Alaska Hatchery Research Program **Technical Document:**[1]

Title: Thermal Mark Recovery Data Quality Assurance and Quality Control Procedures by the ADF&G Mark, Tag and Age Laboratory **Version:** 1.0
Authors: Agler, B., L. Wilson, and M. Lovejoy
Date: May 25, 2017

Abstract

Origin of Pacific salmon (*Oncorhynchus* spp.) sampled for the Alaska Hatchery Research Program can be determined by examining otoliths (ear stones) for thermal marks. Thermal mark presence indicates that a fish originated from a hatchery; whereas, thermal mark absence indicates wild origin. Identification of such marks provides information about a fish's age, hatchery of origin, and release location. The Mark, Tag and Age Lab, Alaska Department of Fish and Game is responsible for conducting mark recovery operations for a variety of statewide management and research projects. Thermal-marked fish typically are not given a secondary mark, so multiple readings among readers and across geographic areas are used to estimate reader ability to detect a thermal mark and to calculate agreement of thermal mark identifications. Thus, we compare first and second reads with an agreement matrix to determine whether there are any significant problems in reader training or challenging marks that might be re-examined. We then use the *kappa* statistic to examine overall agreement between readers as well as agreement by specific thermal mark. At the end of each project, we estimate the error rates of each reader using latent class models, because although useful, *kappa* statistics are influenced by the true proportion of marked fish. Analyzing the thermal mark read results in this manner provides a method to ensure quality control among projects and a measure of accuracy of thermal mark recoveries of fish sampled for the Alaska Hatchery Research Program.

Background of AHRP

Extensive ocean-ranching salmon aquaculture is practiced in Alaska by private non-profit corporations (PNP) to enhance common property fisheries. Most of the approximately 1.7B juvenile salmon that PNP hatcheries release annually are pink salmon in Prince William Sound (PWS) and chum salmon in Southeast Alaska (SEAK; Vercessi 2014). The large scale of these hatchery programs has raised concerns among some that hatchery fish may have a detrimental impact on the productivity and sustainability of natural stocks. Others maintain that the potential for positive effects exists. To address these concerns ADF&G convened a Science Panel for the

[1] This document serves as a record of communication between the Alaska Department of Fish and Game Commercial Fisheries Division and other members of the Science Panel of the Alaska Hatchery Research Program. As such, these documents serve diverse ad hoc information purposes and may contain basic, uninterpreted data. The contents of this document have not been subjected to review and should not be cited or distributed without the permission of the authors or the Commercial Fisheries Division

Alaska Hatchery Research Program (AHRP) whose members have broad experience in salmon enhancement, management, and natural and hatchery fish interactions. The AHRP was tasked with answering three priority questions:

I. *What is the genetic stock structure of pink and chum salmon in each region (PWS and SEAK)*?;
II. *What is the extent and annual variability in straying of hatchery pink salmon in PWS and chum salmon in PWS and SEAK*?; and
III. *What is the impact on fitness (productivity) of natural pink and chum salmon stocks due to straying of hatchery pink and chum salmon*?

Introduction

An important consideration in fisheries management is the ability to identify the origins of captured and harvested fish. The development of mass-marking techniques, such as thermal manipulation of water temperature to mark otoliths, permits millions of hatchery-incubated juvenile salmon to be marked simultaneously. These techniques have successfully applied species-specific, thermal mark patterns to otoliths (ear stones) of hatchery-reared salmon throughout Alaska and the Pacific Rim over the past 26 years (Hagen et al. 1995; Volk et al. 1990). For the AHRP, accurate mark interpretation is vital to the assessment of stray rates associated with hatchery-reared salmon and provides validation of genetic stock identifications.

There are many potential sources of error in any research project, and the extent that these errors can be minimized increases confidence in a study's findings and conclusions. There are two categories of reliability with respect to data collectors: reliability across multiple data collectors, or *inter-rater* reliability, and reliability of a single data collector, or *intra-rater* reliability. Presented with the same situation and phenomenon every time, the assumption is that a laboratory staff would react the same way every time; however, Gwet (2014) provided examples of where this was false and affected intra-rater reliability. Reader reliability is affected by the fineness of discriminations required by the samples. If a variable only has two possible states, and the states are sharply differentiated, reliability is likely to be high. For example, if the outcome variable is that a fish either survived or did not, or the otolith is marked or not marked, there is likely to be high reliability in data comparisons between readers. On the other hand, if readers are required to make judgements or determinations regarding the width and amount of thermal mark rings, both inter- and intra-rater reliability declines. Careful training of laboratory staff is critical to reader reliability.

To determine the presence or absence of a thermal mark in an otolith, laboratory staff use pattern recognition and image matching (Blick and Hagen 1998). Although hatcheries follow strict rearing protocols to produce consistent thermal marks, natural variation in otolith development and growth patterns can obscure these patterns and interfere with the ability to detect a mark, reducing mark identification. In addition, stress on fish at the hatchery caused by temperature fluctuations, water quality, rearing density, noise and light fluctuations, lot size, maintenance procedures, and handling protocols can affect mark consistency and clarity (Hagen et al. 1995**)**.

Examination of accuracy rates of thermal mark identifications, including correct assignment of age, hatchery, and release site, provide useful knowledge regarding reliability of mark recoveries to assess stray rates and validate genetic stock identifications. However, otolith thermal marks are typically applied without a secondary mark, such as a coded-wire tag or a passive integrated transponder (PIT) tag, thus there is no reliable method to assess the true accuracy of thermal mark presence and identification. Consequently, the Mark, Tag, and Age (MTA) Lab uses latent class models (LCMs) to estimate a reader's ability to distinguish between hatchery and wild fish. In addition, *kappa* statistics (Cohen 1960) are used to assess reader agreement among individual mark patterns. Agreement matrices combined with the *kappa* statistic assist in identifying problematic mark patterns.

Goal

Our goal is to describe the methods used by the MTA Lab in Juneau, Alaska to find errors in thermal mark classification and correct them. We describe the methods used to assess the accuracy of reader's ability to correctly ascertain presence and absence of a thermal mark and to identify specific mark patterns.

Methods

Hatcheries apply thermal marks to incubating salmon eggs and fry by raising and lowering the water temperature at set intervals. Cycling temperature, or thermal marking, leaves patterns of optically-dense rings in the otolith (Volk et al. 1990). A thermal mark consists of rings, which are optically dark circles visible in the otolith and bands, which consist of one or more rings separated by a space from other rings (Figure 1). We describe the thermal mark with a specialized notation termed the "hatch code" (Josephson et al. 2006). For example, a hatch code of 4,2,2H describes a set of three bands: the first band is composed of four rings, the second band includes two rings, and the last band contains two rings. The capital "H" indicates the mark was applied before hatching. In this example, all three bands occur prior to the hatch mark (Figure 1). Varying the number and spacing of the induced rings produces unique patterns used to distinguish among similarly treated hatchery fish and wild stock (Hagen et al. 1995).

Thermal mark reference collection

Initially, laboratory personnel are trained to dissect, prepare, and process otoliths from reference specimens or representative samples of salmon eggs, fry, and smolt obtained from the hatchery and preserved in alcohol before release. Upon receipt, five otoliths from each sample are dissected, mounted to glass slides (see "Otolith mounting" section of AHRP MTA Processing Tech Doc 7), and examined with a compound microscope. These reference specimens become the standard or authoritative mark pattern for that thermal mark after laboratory staff compare the observed marked with the assigned mark. Laboratory staff then measures the specimens, because mark locations and ring spacing can vary among individuals of the same thermal mark group due to variability of fry developmental stage during marking. When a thermal mark is applied during different developmental stages, the distance from the core of the otolith to the initial band varies among fish making the mark challenging to identify. Successful thermal mark application at hatcheries is the first step to correct determination of fish origin and is fundamental to the success of this project.

Thermal marks are described in the Mark Characteristic Report (available online – see below). This report includes: brood year, release year, thermal mark identification, species, brood stock, release site(s), assigned (target) mark, actual mark observed, mark quality assessments, number of samples received, measurements of the otolith, information about temperature profile (if available), various comments, and an authoritative image of the mark as well as images of any variants of the mark. Measurements include minimum, maximum, and average distance (µm) from the core of the otolith to the first band; minimum, maximum and average width of each band; and distance among bands (Figure 2). The authoritative image, which represents the mark pattern observed in the majority of voucher samples, is annotated with measurements and a comment about the thermal mark (Figure 1). Occasionally, thermal marking procedures can produce errant mark patterns or multiple pattern variants of the planned mark (Figure 3). When this occurs, images of these mark variants are included in the reference collection. The Mark Characteristic Report and the thermal mark reference collection are both available online:

http://www.taglab.org/OTO/reports/VoucherSummary.asp

Reader Training

Prior to each field season, laboratory staff (or "readers") gain familiarity with the thermal mark patterns likely to appear in AHRP samples by studying the physical and online reference collection of marked otoliths maintained at the MTA Lab. Familiarization with thermal mark patterns is important because growth rings in otoliths of wild salmon can occasionally appear to be similar to marks create during the thermal marking process. This review of known marks helps to minimize the chance of labeling an otolith as marked when it is actually wild as well as helps to increase reader accuracy and precision with regards to mark identification.

Laboratory personnel are trained to process adult otoliths using surplus otoliths to practice grinding to visually enhance the core or the "primordia" of the otolith. Staff learns to reduce processing time by controlling the pressure exerted during grinding and by becoming familiar with variations in otolith patterns and shapes. After approximately two to four weeks of training, laboratory staff begins to examine samples containing a mixture of marked and unmarked otoliths. Experienced personnel work with new staff members until their reader agreement is at least 95%.

First and Second Reads

All chum salmon (*O. keta*) otoliths are examined twice. In other words, these samples are read independently by a first reader and then read a second time by a different reader. The second reader typically knows who read the first sample but has no knowledge of the previous read results. Thus, we consider these to be a blind second read. The AHRP stream and pedigree samples are stratified into four areas (Figure 4). Disagreements between first and second readers are resolved by a third reader examining the otolith. The third read is not independent. The third reader knows who conducted both first and second reads and is cognizant of the results of each read. Second reads are performed as first reads are completed, and readers review the results. If disagreements occur, these are discussed, increasing familiarity with challenging patterns.

Study Design

Samples are assigned to readers by sample location (area) and over time. The MTA Lab currently uses four readers, thus there are six reader-pair combinations, which is critical for data analysis using a latent class model (see below). For the AHRP, the stream strata include four geographic areas in Southeast Alaska (Figure 4). Four streams were chosen for the pedigree sites, and each pedigree stream is treated as one stratum.

Read Assessment Methods

The MTA Lab uses three methods to assess a reader's ability to determine the presence or absence of a thermal mark. These methods include two agreement measures (agreement matrix and *Kappa*) and a latent class model, part of a family of models that allow estimation of reader classification error through the use of spatial data and multiple independent readings.

1) Agreement Matrices

As otoliths are examined, a preliminary review of results is conducted by cross-tabulating the first read and second read results (Table 1). Common in reliability studies (Blick and Hagen 1998), this matrix highlights results to review in detail. The matrix also highlights thermal marks that are mistakenly termed wild fish, as well as thermal mark identifications with a high percentage of disagreement. The first reader's results are listed on the rows, while the second reader's results are listed in the columns. Table 1 shows the number of thermal marked fish as well as the number not marked (e.g. wild) and unreadable. The numbers on the diagonal between the rows and columns indicates the number of thermal marks upon which the two readers agreed. Numbers off the diagonal highlight the disagreements (Table 1). For example, reader one and two agreed that 34 otoliths were TM3, but reader one called two otoliths TM4 and reader two labeled them TM3. Discrepancies in whether the otoliths are marked or unmarked are located on the edge of the matrix, and differences in readability may also be found by examining the matrix. For example, six otoliths were labeled TM4 by reader one but were called "wild" by reader two, and three otolith were called wild by reader one but labeled TM3 by reader two. Examination of the matrix provides a preliminary analysis during a project and allows biologists to target areas for review. Deviations from the diagonal are reviewed, and sometimes otoliths are read a third time to ensure consistency. This matrix has been a useful tool for highlighting when a reader missed a mark. Often such errors are caused by incorrect sample preparations. If an otolith is not ground enough, the thermal mark will not be visible. In such cases, the sample is simply ground some more until the core is visible. Conversely, if an otolith is ground too much, the mark will be removed. In this instance, the other otolith can be prepared for mark recovery since both left and right otoliths will exhibit a thermal mark.

2) Latent Class Model

Latent class models (LCMs) provide an alternative approach to estimating agreement (Hui and Walter 1980). LCMs incorporate an estimate of reader classification error, so that the variability of reader agreement may be estimated. These models hypothesize the existence of unobservable (i.e. “latent”) variables about which information can only be obtained through measurements on observable (i.e. “manifest”) variables (Blick and Hagen 1998). LCMs use categorical variables for the latent and manifest variables. For the AHRP, the latent variable is whether an otolith is hatchery or wild; whereas, the manifest variables are a reader’s classifications. Because the true error rate for each reader is unknown, latent class models provide a method to assess the accuracy of thermal mark results. Blick and Hagen (1998) demonstrated that LCMs could be successfully applied to thermal mark results by setting additional constraints or collecting additional information.

The most economical LCM method is to separate the study area into strata and use two readers. Use of three or more readers would give more degrees of freedom (*df*) and improve model results, but the cost of the project would increase. Maximum likelihood models are the preferred method for estimating LCMs. Assuming readings are independent among readers and among otoliths, the likelihood function is as follows:

$$\prod_{i=H,W}\prod_{j=H,W}\prod_{k=H,W}\left\{p\pi_{i|H}^{(1)}\pi_{j|H}^{(2)}\pi_{k|H}^{(3)}+(1-p)\pi_{i|W}^{(1)}\pi_{j|W}^{(2)}\pi_{k|W}^{(3)}\right\}^{n_{ijk}}$$

$where$

H	=	hatchery (thermal marked)
W	=	wild (unmarked)
n	=	sample size
$\pi_{i\|j}{}^{(k)}$	=	probability that reader *k* classifies an otolith as *i* when its true state is *j*
p	=	proportion of hatchery fish

The likelihood functions used to estimate the above parameters are maximized using Solver in Microsoft Excel. Standard errors are estimated using the jackknife method (Haddon 2001).

When there are only two readers, neither is a standard, and there are five parameters to estimate $\pi_{H|H}^{(1)}$, $\pi_{H|H}^{(2)}$, $\pi_{W|W}^{(1)}$, $\pi_{W|W}^{(2)}$ and *p*, which gives only three *df* (four data points – one due to fixed sample size, *n*). To prevent overparameterization, constraints on the parameters or more data are needed. Possible constraints include: 1) considering two parameters as known (e.g.; $\pi_{W|W}^{(1)} = \pi_{W|W}^{(2)} = 1$, both readers will call a wild stock correctly); or 2) considering two sets of parameters equal (e.g.; $\pi_{H|H}^{(1)} = \pi_{H|H}^{(2)} = \pi_{W|W}^{(1)} = \pi_{W|W}^{(2)}$, the accuracy rates are the same for both readers). These constraints are likely unrealistic, thus more data are necessary. One way to generate more information is to have a third independent reader (Walter 1984). Three readers provide seven parameters: $\pi_{H|H}^{(1)(2)(3)}$, $\pi_{W|W}^{(1)(2)(3)}$, and *p,* thus there are $2^3 - 1 = 7$ *df*, so all parameters may be estimated. On the other hand, adding a third reader is usually logistically unfeasible given the financial constraints of a project.

Hui and Walter (1980) proposed an alternative method to generate information. They suggested that if there are two or more strata with different hatchery proportions in each strata (Blick and Hagen 1998), then reader results could be stratified temporally or spatially. We can then assume that $\pi_{H|H}^{(k)}$ and $\pi_{W|W}^{(k)}$ remains constant across strata (Blick and Hagen 1998), reducing model parameters to eight with 12 *df*. Thus, a two reader – four strata model would have 4 *df* extra for goodness-of-fit, preventing overparameterization of the model.

The following is the likelihood function for the two independent reads with *S* strata (Hui and Walter 1980):

$$\prod_{g=1}^{S}\prod_{i=H,W}\prod_{j=H,W}\left\{p_g\pi_{i|H}^{(1)}\pi_{j|H}^{(2)} + (1-p)\pi_{i|W}^{(1)}\pi_{j|W}^{(2)}\right\}^{n_{gij}}$$

To estimate the latent variable for each reader, the stream samples collected during the AHRP project were separated into four spatial strata (Figure 4). These spatial strata included: (1) Southern Southeast waters; (2) Lynn Canal and Stephens Passage; (3) Chatham and Icy Straits; and (4) Northern Outside waters. Samples were apportioned fairly equally across area. In addition, these areas provided both geographic coverage and geospatial separation. Pedigree samples were separated into strata based on the four creeks used in the project: Fish, Prospect, Admiralty, and Sawmill creeks. Care was taken to distribute readings evenly among readers, across areas, and by time. Samples were distributed among readers equally because we have observed that when the LCM was heavily weighted by one individual, it performed poorly.

We have also observed that “reader drift” can occur over time as readers observe more marks and sometimes altered their initial perception of a mark pattern (intra-rater reliability). To ensure that the LCM analyses included this potential scenario, we assigned readers samples from across the entire study period.

A critical assumption for both the LCM estimates of reader ability to detect a mark and *kappa* agreement values (see below) is that readings are independent, meaning that the reading of each otolith by a reader is independent of any other reading by the same reader and independent of readings by other readers for a given otolith. To support these assumptions, otolith first and second reads are provided to readers in random order by box. Another assumption is that individual accuracy rates are known to be greater than the error rates (Blick and Hagen 1998). Historically, reader agreement associated with mark recoveries conducted during the commercial sockeye fishery exceed 95%, so we believe this assumption is likely valid for the MTA Lab.

3) Kappa

The *kappa* statistic (Fleiss 1981) is frequently used to test inter-rater reliability. Rater reliability represents the extent to which the data collected in a study represent the variables measured. The *kappa* statistic provides examination of overall agreement between readers as well as agreement by specific thermal mark and an associated standard error (Fleiss 1981). Individual *kappa* statistics can be calculated for each category and pooled from different trials. Traditionally, inter-rater reliability was measured as percent agreement, calculated as the number of agreement scores divided by the total number of scores. Cohen (1960) critiqued the use of percent agreement due to its inability to account for chance, thus percent agreement tends to be higher when a category being rated has a high probability of occurrence. He introduced the Cohen's *kappa* (1960), which is chance corrected or accounts for the possibility that raters guess on some variables due to uncertainty.

Kappa is calculated by correcting the observed agreement for the degree of agreement expected by chance alone (P_O = $(n_{HH} + n_{WW})/n$). Overall *kappa* is weighted and is defined as:

$$\hat{\kappa}_w = \frac{P_o - P_e}{1 - P_e} \tag{3}$$

where P_e is the proportion of expected agreement = $(n_H n_H + n_W n_W)/n^2$ (Cohen 1960; Blick and Hagen 1998; Fleiss 1981). The weighted version of *kappa* has the same properties discussed above, but it is adjusted by giving lower weight to disagreements over marks with small numbers and full weight to disagreements over marks where agreement is high (Hagen et al. 1995). This better reflects agreement on what is marked and unmarked and reduces the influence of mark identifications with only one or two otoliths. Overall $\hat{\kappa}$, which assesses overall agreement between readers, is a weighted average of individual $\hat{\kappa}$ for each individual thermal mark identified and is equal to the sum of the individual p_0 - p_e (i.e., the sum of the numerators of the individual $\hat{\kappa}$) divided by the sum of the individual 1 - p_e differences (i.e., the sum of the denominators of individual $\hat{\kappa}$, Fleiss 1981).

The standard error for $\hat{\kappa}_w$ is estimated by:

$$SE(\hat{\kappa}_w) = \frac{\sqrt{A+B-C}}{(1-p_e)\sqrt{n}} \quad (4)$$

where

$$A = \sum_{i=1}^{n} p_{ij}\left[1-\left(p_i+p_j\right)+\ (1-\hat{\kappa}_w\ )\right]^2, \quad (5)$$

$$B = (1-\hat{\kappa}_w\ )^2 \sum\sum p_{ij}(p_i+p_j)^2\ , \quad (6)$$

and

$$C = [\hat{\kappa}_w - \ p_e(1-\hat{\kappa}_w)]^2 \quad (7)$$

for readers *i* and *j* who have read *n* samples.

Although *kappa* is a commonly used inter-rater reliability statistical test, it has limitations. Judgments about what level of *kappa* is acceptable are often questioned. As in most correlation statistics, *kappa* values range from -1 to +1, where $\hat{\kappa}_w$ = 1 indicates complete agreement and $\hat{\kappa}_w$ = -1 indicates complete disagreement. If observed agreement is greater than or equal to chance agreement, $\hat{\kappa}_w \geq 0$, and if observed agreement is less than or equal to chance alone, $\hat{\kappa}_w \leq 0$ (Landis and Koch 1977). Landis and Koch (1977) suggested that $\hat{\kappa}_w$ > 0.61 indicates substantial agreement beyond chance. Values between 0.41 and 0.60 represent moderate agreement, and $\hat{\kappa}_w$ < 0.40 represent slight to poor agreement (Landis and Koch 1977). Although Landis and Koch (1977) interpreted a *kappa* score of 0.41 as acceptable, this might be considered too lenient for a project like AHRP.

At the MTA Lab, we use *kappa* to ascertain amount of agreement among marks between readers. Overall *kappa* among a suite of marks can be high (>0.80), but sometimes *kappa* scores for individual marks can be low (<0.50). This occurs for a variety of reasons: 1) the mark was rarely observed in a sample, usually older-aged fish; 2) otoliths were over- or underground; 3) mark application was incomplete or differed among incubation groups, causing recovering to be challenging; and 4) duplication of mark patterns among brood years required that otoliths be aged to differentiate between years. Once we have determined why errors occurred, we determine whether a higher proportion of the sample need to be second read or whether we need to have some samples re-examined to determine whether marks were missed (i.e.; mount right side of otolith and examine for thermal mark by a third reader). In the last instance, we work with staff to improve thermal mark identification proficiency.

Thermal marks with poor *kappa* values are examined and discussed among readers during each year of the project. They are also targeted for study prior to each project year. If a sample has a poor overall *kappa* value, then those otoliths are examined further to determine the cause (i.e.;

multiple poor marks or a sample coordination errors). *Kappa* values are archived on the local network.

Because *Kappa* is an index, it is important to remember that interpretation can be affected by the values of the underlying parameters (Blick and Hagen 1998). Thus, direct comparison of $\hat{\kappa}$ across populations with different underlying proportions is not appropriate. Although agreement measures may be subject to some ambiguity, they are useful in monitoring results for potential errors and pinpointing areas for the Lab to re-examine.

Discussion

Fisheries research often requires that trained individuals classify data according to a strict but somewhat subjective set of rules. In many situations, there is no standard available with which to confirm classifications, and it is necessary to apply some other method to determine the accuracy of the determinations. Distinguishing thermal-marked fish from wild fish is a good example of this type of problem because: 1) most thermal-marked salmon do not receive a secondary mark, so cross-validation is not possible; and 2) the ability to read otoliths for thermal mark presence and identification requires training and experience because natural variation in growth rings observed in chum salmon otoliths can appear similar to thermal mark patterns. In the absence of samples of known origin, it is common to collect multiple, independent observations of the same samples and assume that percent agreement among readers serves as a proxy for read accuracy. Agreement indices (matrices and *kappa*) are easy to compute and indicate read discrepancies in mark recovery and identifications**.** For the AHRP project, these QA/QC methods provide additional direction for validation of reader accuracy and precision. They also provide some quantitative indication of reader accuracy.

In addition, we use the agreement measures described above to highlight results in need of closer examination and suggest potential areas for critical review. When agreement measures indicate that results require evaluation, we examine the data to determine whether we need to: 1) conduct additional reader training when an individual is under- or over-grinding and missing marks, 2) read samples a third time by another independent reader when marks are especially difficult to discern, and 3) examine potential issues in greater detail during the next season’s training period if a particular mark or brood year is expected to return.

Although these indices are fairly easy to calculate and are useful indicators of reading problems, it is important to remember that some of these indices are not directly comparable. It is difficult to compare *kappa* statistics across populations with different underlying proportions. Because of this, even when a suite of *kappas* is consistent, it may not be clear how reader agreement/disagreement influences the contribution estimate. In addition, these indices do not provide inferences about the relative ability of one reader over another to determine a particular set of patterns. Latent class models, however, provide readily interpretable qualities that can be easily calculated. Classification accuracies or errors provide direct, meaningful parameters, unlike the use of an index of agreement alone. In addition, LCMs provide estimates of hatchery proportions (*p*).

We feel that the procedures described above provide a combination of approaches to provide a comprehensive examination of error rates and accuracy of reads conducted in the MTA Lab. The matrices and *kappa* statistics point out areas for review, and the LCM provides direct, meaningful parameters that can be compared from year-to-year.

References

Blick, D. J., and P. T. Hagen. 1998. The use of agreement measures and latent class models to assess the reliability of thermally marked otolith classifications. NPAFC Doc 370:1-15.

Campana, S. E. 1983. Calcium deposition and otolith check formation during periods of stress in coho salmon, *Oncorhynchus kisutch*. Comparative Biochemistry and Physiology 75A.(2):215-220.

Cohen, J. 1960. A coeffecient of agreement for nominal scales. Educational and Psychological Measurement XX(1):37-46.

Fleiss, J. L. 1981. Statistical methods for rates and proportions 2nd edition. John Wiley and Sons, New York, NY.

Gwet, K. L. 2014. Handbook of inter-rater reliability, 4th edition. Advanced Analytics, Gaithersburg, MD.

Haddon, M. 2001. Modelling and quantitative methods in fisheries. Chapman and Hall, New York, NY.

Hagen, P., K. Munk, B. W. Van Alen, and B. White. 1995. Thermal mark technology for inseason fisheries management: a case study. Alaska Fishery Research Bulletin 2(2):143-155.

Hui, S. L., and S. D. Walter. 1980. Estimating the error rates of diagnostic tests. Biometrics 36:167-171.

Josephson, R., B. A. Agler, K. F. Van Kirk, and D. S. Oxman. 2006. A proposal to simplify the thermal mark code notation. NPAFC Doc 944:1-4.

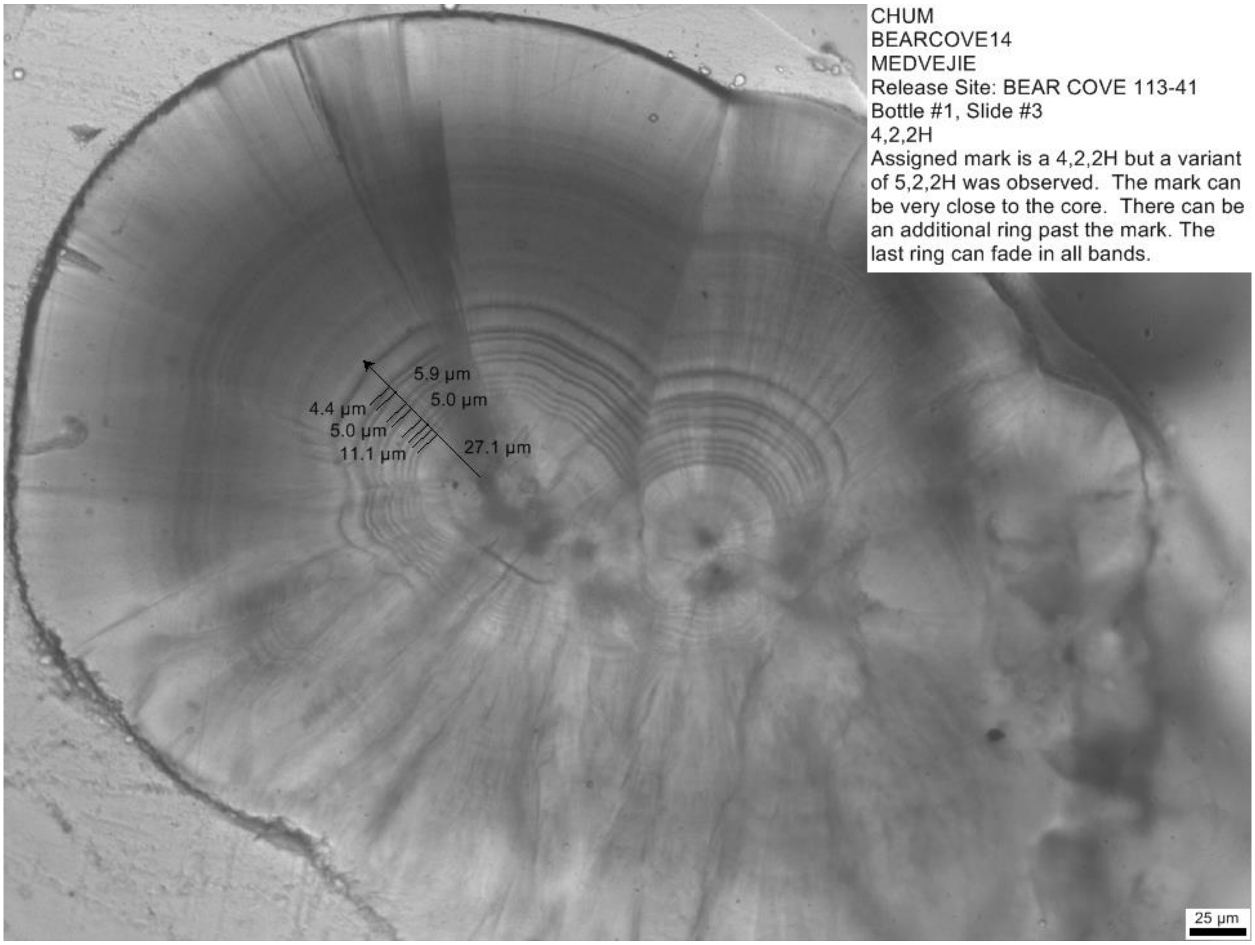
Landis, J. R., and G. G. Koch. 1977. The measurement of observer agreement for categorical data. Biometrics 33:159-174.

Vercessi, L. 2014. Alaska salmon fisheries enhancement program 2013 annual report. Alaska Department of Fish and Game, Anchorage.

Volk, E. C., S. L. Schroder, and K. L. Fresh. 1990. Inducement of unique otolith banding patterns as a practical means to mass-mark juvenile Pacific salmon. American Fisheries Society Symposium 7:203-215.

Walter, S. D. 1984. Measuring the reliability of clinical data: the case for using three observers. Revue d'épidémiologie et de santé publique 32(3-4):206-211.

Figures



CHUM
BEARCOVE14
MEDVEJIE
Release Site: BEAR COVE 113-41
Bottle #1, Slide #3
4,2,2H

Assigned mark is a 4,2,2H but a variant of 5,2,2H was observed. The mark can be very close to the core. There can be an additional ring past the mark. The last ring can fade in all bands.

Figure 1. Image of a thermal mark reference specimen. From Medvejie Hatchery, this brood year 2014 mark (BearCove14) has a hatch code of 4,2,2H. The code indicates that the first band from the otolith's core contains four dark rings, then there is a space, followed by a band with two rings, followed by another space and a final band with two rings prior to the hatch mark (blurry, wide, dark area). Annotated measurements on the transect line include distance from otolith core (primordia) to first band, width of first band, space between first and second bands, and average distance between rings in each band. All thermal mark images are available online through the North Pacific Anadromous Fish Commission (NPAFC) Working Group on Salmon Marking (WGOSM) website: http://wgosm.npafc.org/MarkSummary.asp

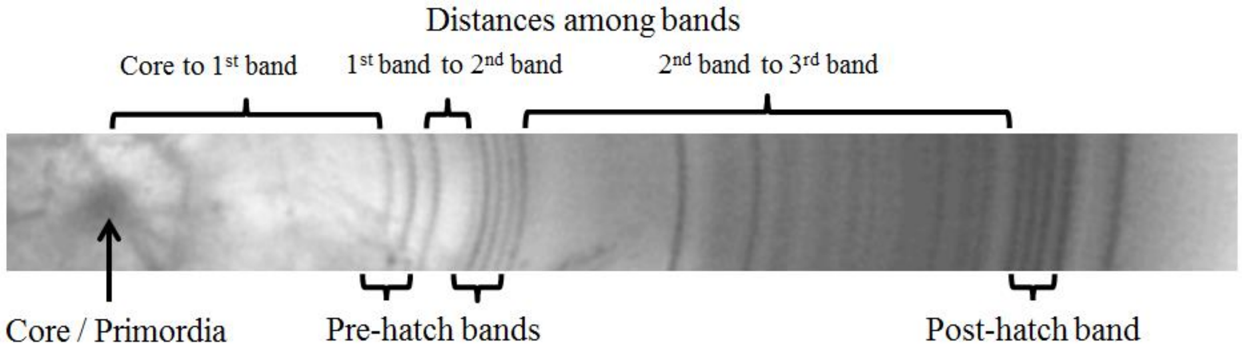
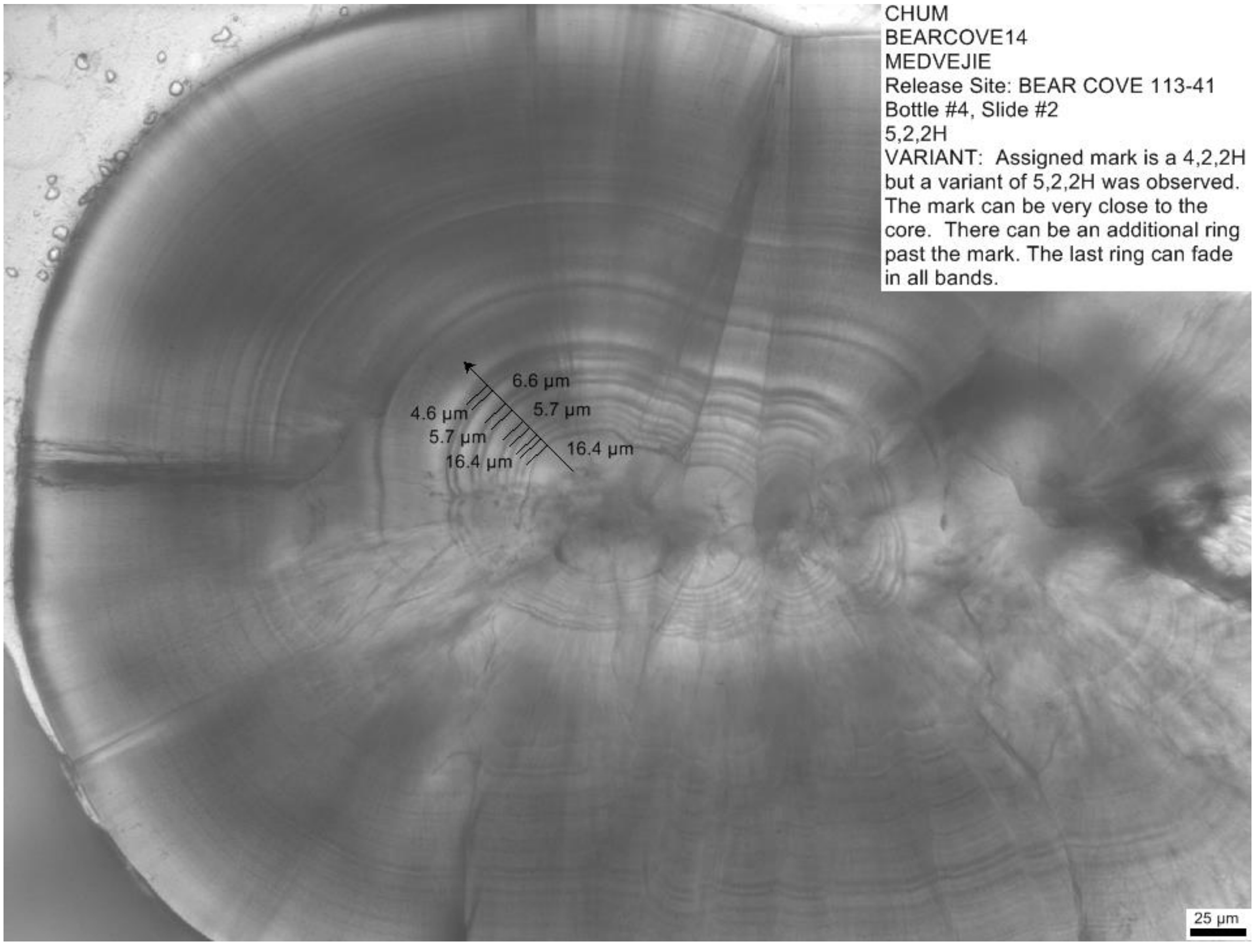


Figure 2. Thermal mark image with measurements shown in the Mark Characteristic Report. This figure shows a 3,5nH4 mark with a pre- and post-hatch mark. Thus this mark has two bands prior to hatch (the first with three rings and the second with 5 rings) and one band after the hatch containing 4 rings. The individual rings are the dark lines in each band, and in the second band, the spacing among the rings is narrower than that in the other bands so the 5 is followed by an “n.”



CHUM
BEARCOVE14
MEDVEJIE
Release Site: BEAR COVE 113-41
Bottle #4, Slide #2
5,2,2H
VARIANT: Assigned mark is a 4,2,2H but a variant of 5,2,2H was observed. The mark can be very close to the core. There can be an additional ring past the mark. The last ring can fade in all bands.

Figure 3. Image of a thermal mark variant of a reference specimen. This figure shows another image of Figure 1, thermal mark ID BearCove14. This fish, assigned a target thermal mark of 4,2,2H, which indicates that the first band from the otolith core contains four dark rings, a space, then a band with two rings, a space, and a band with two rings followed by the hatch mark (the blurry, wider, dark area). Instead, this otolith shows a 5,2,2H or a variant, meaning that the first band has five rings instead of the planned four rings.

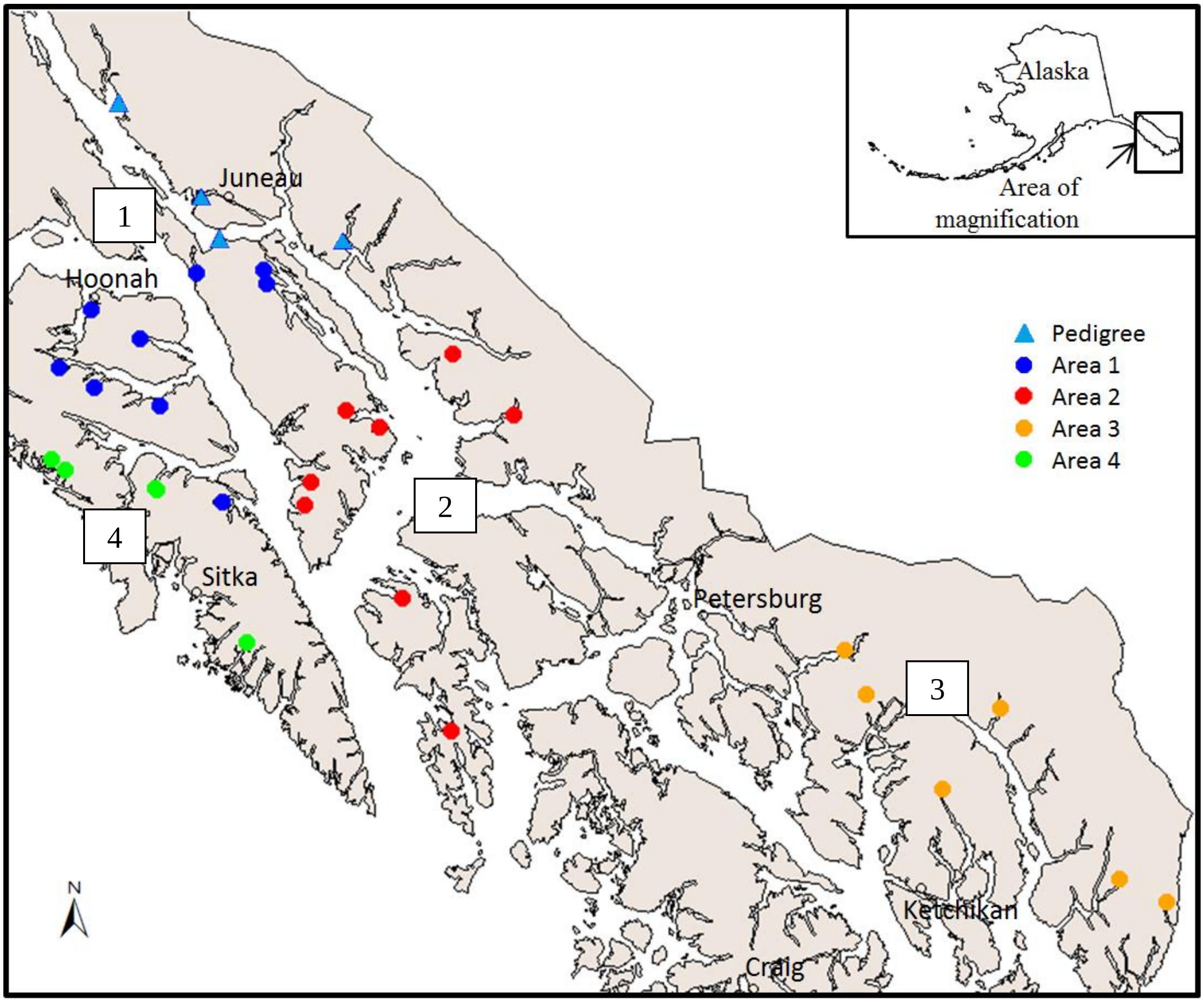


Figure 4. Four strata used for assessing the accuracy of thermal mark readings of chum salmon otoliths recovered from streams in Southeast Alaska during 2013 and 2014 for the Alaska Hatchery Research Project**.**

Tables

Table 1. Example matrix comparing thermal mark reader agreement. Row and column names represent potential thermal marks identified by each reader (TM1 through TM6), otoliths classified as wild, and otoliths classified as unreadable (ND). The number of otoliths where both readers agree is in bold font along the diagonal between the row and columns.

	2nd Reads								
1st Reads	TM 1	TM 2	TM 3	TM 4	TM 5	TM 6	Wild	ND	Total
TM 1	**0**	1							1
TM 2	1	**12**							13
TM 3			**34**						34
TM 4			2	**9**			6		11
TM 5					**26**				26
TM 6						**4**			4
Wild			3				**357**	1	358
ND							1	**3**	4
Total	1	13	36	9	26	4	358	4	451