Modeling the Survival of Chinook Salmon Smolts Outmigrating Through the Lower Sacramento River System

Ken B. Newman, Division of Statistics, University of Idaho

John Rice, Department of Statistics, University of California, Berkeley

October 7, 1998

Abstract

A quasi-likelihood model with a ridge parameter was developed to understand the factors possibly associated with the survival of juvenile chinook salmon smolts outmigrating through the lower portions of the Sacramento river system. Coded-wire-tagged (CWT) chinook salmon smolts were released at various locations within the river between the years 1979 and 1995. Recoveries of these juvenile salmon in a lower river trawl fishery and later recoveries of adults from samples of ocean catches provided the basic data. Due to the number of interested parties with differing *a priori* opinions as to which factors most affected survival, a large number of covariates were required relative to the number of cases. To stabilize the parameter estimates and to improve predictive ability, a ridge parameter was included. Given the complexity of the processes generating recoveries, including possible dependencies between fish, and the additional sources of variation experienced by ocean recoveries relative to river recoveries, separate dispersion parameters were applied to the river and ocean recoveries. Interpretation of estimated coefficients was delicate given correlation between some of the covariates, the biases introduced by the ridge parameter, and possible confounding factors. With these caveats in mind, we found the most influential covariates to be the water temperature covariates and a measure of regional pesticide application, with increasing temperatures and increasing pesticide levels having a negative association with recoveries. Two covariates were of particular interest to the biologists and water managers: the position of a water diversion gate (open or closed) separating the mainstem from the central delta and the relative fraction of water exported for irrigation and urban consumption. Of these two, only gate position suggested a strong effect. When the gate was open, fish released upstream of the gate suffered increased mortality but survival increased for fish released in the central delta region (on the other side of the gate). Over the range of export levels observed, there was no strong evidence for either adverse or beneficial effects of increasing water exports.

KEYWORDS: overdispersion, release-recovery, ridge regression.

1 Introduction

The Sacramento-San Joaquin River system is the southern limit for chinook salmon (*Oncorhynchus tshawytscha*) and until the early portion of this century supported returns of chinook salmon numbering a million or more (Healey 1991;Clark 1929). Since then the number of adult salmon returning to spawn has decreased dramatically; during the mid-1970s returns to the San Joaquin river averaged less than 4,000 (Kjelson, Raquel, and Fisher 1982). There are a variety of reasons for the decline, including freshwater habitat loss and degradation, increased ocean fishing, and the export of water for human use. Water export from the lower portion of the river system, including the delta, is facilitated by water pumping stations, diversion gates, and hundreds of man-made canals. The diversion and exporting of water has drastically altered historical outmigration routes and lowered the chance of juvenile salmon successfully reaching the estuary and the ocean.

To identify water management schemes that will have less adverse effects on juvenile salmon survival, the U.S. Fish and Wildlife Service has conducted numerous release-recovery studies in the lower portions of the Sacramento-San Joaquin River system for the past twenty-plus years. Juvenile chinook salmon, raised and tagged at a hatchery, are released at various locations throughout the system, under varying water conditions (e.g., a major diversion gate is open or closed, flows are low or high, export levels are low or high), and recoveries are made downstream of the release point, usually by a mid-water trawl.

The release-recovery data from these studies have been the basis for several statistical models for survival through the delta developed by the USFWS, California Department of Water Resources, and other interested agencies. A recent approach (Kjelson, Greene, and Brandes 1989) was based on releases during the years 1978-1989 made at various locations near and downstream of Sacramento with subsequent recoveries by a trawl operating near Chipps Island (Figure 1). Kjelson, et al. fit separate multiple regression models for reach-specific mortality through three geographic areas of the river system. The relevant geography can be roughly characterized as a line from Sacramento to Courtland (reach 1), where the line branches into two segments, one arcing through the 'central delta' (reach 2) and the other staying in the main river (reach 3), and the two segments then come back together just above Chipps Island. Where the line branches into two segments there is a removeable diversion called the cross-channel gate which, when open, diverts water from reach 1 into reach 3. The three dependent variables, reach specific mortality indices scaled to [0,1], could not be directly observed and were estimated using a somewhat involved procedure. The details of the estimation procedure are not discussed here, but the accuracy of the estimates hinged upon a critical assumption that where reach 1 branches into reaches 2 and 3, the percentage of fish travelling through reach 2 equalled the estimated percentage of water entering that reach, i.e., fish 'go with the flow'. The indices estimation procedure also required that paired releases be available for various endpoints of the reaches (with some *ad hoc* methods to deal with a lack of pairs for reach 1 estimation in particular). Covariates examined for inclusion into the model included measures of river flow, amount of water extracted from the river by pumping stations in the delta (delta exports), water temperatures, fish size at time of release, and two different tide-related variables. Stepwise multiple regressions were used in each reach. Kjelson, et al. concluded that increases in water temperature, fraction of water diverted from the Sacramento River, and total exports adversely affected juvenile chinook salmon survival. They also recommended that smolts be kept out of the central delta (reach 2) because the greatest mortality (based on the estimated mortality index) was observed there.

The Kjelson, et al. (1989) work was closely scrutinized by numerous interested parties and their methodology was criticized on a number of grounds. The assumptions and methods for estimating the indices, the application of standard linear regression to dependent variables ranging between 0 and 1, and selection of covariates were major criticisms. In light of these criticisms, the interested parties chose to bring in statisticians previously unaffiliated with this work, namely the authors, to attempt to develop an alternative approach for modeling the release-recovery data. This paper describes our resulting model and the process leading to its formulation. Although our approach was quite different from that of Kjelson, et al., the essence of some of our conclusions is quite similar, for example, the effect of water temperature.

The time it took to decide upon the approach to take was perhaps longer than the actual model fitting, diagnostics, and interpretation. The process leading to model formulation is a key component of much applied statistics that is often neglected in the eventual presentation (c.f., "The Zeroth Problem" (Mallows 1998)). Thus, and especially in light of the difficulty of this particular problem, we devote Section 2 to the process we went through. Next, sections 3 and 4 describe our methods and results followed by some discussion and concluding remarks.

2 Formulating an approach

To three problems identified by R.A. Fisher that arise in the course of reduction of data (1. specifying the mathematical form of the population giving rise to the data; 2. estimating parameters; 3. determining sampling distributions for parameter estimates), Mallows (1998) adds a 'zeroth' problem: "considering the relevance of the observed data, and other data that might be observed, to the substantive problem".

Before addressing the 'zeroth' problem, an earlier problem, that Mallows makes passing comment about, was also present— choosing what problem to study. This required numerous discussions with the fisheries biologists to clarify what it was they hoped to be able to learn and to do. The answers were to identify which variables were most closely associated with salmon survival, including both the somewhat non-manipulable variables like fish size, mainstream river flow rate, and water temperature as well as the manipulable variables like position, open or closed, of a main diversion gate into the central delta and water export volume. Besides identifying the more important factors, the biologists wanted a quantitative tool developed that could be used to assist in making water management decisions.

There were several parts to our zeroth problem and solutions were gradually arrived at over the course of several meetings of fisheries biologists, water resource experts, and statisticians. The group included representatives of institutions that had often been somewhat at odds—for example USFWS and consultants from urban water districts. Newman was a consultant to USFWS and Rice was a consultant to the California Urban Water Agencies. The meetings educated us about the data collection and biological processes.

An initial problem in determining the relevant data was selecting the release groups to use and whether to combine some release groups. We wanted to use as much data as possible and elected to use releases from 1979 through 1995, a large number of which were unpaired releases (in contrast to Kjelsen, et al.'s data). With the help of knowledgeable biologists several groups were combined, differing only in what were considered minor ways (perhaps reared in different ponds at the hatchery,) but otherwise identical in age, size at release, location of release, etc. A total of 86 release group(ing)s resulted.

A second issue was choice of dependent variables. It seemed clear from the beginning that a primary dependent variable should be the number of tags recovered by the midwater trawl at Chipps Island, but a number of questions were raised regarding the nature of the recovery effort. The trawl makes ten 20-min tows during the daylight during the days in which the salmon are passing by (Kjelson and Brandes 1989). The width of the river at this location is around 1200 m, the trawl mouth opening is 9.1 by 7.9 m, the trawl fishes to a depth of 8 to 15 m, and a roughly equal number of sweeps are made in the south, middle, and north portions of the river. The trawl is thus covering a relatively small portion of the river spatially and temporally and the capture probabilities are low. For example, two replicate groups of 53,430 and 51,086 fish were released near Sacramento in 1989 and only 21 and 34 fish were recovered by the trawl. Given that 39 and 44 fish were later recovered at age three in the ocean fishery, these low numbers are likely more a function of low capture efficiency than terrible freshwater suvival. Besides typically low recovery rates, concerns were voiced over where in the water column fish may be passing through, whether or not fish might be selecting one side of the river over another, changes in gear effectiveness as turbidity changed, how differences in fish size may be affecting gear effectiveness, how much of the outmigration is missed by not sampling at night, and how many fish may be missed by failing to start sampling soon enough. Results of these discussions led to decisions to include measures of trawl effort, fish size, and turbidity in an attempt to partially control for variation in the capture probabilities between release groups.

Two other potential dependent variables were tag recoveries of adult fish in ocean fisheries and from those returning to the hatchery (or spawning in natural areas, so-called strays).

Primary interest was in factors affecting the freshwater survival, but differences in ocean recovery rates could be attributed to factors other than freshwater survival. Chinook salmon can stay in the ocean for three to four years and be caught at age two, three, four, or (occasionally) five. Tagged fish are subsequently recovered in samples of the catch taken at the ports of landing. Thus the number of recoveries of tagged fish in the ocean fishery depends not only on freshwater survival, but also on marine survival rates, the spatial distribution of and level of fishing effort in the ocean, ocean migration patterns, maturation schedules (a probability vector for an adult chinook salmon returning to spawn at age two, age three, etc.), catch sampling levels, and year to year variation in ocean conditions.

To help determine whether or not the ocean recoveries could prove useful for assessing freshwater survival, we examined the spatial-temporal distributions of the *estimated* ocean recoveries for fish. (The process by which this estimation was done will be described below.) We knew *a priori* that variation between years in fishing patterns and ocean conditions was great enough that indicator variables for release year effect, at least, would be necessary. Our hope, however, was that releases made within the same year would have similar ocean migration patterns and maturation schedules; if not, such differences would confound differences in freshwater survival. We examined estimated recoveries, which are expansions of the numbers actually observed in catch samples, rather than observed recoveries because catch sampling fractions can vary considerably between ports and time of season: thus using observed recoveries can be misleading. To study the degree of similarity in ocean migration patterns and maturation schedules, we conducted a cluster analysis of all release groups simultaneously using the relative number of recoveries at ages two, three, and four by time and location of recovery as the variables. The relative numbers, number per time-location stratum divided by total recoveries, were used instead of absolute numbers to partially control for differences in freshwater survival rates (as well as number released). On the basis of much greater similarities within release years than between release years, we collectively agreed to assume similar marine survival, ocean migration patterns, and maturation schedules for fish released the same year and to use the ocean recoveries as a dependent variable in addition to recoveries by the Chipps Island trawl.

We decided against using freshwater recoveries of returning adult salmon, however. Some fishery biologists argued that the further a fish is released downstream of its natal area (the hatchery), the more its freshwater homing ability is adversely affected. Thus straying rate, or the probability of an adult fish not returning to the hatchery, increases as the distance between the hatchery and the release location increases. This would lead to fewer recoveries for the downstream releases than upstream releases even if survival rates were identical. The sampling of natural spawning areas is much more sporadic than that of hatcheries and the methods employed are non-systematic and largely undocumented. In particular, the level of sampling "effort" could not be quantified and the potential for subsequent bias too large.

An issue intermediate between the selection of release groups and choice of dependent variables was whether to model absolute survival rates through the delta or relative survival rates—survival up to an unknown constant of proportionality. Given the relative survival rates for two release groups, the ratio of the survival rates can be estimated. Absolute survival rate is of course more informative, but is difficult to estimate because the probability of recovery is a product of the absolute survival and capture rates.

One approach to disentangling absolute survival rate to Chipps Island from capture rate would be to optimistically assume a deterministic relationship between effort and capture rate, with the "residual" being survival. Measurement of effort, however, is inexact and even with fixed effort levels, changes in water conditions certainly affect capture rate. Another approach to this non-identifiability problem is to use ocean recoveries of releases that were made near Chipps Island. Given upstream releases that could be paired with these releases, both absolute survival to and capture rate at Chipps Island can be separately estimated, assuming an equal ocean recovery rate for upstream and downstream releases. Another necessary assumption is that the subsequent survival rate for Chipps Island releases equals the rate for the upstream releases surviving to Chipps Island and eluding the trawl. The non-survivors from the upstream releases have been culled, however, and the survivors may in fact be a hardier subgroup than the downstream releases. Finally, restricting attention to only those upstream releases that can be paired with the downstream releases greatly reduces the data set. Although future work may be directed towards this approach, the parties involved agreed that modeling relative survival to Chipps Island would be sufficient for now.

A fourth issue was choosing amongst the more than 30 variables that could possibly influence recovery rates. Many were viewed as redundant by biologists, e.g., three salinity measurement made at different locations, and these opinions were quickly checked by simple statistical analyses, such as scatterplots. Strong differences of opinion remained, however, within the group as to which of the remaining non-redundant variables needed to be in the model. We (the statisticians) recommended against relying upon standard variable selection procedures, such as backward elimination, because of potential instability in the parameter estimates (given the large number of covariates relative to number of cases). Another reason was that it seemed quite possible that several different covariate combinations could lead to comparable fits. Anoiting one particular combination arrived at by a somewhat arbitrary procedure seemed a dangerous practice to us. Furthermore, it would be quite difficult to arrive at plausible measures of uncertainty after extensive variable selection. Given a lack of consensus on which of these remaining variables should be used as covariates in the model, we decided to use all of them and rely upon a statistical procedure to dampen the variability of the parameter estimates (using a ridge parameter to shrink the estimates). A couple covariates were later dropped or modified on the basis of initial model fitting results, but the original selection remained largely intact.

Finally there was the problem of how to deal with the spatial geometry of the release sites. The approach of Kjelsen, et al. (1989), modeling reach specific mortalities, was initially considered but rejected. The primary drawbacks were the difficulty in dealing with releases upstream of the branch to the central delta and the need for paired releases. Using only paired releases greatly decreased the amount of available data. Furthermore, many of the biologists were unwilling to accept the *go with the flow* assumption. The spatial geometry is much more complicated than the three reach structure used by Kjelsen, et al. (1989) and to use several of the releases (e.g., those from Sutter and Steamboat Sloughs (Figure 1)) would require adding more reaches and further controversial assumptions. Finally crucial, but dubious assumptions of conditional independence are needed to paste together survival estimates from various reaches. The next thought was to develop several models, on a per release site basis with some sharing of covariates amongst different release sites. This led to what became a final formulation of a single model, using indicator variables for separate release sites (or groups of release sites) and allowing for potential interactions between release site and some of the covariates. For example, the effect of the diversion gate position should be different for salmon released upstream of the gate than for those released downstream.

The eventual solution to our zeroth problem was arrived at in a somewhat iterative fashion, some probability model selection and model fitting was done concurrently, but the essence of the solution was arrived at before a single parameter was estimated. This was desirable for a couple of reasons. First, we thought it important for all parties to the discussion to agree in principle to an approach before seeing what it actually produced. Second, the more highly interactive and iterative the model building procedure, the more difficult it would be to assess statistical variability.

3 Methods

The model for the recoveries of tagged fish at Chipps Island and in the samples of ocean fishery catches can be broadly characterized as a generalized linear model with an underlying overdispersed Poisson distribution for recoveries, where ridge regression is used to stabilize the variances of the estimated coefficients. After introducing some notation and key assumptions, details of the model structure and fitting are given.

3.1 Notation

S, r, and f represent survival rate, recapture probability conditional on survival, and recovery effort level, respectively. The product Sr is referred to as a recovery rate. Subscripts C or O refer to Chipps Island or the ocean.

S_C = probability an upstream fish survives to Chipps Island

r_C = capture probability at Chipps Island, assuming that the fish is alive at that location

f_C = proportion of space-time sampled at Chipps Island

S_O = probability that a fish passing Chipps Island safely then survives and is vulnerable to the ocean fisheries

r_O = probability that a fish vulnerable to the ocean fisheries is caught

f_O = average expansion factor for ocean recoveries over all catches sampled over the three year periods during which each cohort is being recovered; equals #estimated/#observed and is roughly the average fraction of each catch being sampled

π_{CO} = probability that an upstream fish surviving to Chipps Island is later caught in the ocean and recovered in a catch sample (roughly $\equiv S_O r_O / f_O$)

R = number of fish released

Y_C = number of fish recaptured at Chipps Island

Y_O = (estimated) number of fish recaptured in ocean fishery (at ages two, three, or four)

x_i = covariate i

3.2 Assumptions

The statistical model rests on certain assumptions:

A1 : For a given release group the expected number of recoveries at Chipps Island is proportional to the product of: (1) the number of smolts released, (2) the probability of a smolt surviving from the release point to Chipps Island, and (3) the reported effort at Chipps Island. The constant of proportionality is an unknown "catchability coefficient", q_C, which is assumed to be independent of the covariates and constant for all release groups. I.e., for a given release group of R fish,

$$\begin{aligned} \mathrm{E}[Y_C] &= R S_C r_C \\ &= R S_C f_C q_C \qquad (1) \end{aligned}$$

A2 : For such a release group, the expected estimated number of fish recaptured in the ocean fisheries at ages two, three, and four is equal to the product of: (1) the number released, (2) the probability of survival to Chipps Island, (3) the probability that a smolt surviving to Chipps Island is later caught in the ocean and recovered in a catch sample, and (4) an expansion factor that reflects catch sampling effort. It is assumed that (3) is constant for all groups released within the same year. Namely,

$$\begin{aligned} \mathrm{E}[Y_O] &= RS_C(1-r_C)\pi_{CO}f_O && (2) \\ &\approx RS_C\pi_{CO}f_O && (3) \end{aligned}$$

where within a given year π_{CO} is constant for all release groups. An implicit assumption in going from equation (2) to equation (3) is that r_C is so small that $1-r_C \approx 1$, in other words that only a small fraction of the fish surviving to Chipps Island are captured there.

A3 : The variance of the number recovered is proportional to the expected number recovered, or

$$\begin{aligned} \mathrm{Var}[Y_C] &= \phi_C \mathrm{E}[Y_C] && (4) \\ \mathrm{Var}[Y_O] &= \phi_O \mathrm{E}[Y_O] && (5) \end{aligned}$$

The parameters ϕ_C and ϕ_O are dispersion parameters (McCullagh and Nelder 1989). For fisheries data, dispersion parameters greater than 1.0, evidence of overdispersion, are common and can result from shoaling behavior and from fish in a release group having a common response to environmental factors (Cormack 1993).

A4 : Recoveries are related to covariates through a model in which the logarithm of the expected number of recoveries is a linear function of covariate values, with coefficients to be determined by fitting. In particular, release size R and the measures of recovery effort, f_C and f_O, are treated as known constants. From equations (1) and (3), using p covariates for Chipps Island recoveries and s covariates for ocean recoveries

$$\begin{aligned} \log(E[Y_C]) &= \log(Rf_C) + \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \\ \log(E[Y_O]) &= \log(Rf_O) + \beta_0' + \beta_1' x_1' + \beta_2' x_2' + \ldots + \beta_s' x_s' \end{aligned}$$

Observe from equations (1) and (3) that the logarithms of S_C and π_{CO} are being modeled as functions of covariates, while the logarithm of q_c, assumed constant, is included in the intercept β_0. Some of the covariates are shared by both recovery locations as are some of the coefficients and the relationships between β and β' are described later.

A5 : The coefficients of some covariates are assumed to be the same for all release sites and those of others are allowed to differ from release site to release site; i.e., there is a release site interaction with these covariates.

A6 : The relationships between the covariates and mortality have not changed during the period of time for which the model is fitted.

3.3 Covariates

Primary interest is in S_C, which is assumed to be a function of two types of covariates: (1) factors whose influences are independent of release site, and (2) release site dependent factors. The effects of the site dependent covariates, export levels, gate position, and turbidity, were allowed to vary between groups of sites; i.e., an interaction between release site and these three covariates was modeled. For example, the coefficient for export level could differ between fish released at Courtland and fish released at Mokelumne. In order to cut down on the number of coefficients to be estimated, we only allowed covariates to be site dependent when there was clear *a priori* reason to do so.

The covariates are categorized below, with variable name given in italics.

Site Independent (SI) :

1. *Size* = (estimated) average length of fish (mm)
2. *Log.Flow* = natural logarithm of the median flow (cfs) at a location (Freeport) on the Sacramento river, with median calculated over the period between release date and last day of recoveries at Chipps Island

3. *Salinity* = salinity (micro mho/cm) measured at a location (Collinsville) on the Sacramento River

4. *Pesticide* = annual amount (pounds) of applied rice pesticide

5. *Trend* = linear annual trend (last 2 digits of year of release)

6. *Release.Temp* = river temperature (degrees Fahrenheit) taken at time of release near the shore and at the surface

7. *Hatch.Temp* = mean temperature (degrees Fahrenheit) at the Feather River Hatchery on day of release

8. *Shock* = truck temperature - release temperature, where truck temperature is the water temperature in the truck carrying the smolts to the release site and is measured at time of release

9. *Tide.Var* = a measure of the magnitude of the change in low-low and high-low tides and whether the delta is generally filling or draining (feet)

Site Dependent (SD) : The export, gate position, and turbidity covariates listed below were allowed to have different effects depending upon the release location. The actual covariates used for these three classes of covariates were a cross-product of the release site indicator variables and the export, gate position, or turbidity values.

1. Release Location: Release sites were grouped into the following seven categories with corresponding indicator variables used for all but Jersey Point.

   - *FRH.ind* = Feather River Hatchery
   - *Sac.ind* = Discovery Park, Miller Park, and Sacramento
   - *Slo.ind* = Steamboat and Sutter Sloughs
   - *Crt.ind* = Courtland
   - *Ryd.ind* = Ryde, Rio Vista, and Isleton
   - *Mkg.ind* = Lower Mokelumne, North Fork Mokelumne, South Fork Mokelumne, and Georgiana Slough
   - Jersey Point (the default location, thus no variable)

2. Exports/Inflow Ratio (cfs/cfs): the ratio of amount of water exported to amount of water flowing in the mainstem

   - *Upper.Exp.Inflow* = export/inflow value crossed with indicators for releases from Feather River Hatchery, Discovery Park, Miller Park, Sacramento, and Courtland
   - *Delta.Exp.Inflow* = export/inflow value crossed with indicators for Lower Mokelumne, North Fork Mokelumne, South Fork Mokelumne, Georgianna Slough, and Jersey Point

3. (Cross-channel) Gate Position: coded as 1 for open and 0 for closed;

   - *Upper.Gate* = gate position value crossed with indicators for releases from Feather River Hatchery, Discovery Park, Miller Park, Sacramento, and Courtland
   - *Delta.Gate* = gate position value crossed with indicators for releases from Lower Mokelumne, North Fork Mokelumne, South Fork Mokelumne, Georgianna Slough, and Jersey Point

4. Turbidity (Formazine turbidity units)

   - *Mainstem.Turbid* = mainstem turbidity value used for releases from Feather River Hatchery, Discovery Park, Miller Park, Sacramento, Courtland, and Ryde (including Rio Vista and Isleton)
   - *Delta.Turbid* = central delta turbidity value used for Lower Mokelumne, North Fork Mokelumne, South Fork Mokelumne, Georgianna Slough, and Jersey Point

Note that releases from Steamboat Slough or Sutter Slough are assumed unaffected by the export/inflow ratio, cross-channel gate position, and mainstem (or central delta) turbidity. Releases at these locations are geographically removed from the primary water pumping locations. The fish are unlikely to travel upstream into the mainstem and thus be affected by the gate position. The sloughs empty back into the mainstem about midway between the release points on the sloughs and Chipps Island, so for survivors that do reach the mainstem, we are assuming the effect of variations in mainstem turbidity is relatively negligible.

Similarly, releases from Ryde are assumed unaffected by the export/inflow ratio and gate position, but given its location on the mainstem, mainstem turbidity could have some effect. Whatever effect there may be is assumed identical for all other mainstem releases.

Another covariate was an indicator variable for Chipps Island recoveries, the intercept could differ between Chipps Island recoveries and ocean recoveries. The effect of the catchability coefficient at Chipps Island, q_C, is partially modeled through the Chipps Island indicator variable.

Finally a set of indicator covariates was included for ocean recoveries representing the year of release (*1979.ind 1980.ind ... 1993.ind*) with 1994 the default year. There was no indicator for 1995 releases because those releases did not have ocean recoveries at the time the data set was created. The ocean survival-capture combination π_{CO} for each year is thus modeled, to some degree, by these indicators with the intercept shifting up or down between years.

3.4 Model structure

The model formulation for the expected Chipps Island and ocean recoveries follows from (1) and (3) and the above covariates (symbolically):

$$\begin{aligned}
\log(E[Y_C]) &= \log(Rf_C) + \log(q_C) + \log(S_C) \\
&= \log(Rf_C) + \beta_0 + \beta_1 \text{Chipps Island Indicator} + \beta_2 \text{SI} + \beta_3 \text{SD} \\
\log(E[Y_O]) &= \log(Rf_O) + \log(S_C) + \log(\pi_{CO}) \\
&= \log(Rf_O) + \beta_0 + \beta_2 \text{SI} + \beta_3 \text{SD} + \beta_4 \text{Year Dummies}
\end{aligned}$$

The Chipps Island and ocean recoveries share common coefficients for the site indicator and site dependent covariates because these fish have both survived to Chipps Island; i.e., this reflects the common S_C component in probability of recovery. The two recovery ‘areas’ are distinguished, however, by the use of an indicator variables for Chipps Island recovery, partially reflecting q_C, and the indicators for year of recovery, partially reflecting $\pi_{CO} f_O$.

The detailed model structure is shown in equation (6).

$$\begin{aligned}\log(E[Y]) \quad = \quad & \log(Rf) + \beta_0 + \beta_1 Chipps.ind + \beta_2 Size + \beta_3 Log.Flow \\ & + \beta_4 Salinity + \beta_5 Pesticide + \beta_6 Trend + \beta_7 Release.Temp \\ & + \beta_8 Hatchery.Temp + \beta_9 Shock + \beta_{10} Tide.Var \\ & + \beta_{11} FRH.ind + \beta_{12} Sac.ind + \beta_{13} Slo.ind + \beta_{14} Crt.ind \\ & + \beta_{15} Ryd.ind + \beta_{16} Mkg.ind + \beta_{17} Upper.Exp.Inflow + \beta_{18} Delta.Exp.Inflow \\ & + \beta_{19} Upper.Gate + \beta_{20} Delta.Gate + \beta_{21} Mainstem.Turbid \\ & + \beta_{22} Delta.Turbid + \beta_{23} 1979.ind + \ldots + \beta_{37} 1993.ind\end{aligned} \tag{6}$$

3.5 Parameter estimation

The final formulation of the model described by equation (6) required estimating $p = 38$ coefficients corresponding to the covariates and the two dispersion parameters, ϕ_C and ϕ_O. There were 86 observations for the Chipps Island recoveries, but only 84 observations for the ocean recoveries due to the lack of complete information for the 1995 releases. The observation vector thus was of length 170 and the design matrix had dimension 170 by 38. The design matrix (for the raw data) had an 86 by 15 submatrix of zeros corresponding to the release year indicator variable values for Chipps Island recoveries. The design matrix used for parameter estimation, however, was based on standardized covariates to minimize numerical errors as well as to make the estimated coefficients more directly comparable.

The coefficients, β_0, …, β_{37}, and the dispersion parameters were estimated using iterated weighted least squares (IWLS) with a ridge regression parameter, λ. Ignoring the ridge aspect momentarily, the estimated coefficients correspond to those arising from maximizing a quasi-likelihood. In our formulation if both dispersion parameters had been fixed at 1.0, then the equality of the mean and the variance (equations (4) and (5)) would imply a Poisson distribution. The resulting estimating equations were

$$(X^t W X + \Lambda)\beta \quad = \quad X^t W Z$$

where X is the design matrix of (standardized) covariates (n by p), W is a diagonal matrix of weights for each observation, Λ is a diagonal matrix of the ridge parameter values, and

Z is a vector of 'transformed' observations. The components of Z were decremented by an offset for the release number R and the recovery effort measures (f_C or f_O); i.e., for the ith Chipps Island and jth ocean recoveries

$$\begin{aligned} z_{C,i} &= \log(\hat{\mu}_{C,i}) + \frac{Y_{C,i} - \hat{\mu}_{C,i}}{\hat{\mu}_{C,i}} - \log(R_i) - \log(f_{C,i}) \\ z_{O,j} &= \log(\hat{\mu}_{C,j}) + \frac{Y_{O,j} - \hat{\mu}_{C,j}}{\hat{\mu}_{C,j}} - \log(R_j) - \log(f_{O,j}) \end{aligned}$$

where the fitted values, $\hat{\mu}$, were estimated from (6) (using standardized covariates), substituting current estimates, $\hat{\beta}$, for β.

The elements on the diagonal of the weight matrix were functions of the estimated variance for the corresponding observation, equations (4) and (5) for Chipps Island and ocean recoveries. The dispersion weights, ϕ_C and ϕ_O, were estimated iteratively based on squared residuals,

$$\begin{aligned} \hat{\phi}_C &= \sum_{i=1}^{86} \frac{(y_{C,i} - \hat{y}_{C,i})^2}{\hat{y}_{C,i}} / (86 - p) \\ \hat{\phi}_O &= \sum_{i=87}^{170} \frac{(y_{O,i} - \hat{y}_{O,i})^2}{\hat{y}_{O,i}} / (84 - p) \end{aligned}$$

where p is the number of coefficients (p=38). The weighting influence of the dispersion parameters was adjusted relative to the Chipps Island recoveries; for the ith Chipps Island observation and the jth ocean observation

$$\begin{aligned} w_{i,C} &= \hat{\mu}_{i,C} / \hat{\phi}_C \\ w_{j,O} &= \hat{\mu}_{j,O} / \hat{\phi}_O \end{aligned}$$

The reason for the ridge parameter is the large number of coefficients to estimate, 38, relative to the number of observations, 170, and the consequent potential for large variances for the estimated coefficients. A ridge parameter was not used for the intercept and the Chipps Island indicator variable, the terms distinguishing the intercepts for the ocean and for the Chipps Island subsets, since the overall averages should not be shrunk toward zero. The ridge parameter was selected using a combination of measures, but primarily the mean square of cross-validated prediction errors, MSPE. The MSPE was calculated using a leave-one-out approach, for each λ considered,

1. Leave out observation i and estimate the coefficients $\beta_{-i,\lambda}$.

2. Predict the number of (Chipps Island or Ocean) recoveries for observation i, denoted $\hat{y}_{i,\lambda}$, using $\hat{\beta}_{-i,\lambda}$ and the covariate values for observation i.

3. Calculate a squared Pearson residual for each observation, $e^2_{i,\lambda} = (y_i - \hat{y}_{i,\lambda})^2/\hat{y}_{i,\lambda}$.

4. Calculate the mean score for the particular λ, namely, $\frac{1}{170}\sum_{i=1}^{170} e^2_{i,\lambda}$. (Trimmed means were examined as well.)

Standard errors for the estimated coefficients were estimated using an analytical approximation. The approximation ignores possible time dependence between observations, for example, the correlation of environmental conditions in adjacent years, and as such may yield slight underestimates. The standard errors also ignore the data-based choice of the ridge parameter λ. Details are given in Appendix A.

Ridge estimators have been applied to generalized linear models, particularly logistic models (Schaefer, Roi, and Wolfe 1984; Shaefer 1986; Lee and Silvapulle 1988; Duffy and Santner 1989; le Cessie and van Houwelingen 1992) and at least once to a Poisson-log link model (Segerstedt 1992). The authors are unaware of any applications to quasi-likelihood models, however, especially in a situation with differing dispersion parameters.

4 Results

4.1 Choice of ridge parameter

The ridge parameter was set equal to 40 on the basis primarily of the cross-validation scores for prediction errors and secondarily on the variances for the $\hat{\beta}$'s and ridge traces. Figure 4 shows the cross-validation scores over ridge parameter values ranging from 0 to 60. There are a few extremely large prediction errors for the ocean observations that greatly influence the mean score. Emphasis was placed on prediction errors for the Chipps Island observations (plot (b)) which were minimized at λ=35. Ridge traces (plots (d)-(f)), the estimated coefficients versus the size of the ridge parameter, were examined as well. The changes in coefficients were relatively minor for $\lambda > 40$. The sum of the estimated variances for the $\hat{\beta}$ for the same range of λ are listed in Table 1. The relative decrease in total variance from λ=0 to λ=40 was 78%; the decrease in total variance is only 7% as λ increases from 40 to 45.

4.2 Estimates of β's and ϕ's

Estimates of the coefficients, β_0, ..., β_{37}, and their standard errors are given in Table 2. The estimated dispersion parameters were $\hat{\phi}_C$=15.35 and $\hat{\phi}_O$=85.82. Recall that the Site Dependent coefficients corresponding to the export, gate position, and turbidity covariates are release site specific. For example, the site dependent effects for a Feather River Hatchery release are reflected by *FRH.ind*, *Upper.Exp.Inflow*, *Upper.Gate*, and *Mainstem.Turbid*.

The estimated coefficients for the site independent and site dependent covariates are plotted in descending order in Figure 2 along with ± 2 standard errors. Ignoring bias in the estimates, this provides an approximate means of visually separating strong from weak effects in that coefficients with these intervals including zero would be considered weak. Given that the bias in positive estimates is probably negative and bias for negative estimates probably positive (with the ridge parameter shrinking estimates towards zero), these intervals are likely shifted more to the origin than correct 95% confidence intervals.

Table 2 also includes the non-ridge estimates of the coefficients and estimated standard errors. For the site independent covariates, the inclusion of the ridge parameter diminished the influence of all but *Trend* and *Shock*, which marginally increased in absolute size. With the exception of releases from Feather River Hatchery and the Mokelumne-Georgianna region, the between site distinctions were all diminished by the ridge parameter as were the site specific export, gate, and turbidity effects. The reduction in (estimated) standard error, a primary objective for including the ridge parameter, was sizeable for estimated coefficients—an average 32% reduction for site independent coefficients and average 48% reduction for site dependent coefficients.

4.3 Residual analyses

The fitted values were plotted against the observed values (Figure 3) for all 170 observations, for the 86 Chipps Island observations alone, and for the 84 ocean recoveries alone. Variation in observed values increases as the fitted or expected value increases, consistent with the assumption that variances are proportional to means. The need for different dispersion parameters for Chipps Island and ocean recoveries, and the explanation for the large difference between $\hat{\phi}_C$ and $\hat{\phi}_O$, can be seen when the Pearson residuals, $(y - \hat{y})/\sqrt{\hat{y}}$, are plotted by observation order (plot (d)) with the first 86 values corresponding to the Chipps Island observations.

Additional residual analyses focused on Chipps Island recoveries. Plots of these residuals against the Site Independent and Site Dependent covariates (not shown) revealed a possible non-linearity with *Shock*, in particular a threshold effect, in that for small shocks the the model is overestimating survival. Namely, for shocks less than 7 degrees ther are no positive residuals. Otherwise the balance between positive and negative residuals was good and no remaining associations were evident.

4.4 Individual case influence

The influence of individual observations on the estimated coefficients was examined using a variation on Cook's distance measure (Weisberg 1985). The measure for obervation i was calculated by

$$D[i] \quad = \quad \frac{1}{p}\sum_{j=1}^{38}(\hat{\beta}_{\lambda,j,[-i]} - \hat{\beta}_{\lambda,j})^2.$$

The maximum value of D is for a 1986 ocean recovery observation for a Ryde area release, is 0.00022, indicating a very small change in estimates of $\hat{\beta}$ when this observation is deleted.

4.5 Effect of ocean observations

The inclusion of the ocean recoveries into the analysis of the effects of the various covariates was a departure from previous work and the consequent effect on the resulting estimates was of particular interest. It is informative to examine the consistency of the two sources of data. There are unknown biases in both data sets and without basic consistency in results, quantitative modeling becomes problematic.

Note that the weighting given to observations from each data set differed due to differences in number of recoveries and overdispersion parameters. With a Poisson-log link model, increases in the observed numbers translate into increase weighting of the observations in the design matrix, X^tWX. For the quasi-likelihood model the weights are scaled by the dispersion; i.e., the ith case is given weight μ/ϕ. Thus the much larger number of (expanded) recoveries for the ocean data was diminished by the much larger dispersion parameter, but the average weighting for the ocean cases was 5.0 compared to 3.0 for the Chipps Island cases.

We evaluated the effect of the ocean data three different ways:

1. The difference in estimated β's for Chipps Island data alone and for Ocean data alone.

2. The change in estimated β's when Ocean data is added to the Chipps Island data.

3. The ability of the Ocean data-based model to estimate the Chipps Island recoveries (up a constant of proportionality).

Figure 5 contrasts the estimated coefficients for the Site Independent and Site Dependent covariates based on using either the Chipps Island or Ocean data alone. To fit the models the Chipps Island indicator variable was dropped from both data sets and the year indicators were dropped from the Chipps Island data set. The numbering for the coefficients is the sorted order, from smallest (most negative) to largest, of the Chipps Island estimates. The intersecting vertical and horizontal lines are drawn between the point estimates ± one standard error. The greater the distance from the 45 degree line, the greater the difference in estimated values. Points in the upper left and lower right quadrants indicate coefficients that changed sign between the two data sets. The greatest difference is for the *Salinity* coefficient— this is confounded somewhat by the change in the *Log.Flow* coefficient; these two variables are the most strongly correlated in the data set. Overall the estimated coefficients are relatively consistent between the two data sets. Many of the coefficients based on the Ocean only data are exaggerations of the coefficients based on Chipps Island data only— negative coefficients become even more negative, and positive coefficients become more positive. At the same time the standard errors for the ocean-only coefficients remain larger, partly a reflection of the larger dispersion parameter.

Adding the Ocean data to Chipps Island data has little effect on the estimated coefficients with relatively large values (Figure 6). Some of the smaller estimated coefficients such as for *Size*, *Tide.Var*, and *Delta.Turbid*, change in sign. The decrease in standard errors is minor, however; there is an average relative decrease of slightly less than 5%.

The degree of agreement between models based on Ocean data alone or Chipps Island data alone in terms of fitted Chipps Island recoveries can be seen in Figure 7. The fit from the Ocean data-based model should not be as good, of course, and because of different intercepts (affected by different magnitudes of offsets), the fitted values will be at best proportional. Even with the coefficients differing between the two models, however, the fitted values can be similar because of tradeoffs in coefficient values. With a handful of notable exceptions the fitted values do look roughly proportional— the straight line drawn across the plot is the least squares line. Combining the two data sets for parameter estimation seems appropriate.

5 Discussion

As discussed in Section 2 the two primary objectives of the biologists were to identify the factors with the apparently greatest impact on fish survival and to develop a quantitative tool for assisting them in making decisions about water management.

5.1 Interpretation of covariates' effect

Clearcut statements as to the covariates' effect on survival through the delta as measured by the estimated coefficients are difficult for several reasons. First, the nature of an observational study makes assertions as to causality problematic, although the inclusion of so many covariates was partially an attempt to account for factors, that if omitted, would confound the interpretation of the estimated coefficients. For example, if release year effects were not accounted for, site location effects could definitely be confounded with year effects given the lack of balance in the data set. For example releases were not made from Ryde for the years 1979 through 1982. If an interaction between year and site does not exist, then the inclusion of both terms partially alleviates the problem of confounding. The tentative interpretations made below are dependent upon assuming that most of the potentially confounding factors have been included in the model and no important interactions have been omitted.

In addition to the limitations of an observational study, correlation, both real and that due to the configuration of the given data set, complicates the interpretation. The most extreme example is the real correlation that exists between flow and salinity, as flow increases salinity decreases. With flows about 14,000 cfs, the salinity is nearly zero. Both the estimated coefficients are large, but a relatively large flow is always paired with a low salinity and the large 'positive' effect of flow is offset by the relatively large 'negative' effect of a low salinity that is negative in standardized units. The end result is that for flows 14,000 cfs or under, the effects of flow and salinity combined are negligible. Other covariates, besides flow and salinity, that are known to have real correlation include flow and turbidity, and hatchery and release temperatures. With these exceptions the degree of multicollinearity in this data set is low.

Dataset imbalances that complicate the interpretation include the relationship between flow and gate position. At flows at or about 20,000 cfs the gate was always closed—hence the apparent positive effect of these high flows (salinity is not offsetting flow at this cfs) for mainstem releases could be confounded with a positive effect of the gate necessarily being closed (recall that the negatively valued *Upper.Gate* coefficient only enters into the estimation when the gate is *open*).

Besides looking at pairwise scatterplots and the correlation of estimated coefficients, the sensitivity of the Site Independent estimates to the inclusion of other Site Independent covariates was examined. Table 3 shows the change in estimated coefficients when one of the Site Independent covariates is left out of the model. The flow-salinity trade-off is clear; dropping salinity in particular yields a sizeable decrease in the flow coefficient, thus requiring relatively high flows to yield a noticeable effect. Dropping $Release.Temp$ increases the negative magnitude of the other temperature coefficients. Finally, the deletion of the $Pesticide$ covariate leads to an decrease in the slope for $Trend$.

A fourth restriction on interpretation of the estimated coefficients is the bias in the ridge estimates likely shrunk towards zero.

With these caveats in mind we will draw a few tentative conclusions about the Site Independent and Site Dependent covariates.

Site Independent covariates

As alluded to above high flows, above 14,000 cfs, are positively associated with improved survival. Increases in flow yielding increases in survival are consistent with conventional wisdom. As flow increases from 2,000, say, to 14,000 cfs, however, the corresponding decrease in salinity seems to offset the positive effect of greater flow.

As release temperature and hatchery temperature increase over the ranges observed, fish survival decreases. The two temperatures are highly correlated since hatchery water and river water usually come from the same source; so it is simplest to say that as the river water warms, mortality increases. This is qualitatively in agreement with the findings of Baker, Speed, and Ligon (1995), who analyzed the effect of release temperature on survival of releases at Ryde from 1983 through 1990.

Increased levels of (rice) pesticides are associated with increased mortality. Pesticides are yearly values and as such may be proxies for some unmeasured variable, although the inclusion of the pesticide covariate did account for variation over and above that captured by an annual linear trend term.

Site Dependent covariates

The effect of gate position, export levels, and turbidity need to be considered in light of the possible interaction with release site. The export effect is just mildly negative, over the range of export to flow ratios observed. The effect is slightly more harmful for fish released in the central delta, which is where the two major export pumping facilities are located. Mainstem releases that do not stray into the central delta are not at risk of being sucked into the pumps unlike the central delta releases. When the gate is open, releases made upstream seem to suffer while fish released in the central delta appear to benefit. With the gate open there is a greater probability that mainstem releases above the gate, namely from Feather River Hatchery, Sacramento, and Courtland, will enter the central delta and then do risk mortality from the pumpting facilities. On the other hand, fish released inthe delta when the gate is open may be benefitting from the increased water flow.

5.2 Comparing releases with lowest and highest fitted recovery rates

Given potential instability in estimates of individual coefficients, it is useful to compare the worst and best fitted rates at Chipps Island in terms of the corresponding covariate values. (Fitted recovery rates were defined as fitted recoveries divided by release number and f_c.) In other words, one would like to see if there are certain patterns in covariate values for which the model predicts very poor or very good survival.

An initial analysis (not presented here), that compared the best ten and the worst ten fitted rates, indicated that site effects were quite influential; e.g., four of the best ten groups came from Ryde and *Ryde.ind* had the largest positive site indicator coefficient. To focus attention on other factors, the release site effects were partially removed by dividing fitted recovery rates by $\exp(\hat{\beta}_{11} FRH.ind + \ldots + \hat{\beta}_{16} Mkg.ind)$ and the best and worst ten by this measure were compared (Table 4 and Figure 8). Due to interactions between release site and other site dependent covariates (export/inflow ratio, gate position, turbidity) this attempt to control for site effect is only partially effective but nonetheless revealing.

Figure 8 shows the impossibility of separating salinity from flow. For example, the highest fitted recovery rate was with a very low flow but very high salinity, while the sixth highest fitted recovery rate was with a very high flow but very low salinity. The most noteable separation is from the *Shock* covariate and relatedly from $Release.Temp$ and $Hatchery.Temp$. Somewhat lower pesticide levels are evident for the best compared to the worst. In Table 4 the contributions per covariate, namely $\hat{\beta}_i x_i$, are given as well to highlight which terms were most influential. Again the large influence of hatchery and release temperatures is evident.

5.3 Evaluating management strategies

For management purposes the model serves as a means of estimating, or predicting, the relative survival rates of two management strategies. It is not a tool for estimating the absolute survival rate of a single strategy for several reasons. A primary reason is the degree of non-identifiability of the probability of surviving and the probability of recapture conditional on survival. Secondary reasons included the related fact that recovery rate is just estimated up to an unknown proportionality constant (and predicted recoveries for some combinations of covariates can exceed number released) the problematic choice of the $Trend$ term.

The application of the model for comparing survival strategies is in principle straightforward. Because of cancellations in the numerator and denominator, only those covariates for which the groups have differing values, and those coefficients that differ for identical covariates, need to be considered.

A problem in practical application, however, is that of not extrapolating beyond the data used to fit the model. With this many covariates this is not a trivial matter. Besides not inputing values outside the joint range of the covariates, e.g., not inputting a flow of 100,000 cfs, one must avoid selecting combinations that have not occured or cannot occur. Examples of the latter are high flows and high salinities, or high flows and an open gate.

As an example where the input values do fall within the historical set, suppose interest is in comparing the survival of fish under differing gate positions and export/inflow ratios. The fish will be released from Sacramento and all other covariates, release numbers, fish sizes, flows, etc, are identical.

- Strategy I: the gate is open and export/inflow ratio is 0.4
- Strategy II: the gate is closed and export/inflow ratio is 0.2

The only relevant coefficients are those for *Upper.Exp.Inflow* and *Upper.Gate*, where the covariate for *Upper.Gate* when the gate is closed is zero. The calculations can be eased by simply using the difference in covariate values as new covariates. The relative survival of the first strategy to the second strategy (using the coefficients for unstandardized variables):

$$\begin{aligned}\hat{\mathrm{E}}\left[\frac{S_{\text{Strategy I}}}{S_{\text{Strategy II}}}\right] &= \exp((-0.219 \times (0.4-0.2))[Upper.Exp.Inflow] + (-0.348 \times (1-0))[Upper.Gate] \\ &= 0.676\end{aligned}$$

The estimated standard error in this case is 0.0827 and the estimated prediction error is 0.825 (see Appendix A). On average, for every 10 fish reaching Chipps Island under Strategy II, only 6.76 ± 2 × 0.827, or 5.11 to 8.41 (the latter a crude 95% confidence interval ignoring bias in the ridge estimates), should reach Chipps Island under Strategy I.

Comparisons such as this must be tempered with caution, however. The model we have developed summarizes historical relationships and is relevant to prediction in such a passively observed system. Because a number of unmeasured variables may well be important, it is much less suited to predicting what would happen if the system were directly manipulated (Box, 1966). Thus it would be a mistake to take quite literally the numerical predictions of the model in the latter case; a more modest and realistic hope is that they point to beneficial management strategies.

Acknowledgments

Financial support for this work has been provided by the US Fish and Wildlife Service (for Newman) and California Urban Water Agencies (for Rice). We thank Pat Brandes (US Fish and Wildlife Service) and Phyllis Fox (California Urban Water Agencies (CUWA)) for their substantial input. Other key individuals were Elaine Archibald (consultant to CUWA), Randy Bailey (consultant to Metropolitan Water District), Jim Buell (consultant to Metropolitan Water District), Wim Kimmerer (Romberg Tiburon Center of San Francisco State University), and Tom Taylor (consultant to City of San Francisco). Numerous others provided suggestions and advice, including Alan Baracco, Sheila Greene, Martin Kjelson, Sharon Kramer, Pete Rhoads, and Terry Speed. Allison Britton carried out the vital work of putting together in a useful manner data coming from a wide variety of sources.

References

Baker, P. F., Speed, T. P., and Ligon, F. K. (1995.) Estimating the influence of temperature on the survival of chinook salmon smolts (*Oncorhynchus tshawytscha*) migrating through the Sacramento-San Joaquin River Delta of California. *Canadian Journal of Fisheries and Aquatic Sciences*, **52**:855–863.

Box, G. E. P. (1966.) Use and abuse of regression. *Technometrics*, **8**:625–629.

le Cessie, S., and van Houwelingen, J. C. (1992.) Ridge estimators in logistic regression. *Applied Statistics*, **41**:191–201.

Clark, G.H. (1929.) Sacramento-San Joaquin salmon (*Oncorhynchus tschawytscha*) fishery of California. *California Department of Fish Game Fisheries Bulletin.* **17**: 73 p.

Cormack, R. M. (1993.) Recapture data on tagged fish: some questions of analysis and interpretation. **In** *Statistics for the Environment*, V. Barnett and K. F. Turkman (eds), 311–331.

Duffy, D. E., and Santner, T. J. (1989.) On the small sample properties of norm-restricted maximum likelihood estimators for logistic regression models. *Communications in Statistical Theory and Methods*, **18**:959–980.

Healey, M.C. (1991.) Life history of chinook salmon. **In** *Pacific Salmon Life Histories*, C. Groot and L. Margolis (eds), 311–393. Vancouver: UBC Press.

Kjelson, M. and Brandes, P. (1989.) The use of smolt survival estimates to quantify the effects of habitat changes on salmonid stocks in the Sacramento-San Joaquin Rivers, California. **In** *Proceedings of the National Workshop on Effects of Habitat Alteration on Salmonid Stocks*, C.D. Levings, L.B. Holtby, and M.A. Henderson (ed), 100-115. Canadian Special Publication of Fisheries and Aquatic Sciences 105.

Kjelson, M., Greene, S., and Brandes, P. (1989.) A model for estimating mortality and survival of fall-run chinook salmon smolts in the Sacramento river delta between Sacramento and Chipps Island. *USFWS Technical Report, WQCP-USFWS-1*, 50 p.

Kjelson, MA., Raquel, P.F., and Fisher, F.W. (1982.) Life history of fall-run juvenile chinook salmon, *Oncorhynchus tshawytscha*, in the Sacramento-San-Joaquin estuary, California. **In** *Estuarine comparisons*, V.S. Kennedy (ed), 393–411. New York: Academic Press.

Lee, A. H., and Silvapulle, M. J. (1988.) Ridge estimation in logistic regression. *Communications in Statistical Simulation*, **17**: 1231–1257.

Mallows, C. (1998.) The Zeroth Problem. *The American Statistician*, **52**, 1–9.

McCullagh, P., and Nelder, J.A. (1989.) *Generalized linear models, 2nd Ed.* New York: Wiley and Sons.

Shaefer, R. L. (1986.) Alternative estimators in logistic regression when the data are collinear. *Journal of Statistical Computing and Simulation*, **25**: 75–91.

Schaefer, R. L., Roi, L. D., and Wolfe, R. A. (1984.) A ridge logistic estimate. *Communications in Statistical Theory and Methods*, **13**: 99–113.

Segerstedt, B. (1992.) On ordinary ridge regression in generalized linear models. *Communications in Statistical Theory and Methods*, **21**: 2227–2246.

Weisberg, S. (1985.) *Applied linear regression, 2nd Ed.* New York: Wiley and Sons.

A Standard errors

To estimate the standard errors of the estimated coefficients as well as the fitted values of Y, an analytical approximation was used Formally, since

$$\hat{\beta} \quad = \quad (X^tWX + \lambda I)^{-1}X^tWZ$$

the variance of $\hat{\beta}$ can be shown to be approximately:

$$\mathrm{Var}[\hat{\beta}] \quad \approx \quad (X^tWX + \lambda I)^{-1}X^tW\mathrm{Var}[Z]W^tX(X^tWX + \lambda I)^{-1}$$

where $\mathrm{Var}[Z] = \Phi W^{-1}$, with Φ being a diagonal matrix of dispersion values.

Estimates of the expected value and the predicted value for number of recoveries given standardized covariates $x_2^*, \ldots, x_p^*$, are the same:

$$\hat{\mathrm{E}}[Y] = \hat{Y} \quad = \quad Rfe^{\hat{\beta}_1 + \hat{\beta}_2 x_2^* + \ldots + \hat{\beta}_p x_p^*}$$

but the variances differ because of the additional variation in predicting a particular return number as opposed to just estimating the average return number.

Let $\mathbf{x}^*$ be a p by 1 vector of standardized covariate values for a single observation. The variance of a single expected value is:

$$\begin{aligned} \mathrm{Var}[\hat{\mathrm{E}}[Y]] &= \mathrm{Var}[Rf\exp(\mathbf{x}^*\hat{\beta})] \\ &= (Rf)^2\mathrm{Var}[\exp(\mathbf{x}^*\hat{\beta})] \\ &\approx (Rf\exp(\mathbf{x}^*\hat{\beta}))^2\mathbf{x}^{*t}\mathrm{Var}[\hat{\beta}]\mathbf{x}^* \\ &= \hat{\mathrm{E}}[Y]^2\mathbf{x}^{*t}\mathrm{Var}[\hat{\beta}]\mathbf{x}^* \qquad (7) \end{aligned}$$

The variance of a predicted value follows from the 'double variance' formula and the assumed overdispersed Poisson distribution:

$$\begin{aligned} \mathrm{Var}[\hat{Y}] &= \mathrm{Var}_{\hat{\beta}}\mathrm{E}_{Y|\hat{\beta}}[\hat{Y}] + \mathrm{E}_{\hat{\beta}}\mathrm{Var}_{Y|\hat{\beta}}[\hat{Y}] \\ &= \mathrm{Var}_{\hat{\beta}}[\hat{\mathrm{E}}[Y]] + \mathrm{E}_{\hat{\beta}}[\hat{\phi}\hat{Y}] \qquad (8) \end{aligned}$$

where the first term of (8) is the variance in (7).

In the case of estimating ratios of survival rates, the point estimate can be found most simply by substituting a vector of differences in covariate values and dividing the differences by the standard deviation vector from the baseline dataset. The variance of the estimated expected ratio can be estimated using equation (7), substituting the differenced covariance vector divided by the standard deviation vector for $\mathbf{x}^*$ and the estimated ratio for $\hat{\mathrm{E}}[Y]$. This variance estimate is based on a first order Taylor series approximation to the ratio estimate written as a function of the estimated coefficients. A first order approximation to the variance of a prediction is to add the estimated ratio to the variance for the expected ratio (assuming equal dispersion parameters).

Table 1: Sum of variances for estimated coefficients as function of ridge parameter value.

λ	0	10	20	30	35	40	45	50	60
$\sum_{i=1}^{38} \hat{V}[\hat{\beta}_{\lambda,i}]$	0.473	0.209	0.152	0.123	0.113	0.105	0.098	0.092	0.083

Table 2: Estimated coefficients and standard errors (se) for the standardized covariates. $\hat{\beta}_\lambda$ is the coefficient with ridge parameter set equal to 40. $\hat{\beta}$ is the coefficient without a ridge parameter. Default site location is Jersey Point and default release year is 1994.

Covariate	$\hat{\beta}_\lambda$	se($\hat{\beta}_\lambda$)	$\hat{\beta}$	se($\hat{\beta}$)
Intercept	-5.978	0.046	-6.049	0.050
Chipp.ind	1.013	0.069	1.594	0.264
Site Independent Variables				
Size	0.074	0.050	0.111	0.060
Log.Flow	0.153	0.062	0.251	0.110
Salinity	0.269	0.060	0.365	0.099
Pesticide	-0.170	0.054	-0.186	0.076
Trend	-0.069	0.062	-0.042	0.101
Release.Temp	-0.278	0.061	-0.392	0.097
Hatchery.Temp	-0.088	0.061	-0.119	0.094
Shock	-0.052	0.058	-0.010	0.087
Tide.Var	-0.026	0.045	-0.077	0.052
Site Dependent Variables				
FRH.ind	-0.188	0.047	-0.174	0.080
Sac.ind	0.039	0.049	0.161	0.171
Slo.ind	0.068	0.048	0.096	0.097
Crt.ind	0.040	0.048	0.126	0.151
Ryd.ind	0.144	0.051	0.208	0.173
Mkg.ind	-0.126	0.059	-0.111	0.084
Upper.Exp.Inflow	-0.036	0.061	-0.080	0.107
Delta.Exp.Inflow	-0.067	0.066	-0.096	0.120
Upper.Gate	-0.144	0.051	-0.151	0.059
Delta.Gate	0.127	0.066	0.196	0.125
Mainstem.Turbid	-0.042	0.057	-0.146	0.094
Delta.Turbid	-0.034	0.064	-0.084	0.133
Ocean Year Effects				
1979.*ind*	-0.103	0.056	-0.017	0.077
1980.*ind*	0.083	0.041	0.253	0.090
1981.*ind*	-0.060	0.068	0.026	0.101
1982.*ind*	0.060	0.037	0.156	0.055
1983.*ind*	-0.072	0.042	0.083	0.085
1984.*ind*	0.120	0.048	0.358	0.101
1985.*ind*	0.045	0.035	0.233	0.086
1986.*ind*	0.230	0.032	0.436	0.079
1987.*ind*	0.238	0.028	0.422	0.082
1988.*ind*	0.239	0.042	0.535	0.134
1989.*ind*	-0.047	0.057	0.279	0.155
1990.*ind*	-0.038	0.048	0.166	0.116
1991.*ind*	-0.015	0.045	0.185	0.107
1992.*ind*	0.022	0.047	0.231	0.104
1993.*ind*	0.232	0.041	0.453	0.109

Table 3: Change in estimated Site Independent coefficients when a single Site Independent covariate is omitted. Entries are $\hat{\beta}_{all}$ - $\hat{\beta}_{omit}$.

Covariate	Value with	Omitted covariate								
	all SI covs	*-Size*	*-Log.Flow*	*-Salinity*	*-Pesticide*	*-Trend*	*-Rel.* *Temp*	*-Hatch.* *Temp*	*-Shock*	*-Tide.Var*
Size	0.074	—	0.013	-0.034	-0.010	-0.011	0.011	0.007	0.000	0.007
Log.Flow	0.153	0.013	—	0.122	0.022	-0.011	-0.008	0.006	0.001	0.000
Salinity	0.269	-0.019	0.069	—	0.011	0.017	0.021	0.012	0.003	-0.004
Pesticide	-0.170	0.000	-0.011	0.000	—	0.023	-0.008	-0.007	0.011	0.002
Trend	-0.069	0.020	0.021	-0.061	0.086	—	0.027	-0.025	-0.010	-0.003
Release.Temp	-0.278	-0.005	0.006	-0.019	-0.003	0.009	—	0.036	0.036	-0.009
Hatchery.Temp	-0.088	-0.010	-0.010	-0.032	-0.021	-0.017	0.096	—	0.003	0.005
Shock	-0.052	-0.001	-0.006	-0.017	0.042	-0.012	0.161	0.004	—	0.003
Tide.Var	-0.026	-0.015	-0.002	0.011	0.009	-0.003	-0.039	0.007	0.004	—

Table 4: Covariate values for the 10 best (listed first) and 10 worst release groups based on model estimates of recovery 'rate' at Chipps Island with site effect removed. Numbers in smaller, italicized type are the contribution to each estimate, the coefficient times covariate value.

Recovery 'Rate'	Fish Size	(log) Sac Flow	Collin. Salin.	Pest.	Trend	Rel. Temp.	Hatch. Temp.	Shock	Tide Var	Site	Export	Gate Pos.	Turbidity
1.87	86	8.85	13301	2982796	91	61	53	5	2.8	Jers	0.20	Open	8.5
	(*1.01*	*2.79*	*0.96*	*-0.80*	*-1.63*	*-3.87*	*-1.55*	*-0.05*	*-0.16*		*-0.10*	*0.34*	*-0.06*)
1.54	79	8.92	13404	2982796	91	61	52	8	2.3	Mk-G	0.52	Open	10.0
	(*0.93*	*2.81*	*0.97*	*-0.80*	*-1.63*	*-3.87*	*-1.52*	*-0.08*	*-0.13*		*-0.25*	*0.34*	*-0.07*)
1.52	93	9.75	422	2940220	80	55	55	0	1.8	FRH	0.24	Closed	8.5
	(*1.09*	*3.07*	*0.03*	*-0.79*	*-1.43*	*-3.49*	*-1.59*	*0.00*	*-0.10*		*-0.05*	*0.00*	*-0.06*)
1.16	82	8.89	10846	2982796	91	63	52	8	2.3	Jers	0.52	Open	10.0
	(*0.96*	*2.81*	*0.79*	*-0.80*	*-1.63*	*-3.99*	*-1.52*	*-0.08*	*-0.13*		*-0.25*	*0.34*	*-0.07*)
1.14	84	8.81	9843	2982796	91	65	51	6	1.6	Mk-G	0.34	Open	8.5
	(*0.98*	*2.78*	*0.71*	*-0.80*	*-1.63*	*-4.12*	*-1.49*	*-0.06*	*-0.09*		*-0.17*	*0.34*	*-0.06*)
1.13	79	10.84	164	2821676	83	60	52	8	2.7	Crt	0.04	Closed	25.0
	(*0.93*	*3.42*	*0.01*	*-0.75*	*-1.48*	*-3.80*	*-1.52*	*-0.08*	*-0.15*		*-0.01*	*0.00*	*-0.17*)
1.08	81	10.82	166	2821676	83	61	55	4	2.0	Ryde	—	—	25.0
	(*0.95*	*3.41*	*0.01*	*-0.75*	*-1.48*	*-3.87*	*-1.60*	*-0.04*	*-0.11*		—	—	*-0.17*)
1.07	71	9.14	11897	3484814	90	63	53	8	2.0	Jers	0.48	Open	10.0
	(*0.83*	*2.88*	*0.86*	*-0.93*	*-1.61*	*-3.99*	*-1.55*	*-0.08*	*-0.11*		*-0.24*	*0.34*	*-0.07*)
1.03	94	9.69	422	2940220	80	55	56	0	1.7	FRH	0.24	Open	8.5
	(*1.10*	*3.06*	*0.03*	*-0.79*	*-1.43*	*-3.49*	*-1.63*	*0.00*	*-0.10*		*-0.05*	*-0.35*	*-0.06*)
1.02	75	10.84	168	2821676	83	63	53	7	1.4	Mk-G	0.05	Closed	9.5
	(*0.88*	*3.42*	*0.01*	*-0.75*	*-1.48*	*-3.99*	*-1.55*	*-0.07*	*-0.08*		*-0.02*	*0.00*	*-0.07*)
0.12	74	9.36	7752	4613002	88	76	60	23	2.7	Crt	0.46	Open	6.7
	(*0.87*	*2.95*	*0.56*	*-1.23*	*-1.57*	*-4.82*	*-1.75*	*-0.22*	*-0.15*		*-0.10*	*-0.35*	*-0.05*)
0.17	89	9.42	7752	4613002	88	74	60	19	2.4	Sac	0.46	Open	6.7
	(*1.04*	*2.97*	*0.56*	*-1.23*	*-1.57*	*-4.69*	*-1.75*	*-0.18*	*-0.14*		*-0.10*	*-0.35*	*-0.05*)
0.24	88	9.43	7752	4613002	88	76	60	21	2.4	Slo	—	—	—
	(*1.03*	*2.98*	*0.56*	*-1.23*	*-1.57*	*-4.82*	*-1.75*	*-0.20*	*-0.14*		—	—	—)
0.27	85	9.50	5113	3559856	89	71	57	11	2.5	Crt	0.26	Open	7.0
	(*1.00*	*3.00*	*0.37*	*-0.95*	*-1.59*	*-4.50*	*-1.66*	*-0.10*	*-0.14*		*-0.06*	*-0.35*	*-0.05*)
0.27	81	9.54	951	2921654	86	74	58	18	2.1	Ryde	—	—	11.0
	(*0.95*	*3.01*	*0.07*	*-0.78*	*-1.54*	*-4.69*	*-1.69*	*-0.17*	*-0.12*		—	—	*-0.07*)
0.28	88	9.31	7752	4613002	88	74	55	24	2.7	Ryde	—	—	6.7
	(*1.03*	*2.94*	*0.56*	*-1.23*	*-1.57*	*-4.69*	*-1.60*	*-0.23*	*-0.15*		—	—	*-0.05*)
0.28	83	9.50	5334	3559856	89	70	57	12	2.5	Sac	0.26	Open	7.0
	(*0.98*	*3.00*	*0.39*	*-0.95*	*-1.59*	*-4.44*	*-1.66*	*-0.11*	*-0.14*		*-0.06*	*-0.35*	*-0.05*)
0.30	82	9.60	2902	3820103	84	66	58	12	2.6	Crt	0.32	Open	6.5
	(*0.96*	*3.03*	*0.21*	*-1.02*	*-1.50*	*-4.18*	*-1.69*	*-0.11*	*-0.15*		*-0.07*	*-0.35*	*-0.04*)
0.34	84	9.47	11721	4613002	88	75	59	24	1.3	Ryde	—	—	6.7
	(*0.98*	*2.99*	*0.85*	*-1.23*	*-1.57*	*-4.76*	*-1.72*	*-0.23*	*-0.07*		—	—	*-0.05*)
0.34	83	9.15	3161	4713055	94	67	50	15	0.8	Sac	0.11	Closed	5.0
	(*0.97*	*2.89*	*0.23*	*-1.26*	*-1.68*	*-4.25*	*-1.46*	*-0.14*	*-0.05*		*-0.02*	*0.00*	*-0.03*)

Figure 1: Release locations in lower Sacramento river system.

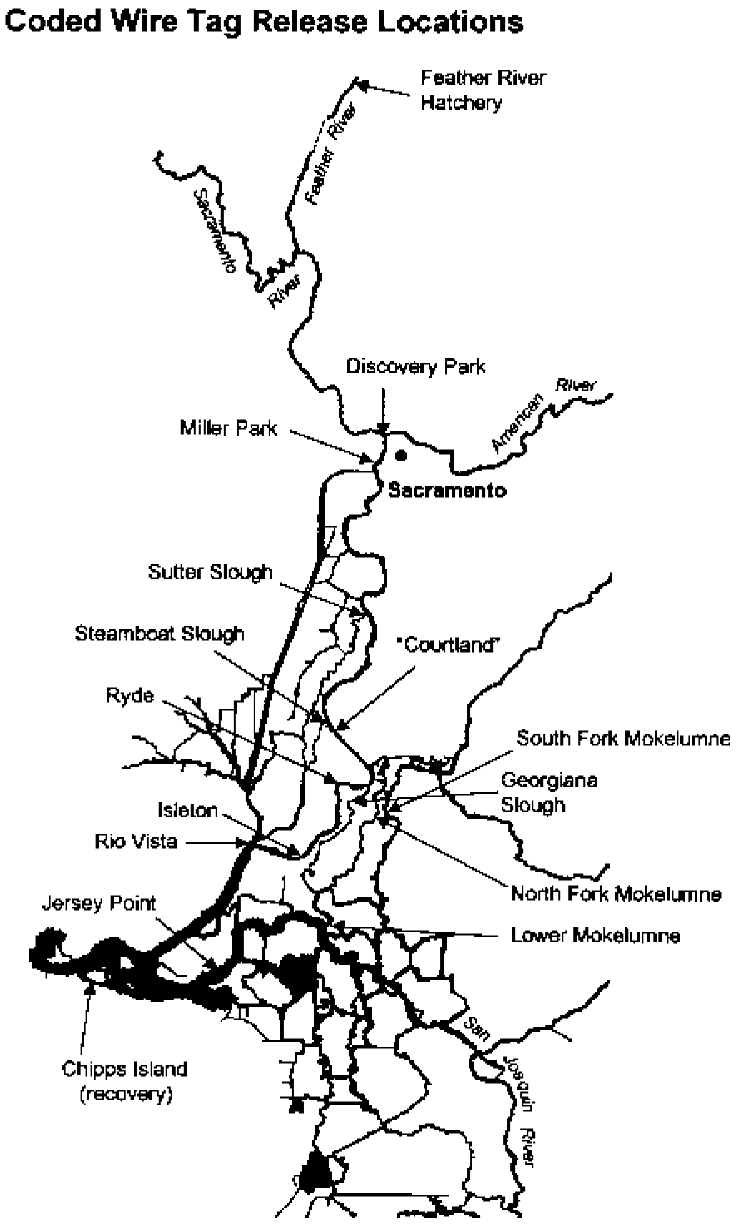


Figure 2: Estimated coefficients for site independent and site dependent covariates ± 2 se's.

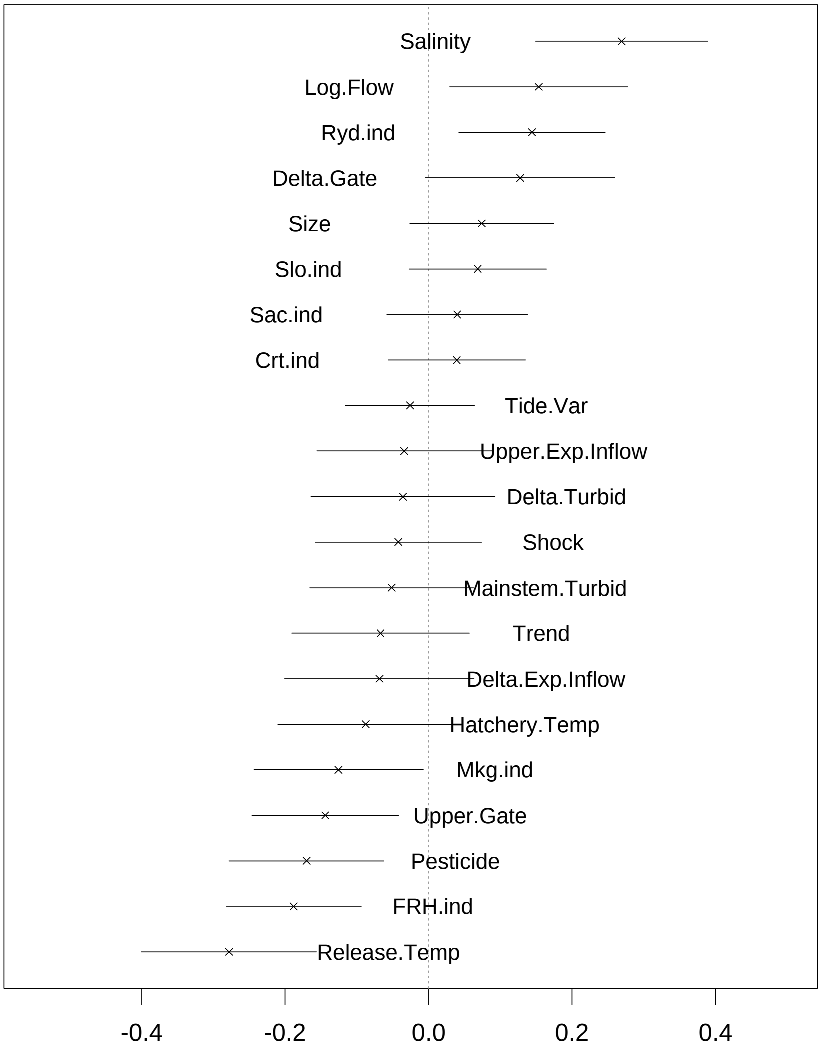
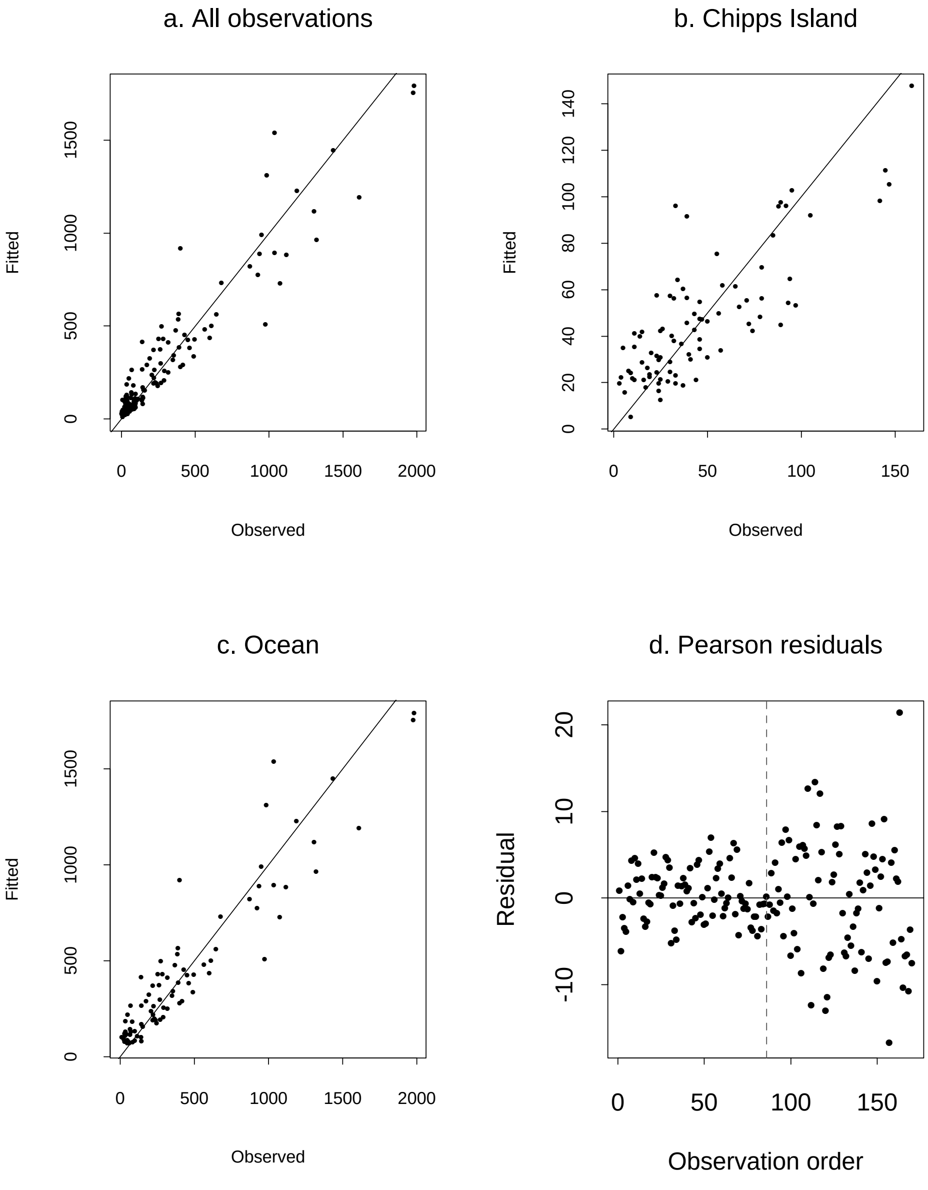


Figure 3: Fitted versus observed recoveries plus residual plot.



a. All observations

b. Chipps Island

c. Ocean

d. Pearson residuals

Figure 4: Cross-validated prediction errors and estimated coefficients for ridge parameter values ranging from 0 to 60.

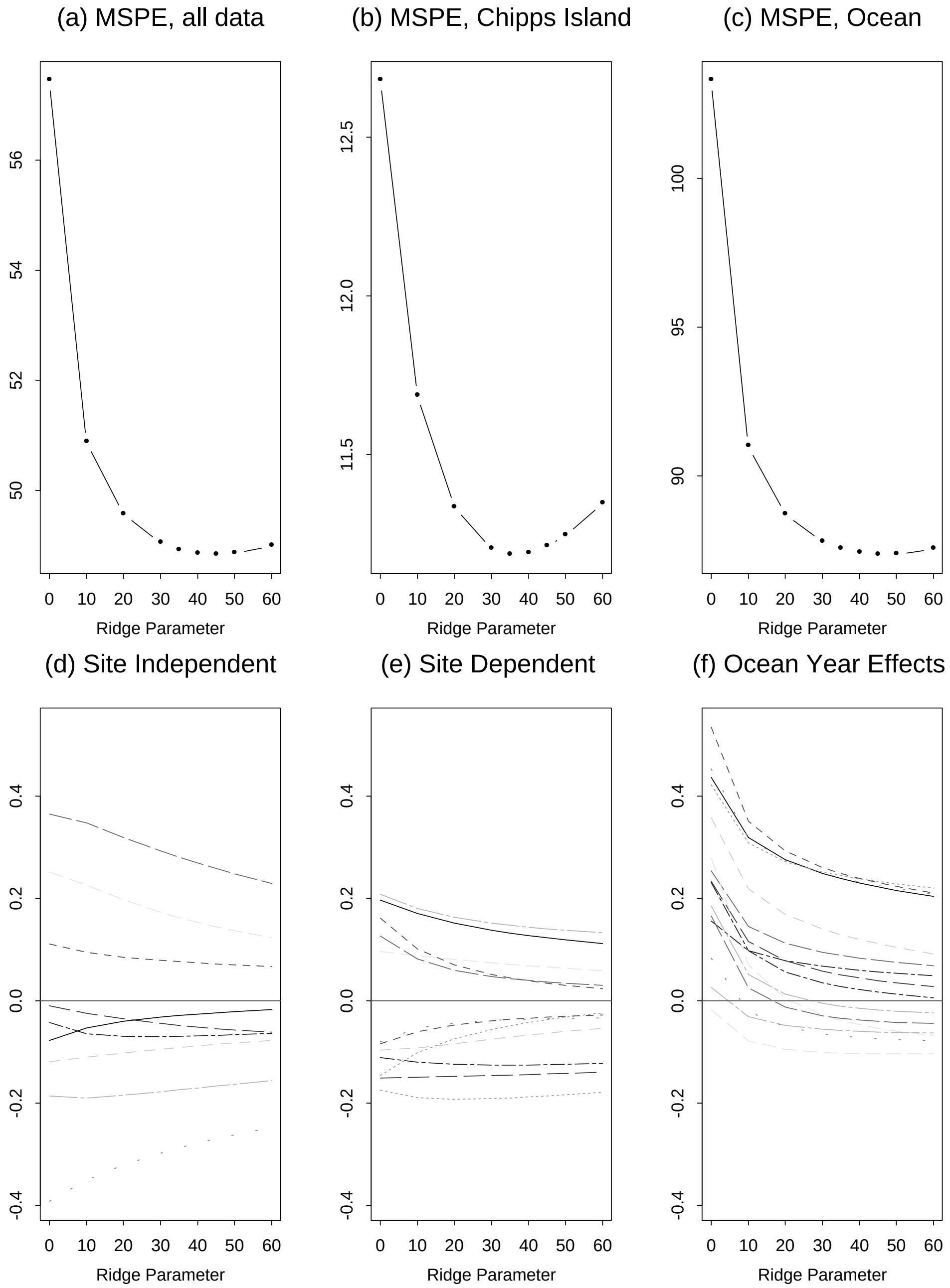


Figure 5: Estimated coefficients (± 1 se) based on Ocean data versus those based on Chipps Island data.

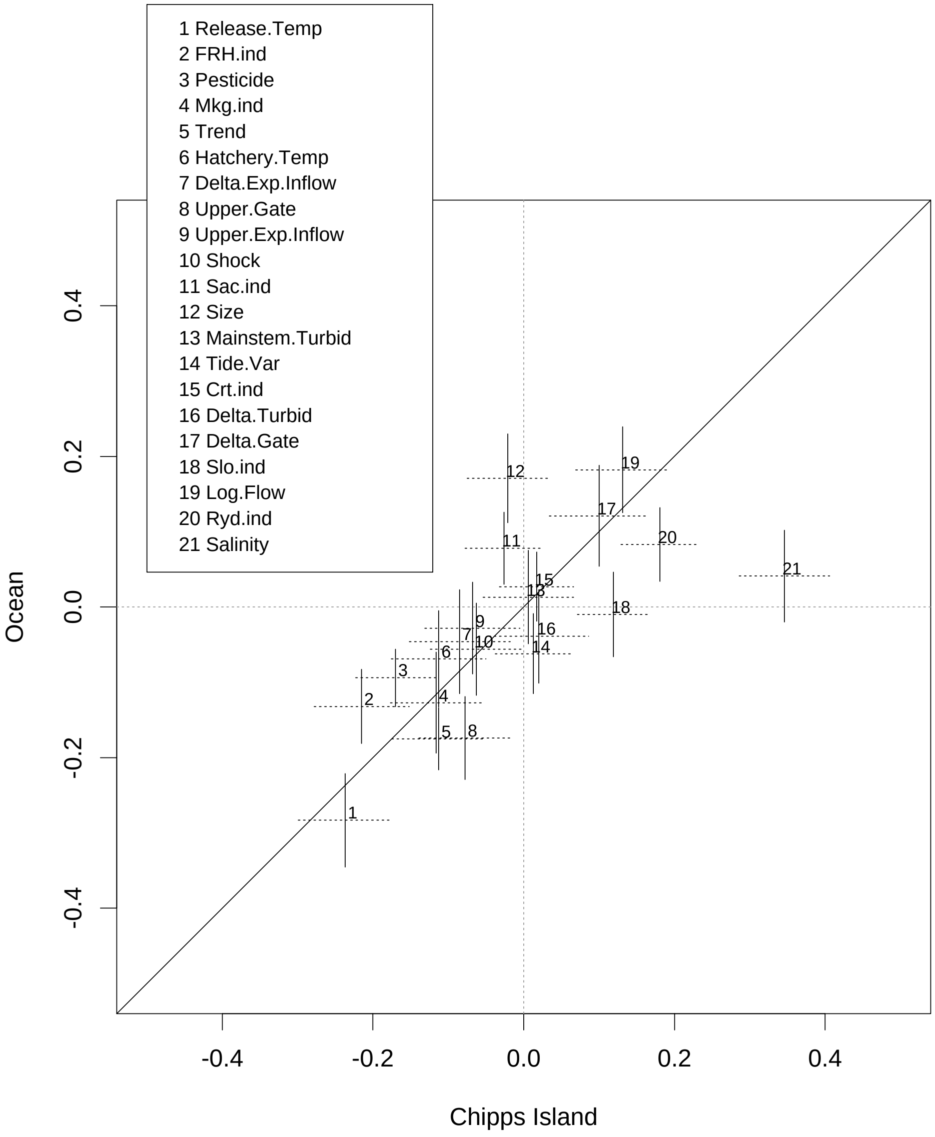


Figure 6: Estimated coefficients (± 1 se) based on both Chipps Island and Ocean data sets combined versus those based on Chipps Island data alone.

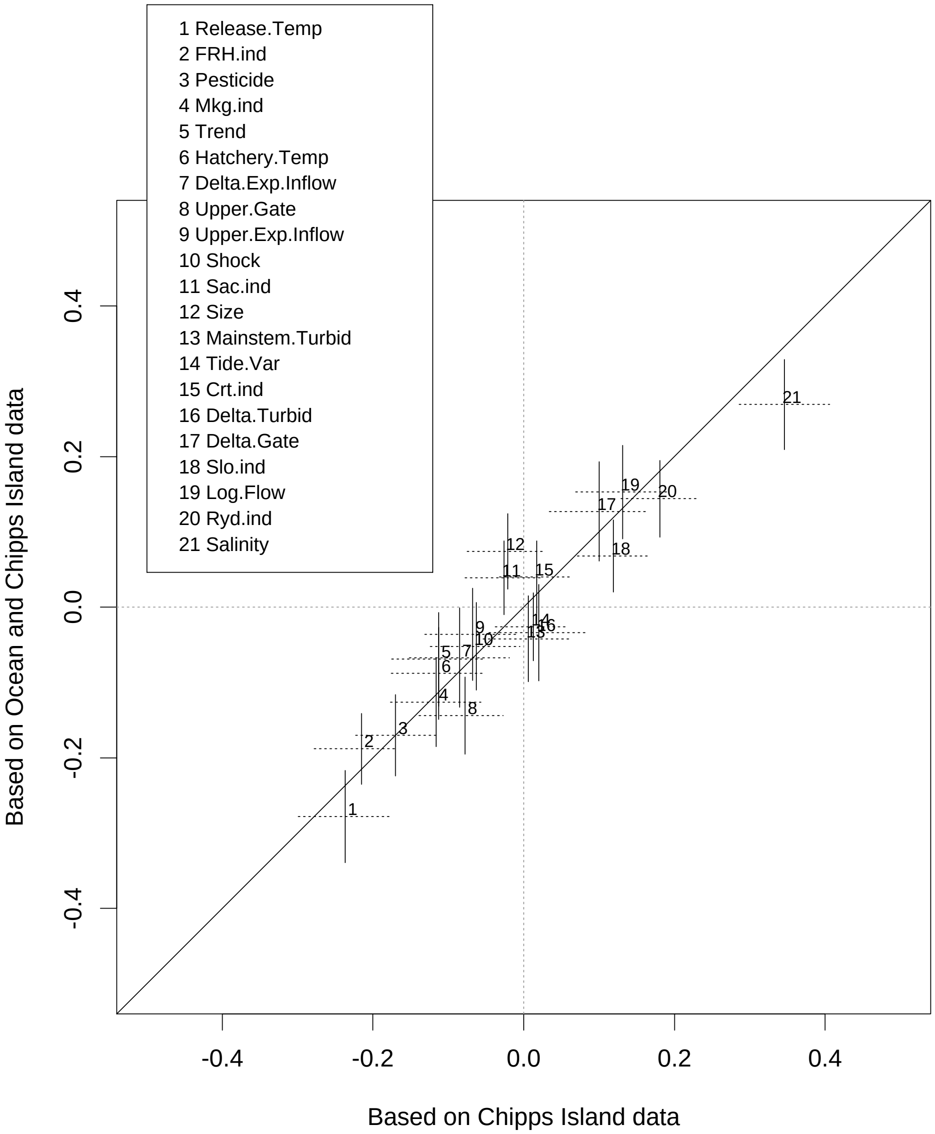


Figure 7: Fitted values for Chipps Island recoveries based on models constructed from two data sets, Ocean data alone versus Chipps Island data alone. (Straight line across plot is the least squares line.)

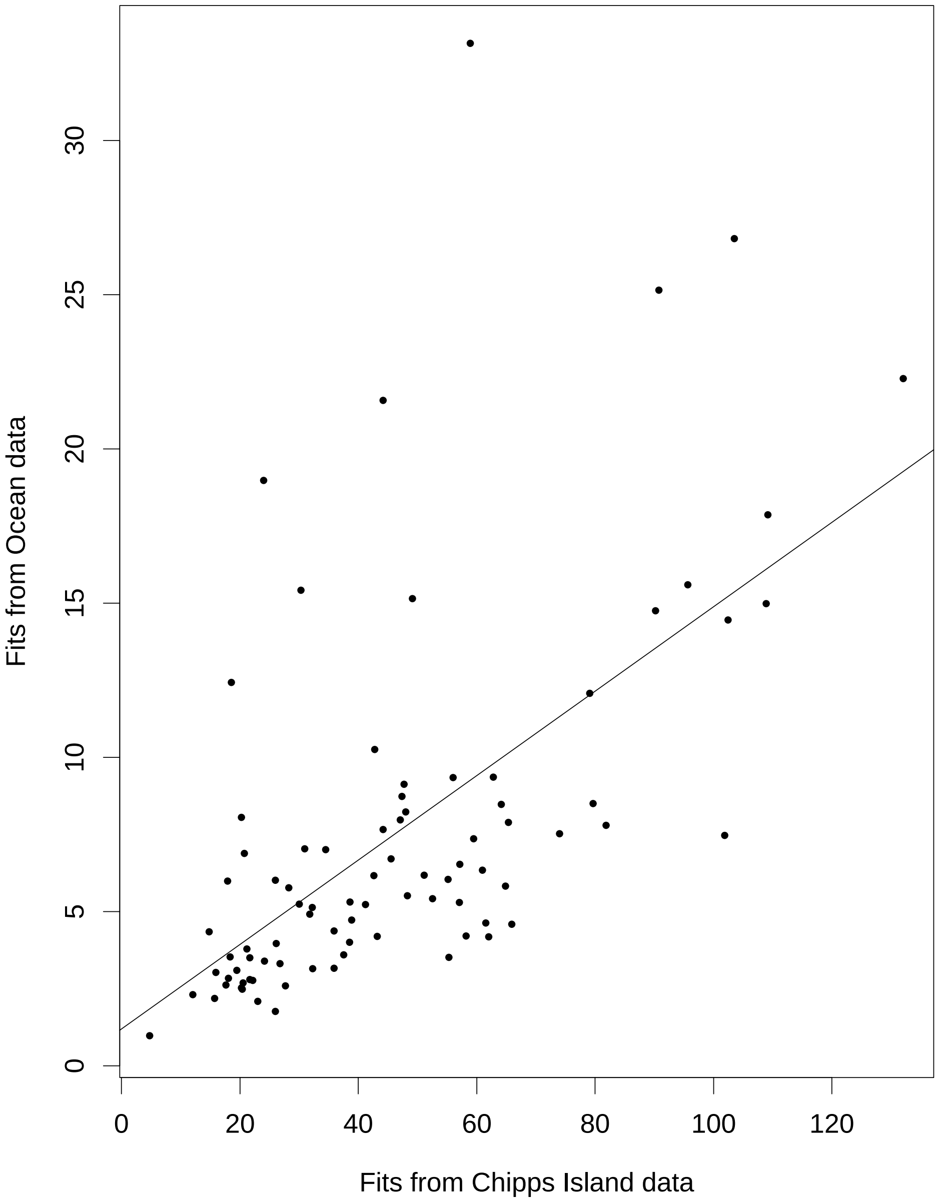
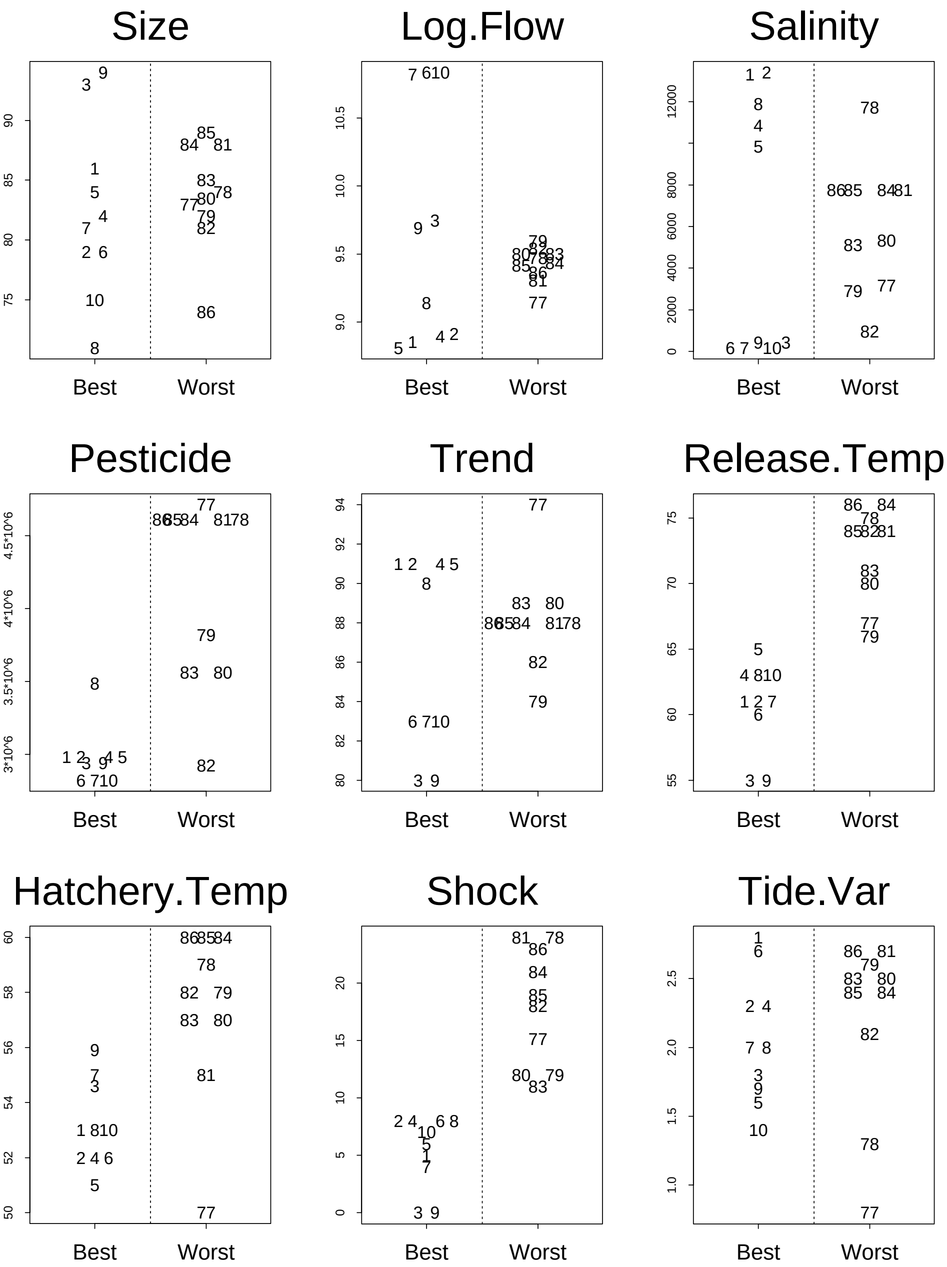


Figure 8: Dot plots of SI covariates for the ten best and ten worst releases, i.e., releases with highest and lowest estimated Chipps Island recovery rates (site effects removed). The numbers in the plots indicate ranks, 1 to 10 for top ten, and 77 to 86 for bottom ten.


Size
Log.Flow
Salinity
Pesticide
Trend
Release.Temp
Hatchery.Temp
Shock
Tide.Var
Best
Worst
3*10^6
3.5*10^6
4*10^6
4.5*10^6