical week. The statistical week (0 to 5) indicates the week within a month. When week = 0, the week is not known. These three components are usually known only for commercial recovery times. For the sport fisheries, the recoveries are mostly monthly data, and the escapement has only yearly bulk data.

A term called 'period', which represents the time of recovery in each year, is defined using the month and statistical week components. This 'period' is a number ranging between 1 and 40 representing a one week time period during which salmon fishing may occur. Table 3.4 is a reference table for computing 'period' from month and statistical week in each calendar year. Note that period 40 actually covers three months of the winter season when there is no salmon fishing.

Using the definition of 'period', the total observed number of recoveries from sport and commercial fisheries over the entire recovery period, ignoring the catch regions, are now inspected. The plot of chinook, as shown in Figure 3.2a, shows that the tagged chinook have a recovery period spanning four years, and that most recoveries are concentrated between May and September in each year, but larger observations are obtained during the second and the third year. For the tagged coho, the sport recoveries are monthly data; therefore, they cannot be combined with the commercial recoveries. The plots in Figure 3.3a(commercial) and 3.3b(sport) indicate that only a few return in later months of the first recovery year, and most data from the commercial fisheries are observed between June and September in the second recovery year. A point that is not demonstrated by these plots is that most marked recoveries found during the first year for both species are from escapement.

The variation of commercial and sport chinook recoveries can clearly be identified when they are plotted separately and this is done in Figures 3.2b and 3.2c, respectively. Note that the plots for the commercial recovery data are similar to those for sport recovery. Also note from Figure 3.2c the relative paucity of points for sport recovery. The results from the chinook and coho plots indicate that the sport fisheries contribute little to the total number of recoveries. As mentioned earlier, the escapement data are yearly bulk data so we cannot compare them to the commercial data. Therefore, we have chosen to concentrate our study on the commercial recoveries with little apparent loss of information obtainable from these data.

The plots of the cumulative sum of chinook and coho commercial recoveries observed over time are also examined. These plots, as presented in Figures 3.4a and 3.4b , show that the recoveries are a step function of time. The jumps that are of visible size in August in later recovery years are especially obvious for coho. These observations provide more information on the peak season of salmon return.

3.3 Other Related Information and Data Sets

We first investigate the 37 commercial and sport catch regions. Table 3.5 is a list of old and new catch region codes. The old codes are the originals used by DFO, and the new ones are created for ease of programming. The results in Table 3.6 show that only 25 of these regions have recoveries among the approximately thousand records considered, and very few of them have more than 100 recoveries over the entire recovery period.

The sampling period data are the sampling schedules for different catch regions. In each year, each catch region has its own sampling scheme which may be different from previous years. For most commercial catch regions, there are usually consistent sampling periods during the fishing season. However, for the sport catch regions, often there are few samples taken each year and this results in a long period of no information (missing values). This is another reason why it is difficult to analyze the sport data.

Using the data on sampling periods and catch regions, the commercial observed recoveries may be examined further. As a result, three commercial catch regions are found to have a substantial number of observations for chinook, and four for coho over the corresponding recovery period. The three regions for chinook are: Northwest Vancouver Island Troll, Northern Troll and North Central Troll. The four regions for coho are: Southwest Vancouver Island Troll, Georgia Strait Troll, South Central Troll and Johnstone Strait Net.

For chinook, these results indicate that the length of the recovery period in these three catch regions is about three years long instead of four for each particular tag code. Also, the samples are mostly taken from the beginning of May to the end of October with some missing ones in between, and no sampling was done beyond these time limits. For this reason, there are about 63 sampling periods, which will hereafter be called the adjusted period, in each catch region over the three-year period. The observed recoveries over the adjusted periods are now plotted for each catch region. Note that the blanks between lines on the graphs, as shown in Figure 3.5, indicate missing values and that most observations are obtained at the beginning of the recovery period.

For coho, the recovery period is only a year long for each tag code in the four regions mentioned. In addition, the sampling period, which starts in mid-June and ends at the end of October, is shorter than that of chinook. Plots of observations over the 17 periods in

Figure 3.6 reveals that few tagged coho are found during sampling.

Data is available in the MRP database on replicated tag codes for both coho and chinook. This enables us to pool information to estimate the trend of recoveries over time. In addition, the model developed for these replicates can be used as a guide in modelling other individual tag code or pseudo-replicates. This set of data consists of statistical replicates, in that each set of replicated tag codes of various time-size combinations come from a single pond representing one treatment. The replicates in each group of three can be further classified according to three relative size groups: large, medium, and small. Table 3.7 gives a sample list of these size groups.

Examination of both chinook and coho replicates shows that there are not many observed coho recoveries over the 17 periods in each catch region for each group and there are even fewer observations over the 63 adjusted periods for the chinook groups. The sparse data suggest that pooling observations appropriately may be helpful in obtaining meaningful results. In addition, since there is so much sampling variation and noise in the recovery data, some smoothing techniques may be useful when modelling this data.

4. EMPIRICAL BAYES APPROACH FOR MODELLING COUNT DATA

4.1 Introduction

The investigation of the trends in recovery rate for each catch region and species of salmon is the main interest of this study. The results of exploratory analysis indicate that the tagged salmon recovery data have not only noise but also huge variation due to other factors. Further, data on individual tag codes are sparse given that, in fact, there are few recoveries over the entire recovery period. However, replicated tag codes have a somewhat similar pattern of recoveries, sampling scheme and set of recovery regions. Thus, pooling the data for tagged salmon with similar characteristics provide more informative data for further statistical analysis. Some smoothing techniques for removing noise and sampling variation in the data also prove useful.

Since no prior information about the distribution of the rates of observing marked salmon are available, we can only use the recovery data to suggest possible ways to estimate these rates and their corresponding confidence intervals. We now develop models that can aggregate data, remove most of the sampling variation and noise from data while reflecting the mechanisms of the underlying process and taking account of the source.

The Poisson process is a common model for count data. However, a global Poisson model is inadequate for these data because of the heterogeneity in these data. An empirical Bayes approach to fitting local Poisson models to these counts enables us to incorporate this heterogeneity.

4.2 Local Parametric Poisson Models with Smoothing Techniques

Let $N_i(t)$ denote the count for the ith process up to time t, where i = $1,\ldots,I$. Assume that for a fixed period, $N_i(t)$ has been observed at a sequence of time-intervals, (t_j, t_{j+1}), $j = 1,\ldots,J-1$ all equally long. Each of these I processes is serially correlated, and furthermore, all of them are interrelated. For a fixed i, let $E[N_i(t)] = \lambda_i(t)$. This mean function reflects how the arrival rate of $N_i(t)$ changes over time. Suppose $\lambda_i(t)$ has a prior distribution which is exponentially distributed with mean $\beta_i(t)$. Then β_i represents our prior expectation about the size of λ_i. However, the β_i-values are themselves uncertain so we put a second stage hyperprior on β_i to incorporate this uncertainty and also to remedy the possible over-dispersion effect relative to the Poisson distribution. This prior is an inverse exponential distribution with parameter $\zeta(t)$.

Note that in choosing our prior distribution, we seek a distribution which is both noninformative and easy to handle. Further, to reflect our a priori view of these β_i's as different in unspecified ways, we postulate that they are exchangeable random variables. More precisely, at every fixed time t, $\beta_1,\ldots,\beta_I$ are regarded as a random sample taken from a common inverse exponential distribution so these β_i's are independent and identically distributed(i.i.d.). Also, $E(\lambda_i) = \beta_i$ for i = 1,...,I. Once the λ_i's are given, the N_i's are independent, but the I processes are related indirectly through ζ.

Recalling our convention of letting U | V denote the conditional distribution of U given V for any two random variables U and V having a joint distribution, we assume

i) $N_i(t)|\lambda_i(t),\beta_i(t),\zeta(t)$ is Poisson$(\lambda_i(t))$,

where $\beta_i(t)$ and $\zeta(t)$ are non-negative,

ii) $\lambda_i(t)|\beta_i(t),\zeta(t)$ is exponential$(\beta_i(t))$, and

iii) $\beta_i(t)|\zeta(t)$ is inverse exponential$(\zeta(t))$.

We now develop the model further by specifying the densities of these non-negative real variables. From now on, the time t is assumed to be fixed and is omitted for clarity. Then, for the ith process, we have for λ_i, β_i, $\zeta \geq 0$:

$$f(N_i = n_i|\lambda_i,\beta_i,\zeta) = \frac{e^{-\lambda_i}\lambda_i^{n_i}}{n_i!}, \qquad (4.1)$$

$$f(\lambda_i|\beta_i,\zeta) = \frac{1}{\beta_i}e^{-\lambda_i/\beta_i}, \qquad (4.2)$$

$$f(\beta_i|\zeta) \quad = \quad \frac{\zeta}{\beta_i^2} e^{-\zeta/\beta_i} \quad . \qquad (4.3)$$

Note that because of our choice of a noninformative prior for β_i, the moments of $f(\beta_i|\zeta)$ do no exist. We can rewrite (4.3) as follows

$$f(\beta_i|\zeta) \quad = \quad \frac{1}{\zeta} \left[e^{-(\beta_i/\zeta)} \, (\beta_i/\zeta)^{-2} \right].$$

That is,

$$f(\beta_i|\zeta) \quad = \quad \frac{1}{\zeta} \, g(\beta_i/\zeta), \qquad (4.4)$$

where $g(\beta_i/\zeta) \quad = \quad \exp(-\beta_i/\zeta) \; (\beta_i/\zeta)^{-2}$.

We can easily identify ζ in equation (4.4) as the scale parameter of the density $f(\beta_i|\zeta)$ and $1/\zeta$ as the precision parameter. Thus, ζ indicates the spread of the β population from which the β_i's are picked, and $1/\zeta$ expresses the degree of equality among the β_i's. In particular, the β_i's are identical when ζ is zero and the β_i's are very different when ζ is large.

Now the joint density of λ_i and β_i given ζ is

$$f(\lambda_i,\beta_i|\zeta) \quad = \quad f(\lambda_i|\beta_i,\zeta) \; f(\beta_i|\zeta)$$

$$= \quad \frac{1}{\beta_i} e^{-\lambda_i/\beta_i} \left[\frac{\zeta}{\beta_i^2} e^{-\zeta/\beta_i} \right].$$

That is,

$$f(\lambda_i,\beta_i|\zeta) = \frac{\zeta}{\beta_i^3} e^{-(\lambda_i+\zeta)/\beta_i} . \qquad (4.5)$$

Then the prior for λ_i given ζ is

$$f(\lambda_i|\zeta) = \int_0^\infty f(\lambda_i,\beta_i|\zeta)\, d\beta_i .$$

That is,

$$f(\lambda_i|\zeta) = \zeta \int_0^\infty \frac{e^{-(\lambda_i+\zeta)/\beta_i}}{\beta_i^3}\, d\beta_i . \qquad (4.6)$$

After the change of variable, $u = 1/\beta_i$, equation (4.6) becomes

$$f(\lambda_i|\zeta) = \zeta \int_0^\infty u\, e^{-(\lambda_i+\zeta)\,u} du.$$

or

$$f(\lambda_i|\zeta) = \frac{\zeta}{(\lambda_i+\zeta)^2}. \qquad (4.7)$$

It is easy to show that $f(\beta_i|\zeta)$ and $f(\lambda_i|\zeta)$ are both unimodal functions with respect to ζ and their unique modes are at $\zeta = 2\beta_i$ and $\zeta = \lambda_i$, respectively. Thus, a priori, most of the β_i's are concentrated near $\zeta/2$.

We now determine the joint density of N_i and λ_i given ζ which is

$$f(N_i,\lambda_i|\zeta) = f(N_i|\lambda_i)\, f(\lambda_i|\zeta),$$

that is,

$$f(N_i,\lambda_i|\zeta) = \frac{e^{-\lambda_i}\lambda_i^{n_i}}{n_i!}\,\frac{\zeta}{(\lambda_i+\zeta)^2}. \tag{4.8}$$

After integrating out λ_i, we obtain

$$f(N_i|\zeta) = \int_0^{\infty} \frac{e^{-\lambda_i}\lambda_i^{n_i}}{n_i!}\,\frac{\zeta}{(\lambda_i+\zeta)^2}\,d\lambda_i. \tag{4.9}$$

Finally, to obtain the conditional posterior joint distribution of λ_i and β_i, we require an estimate of $\zeta = \zeta(t)$. The value of ζ can be estimated by maximizing an expression which involves the integral of (4.9). Then, using (4.1) ,(4.5) ,(4.9) and $\hat{\zeta}$, the above mentioned estimate, the posterior is

$$f(\lambda_i,\beta_i|N_i,\hat{\zeta}) = \frac{f(N_i|\lambda_i,\beta_i,\hat{\zeta})\ f(\lambda_i,\beta_i|\hat{\zeta})}{f(N_i|\hat{\zeta})}.$$

That is,

$$f(\lambda_i,\beta_i|N_i,\hat{\zeta}) = \frac{\dfrac{e^{-\lambda_i}\lambda_i^{n_i}}{n_i!}\,\dfrac{\hat{\zeta}}{\beta_i^{3}}\,e^{-(\lambda_i+\hat{\zeta})/\beta_i}}{\displaystyle\int_0^{\infty}\frac{e^{-\lambda_i}\lambda_i^{n_i}}{n_i!}\,\frac{\hat{\zeta}}{(\lambda_i+\hat{\zeta})^{2}}\,d\lambda_i}, \tag{4.10}$$

where β and λ are non-negative.

The integral in the denominator of (4.10), which is just (4.9), is essentially a constant once ζ is estimated. It remains to estimate ζ using (4.9). For clarity, the subscripts in (4.9) are dropped and we let $P = n_i!$. Then (4.9) becomes

$$g(\zeta) = P^{-1} \int_0^{\infty} e^{-\lambda} \lambda^{n} \zeta (\lambda + \zeta)^{-2} d\lambda \qquad (4.11)$$

where $\lambda \geq 0$, n and P are positive integers.

To prove that $g(\zeta)$ has a unique maximum, we appeal to a lemma of Brewster and Zidek(1974). First, suppose $W(u)$ is a continuous non-negative function whose domain is either $(0,\infty)$ or $(-\infty,\infty)$, and it is strictly bowl-shaped. Thus, W is differentiable almost elsewhere. In addition, assume that, whenever necessary for integrals involving W, the interchange of integral and derivative is permissible. The lemma is:

If f is a density on $(0, \infty)$ $[(-\infty, \infty)]$ and $\{ f(xc^{-1}): c > 0 \}$ $[\{f(x-c): -\infty < c < \infty \}]$ has monotone likelihood ratio property (MLRP), then

$$c \to \int x\, W'(cx)\, f(x)\, dx \qquad \left[\int W'(x+c)\, f(x)\, dx \right]$$

has at most one sign change and

$$c \to \int W(cx)\, f(x)\, dx \qquad \left[\int W(x+c)\, f(x)\, dx \right]$$

is strictly bowl-shaped (or monotone).

We first show that $g(\zeta)$ in (4.11) satisfies the conditions stated in this lemma. From (4.11), we have $W(\lambda) = P^{-1} e^{-\lambda} \lambda^{n}$, where n and P are positive integers and $\lambda \geq 0$. It can easily be shown that $W(\lambda)$ is strictly bowl-shaped (opening down) and $f(\lambda|\zeta)$ is a scale density which can be written as

$$f(\lambda|\zeta) = \frac{1}{\zeta\ (1 + \lambda/\zeta)^{2}} = \frac{1}{\zeta} f_{o}(\lambda/\zeta),$$

where $f_{o}(y) = f(y|1) = (1 + y)^{-2}\, dy$.

In addition, $\langle f(\lambda|\zeta) \rangle$ has the MLRP since if $\lambda_{1} < \lambda_{2}$ and $\zeta_{1} < \zeta_{2}$, then

$$\begin{vmatrix} f(\lambda_{1}|\zeta_{1}) & f(\lambda_{1}|\zeta_{2}) \\ f(\lambda_{2}|\zeta_{1}) & f(\lambda_{2}|\zeta_{2}) \end{vmatrix} > 0.$$

Since $g(\zeta)$ satisfies all the conditions in the lemma, we conclude that

$$\zeta \rightarrow \int \lambda\, W'(\lambda\zeta)\, f_{o}(\lambda)\, d\lambda$$

has at most one sign change, and

$$\zeta \rightarrow \int W(\lambda)\, f(\lambda|\zeta)\, d\lambda$$

is strictly bowl-shaped (opening down). Thus, $g(\zeta)$ in (4.11) has a unique maximum which implies the same for expression (4.9).

Since at each time point t, $\beta_{1},\ldots,\beta_{I}$ are assumed to be a random sample from the inverse exponential distribution with parameter ζ, we can use all the data from these I processes to estimate $\zeta(t)$. An advantage of this approach is that we can aggregate data indirectly instead of simply summing the data without considering their sources, degree of association, and so on. Further, if observations obtained for each process are from replicated experiments, we can also use them together so that n_i (the observed count) is a vector of counts from the replicates.

To remove noise and other extraneous variation from the data, we use moving averages with appropriate window sizes (of at least 3). Other more complicated smoothing techniques may be used, but this simple method has the advantage that it can easily be incorporated into the models we are developing. One assumption underlying all smoothing methods is that the counts within a window are homogeneous. So the size of the window is restricted since the data points in a small neighborhood are expected to be more similar than those far apart.

As an example, suppose there are 4 processes (I = 4) with 2 replicates each, and a symmetric neighborhood with a 3-point window is used for smoothing. Then, for a given time t and for each i = 1,...,4, $N_i = (N_{i1}, N_{i2})'$ is a vector of the 2 replicates. The joint density of these two replicates given λ_i is

$$f(N_i | \lambda_i) \quad = \quad \prod_{j=t-1}^{t+1} \prod_{k=1}^{2} f(N_{ijk} | \lambda_i, \beta_i). \qquad (4.12)$$

Note that $f(N_i,\lambda_i|\beta_i) = f(N_i|\lambda_i)\ f(\lambda_i|\beta_i)$ and

$$f(N_i,\lambda_i,\beta_i|\zeta) = f(N_i|\lambda_i)\ f(\lambda_i|\beta_i)\ f(\beta_i|\zeta).$$

Thus, $f(N_i,\lambda_i|\zeta)$ is given by

$$\int_0^\infty f(N_i,\lambda_i,\beta_i|\zeta)\ d\beta_i = f(N_i|\lambda_i) \int_0^\infty f(\lambda_i|\beta_i)\ f(\beta_i|\zeta)\ d\beta_i$$

$$= f(N_i|\lambda_i) \int_0^\infty f(\lambda_i,\beta_i|\zeta)\ d\beta_i .$$

That is,

$$f(N_i,\lambda_i|\zeta) = f(N_i|\lambda_i)\ f(\lambda_i|\zeta), \qquad (4.13)$$

which is a product of (4.12) and (4.6).

4.3 Inference on the Parameters of Interest

To estimate ζ using data from the 4 processes, we require $f(N_1,N_2,N_3,N_4|\zeta)$. First, note that

$$f(N_i,\lambda_i,\beta_i|\zeta) = f(N_i|(\lambda_i,\beta_i),\zeta)\ f(\lambda_i,\beta_i|\zeta).$$

In section 4.2, we assumed that the λ_i's are conditionally independent given the β_i's and ζ, and that the N_i's are independent given the λ_i's. Then:

1) the conditional joint distribution of the 4 processes at time t is given by:

$$f(N_1,N_2,N_3,N_4|(\lambda_1,\beta_1),(\lambda_2,\beta_2),(\lambda_3,\beta_3),(\lambda_4,\beta_4),\zeta)$$

$$= \prod_{i=1}^{4} f(N_i|(\lambda_i,\beta_i),\zeta);$$

2) $f(N_1,N_2,N_3,N_4,(\lambda_1,\beta_1),(\lambda_2,\beta_2),(\lambda_3,\beta_3),(\lambda_4,\beta_4)\,|\,\zeta)$

$$= \prod_{i=1}^{4} f(N_i\,|\,(\lambda_i,\beta_i),\zeta)\;\; f(\lambda_i,\beta_i\,|\,\zeta). \qquad (4.14)$$

Thus, to obtain $f(N_1,N_2,N_3,N_4\,|\,\zeta)$, we eliminate the λ_i's and β_i's in (4.14) by integrating them over $(0,\infty)$. We first integrate out the λ's and obtain

$$\int_0^\infty \cdots \int_0^\infty \prod_{i=1}^{4} f(N_i\,|\,(\lambda_i,\beta_i),\zeta)\;\; f(\lambda_i,\beta_i\,|\,\zeta)\; d\lambda_1 \ldots d\lambda_4$$

$$= \prod_{i=1}^{4} f(N_i\,|\,\beta_i,\zeta)\;\; f(\beta_i\,|\,\zeta). \qquad (4.15)$$

Then, we integrate the β_i's in (4.15) over its entire domain to get

$$\int_0^\infty \cdots \int_0^\infty \prod_{i=1}^{4} f(N_i\,|\,\beta_i,\zeta)\quad f(\beta_i\,|\,\zeta)\; d\beta_1 \ldots d\beta_4$$

$$= f(N_1,N_2,N_3,N_4\,|\,\zeta). \qquad (4.16)$$

Using the result in equation (4.13) and the assumptions of conditional independence, we have

$$f(N_1,N_2,N_3,N_4\,|\,\zeta)$$

$$= \int_0^\infty \cdots \int_0^\infty \prod_{i=1}^{4} \;\; f(N_i\,|\,\lambda_i)\; f(\lambda_i\,|\,\zeta)\; d\lambda_i$$

$$= \int_0^\infty \cdots \int_0^\infty \prod_{i=1}^{4} \left[\prod_{j=t-1}^{t+1} \prod_{k=1}^{2} \frac{e^{-\lambda_i}\;\lambda_i^{n_{ijk}}}{n_{ijk}!} \right] \frac{\zeta}{(\lambda_i+\zeta)^2}\, d\lambda_i,$$

that is,

$$f(N_1,N_2,N_3,N_4|\zeta)$$

$$= \int_0^\infty \cdots \int_0^\infty \prod_{i=1}^{4} \frac{e^{-6\lambda_i} \lambda_i^{n_{i++}}}{P_i} \frac{\zeta}{(\lambda_i + \zeta)^2} d\lambda_i, \qquad (4.17)$$

where

$$n_{i++} = \sum_{j=t-1}^{t+1} \sum_{k=1}^{2} n_{ijk}$$

and

$$P_i = \prod_{j=t-1}^{t+1} \prod_{k=1}^{2} n_{ijk}!.$$

Note that $f(N_1,N_2,N_3,N_4|\zeta)$ given in equation (4.17) is a product of $f(N_i|\zeta)$ and is similar to the result for the unreplicated case given in equation (4.9). Though the function in equation (4.9) itself is unimodal, the analogous result in the present case has not been established. The shape of this product integral evaluated in equation (4.17) with numerous different combinations of n_{ijk}'s at various values of ζ's suggest the result is true. Numerical methods are used to compute the estimate of ζ that maximizes the function in equation (4.17) in spite of the potential risk of multi-modality.

The smoothing window used for estimating ζ is small but the $\hat{\zeta}(t)$ obtained may still contain a fair amount of noise and sampling variation. A numerical method called SABL, described in section 2.2.2, can be used to decompose this $\hat{\zeta}$-series into trend, seasonal and irregular components. The seasonal and irregular components, which are usually of much smaller magnitude when compared with the trend, reflect some of the sampling variation and noise still in the $\hat{\zeta}$-series. The trend is smoother than the $\hat{\zeta}$-series, and is closely related to the trends of $\hat{\lambda}_i(t)$'s for the 4 processes so we will use it in this study as a general summary of the data in the domains for which $\hat{\zeta}$ is computed.

With the smoothed version of $\hat{\zeta}(t)$, we can compute the posterior point estimates for λ_i at each t by maximizing an equation similar to that of (4.9) with respect to λ_i. So for each i = 1,...,4 and each time t,

$$\max_{\lambda_i} f(\lambda_i | N_i, \hat{\zeta}) = \max_{\lambda_i} \frac{e^{-6\lambda_i} \lambda_i^{n_{i++}}}{P_i} \frac{\hat{\zeta}}{(\lambda_i + \hat{\zeta})^2}, \qquad (4.18)$$

where

$$n_{i++} = \sum_{j=t-1}^{t+1} \sum_{k=1}^{2} n_{ijk} \text{ and } P_i = \prod_{j=t-1}^{t+1} \prod_{k=1}^{2} n_{ijk}!.$$

This maximization problem is easier to handle when we take the natural logarithm of (4.18) to obtain a quadratic expression of λ_i. The maximum can be explicitly evaluated in this case and is found to be at

$$\hat{\lambda}_i^* = \frac{(n_{i++} - 2 - 6\hat{\zeta}) + [(n_{i++} - 2 - 6\hat{\zeta})^2 + 24\, n_{i++}\hat{\zeta}\,]^{1/2}}{12}, \qquad (4.19)$$

where $n_{i++} = \sum_{j=t-1}^{t+1} \sum_{k=1}^{K} n_{ijk}$.

The $\hat{\lambda}_i^*$ in equation (4.19) is only 1/6 of the actual λ_i so it is multiplied by 6 (2 for the number of replicates times 3 for the window size). Now, let $\hat{\lambda}_i = 6\hat{\lambda}_i^*$ which is comparable in magnitude with the observations of the ith process. It would be desirable to present a $100(1-\alpha)\%$ credibility interval for λ_i at each t. Each interval is a subset, C, of the parameter space which gives the probability that λ_i is in C. This amounts to choosing a pair of upper and lower limits, (a,b), such that

$$F(\hat{\zeta})^{-1}\int_a^b f(\lambda_i|N_i,\hat{\zeta})\,d\lambda_i = 1-\alpha, \tag{4.20}$$

where $a \geq 0$ and $b > 0$, and

$$F(\hat{\zeta}) = P_i^{-1}\int_0^\infty e^{-6\lambda_i}\,\lambda_i^{n_{i++}}\,\hat{\zeta}\,(\lambda_i+\hat{\zeta})^{-2}\,d\lambda_i,$$

$$n_{i++} = \sum_{j=t-1}^{t+1}\sum_{k=1}^{2} n_{ijk} \quad\text{and}\quad P_i = \prod_{j=t-1}^{t+1}\prod_{k=1}^{2} n_{ijk}!.$$

In choosing a credibility interval for λ_i at time t, the usual approach is to minimize its length. This may be done by using the highest posterior density(HPD) criterion which is to include in the interval only those points with HPD, that is, the 'most likely' values of λ_i.

To evaluate the HPD credibility interval in equation (4.20), we set up the following program:

1) set the lower limit $a = 0$;

2) create a subroutine which, for a given time t, computes the

maximum of $f(\lambda_i|N_i,\hat{\zeta})$ as a function of λ_i at, say, $\lambda_i = m$;

3) set $x = (a+m)/2$ and evaluate $f(\lambda_i|N_i,\hat{\zeta})$ at $\lambda_i = x$;

4) create a subroutine which find the value of b such that $f(\lambda_i = b\ |N_i,\hat{\zeta}) = f(\lambda_i = x\ |N_i,\hat{\zeta})$;

5) numerically integrate

$$P(a,b) \quad = \quad F(\hat{\zeta})^{-1} \int_a^b f(\lambda_i|N_i,\hat{\zeta})\ d\lambda_i,$$

where a and b are the values from steps (3) and (4).

If this value is approximately $(1-\alpha)$, then stop. If not and if:

i) the value is larger than $(1-\alpha)$, set $m = x$ (a remains unchanged) and go back to steps (3) to (5);

ii) the value is smaller than $(1-\alpha)$, set $a = x$ (m remains unchanged) and return to steps (3) to (5).

It can happen that the integral $P(a,b)$ evaluated from $a = 0$ to the point where $f(\lambda_i|N_i,\hat{\zeta})$ has its maximum is very small so that the above procedure cannot be used. In such situations, we abandon the HPD criterion and find b such that $P(a,b)$ evaluated at $(0,b)$ is approximately equal to $(1-\alpha)$.

We can plot all these results with lines connecting the point estimates of λ_i and their corresponding limits over time for visual inspection. These graphs should roughly indicate the trend of λ_i's and the size of their possible fluctuations.

5. APPLICATION

5.1 Introduction

In this section, the method developed in the last chapter is applied to the marked salmon recovery data. For coho, data from replicated tag codes observed in 4 catch regions are used. These data are further grouped according to hatchery, brood year and size at release of the tagged smolts. The hatcheries of interest are Quinsam with broods from 1978 and 1979, and Capilano with broods from 1979 and 1980. An example of such a grouping is (Quinsam, 1979, large) which refers to smolts which are raised in the Quinsam hatchery starting in 1979 with a large average size at release. For each grouping, there are 4 similar sets of replicated tag codes. Thus, in each catch region, a total of 12 codes are included for each possible hatchery, brood year and size combination. Details of these tag codes can be found in Appendix F of 'A Canadian MRP Data Benchmark'(1986).

Overall, the recovery data set for replicated tag codes of chinook has very few observations. Even with 4 sets of replicates for each hatchery, brood year and size combination, few recoveries are found over the 63 adjusted sampling periods for the 3 catch regions considered. Therefore, over almost the entire recovery period, the estimated intensity, $\hat{\lambda}_i$, is near zero for each catch region. Three tag codes (021827, 021829 and 021661) which have substantial recoveries are thus selected from the benchmark rollup data subset. The method described is then applied to these data.

The local Poisson model allows us to aggregate data from related tag codes when there is a paucity of information for a single tag. The use of this approach on two different sets of data, one from replicated tag codes and the other from individual codes, illustrates the flexibility and advantage of our method.

5.2 Problems Encountered when Modelling the Salmon Recovery Data

5.2.1 Missing values

In fitting local Poisson models to the recovery data, the ζ's are estimated from small neighborhoods where counts are homogeneous. As indicated in subsection 3.2.2, usually the first 10 and the last 10 periods for all catch regions in each recovery year contain missing values so our approach cannot be used to obtain ζ-estimates over these long intervals. We thus include at most two periods with missing values to open or to close a sampling interval. As a result, there are 18 periods for coho and 63 adjusted periods for chinook in which the interval between these periods is more or less equally spaced.

Suppose there are a few missing values in the recovery data over m periods of interest. An ad hoc method is used to obtain a number for each period with missing values. Assume that a symmetric neighborhood with a 3-point window is used for smoothing. For a fixed catch region, let the 3 counts in a window be denoted by n_{t-1}, n_t and n_{t+1}, where t = 1,...,m, $n_0 = 0$ and $n_{m+1} = 0$. Then, the procedure adopted to fill in missing value in that window is as follows:

1) if n_{t-1} is missing and there are data on the left of n_{t-1}, search backward for at most 5 steps to obtain the first data value n_{t-j}, where j = 2,..., or 6, for n_{t-1}. When n_{t-1} is in the first position or when there are only missing values before it, set $n_{t-1} = 0$.
2) if n_t is missing and there are data close to the right of n_t, search forward for at most 5 steps to find the first available count n_{t+j}, where j = 2,..., or 6, for n_t. When there are only missing values following n_t, set $n_t = 0$.
3) if n_{t+1} is missing and there are data points just beyond n_{t+1}, search forward for at most 5 steps and set $n_{t+1} = n_{t+j}$ for the first j = 2,..., or 6 which is not missing. If n_{t+1} is at the end or there are only missing values after it, set $n_{t+1} = 0$.

In a few cases when there is more than one missing value in a window, we use an appropriate combination of the above three cases to obtain counts for these periods. The decision to search for at most 5 steps is based on the general pattern of missing values in the sampling periods for two of the species of salmon. Having 3 or more consecutive missing values is rare, except for one chinook tag code.

5.2.2 The edge effect

The *edge effect*, which occurs at the two extremes of the $\hat{\zeta}$-series, is a byproduct of the local smoothing technique in our model. To obtain $\hat{\zeta}$'s for m periods, we require counts for 2 extra periods, n_o and n_{m+1}, to obtain estimates for the first and the last windows. Here, we set n_o and n_{m+1} to zero since in most cases no more than one recovery was observed in the first and last periods. These two estimates must therefore be regarded with caution.

5.3 Fitting the Proposed Models to the Selected Data Sets

For a particular hatchery, brood year and size combination, consider the ith catch region at time t, where $t = 1,\ldots,m$. Then $N_i = (N_{ikl}(t))$ is a matrix of counts where $k = 1,\ldots,K$ denotes the number of similar sets of replicated tag codes, and $l = 1,\ldots,L$ denotes the number of replicates in each of the K sets. For coho, $K = 4$, $L = 3$, and $t = 16,\ldots,33$ ($m = 18$ actual periods) for each $i = 1,\ldots,4$. For chinook, $K = 1$, $L = 1$, and $t = 1,\ldots,63$ ($m = 63$ adjusted periods) for each $i = 1,2,3$. In each case, the smoothing technique uses a symmetric neighbourhood with a 3-point window, and the sampling periods are the same for the catch regions considered.

Now, following equation (4.12), the joint density of these $N_{ikl}(t)$'s given λ_i for a fixed catch region i at time t is

$$f(N_i|\lambda_i) = \prod_{j=t-1}^{t+1} \prod_{k=1}^{K} \prod_{l=1}^{L} f(N_{ikl}(j)|\lambda_i). \tag{5.1}$$

The ζ's are estimated for each group of replicates by maximizing a modified equation similar to that of (4.17). That is, we maximize the following function for I catch regions with respect to ζ:

$$f(N_1,\ldots,N_I|\zeta)$$

$$= \prod_{i=1}^{I} \int_0^{\infty} \Bigg[\prod_{j=t-1}^{t+1} \prod_{k=1}^{K} \prod_{l=1}^{L} \frac{e^{-\lambda_i} \lambda_i^{n_{ijkl}}}{n_{ijkl}!} \Bigg] \frac{\zeta}{(\lambda_i + \zeta)^2} \, d\lambda_i, \tag{5.2}$$

where t = 1,...,m.

The $\hat{\zeta}$-series obtained from each group of codes are first plotted over the periods for visual inspection. Figures 5.1a is an example for coho of the Quinsam, 1979 and medium combination, Figures 5.2a is for a similar groupings of Capilano coho, and Figure 5.7a is the $\hat{\zeta}$-plot for chinook tag code 021827. These plots show for each group, that the $\hat{\zeta}$-series is not very smooth and has a different trend for each group.

The computer routine SABL, described in subsection 2.2.2, is used to smooth these $\hat{\zeta}$'s. The $\hat{\zeta}$'s associated with tag codes from the Quinsam hatchery are transformed by taking their fourth root. A plot of these results is illustrated in Figures 5.1b. For the remaining tag codes, no transformation is used on the $\hat{\zeta}$'s. Examples of the resulting trends of ζ's are then given in Figures 5.1c, 5.2b and 5.3b.

From now on, we will use the smoothed version of $\hat{\zeta}(t)$, denoted by $\tilde{\zeta}(t)$, as a general summary of the data. The $\tilde{\zeta}$ curves for all brood year and size combinations of Quinsam coho are plotted against time in Figure 5.4. In general, the shapes of these ζ-plots exhibit two humps connected by a small dip. The second hump, which is between periods 26 and 33, is usually flatter than the first one, which is between periods 16 and 22. For brood year 1978, the size of the $\tilde{\zeta}$'s for each size-group is about the same. However, for brood year 1979, the medium size-group has larger $\tilde{\zeta}$'s, which indicates that there is more β dispersion among the 4 different catch regions. For Capilano coho, as shown in Figure 5.5, the shape of the $\tilde{\zeta}$ curves is unimodal, with the mode usually in period 21. The $\tilde{\zeta}$-plots of the medium size coho for both brood years 1979 and 1980 have a similar magnitude. However, for the other two sizes, there are big differences between the two broods.

Note that for each brood year and hatchery, if the $\tilde{\zeta}$-series of one size group is fixed, the $\tilde{\zeta}$ curves of each remaining group is only a constant multiple of it. Also, in general, larger $\tilde{\zeta}$-values indicate that there is more β dispersion among different catch regions while smaller $\tilde{\zeta}$-values reflect that the β_i's are more similar.

The $\tilde{\zeta}$-plots for the 3 chinook tag codes are given in Figure 5.6. Clearly, we can observe that there are 3 distinct recovery intervals, with each interval of 21 periods representing a fishing season. Within each interval, an increasing trend in the $\tilde{\zeta}$-series is always followed by a decrease, and the largest peak comes near the end. Further, the $\tilde{\zeta}$'s for the first two intervals are larger in magnitude than the third. When comparing the $\tilde{\zeta}$'s of the three tag codes, those of 021827 are usually smaller even during the peak periods. This result reflects that the β_i's are more similar among the 3 catch regions for tag code 021827.

With the smoothed version of $\hat{\zeta}$-series, $\hat{\lambda}^*_i$ can be obtained for each catch region using the modified equation (4.19) that includes all replicates with similar characteristics. We then have

$$\hat{\lambda}^*_i = \frac{(n_{i+++} - 2 - A\tilde{\zeta}) + [(n_{i+++} - 2 - A\tilde{\zeta})^2 + 4A\, n_{i+++}\tilde{\zeta}\ ]^{1/2}}{2A}, \qquad (5.3)$$

where

$$n_{i+++} = \sum_{j=t-1}^{t+1} \sum_{k=1}^{K} \sum_{l=1}^{L} n_{ikl}(j),$$

and

A = 3 × K × L (3 is the window size).

The final result is multiplied by A to obtain $\hat{\lambda}_i$ for the ith region.

Figures 5.7 to 5.9 are plots of the estimated recovery intensities, $\hat{\lambda}_i$'s, for the 3 sizes of Quinsam coho. Note that $\hat{\lambda}(t)$'s, that is, the recovery intensity levels for the three trolling regions: Southwest Vancouver Island Troll, Georgia Strait Troll and South Central Troll, peak much earlier than the Johnstone Strait Net. In addition, the recovery intensity for the Johnstone Strait Net is the largest among the 4 catch regions for all sizes.

For Capilano coho the $\hat{\lambda}$-plots are shown in Figures 5.10 to 5.12. Like the Quinsam coho, all trolling regions here have their $\hat{\lambda}$ curves peaking earlier than the Johnstone Strait Net. Nevertheless, the general pattern for each of the four catch regions is different from that of Quinsam coho. First, the recovery intensity is often the largest for Southwest Vancouver Island Troll and the smallest for Johnstone Strait Net. Second, from periods 28 to 33, the intensity is usually zero for all catch regions.

Now, we turn to examine the relationship between the smolts' size at release and their recovery intensity at maturity, and the difference in recovery rate between different brood years. For the first case, we have to study the estimated recovery intensities of the 3 sizes: small, medium and large, for each brood year separately. However, no obvious conclusions are suggested since there is so much variation among the 4 catch regions. For the second case, we compare the estimated intensity for different brood years and observe that:

1) for Quinsam coho, the recovery intensities for coho raised in 1979 are larger than those from 1978; and

2) for Capilano coho, the recovery intensities for brood year 1980 are much larger than for 1979.

These results indicate that for coho from a particular hatchery more recoveries are observed from one brood than the other on average.

The $\hat{\lambda}$-plots for the 3 chinook tag codes in each catch region are presented in Figure 5.13. Here, the trolling regions are: Northwest Vancouver Island Troll, Northern Troll, and North Central Troll. Again, as in the $\tilde{\zeta}$-plots, the recovery intensities are distributed in 3 intervals, with larger intensities found in the first two intervals. Among the 3 catch regions, the Northern Troll has larger recovery rate on average. When comparing the estimated intensities for the 3 tag codes, overall 021661 and 021827 has better recovery rate than 021827. This indicates that on average fewer recoveries are observed with tag code 021827.

After computing the $\hat{\lambda}$-series, we derived the 95% credibility intervals for the λ's using the procedure described in section 4.3. That is, we found the upper and lower limits (a,b) such that

$$G(a,b,\tilde{\zeta}) = F(\tilde{\zeta})^{-1} \int_{a}^{b} \frac{e^{-A\lambda_i} \lambda_i^{n_{i+++}}}{P_i} \frac{\tilde{\zeta}}{(\lambda_i + \tilde{\zeta})^2} d\lambda_i = 0.95, \qquad (5.4)$$

where

$A = 3 \times K \times L$ (3 is the window size),

$$n_{i+++} = \sum_{j=t-1}^{t+1} \sum_{k=1}^{K} \sum_{l=1}^{L} n_{ikl}(j),$$

$$P_i = \prod_{j=t-1}^{t+1} \prod_{k=1}^{K} \prod_{l=1}^{L} n_{ikl}(j)!,$$

and

$$F(\tilde{\zeta}) = P_i^{-1} \int_0^{\infty} e^{-A\lambda_i} \lambda_i^{n_{i+++}} \tilde{\zeta} (\lambda_i + \tilde{\zeta})^{-2}.$$

Note that in the case when the total count n_{i+++} and $\hat{\lambda}_i$ are zero, and $\tilde{\zeta}$ is close to zero, the corresponding credibility interval for λ_i is $(0,\epsilon)$ where $\epsilon > 0$ is arbitrarily small. In this case, we are sure that there is no recovery in the ith catch region. The result is obtained to a good approximation by taking the limit of $G(a,b,\zeta)$ in (5.4) as $\zeta \to 0$. The proof of the result $\lim_{\zeta \to 0} G(a,b,\zeta) = 1$ is shown in the following.

First, note that

$$\lim_{\zeta \to 0} G(a,b,\zeta) = \lim_{\zeta \to 0} \frac{\int_0^b e^{-\lambda_i} / (\lambda_i + \zeta)^2 \; d\lambda_i}{\int_0^{\infty} e^{-\lambda_i} / (\lambda_i + \zeta)^2 \; d\lambda_i}. \tag{5.5}$$

Letting $u_i = \lambda_i + \zeta$, (5.5) becomes

$$\lim_{\zeta \to 0} \frac{\int_{\zeta}^{b+\zeta} \left[e^{-(u_i - \zeta)} / u_i^2 \right] du_i}{\int_{\zeta}^{\infty} \left[e^{-(u_i - \zeta)} / u_i^2 \right] du_i}. \tag{5.6a}$$

The expression in (5.6a) is just

$$\lim_{\zeta \to 0} \frac{\int_{\zeta}^{b+\zeta} e^{-u_i} u_i^{-z} du_i}{\int_{\zeta}^{\infty} e^{-u_i} u_i^{-z} du_i}. \qquad (5.6b)$$

Since in equation (5.6b), both numerator and denominator are analytical functions, they are differentiable. Further, they are both infinite in the limit; therefore, we can apply the L'Hospital rule to (5.6b) and obtain

$$\lim_{\zeta \to 0} \frac{e^{-(b + \zeta)} (b + \zeta)^{-z} - e^{-\zeta} \zeta^{-z}}{-(e^{-\zeta} \zeta^{-z})}$$

$$= \lim_{\zeta \to 0} \zeta^{-z} e^{-b} (b + \zeta)^{-z} + 1$$

$$= 1.$$

Sample plots of these credibility intervals for coho and chinook and their corresponding λ-estimates are portrayed in Figures 5.14 to 5.16. We must be careful in interpreting these credibility intervals. The bands in Figures 5.14 to 5.16 are not simultaneous interval estimates. They merely indicate the pointwise credibility interval for the recovery intensity, λ(t), at each period t. Note that during some periods when a region has substantial recoveries, the ratio of the $\hat{\lambda}$ and the width of theses intervals on average for that region is about two or three times smaller than regions with much smaller recoveries. This indicates that large recoveries are in general more informative for estimating the recovery intensities and computing the corresponding credibility interval just as intuition would suggest.

5.4 Conclusion

To identify an appropriate model for the salmon recovery data, we have examined the raw data in detail. We learned that the salmon recovery data set has missing values, huge extraneous variations and noise. An empirical Bayes model was thus developed to handle these data. The method fitted local parametric, Poisson models to be precise, and incorporated this approach with a conventional smoothing technique to obtain the overall recovery patterns and the corresponding credibility intervals for the underlying salmon recoveries.

The resulting conclusions are:

1) there are huge variations in the pattern of recovery intensities for both species of salmon from different brood years;
2) in all catch regions, the overall recovery trends for coho from the two hatcheries are different in shape;
3) no overall difference is observed for the three sizes of coho;
4) in general the Quinsam coho recovery is usually the largest in the Johnstone Strait Net area while for Capilano this is largest in Southwest Vancouver Island Troll area; and

5) among the 3 intervals of recovery for the chinook, the recovery rate is always the largest during the first two periods.

Our method has demonstrated its usefulness in aggregating information from individual sampling periods to overcome the problem of sparse data. However, we suggest that further investigations be carried out, including

1) extending our method to include general parametric models, and
2) finding a weighting scheme for smoothing so that the weight given to a point depends on the distance it is from the 'period' of interest.

REFERENCES

Becker, R. A. and Chambers, J. M. (1984). *S: An Interactive Environment for Data Analysis and Graphics*. California: Wadsworth.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag.

Brewster, J. F. and Zidek, J. V. (1974). Improving on equivariant estimators. *Annals of Statistics*, **2**, 21-38.

A Canadian MRP Data Benchmark (1986). Fisheries Research Branch, Department of Fisheries and Oceans, Canada.

English, K. K. (1985). The contribution of hatchery produced chinook and coho to west coast fisheries: preliminary analysis. Department of Fisheries and Oceans, Canada.

Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, **1**, 297-318.

Holden, R. T. (1987). Time Series Analysis of a Contagious Process. *J. Amer. Statist. Assoc.*, **2**, 1019-1026.

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics*, **15**, 209-225.

Ma, H., Joe, H. and Zidek, J. (1986). A Bayesian Nonparametric Univariate Smoothing Method, with Applications to Acid Rain Data Analysis. University of British Columbia Statistics Department Technical Report No. 47.

Nicholls, D. F., Heathcote, C. R. and Cunningham, R. B. (1986). The Evaluation of Long Term Trend I. *Austral. J. Statist.*, **28**, 294-313.

Salmon Stock Interpretation Unit (1984). The mark recovery program as an assessment tool for the hatchery chinook and coho salmon enhancement program. Fisheries Research Branch, Department of Fisheries and Oceans, Canada.

Salmon Stock Interpretation and Assessment Unit (1986). Development of a Pacific coastal database for assessing the contribution of B.C. hatchery chinook and coho salmon production to the Canadian commercial catch. Fisheries Research Branch, Department of Fisheries and Oceans, Canada.

Weerahandi, S. and Zidek, J.V. (1988). Bayesian nonparametric smoothers for regular processes. *The Canadian Journal of Statistics*, **16**, 61-74.

APPENDIX

The Mark Recovery Program(MRP) Database

A two-phase marking program, which includes tagging fish in the hatchery and recovering tags in the fishery, is currently the best known method for providing information to assess the benefits of artificially reared fish. The most popular marking technique for hatchery coho and chinook is the use of coded wire tags (CWTs), which are usually inserted in the snouts of juvenile fish as an indicator of the fish's origin. In addition, the adipose fins of these tagged fish are clipped to allow detection of them later as adults.

Two main types of data are thus available from this marking program. One is the hatchery release data and the other is the recovery data. The first type is of two broad categories:

1) fish released with a CWT and clipped adipose fin (marked), and

2) fish released without a CWT (unmarked).

The second type of data include fishery data giving recoveries from commercial and sport fishing, and escapement data which are recoveries not from any fishery.

In each category of the release data, there are time variables, such as the brood year (the year in which the eggs were spawned) and the date of release. There are also geographic variables which include stock site (location from which eggs are taken), hatchery site and release site. In addition, an average size (gram/fish) and the actual number of releases are given to each released group of fish. The above variables represent only a small subset in the release data. Many others variables related to the survival of fish are also available.

Note that in order to calculate the total contribution of hatchery fish to fisheries, each hatchery release group must be represented by a group of marked fish. A method has been developed to determine the marked release group that would represent the release of unmarked juveniles. Thus, the different variables in the release data are also important for associating marked and unmarked releases.

In the fishery data, there are information about recoveries of marked fish and the corresponding sampling program. Whenever possible, the recovery time is recorded as year, month and statistical weeks (about 5 per calendar month), and the recovery region corresponding to the fishing method is also recorded. Not every fish caught by a commercial fishery is inspected for tag since this is costly. Thus, a sample is taken for mark inspection and only those marks detected in the sample are recorded as data. The sport recovery data usually come from voluntary returns of salmon heads by fisherman. Thus, these recoveries depend very much on the fishermen's awareness of the clipped adipose fins. The sport catch size is obtained differently from commercial catch, which is estimated based on sales slips.

The escapement data include tagged fish escapement to the hatchery or escapement to rivers or lakes near the hatchery. Detailed recovery information on a single tagged fish is difficult to obtain so only yearly bulk data are available.

A brief review of the life cycle of chinook and coho would show us how recoveries from a particular brood are distributed over time. A coho's or chinook's life begins as an egg in the year of spawning, the 'brood year'. Eggs are hatched, and juvenile fish are reared until the 'release year', when they are allowed to leave the hatchery and begin life in the ocean. Eventually, in the 'recovery year', adults are captured by the fishery or escape to their spawning ground. Therefore, if the brood year is defined to be year 0, then the following table from the Salmon Stock Interpretation and Assessment Units(1986) report is a summary for the majority of coho and chinook.

Year	chinook	coho
0	brood year	brood year
1	release year	release year (fry)
2		release year (smolts)
3	recovery year (age 3)	recovery year (age 3)
4	recovery year (age 4)	
5	recovery year (age 5)	

In some cases, some fish of both species are recovered in year 2 as jacks and chinook are sometimes recovered in years 6 and 7.

These release and recovery data of tagged and untagged salmon have been collected over years, but they were unorganized and scattered among different agencies. Thus, it is difficult to have an analysis on a complete set of data. In 1983, the Canadian Department of Fisheries and Oceans(DFO) decided to construct a Mark Recovery Program(MRP) database on the VAX computer at the Pacific Biological Station(PBS) in Nanaimo. As a result, many interesting questions can now be addressed. When this database is completed, valuable information will be available for assessing the coho and chinook salmon enhancement program.

Table 3.1. The tag codes found in the benchmark data subset.

code used in plots	original code	species Hart code
1	020408	124 (chinook)
2	020409	124
3	021615	124
4	021635	124
5	021661	124
6	021827	124
7	021829	124
8	022202	124
9	022405	124
1	081810	115 (coho)
2	081811	115
3	081812	115
4	081813	115
5	081841	115
6	081842	115
7	081843	115
8	081844	115
9	081845	115
10	082001	115
11	082002	115
12	082003	115
13	082004	115
14	082005	115
15	082006	115
16	082007	115
17	082008	115
18	082009	115
19	082019	115
20	082020	115
21	082021	115
22	082022	115
23	082023	115
24	082024	115
25	082025	115
26	082026	115
27	082027	115

Table 3.2. Summary of data fields for the benchmark release data subset.

field	description	number of zeros	% of zeros
1	tag code	0	0
2	species Hart code	0	0
3	brood year	0	0
4	run type code	27	75.00
5	day first released	29	80.56
6	month first released	29	80.56
7	year first released	29	80.56
8	day last released	0	0
9	month last released	0	0
10	year last released	0	0
11	number tagged	0	0
12	adipose only	0	0
13	unclipped	0	0
14	total released	0	0
15	number of days held	5	13.89
16	size code	0	0
17	size at release	0	0
18	percentage tag loss	0	0
19	expected survival	36	100
20	stage code	0	0
21	study type	0	0
22	hatchery code	0	0
23	release site code	0	0
24	stock site code	0	0
25	agency code	0	0
26	co-ordinator code	0	0
27	production area code	0	0
28	province/state code	0	0
29	years with recoveries	0	0
30	release type	0	0
31	total associated release	0	0

Table 3.3. Summary list of chinook data fields for the benchmark rollup recovery subset.

(There are four possible recovery methods: troll, net, sport(S), or escapement(E). The letter in square brackets indicates that only one of these methods applies to the field. Those fields without any letters apply to all methods. The number and percentage of NAs were calculated according to the number of records corresponding to a particular catching method.)

* NAs = missing values

field	description	number of NAs	% of NAs
1	tag code	0	0
2	recovery year	0	0
3	gear	0	0
4	catch region	0	0
5	brood year	0	0
6	non-tag indicator	0	0
7	species Hart code	0	0
8	statistical week	0	0
9	average fork length (mm)	1361	98.27
10	average hyperal length (mm)	1385	100
11	average total length (mm)	1385	100
12	average dress weight (kg)	1385	100
13	average round weight (kg)	1385	100
14	% immature female	1385	100
15	% mature female	1385	100
16	% immature male	1385	100
17	% mature male	1385	100
18	% unknown sexual maturity	35	2.53
19	recovery site code [E]	0 out of 35	0
20	recovery site number [E]	0 "	0
21	run type [E]	0 "	0
22	sample age type [E]	0 "	0
23	number of observed recoveries	0	0
24	catch or escapement	67	4.84
25	sample size	67	4.84
26	sum of known tags	5	0.36
27	number of no-pins	138	9.96
28	number of lost-pins	347	25.05
29	number with no data	614 out of 1350	45.48
30	number of sport marks observed [S]	68 out of 89	76.40
31	est. marks in the est.sport catch [S]	68 "	76.40
32	number observed sport recoveries [S]	0	0
33	sum of escapement non-tags [E]	3 out of 35	8.57

Table 3.4. Table for computing "Period" from statistical week (MMW).

Week	Jan 1	Feb 2	Mar 3	Apr 4	May 5	Jun 6	Jul 7	Aug 8	Sep 9	Oct 10	Nov 11	Dec 12
1	40	40	1	5	10	14	18	23	27	31	36	40
2	40	40	2	6	11	15	19	24	28	32	37	40
3	40	40	3	7	12	16	20	25	29	33	38	40
4	40	40	4	8	13	17	21	26	30	34	39	40
5	40	0	0	9	0	0	22	0	0	35	0	0

"Period" is a number ranging between 1 and 40 representing a one week time period during which salmon fishing may occur.

Table 3.5. List of catch region codes, and names.

new code	old code	Name
1	1	NW Vancouver Is. Troll
2	2	SW Vancouver Is. Troll
3	3	Washington/Oregon Troll
4	4	Georgia Strait Troll
5	5	Central Troll
6	6	Northern Troll
7	7	Alaska Troll
8	14	Juan de Fuca Troll
9	15	NW Vanc. Is. and Central Troll
10	17	NW Vanc. Is. and SW Vanc. Is. Troll
11	18	Northern and Central Troll
12	34	Georgia Strait and Central Troll
13	53	Georgia Strait and SW Vanc. Is. Troll
14	56	North Central Troll
15	57	South Central Troll
16	8	Fraser Gillnet
17	9	Northern Net
18	10	Georgia Strait Net
19	11	Johnstone Strait Net
20	12	Central Net
21	13	Juan de Fuca Net
22	19	Johnstone Strait and Central Net
23	20	NW Vancouver Is. Net
24	21	SW Vancouver Is. Net
25	33	Northern and Central Net
26	36	Yukon Net
27	37	Juan de Fuca and Georgia Strait Net
28	45	Johnstone Strait and Georgia Strait Net
29	46	Fraser Gillnet and Georgia Strait Net
30	47	Alaska Net
31	48	British Columbia Net
32	58	Fraser Seine Net
33	25	Northern Sport
34	26	Central Sport
35	27	Washington Sport
36	28	Georgia Strait Sport
37	29	Freshwater Sport
38	99	Canadian Escapement

Table 3.6. A summary of catch regions with observed recoveries.

(New catch region codes from Table 3.5 are used here.)

Description	COHO	CHINOOK
# Troll catch regions	10	6
Catch region codes	1,2,4,5,6,9,10, 11,14,15*	1,4,6,11,14,15
# Net catch regions	9	6
Catch region codes	16,17,18,19*,20, 21,22,23,24	17*,18,19,20,22, 24
# Sport catch regions	4	4
Catch region codes	34,35,36*,37	33,34,35,36*
# tags considered	153	69
# records examined	2989	940

* : Catch region with more than a hundred observed recoveries.

Table 3.7. A sample list of coho release replicates that are classified according to size.

(All the replicates are in groups of three.)

release date	tag code	release site	brood year	size	total rel.	total obs.	% rec.	total esp.
10/5/81	081855	Quinsam	1979	small	7189	123	1.71	87
	59				7191	130	1.81	86
	62				7192	111	1.54	71
10/5/81	081856	Quinsam	1979	medium	7192	144	2.00	108
	58				7210	114	1.58	88
	61				7193	115	1.60	86
10/5/81	081857	Quinsam	1979	large	7202	148	2.05	110
	60				7192	146	2.03	116
	63				7207	134	1.86	99
26/5/81	081910	Capilano	1979	small	4098	135	3.29	49
	11				4093	115	2.81	48
	12				3845	103	2.68	42
26/5/81	081913	Capilano	1979	medium	3983	123	3.09	51
	14				4038	127	3.15	54
	42				4208	139	3.30	67
26/5/81	081943	Capilano	1979	large	3516	91	2.59	46
	44				3570	102	2.86	57
	45				3565	81	2.27	46

Figure 3.1a. Size of chinook release for tag codes from the benchmark release data subset.

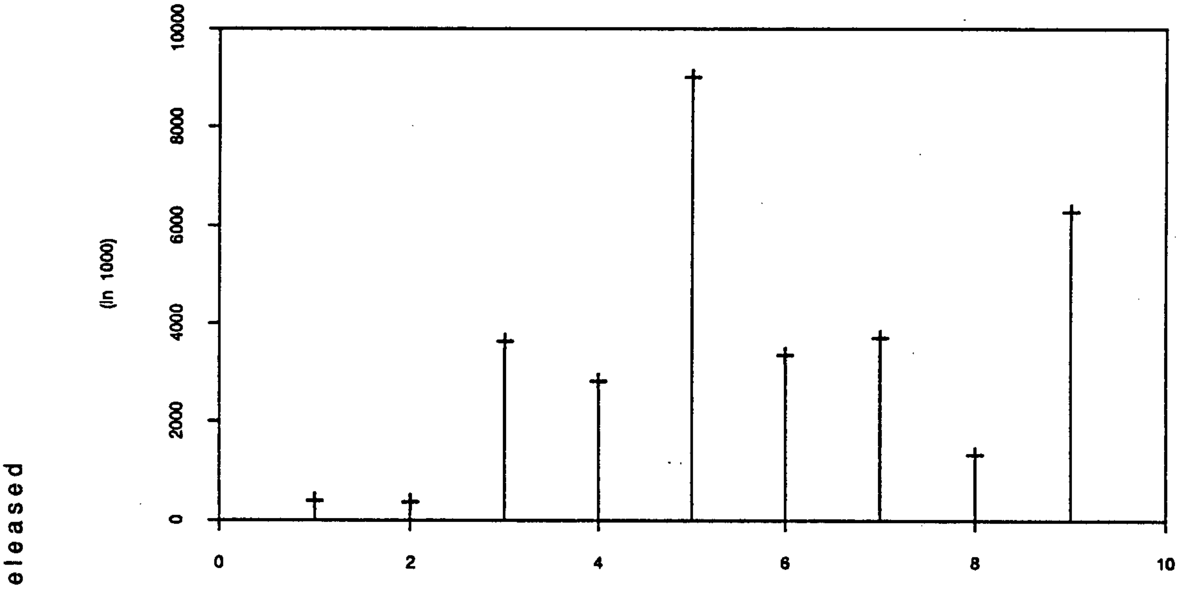


Figure 3.1b. Size of coho release for tag codes from the benchmark release data subset.

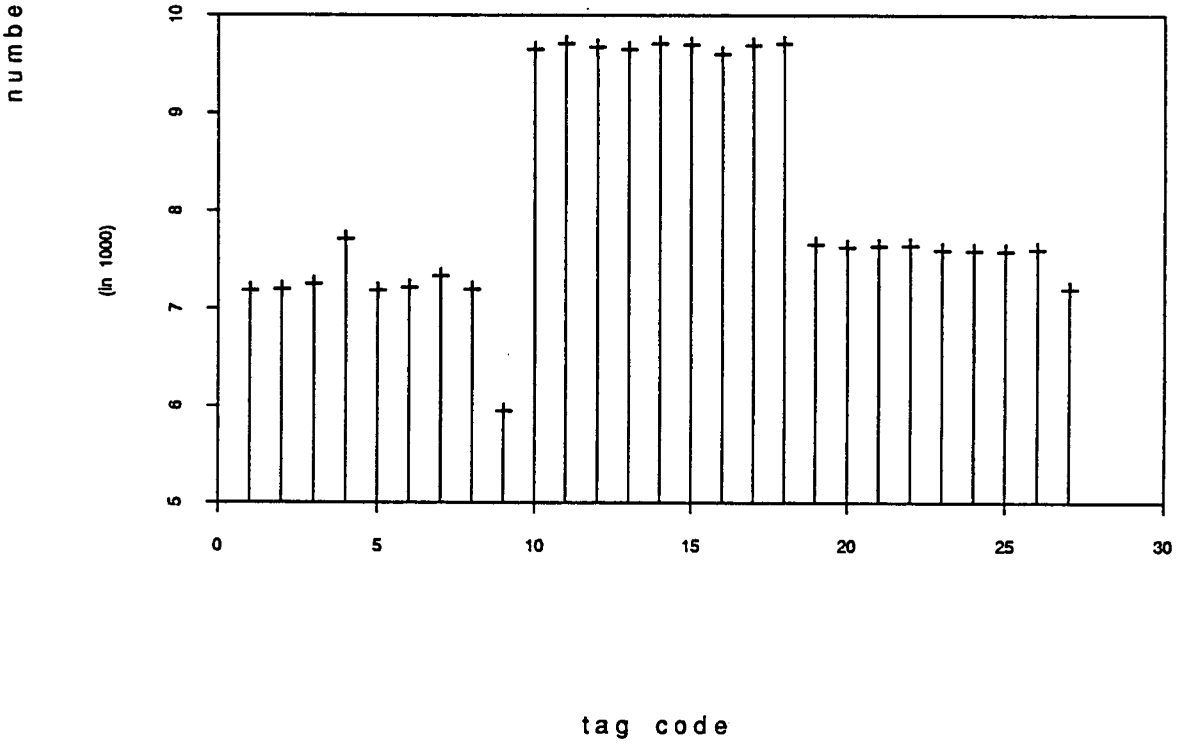


Figure 3.2. Chinook observed recoveries over the recovery period considered.
(tag code: 021827 brood year: 1979 recovery year: 1981 to 1984)

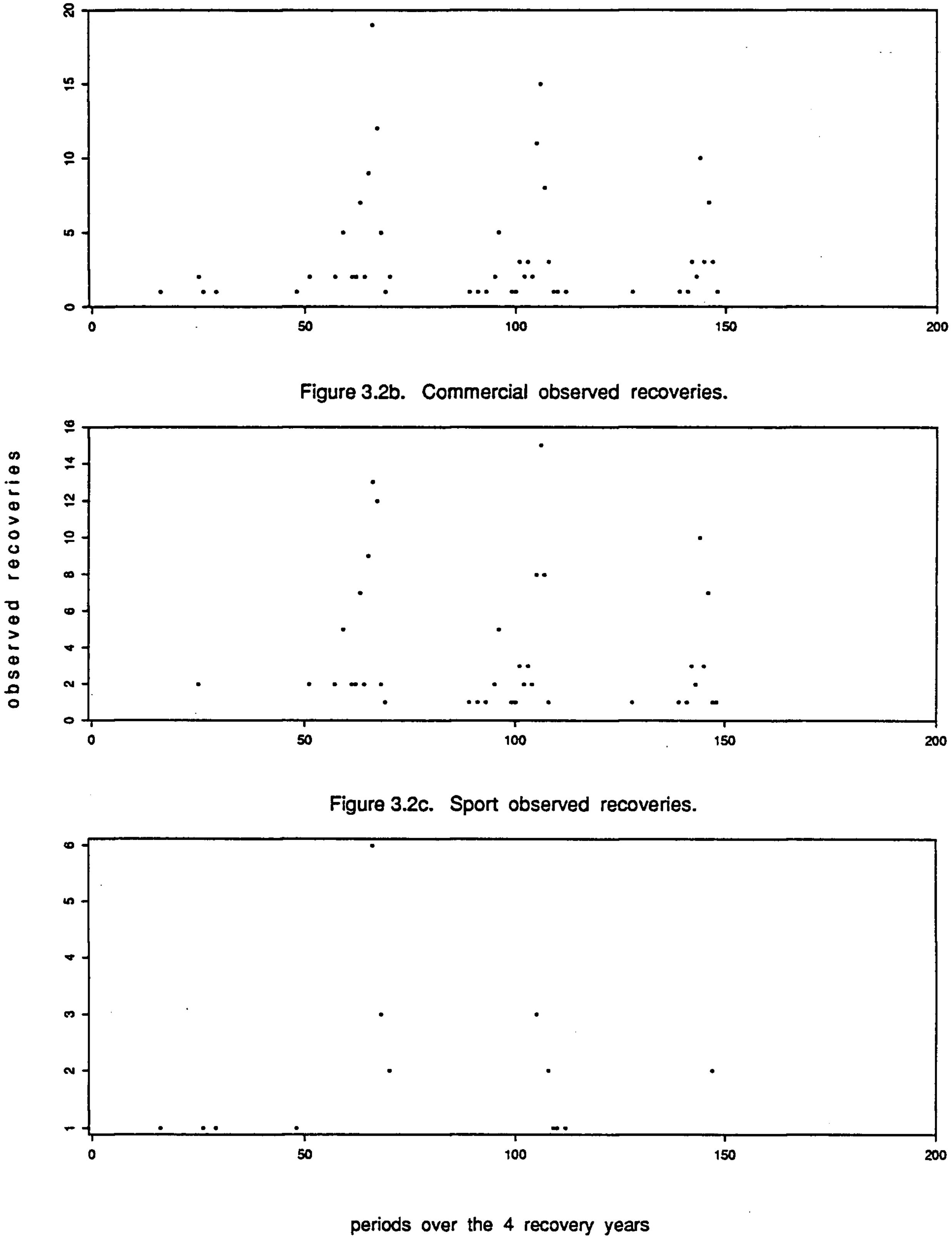


Figure 3.2a. Commercial and sport observed recoveries.

Figure 3.2b. Commercial observed recoveries.

Figure 3.2c. Sport observed recoveries.

Figure 3.3. Coho observed recoveries over the recovery period considered.
(tag code: 081842 brood year: 1979 recovery year: 1981 to 1982)

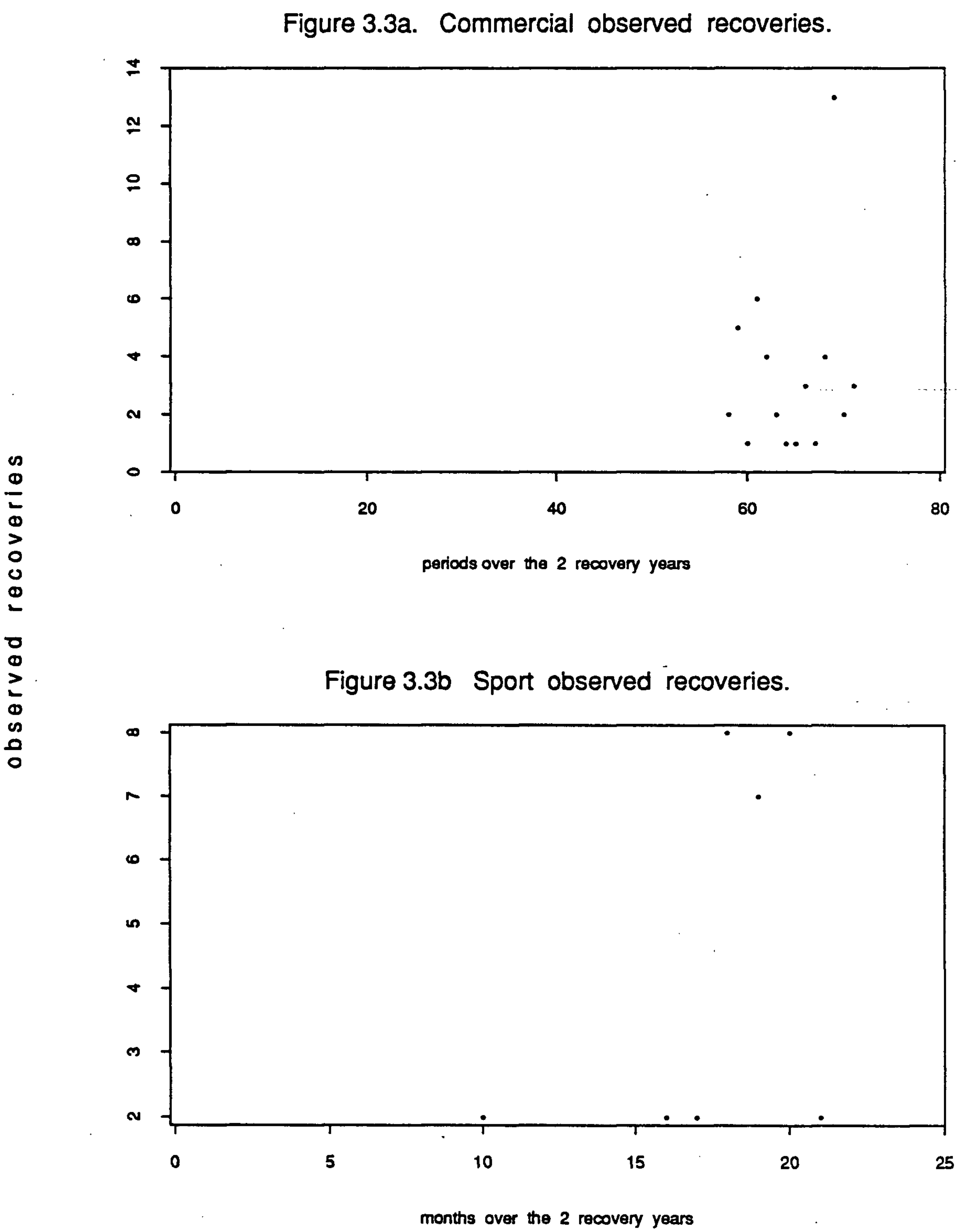


Figure 3.3a. Commercial observed recoveries.

Figure 3.3b Sport observed recoveries.

Figure 3.4a. Plot of cumulative sum of chinook commercial observed recoveries over time.

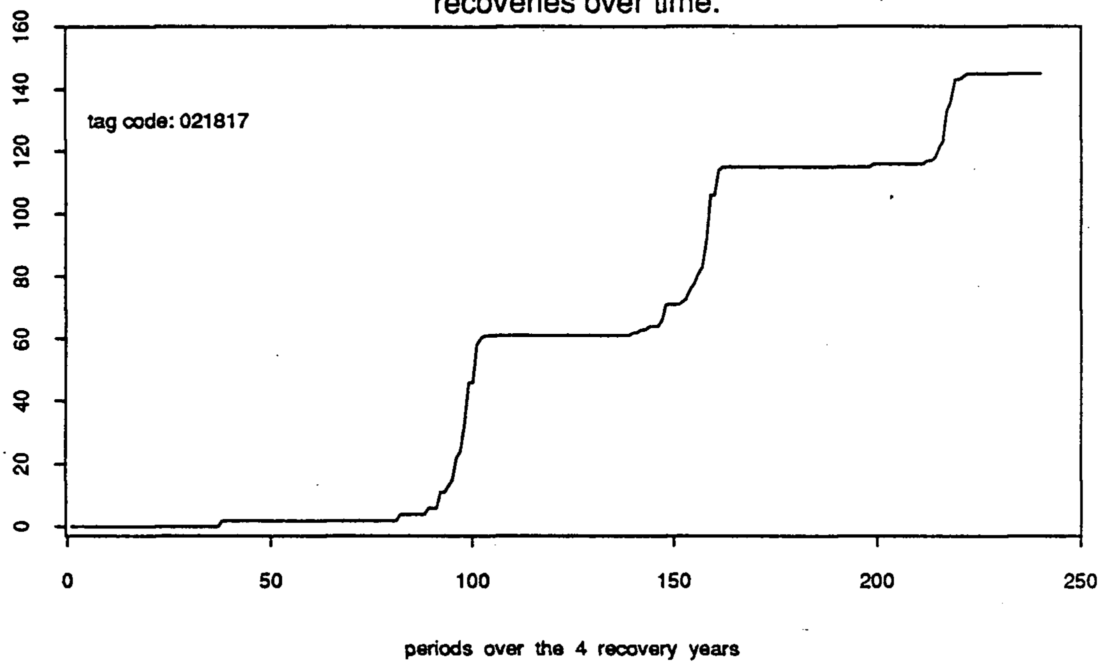


Figure 3.4b. Plot of cumulative sum of coho commercial observed recoveries over time.

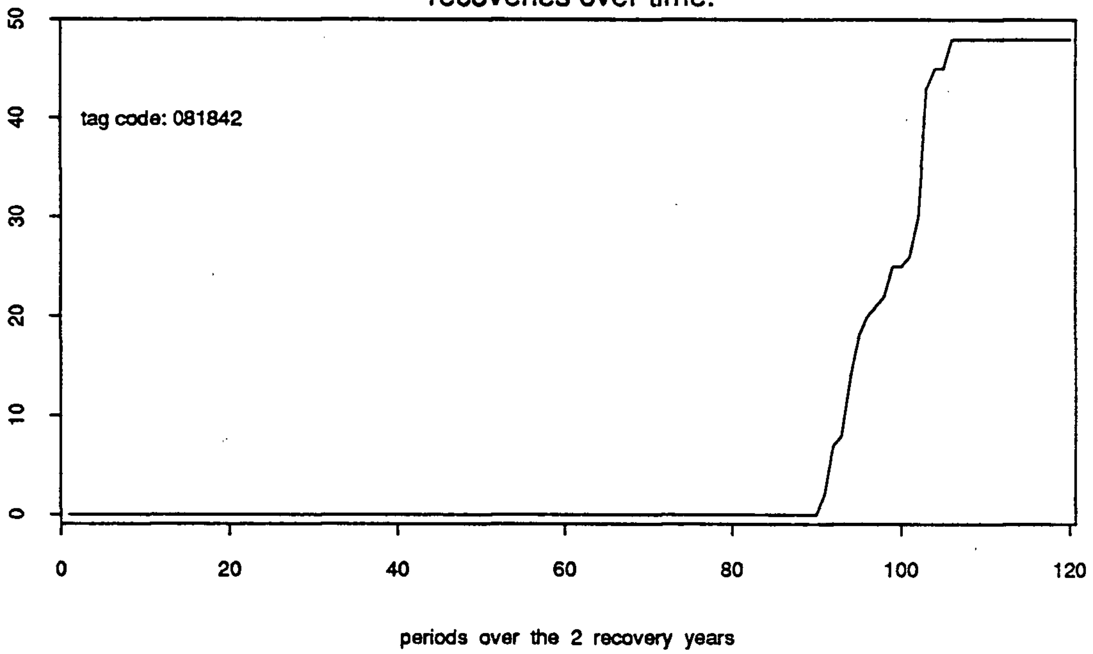
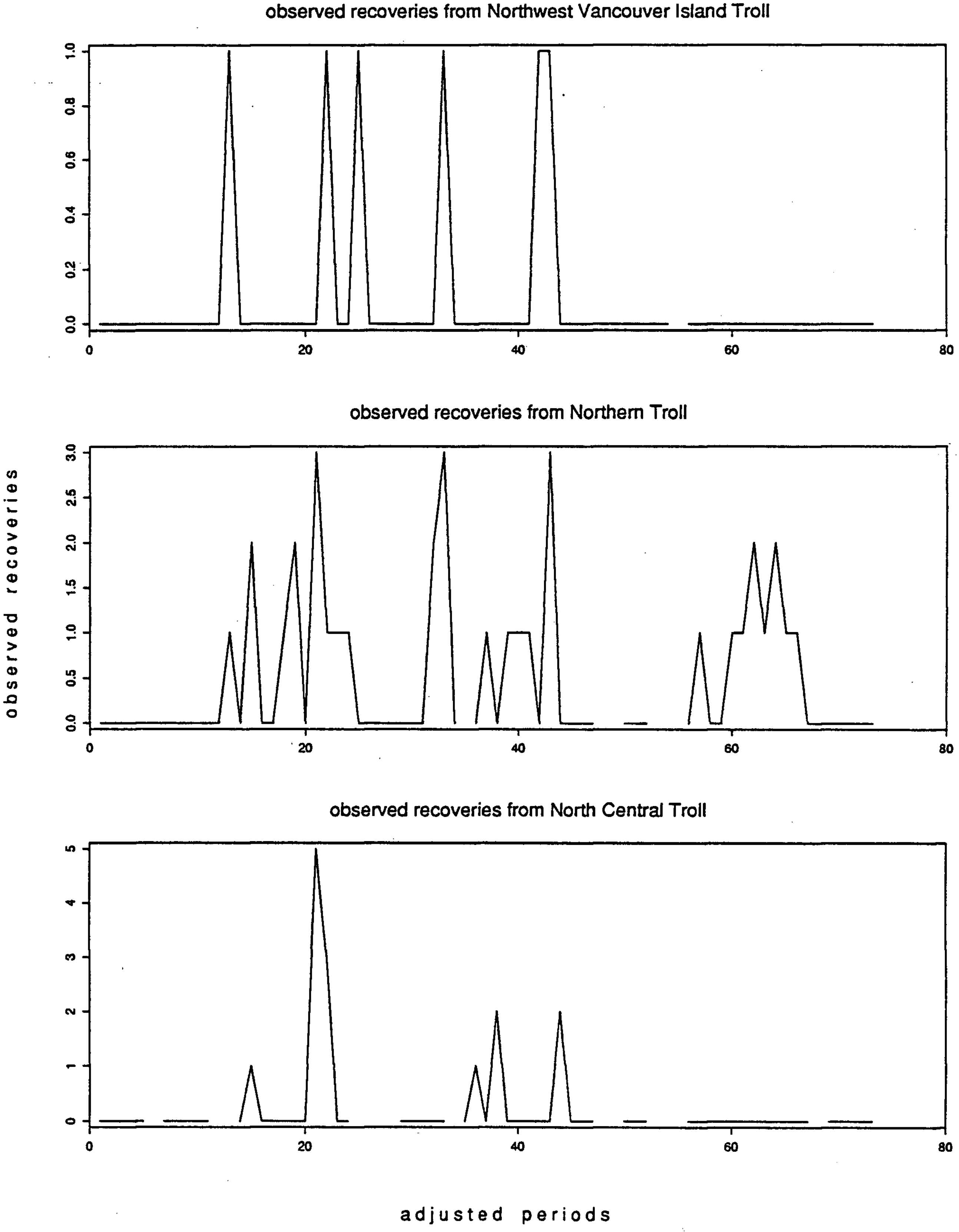


Figure 3.5. Plots of chinook commercial observed recoveries over the sampling period. (tag code: 021827 brood year: 1979)

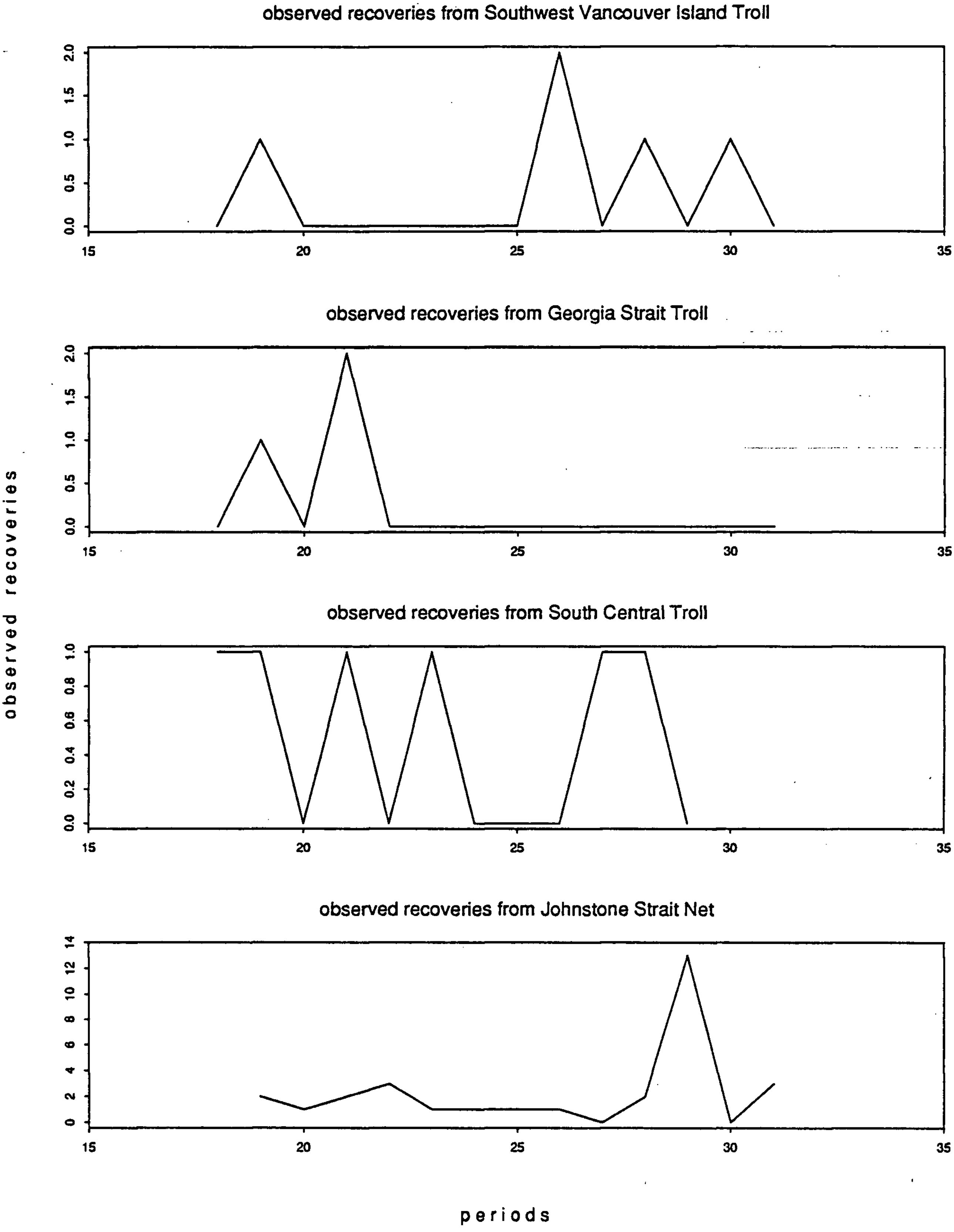


observed recoveries from Northwest Vancouver Island Troll

observed recoveries from Northern Troll

observed recoveries from North Central Troll

Figure 3.6. Plots of coho commercial observed recoveries over the sampling period.
(tag code: 081842 brood year: 1979)



observed recoveries from Southwest Vancouver Island Troll

observed recoveries from Georgia Strait Troll

observed recoveries from South Central Troll

observed recoveries from Johnstone Strait Net

Figure 5.1a. Zeta(t) for coho.
(Hatchery: Quinsam brood year: 1979 size at release: medium)

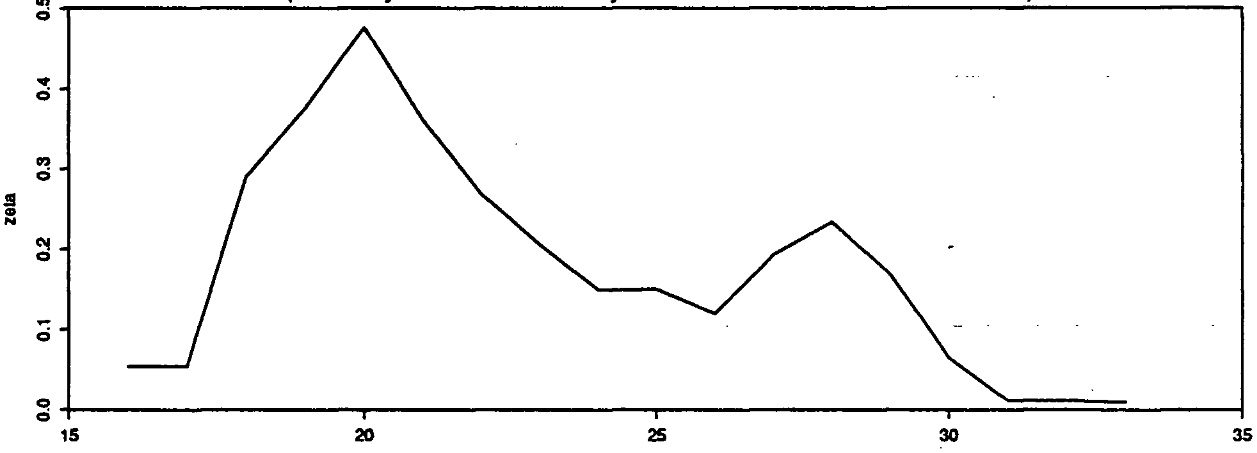


Figure 5.1b. Transformed zeta(t) (power = 0.25).

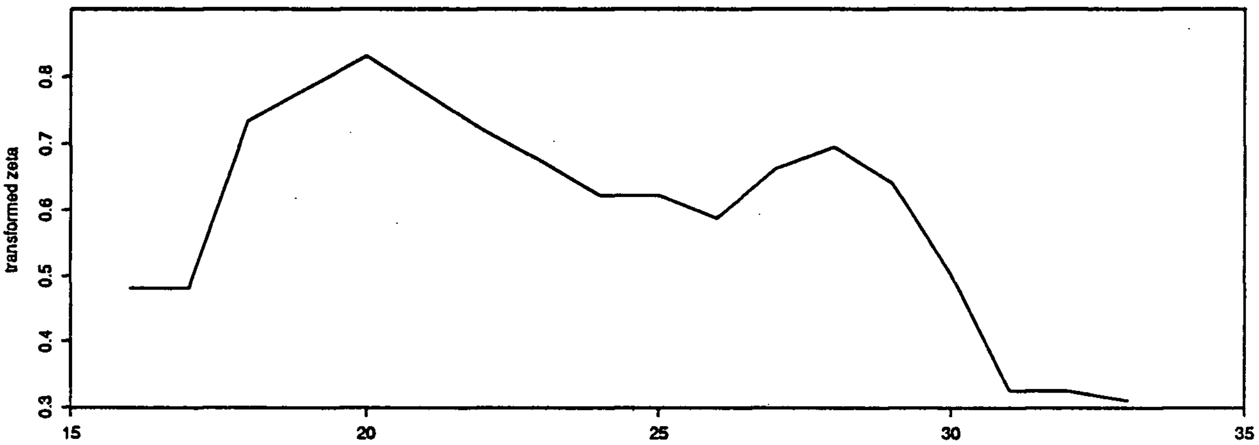


Figure 5.1c. Trend for the zeta(t).

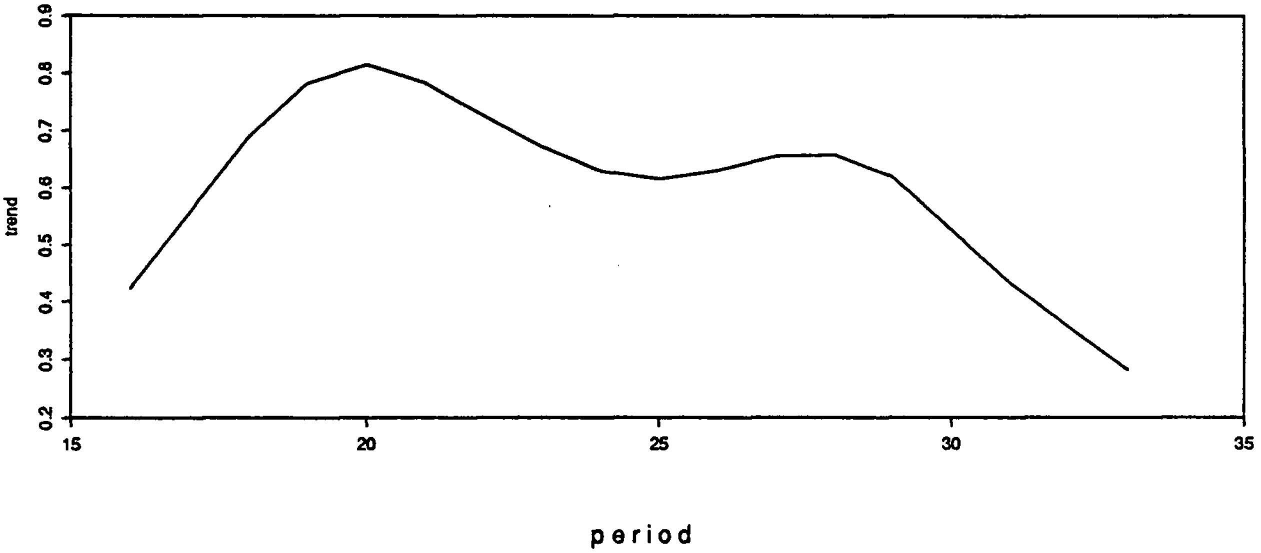


Figure 5.2a. Zeta(t) for coho.
(Hatchery: Capilano brood year: 1980 size at release: medium)

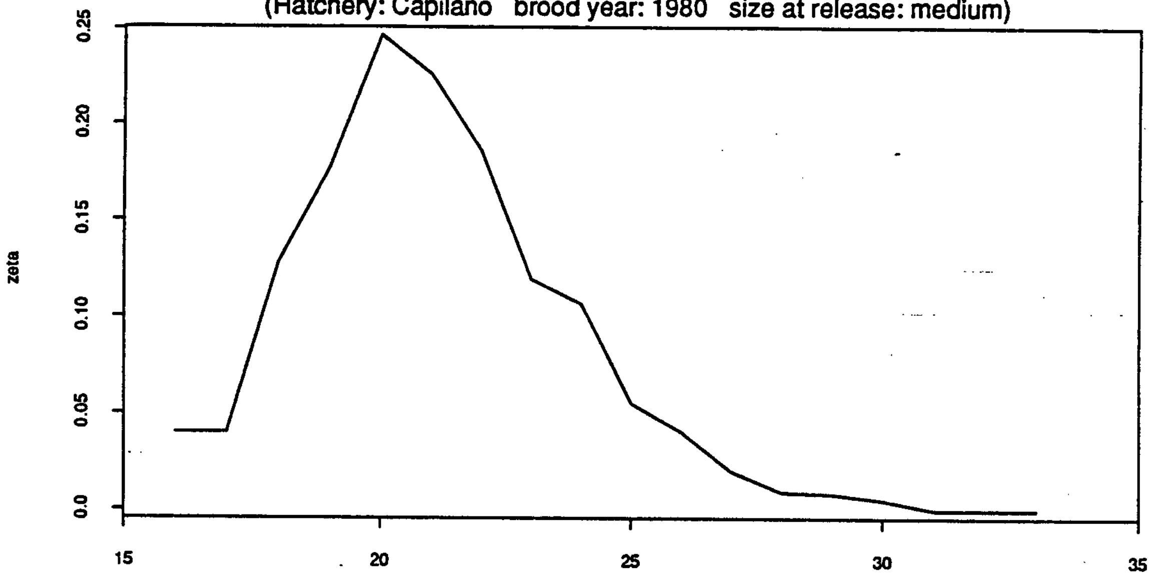


Figure 5.2b. Trend for the zeta(t).

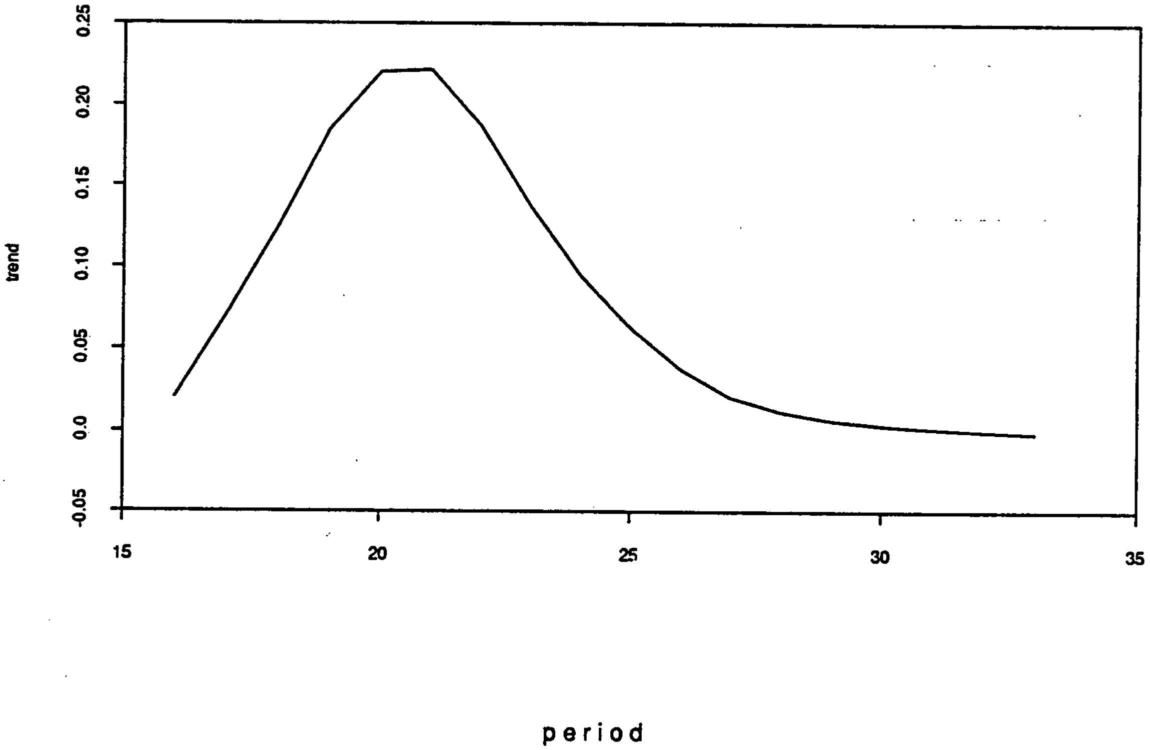


Figure 5.3a. Zeta(t) for chinook tag code: 021827.

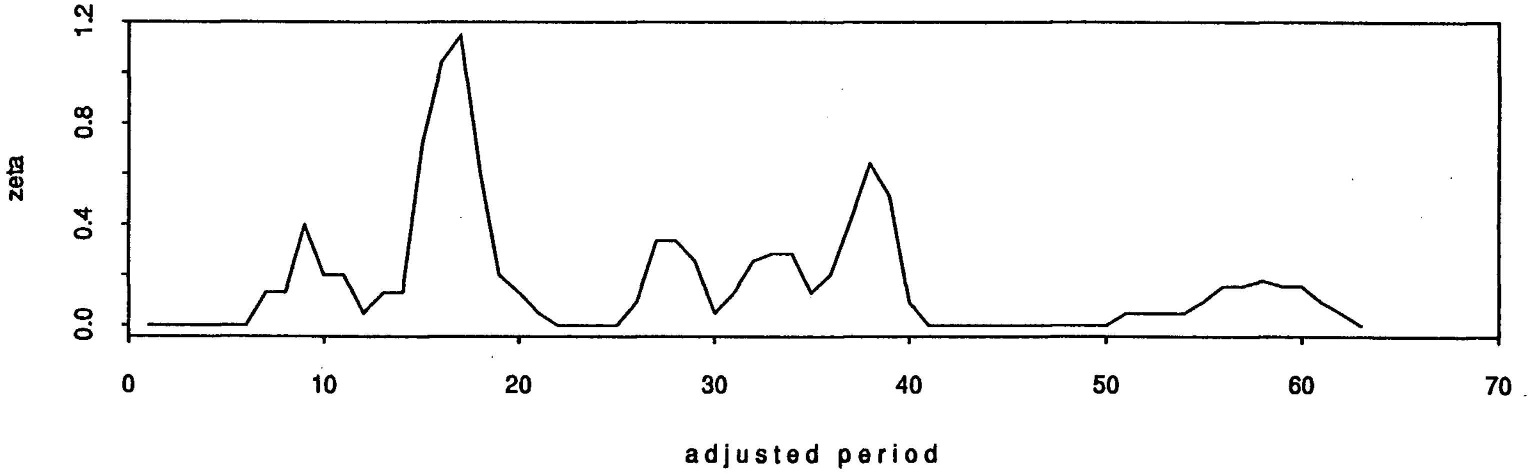


Figure 5.3b. Trend for the zeta(t).

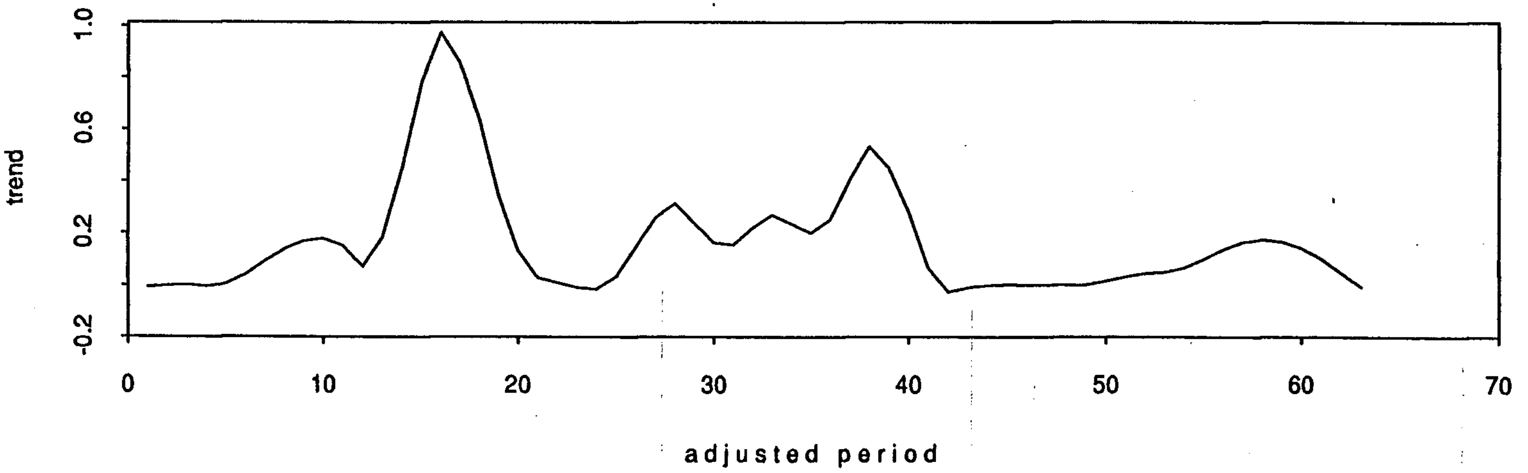
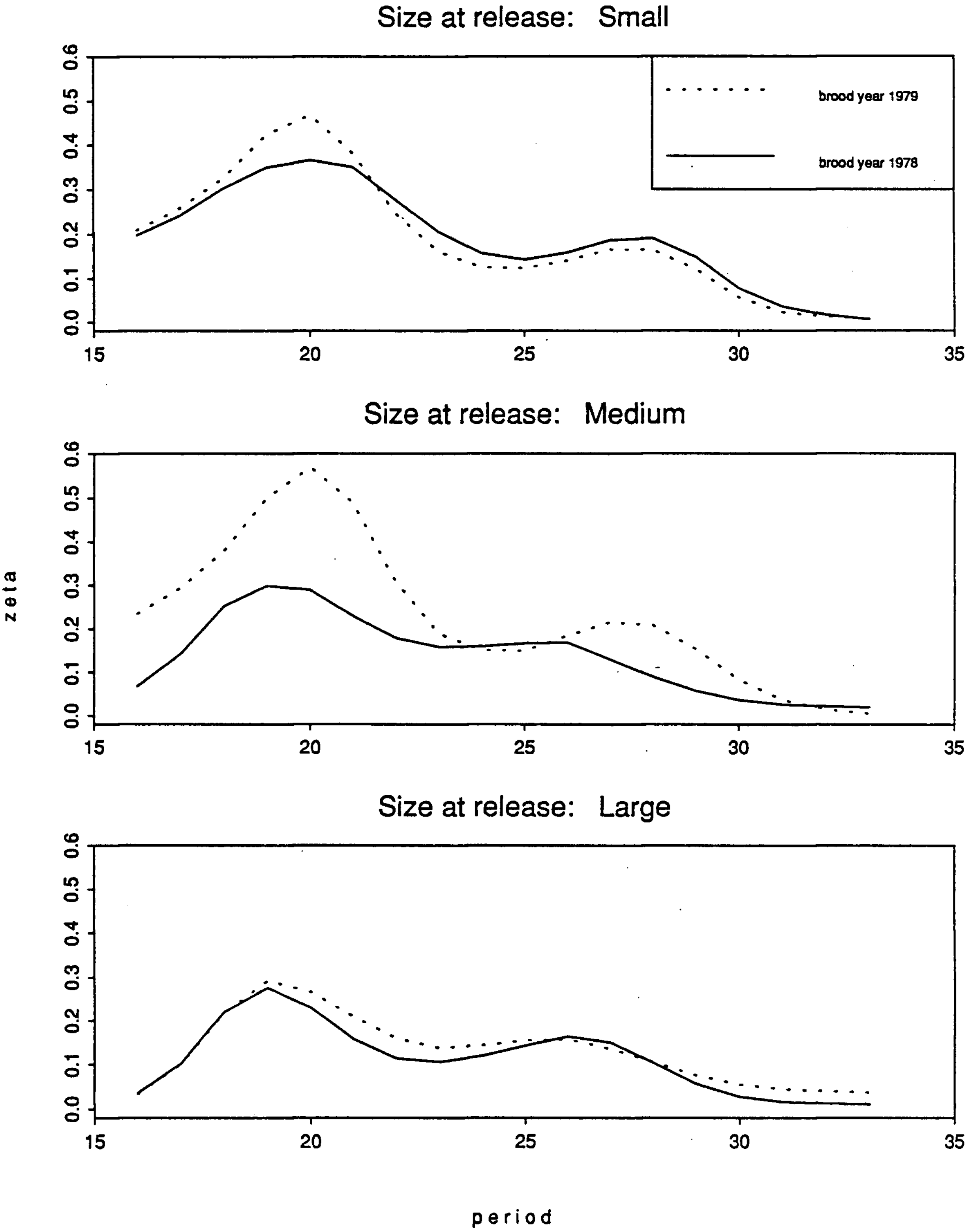


Figure 5.4. The trends of zeta's for Quinsam coho from different brood years.



Size at release: Small

Size at release: Medium

Size at release: Large

Figure 5.5. The trends of zeta's for Capilano coho from different brood years.

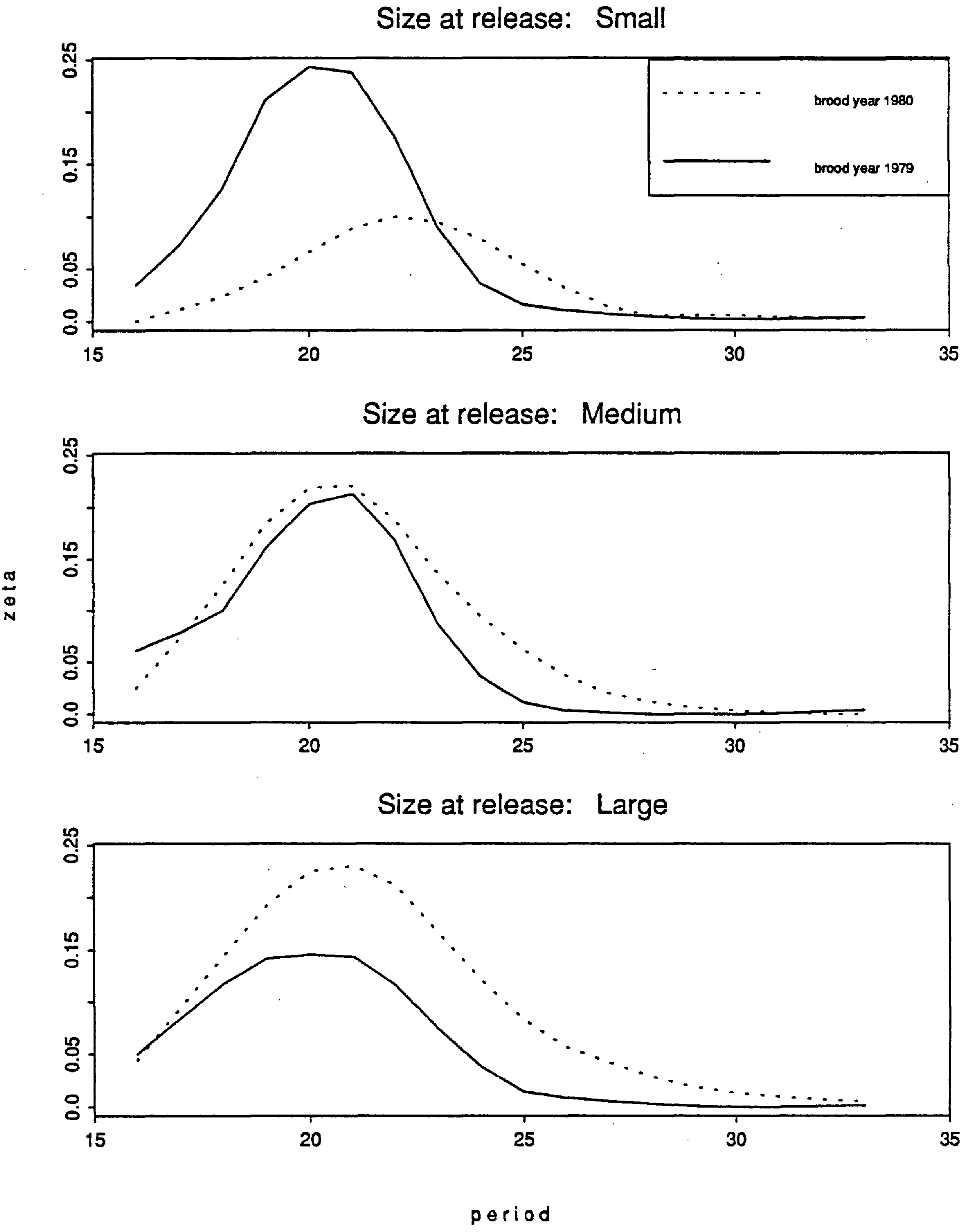


Figure 5.6. The trends of zeta's for the three chinook tag codes.

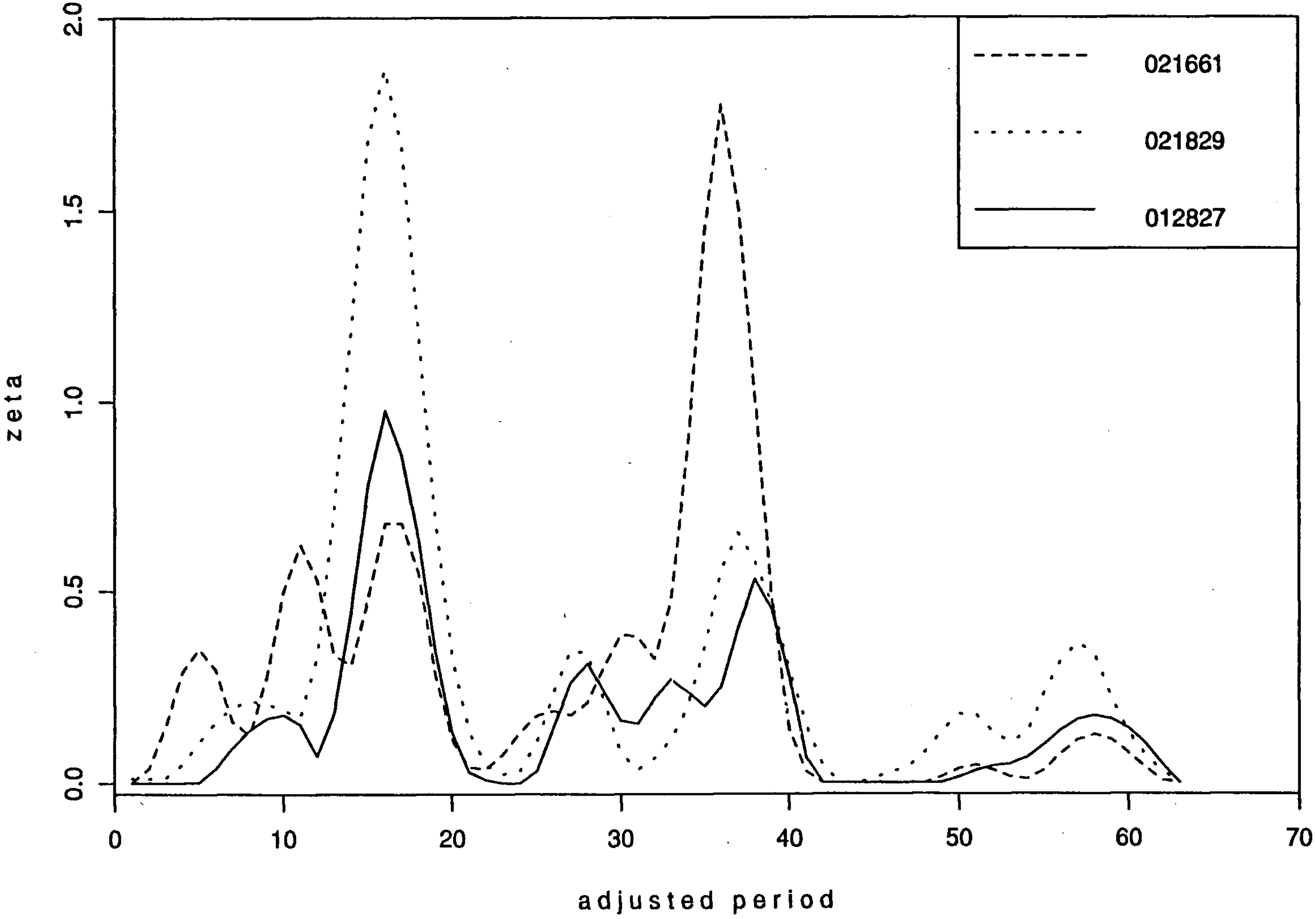


Figure 5.7. The estimated recovery intensity of coho for each of the 4 catch regions. (Hatchery: Quinsam size at release: large)

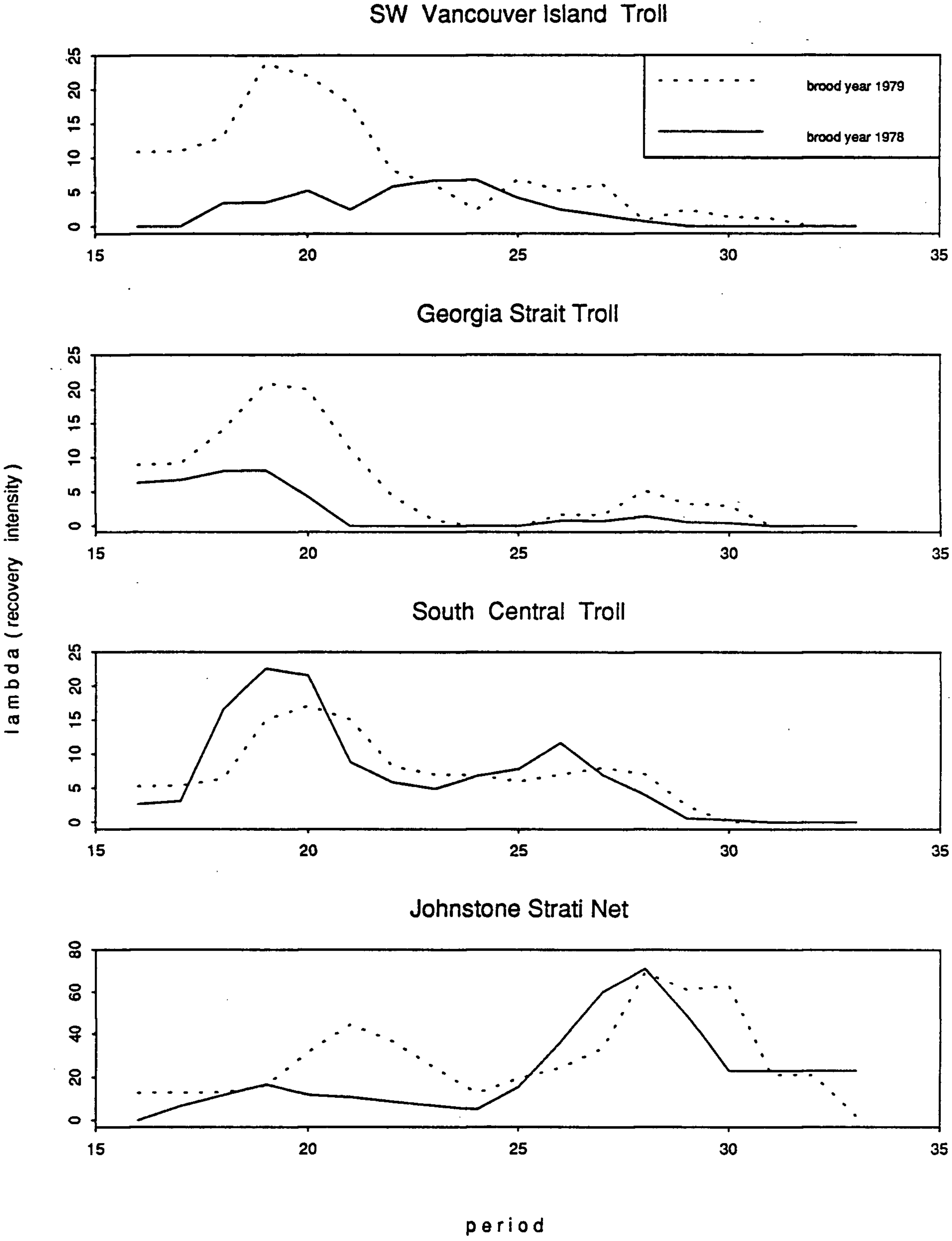
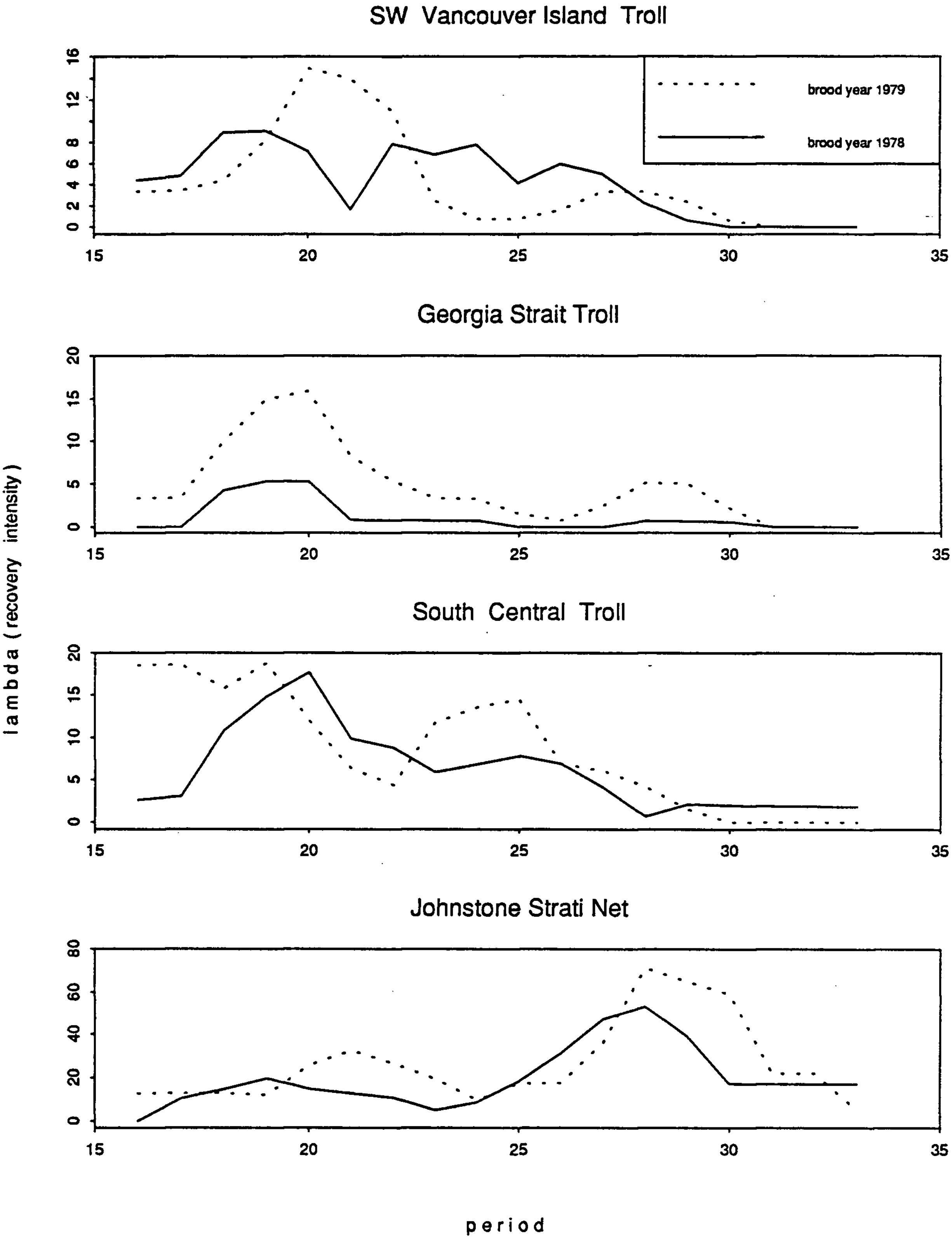


Figure 5.8. The estimated recovery intensity of coho for each of the 4 catch regions. (Hatchery: Quinsam size at release: medium)



SW Vancouver Island Troll

Georgia Strait Troll

South Central Troll

Johnstone Strati Net

Figure 5.9. The estimated recovery intensity of coho for each of the 4 catch regions.
(Hatchery: Quinsam size at release: small)

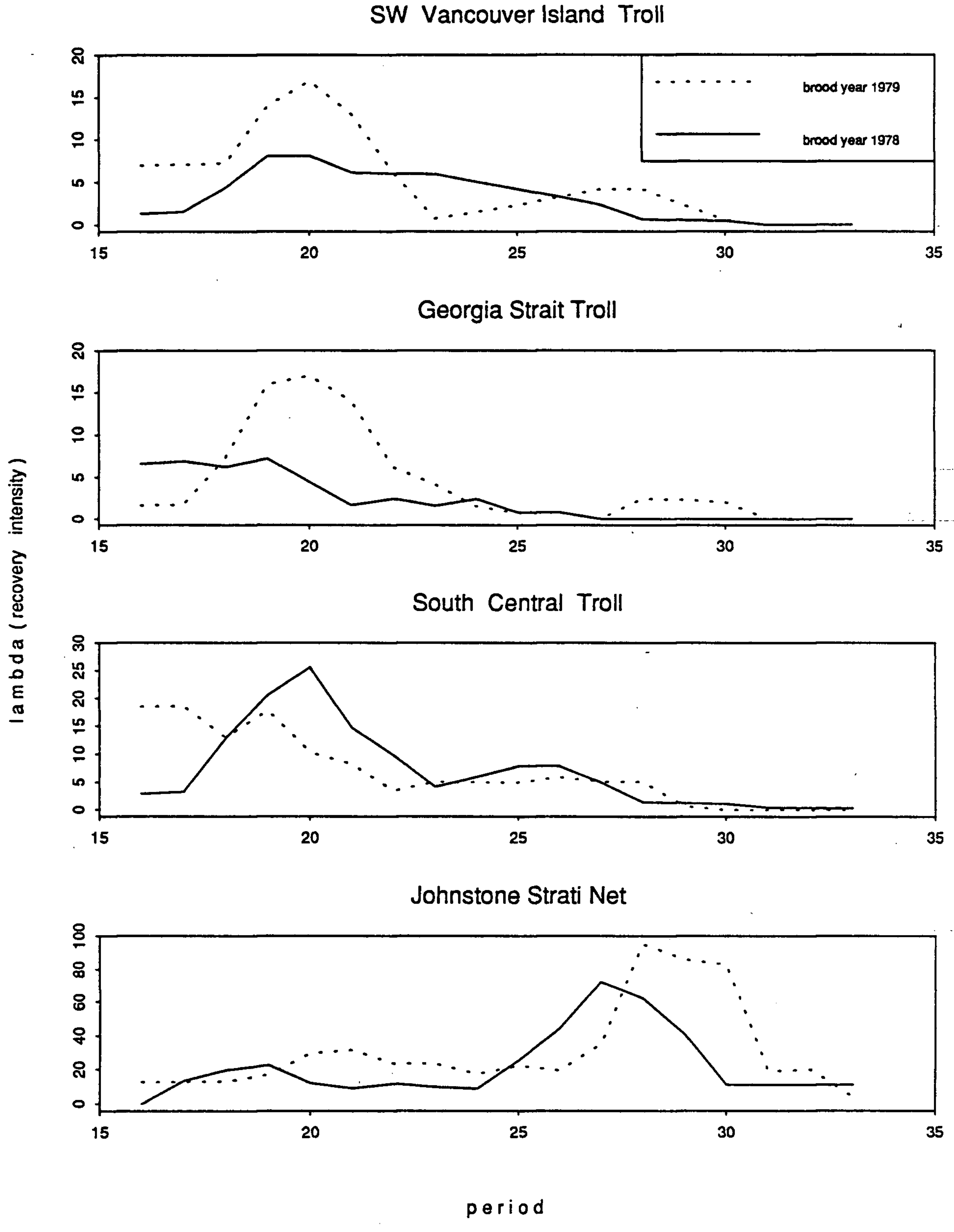


Figure 5.10. The estimated recovery intensity of coho for each of the 4 catch regions.
(Hatchery: Capilano size at release: large)

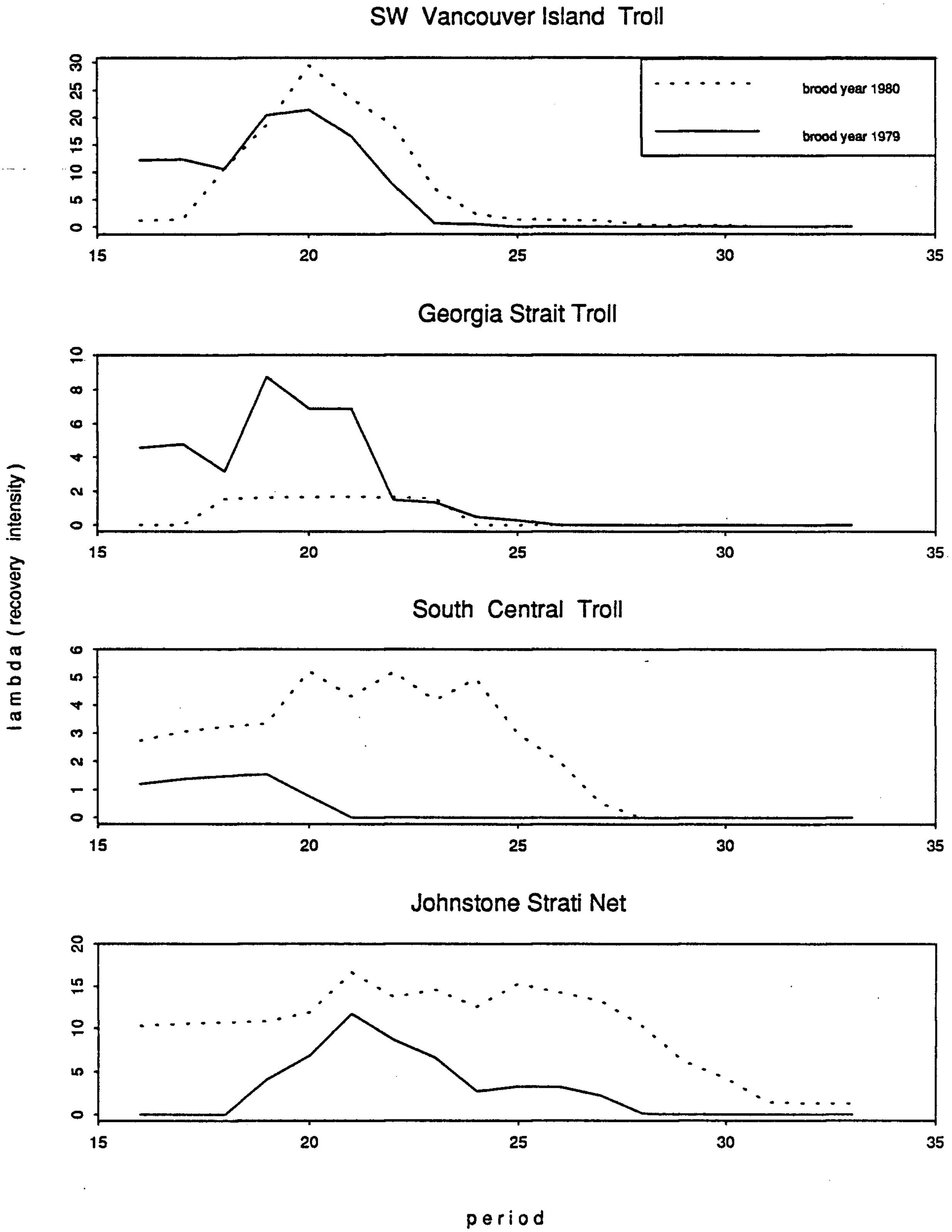
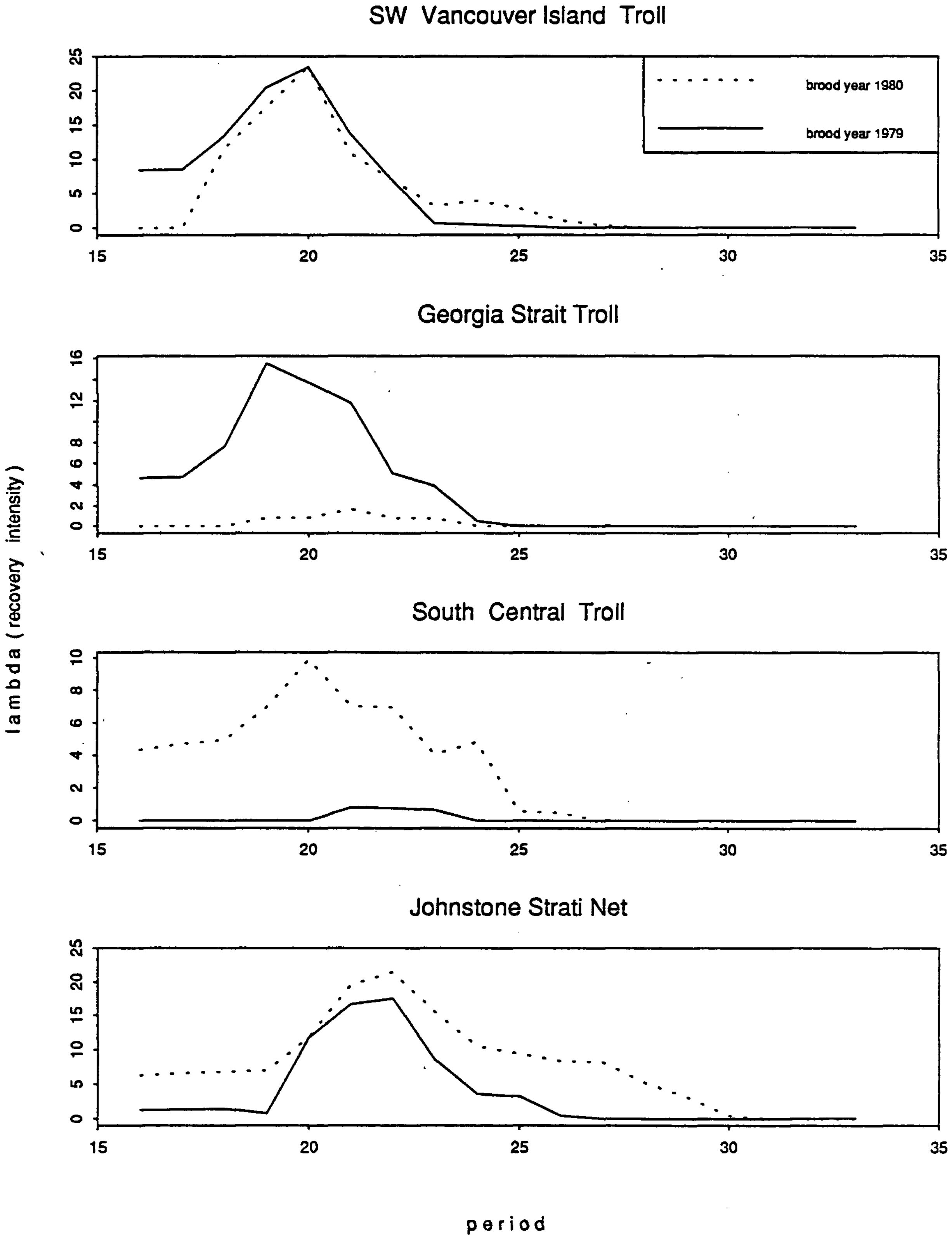


Figure 5.11. The estimated recovery intensity of coho for each of the 4 catch regions.
(Hatchery: Capilano size at release: medium)



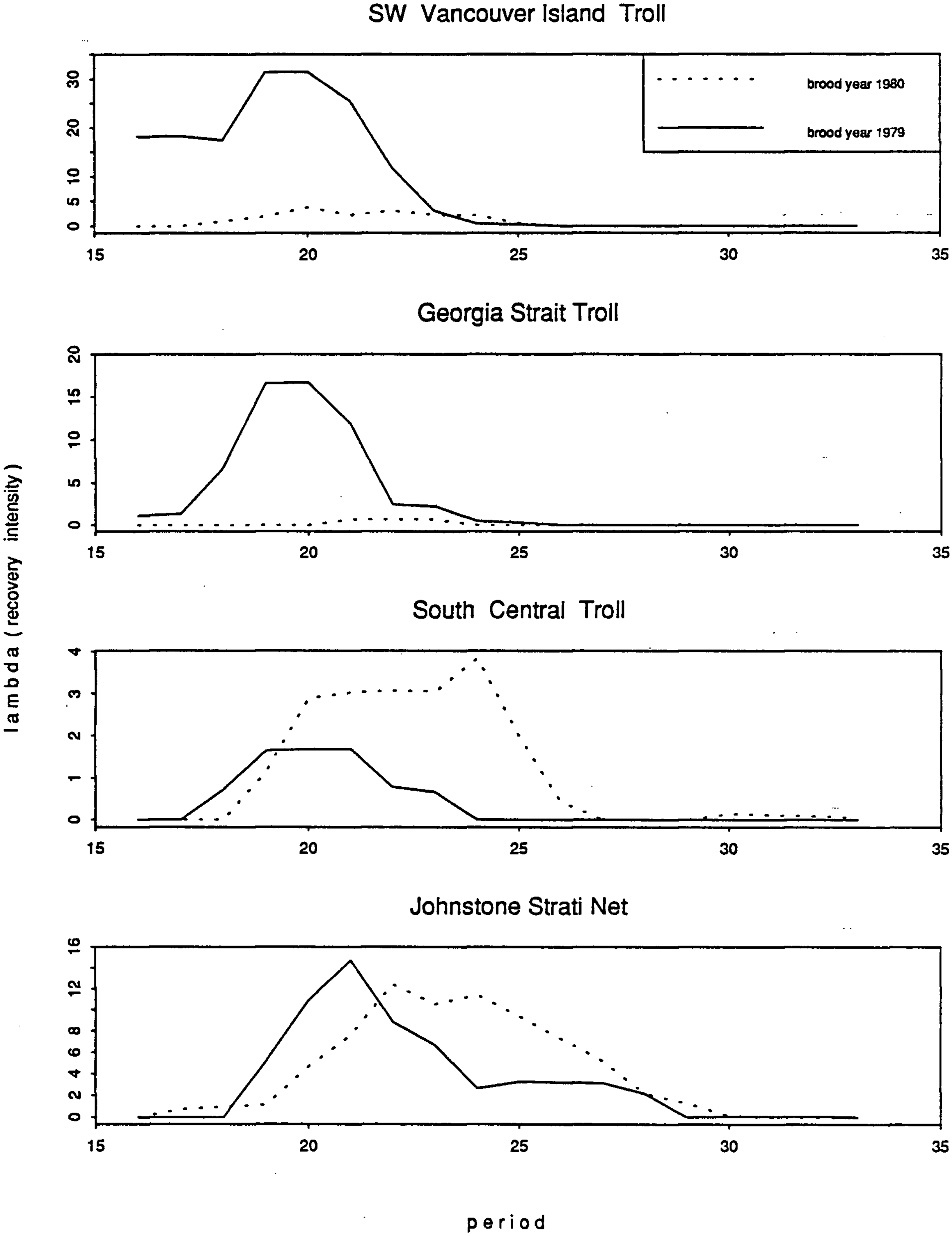
SW Vancouver Island Troll

Georgia Strait Troll

South Central Troll

Johnstone Strati Net

Figure 5.12. The estimated recovery intensity of coho for each of the 4 catch regions. (Hatchery: Capilano size at release: small)



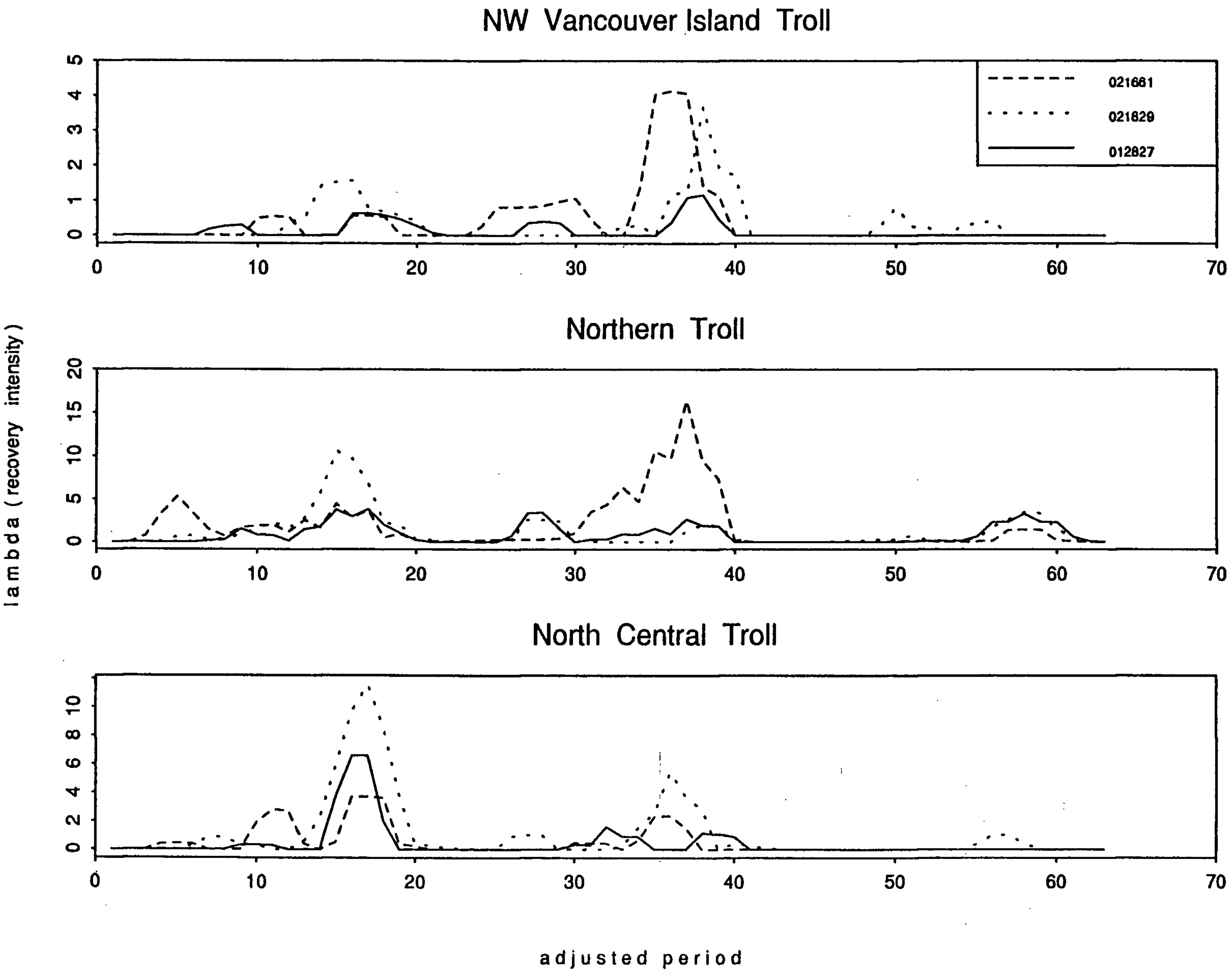
SW Vancouver Island Troll

Georgia Strait Troll

South Central Troll

Johnstone Strati Net

Figure 5.13. The estimated recovery intensity of chinook for each of the 3 trolling regions.



NW Vancouver Island Troll

Northern Troll

North Central Troll

Figure 5.14. Estimated recovery intensities of coho and the corresponding 95% credibility intervals. (Hatchery: Quinsam brood year: 1979 size at release: medium)

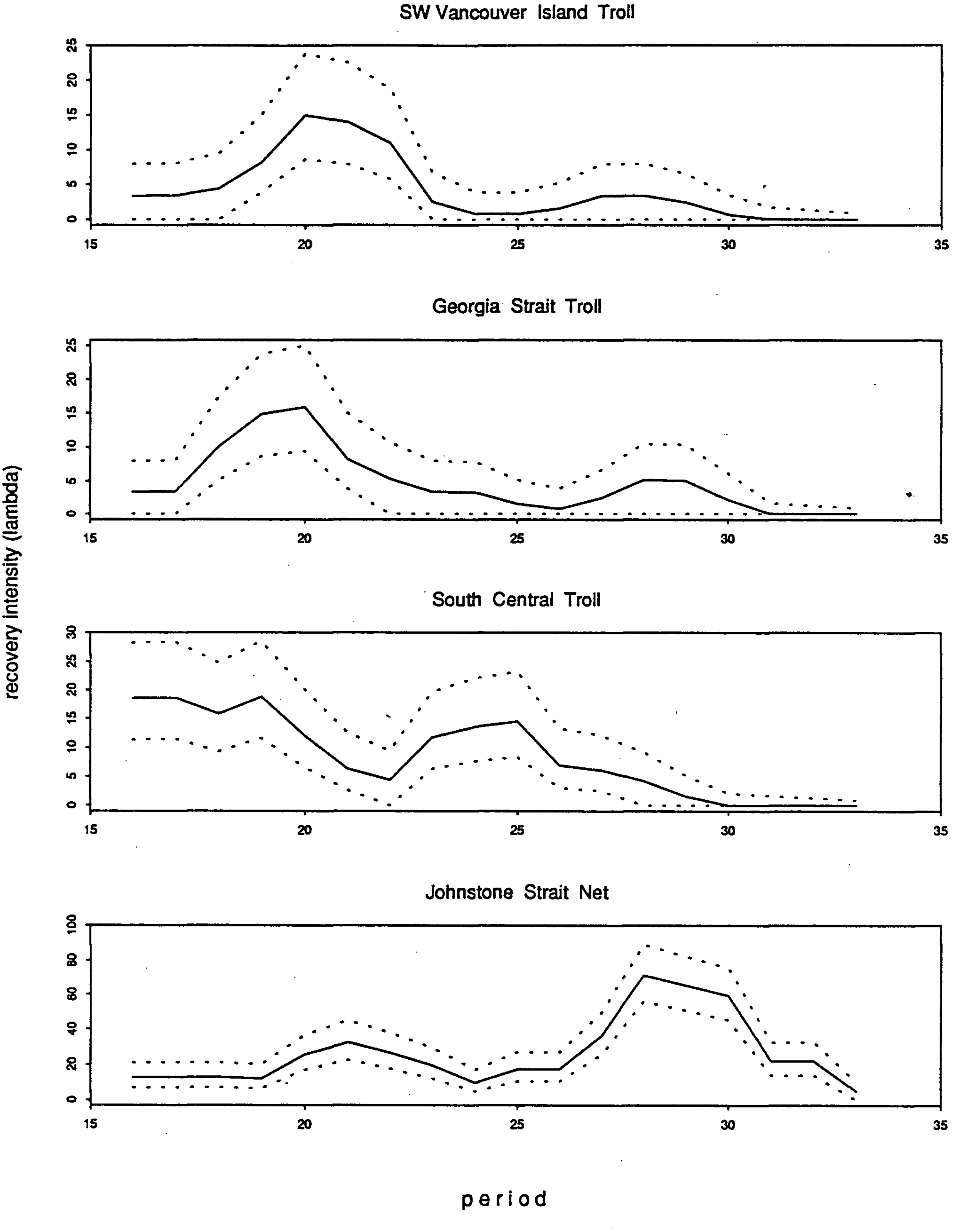
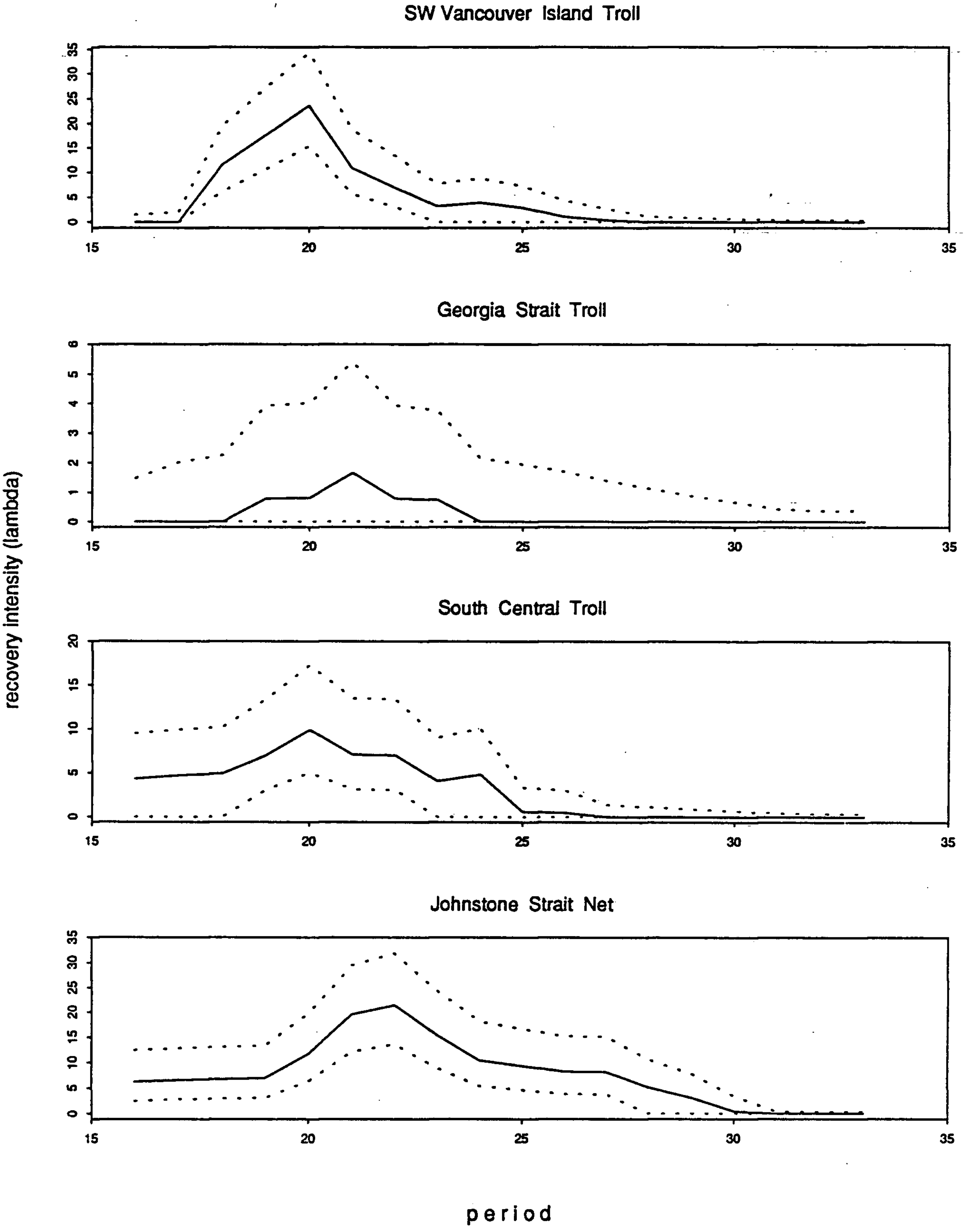


Figure 5.15. Estimated recovery intensities of coho and the corresponding 95% credibility intervals.
(Hatchery: Capilano brood year: 1980 size at release: medium)



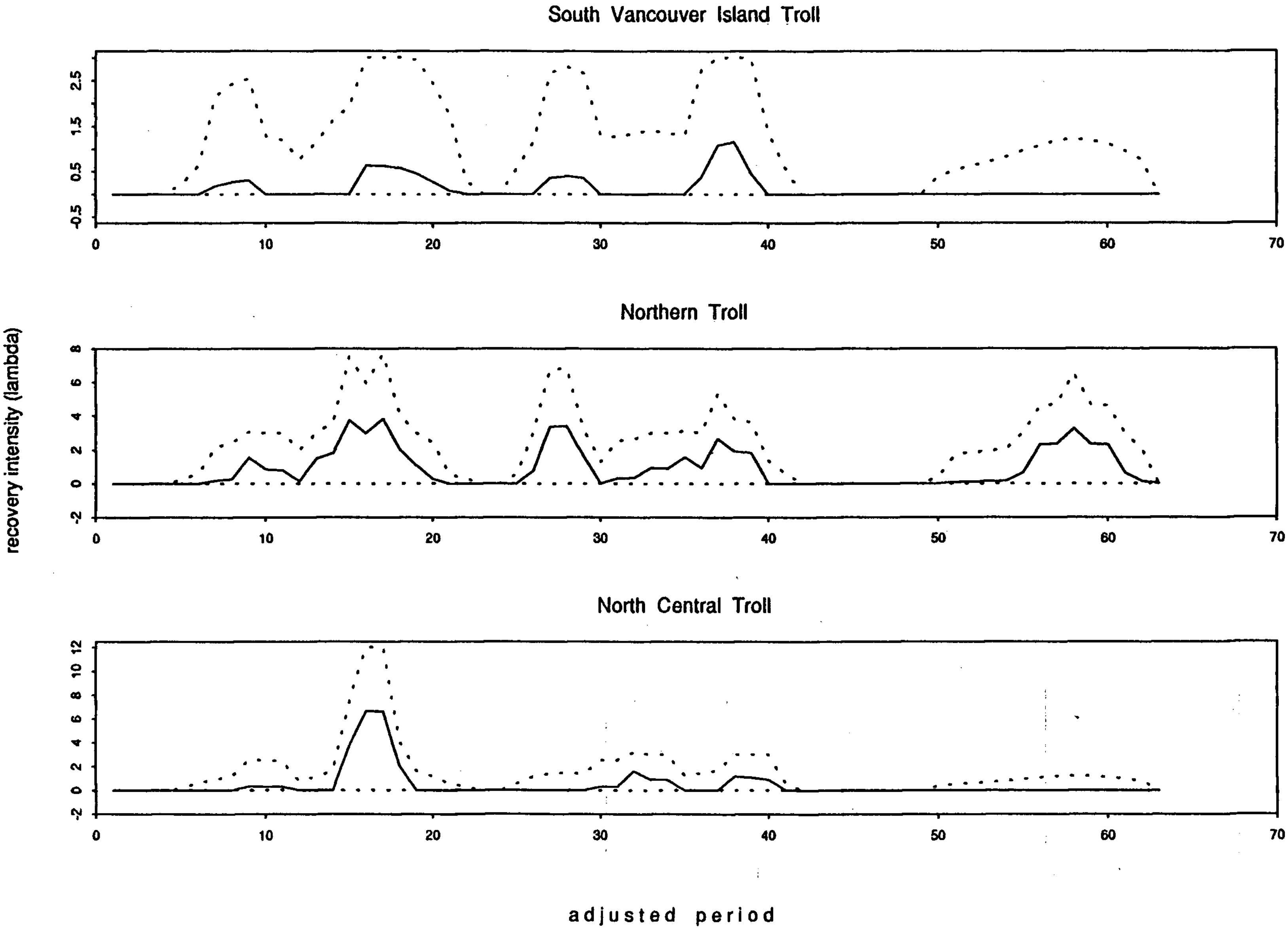
SW Vancouver Island Troll

Georgia Strait Troll

South Central Troll

Johnstone Strait Net

Figure 5.16. Estimated recovery intensities of chinook and the corresponding 95% credibility intervals.
(tag code : 021827 brood year : 1979)


South Vancouver Island Troll
Northern Troll
North Central Troll
recovery intensity (lambda)
adjusted period