### Classification of known-age coho salmon smolt to freshwater age using linear discriminant function analysis of scale measurement characters

**Intro…**

Determination of freshwater age for coho salmon using scales has long been recognized as problematic and questionably reliable. This is due primarily to the difficulties scale readers have in distinguishing between freshwater annuli and false checks for this species. This is generally believed to result from the complex interplay of highly variable environmental conditions all coho fry experience during their one to three years of freshwater residence in a wide variety of habitats. Experienced scale readers are assumed to be more accurate in their interpretation of scale ages, but even the most skilled readers often express lack of confidence in their aging accuracy for this species. The task requires a lot of subjective interpretation, and is often described as more art than science. Loss of experienced readers and the need to train new readers presents serious challenges to maintaining long term accuracy and reliability of age compostions.

[Preliminary results from a set of known aged smolt at Auke Lake (Farington?) indicated that ages assigned by scale readers were less accurate than even previously believed. ](I don’t have any of the literature to cite for this section). Primary use = brood tables for S/R analysis Blah blah…..

This study was initiated to determine if more objective means of determining scale age could be developed using typical scale patterns analysis (SPA) techniques. Secondarily, could the results be used to develop easily applied methods to aid in training new scale readers, while reducing the subjectivity of scale age interpretations.

**Methods and Materials**

Coho salmon smolt, marked with coded wire tags, were sampled from three Southeast Alaskan locations over a number of years at various life history stages. Auke Lake (AL) was sampled in 1994-1998 and 2000. Bering River (BR) and Hugh Smith Lake (HS) were both sampled in 1998-2005. Captured smolt were sampled for scales, length, and recovery of coded wire tags (CWT). Coded wire tags were used to determine actual age at the time of capture (age 1, 2, or 3). Scales were initially examined by an experienced scale reader and subjectively assigned to perceived age groups. Due to low sample sizes, age 3 scales were excluded from further analysis. Scales images were captured using a recording microfiche reader, and later examined with standard scale pattern analysis (SPA) techniques to classify scales to age at time of capture using the known CWT ages.

We modified standard SPA techniques, regularly used to classify adult fish scales to stock of origin, to classify smolt scale images to age group. Scales images were first measured using image analysis software to obtain distances between scale circuli in individual first year and second year freshwater growth zones (as subjectively identified by an experienced scale reader). Individual circulus measurements were later combined and counted to calculate standardized sets of quantitative scale growth characters. Scale growth characters were then examined using linear discriminant function analyses to determine their suitability for classifying scales to actual ages at the time of capture.

Initially, we obtained a set of 50 quantitative variables (Table 1) for each individual fish characterizing growth characteristics the total annular freshwater growth zones. These variables are similar to those typically used in stock of origin analyses for adults. However, some variables frequently cannot be calculated for specific ages due to their younger age or slow growth (e.g. too few circuli in a zone to calculate a value). Although these null scale character variables can often intuitively seem to provide information helpful to classify a scale to the correct category of interest (stock or age), they must be excluded from LDF analyses because they violate a key constraint for univariate normality. The remaining variables (bold in Table 1) were applied on repeated iterations of STEPDISC (SAS, Inc.) to examine suitability for use in subsets of variables that best differentiated the ages. This selection process was applied separately for each stock to select subsets of variables specific to each stock. The process was also applied for all stocks pooled to select a separate subset of variables to determine suitability of generalized non stock-specific analysis.

To address the central question of whether a simplified and objective means of estimating scale age could be developed, a restricted set of scale character variables was devised. As in the original set, the restricted set of characters considered the first and second annular zones as a single freshwater annular zone excluding any additional growth in a freshwater “plus growth” transition zone. This approximates a scale reader unable to distinguish annuli from false checks in the annular zones, but reliably able to identify the last freshwater annulus from “plus growth,” an experience level reasonably obtainable through training.

All valid remaining variables were examined to determine adherence to univariate normality (a constraint of parametric LDF analysis), and to identify outliers from incorrectly measured images or obviously atypical scales. The variable selection procedures were iteratively run, including and excluding variables to include only non-covariant variables (or variables believed to be minimally covariant) that maximized classification accuracy. Eventually, a subset consisting of only three general, and easily quantified, variables was considered for analysis (starred and bold in Table 1): 1) total count of freshwater annular zone circuli, 2) total width of freshwater annular zone, and 3) count of circuli in the first half of the annular zone).

LDF classification routines for three combinations from the restricted variable subset were used to classify scales to age using variables calculated for a single freshwater annular growth zone. Preliminary descriptive statistical analyses generally indicated large differences in scale measurements between the three different stock locations. Preliminary LDF examinations also indicated that generalized classification models for pooled stocks were not as reliably accurate as stock-specific models. Consequently, discriminant function criteria were developed only for individual locations for the final classification runs. For each stock location, each year of sampling was iteratively designated as a test set and classified to age of origin using all other years of sampling as the learning set. At Auke Lake, age 1 and age 2 samples were only collected in alternating years. For this location, the two ages from two successive sampling years were also pooled to simulate single multiyear test sets.

**Results**

Preliminary examination of LDF classification of scale images to age indicated relatively low accuracy for generalized models using all stocks combined (example: Table 2) (Could also cite figures showing big differences in frequency historgrams for Circuli counts by age for each stock). Stepwise classification usually produced modest accuracy with only one variable included, declining with each additional variable until more than twice as many variables were included than the number of categories. This usually indicates a simple relationship, but confounding issues within the scales included in the learning set. Restricting the analysis to single stocks produced much better accuracies.

Restriction of variables to include in models to only the most rigorously valid scale characters and using only stock-specific subsets generally improved accuracies and behavior during stepwise variable selections (example, Table 3). Accuracies were typically in the mid to high 80%’s with one variable, improved or slightly improved with a second variable, and with little or no change on successive variables. Smolt length was included as a variable during early examinations, and was the most frequently chosen first variable. No surprise, considering the linear relationships between length and scale width frequently used to back calculate length at age. Since a scale reader would not have smolt length available at the time of reading adult scales, scale measurements should be feasible to use as an alternative, and we excluded length from final investigations.

After determining that all other variables seldom (if ever) provided any consequential improvement in model accuracy through iterative runs of the variable selection routine, we restricted the list of variables to three of the most robust and easily accessible variables (starred in Table 2). Three combinations of this restricted list were iteratively run through the LDF classification routine to estimate potential accuracies for stock-specific models to classify scales to age (Tables 4-6). Each year of each stock was iteratively classified to age using a learning set made up of all other years for that stock. Average accuracies (weighted average percentage of each age correctly classified) for learning sets (“expected” model accuracies) and test sets (“observed” model accuracies) from all iterations were tabulated (Table 7).

The first combination of variables (z1:count of circuli in FW annular zone, and z2:size of FW annular zone), was an attempt to combine the statistically related, but intuitively different, observations of counts and width measurements (Table 4). Accuracies were generally high (85-90%) for both Berners River and Hugh Smith, and lower and more variable for Auke Lake (39-100%).

The second combination (z1 and z48:count of circuli in 1st half of FW annular zone) was intended to combine the demonstrated usefulness of the whole zone count with age information that would be visually available to a scale reader either as 1) consistently fast summer growth in the first half an age 1 scale, or as 2) the variably spaced whole year growth pattern of an age 2 scale. This combination improved overall accuracy for most models, especially some of the most problematic Auke Lake models, but reduced accuracy for some and only slightly improved accuracies overall (Table 5).

The last trial included only variable z1–total count of circuli in the FW annular zone. This scale feature requires no specialized equipment and only minimal training for a scale reader to obtain. Average accuracies using this one variable alone were only slightly lower than the other two variable combinations. Although not tested for statistical significance, I suspect the resulting variability on age determination would be considerably less than that for even the best skilled human scale reader.

|  |  |  |
| --- | --- | --- |
| **Table 1. Scale measurement and count characters calculated from intercirculus distances** | | |
| Note: variables in bold are selected for discriminant analysis. | | |
| Variable |  | Total Freshwater Annular Zone |
| **\*Z1** | **v1** | **Number of circuli in FW Annular Zone (NCFAZ)** |
| **\*Z2** | **v2** | **Width of FW Annular Zone (SFAZ)** |
| **Z3** | z2/z1 | Average interval between circuli (SFAZ/NCFAZ) |
| **Z4** | v3 | Number of circuli in Plus Growth Zone (NCPGZ) |
| **Z5** | v4 | Width of Plus Growth Zone (SPGZ) |
| **Z6** | z4/z3 | Average interval between circuli (SPGZ/NCPGZ) |
| **Z7** | v5 | Distance from scale focus (C0) to circulus 1 (C1) |
| **Z8** | v6 | Distance from scale focus to circulus 2 (C0 - C2) |
| **Z9** | v7 | Distance from scale focus to circulus 3 (C0 - C3) |
| **Z10** | v8 | Distance from scale focus to circulus 4 (C0 - C4) |
| **Z11** | v9 | Distance from scale focus to circulus 6 (C0 - C6) |
| Z12 | v10 | Distance from scale focus to circulus 7 (C0 - C7) |
| Z13 | v11 | Distance from scale focus to circulus 8 (C0 - C8) |
| Z14 | v12 | Distance from scale focus to circulus 9 (C0 - C9) |
| Z15 | v13 | Distance from scale focus to circulus 12 (C0 - C12) |
| Z16 | v14 | Distance from scale focus to circulus 15 (C0 - C15) |
| Z17 | v15 | Distance from scale focus to circulus 21 (C0 - C21) |
| **Z18** | v8-v6 | Distance from circulus 2 to circulus 4 (C2 - C4) |
| **Z19** | v9-v6 | Distance from circulus 2 to circulus 6 (C2 - C6) |
| Z20 | v11-v6 | Distance from circulus 2 to circulus 8 (C2 - C8) |
| **Z21** | v9-v8 | Distance from circulus 4 to circulus 6 (C4 - C6) |
| Z22 | v11-v8 | Distance from circulus 4 to circulus 8 (C4 - C8) |
| **Z23** | v9-v7 | Distance from circulus 3 to circulus 6 (C3 - C6) |
| Z24 | v12-v7 | Distance from circulus 3 to circulus 9 (C3 - C9) |
| Z25 | v13-v7 | Distance from circulus 3 to circulus 12 (C3 - C12) |
| Z26 | v14-v7 | Distance from circulus 3 to circulus 15 (C3 - C15) |
| Z27 | v12-v9 | Distance from circulus 6 to circulus 9 (C6 - C9) |
| Z28 | v13-v9 | Distance from circulus 6 to circulus 12 (C6 - C12) |
| Z29 | v14-v9 | Distance from circulus 6 to circulus 15 (C6 - C15) |
| Z30 | v14-v12 | Distance from circulus 9 to circulus 15 (C9 - C15) |
| Z31 | v15-v12 | Distance from circulus 9 to circulus 21 (C9 - C21) |
| Z32 | v14-v13 | Distance from circulus 12 to circulus 15 (C12 - C15) |
| Z33 | v15-v13 | Distance from circulus 12 to circulus 21 (C12 - C21) |
| Z34 | v15-v14 | Distance from circulus 15 to circulus 215 (C15 - C21) |
| Z35 | v16 | Distance from sixth-to-last circulus to end of zone, C(NCFAZ-6) - EOZ |
| Z36 | v17 | Distance from third-to-last circulus to end of zone, C(NCFAZ-3) - EOZ |
| **Z37** | v2-v7 | Distance from circulus 3 to end of zone (C3 - EOZ) |
| Z38 | v2-v12 | Distance from circulus 9 to end of zone (C9 - EOZ) |
| Z39 | v2-v14 | Distance from circulus 15 to end of zone (C15 - EOZ) |
| Z40 | z25/z2 | Relative width, (variable 25)/SFAZ |
| Z41 | z26/z2 | Relative width, (variable 26)/SFAZ |
| Z42 | z27/z2 | Relative width, (variable 27)/SFAZ |
| Z43 | z28/z2 | Relative width, (variable 28)/SFAZ |
| Z44 | z29/z2 | Relative width, (variable 29)/SFAZ |
| Z45 | z30/z2 | Relative width, (variable 30)/SFAZ |
| Z46 | z32/z2 | Relative width, (variable 32)/SFAZ |
| Z47 | z34/z2 | Relative width, (variable 34)/SFAZ |
| **\*Z48** | **v18** | **Number of circuli in first 1/2 of zone** |
| **Z49** | v19 | Number of circuli in first 3/4 of zone |
| **Z50** | v20 | Maximum distance between two consecutive circuli |