ASSESSING INFILLING METHODS FOR MISSING DATA IN SALMON SPAWNING ESTIMATES

by

Ruth Joy

B. Sc., University of Victoria, 1996

a Master's Project submitted in partial fulfillment
of the requirements for the degree of

Masters of Science

in the

Department of Mathematics and Statistics
Faculty of Science

© Ruth Joy 2002
SIMON FRASER UNIVERSITY
June 2002

All rights reserved.

However, in accordance with the *Copyright Act of Canada*, this work may be reproduced without authorization under the conditions for "Fair Dealing." Therefore, limited reproduction of this work for the purposes of private study, research, criticism, review and news reporting is likely to be in accordance with the law, particularly if cited appropriately.

APPROVAL

Name: Ruth Joy

Degree: Masters of Science

Master's Project: Assessing infilling methods for missing data in salmon spawning estimates

Examining Committee: Dr. R. Lockhart
Chair

Dr. Richard Routledge, Senior Supervisor

Dr. Carl Schwarz, SFU Examiner

Dr. James Richard Irvine, External Examiner
Research Scientist
Department of Fisheries and Oceans.

Date Approved: ---

June, 2002

Acknowledgments

I would like to thank Dr. Rick Routledge for his patience and guidance with each step of this project, and for his comedy and wit about all matters of importance. Thank you to my friends and office mates in K-9501, with a special thanks to Laurie Ainsworth for showing me how to be a fun statistician! Lastly, I would like to thank my husband, Jeff Joy, and my son, Timothy whose humour and encouragement give me strength in everthing I do.

Of all the science that seeks to learn, we know of only one planet that is so complex and intricate that it contains life. Paradise only happens once.

Assessing infilling methods for missing data in spawning salmon estimates

Monitoring of populations is a key component of an effective conservation program. Trends in abundance must be monitored to ensure that timely action is taken before conservation risks become too severe. Unfortunately the monitoring is expensive and in many instances, only a portion of widespread species can feasibly be estimated in a given year. In the case of British Columbia's Thompson River coho salmon, 41% of the data are missing. Accurate abundance estimates for this aggregate are particularly important as the abundance of these salmon had declined so severely by the late 1990's that a major, continuing conservation effort was initiated.

This project presents an examination of seven imputation methods for infilling missing data in such records. We assessed the performance of these methods through a simulation study that modeled widely accepted features of the population dynamics of Thompson River coho, specifically including the recent decline. The study also incorporated the historical record of missing estimates. Performance was measured through jackknifed sums-of-squares estimates to evaluate bias, chance error and total error of the infilled values. We found that the infilling methods that use a multiplicative analysis-of-variance-style model outperfomed the others, with the preferred version within this class of methods application-dependent.

We also investigated a sockeye salmon population where the missing data pattern was extreme (72%). In this extreme case, with little time overlap between data records for different subsets of spawning areas, no method for imputing missing values will

work well. For methods based on modified analyses of variance, this difficulty can be related to the concept of balance in an experimental design. The project concludes with an exploration of the advantages of sampling schemes that promote this sort of balance.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

A major component of a conservation program is monitoring of populations. Trends in abundance need to be monitored if timely action is to be taken. Unfortunately, the monitoring is almost always expensive and complicated. This is particularly so for populations such as Pacific salmon (*Oncorhynchus* spp.) that occupy numerous, more or less discrete habitat units. Attempts to census entire species are doomed to failure. At best, accurate estimates can be obtained for relatively few local populations in any given year. In the case of British Columbia's coho salmon (*O. kisutch*), the existing record of spawner-abundance data is irregular.

As a result, it is typically difficult to obtain clear, unambiguous evidence of abundance trends from the irregular records. Here we examine two such datasets: Thompson River coho and North Coast sockeye (*O. nerka*). Both coho and sockeye salmon are protected in areas of Washington, Oregon and California by the US Endangered Species Act. In British Columbia, Thompson River coho salmon have become the recent focus of a major conservation effort since these salmon are recognised as both genetically unique and severely depressed by a decline in marine survival in the last decade.

Clear evidence of abundance trends is key to the implementation of management actions. Without such evidence, declines may go undetected before a crisis ensues.

This project addresses potential improvements to the estimation procedures for assessing trends in spawner abundance. The questions addressed specifically are:

1. How might overall abundance estimates be constructed from partial records of abundance with estimates for individual spawning populations for some creeks missing in some years? This question will be addressed by considering several methods for imputing the missing elements in the data record (Chapter 2).

2. Which of these imputation methods would make this task more reliable, and for what sorts of patterns of missing data can a reliable estimate not be constructed (Chapter 3)?

3. What sorts of sampling schemes for deliberately generating partial records would make this task easier (Chapter 4)?

Chapter 2

Missing Data Methods

In this chapter, seven methods for imputing values for missing data in spawning salmon records are presented. This kind of data containing information about an entire river system, is typical for coastal North America. Numbers of spawning salmon are recorded in each of the creek tributaries for several years. Typically spawning salmon numbers are not recorded for every year, but instead creeks are irregularly sampled, especially for minor but potentially important subpopulations. Likewise, intensity of sampling varies between years, depending on fluctuating budgets and changing government priorities. The following imputation methods will be discussed with records for a single creek running across a row (rows $1, \ldots, c$), and records for a single year running down a column (columns $1, \ldots, t$), such that the data are in a $c \times t$ matrix.

2.1 Seven Imputation Methods

2.1.1 Zero-Infilled

The zero-infilled method replaces all missing values with zeros.

2.1.2 Nearest-Value

The nearest-value method imputes from the same creek from the nearest year for which there is an entry (i.e., the closest entry in the same row). In the case of a tie, choose the entry that is for the closest preceding year (i.e., the closest right-hand value).

2.1.3 Interpolation/Extrapolation

This method calculates the slope and intercept of a line between the closest left and closest right values for a creek and then infills based on this equation. If there are no left-hand values then the equation is extrapolated based on the closest two right-hand values. Likewise, if there are no right-hand values then the equation is based on the closest two left-hand values.

2.1.4 Averaging of scaled values

The averaging of scaled values is a method that has been used by the Department of Fisheries and Oceans to infill missing values. The method begins by scaling all the observed values by dividing each by the maximum observation for the same creek across all years. The missing values in this rescaled record are then filled in by the average of all the observed scaled values for that year. The missing values are then rescaled by multiplying by the creek maxima. For example: a single missing value in the j^{th} column (i^{th} row) is replaced with:

$$n_{ij} = n_{i.max} \times mean(\frac{n_{1j}}{n_{1.max}} + \frac{n_{2j}}{n_{2.max}} + ... + \frac{n_{(i-1)j}}{n_{(i-1).max}} + \frac{n_{(i+1)j}}{n_{(i+1).max}} + ... + \frac{n_{cj}}{n_{c.max}})$$

where $N = [n_{ij}]$ is the data matrix, and $n_{i.max} = \max_{j=1,...,t}(n_{ij})$.

2.1.5 Transformed Linear Model

This method fits an analysis of variance model to the observed data to estimate creek and year effects. One assumption to this analysis of variance model is that the responses are normally distributed and have constant variance independent of the mean. However, this data is count data and the variance is a function of the mean. One approach to this problem is to perform a variance stabilizing transformation before fitting the model. The natural log transform (n_{ij} transformed to $\ln\, n_{ij} + 0.5$) suits this purpose, where the addition of 0.5 is to avoid problems with observed zero counts. The second assumption is that the systematic effects combine additively with no interactions. Only main effects (1, ..., c creeks and 1, ..., t years) are considered, and interactions between creek and year are assumed negligible as they are totally confounded with the error. The log transformation is also a good choice as it helps linearize the fit. Missing values are infilled by applying the fitted linear model and back transforming.

The linear model may be represented by:

$$\ln(N_{ij} + .5) \quad = \quad \mu + c_i + t_j + \epsilon_{ij} \tag{2.1}$$

where μ is the intercept or grand mean; c_i is the i^{th} creek effect: $i = 1, ..., c$ and $\sum_{i=1}^{c} c_i = 0$; t_j is the j^{th} year effect: $j = 1, ..., t$ and $\sum_{j=1}^{t} t_j = 0$; and ϵ_{ij} is the error term.

2.1.6 Model Using Poisson Distribution

This method assumes that the data are Poisson-distributed; thus the variance is propotional to the mean. We used the following model:

$$Y_{ij} = \mu c_i t_j (1 + \epsilon_{ij}), \tag{2.2}$$

where μ is an arbitrary constant that could be set to 1 since there are no restrictions placed on the c_i's and t_j's; c_i is the i^{th} creek effect: $i = 1, ..., c$; t_j is the j^{th} year effect: $j = 1, ..., t$; and ϵ_{ij} is the error term. Interaction effects between creeks and years are

not considered as they are confounded with the error. This gives $E(y_{ij}) = c_i t_j$ and $Var(y_{ij}) = c_i t_j$, thus $Y_{ij} \sim Poisson(c_i t_j)$ with

$$P(Y_{ij} \quad = \quad y_{ij}) = \frac{e^{-c_i t_j}(c_i t_j)^{y_{ij}}}{y_{ij}!}$$

and the constraint $\prod_i^c c_i = 1$. Then up to a constant the log likelihood function is:

$$l(\mathbf{c}, \mathbf{t}|y_{ij}) = \sum_i^c \sum_j^t (y_{ij} \ln(c_i t_j) - c_i t_j) \tag{2.3}$$

where c and t indicate that the sum was taken over all observed creeks and all observed years such that $c \leq c$ and $t \leq t$. To maximize this function (2.3) under the constraint that $\prod_i^c c_i = 1$ or equivalently that $\sum_i^c \ln(c_i) = 0$, we can form the Lagrangian equation:

$$Ł(\mathbf{c}, \mathbf{t}, \lambda) = l(\mathbf{c}, \mathbf{t}|y_{ij}) + \lambda \sum_i^c \ln(c_i)$$

and set the partial derivatives with respect to c_i and t_j equal to zero.

$$\frac{\partial Ł}{\partial c_i} = \sum_j^t \left(\frac{y_{ij}}{c_i} - t_j \right) + \frac{\lambda}{c_i} = 0$$

$$\frac{\partial Ł}{\partial t_j} = \sum_i^c \left(\frac{y_{ij}}{t_j} - c_i \right) = 0$$

Lagrange multipliers indicate the rate at which the maximum value increases as the constraint is relaxed. Here, because the maximum value is independent of the constraint, λ must be zero, and therefore the maximum likelihood estimates are simply:

$$c_i = \frac{\sum_j^t y_{ij}}{\sum_j^t t_j}$$

$$t_j = \frac{\sum_i^c y_{ij}}{\sum_i^c c_i}$$

For this Poisson distribution, the parameters can be approximated by the Iterated Reweighted Least Squares (IRLS) generalized linear model. The following problem is

minimized through IRLS:

$$\sum_{i}^{c}\sum_{j}^{t}\frac{(y_{ij}-c_it_j)^2}{c_it_j} \qquad \text{subject to } \prod_{i}^{c} c_i = 1$$

which converges to the same solution as the maximum likelihood equations. In this context, the generalized linear model (GLM) has the intuitive appeal of depending on means and sums-of-squares about supposed means. We therefore fit the model using the IRLS method from SPLUS.

As in the previous method, only main effects are considered and interactions between creek and year are assumed to be negligible as they are inseparable from the error component. Missing values are infilled by applying the fitted Poisson model and back transforming.

2.1.7 Model Using Gamma Distribution

The final imputation method explored here uses a Gamma distribution in which the variance (ν) is proportional to the square of the mean; thus allowing a relatively larger variance to mean relation. We used the model as in (2.2) with $E(y_{ij}) = c_it_j$ and $\text{Var}(y_{ij}) \propto (c_it_j)^2$. Thus $Y_{ij} \sim Gamma(c_it_j, \nu)$ with

$$P(Y_{ij}=y_{ij}) = \frac{1}{\Gamma(\nu)}\left(\frac{\nu y_{ij}}{c_it_j}\right)^{\nu}\frac{1}{y_{ij}}e^{-\nu\frac{y_{ij}}{c_it_j}}$$

and the constraint $\prod_i^c c_i = 1$. Interaction effects between creeks and years are not considered as they are confounded with the error.
The log-likelihood is therefore a fixed constant plus:

$$l(\mathbf{c},\mathbf{t}|y_{ij}) = \sum_{i}^{c}\sum_{j}^{t}\left[\ln\Gamma(\nu)+\nu\ln\left(\frac{\nu y_{ij}}{c_it_j}\right)-\frac{\nu y_{ij}}{c_it_j}\right] .$$

As in the Poisson model, to maximize the likelihood function under the constraint $\prod_i^c c_i$=1, we form the Lagrangian equation:

$$Ł(\mathbf{c}, \mathbf{t}, \lambda) = l(\mathbf{c}, \mathbf{t}|y_{ij}) + \lambda \sum_i^c \ln(c_i).$$

When the first derivative of the Lagrangian equation is taken with respect to c_i and t_j, and setting these derivatives equal to zero, solutions for the maximum likelihood estimators can be found.

$$\frac{\partial Ł}{\partial c_i} = \sum_j^t \left[-\frac{\nu}{c_i} + \frac{\nu y_{ij}}{t_j c_i^2} \right] + \frac{\lambda}{c_i} = 0.$$

As in the case of the Poisson model, the Lagrange multiplier equals zero, and the maximum likelihood equations simplify to:

$$c_i = \frac{1}{t} \sum_j^t \frac{y_{ij}}{t_j} \quad \text{for all } i.$$

Similarly

$$\frac{\partial Ł}{\partial t_j} = 0 \qquad \text{gives } \; t_j = \frac{1}{c} \sum_i^c \frac{y_{ij}}{c_i} \; \text{ for all } j\,.$$

Similar to the Poisson model, the GLM solution to the IRLS problem is approximated by the maximum likelihood equations for the Gamma model. Therefore instead of using the IRLS solutions from SPLUS for the variance proportional to the mean squared, we used the maximum likelihood equations to numerically find the solutions.

The maximum likelihood algorithm starts by making reasonable guesses for the starting values of $c_1, \ldots, c_c$. Each unknown year parameter (t_j) is considered separately and the estimation problem is reduced to an estimation of a mean:

$$t_j = mean\left(\frac{y_{1,j}}{c_1}, \frac{y_{2,j}}{c_2}, ..., \frac{y_{c,j}}{c_c} \right)$$

To update the estimate of c_i, the t_j's from above are used to solve for each creek parameter (c_i):

$$c_i = mean\left(\frac{y_{i,1}}{t_1}, \frac{y_{i,2}}{t_2}, ..., \frac{y_{i,t}}{t_t}\right) ,$$

and the maximum likelihood equations are iterated until convergence. Once the parameters are determined, the model is then used to fill in the gaps in the abundance record.

Chapter 3

Comparison of imputation methods

This chapter discusses the strengths and weaknesses of the seven imputation methods described in the previous chapter. The chapter contains a simulation study that investigates performance in the context of the Thompson River coho population with 40.7% of the observations missing. We then examine performance when the missing value pattern is extreme (72.0% missing) by evaluating the sockeye salmon abundance record for an area on the British Columbia North Coast.

3.1 Simulation Study

3.1.1 Review of Coho Life Cycle Ecology

Coho salmon have a 3-year life cycle in which they reproduce only once and therefore have a semelparous life history strategy (Sandercock 1991). From November to January, adults migrate from the ocean to their natal streams, where spawning occurs. After spawning the adult salmon die. Coho fry emerge the next spring and remain in freshwater usually for one year before migrating to the sea as smolts. This once year residency in creeks is a potential bottleneck through limited carrying capacity specific to each creek. The majority of these fish remain in the ocean for 18 months before returning to freshwater to begin the three year cycle again.

Between 1976 and 1990, spawning coho populations in British Columbia were healthy. In the past decade, however, there has been a considerable decline in numbers of spawning coho salmon. This decline is thought to be in large part due to poor marine survival (Nickelson et al. 1994). Specifically, these declines have been correlated with various ocean parameters including upwelling and nearshore temperatures (Nickelson 1986, Fisher and Pearcy 1988). We incorporated the three year life cycle and the population trend of this species into the design of the simulation study with spawning abundances based on the Thompson River coho records.

3.1.2 Details of Simulation

The goal of the simulation study was to evaluate performance of the infilling methods by simulating data to approximate that of the Thompson River coho salmon records. We generated a data array of 100 matrices based on a stock-recruitment curve from Black Creek on Vancouver Island, the only reliably observed wild coho salmon population in B.C. (Routledge and Irvine 1999). The stock-recruitment curve was generated through three parameters chosen to specify:

r the ratio of number in year j to number in year $j-3$,

k the average number of fish in the creek during the first 10 years of records (from 1976 to 1985) when coho salmon populations were considered to be stable,

rk the carrying capacity, and

τ an extra variance parameter.

Specifically if y_{ij} is the number in creek i in year j, then the stock-recruitment curve is modeled by

$$\begin{aligned} y_{ij} &= r y_{i,j-3} \text{ for } y_{i,j-3} \leq k, \\ &= rk \quad \text{for } y_{i,j-3} > k, \end{aligned}$$

where $y_{i,j-3}$ is the number of salmon spawning in the previous generation. Thus the parameter r is the ratio of population from one parent generation to the next, given

that the population size is below the carrying capacity of the stream and that resources are unlimited. This parameter was set at 1.9 which was based on data collected from Black Creek, a coho salmon stream on Vancouver Island (Routledge and Irvine 1999), and was set at the same level for all 89 creeks in the simulation.

The carrying capacity parameter (rk) is the maximum sustainable population size for each creek. If a creek's population was below the carrying capacity, it would approach rk in the subsequent generations at the rate of r=1.9. If more coho salmon were in the spawning channels than k, then the number of salmon in the following generation can be no more than if k salmon had been on the spawning grounds. This number k is fixed for each creek and specifies the number of spawning salmon required to fully stock the creek with juvenile salmon.

For this simulation study we introduced chance fluctuations about the recruitment curve using the Poisson-Inverse Gaussian distribution. The PIG distribution is widely used as a parametric model for extra-Poisson variability. Unlike its major competitor, the negative binomial, it can have a long right-hand tail without a sharp spike at zero (Dean et al. 19??). It is also easier to manipulate analytically than the Poisson log-normal distribution.

Consider the mixed Poisson model where:

$$f(Y = y|\mathbf{x}) = \int_0^\infty e^{-\nu\mu}\frac{[\nu\mu]^y}{y!}g(\nu)d\nu, \quad y = 0, 1, 2, ..., \tag{3.1}$$

where Y is the number of spawning salmon which has a Poisson distribution with mean $\nu\mu$. ν is a random effect, and $g(\nu)$ is a probability density function such as an Inverse Gaussian density:

$$g(\nu) = \frac{1}{\sqrt{2\pi\tau\nu^3}}e^{\frac{-(\nu-1)^2}{2\tau\nu}} \quad , \nu > 0 \tag{3.2}$$

The distribution of N given $\mathbf{x}$ (3.1) then has a Poisson-Inverse Gaussian distribution with mean and variance functions: μ and $\mu(1 + \mu\tau)$, respectively. For a PIG distribution, τ is the variance of the random effect ν and dictates the amount of extra Poisson variation. The value of τ was determined for each creek from the variability of the

first 10 years of data (σ^2) and the number of spawning salmon three years before $y_{i,j-3}$.

$$\tau = \frac{\sigma^2}{y_{i,j-3}^2}.$$

This extra-Poisson parameter is desirable here because of the considerable environmental variability inherent in the system.

Because of the three-year generation time for coho salmon, values in the time series for each creek were simulated based on the observation three years before ($y_{i,j-3}$) and the extra variance parameter τ. The number of fish at time t_j was generated by multiplying the value three years before $y_{i,j-3}$ by the parameter r. If the value three years before was greater than the carrying capacity, then rk was used instead of $y_{i,j-3}$.

$$y_{i,j} \sim \text{PIG}(ry_{i,j-3}, \tau_i) \quad or \quad \sim \text{PIG}(rk_i, \tau_i)$$

An extra 20 years of data prior to the actual start of the record were generated. This was to ensure that the population had reached an equilibrium around rk at the start of the comparative analysis.

Impact of declining marine survival

Ocean conditions appear to have negatively affected coho salmon survival. Lower ocean temperatures have been correlated with higher ocean survival, and higher ocean temperatures with lower survival (Johnson 1988). If marine temperatures and oceanic conditions (currents and upwellings) are being altered through global warming, coho survival will be affected. A progressive loss in marine survival could account for the observed population decline in coho spawning numbers.

We reproduced the decline in spawning coho salmon in the past decade by using a logistic deline equation. We used the Verhulst-Pearl equation to introduce a decline in the carrying capacity and the intrinsic growth rate of each creek. These changes were introduced to approximate the decline in survival of Thompson coho salmon of the past decade and to evaluate the performance of the different imputation methods in

detecting the trend and in imputing reasonable values for missing data. The decline followed a deterministic logistic curve with a rate of decline:

$$\frac{dN(t)}{dt} = -RN\left(1 - \frac{N}{rk}\right)$$

where R gives the rate of decline of the spawning population $N(t)$, rk is the carrying capacity and

$$N(t_0) = \frac{rk}{1 + e^{-C}} \approx rk \quad \text{(for some large positive } C\text{)},$$

The number of fish at time t is given by integrating with respect to t:.

$$N(t) = \frac{rk}{1 + e^{-C}e^{R(t-t_0)}} \tag{3.3}$$

For a declining curve, the intrinsic growth parameter was arbitrarily set at $R = 0.5$, and the onset of decline was set at $t_0 = 11.5$, the constant was set at $C = 6$ to give from (3.3) the following logistic equation for N:

$$N(t) = \frac{rk}{1 + e^{-6}e^{0.5(t-11.5)}}$$

During the decline phase, the extra-variance component τ was held constant. This was to ensure that the PIG variability declined as the population declined.

The simulation was now set to compare the seven imputation methods. We used the same missing data pattern as existed in the Thompson River dataset for each of the 100 simulated data matrices. We then infilled the gaps with values from the seven different imputation methods.

3.2 Evaluating Performance of Methods from the Simulation Study

3.2.1 Methods

Plotting Annual Totals

We plotted the yearly totals averaged over all creeks and all 100 simulations for each of the seven imputation methods against the average yearly totals from the known model ($fig.3.1a : f$). This was done to visually assess how the imputation methods were performing on average. In particular, we were interested to see if there were consistent biases associated with any of the methods.

Jackknifing Dependent Samples

We used a jackknifed sums-of-squares estimate of dependent samples as described below to evaluate three components of error from the infilling methods relative to the known model. In this simulation, we know the actual model from which the data were generated. Thus it is straightforward to compare performance of imputation methods.

We used the following notation in the sums-of-squares equations:

Y_{ijk} is the number of fish in the i^{th} creek, j^{th} year and k^{th} simulation, and is an element in the $89 \times 26 \times 100$ imputed array.
Therefore, $\bar{Y}_{.jk} = \frac{1}{89}\sum_{i=1}^{89} Y_{ijk}$ is the average number of fish across all 89 creeks for the j^{th} year and the k^{th} simulation and $\bar{Y}_{.j.} = \frac{1}{100}\sum_{k=1}^{n} \bar{Y}_{.jk}$ tracks the average decline across years in the imputed array.

Z_{ijk} is the number of fish in the i^{th} creek, j^{th} year and k^{th} simulation, and is an element in the $89 \times 26 \times 100$ array generated from the known model.
Likewise, $\bar{Z}_{.jk} = \frac{1}{89}\sum_{i=1}^{89} Z_{ijk}$ is the average number of fish across all 89 creeks for the j^{th} year and the k^{th} simulation and $\bar{Z}_{.j.} = \frac{1}{100}\sum_{k=1}^{n} \bar{Z}_{.jk}$ tracks the average decline across all years in the known model.

We used the following equations in calculating three components to the error sums-of-squares (bias, chance error, and total error) for each imputation method.

$$SS_{bias} = \sum_{j=1}^{26} \frac{(\bar{Z}_{.j.} - \bar{Y}_{.j.})^2}{(\bar{Y}_{.j.})^2}$$

where $\bar{Y}_{.j.}$ and $\bar{Z}_{.j.}$ should be very close and SS_{bias} small if there is little systematic bias. The denominator is a weighting to account for the greater variability in years with more abundant fish.

$$SS_{chance} = \frac{1}{100} \times \sum_{j=1}^{26} \sum_{k=1}^{100} \frac{(\bar{Z}_{.jk} - \bar{Z}_{.j.})^2}{(\bar{Y}_{.j.})^2}$$

where SS_{chance} gauges how much chance variation there is about the means.

$$SS_{total} = \frac{1}{100} \times \sum_{j=1}^{26} \sum_{k=1}^{100} \frac{(\bar{Z}_{.jk} - \bar{Y}_{.j.})^2}{(\bar{Y}_{.j.})^2}$$

where SS_{total} is algebraically the sum of the above two equations, and is a gauge of the overall error.

Because the imputed arrays for each of the seven imputation methods are based on the same simulated dataset, the arrays are not independent samples. Therefore we made pairwise comparisons and tested the null hypothesis that these limiting values are the same for any pair of imputation methods, i.e., that their differences is zero.

The sums-of-squares are calculated from only 100 simulations, therefore we were concerned that the sum-of-squares estimates may be biased estimates for an indefinitely large number of simulations. Therefore we elected to obtain numerical approximations for the sums-of-squares estimates using a jackknife estimation procedure. This jackknife procedure is now described before relating back to the sums-of-squares procedure.

The jackknife procedure consists of taking repeated subsamples of the original sample of n independent observations by omitting a single observation at a time. Thus, each subsample consists of $(n-1)$ observations formed by deleting a different observation from the sample. The jackknife estimate and its standard error are then calculated from these truncated subsamples. For example, suppose θ is the parameter of interest and $\hat{\theta}_{(1)}$,$\hat{\theta}_{(2)}$, ..., $\hat{\theta}_{(n)}$ are estimates of θ based on n subsamples, each of size $(n-1)$. The jackknife estimate of θ is calculated as the mean of the subsample estimates of θ:

$$\hat{\theta}_{(.)} = \frac{\sum_{k=1}^{n} \hat{\theta}_{(k)}}{n} \tag{3.4}$$

The jackknife estimate of the standard error of $\hat{\theta}_{(.)}$ is

$$SE\,(\,\hat{\theta}_{(.)}\,) = \frac{1}{n}\sqrt{var\ \left[\ \hat{\theta}_{(k)}\,'s\ \right]} \tag{3.5}$$

In our jackknife procedure, we tested the hypothesis that the expected difference in sums of squares was equal to zero: $H_0 = E(\theta) = 0$, where $\theta = SS_{method\,A} - SS_{method\,B}$, and A and B are from methods 1 through 7 as described in Chapter 2. The parameter of interest is then θ and the asymptotic expansion of its expectation is:

$$E(\theta_n) = \theta + \frac{a}{n} + \frac{b}{n^2} + ...$$

and of its jackknifed expectation ($E(\theta_{n-1})$) is:

$$E(\theta_{n-1,k}) = \theta + \frac{a}{n-1} + \frac{b}{(n-1)^2} + ...$$

Assuming that terms higher than first order are negligible, this gives two equations:

$$\begin{aligned} nE(T_n) &= n\theta + a \\ (n-1)E(T_{n-1,k}) &= (n-1)\theta + a \end{aligned}$$

These can be subtracted to obtain $\hat{\theta}_{(k)}$ an unbiased estimate of θ with the bias term of $\frac{1}{n}$ eliminated:

$$\theta_{(k)} = n(T_n) - (n-1)(T_{n-1,k}) \tag{3.6}$$

We calculated this for the $n = 100$ jackknifed estimates of $\theta_{(k)}$ for $k = 1, \ldots, 100$. The $\theta_{(k)}$ are functions of U-statistics and by jackknifing a U-statistic, we get an asymptotically normally distributed random variable with a mean as in (3.4) and standard error as in (3.5) (ref from sir R.).

The difference in sums-of-squares between methods A and B in bias, chance error and total error following (3.6), are estimated by:

$$\hat{\theta}_{(.)} = \frac{1}{100} \times \sum_{k=1}^{100} \Big[100 \left(SS_{A;\, non-jack.} - SS_{B;\, non-jack.} \right) - 99 (SS_{A;\, jack._{[k]}} - SS_{B;\, jack._{[k]}}) \Big]$$

Also, the relationship between the jackknifed difference in sums of squares is:

$$\hat{\theta}_{(.)\ bias} + \hat{\theta}_{(.)\ chance} = \hat{\theta}_{(.)\ total}.$$

B ↓	**A →**	**ZI**	**NV**	**I/E**	**AS**	**TL**	**GL**	**NL**
ZI	$\hat{\theta}_{(.)bias}$	0						
	$(st.error)$	(0)						
	$\hat{\theta}_{(.)chance}$	0						
	$(st.error)$	(0)						
NV	$\hat{\theta}_{(.)bias}$	**1.45**	0					
	$(st.error)$	(0.019)	(0)					
	$\hat{\theta}_{(.)chance}$	-0.0376	0					
	$(st.error)$	(0.022)	(0)					
I/E	$\hat{\theta}_{(.)bias}$	**1.47**	**0.0238**	0				
	$(st.error)$	(0.019)	(0.0061)	(0)				
	$\hat{\theta}_{(.)chance}$	-0.290	-0.252	0				
	$(st.error)$	(0.13)	(0.13)	(0)				
AS	$\hat{\theta}_{(.)bias}$	**1.46**	0.00508	-0.0188	0			
	$(st.error)$	(0.018)	(0.069)	(0.0065)	(0)			
	$\hat{\theta}_{(.)chance}$	0.0189	0.0187	0.271	0			
	$(st.error)$	(0.047)	(0.049)	(0.096)	(0)			
TL	$\hat{\theta}_{(.)bias}$	**1.42**	**-0.0347**	**-0.0585**	**0.0398**	0		
	$(st.error)$	(0.018)	(0.0051)	(0.0065)	(0.0040)	(0)		
	$\hat{\theta}_{(.)chance}$	-0.0647	**0.102**	0.354	-0.0836	0		
	$(st.error)$	(0.015)	(0.020)	(0.13)	(0.044)	(0)		
GL	$\hat{\theta}_{(.)bias}$	**1.48**	**0.0247**	0.000084	**0.0196**	**0.0594**	0	
	$(st.error)$	(0.018)	(0.0045)	(0.0069)	(0.0046)	(0.0034)	(0)	
	$\hat{\theta}_{(.)chance}$	**-0.235**	**-0.197**	0.0547	**-0.216**	**-0.299**	0	
	$(st.error)$	(0.028)	(0.022)	(0.012)	(0.037)	(0.023)	(0)	
NL	$\hat{\theta}_{(.)bias}$	**1.47**	0.015	-0.00878	0.00998	**0.0497**	**-0.0096**	0
	$(st.error)$	(0.018)	(0.0049)	(0.0070)	(0.0045)	(0.0031)	(0.0013)	(0)
	$\hat{\theta}_{(.)chance}$	0.0289	0.0664	0.318	0.0478	**-0.0358**	**0.264**	0
	$(st.error)$	(0.015)	(0.020)	(0.13)	(0.042)	(0.0047)	(0.021)	(0)

Table 3.1: Jackknifed estimates of differences in error sums-of-squares between methods A and B. Zero-infilled (ZI), Nearest-value (NV), Interp./Extrap. (I/E), Average of scaled values (AS), Transformed linear model (TL), Poisson model (PM), Gamma model (GL). **Bold-face text** represents significant differences at the 5% level with Bonferroni adjusted p-values. Positive values indicate method A has greater sum-of-squares error, negative values indicate method B has greater error.

Figure 3.1: Annual averages of numbers of spawning coho salmon in the Thompson River system of British Columbia; Zero-infilled, Nearest-value, Interpolation/extrapolation, Averaging scaled values. Standard error bars were calculated as the square root of the variance divided by the number of data points for that year.

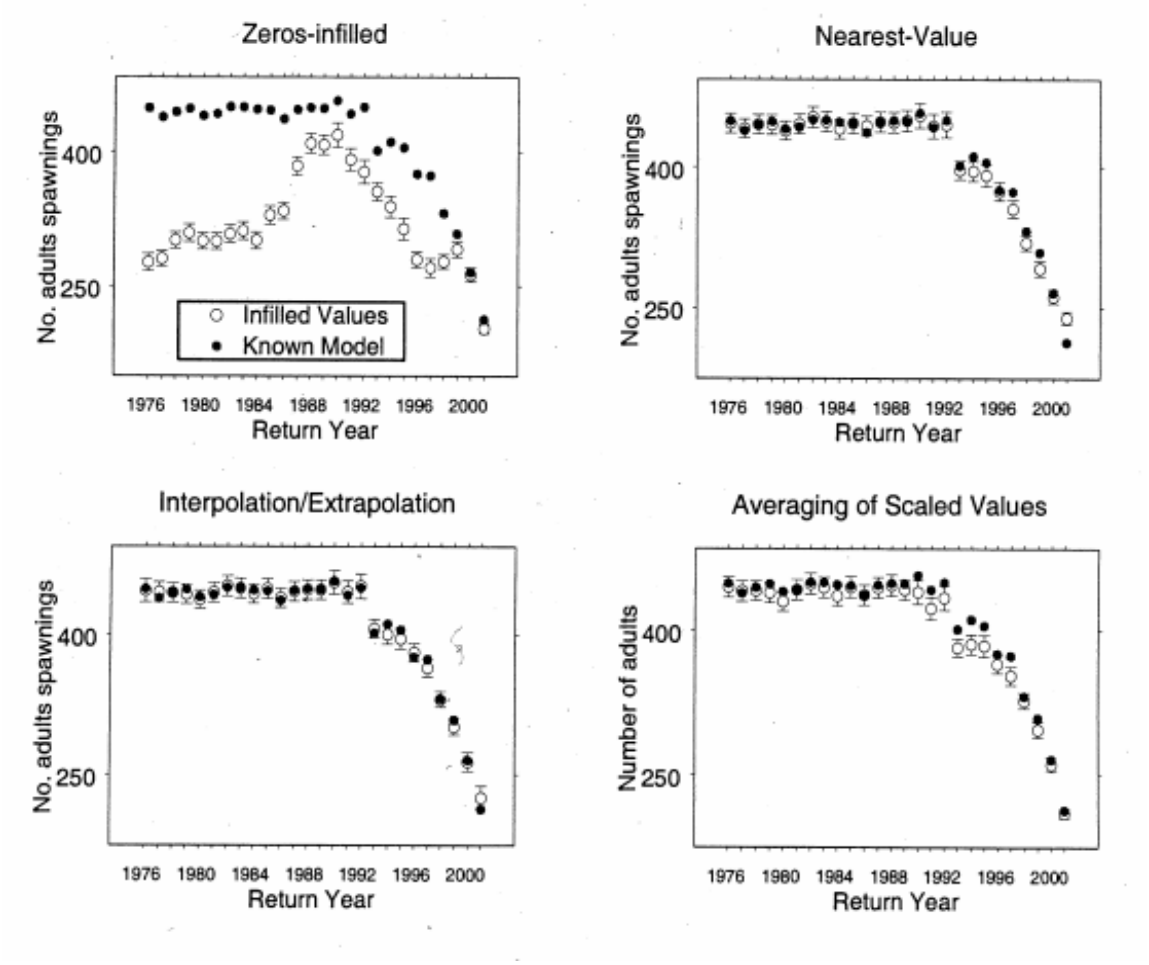


Figure 3.2: Annual averages of numbers of spawning coho salmon in the Thompson River system of British Columbia; ANOVA-based methods. Standard error bars were calculated as the square root of the variance divided by the number of data points for that year.

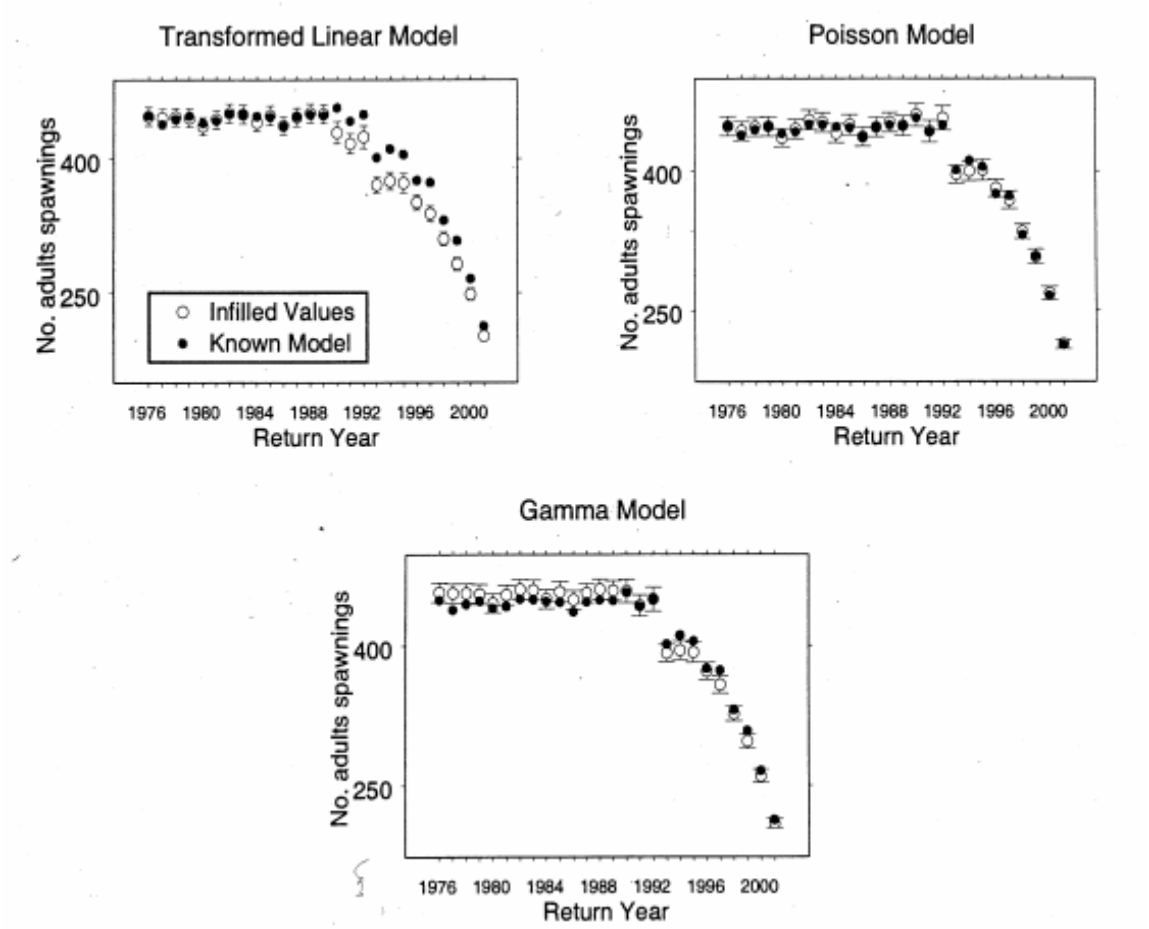


Figure 3.3: Plot of differences between the infilled values and the known values for the seven imputation methods

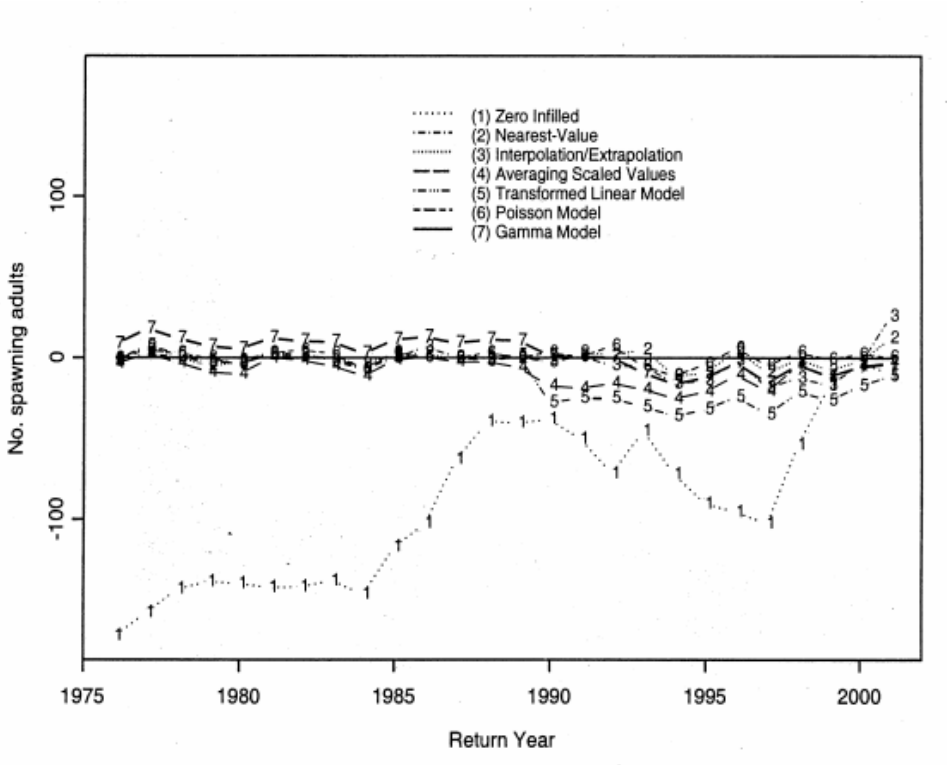
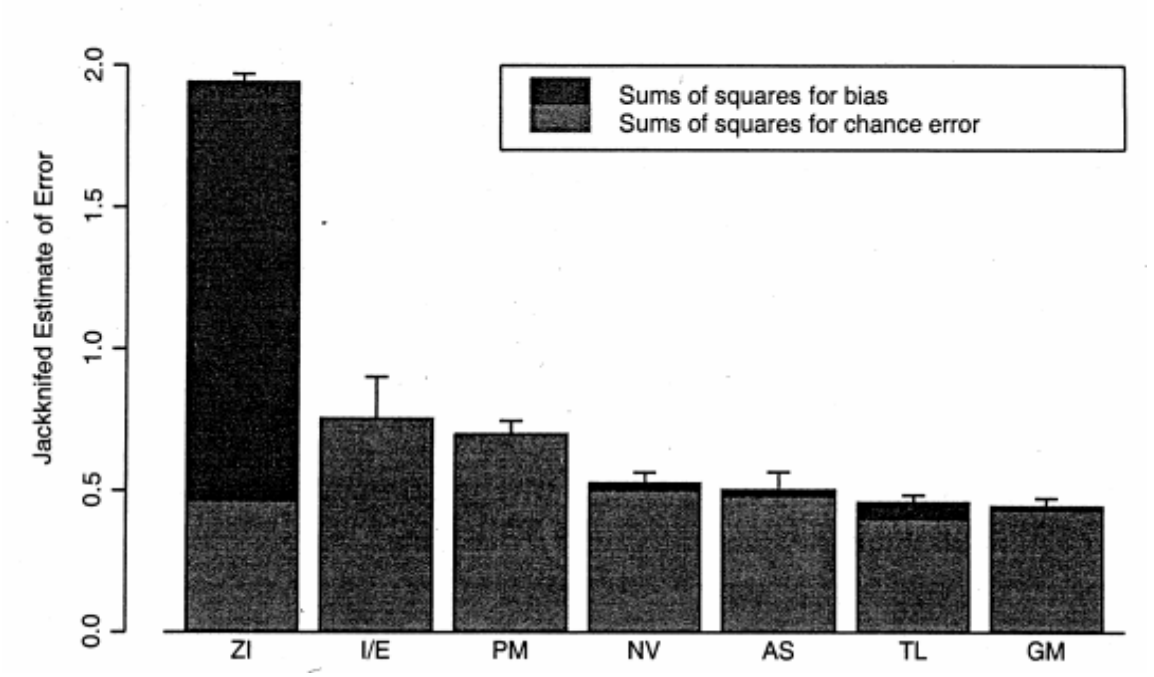


Figure 3.4: Jackknifed Bias and Chance Error Sums of Squares with Standard Error bars calculated from the Overall Error Sums of Squares. Zero-infilled (ZI), Nearest-value (NV), Interp./Extrap. (I/E), Average of scaled values (AS), Transformed linear model (TL), Poisson model (PM), Gamma model (GM).



3.2.2 Results

Of the seven methods, and for the pattern of missing values observed in the Thompson coho spawning records, three methods are not as effective interpolators as the remaining four. We found the zero-infilled, nearest value, and interpolation/extrapolation methods unsuitable, whereas the averaging of scaled values, tranformed linear, Poisson, Gamma models seem to perform better. Table 3.1 gives the jackknifed sums-of-squares estimates of bias and chance error for each pair of the seven imputation methods. Insight from this error analysis and a comparative data exploration aided in selecting the better methods. Interpretation of results is presented below.

Zero-infilled

The zero-infilled method is clearly inadequate. The assumption that when data are missing there were no fish, is clearly untrue and introduces an obvious and unacceptable bias. The jackknifed sums-of-squares analysis ($fig.3.1$) and the graph of annual totals ($fig.3.2a$) show clearly the degree of bias introduced by this method.

Nearest value

Although the performance of the nearest-value method as indicated by the jackknifed sums of squares analysis is not unreasonable, we have concerns about bias in this method. A major concern is that this method does not allow for trends to be observed within individual creeks when a gap is to be filled. Only horizontal shifts over time are possible. This may be the cause of the bias evident in the decline phase of $fig.3.2b$. Furthermore, because infilled values rely only on one other data point, an unrepresentative observation combined with large data gaps will cause this method to work poorly. Therefore there are better tools for monitoring trends in populations.

Interpolation/Extrapolation

Although the interpolation/extrapolation method has a small bias component to the total error, the standard error of the method is very large. Thus the method is unreliable. Furthermore, because of "extrapolation", impossibly large, or worse, negative expected values are possible. As a result, we regard this method as unsuitable.

Averaging of Scaled Values

This method performs adequately in this sums of squares analysis, however, there are some concerns about this method. In particular, infilled values are sensitive to the maximum observed value in that they can never be greater than this observed maximum. This may be why this method consistently underestimates the population in the decline phase and introduces a non-trivial source of bias ($fig.3.2d$).

Gamma Model

Although this method tends to have smaller bias and chance error sums of squares than the Averaging of Scaled Values method, it is not statistically different from this method in any of the three sums of squares components (fig. 3??). However, according to the graphs of annual averages ($fig.3.2f$), imputed values are consistently higher than the known data during the stable population period, and consistently lower during the decline phase. The algorithm for this model, if it converges, will converge under the gamma distribution. There are two problems with fitting these data to the gamma distribution. The first is that this kind of data is discrete and the Gamma distribution is continuous. The second is our data contain zeros and these zeros have a dominating effect on our parameter estimates. Therefore there are problems inherent to the underlying fitted distribution with this Gamma model. This method, then, does not improve the imputation in this instance over the simpler Averaging of Scaled Values method.

Transformed Linear Model

The responses after the transformation were assumed to be normally distributed, have constant variance independent of the mean, and have systematic effects that combine additively. A single transformation is asked to produce three things simultaneously. It should not be surprising then that this method is less effective than the GLM method described below. The sums-of-squares analysis showed that a large component of the error in this infilling method is attributed to bias. The graph of annual totals (fig3.1?) shows that the bias is substantial in the decline phase indicating that this method is particularly poor at following downward trends.

Poisson Model

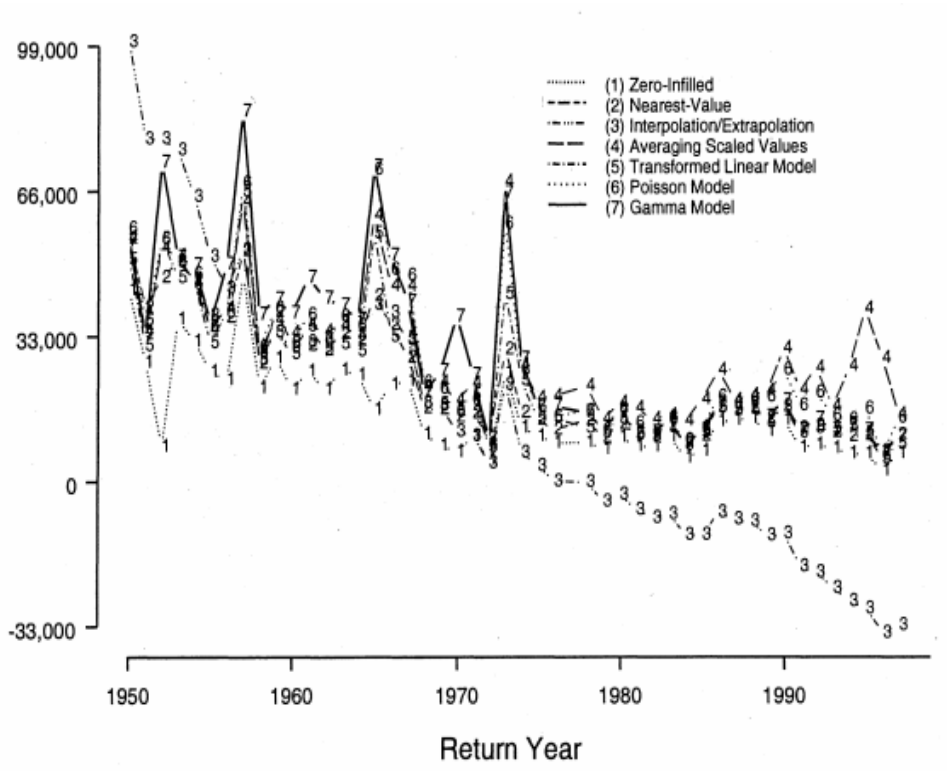
A big advantage over the traditional transformation approach in the Poisson generalized linear model approach is the freedom to specify the variance to mean relation and the error distribution. Therefore, when GLM's are used to fit the data, a single transformation is not trying to do several jobs. The sums of squares analysis shows that a small component of the error is attributable to bias, but GLM gives a larger chance error component than the linear model method. We would preferentially select a method with a larger amount of chance error over one with a large bias component particularly as the graphs of annual totals ($fig.3.2g$) suggests this method works the best on average.

3.3 Evaluating methods when the missing data pattern is extreme: North Coast sockeye

These sockeye salmon (*O. nerka*) spawning records were collected between 1950 and 1997 for 68 creeks on the North Coast of British Columbia. Sampling intensity could be described at best as sporadic with 72.0% of the records missing. The scarcity of records is particularly poor in the latter half of the series. This large number of missing values and the lack of balance in this pattern has caused problems in many

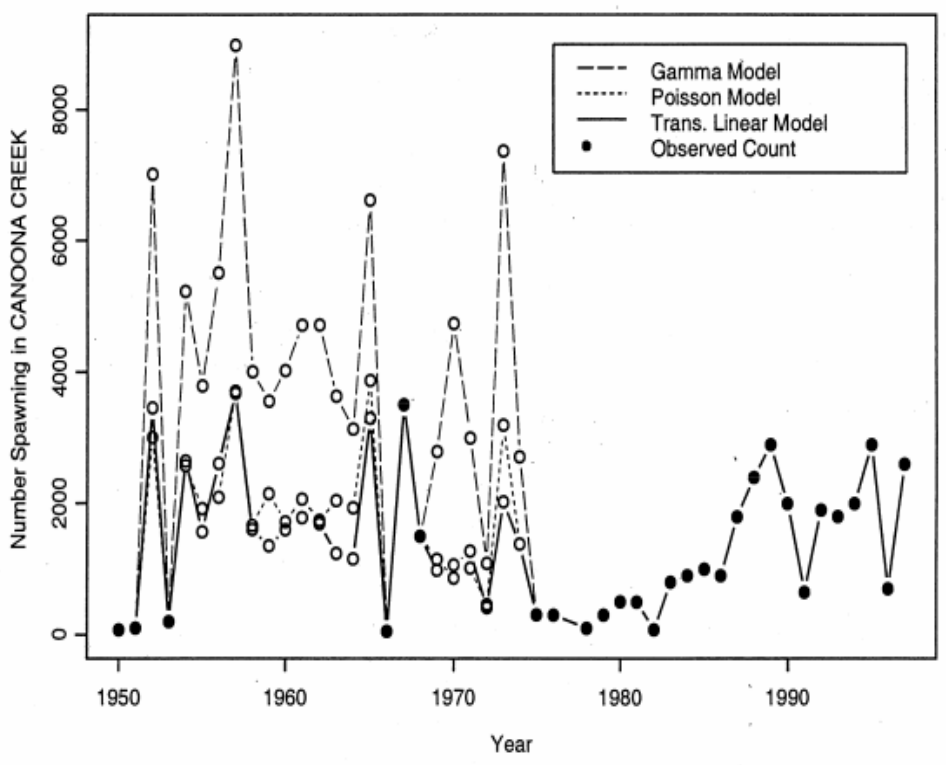
imputation methods. In addition, the lack of area-wide consistency in trends across creeks and years caused instabilities with the more sophisticated methods. These instabilities are examined here.

Figure 3.5: Annual totals of spawning salmon for 68 creeks on the North Coast of British Columbia.



We plotted the total numbers of sockeye salmon for the North Coast area as estimated by the seven imputation methods (fig.3.3). From this plot, it was obvious that three of the methods do not perform well.

Figure 3.6: Annual numbers of spawning sockeye salmon for Canoona Creeks on the North Coast of British Columbia.



Zero-infilled, Interpolation/extrapolation, Averaging of scaled values Methods

The zero-infilled method substantially underestimates the number of fish spawning, particularly in the later years where the number of observations was lowest.

The weakness of the interpolation/extrapolation method is clearly seen in fig. ??. At either end of the data record where the method extrapolates rather than interpolates, absurdly large and negative annual totals are obtained.

The right hand section of(fig.3.3?) shows that the Averaging Scaled Values method predicts substantially higher totals than any other method. As most of the data are more complete for the first half of the record, and most of the data have a roughly decreasing trend, the creeks that contain data in the latter years have a big influence on the infilled values. This method fails here because some of the creeks that have more complete data records for the later dates, also have an increasing trend. This causes the average scaled column value to be unduly large, which results in the missing values being inflated without much apparent justification. This is clearly seen in the latter half of the series in (fig.3.3).

Transformed Linear, Poisson and Gamma Models

The problem of having inconsistent trends across all the creeks is made evident with those methods that model creek and year effects but ignore the interaction effects (transformed linear, Poisson and Gamma models). Inconsistent trends across creeks are a result of interactions between year and creeks, and because these interactions cannot be modeled, the imputations are untrustworthy. To demonstrate, we plotted the imputed values of these three methods for Canoona Creek (fig.3.4). Canoona Creek is a creek unlike most of the other creeks in the North Coast area as the sockeye population does not decrease over the time period and missing values are mostly in the first half of the time series. In particular, the Gamma model method does a poor job.

The transformed linear model method acts as if the residuals are lognormal. Hence,

a large, positive residual will not be unexpected as the lognormal distribution has a long, right-hand tail. The Poisson and Gamma models act as if the errors are normal, and hence a large, positive residual will be less consistent with the model. These two methods will react to a large positive residual by increasing the estimated "year" effect, and hence produce the larger peak. The variance of the Gamma model is proportional to the square of the mean, and therefore the compensation is greater in this model, and the peaks are higher in Cannoona Creek.

Nearest Value

The nearest value method considers only the information from the colsest value for that creek. It cannot include information from other creeks for a given year to help with the imputation. In this circumstance however, when there are conflicting trends occuring within the same area, the nearest value method performs moderately well.

3.3.1 Chapter Summary

If the pattern of missing values is very extreme, even the best methods fail. This emphasizes the importance of a good long-term sampling design, an issue to which we now turn.

Chapter 4

Survey Design

Accurately monitoring salmon populations is a critical step in ensuring the persistence of a sustainable population of a species. This chapter is a discussion of the meaning of balance in a design and the optimal properties of Balance Incomplete Block Designs with reference to missing data in salmon spawning surveys. In addition, due to the 3-year life cycle of coho salmon, we consider existing theory on rotation designs as a rough guideline for recommendations. This chapter focuses on a survey protocol for a population of coho salmon, although the same issues are relevent for all pacific anadromous salmon species. The chapter end by incorporating conclusions fromn the simulation studies and existing theory.

We designed a simple scenario to demonstrate the importance of balance in survey design. Using the same 100 data matrices as were generated for the Thompson River coho simulation in Chapter 3, we removed the same proportion of observations from the record but redistributed the missing pattern such that the pattern was balanced throughout the creeks and years. We removed 38 of the 89 observations for each of the 26 years and for all of the 100 simulated data matrices. Furthermore, we assigned weights to the sampling scheme with the idea that some creeks have more important populations than others. Three weight categories were introduced based on creek size: small, medium and large to which the weights 1, 2, and 3 were assigned respectively. The proportion of creeks within each category was roughly equivalent. We then used a random number generator to take a weighted random sample from the 89 creeks.

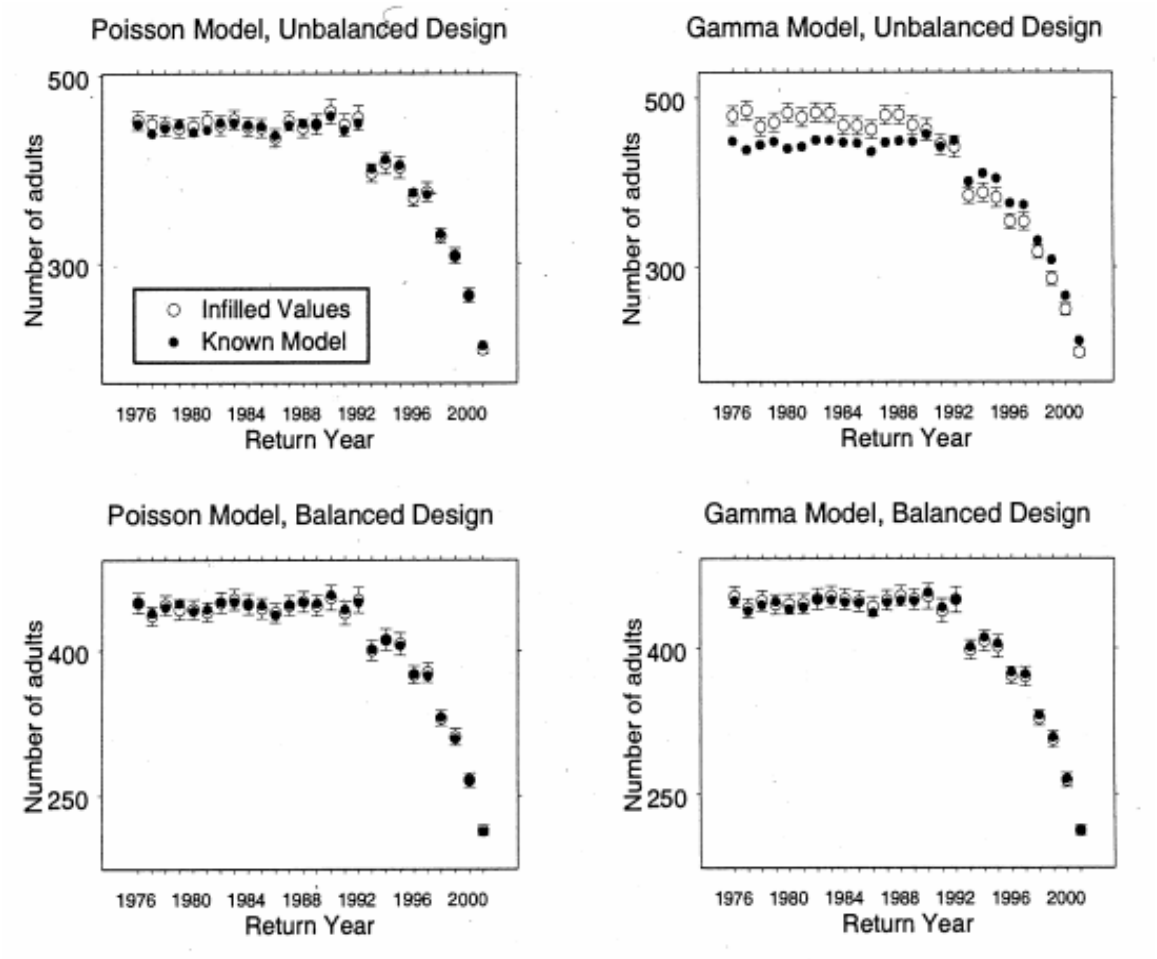
	1^{st} 13 years	2^{nd} 13 years
weight 1	no missing values	94% missing
	94% missing	no missing values
weight 2	no missing values	85% missing
	85% missing	no missing values
weight 3	no missing values	71% missing
	71% missing	no missing values

Table 4.1: Missing pattern used in the simulation of an unbalanced sampling design

To contrast with the balanced sampling scenario, we devised another sampling scheme with the same proportion of missing values per year (38/89) but with an unbalanced pattern. We divided each of the three abundance-based categories of creeks into two, roughly equal subsets (4). We also divided the range of years for which the data are used into two roughly equal halves. Within each abundance-based category of creeks, we selected one half of the creeks to have a high portion of the values missing from the first half of the year range, and no missing values from the second half of the year range. Then for the other half of the creeks in the abundance category, we reversed the pattern such that there were no missing values in the first half of the year range and a high portion of missing values in the second half of the year range. The weighting scheme was observed as best as possible.

The changes in jackknifed sums-of-squares bias and chance error were substantial for infilled values from the Poisson and Gamma models. It is evident that balance affects the degree of chance error more in the Poisson model (almost 3×) than for the Gamma model (less than 2×; table 4.2) and bias is affected more in the Gamma model than the Poisson model (figure 4). If the design is balanced, then the amount of bias in the Gamma model becomes negligible and it is better model to use as the precision of the infilled methods is better due to chance error component being almost a third that of the Poisson model. However, if the design is unbalanced then the GLM method has amaller bias, and is a relatively better model to select.

Figure 4.1: Comparing bias in the infilled values from a GLM using the Gamma distribution when the design is balanced and unbalanced.



		$\hat{\theta}_{bias}$ *(st. error)*	$\hat{\theta}_{chance}$ *(st. error)*
Balanced	**Poisson Model**	0.00136 (0.013)	0.237 (0.022)
	Gamma Model	0.0009 (0.0011)	0.134 (0.018)
Unbalanced	**Poisson Model**	-0.0009 (0.0020)	0.623 (0.048)
	Gamma Model	0.0838 (0.013)	0.218 (0.023)

Table 4.2: Comparison of jackknifed sums-of-squares errors from a balanced and unbalanced sampling design

The sampling pattern for the Thompson River coho dataset had an added dimension of imbalance in the above simulation. As in some years, creeks were sampled intensively and in others not, such that there might be 89 creeks sampled in one year and 10 the next. Our simulations did not consider this added dimension of imbalance, but it is fair to say that this would have produced even more substantial errors in the infilled data estimates. This simple demonstrations highlights the serious need for better design criteria.

In the simulations discussed here, the balanced and unbalanced designs have the same efficiency. Improvements in the efficiency of sampling design could be made if a balanced incomplete block (BIBD) design could be selected. If E is the efficiency factor of an incomplete block design (IBD), and is measured by

$$E = \frac{\lambda t}{mn}$$

where:

λ is the number of times each pair of creeks appear in the same year and must be an integer,

t is the number of years,

n is the number of creeks measured in a given year

m is the number of years measured for a given creek; $m < t$

then for all incomplete block designs E is less than one. But in the class of designs of block size m, and t years (treatments), the most efficient design is the balanced incomplete block design if one exists (Kempthorne 1951). The efficiency of a BIBD (John 1971) is:

$$E = \frac{\lambda t}{mn} = \frac{t(m-1)}{m(t-1)}$$

where efficiency only depends on t and m. For a given t, efficiency is an increasing fuction of m and therefore the number of measurements per creek (m) should be as large as is reasonable (John 1971). Values of E may be useful in deciding on the best model.

Connectivity

Related to the notion of balance is the concept of connectivity. The analysis of the North Coast sockeye dataset (Section 3.3) was an example of how lack of connectivity combined with inconsistent trends can cause the best methods to fail. For example, if creeks A & B are censused in year 1, and creeks B & C are censused in year 2, then creeks A & C are connected. The relationship A connected to C is an equivalence relation which forms disjoint equivalence classes for the treatments. A design is said to be connected if there is one equivalence class (i.e., if every pair of treatments is connected). Problems arise when creeks A & B are censused in year 1 and creeks C & D in year 2, then creeks A & C, A & D, B & C, and B & D are disjoint and the design is then disconnected. This means that creek effects are confounded with year effects, such that time trends will be mixed up with either more or less productive set of creeks. This implies that the analysis of variance methods will breakdown without connectivity and balance. Therefore, the ultimate goal of survey design should be to maintain conectivity and balance.

Rotating Panel Surveys

Incorporation of the 3-year life cycle of this species into a rotation design would improve infilled estimates by reducing within creek variability. The variance can often be reduced by using the same sampling units (creeks) in the two successive salmon generations (Kish 1965). The variance tends to be least when the overlap is about one third, with the largest reduction comes from using the same elements, whose correlations are the highest (Kish 1965).

For those Thompson River creeks, in which there was enough observed data to do a time series analysis, a lag-3 serial correlation for 11 of the 13 creeks was demonstrated. If this correlation had been incorporated into the survey design in the form of some kind of rotation panel design, then significant improvements could be made in the infilled estimates due to reduced variability. A rotation panel that considers the correlation structure between succesive generations of salmon is recommended for monitoring populations and ensuring proper conservation methods can be implemented if required.

Chapter 5

Summary of Conclusions

This thesis presents seven imputation methods for infilling missing data into spawning salmon records and examines their performance in three patterns of missing data. For a reasonable amount of randomly missing data, we found the zero-infilled, nearest value, and interpolation/extrapolation methods to be inferior methods compared to the averaging of scaled values, transformed linear Poisson and gamma models to perform better. In particular, our simulation study shows the Poisson model which assumes the variance proportional to the mean to be the most reliable in this context.

When the missing data pattern is extreme and trends are not consistent across all creeks, problems arise for those methods that model creek and year effects but ignore interactions. In these cases, inconsistent trends across creeks correspond to interactions between year and creeks and because these interactions cannot be modelled, imputations are untrustworthy. In this context of imbalance and inconsistent trends, the nearest value method is the most reliable.

Balance, connectivity and consideration of life cycle ecology are key components in designing an area-wide survey. In the final chapter, we show that the degree of balance is critical in reducing bias and chance error and the reduction appears not to be uniform across all infilling methods. If balance in incorporated into the design, the

gamma model outperforms the Poisson model. In light of this, further investigation into modeling using the Poisson Inverse Gaussian distribution, which is a mixed distribution that is more flexible in its capacity to model the variance component, may prove instructive.

Chapter 6

Bibliography

(1) ARVESON, W. (1969). Jackknifing U-statistics. *Annals of Math. Stat.* 40: 2076-2100

(2) BRADFORD, M.J., J. IRVINE. (2000). Land use, fishing, climate change, and the decline of Thompson River, British Columbia, coho salmon. *Can. J. Fish. Aquat. Sci.* 57: 13-16.

(3) COSEWIC. (2000) Canadian Species at Risk, May 2002. *Committee on the Status of Endangered Wildlife in Canada.*

(4) DEAN, C. and LAWLESS, J.F., and WILLMOT, G.E. (1989). A mixed Poisson-inverse Gaussian regression model. *Can. J. Statist.* 17: 171-181.

(5) JOHN, P. W.M. (1971). Statistical Design and Analysis of Experiments. *The Macmillan Company, New York.*

(6) KEMPTHORNE, 0. (1956). The efficiency factor of an Incomplete Block Design. *Ann. Math. Statist.* 27: 846-849. -4

(7) KISH, L. (1965). Survey Sampling. *John Wiley B Sons, Inc., New York.*

(8) NICKELSON, T. E., J. NICHOLAS, H. WEEKS, and K. KOSTOW. (1994). Oregon coho salmon biological status assessment. *Oreg. Dep. Fish Wildl., Corvallis, OR.*

(9) RENSHAW, E. (1991). Modelling biological populations in space and time. *Cambridge University Press.*

(10) ROUTLEDGE, R.D., and IRVINE, J.R. (1999). Chance fluctuations and the survival of small salmon stocks. *Can. J. Fish. Aquat. Sci.* 56: 1512-1519. -

(11) SANDERCOCK, F.K. (1991). Life history of coho salmon (*Oncorhynchus kisutch*). p.357-455. in *C. Groot and L. Margolis, editors. Pacific salmon life histories. University of British Columbia Press, Vancouver, B. C.*