Feasibility of a Comprehensive Database and Analysis System for Rotary Screw Trap Data in the California Central Valley

Pacific States Marine Fisheries Commission
205 SE Spokane Street
Portland, OR 97206

April 28, 2010

Funded by the U.S. Fish and Wildlife Service's
Comprehensive Assessment and Monitoring Program (CAMP)
USFWS Cooperative Agreement No. 81420-8-J163

TABLE OF CONTENTS

Executive Summary

The U.S. Fish and Wildlife Service's (USFWS) Comprehensive Assessment and Monitoring Program (CAMP) was created pursuant to the Central Valley Project Improvement Act (U.S. Fish and Wildlife Service 2008). CAMP produces a variety of reports that summarize and tabulate Chinook salmon (*Oncorhynchus tshawytscha*) data from sources in California's Central Valley. To prepare certain of these reports, in-depth statistical analyses and the development of complex databases are required. Through a cooperative agreement, CAMP funded the Pacific States Marine Fisheries Commission (PSMFC) and a statistical subcontractor (Western EcoSystems Technology, Inc.) to assist in evaluating the feasibility of developing a comprehensive data storage and analysis system for Chinook salmon data collected with rotary screw traps (RSTs) in the Central Valley.

Because the development of such a system is inherently challenging, CAMP determined that a phased approach was appropriate, beginning with a feasibility evaluation as the first of three or more phases. In this Phase I feasibility evaluation field methods from several RST studies (as detailed in annual progress reports), the data capture procedures and databases currently in use, and current data analysis routines were examined. This first phase was intended to determine if a central database, common analysis procedures, and user interfaces could be created for use with data that have already been collected in the Central Valley. Phase I was meant to be a low-cost review leading to an indication of the likelihood of success of later phases. This Phase I report addresses the feasibility of developing a comprehensive central database that will meet CAMP's needs for capturing existing data from all sources and providing data analysis routines to produce defensible estimates of juvenile Chinook salmon production in the California Central Valley both now and in the future.

Nineteen individuals who participate in or are knowledgeable of RST studies in the Central Valley were contacted in regard to their knowledge of RST operations and data available.

Field method and computer system examinations began with on-site visits to three RST operations run by staff from the USFWS Red Bluff Fish and Wildlife Office (USFWS-RB), where USFWS-RB fisheries biologists demonstrated field procedures, provided information, and answered questions about field methods, data input, data management, and data analysis procedures. In addition to the on-site visits, field methods as described in the most recent annual reports for 20 of the RST operations in the Central Valley that target Chinook salmon were reviewed to determine the degree of similarity in field and statistical methods employed.

The structure and capabilities of existing databases, and their associated user interfaces and analysis procedures, were evaluated from the perspective of providing the functionality required for CAMP's needs.

Seven databases were obtained and their structures (tables, table relationships, fields, codes) examined. In addition to examining each database individually, comparisons were made among databases structures — similarities and differences in tables, table relationships, fields, and codes — to determine the degree of standardization already implemented and the level of difficulty

expected in attempting to compile data in a single database. As well, the data in these databases were examined for suitability to meet CAMP's needs.

In general, almost all studies appeared to use similar field methods and collected similar information (except for capture efficiency data, which were not collected in some studies). All of the reports lacked some information of interest.

Field methodologies employed by the various Central Valley RST studies were evaluated in relation to three categories of complications that could influence CAMP's ability to utilize RST data:

1. Potential complications in capturing data in a central database
2. Potential complications in performing statistical analyses on the data
3. Potential complications in interpreting the statistical analyses

Most field methods described were straight forward and should generate data that can be captured fairly easily in a standardized data structure. Only a few field methods were identified that could cause complications for capturing data in a central database. None of these were considered a serious impediment to capturing data in a central database, although the total of all variations could present moderate difficulties.

The field visits and literature review identified several field methods that may complicate statistical analyses of the data collected. These included changes in trapping site or RST position at a site, use of variable numbers of traps at a site, RSTs not run continuously through the sampling season, lack of appropriate capture efficiency tests, and variations on environmental covariates measured.

Several field methods were identified that may be problematic for interpretation of statistical tests. These included using non-target fish (hatchery origin or of a different run) for capture efficiency tests, and the release of marked test fish in one study at non-optimal distances above the trap site for capture efficiency tests.

An examination of the database structures revealed that the databases were similar due to a common origin, but not identical. All databases were incomplete from the perspective of capturing all information necessary for production estimates. Important information was missing, including such basic items as the stream name where sampling occurred.

Little specific information was gathered regarding the data analysis routines and tools used at the various RST operations. It appeared that capture efficiency data were entered into spreadsheets rather than in the databases that house catch data. However, the full data set necessary for calculating Chinook salmon production was not provided by any of the entities conducting RST studies in the Central Valley.

The evaluation by Shannon (2009) of the nine data sets obtained found that several types of shortcomings were common. Most significantly, the full set of data necessary to calculate production estimates was not provided for any of the data sets. While catch data were available in some cases, the suite of data necessary to calculate capture efficiency were not included in any database or were very incomplete. For those databases which contained some mark and release information, these data were very incomplete, and did not indicate how many fish were released for capture efficiency tests. For several of the databases, deriving release data may be possible with significant effort. Linkages between catch data and capture efficiency data were not explicit within any of the databases.

This evaluation uncovered little that should cause significant difficulties for compiling Chinook salmon RST catch data into a single database for the Central Valley. However, since the objective of this Phase I analysis is to evaluate the feasibility of estimating juvenile production from data that were sometimes collected for other purposes, it is not surprising that a number of data related limitations and concerns were identified.

Of greatest concern are the 5 sites that did not include calculation of production and therefore no capture efficiency tests were conducted (Big Chico Creek, Butte Creek, Deer Creek, Mill Creek, and Yuba River). If these studies continue in the future, it may be possible to calculate capture efficiencies that could be applied retroactively to estimate past production, but the accuracy of those estimates would be dependent on consistent trap placement and a determination that the efficiency estimates are consistent year to year at those sites. Without such consistency, it will not be possible to calculate juvenile Chinook salmon production accurately at these five sites.

Several field methods were discovered that may rely on problematic assumptions, which could bring the accuracy of statistical results into question. Accepting the inherent assumptions of these methods, such as using a different run of fish for capture efficiency tests when the target run of juvenile salmon is rare, or mixing fish from different runs, may be necessary and acceptable for the original purposes of the studies, but when trying to use these data for a different purpose they may present problems if the new purpose requires more stringent assumptions. The CAMP program will need to evaluate their needs relative to the quality and completeness of RST data available before deciding to move forward. It appears over half of the studies evaluated used non-target fish for capture efficiency tests.

The key question addressed in this report is whether a single database can be created to capture the data from all Chinook salmon RST studies in the California Central Valley, and then provide these data in a format suitable for calculating statistically-derived estimates of Chinook salmon production. Although there would be challenges, this is feasible from the computer technology perspective. The limiting factor is the availability of appropriate data from RST studies that are complete enough, comprehensible enough, and of the right format to make the effort worthwhile. Whether data of sufficient quality and completeness exist for each of the RST studies can only be answered by obtaining and examining the data sets, accompanied by a more thorough examination of field methods. But based on the findings here and by Shannon (2009), it is apparent that calculating juvenile Chinook salmon production estimates is only feasible for some sites in the Central Valley.

If it is decided to move forward, then creation of a database, or selection of an existing one, will be a primary concern. The existing Central Valley databases examined were not sufficient. Their shortcomings are great enough that a wholly new database should be designed or a different existing database be found, rather than trying to modify an existing one. Several alternatives exist. A benefit of adopting an existing database is the ability to more easily share data with other programs along the West Coast. The components of a data storage and analysis system are listed, and recommendations are provided for software selection if further efforts are pursued.

Based on information acquired during this review, fewer than half of the studies generated capture efficiency data appropriate for CAMP's needs, and none of them provided for this review the complete raw data needed for production estimates. Without appropriate capture efficiency data, recalculating production estimates may not be feasible.

A remaining challenge, should this effort continue, will be to cultivate the support of biologists to participate in this effort to the degree necessary for them to provide their data and assist with understanding them well enough for capturing them in the central database.

Introduction

The U.S. Fish and Wildlife Service's (USFWS) Comprehensive Assessment and Monitoring Program (CAMP) was created pursuant to the Central Valley Project Improvement Act (U.S. Fish and Wildlife Service 2008). CAMP produces a variety of reports that summarize and tabulate Chinook salmon (*Oncorhynchus tshawytscha*) data from sources in California's Central Valley. To prepare certain of these reports, in-depth statistical analyses and the development of complex databases are required. Through a cooperative agreement, CAMP funded the Pacific States Marine Fisheries Commission (PSMFC) and a statistical subcontractor (Western EcoSystems Technology, Inc.) to assist in evaluating the feasibility of developing a comprehensive data storage and analysis system for Chinook salmon data collected with rotary screw traps (RSTs) in the Central Valley.

The main purpose of such a system would be for CAMP to document and understand changes in the catch (number of fish caught) and production (estimated number of fish migrating downstream past a specific point in a stream) of juvenile Chinook salmon in the Central Valley among years, among time periods within a year, and among locations. One of CAMP's goals is to assess the relative effectiveness of various categories of habitat restoration actions that are implemented to increase the number of naturally produced Chinook salmon in the Central Valley, and juvenile production data are a potentially rich source of information for these assessments.

Down-migrating juvenile Chinook salmon were studied using RSTs for at least one season at no fewer than 26 sites in the Central Valley since 1991 (Table 1). Currently trapping occurs in 12 watersheds, and at more than one location in several watersheds. Trapping at some of these sites has been conducted for as long as 18 years, but using the results of these studies is often difficult for a variety of reasons. Analyses from many of the trapping efforts have never been presented in report form. Project objectives and field methods varied among studies. Data were stored in different formats. Different analytical techniques were used to estimate production when that was a project objective. At some locations capture efficiency tests were not conducted, making it impossible to calculate production estimates for those locations. Separate or non-existent reports and different analytical techniques make it difficult, if not impossible, to understand valley-wide long-term trends in juvenile salmon production, and confound CAMP's ability to understand how restoration activities have influenced juvenile and adult salmon production.

Table 1. Summary information for rotary screw trapping operations that have occurred in the Central Valley since 1991. USFWS=U.S. Fish & Wildlife Service; CDFG=California Department of Fish and Game.

Watershed	**Runs of Chinook salmon present**	**Affiliation**	**Years when trapping is known to have occurred**
American River	Spring, Fall	CDFG, Rancho Cordova	1993-2007
Battle Creek (2 sites)	Spring, Fall, Late fall, Winter	USFWS, Red Bluff	1998-present

Big Chico Creek	Spring, Fall	CDFG, Chico	1998-2003
Butte Creek (2 sites)	Spring, Fall	CDFG, Chico	1995-2008
Clear Creek (2 sites)	Spring, Fall, Late fall, (Presence of winter run uncertain)	USFWS, Red Bluff	1998-present
Cosumnes River	Fall	CDFG, Rancho Cordova	1999
Deer Creek	Spring, Fall	CDFG, Red Bluff	1994-present
Feather River (2 sites)	Spring, Fall, Late fall	California Dept. of Water Resources, Oroville	1998-present
Merced River (Hagaman State Park)	Fall	CDFG, La Grange	1998-2002
Merced River (Hatfield State Park)	Fall	Cramer Fish Sciences, Oakdale	2007-2009
Merced River (near Hopeton)	Fall	Natural Resource Scientists, Inc., Red Bluff	1999-present
Mill Creek	Spring, Fall	CDFG, Red Bluff	1995-present
Mokelumne River	Fall	East Bay Municipal Utility District, Lodi	1993-present
Sacramento River (Glenn-Colusa Irrigation Diversion Fish Screen Facility)	Spring, Fall, Late fall, Winter	CDFG, Hamilton City	1991-2008
Sacramento River (Knights Landing)	Spring, Fall, Late fall, Winter	CDFG, Sacramento	1995-present
Sacramento River (Red Bluff Diversion Dam)	Spring, Fall, Late fall, Winter	USFWS, Red Bluff	1994-present
Stanislaus River (Caswell State Park)	Fall	Cramer Fish Sciences	1994-present
Stanislaus River (Oakdale)	Fall	FISHBIO, Environmental, Oakdale	1993-present
Tuolumne River (Grayson Ranch)	Fall	FISHBIO Environmental, Oakdale	1995-present
Tuolumne River (Shiloh Bridge)	Fall	CDFG, La Grange	1995-1998
Tuolumne River (near Waterford)	Fall	FISHBIO Environmental, Oakdale	2006-present
Yuba River	Spring, Fall, Late fall	CDFG, Rancho Cordova	1999-present

For CAMP to address the difficulties inherent in trend detection under these circumstances, a single data storage and analysis system would be needed to consolidate the previously collected RST catch data and what capture efficiency data exist. The ability to enter newly collected RST catch and efficiency data into such a system would ensure that data collected in the future are compatible with and can be analyzed with existing data. This system, if built, would be designed to store and manage RST data from across the Central Valley, as well as produce statistically robust and repeatable estimates of juvenile Chinook salmon production where complete data sets — catch and efficiency — exist.

Because the development of such a system is inherently challenging, CAMP determined that a phased approach was appropriate, beginning with a feasibility evaluation as the first of three or more phases. In this Phase I feasibility evaluation field methods from several RST studies (as detailed in annual progress reports), the data capture procedures and databases currently in use, and current data analysis routines were examined. This first phase was intended to determine if a central database, common analysis procedures, and user interfaces could be created for use with data that have already been collected in the Central Valley.

Rather than an in-depth analysis of all possible issues that might be encountered in subsequent phases, Phase I was meant to be a low-cost review leading to an indication of the likelihood of success of those later phases. Phase II would include design and creation of a central database that could capture existing and future data from all sources, and then compilation of RST data from the various sources into this database. To the extent possible, data conversion routines would be created for loading existing data into the new database, and these routines could serve when capturing future data as well. During Phase III a data analysis routine would be written to produce consistently-generated, statistically robust production estimates. If it was determined in Phase I or II that producing a system to meet CAMP's needs was impractical then Phase III would not proceed. Field databases and interfaces could be built for use by the data collectors as a Phase IV of the project. The user interfaces would be for data entry and editing. Routines and procedures for sharing data with the central database would also be created.

This Phase I report addresses the feasibility of developing a comprehensive central database that will meet CAMP's needs for capturing existing data from all sources and providing data analysis routines to produce defensible estimates of juvenile Chinook salmon production in the California Central Valley both now and in the future.

Definitions

Numbers of fish

<u>Catch</u>
The number of fish caught at a trapping site over a defined period of time — from as little as a day to as long as a year.

Production
An estimate of the total number of fish that moved downstream past a trapping site over a defined period of time. This time period is usually long — e.g. a year, season, or month.

Index of abundance
A measure used to detect trends in the number of fish that moved downstream past a trapping site over a defined period of time, but which is not an actual estimate. The time period of an index of abundance may be as short as a day or as long as a year.

Relative abundance
Comparisons of the index of abundance between time periods.

While all these terms relate to the number of fish, in this document the term "catch" is used for a measured datum; "production" is a statistically-derived estimate of a parameter and has associated confidence intervals; "index of abundance" is a simple arithmetically derived value meant to summarize and standardize the catch values, but is not rigorously derived as is the production estimate.

Production is derived from catch and estimated capture efficiency. The catch alone, perhaps standardized to account for changes in the number or configuration of traps at a site, is used as the index of abundance.

Efficiency

The efficiency of trapping is measured using two related terms.

Capture efficiency
The probability that a fish will be captured as it moves downstream past a sampling site.

Trap efficiency
The probability that a fish will be captured in a particular trap as it moves downstream past a sampling site.

Capture efficiency is equal to trap efficiency when a single trap is used at a site. When multiple traps are used at a site, capture efficiency is affected by the efficiency of all traps at the site.

Back end; front end

These are computer programming terms used to characterize program interfaces and services relative to the user. In this document, the user is a human being interacting with a computer.

Front end

A computer program (application) which the user interacts with for data input, data quality control, data analysis, or other purpose. The front end displays the boxes, buttons, menus, and other features of programs which people are familiar with.

Back end

A computer program that the user does not generally interact with directly. Rather, the user interacts with the back end indirectly through the front end application. For the purposes of this paper the back end is a relational database management system providing data capture, storage, management, and retrieval services. The back end stores the data.

Central database; field database

Central database

A database used to house data obtained from multiple other databases.

Field database

A database used by field workers to capture data in electronic form.

The central database houses data contributed by the field databases. For both the central and field databases, it is assumed that a relational database management system will be employed.

Normalized; denormalized

Normalized

A relational database design term used to indicate that data are in separate two-dimensional tables which interact with each other through parent-child relationships. Normalization of data is done to reduce file size, to ease data management, to provide consistency in data, and particularly to prevent duplication and other types of data integrity errors.

Denormalized

A relational database design term generally used to indicate that data are in a single two-dimensional table. Data stored in a normalized database are usually queried into a denormalized form for use in a spreadsheet.

Methods

Contacts

The following individuals were contacted by Shannon (2009) or by the author in regard to their knowledge and expertise related to RST operations in the Central Valley:

- Clint Garmin (California Department of Fish and Game)
- Colleen Harvey Arrison (California Department of Fish and Game)
- Dennis Blakeman (California Department of Fish and Game)

- Douglas Burch (California Department of Fish and Game)
- Michael Healey (California Department of Fish and Game)
- Robert Vincik (California Department of Fish and Game)
- Jason Kindopp (California Department of Water Resources)
- Ayesha Gray (Cramer Fish Sciences)
- Clark Watry (Cramer Fish Sciences)
- Chrissy Sonke (Cramer Fish Sciences)
- Robyn Bilski (East Bay Municipal Utility District)
- Duane Massa (Pacific States Marine Fisheries Commission)
- Jennifer Bergmen (Pacific States Marine Fisheries Commission)
- Douglas Threloff (USFWS-Sacramento, CAMP Program Manager)
- James Earley (USFWS-Red Bluff)
- Kellie Whitton (USFWS-Red Bluff)
- William Poytress (USFWS-Red Bluff)
- Michelle Workman (formerly with East Bay Municipal Utility District, now with USFWS)
- Elizabeth Cook (formerly with California Department of Water Resources), who organized the Interagency Ecological Program's Bay-Delta and Tributaries project (BDAT) database effort and created the BDAT RST field databases in use at many operations in the Central Valley.

Examination of field methods

Field method examinations began with on-site visits December 15 and 16, 2008 to three RST operations run by staff from the USFWS Red Bluff Fish and Wildlife Office (USFWS-RB). Staff from PSMFC (fisheries biologist; computer programmer), USFWS Pacific Southwest Region in Sacramento (CAMP Program Manager), and Western EcoSystems Technology, Inc. (statistician/statistical computer programmer) met on-site with five USFWS-RB fisheries biologists who demonstrated field procedures, provided information, and answered questions about field methods, data input, data management, and data analysis procedures. Sites visited were Battle Creek, Clear Creek, and the Red Bluff Diversion Dam on the Sacramento River. At each site the general layout of the drainage basin, stream, and trapping site were discussed, as were general observations about trapping at that site, and the fish collected.

In addition to the on-site visits, field methods as described in the most recent annual reports for 20 of the RST operations in the Central Valley that target Chinook salmon were reviewed to determine the degree of similarity in field and statistical methods employed. Several reports referred to earlier years' reports for field method details; these reports were reviewed also but were considered part of the same review and thus are not enumerated. To the degree possible the field methods used at each site were determined. Reports were reviewed for RST studies on the following streams:

- American River
- Battle Creek (2 sites)
- Big Chico Creek
- Butte Creek

- Clear Creek (2 sites)
- Cosumnes River
- Deer Creek
- Feather River (2 sites)
- Merced River at Hatfield State Park
- Mill Creek
- Mokelumne River
- Sacramento River at Knights Landing
- Sacramento River at Red Bluff Diversion Dam
- Stanislaus River (2 sites)
- Tuolumne River at Grayson ranch
- Yuba River

Examination of existing computer systems

The structure and capabilities of existing databases, and their associated user interfaces and analysis procedures, were evaluated from the perspective of providing the functionality required for CAMP's needs. If an existing system 1) was able to capture all necessary data and store them in an appropriate structure for use, and 2) had the necessary data reporting and analysis capabilities, then it could potentially be used as is, or modified, to meet CAMP's needs for a valley-wide database.

During the December 2008 on-site visit, a data input front end used at Battle and Clear creeks was demonstrated. The back end was a Microsoft Access database developed by the Interagency Ecological Program (IEP) for the Bay-Delta and Tributaries project (BDAT) system. The front end was a Microsoft Access database file with user interface forms and Visual Basic for Applications programming providing additional functionality. Also demonstrated was the ability to query data from the database for transfer to a spreadsheet, where data analysis was performed. The back end and front end databases were later obtained and examined directly. The programming code's documentation indicated the front end and back end were designed and created by the California Department of Water Resources for the IEP's BDAT system.

Also during the December 2008 on-site visit, USFWS-RB personnel demonstrated their data analysis procedures for the Sacramento River - Red Bluff Diversion Dam study. A minimum of 27 steps were used to produce the production estimates which were shared with other agencies in the basin via the BDAT database and web site. This sophisticated process involved downloading stream discharge data from a U.S. Geological Survey web site, data quality checks, and other procedures necessary to create the final production estimate.

The front end and back end database used for the Merced and Stanislaus rivers were obtained from Cramer Fish Sciences, and those used for Clear and Battle creeks were obtained from USFWS-RB. These front and back ends were designed and created by the California Department of Water Resources, with refinements made by Cramer Fish Sciences or USFWS-RB, respectively. Along with the Merced/Stanislaus and Clear/Battle databases, another five databases were obtained and examined by Shannon (2009) (Sacramento River at the Red Bluff

Diversion Dam, Mokelumne River at Woodbridge, Tuolumne River at Grayson Ranch, Tuolumne river at Shiloh Bridge, and Merced River at Hatfield State Park).

The seven databases obtained were examined, and the following goals were pursued for each:

- Understand the purpose of each table
- Understand the functional relationships between tables
- Determine the specific field(s) used to create table relationships
- Inventory the list of fields in each table
- Inventory the key fields in each table
- Inventory the codes used for each coded field (lookup codes)
- Inventory the types of data populating each database (i.e., catch data; capture efficiency data; environmental covariate data).

In addition to examining each database individually, comparisons were made among databases — similarities and differences in tables, table relationships, fields, and codes — to determine the degree of standardization already implemented and the level of difficulty expected in attempting to compile data in a single database.

Examination of existing data

The Battle/Clear creek and Merced/Stanislaus river databases were examined; Shannon (2009) examined these two databases plus the other five databases she obtained. In addition, the BDAT web site (http://www.bdat.ca.gov) was queried on December 17. 2009 and the full set of RST data that had been submitted to that system was acquired. For each of these databases an initial assessment of the data contents was conducted (2 databases plus BDAT query output during this investigation; seven databases by Shannon (2009)), with an emphasis on determining the completeness of the available data for purposes of producing production estimates. These examinations involved determining whether functional catch data were present and complete, whether functional capture efficiency data were present and complete, and the types of environmental condition data included. Catch data included taxon, rearing type, number of fish, date, etc. Capture efficiency data included taxon, rearing type, fish size, type and location of mark applied, number of marked fish released, number of marked fish recaptured, type and location of marks on recaptured fish, targeted trap for the efficiency tests, dates, etc. Environmental condition data included such things as stream discharge, water temperature, and turbidity.

A sample of lookup tables was briefly compared among databases to determine if the codes and values used were identical, and whether the codes employed would allow for easily combining data in a single database.

Results

Evaluation and comparison of field methods

The field method information in most of the RST annual reports included the purpose(s) of the trapping operations, physical trap operations and their consistency over time, seasonal and weekly timing of trap operations, whether and how capture efficiency tests were conducted, and the environmental conditions measured. Specific results of this evaluation are available in Appendix A. In general, almost all studies appeared to use similar field methods and collected similar information (except for capture efficiency data, which were not collected in some studies). All of the reports lacked some information of interest, however.

Field methodologies employed by the various Central Valley RST studies were evaluated in relation to three categories of complications that could influence CAMP's ability to utilize RST data:

1. Potential complications in capturing data in a central database
2. Potential complications in performing statistical analyses on the data
3. Potential complications in interpreting the statistical analyses

Potential complications discovered within these categories are discussed below, along with explanations where necessary for why they are problematic.

<u>Field methods: potential complications in capturing data in a central database</u>
Most field methods described were straight forward and should generate data that can be captured fairly easily in a standardized data structure. Only a few field methods were identified that could cause complications for capturing data in a central database.

The field visit confirmed that the Battle and Clear creek efforts were in general typical small stream RST operations with no major complications. The only unexpected complication encountered at these locations was the intermittent use of the "half cone configuration" (Appendix A). The term "half cone configuration" refers to a physical modification to the RST so that one half of it is made nonfunctional, returning fish and debris from one half of the trap directly to the river rather than into the live box. A central database would likely need one additional field to capture information regarding when each configuration was in use.

The reports reviewed indicated that most studies were similarly typical, except for five studies that employed a variable number of traps (Appendix A). One of these five, the Sacramento River at the Red Bluff Diversion Dam, was the most atypical operation encountered. Along with a variable number of traps (either three or four RSTs were run concurrently), the traps were moved laterally across the dam and upstream/downstream as river discharge and dam operations changed. In addition, the half cone configuration was employed at this site. Taken together, these created a matrix of RST configuration information that will require a more sophisticated database design than is needed for the simpler cases.

Two other field methods were identified that create uncertainties for capturing data in a central database due to insufficient metadata. First, when a trap ran less than a full day various methods were used to account for the time when the trap was not functioning (Appendix A). A second, similar issue existed when large numbers of fish were caught and subsampling was conducted. When compiling data into a central database it will be necessary to know if these issues are already accounted for in the source databases. To determine whether each field database contains raw or adjusted data will require contacting each database owner. For older data sets, or in other cases such as where staff turnover has occurred, it may be difficult or impossible to answer these questions.

Field methods: potential complications in performing statistical analyses

The field visits and literature review identified several field methods that may complicate statistical analyses of the data collected.

CHANGED TRAP LOCATIONS: The trapping site or trap position within the site changed at several of the operations, either between years or within a season. At Merced River the stream channel changed in 2007, resulting in the trap being moved 40 m from its original placement. At Tuolumne River the trapping site was changed from river mile 3.4 to river mile 5.2. At Deer Creek, the two Feather River sites, Mill Creek, and Yuba River it was not possible to determine if sampling sites changed. At Butte Creek trap placement was frequently adjusted within season. At Sacramento River at the Red Bluff Diversion Dam the traps were moved laterally and upstream/downstream within season in response to river discharge and dam operations. Most of the reports included no indications that efficiency estimates were segregated or stratified by trap site or position.

Other factors were encountered that may also affect capture efficiency estimates and thus complicate statistical analyses.

MULTIPLE TRAPS: At American River, Feather River (lower site), Merced River, Sacramento River at Knights Landing, Sacramento River at Red Bluff Diversion Dam, Stanislaus River at Caswell State Park, Tuolumne River at Grayson Ranch, and Yuba River multiple traps were used. At any site where multiple traps are employed — not only in the Central Valley — the number of traps employed often does not remain constant within or across seasons due to deliberate addition or subtraction of traps, or due to equipment malfunctions. Multiple traps are used at a site to increase capture rate for the site as a whole. Thus varying the number of traps changes the capture efficiency for the site as a whole. For each of these sites data analysis complications can be expected due to changes in the number of traps employed and the resultant effects on capture efficiency. It will be critical to know how capture efficiency tests were conducted in relationship to changes in trap number, and how this information is stored in the databases.

LIMITED TRAP OPERATION: At American River, Mokelumne River, and Stanislaus River at Oakdale, traps were not run seven days per week. This may have resulted in missing marked fish that would have otherwise been caught for capture efficiency tests, as well as requiring that data be imputed for the unsampled days.

LACK OF CAPTURE EFFICIENCY TESTS: Capture efficiency tests were not conducted at five sites. At Cosumnes River calculating a production estimate was a stated objective, but no efficiency tests were conducted.

At Battle Creek, Clear Creek, and the Sacramento River at Red Bluff Diversion Dam the half cone configuration was intermittently employed. Thus the capture efficiency tests conducted do not apply to all capture data without modification, and standardization of measures is required. For periods when the half cone configuration was employed, capture efficiency was assumed to be one half the capture efficiency of normal configuration (Whitton et al. 2008), though biologists at these sites indicated this assumption may not be valid and needed to be tested (Jim Earley, USFWS-RB, pers. comm.).

At Sacramento River at Knights Landing, water is diverted into the Sutter Bypass when discharge exceeds roughly 23,000 cfs. Under these higher flows an unknown proportion of migrating juvenile Chinook salmon may be entrained into the Sutter Bypass flow and diverted around the Knights Landing screw trap operation, thus becoming unavailable for capture in the screw trap (Vincik et al 2006). Capture efficiency under this condition was not determined.

At Yuba River only two efficiency tests were conducted each month, resulting in little information available to build a capture efficiency model. It could not be determined if capture efficiency tests were conducted at Tuolumne River.

Integrating catch and efficiency test data may be challenging in all of these cases.

In addition to the physical sampling issues that may confound determination of capture efficiency, other factors were encountered that may complicate analyses.

NONREPRESENTATIVE CAPTURE EFFICIENCY TESTS: At Clear Creek capture efficiency tests were only run during low to intermediate stream discharge. Therefore capture efficiency at higher flows must be extrapolated beyond the range of observed values. Though the other reports did not specifically address this, this condition is very common for RST studies and probably most of the other studies also have experienced this.

CORRELATED DATA: Stream discharge, or a correlated measure, was recorded at every operation, including those which appeared to have no use for such a measurement (i.e., capture efficiencies were not measured, and the relationship between catch and environmental factors was apparently not determined). Several of these correlated measurements were recorded at some operations. Stream discharge correlates included river stage, water depth at the trap, water velocity at the trap, trap rotation rate (which is dictated by water velocity), and portion of stream discharge sampled. Other measures collected by some Central Valley studies that are influenced by or correlated with stream discharge were amount of debris caught in a trap, turbidity, Secchi disk depth, and conductivity.

OTHER: Many of the differences in methods among sites were likely of minor significance in relation to the ability to collect and analyze the data. Examples included differences in units, in how fish runs were assigned, and in how the expected number of fish was estimated when a trap did not operate for a full day. These can be recalculated from the raw data, and should not be a significant issue.

Field methods: potential complications in interpreting statistical analyses

The capture efficiency tests appeared to be problematic in several of the studies from the perspective of interpretation of the statistical tests.

Twelve sites were reviewed where capture efficiency tests were reportedly conducted. At four of those sites the fish used for capture efficiency tests were at least partially of hatchery origin. Where multiple runs of Chinook salmon are found in the same stream, a different run or a mixture of runs was frequently used. It is likely that for the majority of sites, the fish used for capture efficiency are at least partially non-target fish -- hatchery origin or of a different run. Non-target fish may not represent the capture efficiency of the targeted fish because of different migration patterns laterally across the stream or in depth of travel.

At least one study (the Mokelumne River) released marked test fish at what may be non-optimal distances above the trap site for capture efficiency tests (only 100-500 feet). Optimal distance above a trap for releasing marked fish is not easily determined. Volkhardt et al. (2007) suggested a minimum of 2 riffle/pool sequences, but not so far upstream that predation becomes significant. An optimal distance allows the released fish to redistribute themselves naturally across the stream channel before encountering the trap, without experiencing significant mortality. This distance is unique for each trapping site and may vary with stream discharge, and in the absence of a detailed local study can usually only be guessed at.

Appendix A contains more detailed results for each trap location from the review of the field methods disclosed in the reports.

Evaluation and comparison of existing computer systems: databases

Obtaining databases proved difficult, partially due to time constraints of biologists, and partially due to staff turnover after a study ceased. With the assistance of the CAMP Program Manager two databases were obtained. One was from the USFWS's Red Bluff Fish and Wildlife Office and is used for Battle and Clear creeks (Figure 1). The other was from Cramer Fish Sciences, and is used for the Stanislaus and Merced rivers (Figure 2). Shannon (2009), also with the CAMP Program Manager's assistance, was able to obtain an additional five databases.

The fisheries database created for the BDAT project was discovered during the course of this evaluation. The BDAT fisheries database was from an earlier effort to consolidate and share fisheries and water quality data from the Central Valley, including data from RST studies. (Further information about the BDAT project and the BDAT database can be found at http://bdat.ca.gov.)

Several biologists were reluctant to provide their database, but instead stated that their data were available from the BDAT web site. They stated the BDAT database was the same database they used, and therefore their database structures could be determined by examination of the BDAT database. However, it appears informal use of terms resulted in a misunderstanding of the BDAT central database and the field databases constructed to contribute data to it. Elizabeth Cook (formerly with the California Department of Water Resources (CDWR)) provided the correct context for the BDAT central database and associated field databases. The BDAT database was a central database managed by CDWR that collected data from many sources and could be queried via the BDAT web site. Individualized field databases were created by CDWR for each RST operation. Field databases were used to enter and manage RST data at each site and send data to the BDAT system. However, while based on the BDAT data model, each field database was unique and none was identical to the central BDAT database. Further, the field databases were apparently modified by the biologists running the RST operations. Thus data in the field databases cannot simply be copied into a central database.

It was not possible to obtain a copy of the BDAT central database, either in its native format or converted to a Microsoft Access format, due to staff turnover at the agency where this database resided. Therefore the BDAT online query system was used to obtain data from all sites so that data contents and structure could be examined. This brought to three the number of data sources available for this initial evaluation of databases, each a modified version of the BDAT database. The BDAT website, however, provided a denormalized version of the data, so database design details could not be determined for that data source. The list of fields provided by the BDAT web site is shown in Figure 3. Errors were encountered in the data downloaded from the BDAT web site that reflect on the underlying BDAT database: an apparent lack of strong data typing (text strings were found in one ostensibly numeric field), and lack of controls preventing duplicate data (resulting from the lack of an appropriate key field) allowed these errors to exist.

A cursory comparison of the entity-relationship diagrams for the Battle Creek/Clear Creek (Figure 1) and Stanislaus River/Merced River databases (Figure 2) showed similarities between the two databases due to their common origin. Both databases had similar tables named StationsLookup, MethodsLookup, GearDetailsLookup, Sample, TrapEffort, Catch, OrganismLookup, StagesLookup, and MarksLookup. These tables had many fields in common in the two databases, and the relationships between these tables were nearly the same. A closer examination, however, showed that none of these tables had the same list of fields in both databases, that the relationships between the tables were not always on fields with the same names, that the same information was stored in different field names in the two databases, that the same field name may have been used to house different information in the two databases, that the lookup codes and values often did not match, and that each database had fields added that did not occur in the other database. Thus, while at first glance these databases appeared to be quite similar, some differences were found that in practice may be difficult to reconcile. Shannon's (2009) review of these two plus five additional databases revealed the same pattern of databases with similar, but not identical, information. An example is shown in Table 2, where the manner in which weather was characterized is shown for several of the databases examined. The fields where these issues occurred were in covariates useful for refining capture efficiency estimates — measures of water clarity, descriptions of weather conditions, characterizations of debris gathered by RSTs, and salmon life stage. Major data elements such as taxon and number caught, though in different field names, were not significantly different and should be easily combined.

Table 2. Manner in which weather conditions are characterized in several of the databases reviewed.

Database	Field Name	Values Used
Clear / Battle creeks	WeatherCode	• CLD • CLR • FOG • PCLD • RAN
Sacramento River at Red Bluff Diversion Dam	WeatherCode	• CLD • CLR • FOG • RAN • W • <null>
Stanislaus / Merced rivers	WeatherCode	• CLD • CLO • CLR • FOG • nd • NIT • RAN • <null>
Mokelumne	CloudCoverDescription	• 0% • 1-10% • 11-20% • 21-30% • 31-40% • 41-50% • 51-60% • 61-70% • 71-80% • 81-90% • 91-100% • Not Provided • <null>

Database	Field Name	Values Used
	PrecipitationDescription	• Heavy • Light • Mist • Moderate • No Rain • Not Provided • <null>
	WindSpeedDescription	• Calm • Fresh Breeze • Gentle Breeze • Light Air • Light Breeze • Moderate Breeze • Near Gale • Strong Breeze • Not Provided • <null>
Tuoloumne River at Grayson Ranch	WeatherCodeDescription	• Clear day or night • Cloudy day or night • Foggy day or night • N/P • Partly cloudy day or night • Rainy day or night • <null>
Sacramento River at Knights Landing	[Weather not recorded]	

The downloaded BDAT data set consisted of a single denormalized 2-dimensional table. This table was a query output and did not represent the structure of the back end database (Elizabeth Cook, pers. comm.), so a detailed evaluation of the actual BDAT central database could not be done.

The BDAT central and all the field databases were incomplete from the perspective of capturing all information necessary for production estimates. While the basic approach used to design the field databases appeared appropriate, development of these databases ceased before incorporating the ability to capture all data related to capture efficiency. Tables and fields were difficult to understand, parent-child relationships between tables were apparently circular, and data could not be easily managed without an appropriate front end. The database structures were not ideal, open to variation in interpretation, and undocumented. Important information was missing, including such basic items as the stream name where sampling occurred.

A basic piece of information missing from all databases was an explicit way to indicate rearing type (hatchery origin versus naturally spawned) of fish. While this is presumably known by the biologists who collected each data set, it can not be determined by a secondary user of the data. Of the reports reviewed, only two mentioned identifying rearing type of captured fish (Mokelumne River at Woodbridge; Sacramento River at Knights Landing). An examination of the available databases found that adipose fin clips -- which may indicate hatchery origin fish -- were recorded on fish captured at the following RST operations:

- Feather River (High Flow Channel - Sunset Pumps)
- Feather River (Low Flow Channel - Thermalito)
- Merced River (Hagaman State Park)
- Mokelumne River (Woodbridge Irrigation District Dam)
- Sacramento River (Glenn Colusa Irrigation Diversion)
- Sacramento River (Red Bluff Diversion Dam)
- Tuoloumne River (Grayson Ranch)
- Tuolumne River (Shiloh Bridge)

Evaluation and comparison of existing computer systems: data analysis tools

Due to limited time available for this Phase I analysis, little specific information was gathered regarding the data analysis routines and tools used at the various RST operations. It appeared that, for all operations where they were gathered, capture efficiency data were entered into spreadsheets rather than in the databases that house catch data. Catch data were queried from the database and combined with the efficiency data in the spreadsheet in order to determine Chinook salmon production. This is not an inappropriate use of technology, and may be superior to attempting to run analyses with a program that obtains data directly from the database, as a spreadsheet allows a biologist to easily adjust and apply appropriate capture efficiencies while documenting decisions (if desired). Biologists with the USFWS-RB employed a spreadsheet for applying capture efficiencies to determine production of Chinook salmon at Clear and Battle creeks (James Earley, USFWS-RB, personal communication). Other USFWS-RB biologists operated the RSTs at the Red Bluff Diversion Dam on the Sacramento River. The data obtained from this latter operation were passed through a sophisticated analysis procedure requiring at least 27 steps before production estimates were produced. To generate production estimates for Chinook salmon collected at the Stanislaus and Merced rivers, biologists with Cramer Fish Sciences used a proprietary data analysis system (also a spreadsheet).

At many sites, two or more runs of Chinook salmon were present. In most cases an attempt was made to determine Chinook salmon production for each run independently. Often, the run classification for each particular fish caught was based on fish length at date of capture, and then all fish assigned the same run were analyzed as a group.

A more detailed discussion of data analysis techniques used in the Central Valley is contained in McDonald and Banach (2009).

Examination of existing data sets

BDAT: The downloaded BDAT data set contained many fewer fields of data than the other databases contained, consisting primarily of the basic catch data: agency, site, sample time, and number of each species caught. The table also contained fields for the number of trap revolutions since the previous trap check, although it did not contain a field for stream discharge (or correlate) with which to use this statistic. It also had a field for capture efficiency, but none of the 973,654 records (the entire data set) made use of this field. Implementation of the RST data portion of the BDAT database was incomplete — the most significant item not yet completed was incorporation of capture efficiency information (Elizabeth Cook, pers. comm.). The BDAT database, therefore, contained only catch data and cannot be used alone to determine production. It was unclear whether this database contained data summarized by day, if a single record existed for each time a trap was checked, or if it was a combination of these. In addition, a record may have represented all fish caught at a site, or only a subset (see next paragraph). Duplication errors appeared to exist, and some text strings were found in fields that should have been only numeric values.

The location names from the BDAT database, along with the earliest and latest dates when Chinook salmon were observed, are shown in Table 3. As is apparent in Table 3, location names in the database usually did not indicate the sampling sites as used in the annual reports (compare Tables 1 and 3), and in most cases did not contain the name of the stream being sampled. Typographical errors, duplications in site names, and loosely-defined locations were common. To determine fish caught at a site would require further effort. Also apparent in Table 3 is that the BDAT system is no longer in widespread use. Only two sites, Okie Weir and Red Bluff Diversion Dam (which has multiple "stations"), reported data collected after 2006.

Table 3. Sampling location name, and earliest and latest dates Chinook salmon were observed at each site, as derived from the BDAT database (queried December 16, 2009). Note that station names usually do not correspond well to the site names used in reports.

Station Name	Earliest Observation	Latest Observation	Number of Records
~1/2 upstream of Thermalito Afterbay Outlet	1997/12/23	2005/03/30	11,861
Adams Dam	1997/01/18	1998/05/09	701
American River Fish Hatchery	1960/12/14	1960/12/14	10
Below lower Sacramento Road Bridge	2005/01/04	2005/04/26	1,448
Big Chico Creek, Bidwell Pk.	1999/02/23	2003/05/14	1,458
Caswell north	1996/02/06	2003/06/05	13,513
Caswell South	1996/02/06	2003/06/05	15,600
Caswell Traps for both traps	2003/03/06	2003/03/06	5
Deer Creek near Vina	1997/12/10	1999/12/02	570
Feather River at Live Oak	1997/12/23	2002/01/14	7,133
GCID Fish screen	1997/01/01	2006/10/24	57,292
Herringer RiffleE	2000/02/18	2005/03/30	3,506
Herringer RiffleW	2002/01/17	2005/03/30	6,177
MADDOCK RD. BRIDGE	2001/04/05	2004/06/02	8,312
Merced River at Hagaman	1998/03/15	2002/06/01	9,388
Mill Creek near Los Molinos	1997/11/27	1999/11/18	148
Okie Weir	1995/12/01	2008/04/15	25,135
Red Bluff Diversion Dam Gate 1	1994/08/23	2008/09/01	21,847

Station Name	**Earliest Observation**	**Latest Observation**	**Number of Records**
Red Bluff Diversion Dam Gate 10	1994/08/30	2008/08/08	14,180
Red Bluff Diversion Dam Gate 11	1994/07/19	2008/09/01	35,578
Red Bluff Diversion Dam Gate 2	1995/01/25	2008/09/29	28,917
Red Bluff Diversion Dam Gate 3	1994/10/22	2009/04/24	26,252
Red Bluff Diversion Dam Gate 4	1995/10/31	2008/09/30	14,420
Red Bluff Diversion Dam Gate 5	1994/09/23	2009/03/09	20,905
Red Bluff Diversion Dam Gate 5W	2002/06/04	2002/09/12	2,194
Red Bluff Diversion Dam Gate 6	1999/10/19	2009/04/23	36,144
Red Bluff Diversion Dam Gate 6 E	2002/06/05	2002/09/12	2,818
Red Bluff Diversion Dam Gate 6 W	2002/05/25	2002/09/12	3,368
Red Bluff Diversion Dam Gate 7	1994/10/15	2009/04/24	17,653
Red Bluff Diversion Dam Gate 7 E	2002/05/24	2002/09/12	2,941
Red Bluff Diversion Dam Gate 8	1996/04/16	2009/04/24	27,970
Red Bluff Diversion Dam Gate 9	1994/09/22	2008/09/07	20,071
Screw trap in flood plain	1999/02/13	1999/03/06	18
Screw trap in Toe Drain	1998/01/26	2005/05/31	3,753
Stanislaus River at Oakdale	1996/02/02	2003/06/05	28,280
Stanislaus River island below Oakdale RST	2003/03/12	2003/03/12	38
Sutter Rec. District 1500	2001/01/17	2001/03/29	291
Sutter Rec. District 15000	2001/02/09	2001/03/29	169
Sutter Weir 1, West Borrow trap 1	1996/01/17	2001/06/04	13,840
Sutter Weir 1,West Borrow trap 2	2000/03/23	2000/06/04	1,513
Sutter Weir 2, East Borrow	1999/05/13	2000/06/13	2,190
Tuolumne river @ Grayson (North)	1999/01/23	2002/05/31	5,313
Tuolumne river @ Grayson (South)	1999/01/23	2002/05/23	4,272
Tuolumne River at Shiloh	1998/02/16	1998/07/01	1,116
USDA Sites	1993/01/30	1993/07/21	2,266
Woodbridge Dam 1	1993/04/01	2004/06/30	30,365
Woodbridge Dam 2	1994/01/01	2004/06/29	29,345
Woodbridge Dam Ladder	1990/04/12	2004/06/28	7,956
		TOTAL	568,240

OTHER DATA SETS: The evaluation by Shannon (2009) of the nine data sets obtained (housed in seven databases) found that several types of shortcomings were common. Most significantly, the full set of data necessary to calculate Chinook salmon production estimates was not provided for any of the data sets. While catch data were generally available, the suite of data necessary to calculate capture efficiency were not included in any database or were very incomplete. For example, the number of fish marked and released, release dates, the marks applied, and the number recaptured were not always available. Follow-up conversations with the data source personnel by Shannon (2009) indicated that capture efficiency information, where it was collected, was housed in spreadsheets. Often the capture efficiency information was in summarized form, and the original mark and recapture data may no longer exist.

For those databases which contained some mark and release information, these data were very incomplete, and did not indicate how many fish were released for capture efficiency tests. For several of the databases, deriving release data may be possible with significant effort. A potentially confounding factor at a few sites was marks applied upstream for other traps or for other studies, because the RST databases did not indicate the specific marks applied for capture efficiency tests at specific sites (Shannon 2009). This complication is most likely to affect analysis of data sets from mainstem rivers. Although tributary streams were unlikely to have other studies upstream (Stan Allen, PSMFC, pers. comm.), several RST studies did have upstream and downstream sampling sites which could also lead to this confounding occurrence.

Linkages between catch data and capture efficiency data were not explicit within any of the databases. While these linkages may be discoverable, significant effort may be necessary to create them. Linked summarized catch and capture efficiency data existed within spreadsheets for several of the databases, but the raw data necessary for CAMP to recalculate production estimates did not exist in these spreadsheets (Shannon 2009).

The reports reviewed indicated that twelve of the seventeen studies calculated production estimates (American River, Battle Creek, Clear Creek, Cosumnes River, Feather River, Merced River, Mokelumne River, Sacramento River at Knights Landing, Sacramento River at Red Bluff Diversion Dam, Stanislaus River at Caswell State Park, Stanislaus River at Oakdale, and Tuolumne River at Grayson Ranch). If the original raw data do still exist and can be obtained, the difficulties mentioned above should be surmountable for those twelve. For the remaining five, application of capture efficiency data obtained in future years would be required as a surrogate (McDonald and Banach 2009), though reliability could be compromised. Of these five: Deer Creek and Mill Creek are ongoing studies, so it may be possible to generate capture efficiency data; Butte Creek is not ongoing, but may resume if funding becomes available, so it may be possible to generate capture efficiency data; Big Chico Creek and Cosumnes River are not ongoing and it is unlikely trapping will resume, so there is no opportunity to gather capture efficiency data (Douglas Threloff, USFWS, pers. comm.).

Discussion

This evaluation uncovered little that should cause significant difficulties for compiling Chinook salmon RST catch data into a single database for the Central Valley. However, since the objective of this Phase I analysis is to evaluate the feasibility of estimating the number of naturally produced juvenile Chinook salmon from data that were sometimes collected for other purposes, it is not surprising that a number of data related limitations and concerns were identified.

Of greatest concern are the 5 sites that did not include calculation of production and therefore no capture efficiency tests were conducted (Big Chico Creek, Butte Creek, Deer Creek, Mill Creek, and Yuba River). While data from these studies may provide suitable estimates of run timing, it will not be possible to accurately estimate juvenile production. If these studies continue in the future, it may be possible to calculate capture efficiencies that could be applied retroactively to estimate past production, but the accuracy of those estimates would be dependent on consistent trap placement and a determination that the efficiency estimates are consistent year to year at those sites. Without such consistency, it will not be possible to calculate juvenile Chinook salmon production accurately at these five sites.

For those streams where hatchery Chinook salmon are released upstream from a RST operation, the ability to segregate data by rearing type is necessary if production estimates for naturally produced fish are to be calculated. Of 27 Central Valley RST operations targeting Chinook salmon, 12 are downstream from hatcheries that produce Chinook salmon (Doug Threloff, USFWS, pers. comm.). One database examined included data from an additional two RST sites (Tuolumne River at Grayson Ranch; Tuolumne River at Shiloh Bridge) where Chinook salmon with adipose fin clips -- presumably indicating hatchery origin -- were captured. In all these cases it will be necessary to contact the biologists to confirm whether hatchery origin fish are included in the databases, and if so how they are distinguished from naturally produced fish in the data. This distinction is most likely to be based on adipose fin clips; timing of captures after a release, in conjunction with local information provided by biologists, may be useful if no other method for determining hatchery origin fish is available (Doug Threloff, USFWS, pers. comm.).

Several field methods were discovered that may rely on problematic assumptions, which could bring the accuracy of statistical results into question. Employing such methods (such as using a different run of fish for capture efficiency tests when the target run of juvenile Chinook salmon is rare and listed under the Endangered Species Act, or mixing fish from different runs) and accepting their inherent assumptions may be necessary and acceptable for the original purposes of the studies, but when trying to use these data for a different purpose they may present problems if the new purpose requires more stringent assumptions. The CAMP program will need to evaluate their needs relative to the quality and completeness of RST data available before deciding to move forward.

Significant variation in field methods among RST studies can make combining data into a single database difficult and can make use of standard analysis routines problematic. The RST study annual reports examined provided useful details about field methods that permitted meaningful comparisons among sites and predictions about the ease with which the data from the various studies could be standardized. However, the methods sections in RST reports from different watersheds varied in regard to how much detail they provided, and several were missing information essential to a full understanding of field methods or data analysis procedures. For example: the purpose of trapping was not readily discernible for Big Chico Creek, Butte Creek, and Yuba River; the number of traps at Deer Creek and Mill Creek was not provided. Direct interaction of the biologists with this project will be necessary to answer such questions.

It appears that, for the most part, most of the studies used quite similar field methods, and combining their catch data into a single database should not present unusual difficulties. Because the number of study-specific differences was not very large, capturing the data should not cause significant problems.

The Sacramento River at Red Bluff Diversion Dam used quite different methods, in that the number of traps was higher, varied more, and their placement was more variable. In typical small stream studies, placement of the trap in the thalweg is essential for maximizing capture efficiency. At the Red Bluff Diversion Dam the Sacramento River's thalweg is not well defined, and so proportion of the cross-section of the stream sampled is more meaningful than for small stream sites. Capture efficiency relationships for this and other big river sites may rely on different variables (proportion of stream cross section sampled) than those used in small streams (trap position relative to thalweg; stream discharge or a correlate). The same is true for accounting for traps which fail to function for a full day.

For several studies, trapping site or RST position at a site changed, or the number of traps varied. Both are significant for data analysis because production estimates are based on capture efficiency estimates, which are in turn affected by specific trap position in a stream channel. Each change in trap number or position produces a different capture efficiency, although RST position in large rivers such as the Sacramento River may not greatly affect efficiency for an individual trap. Determining an appropriate set of capture efficiency test data to apply for each day of sampling may be problematic unless efficiency tests are stratified by trap position and configuration.

The half cone configuration used at several sites was assumed to reduce capture efficiency by one half (Whitton et al. 2008). USFWS-RB biologists should test this assumption. When a RST is operating, head builds up in front of the RST. This head may be more easily released into one half of the RST (either the modified half or the standard half), allowing more water and more fish through one of the sides and thus affecting capture efficiency. Whether or not the assumption is valid, the proportion of both marked and unmarked fish captured as they passed the sampling site is reduced when using the half cone configuration, and thus confidence intervals are greater than under normal trap configuration (Ricker 1975). Combining data from both normal configuration and half cone configuration requires standardization of data before applying capture efficiency rates to capture data. Calculation of confidence intervals will be more complicated.

Other such sampling peculiarities should be searched for and their assumptions tested. At sites where insufficient capture efficiency tests have been conducted, additional tests should be conducted in future years and applied to all years if the sampling site has not changed (McDonald and Banach 2009), although the consistency of capture efficiency estimates among years should be evaluated.

At Clear Creek capture efficiency tests were not conducted during high flows. Although not mentioned in the other reports, this is the norm. Extrapolating for periods of stream discharge too high to sample will always be necessary, and should be accounted for appropriately during data analysis. Several studies measured water velocity in front of the trap in place of stream discharge or stream gage height. Because a limited amount of water can pass through a trap, head builds up and water velocity in front of the trap approaches an asymptote as stream discharge increases. The effect of this on the capture efficiency to water velocity relationship should be studied.

It appears over half of the studies evaluated used non-target fish for capture efficiency tests. The preferred source of fish for capture efficiency tests is the fish targeted by the study. That is, if wild spring-run Chinook salmon are being studied, then they are the preferred fish to use for capture efficiency tests (Volkhardt et al. 2007). Other species, hatchery-origin fish, and even fish of the same species but of a different run can be expected to be captured at different rates due to differences in preferred travel path, depth or travel, or other behavior. This being the case, a central assumption of mark-recapture studies is violated if non-target fish are used (Ricker 1975). This should be evaluated more closely and documented during Phase II of this project.

The distance upstream from the RST where marked fish are released for capture efficiency tests is another aspect that should be evaluated and documented. Fish released too far upstream may encounter significant mortality before reaching the trapping site and thus be recaptured at unrepresentatively low rates; fish released too close to the trapping site will not distribute themselves across the stream channel in a natural fashion, potentially resulting in unrepresentatively low or high recapture rates. Both of these situations were encountered in the reports evaluated, and both violate a primary assumption of mark-recapture methodology, which is that marked fish become randomly mixed with the unmarked population (Ricker 1975). This variable should also be more closely examined during Phase II of this project.

The key question addressed in this report is whether a single database can be created to capture the data from all Chinook salmon RST studies in the California Central Valley, and then provide these data in a format suitable for calculating statistically-derived estimates of naturally produced Chinook salmon production. Although there would be challenges, this is feasible from the computer technology perspective. The limiting factor is the availability of appropriate data from RST studies that are complete enough, comprehensible enough, and of the right format to make the effort worthwhile. Determining which locations have adequate data for statistically reliable production estimates can only be answered by obtaining and examining the data sets, accompanied by a more thorough examination of field methods. Appendix B contains a summary of the difficulties identified, possible solutions to these difficulties, and whether these alternate solutions would fully resolve the difficulties. But based on the findings here and by Shannon (2009), it is apparent that calculating juvenile Chinook salmon production estimates is only feasible for some sites in the Central Valley.

To be useful, a database for storing RST data with the aim of determining juvenile salmon production must capture all the data required for the desired statistical analyses to be performed. Thus the utility of a database can be judged by its ability to capture and deliver data sets while avoiding data corruption due to missing information, duplicate data, ambiguous data relationships, and other such issues. Using these criteria, none of the existing BDAT-based databases for the Central Valley were sufficient for capturing all these data.

Anticipated challenges for implementing a central database

McDonald and Banach (2009) provided a statistical analysis framework for estimating the production of salmonids from a basin based on the number of migrating fish caught in RSTs. Their model-assisted production estimation technique is flexible enough to adjust for periods when a RST did not operate, varying trap check intervals, and variable efficiency test schedules.

It also allows for calculating confidence intervals, as well as time-series analysis of production estimates.

Most RST studies in the Central Valley collected similar data, and most collected the minimal data needs identified by McDonald and Banach (2009) for producing estimates of salmon production. These minimum data needs are identified in Appendix C, along with other data fields that are not required elements but helpful in evaluating RST results. For operations where the minimum field data for salmon production estimates were not met, what data are available can be captured, though they will be of limited utility because they cannot be used for estimating production. Capturing data elements that are unique to a specific site or are in addition to the needs identified to produce production estimates can be more challenging; examples included weather conditions, dissolved oxygen concentration, daily qualitative assessments of the functionality of the RSTs, and Secchi disk depth. While most of these are straight-forward measurements and capturing them in a database would not be difficult to implement, they represent a larger issue of whether — and how — CAMP would use these auxiliary measurements in statistical analyses. This would require evaluation and consideration by CAMP.

As opposed to simply adding additional fields as described in the previous paragraph, a more difficult issue is how to accommodate alternative or multiple methods used to measure the same concept. For example, staff gage measure and stream depth below the trap (both measured in cm, m, feet and inches, or tenths of feet), water velocity at the trap (m/second, feet/second, or other units), trap rotation rate (revolutions/minute, revolutions/day, or seconds/revolution), and portion of stream discharge sampled (percent or proportion) are all correlated with stream discharge (measured as either m^3/second or $feet^3$/second). While unit conversion is easily accomplished, capturing alternative measures requires creating a more flexible database. This complicates not only data capture, but also analysis routines because the user would need to be prompted to select from available covariates. Routines may also need to perform transformations of the data.

A major difficulty discovered with Central Valley RST databases was an inability to fully understand the database structures obtained. Because the databases were not documented internally and had no metadata associated with them, it was necessary to interpret the purposes of tables, fields, and codes. Using available front end applications to view the data was only partially successful due to network configuration requirements which could not be duplicated. As a result, interpretation along with limited assistance from the biologists who used the databases was the basis for understanding the structures. However, because the databases obtained were complex and the biologists interacted with them mainly through front end applications, the biologists' understanding of the back end databases appeared incomplete.

The existing Central Valley databases examined were not ideally designed or complete, and each was unique. The same type of information was captured in different ways among these databases, complicating the compilation of data sets. In addition, all the database designs were incomplete, with no way to capture and use all aspects of capture efficiency data that were collected in the field, and thus no way to generate salmon production data with data contained solely within the databases — though the data may exist in separate spreadsheets or other formats. These shortcomings are great enough that a wholly new database should be designed or a different existing database be found, rather than trying to modify an existing BDAT-based database.

There are several alternative databases in existence. One is available from biologists in the Klamath Basin (northern California/southwest Oregon). Another from the Columbia River Basin has been in widespread use for roughly 20 years. The state of Idaho has a juvenile trapping database and application used for RST operations. These efforts, and others, demonstrate that technologically the current project is feasible for the California Central Valley. Additionally, it is likely one of these existing systems can be adopted and modified, reducing the time and resources needed for database creation. One or more are also likely to have existing programming needed for preparing raw data for analysis.

A benefit of adopting an existing database is the ability to more easily share data with other programs along the West Coast. Ideally, all systems along the coast should be interoperable in terms of sharing RST data.

Programming front ends

Obtaining a database to house data is an important early aspect of creating a data capture and analysis system. Data input, management, and analysis can be eased by programming one or more front ends that assist the user in performing necessary tasks.

In the case of California Central Valley Chinook salmon, run of fish is frequently determined post hoc by applying length-at-date criteria to the catch data. As data are being selected for analysis, some degree of summarization will often be necessary — for example, to total the number of fish of the same species and run captured in all traps at a site each day. After summarization, missing data must be imputed for days when sampling for a full 24 hours was unsuccessful. These steps should be straightforward to implement. Total catch estimates are then created by summing actual and imputed daily catch data for the defined taxon, run, and range of dates. Where adequate catch data and capture efficiency data are both available, it should be possible to program a front end to generate production estimates and confidence intervals — or at least to perform those tasks for which human input is not required.

How much user interaction will be necessary is uncertain, but human intelligence will be required when calculating production estimates. The front end should prompt the user when such human input is needed. Besides selecting the data filtering criteria of interest (species, run, rearing type, location, time frame), the user would need to direct the program on how to handle each instance of missing data. The user will need to select the covariates of interest, and perhaps also tell the analysis program the covariates' data types and distributions — e.g., one covariate might be a normally distributed continuous variable, while a second may form a highly skewed distribution, and a third may be categorical. Programming for the variety of independent variables used in the various Central Valley RST studies could become somewhat complex. The data analysis front end, however, should be able to provide high quality tables and graphics for easy incorporation into reports written with standard word processing software.

Software selection

When selecting database software for a back end the limitations of various packages, ease of use, cost, and availability of technical assistance must be weighed against the needs. It is not anticipated that the needs of this program would require high performance database software, which lets low-cost options be considered. High cost options are not discussed here.

While Microsoft Access is a widely used and relatively inexpensive program, its limitations preclude it as the back end software for Central Valley RST data. The most significant limitation is a 2 gigabyte file size limit. This file size is already being approached by the data storage needs for the Sacramento River Red Bluff Diversion Dam operation (William Poytress, USFWS-RB, pers. comm.); to house these data as well as RST and environmental data (stream discharge, temperatures, etc.) data from across the Central Valley would easily surpass the 2 gigabyte limitation. Other limitations of Access (maximum recordset size of 1 GB; maximum fields in primary key is 10) may also present problems.

There are several cost free options for a back end. PostgreSQL is open source object-oriented software. However it is not as widely used as some others, and thus obtaining technical assistance could be difficult. MySQL is a widely-used and very powerful open-source database program. However, for the needs of CAMP this is likely not the preferred option due to an uncommon interface and the higher degree of knowledge necessary to install the software and manage a database with it. OpenOffice.org's Base is another free option, but like Access is not meant for the quantity of data that CAMP will need to manage. Microsoft's SQL Server Express is a somewhat stripped-down version of their high quality (and higher cost) SQL Server database software. SQL Server Express is free, and while it is more complex to use than is Access, it is more user friendly than MySQL.

Perhaps the best option for CAMP is to use the existing Microsoft SQL Server corporate database system of the USFWS. This would allow CAMP to use the existing servers, backup systems, software, and expertise of the USFWS at no additional cost. This option would serve very well the need to capture data from all sources. However, if sharing the database with data collectors is desirable then other software would be needed — most likely a Microsoft Access version of the database that can deliver data to the CAMP central database.

Software selection for programming front ends is an important consideration. It is highly recommended that the front end user interfaces and the back end database be separate, modular portions of a system, interacting through the open database connectivity (ODBC) standard. In this way, front-ends and back-ends can each be created with full capabilities. Modularization also allows replacing or adding to part of the system without affecting other parts.

Front ends can be created within database software — for example, many people are familiar with data input forms created within Access. But employing a full-featured programming language is preferable. Taking Access as the most familiar and most likely example, the Visual Basic for Applications programming capabilities within Access are fairly powerful, but limited when compared to a full-featured programming language such as a standard version of Visual Basic or C. Interfaces made within Access are also visually unappealing and use space inefficiently on the computer monitor. Additionally, at times Microsoft changes the technology used within Access; when this happens, existing interfaces can become inoperable.

Visual Basic, C, and Java are the preferred choices for programming front ends because they are powerful and create high quality interfaces, and also because they are commonly used and thus it will be relatively easy and inexpensive to contract changes in the future. Whichever language is used, USFWS should receive the source code as well as the compiled program (if applicable) and require all code to be internally documented. Either a single front end, or multiple front ends, may be desired to interact with the back end database. Functions that will be needed are data entry and manipulation (if desired), and data filtering, grouping, and analysis. Statistical analysis programming should be written in a stand-alone statistical analysis package. One such package is the open-source R statistical software (http://www.r-project.org) that can be called from within the data analysis interface (Trent McDonald, Western EcoSystems Technology, Inc., pers. comm.).

Components of a data storage and analysis system

Although it is technologically feasible to develop a system for capturing and analyzing RST data, the availability and completeness of data to use with such a system is a separate issue. The four basic activities needed to implement and utilize a functioning centralized database and data analysis system, and the difficulty involved in performing those steps, are as follows.

1. Creation of a database
    a. Alternative 1: Design and create a database. A database design needed for RST data will be of medium complexity, and several iterations are likely. No major problems are anticipated, though the variability of field methods does create some complexity. Medium difficulty.
    b. Alternative 2: Obtain an existing database already in use in another area, and modify it to meet CAMP's needs. Problems are likely to result from a need to modify the source database for local needs or to correct flawed database design features. Medium difficulty, but should take fewer iterations and less time than the first alternative because it should already be debugged to some extent.
2. Data capture
    a. Obtain data as computer files or hard copy. This is mainly a human interaction effort, and can take significant time to locate and acquire existing data. This can range from quite simple to very difficult depending on the availability of biologists' time and the status of their data sets, and may be impossible if raw data no longer exist. For data collected in the future, the preferred approach is to create data sharing agreements and automate data transfer or entry. Obtaining data as computer files will likely require development of data crosswalks to load data to the central database (medium difficulty), while obtaining data as hard copy adds the time and expense of data entry.
    b. Enter data in a central database
        i. Existing data. In cases where significant problems do not occur, data capture is usually of medium difficulty. There are two common problems:

1. Incomplete source data. In some cases — such as applying different years' efficiency tests to capture data in order to calculate production — substitute data could be used. But if not, then it is impossible to overcome the lack of data.
2. Source data set poorly designed and poorly documented. This then becomes a human interaction issue, and sometimes cannot be overcome if the people using the database do not themselves understand the database structure, if they cannot be contacted because they have changed jobs, or if significant time has elapsed since the work was done. If assistance from data providers cannot be obtained, some issues can be debugged if there are a limited number of possibilities as to the meaning and all but a single possibility can be ruled out. Medium difficulty to impossible.

ii. Future data

1. If CAMP wishes to acquire and consolidate data collected in the future, it would be advantageous if data collectors used a field database that is interoperable with the CAMP database. To do so, providing a database and familiar-looking front ends for data entry are of great value, as are contractual obligations or formal agreements to provide the data in a specified format. Providing training is helpful, as is providing data analysis tools that would be attractive to biologists. The field database design and programming portions of this, though potentially time consuming, are not difficult. Providing a database and front ends allows for automated quality assurance practices that can be applied during initial data capture, and for data quality control checks that can be run after data capture. Keeping individual field databases synchronized can be an ongoing challenge; as the needs of local biologists change, the databases and associated programming infrastructure will need to be updated. Easy to medium difficulty.

3. Automating data processing and summarization
   a. Automation, when possible, can be very helpful for frequently-repeated, time-intensive tasks. Examples include pooling data by species, run, rearing type, location, and date; locating and accounting for missing data; applying efficiency data based on trap position and configuration; and determining total number of fish from subsamples. Accounting for the large number of variations can be quite involved and difficult.
4. Data analysis and reporting
   a. Multiple methods exist to analyze RST data. A data analysis front end could help the user perform more than one of these. Creating a data analysis system is an incentive that can be used to gain biologist support for data sharing. Complexity will be added as different sets of environmental covariates are allowed for. Programming for these analyses will likely be of medium difficulty.

Conclusions

This evaluation and that by Shannon (2009) of the field methods and available data from RST operations found significant shortcomings related to recalculating Chinook salmon production from the watersheds of the Central Valley.

A major difficulty exists in that many of the RST operations appear to lack capture efficiency data adequate and appropriate for CAMP's needs. Sufficient capture efficiency data were not provided for any of the databases examined in this review or by Shannon (2009). Five of 17 studies for which field methods were reviewed did not generate capture efficiency data (though it may be possible to generate capture efficiencies in the future for three of these). In such cases the existing catch data can be brought into a database, but generating production estimates is not possible. An additional five studies did not use wild Chinook salmon for capture efficiency tests, adding uncertainty to the production estimates generated. For several additional studies the run of Chinook salmon used for capture efficiency tests may have been different than the targeted study fish. Based on information acquired during this review, fewer than half of the studies generated capture efficiency data appropriate for CAMP's needs, and none of them provided for this review the complete raw data needed for production estimates. Without appropriate capture efficiency data, recalculating production estimates may not be feasible.

It is possible technologically to create a database and data analysis system for capturing RST data and producing statistically robust, consistently generated estimates of the production of juvenile Chinook salmon (and other migratory species) from the watersheds of the California Central Valley, either by initiating a new effort or by modifying an existing system from elsewhere. A number of variations were found in the field methods that, though few of them alone were significant, in aggregate may create the need for a fairly complex database and data analysis system. Going forward, the complexity needed in the database and analysis system could be reduced by increased coordination of field methods used by the various RST projects. None of the existing database systems from the Central Valley proved complete enough or robust enough to serve as the basis for developing a valley wide database system.

The remainder of the minimum field data identified by McDonald and Banach (2009) were generally available. Where the requisite data are available, the production estimation technique identified by McDonald and Banach (2009) is flexible enough to accommodate missing counts, varying trap check intervals, and variable efficiency test schedules.

A remaining challenge, should this effort continue, will be to cultivate the support of biologists to participate in this effort to the degree necessary for them to provide their data and assist with understanding them well enough for capturing them in the central database.

References

Greene, S. 1992. Estimated winter-run Chinook salmon salvage at the state water project and Central Valley Project delta pumping facilities. Memorandum dated May 8, 1992 from Sheila Greene, California Department of Water Resources to Randall Brown, California Department of Water Resources. 3 pages plus 15 table pages.

McDonald, T., and M. Banach. 2009. Feasibility of unified analysis methods for rotary screw trap data in the California Central Valley. Task B Report. USFWS Coop. Agreement #81420-8-J163. 16 pages.

Ricker, W. E. 1975. Computation and interpretation of biological statistics of fish populations. Bull. Fish. Res. Board Can. 191. 382 pages.

Shannon, C. 2009. Status report: Assessment of Central Valley rotary screw trap databases, October 2009. Unpublished report. Pacific States Marine Fisheries Commission, Redding, California.

U.S. Fish and Wildlife Service. 2008. Assessment of anadromous fish production in the Central Valley of California between 1992 and 2007. Report prepared by the U.S. Fish and Wildlife Service and Bureau of Reclamation, Comprehensive Assessment and Monitoring Program. Sacramento, California. 106 pages.

Vincik, R. F., R. G. Titus, and B. Snider. 2006. Timing, composition and abundance of juvenile anadromous salmonid emigration in the Sacramento River near Knights Landing September 1999 - September 2000. California Department of Fish and Game, Sacramento Valley-Central Sierra Region. 23 pages.

Volkhardt, G. C., S. L. Johnson, B. A. Miller, T. E. Nickelson, and D. E. Seiler. 2007. Rotary screw traps and inclined plane screen traps. Pages 235-266 *in* D. H. Johnson, B. M. Shrier, J. S. O'Neal, J. A. Knutzen, X. Augerot, T. A. O'Neill, and T. N. Pearsons. Salmonid field protocols handbook: Techniques for assessing status and trends in salmon and trout populations. American Fisheries Society, Bethesda, Maryland.

Whitton, K. S., D. J. Colby, J. M. Newton, and M. R. Brown. 2008. Juvenile salmonid monitoring in Battle Creek, California, November 2007 through June 2008.

Figures

Figure 1. Entity-relationship diagram for the main tables in the Battle Creek / Clear Creek screw trap database.

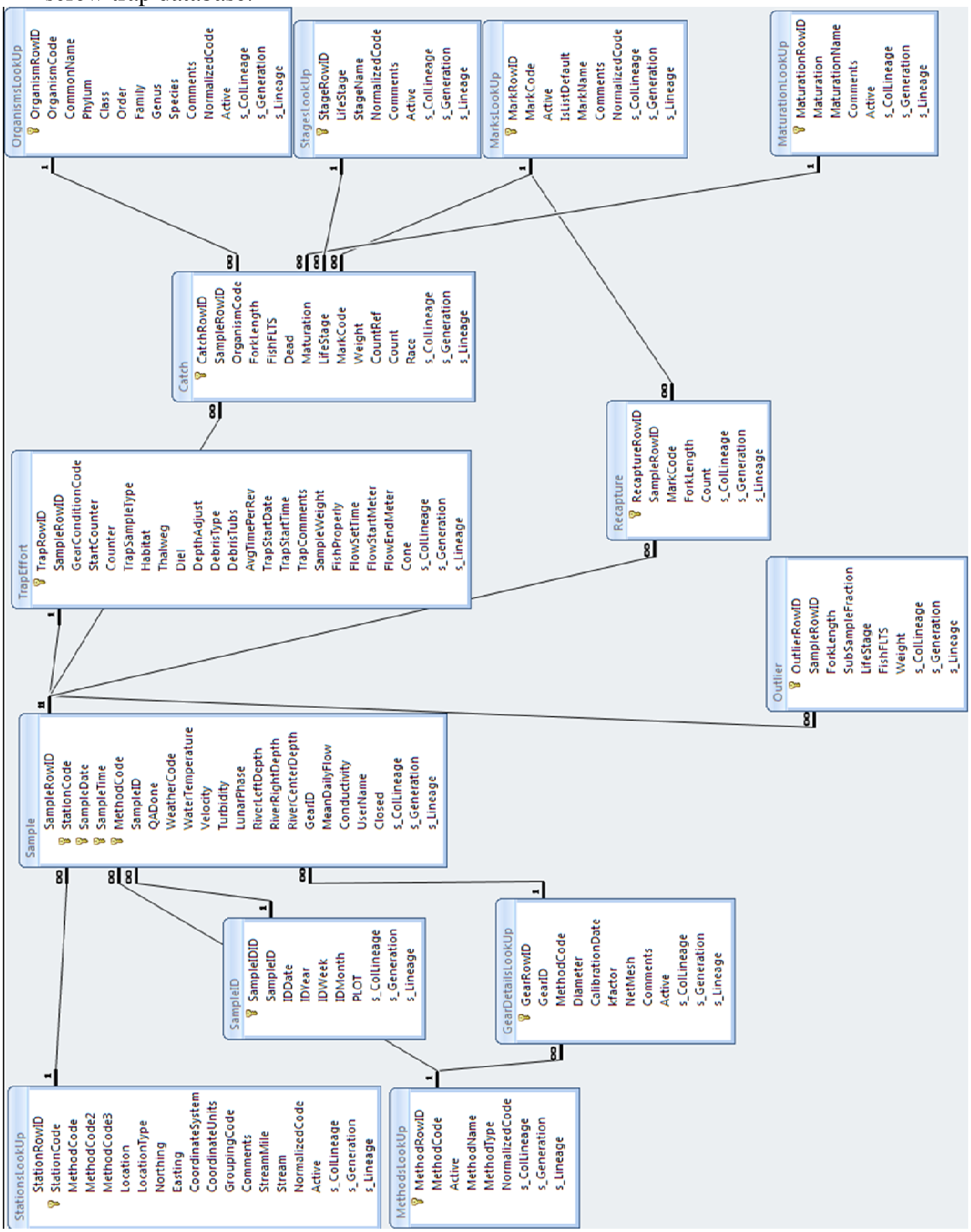

Figure 2. Entity-relationship diagram for the main tables in the Stanislaus River / Merced River screw trap database.

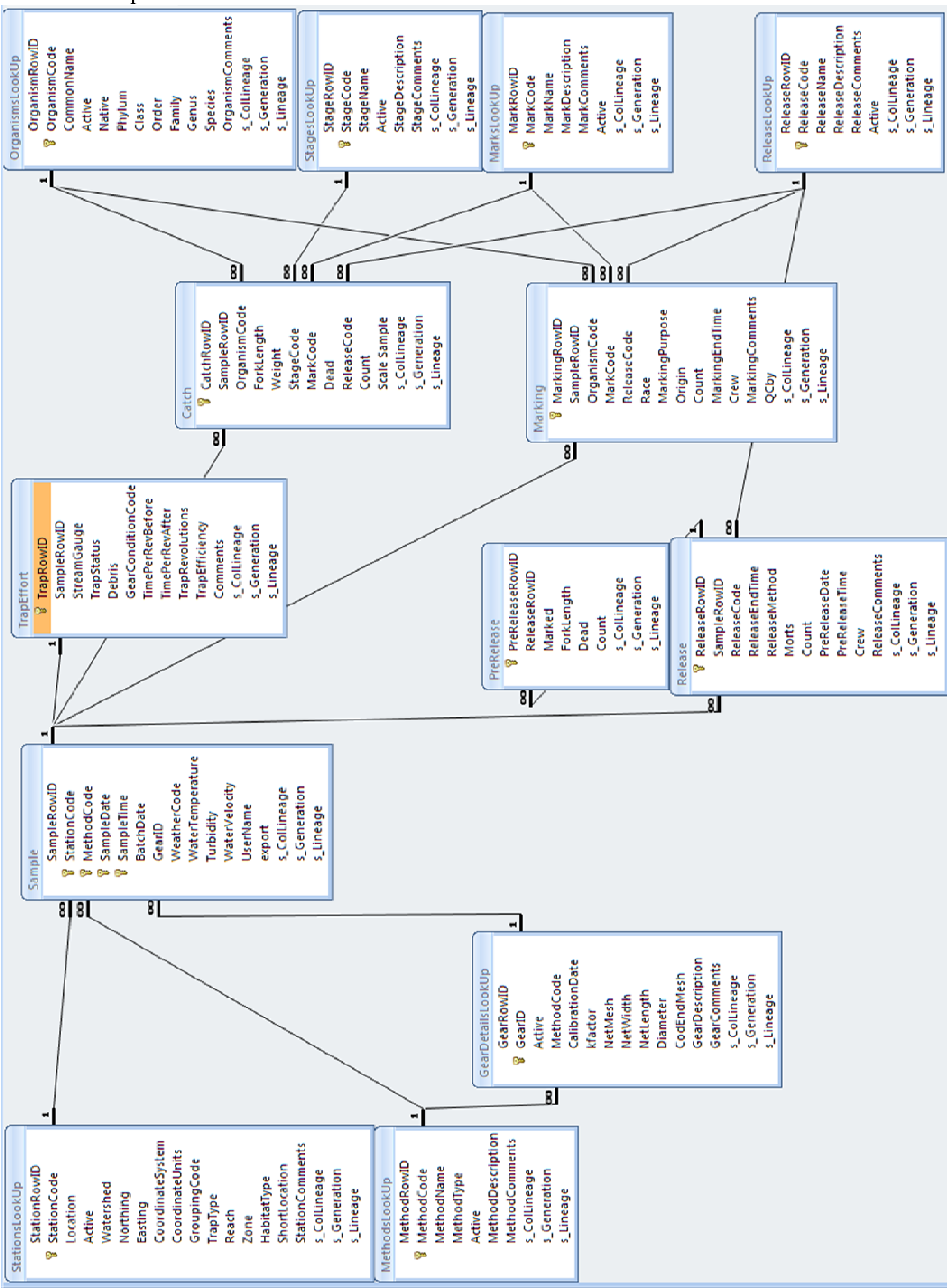

Figure 3. Table structure for the Bay-Delta and Tributaries (BDAT) data as obtained from the BDAT web site (http://bdat.ca.gov).

Project
Agency
Agency link
Metadata link
Station Code
Station Name
Region
River km Index
x coord
y coord
Sample Date
Sample Time
Sample ID
Subsample Number
Survey Type
Common Name
Phylum
Class
Order
Family
Genus
Species
Count
Length (mm)
Weight (g)
Life Stage
Mark Code
Gear Status
Comments
Trap Revolutions
Trap Comments
Trap Period (min)
Efficiency Percent
QAQC status
Unit Symbol

Appendix A

Table A-1. Field operations at the 19 California Central Valley screw trapping sites, as determined from a cursory review the most recent annual progress report for each site. Some items are summarized. For additional details see table A-2.

Number of traps at site has been constant	Purpose: environmental factors' effects on abundance	Purpose: outmigration timing	Purpose: relative abundance	Purpose: quantify production	Sampling has always taken place at the same location.	**Item** / **Site**
No	Yes	Yes	Yes	Yes	Yes	**American River**
Yes	?	Yes	Yes	Yes	Yes	**Battle Creek (2 sites)**
Yes	?	?	?	?	Yes	**Big Chico Creek**
Yes	?	?	?	?	No	**Butte Creek**
Yes	?	Yes	Yes	Yes	Yes	**Clear Creek (2 sites)**
Yes	No	Yes	Yes	Yes	Yes	**Cosumnes River**
?	?	Yes	?	No	?	**Deer Creek**
Yes	?	Yes	Yes	Yes	?	**Feather River (2 sites)**
No	?	Yes	Yes	Yes	No	**Merced River**
?	?	Yes	?	No	?	**Mill Creek**
Yes	Yes	Yes	Yes	Yes	No	**Mokelumne River**
No	Yes	Yes	Yes	Yes	Yes	**Sacramento River at Knights Landing**
No	?	Yes	Yes	Yes	Yes	**Sacramento River at Red Bluff Diversion Dam**
No	?	Yes	Yes	Yes	Yes	**Stanislaus River at Caswell State Park**
Yes	?	Yes	Yes	Yes	Yes	**Stanislaus River at Oakdale**
Yes	?	Yes	Yes	Yes	No	**Tuolumne River at Grayson Ranch**
Yes	?	?	?	?	?	**Yuba River**

Typical capture efficiency (percent)	Sufficient number of fish used for efficiency tests	Target fish used in capture efficiency tests	Capture efficiency test frequency	Capture efficiency tests are conducted	Trap revolutions are monitored & reported	Attempt to trap 7 days each week	Item / Site
0.5 - 1.5	Yes	Yes	9 / year	Yes	?	No	**American River**
3 - 7	Yes	Yes	2 / week	Yes	Yes	Yes	**Battle Creek (2 sites)**
N/A	N/A	N/A	N/A	No	Yes	Yes	**Big Chico Creek**
N/A	N/A	N/A	N/A	No	Yes	Yes	**Butte Creek**
2 - 17	Yes	Yes	2 / week	Yes	Yes	Yes	**Clear Creek (2 sites)**
N/A	N/A	N/A	N/A	No	No	Yes	**Cosumnes River**
N/A	N/A	N/A	N/A	No	No	Yes	**Deer Creek**
?	Yes	Yes	8 / year	Yes	Yes	Yes	**Feather River (2 sites)**
?	Yes	No	7 / year	Yes	Yes	Yes	**Merced River**
N/A	N/A	N/A	N/A	No	No	Yes	**Mill Creek**
0.1 - 7.2	Yes	No	18 / year	Yes	Yes	No	**Mokelumne River**
0.25	Maybe	No	Weekly.	Yes	Yes	Yes	**Sacramento River at Knights Landing**
0.9 - 3.4	Yes	Yes	8 / year	Yes	No	Yes	**Sacramento River at Red Bluff Diversion Dam**
2 - 20	Yes	Yes	2 - 35 / year	Yes	Yes	Yes	**Stanislaus River at Caswell State Park**
2 - 20	Yes	Yes	2 - 35 / year	Yes	Yes	No	**Stanislaus River at Oakdale**
?	Yes	No	?	Yes	?	Yes	**Tuolumne River at Grayson Ranch**
?	Yes	Yes	2 / month	Yes	Yes	Yes	**Yuba River**

What is done when trap operates <24 hours in a day?	Factors used to refine efficiency tests	Is trap run modified at times? (Such as half cone configuration)	How is # of fish estimated when trap does not operate for a full day?	**Item** / **Site**
N/A	Multiple	No	Straight proportion	**American River**
?	Multiple	Yes	Not conducted	**Battle Creek (2 sites)**
?	Multiple	?	N/A	**Big Chico Creek**
?	Multiple	Yes	N/A	**Butte Creek**
Proportional expansion	Multiple	Yes	mean of +/- number of days missed.	**Clear Creek (2 sites)**
N/A	Multiple	No	N/A	**Cosumnes River**
N/A	Multiple	No	N/A	**Deer Creek**
?	Multiple	No	Three methods mentioned.	**Feather River (2 sites)**
?	Multiple	?	?	**Merced River**
N/A	Multiple	No	N/A	**Mill Creek**
?	Multiple	?	7-day moving mean	**Mokelumne River**
?	Multiple		?	**Sacramento River at Knights Landing**
Proportional expansion	Multiple	Yes	?	**Sacramento River at Red Bluff Diversion Dam**
?	Multiple	Yes	Weighted 11-day moving mean	**Stanislaus River at Caswell State Park**
?	Multiple	No	Weighted 11-day moving mean	**Stanislaus River at Oakdale**
Proportional expansion	Multiple	?	?	**Tuolumne River at Grayson Ranch**
?	Multiple	?	?	**Yuba River**

Table A-2. Field operations at the 19 California Central Valley screw trapping sites, as determined from a cursory review the most recent annual progress report for each site. Items highlighted in yellow may cause difficulties in performing statistical analyses. Items in red text may cause difficulties in interpreting statistical analyses that are performed. (Information in this table is summarized in Table A-1.)

Item	**Trapping Site**		
	American River	**Battle Creek (2 sites)**	**Big Chico Creek**
Sampling has always taken place at the same location.	Yes (river mile 9)	Yes (river miles 3 and 6)	Yes
Purpose: quantify production	Yes	Yes	?
Purpose: relative abundance	Yes	Yes	?
Purpose: outmigration timing	Yes	Yes	?
Purpose: environmental factors' effects on abundance	Yes	?	?
Number of traps at site	1 to 2 (8-foot trap, at times supplemented by a 5-foot trap)	1	1
Attempt to trap 7 days each week	No	Yes	Yes
Trap revolutions are monitored & reported	?	Yes	Yes
Capture efficiency tests are conducted	Yes	Yes	No
Capture efficiency test frequency	9 / year	2 / week	N/A
Fish used in capture efficiency tests	Chinook salmon	Wild fall Chinook salmon	N/A
Number of fish used for efficiency tests	All fish caught: as high as 6,012	70-500	N/A
Typical capture efficiency (percent)	0.5 - 1.5 %	3 - 7 %	N/A
How is # of fish estimated when trap does not operate for a full day?	Straight proportion	Not conducted	N/A
Is trap run modified at times? (Such as half cone configuration)	No	half cone configuration employed at times	?

	Trapping Site		
Factors used to refine efficiency tests. [In square brackets means these are recorded but not for efficiency tests.]	Stream discharge Turbidity Temperature Fish fork length Fish life stage Fish weight	Water depth Cone depth Trap rotation rate Amount of debris Weather Temperature Water velocity Turbidity Stream discharge Fish fork length Fish life stage	[Water velocity Temperature Turbidity Fish fork length Fish weight]
What is done when trap operates <24 hours in a day?	At times run >24 hours between checks. Unit of analysis is a week.	?	?

Table A-2. (continued)

Item	Trapping Site		
	Butte Creek	**Clear Creek (2 sites)**	**Cosumnes River**
Sampling has always taken place at the same location.	Same site, but trap is moved within the site	Yes (river mile 1.7 and river mile 8.3)	Only occurred 1999
Purpose: quantify production	?	Yes	Yes
Purpose: relative abundance	?	Yes	Yes
Purpose: outmigration timing	?	Yes	Yes
Purpose: environmental factors' effects on abundance	?	?	No
Number of traps at site	1	1	1
Attempt to trap 7 days each week	Yes	Yes	Yes
Trap revolutions are monitored & reported	Yes	Yes	No
Capture efficiency tests are conducted	No	Yes	No
Capture efficiency test frequency	N/A	2 / week Only during low to mid-level flows	N/A
Fish used in capture efficiency tests	N/A	Wild Chinook salmon	N/A
Number of fish used for efficiency tests	N/A	>400	N/A
Typical capture efficiency (percent)	N/A	2 - 17 %	N/A
How is # of fish estimated when trap does not operate for a full day?	N/A	$\bar{x}$ of +/- number of days missed. For example, missing 2 days means taking mean of catch from 2 previous and 2 subsequent days	N/A
Is trap run modified at times? (Such as half cone configuration)	Trap placement is changed frequently	half cone configuration employed at times	No

	Trapping Site		
Factors used to refine efficiency tests. [In square brackets means these are recorded but not for efficiency tests.]	[Water velocity Temperature Turbidity Fish fork length Fish weight]	Water depth Cone depth Trap rotation rate Weather Temperature Water velocity Turbidity Fish fork length Fish life stage	[Stream discharge Temperature Fish fork length Fish weight]
What is done when trap operates <24 hours in a day?	?	Proportional expansion if trap ran at least 1/2 day	Trap checked only 2-4 times per week

Table A-2. (continued)

Item	Trapping Site		
	Deer Creek	Feather River (2 sites)	Merced River
Sampling has always taken place at the same location.	?	?	Channel changed in 2007; trap moved 40 m.
Purpose: quantify production	No	Yes	Yes
Purpose: relative abundance	?	Yes	Yes
Purpose: outmigration timing	Yes	Yes	Yes
Purpose: environmental factors' effects on abundance	?	?	?
Number of traps at site	?	1 at upper site. 2 at lower site since 2004.	1 to 2
Attempt to trap 7 days each week	Yes	Yes	Yes
Trap revolutions are monitored & reported	No	Yes	Yes
Capture efficiency tests are conducted	No	Yes	Yes
Capture efficiency test frequency	N/A	8 / year	7 / year
Fish used in capture efficiency tests	N/A	Wild Chinook salmon	Hatchery Chinook salmon
Number of fish used for efficiency tests	N/A	About 1000	About 1000
Typical capture efficiency (percent)	N/A	?	?
How is # of fish estimated when trap does not operate for a full day?	N/A	Unsure. Three different methods mentioned.	?
Is trap run modified at times? (Such as half cone configuration)	No	No	?

	Trapping Site		
Factors used to refine efficiency tests. [In square brackets means these are recorded but not for efficiency tests.]	[Temperature Stream discharge Turbidity Fish fork length Fish weight]	Stream discharge Temperature Turbidity Trap functioning "good/fair/poor" Fish fork length Fish life stage	Stream discharge Water velocity River stage Temperature Turbidity Dissolved oxygen Fish fork length Fish weight Fish life stage
What is done when trap operates <24 hours in a day?	N/A	?	?

Table A-2. (continued)

Item	Trapping Site		
	Mill Creek	**Mokelumne River**	**Sacramento River at Knights Landing**
Sampling has always taken place at the same location.	?	No	Yes
Purpose: quantify production	No	Yes	Yes
Purpose: relative abundance	?	Yes	Yes
Purpose: outmigration timing	Yes	Yes	Yes
Purpose: environmental factors' effects on abundance	?	Yes: day vs. night catch and efficiency	Yes: stream discharge
Number of traps at site	?	1	2 to 3
Attempt to trap 7 days each week	Yes	No	Yes
Trap revolutions are monitored & reported	No	Yes	Yes
Capture efficiency tests are conducted	No	Yes	Yes
Capture efficiency test frequency	N/A	18 / year	Daily when possible. Pooled by week.
Fish used in capture efficiency tests	N/A	Hatchery Chinook salmon	All Chinook salmon caught (hatchery and wild)
Number of fish used for efficiency tests	N/A	About 800 - 2100	> 100 (as many as possible)
Typical capture efficiency (percent)	N/A	0.1 - 7.2 %	0.25 %
How is # of fish estimated when trap does not operate for a full day?	N/A	Moving mean of +/- 3 days	?
Is trap run modified at times? (Such as half cone configuration)	No	?	

	Trapping Site		
Factors used to refine efficiency tests. [In square brackets means these are recorded but not for efficiency tests.]	[Stream discharge Temperature Turbidity Fish fork length Fish weight]	Stream discharge Trap rotation speed Turbidity Temperature Dissolved oxygen Precipitation Fish length Fish weight Fish life stage	Stream discharge Water velocity Turbidity Secchi depth Temperature
What is done when trap operates <24 hours in a day?	N/A	?	?

Table A-2. (continued)

Item	Trapping Site		
	Sacramento River at Red Bluff Diversion Dam	**Stanislaus River at Caswell State Park**	**Stanislaus River at Oakdale**
Sampling has always taken place at the same location.	Yes	Yes	Yes
Purpose: quantify production	Yes	Yes	Yes
Purpose: relative abundance	Yes	Yes	Yes
Purpose: outmigration timing	Yes	Yes	Yes
Purpose: environmental factors' effects on abundance	?	?	?
Number of traps at site	3 to 4	2 to 3	1
Attempt to trap 7 days each week	Yes	Yes	No
Trap revolutions are monitored & reported	No	Yes	Yes
Capture efficiency tests are conducted	Yes	Yes	Yes
Capture efficiency test frequency	8 / year	2 - 35 / year	2 - 35 / year
Fish used in capture efficiency tests	Wild Chinook salmon	Wild Chinook salmon	Wild Chinook salmon
Number of fish used for efficiency tests	800 - 3000 (But may be confounded by releases from upstream traps.)	77 - 3371	77 - 3371
Typical capture efficiency (percent)	0.9 - 3.4 %	2 - 20 %	2 - 20 %
How is # of fish estimated when trap does not operate for a full day?	?	Weighted moving mean of +/- 5 days either side of date missing.	Weighted moving mean of +/- 5 days either side of date missing.

	Trapping Site		
Is trap run modified at times? (Such as half cone configuration)	Number and position; traps moved frequently. half cone. Configuration of dam changes.	Number of traps changes.	No
Factors used to refine efficiency tests. [In square brackets means these are recorded but not for efficiency tests.]	Water velocity Cone depth Percent of discharge sampled Fish fork length	Stream discharge Water velocity Trap rotation rate Temperature Turbidity Dissolve oxygen Fish fork length Fish weight Fish life stage Fish health	Stream discharge Water velocity Trap rotation rate Temperature Turbidity Dissolve oxygen Fish fork length Fish weight Fish life stage Fish health
What is done when trap operates <24 hours in a day?	Proportional expansion	?	?

Table A-2. (continued)

Item	Trapping Site		
	Tuolumne River at Grayson Ranch	Yuba River	
Sampling has always taken place at the same location.	No. River mile 3.4 to river mile 5.2	?	
Purpose: quantify production	Yes	?	
Purpose: relative abundance	Yes	?	
Purpose: outmigration timing	Yes	?	
Purpose: environmental factors' effects on abundance	?	?	
Number of traps at site	2	2	
Attempt to trap 7 days each week	Yes	Yes	
Trap revolutions are monitored & reported	?	Yes	
Capture efficiency tests are conducted	Yes	Yes	
Capture efficiency test frequency	?	2 / month	
Fish used in capture efficiency tests	Mostly hatchery Chinook salmon, some wild	Wild Chinook salmon	
Number of fish used for efficiency tests	About 2000	> 300	
Typical capture efficiency (percent)	?	?	
How is # of fish estimated when trap does not operate for a full day?	?	?	
Is trap run modified at times? (Such as half cone configuration)	?	?	

	Trapping Site		
Factors used to refine efficiency tests. [In square brackets means these are recorded but not for efficiency tests.]	Stream discharge Water velocity Conductivity Temperature Turbidity Fish fork length Fish life stage	Stream discharge Water velocity Temperature Turbidity Fish fork length Fish weight	
What is done when trap operates <24 hours in a day?	Proportional expansion based on trap revolutions	?	

Appendix B

Potentially problematic issues identified in the Results section, potential solutions, and efficacy of the potential solutions.

Difficulty	**Effect of Difficulty**	**Potential Solutions**	**Solution Fully Resolves Issue**
Difficulties identified in "Evaluation and comparison of field methods" section of Results			
"Potential difficulties in capturing data in a central database" subsection			
Unknown if catch data in some existing field databases are already adjusted for days when RST did not operate 24 hours.	Without this information it is impossible to know what the data represent, and thus impossible to know how to store the data.	Communication with the data source personnel to clarify data.	Yes, where communication is possible.
Unknown if catch data in some existing field databases are already adjusted to account for subsampling.	Without this information it is impossible to know what the data represent, and thus impossible to know how to store the data.	Communication with the data source personnel to clarify data.	Yes, where communication is possible.
"Potential difficulties in performing statistical analyses on the data" subsection			
Trapping sites or trap position at a site have changed over time at some operations.	Likely affects trap efficiency and therefore capture efficiency measures do not apply to all catch data.	Stratify catch and capture efficiency data according to trap site and position.	Yes, but only if sufficient capture efficiency data exist for each configuration.
The number of traps used at some sites vary over the course of a field season.	Affects capture efficiency for the site as a whole and therefore capture efficiency measures do not apply to all catch data.	Stratify catch and capture efficiency data according to number of traps and trap position.	Yes, but only if sufficient capture efficiency data exist for each configuration.
Traps are not operated 7 days per week at some operations.	Requires expansion of capture data to account for missing days. Also may result in reducing number of recaptures and thus lowering capture efficiency measures.	Existing data can be used to impute for missed days. Recapture data can be examined to determine likelihood of having missed recaptures.	Only partially. Both effects can be compensated for, but result in increased production estimate confidence intervals.

Difficulty	Effect of Difficulty	Potential Solutions	Solution Fully Resolves Issue
Capture efficiency tests were not done at some sites where production estimation is a stated objective.	Because an estimate of capture efficiency is required, calculating production estimates is impossible.	Two scenarios: 1) Where traps are still operating, capture efficiency tests can be conducted and applied to previous years' data. 2) Where traps are no longer operating or have been moved there is no way to obtain valid capture efficiency estimates.	Under scenario #1 reasonable estimates can be arrived at, but with increased uncertainty. Under scenario #2 there is no way to resolve this issue.
Half-cone configuration was used at some locations.	Affects trap efficiency and therefore capture efficiency measures do not apply to all catch data.	1) Stratify catch and capture efficiency data according to trap configuration. 2) Assume half-cone configuration cut capture of fish by one half, and adjust capture efficiency data accordingly, if not already done.	Yes for option #1, but only if sufficient capture efficiency data exist for each configuration. Yes for option #2, but confidence intervals may increase.
Capture efficiency tests and trapping activities not conducted over the complete range of river discharges that migrating salmon experience.	Leads to an underestimation of salmon production or abundance if large numbers of salmon migrate during high discharge events. Happens only in high-water years. This is an inherent short-coming of methodologies for trapping down-migrating fishes.	Extrapolating outside of the range of observed values is the only possible option.	No. But a range of reasonable extrapolated values can be estimated, and interpreted and employed accordingly.

Difficulty		Effect of Difficulty	Potential Solutions	Solution Fully Resolves Issue
	Too few capture efficiency tests conducted to build a stream discharge-based capture efficiency model at some sites.	Capture efficiency model can be created but will be insensitive to environmental covariates that help refine capture efficiency estimates. Also, substantial variability among available capture efficiency tests (which is common) will result in wide confidence intervals on production estimates.	Employ simple mean of available capture efficiency estimates.	No, unless there is little variation in stream discharge, and also low variability in capture efficiency estimates.
	Units of measure not documented for various variables.	Difficult to report results and to compare among sites.	1) Communication with the data source personnel to clarify data. 2) Examine values and make educated guesses. (Only if units used are not central to analysis. For example, it does not matter if stream discharge is in cubic feet per second or cubic meters per minute if it is used as a covariate for production estimates, but not for further prediction purposes.)	Yes, where communication is possible. Where units remain unknown data analysis is likely still possible, though slightly undefined. Biologists should be asked to prepare metadata for future trapping data.

Difficulty	Effect of Difficulty	Potential Solutions	Solution Fully Resolves Issue
How to account for days when a RST did not operate successfully for 24 hours.	Capture for the day is potentially less than it should have been to account for the fish moving past the trapping site.	1) Treat data as if trap had not operated that day, and impute for the entire day. 2) Estimate proportion of nighttime hours sampled and expand capture to entire nighttime migration period. #2 assumes fish moved little during daylight.	Yes, though both solutions result in increased production estimate confidence intervals.
"Potential difficulties in interpreting the statistical analyses" subsection			
When hatchery Chinook salmon (or different run) are used for capture efficiency tests, their capture efficiencies are usually not the same as for wild Chinook salmon. Inaccurate production estimates for wild Chinook salmon are very likely.	While the capture efficiency estimates provide no difficulty in performing statistical procedures, the assumptions of mark-recapture estimates have been violated and thus the results cannot be readily interpreted.	A review of hatchery and wild fish capture efficiencies at other sites where both are measured may allow for adjusting hatchery capture rates to estimate wild fish capture rate.	No. However, this is an inherent difficulty in sampling fish in some locations. Though exact estimates may be skewed, such measures are still a much better index of abundance than having no efficiency data.

Difficulty		Effect of Difficulty	Potential Solutions	Solution Fully Resolves Issue
	Hatchery versus naturally produced fish not explicitly defined in databases.	The number of naturally produced fish cannot be determined if hatchery fish are included in the data.	Communicate with biologists to ensure data are being correctly interpreted. If this is not possible, make assumption that fish with adipose clips are hatchery fish. Try to confirm designation by querying the coded wire tag database (www.rmis.org).	Yes, if explicit determinations can be made.
Difficulties identified in "Evaluation and Comparison of existing computer systems: databases" section of Results				
	The structure of field databases is difficult to assess.	If a database cannot be understood then the data it contains cannot be reliably captured in a central database.	Communication with the data source personnel to clarify data.	Yes, where communication is possible.
	Each field database is unique in its tables, table relationships, fields in each table, where various data are stored, and lookup codes.	Data from each database must be individually mapped to a central database.	This is normal when compiling data from multiple sources. It had been hoped this was not the case for these databases with a common origin so that labor could be avoided.	This does not cause unanticipated problems, only additional work.

Difficulty		Effect of Difficulty	Potential Solutions	Solution Fully Resolves Issue
	Some databases have fields for data that do not appear in other databases.	Evaluating the value of additional fields is necessary, generally by a biologist rather than a data technician. Evaluating how to use the extra fields is also required.	If the extra fields are thought to be of little value then they can be ignored. If the extra fields are thought to be useful then the central database must be modified to accommodate them and the statistical analyses must be altered to use them when available.	Yes, but could require substantial effort to accomplish.
	The BDAT database structure was not designed to store capture efficiency data; such data are frequently stored in Excel spreadsheets specific to each RST operation.	The field databases based on the BDAT model are incomplete because they do not capture the capture efficiency data needed for estimating production.	Design a way to capture these data in a new central database. Acquire the data from the spreadsheet or other formats.	Yes, where efficiency data are available.
	It is not clear if the BDAT database contains data summarized by day, if a single record exists for each time a trap is checked, or if it is a combination of these	Without this information it is not possible to calculate daily production estimates.	Obtaining data from the field databases, and communicating with the data sources, is the preferred option. If this is not possible then analyzing the data may not be possible because assumptions about the data would be necessary.	Yes, if the information is available from the project staff.

Difficulty		Effect of Difficulty	Potential Solutions	Solution Fully Resolves Issue
	The databases apparently lack strong data typing (text strings were found in one ostensibly numeric field).	If data types are not enforced then data integrity cannot be ensured and analyses may be skewed. Text in a numeric field will cause some programs to generate an error; other programs may treat the text as the number zero and calculate accordingly.	Quality control tests will have to be implemented in data capture and analysis routines.	Yes, if unanticipated data errors can be corrected. Otherwise, some data may need to be ignored, affecting reliability of results.
	Tables in the BDAT and field databases lack key fields that prevent duplicate data.	Duplicate data are possible. Depending on how analysis software handles this situation, incorrect production estimates could result.	1) Write quality control routines to identify and address duplicate data. 2) Design a new central database with appropriate key fields to prevent duplicate data.	Yes, if duplicates found are not ambiguous. (For example, two identical records.) Ambiguous duplication (such as two records for the same day but with different values) will require communication with data sources.
	The BDAT and field databases are apparently circular in referential logic between tables, and open to variation in interpretation.	There is no way to independently analyze the database in order to understand how to query data from it in an appropriate way.	1) Communication with the data source personnel to clarify data. 2) Conduct "what if" test queries against known correct values, if those exist.	The first option potentially resolves the issue, where communication is possible. However, this assumes the data source person is knowledgeable enough about the structure of the database to answer questions that arise. The second potential solution would require more effort, but would work when the data source person interacts with the database only through a front end.

Difficulty	Effect of Difficulty	Potential Solutions	Solution Fully Resolves Issue
Difficulties identified in "Evaluation and Comparison of existing computer systems: data analysis tools" section of Results			
Capture efficiency and daily catch data are not stored in the same database	Using catch and capture efficiency data together will require capturing both in the same location.	1) Capture the capture efficiency data in the central database. 2) Capture the catch data in the spreadsheets where the efficiency data reside. 3) Export the appropriate catch and capture efficiency data to a third location.	While an analysis of which method is preferred will likely require experience to determine, this should be a relatively easy issue to overcome so long as both sets of data exist.
Multiple runs of Chinook salmon occur in a single watershed and migration times overlap.	If there is a desire to estimate the production of each run of Chinook salmon independently, then a mechanism is needed to calculate the number of each run that passed the trapping site. Failure to do so will result in inaccurate results.	1) Reevaluate needs to determine if ignoring runs and analyzing data for all Chinook salmon in a watershed is acceptable. 2) Assign run designations to individual fish by applying an algorithm such as the "length-at-date" criteria currently in use in the Central Valley.	The first option, if acceptable, would fully resolve the issue. The second option leaves analyses open to debate.
Difficulties identified in "Examination of existing data" section of Results			
Data required for calculating production estimates not available in databases; may only exist in summarized form.	Without all the required data it is not possible to recalculate production estimates from the raw data.	1) Use the summarized data available in spreadsheets. 2) Recreate the data set based on data found in reports or spreadsheets.	If the data available are not appropriate for use in the statistical methods developed by McDonald and Banach (2009) then it will not be possible to resolve this issue.

Difficulty		Effect of Difficulty	Potential Solutions	Solution Fully Resolves Issue
	Unknown if fish marked in upstream studies confound capture efficiency data. Linkage between mark/release data and capture data do not exist. Likely only a potential problem for a few RST sites on the major rivers.	If fish marked in upstream studies are interpreted as having been marked for capture efficiency purposes then capture efficiency will be overestimated.	Clearly determine which mark(s) were applied at which sites and when; cross-reference to capture data.	Yes, if the information is available to perform this task.

Appendix C

Shown below is a listing of information pertinent to all RST operations. Items identified as "minimum field data" by McDonald and Banach (2009) are shown in blue text and marked with an asterisk. Two asterisks indicate that the required field may be difficult to obtain for a substantial number of Chinook salmon RST data sets from the Central Valley, based on my evaluation and that of Shannon (2009).

1. Trap placement data (recorded whenever a trap is installed, moved, its configuration modified, or fishing is started or stopped)
    a. *Trap size
    b. *Trapping site
        i. *Stream name
        ii. *Lat/long or other information to identify where on stream
    c. Trap number at a site
    d. Trap size
    e. *Date
    f. *Time
    g. *Whether trap is fishing after the change
    h. **Trap location within the site
    i. Is trap in thalweg?
2. Trap check data: physical
    a. *Trapping site
    b. Trap number at site
    c. *Date
    d. *Time
    e. Whether trap is functioning when trap check begins.
    f. Time check is completed
    g. *Cone rotation counter reading at beginning of trap check (used only when trap is not working when trap check begins)
    h. *Cone rotation counter reading at completion of trap check
    i. *Cone rotation speed (rpm)
    j. **Submerged cone depth
    k. *Is trap operating in standard configuration?
        i. Non-standard: half cone configuration
        ii. Non-standard: drum not fully dropped
3. Trap check data: fish caught (allow multiple instances of the following)
    a. Subsampled fraction (default is 1.0)
        i. Provide mechanism for expanding this before data analysis
    b. *Taxon
    c. Rearing type (wild vs. hatchery)
    d. Life stage
    e. *Length
        i. Type (fork length, total length, etc., with default for each species definable)
        ii. Units

   f. Weight
      i. Units
   g. *Smolt index class
   h. Mortality (did fish die?)
   i. **Existing marks/tags
      i. **Mark/tag type
      ii. **Mark/tag location
      iii. Allow for multiple marks/tags
   j. *Number of fish represented
      i. *Whether number of fish is a count or an estimate
   k. Ability to assign number to individual fish so genetic samples, scales, or other information can be tied to the record.
   l. Ability for user to define other parameters (though they would not be part of standard analysis routines)
4. **Fish marked and released for efficiency test.
   a. Source of fish
   b. *Taxon
   c. *Rearing type (wild vs. hatchery)
   d. Life stage
   e. *Length
      i. *Type (fork length, total length, etc., with default for each species definable)
      ii. Units
   f. Weight
      i. Units
   g. *Marks/tags applied
      i. *Mark/tag type
      ii. *Location
      iii. Allow for multiple marks/tags
   h. *Date and time marks/tags applied
   i. *Date and time fish released
   j. *Release was:
      i. *During daylight
      ii. *Dusk
      iii. *During dark
   k. Whether fish survived to release
   l. *Number of fish represented (each record defaults to 1, since they will probably be measured, but allow for unmeasured groups)
   m. Trapping site(s) release is meant for
   n. *Release location
      i. "Exact"
      ii. *Relative to trapping site being tested (stream distance)
      iii. *Release site in channel (right bank, left bank, center, dispersed, etc.)

5. Environmental covariates
    a. *Stream discharge or analog. Examples include
        i. Stream discharge estimates from USGS stream gage or other source
            1. Location of gage
            2. Gage ID
        ii. Staff gage readings
            1. Gage location
                a. Exact
                b. Relative to trap site
        iii. Water velocity
            1. Location of water velocity measurements (usually done at front of trap, which is questionable in my opinion Approaches an asymptote as flow resistance of the trap increases due to the trap's pores plugging and also as stream discharge continues to increase. Thus not real useful at higher flows or higher debris levels.)
        iv. Proportion of stream cross-sectional area that is sampled.
        v. Proportion of stream discharge that is sampled.
        vi. Water depth at trap
    b. *Water temperature
    c. *Turbidity
    d. *Amount of debris
6. Trap cone rotations per day (closely correlated with stream discharge)
    a. Used to estimate how long trap fished on those days when it stops functioning. Used to impute catches on days when the trap stopped functioning.
    b. Used along with rotation counters on the trap.
7. Nominal capture date (because sometimes trap checked >1 time per day due to high debris load or large number of fish) For example, fish caught when checking trap at 4:00 on February 17 would be applied to February 18 morning nominal date.
8. Trap operations
    a. Need to record because such changes essentially equal a different apparatus used to catch fish and thus the efficiency tests may not apply to these altered configurations.
        i. When trap is fished in "half-cone" or other non-standard configuration.
        ii. When fishing switches from side-by-side dual traps to a single trap.
        iii. Is trap in thalweg?
        iv. Is trap moved within a site?
        v. Is site changed?

Appendix D.

This appendix contains the Shannon (2009) report.

Shannon, C. 2009. Status report: Assessment of Central Valley rotary screw trap databases, October 2009. Unpublished report. Pacific States Marine Fisheries Commission, Redding, California.

Status report: Assessment of Central Valley Rotary Screw Trap Databases
October 2009
Connie Shannon, PSMFC

In fulfillment of Phase II RST database standardization and integration, the following was found to be true with regard to the existing IEP type databases that have been reviewed thus far (see Figure 1).

Efficiency trials and production estimates
The IEP database was never built to store the complete suite of data associated with the capture efficiency tests. The results of efficiency tests are evaluated, and production estimates developed, outside of the IEP database environment (i.e. Excel spreadsheets). The raw efficiency trial data, (i.e. release date, number of test fish released, test mark, and recapture count) are incomplete or absent from the databases examined thus far. This raw efficiency data would be required if production estimates were to be re-evaluated.

In addition, the databases do not indicate which capture efficiency trial was used to expand the raw catch records found within the database (i.e. link from catch to trial number or test result). There is no linkage between the efficiency trial and the time period for which it was applied in order to develop the reported production estimate. Thus, if it was desired to do so, it may not be possible to re-create the production estimates presented in annual reports.

Discussions with various program leads as well as a review of examples provided by some programs indicates that capture efficiency information is being compiled and stored in a variety of formats unique to each project. Data for each program is also typically parsed by year and may vary in format from year to year. Most programs have not forwarded these efficiency/production estimate worksheets but refer to the annual reports instead. Since these worksheets were not intended for anything but internal (and possibly one time) use, they will most likely be difficult to understand and require much time and effort to standardize. The effort and expense required to complete such an effort should not be underestimated.

Regarding the marking and releasing of test fish
In preparation to release fish for efficiency tests, fish are visibly marked so that they can be identified in the catch later. While some of the databases do contain information about fish marking and fish released for efficiency trials, the information appears to be incomplete as there are many more recovered marks than are indicated by records of fish marked and released. In addition, many of the release records were found to contain null count values. These do not indicate how many fish were released.

There is a correlation between records of fish marked (a separate table) for release and the numbers of fish actually released when this information is present in the databases. And it is possible in some cases to decipher the number of fish released in cases where the release record counts are null. This can be done by pairing the release code and year with information in the marking table. Release record counts derived this way, by summarizing the number of fish marked, do not account for mortalities that occurred before at the time of the release; however, in review of release records that contain both a count for live releases and a count for mortalities at the time of release, mortalities appear to be minimal (this assumes mortalities are being reported accurately). Therefore synthesizing counts from the marking table in an effort to complete release records can be done in a few of the databases, but only if we assume no mortality occurred after the pre-release count.

Many of the programs do not record release or marking information in the databases at all but choose instead to maintain them separately, often in Excel. In some cases the Excel spreadsheets combine multiple releases (with various release codes and marks) into a single record for a given time period. Original release codes and marks are often not present in these data. This is apparently because once the efficiency trial is completed; the marks and release codes were not needed to develop the efficiency estimates.

The Excel spreadsheets record a summary of the efficiency trial data together with associated catch for the time period. The efficiency estimates are developed and applied to the RST counts within these Excel spreadsheets. The results are then used for reporting purposes. This summarized data will not work for CAMP's purposes because all of the raw data associated with the efficiency tests is needed if the production estimates are to be re-evaluated.

While there is no direct connection from the re-captured fish to the release records that are included within the databases, there is even less of a connection between the summarized results of the efficiency trials and associated catch data that are stored outside the databases, (i.e., all the steps/data are not in the spreadsheets). So these summarized data cannot be used to fill in the data gaps between release, recapture, and catch data that are incompletely recorded within the databases.

Catch data records

The existing catch data may have value down the road; for example, we may be able to use future efficiency data to estimate past production. There may be some value in compiling these raw catch data into a standardized format.

The raw catch data are more likely to be in a format that can be compiled into a standardized format. Problems with compiling these data include what to do with marked fish in the catch table.

Recaptured fish are indicated in the catch table by the mark that was observed at the trap. Efficiency test fish should not be counted as fresh captures, however, since there is no complete record of releases (and the mark they were given) it is impossible to tell if any of the marked fish were not marked for efficiency tests at the trap where they were recovered. Some of these fish may have been marked for an upstream RST or another study and therefore should not be subtracted from the catch.

The number of fish with marks other than those used for relevant efficiency releases would seem to be relatively few, and so, all records of marked fish in the catch (does not include CWT or other hatchery marks) would most likely need to be removed from the raw catch data. This introduces a small amount of error.

While many of the projects have offered the catch data in their original format (raw catch), other projects have offered summarized catch data. These are catch data that have been grouped by week or by day. Since the data are summarized, the length measures represent averages of length, or are reported as minimum and maximum length and so they are not identical to the detailed 'raw' catch data. However, the numbers are accurate and so the data could still be combined with the raw catch data from other databases.

For catch data we would end up with a rather large flat table that contained both detailed 'raw' catch data as well as catch data summarized by week or some other time period. This would provide 'Catch' for all of the programs that share this type of information with us. Centralizing the data into a single database minimizes database management. The alternative would be to retain the data in watershed specific databases which would require effort to prevent application versioning issues (i.e. additions or modifications to one database would need to duplicated in all databases.)

The CAMP program would prefer the raw unsummarized data including individual fish lengths. The length data may provide the best chance for discriminating between wild and hatchery fish in watersheds where both kinds of fish were present. This assumes that hatchery fish are markedly different in size than wild fish, and we know the size range of released hatchery fish, and when those fish were released.

Standardizing RST data

It may be possible to develop a standardized way to represent the data provided in annual reports used to develop juvenile production estimates. The summarized catch paired with efficiency and calculated production estimates could be represented in a single table by trap, date, species, and race. This table would not include much detail but would indicate the juvenile production estimates as they are currently being reported at end of season.

Because the release (and marking) information in the databases is incomplete at best, and since there is no way to recreate the raw release (and marking) information from the Excel spreadsheets, compiling the summary data from the Excel spreadsheets and annual reports may be our best alternative.

Alternatively we could develop what would be the optimum database structure for storing the complete suite of catch and efficiency test data, review completed annual reports and extract relevant data in a data mining exercise, and then go back to the individuals who collected the data in priority watersheds to fill in any missing data.

This option would require an intensive effort from everyone involved. We may be met with resistance from RST program leads since this would require a lot of their time, or that of their staff. Program leads may also feel slighted that we are attempting to re-evaluate their work. Some of the traps have been in operation for over twenty years. Staff may have retired or moved. Current program staff may not be able to answer questions, or the accuracy of the answers may be questioned. Thus despite much time and expense, we may never have a complete dataset, we may not be able to verify the accuracy of the resulting data, and finally, reevaluating these data may not yield juvenile production estimates that are any more accurate than those developed originally. For these reasons this option is not recommended.

Figure 1

RST Database Review Status Report 10/23/2009								
Priority	**Watershed**	**Received**	**Contact Initiated**	**Triage**	**Documentation**	**Data Summary**	**Flat Catch**	**Flat Release**
1	Sacramento River, Red Bluff Diversion Dam	10/22/2009	Yes	Yes	No	No	No	No
2	Stanislaus River (Caswell SP)	1/7/2009	Yes	Yes	Yes	Yes	Yes	Yes
3	Clear Creek	10/22/2009	Yes	Yes	No	No	No	No
4	American River		Yes	No	No	No	No	No
5	Battle Creek	10/22/2009	Yes	Yes	No	No	No	No
6	Mokelumne River, Woodbridge	9/21/2009	Yes	Yes	Yes	Yes	No	No
7	Feather River	10/22/2009	Yes	No	No	No	No	No
8	Yuba River		Yes	No	No	No	No	No
9	Sacramento River, Knights Landing	10/2/2009	Yes	No	No	No	No	No
10	Butte Creek		Yes	No	No	No	No	No
11	Tuolumne River (Grayson Ranch)	3/25/2009	Yes	Yes	Yes	Yes	Yes	No
12	Tuolumne River (Shiloh Bridge)	3/25/2009	Yes	Yes	Yes	Yes	Yes	No
13	Mill Creek		Yes	No	No	No	No	No
14	Deer Creek		Yes	No	No	No	No	No
15	Stanislaus River (Oakdale)	09/29/09	Yes	Yes	No	Yes	Yes	No
16	Big Chico Creek		Yes	No	No	No	No	No
17	Merced River (Hatfield SP)	1/7/2009	Yes	Yes	Yes	Yes	Yes	Yes

Figure 1 notes: Triage includes an initial assessment of database structure and data content. Documentation includes tables and reports that detail column usage and variable descriptions. A data summary table is produced that totals the number of catch records and release records per year included in each database and tables from each of the databases are compiled into a central database. Flat Catch and Flat Release tables are developed from the original databases in an effort to simplify content. These flat tables are merged into a central database.

Appendix E.

This appendix contains the McDonald and Banach (2009) report.

McDonald, T., and M. Banach. 2009. Feasibility of unified analysis methods for rotary screw trap data in the California Central Valley. Task B Report. USFWS Coop. Agreement #81420-8-J163. 16 pages.

Feasibility of Unified Analysis Methods for Rotary Screw Trap Data in the California Central Valley.

Task B Report

USFWS Coop Agreement #81420-8-J163

Trent McDonald, Ph.D.
Western EcoSystems Technology, Inc.
2003 Central Avenue
Cheyenne, Wyoming 82001

and

Mike Banach
Pacific States Marine Fisheries Commission
205 SE Spokane Street
Portland, OR 97206

2009-12-28

Funded by the U.S. Fish and Wildlife Service's
Comprehensive Assessment and Monitoring Program (CAMP)
USFWS Cooperative Agreement No. 81420-8-J163

Introduction

The U.S. Fish and Wildlife Service's Comprehensive Assessment and Monitoring Program (CAMP) produces a variety of reports that summarize and tabulate salmonid data from collection sources in California's Central Valley. To prepare certain of these reports, in-depth statistical analyses and the development of complex databases are required. Through a cooperative agreement, CAMP contracted the Pacific States Marine Fisheries Commission (PSMFC) and a statistical subcontractor (Western EcoSystems Technology; WEST, Inc.) to assist in evaluating the feasibility of developing a comprehensive data collection, storage, and analysis system for information collected from rotary screw traps in the Central Valley. The ultimate purpose of such a system would be to document and understand changes in the abundance of juvenile salmon in the Central Valley. This feasibility study represents Phase 1 of a planned three phase program which may ultimately result in timely and defensible valley-wide estimates of juvenile salmon abundance.

The abundance of juvenile Chinook salmon (*Oncorhynchus tshawytscha*) has been monitored at 12 or more sites in the Central Valley using rotary screw traps (RSTs) for approximately 13 years. Trapping activities at most RST sites has routinely occurred during all or a part of the year since 1995. Much of the collected data have never been presented in report form, and different analytical techniques have been used to estimate fish numbers passing the traps. Separate or non-existent reports and different analytical techniques make it difficult, if not impossible, to understand valley-wide long-term trends in juvenile salmon production. These factors also confound the ability to understand how restoration activities influence juvenile and adult salmon production. To address the difficulties inherent in trend detection under the current system a single comprehensive, multi-faceted data acquisition, storage, and analysis system is needed. This system, if built, would be designed to collect and manage screw trap data, as well as produce statistically robust and repeatable estimates of juvenile Chinook abundance based on RST catch data.

Because the development of such a system is inherently challenging, CAMP determined that a feasibility evaluation was the appropriate first step. This Phase 1 - Task B report addresses the feasibility of implementing uniform data analysis methods to estimate abundance of juvenile Chinook salmon across California's Central Valley. This report is part of a larger feasibility report that includes recommendations for a comprehensive data entry and management system. The uniform analysis methods include algorithms for estimating the abundance of different life stages (fry, parr, smolts, and yearlings) and runs (fall, late fall, spring, and winter). These algorithms are designed to be applicable to all Central Valley RST data, thus unifying estimation methods and making comparison among sites easier.

Activities

PSMFC personnel spoke or corresponded with: Ayesha Gray (Cramer Fish Sciences), Connie Shannon (PSMFC / California Department of Fish and Game), Doug Burch (California

Department of Fish and Game programmer), Doug Threloff (USFWS-Sacramento, CAMP program coordinator) , Michelle Workman (formerly with the East Bay Municipal Utility District and now with the U.S. Fish and Wildlife Service (USFWS)), Liz Cook (formerly with California Department of Water Resources). Meetings attended by Doug Threloff, Mike Banach (PSMFC fisheries biologist), Greg Wilke (PSMFC programmer), Trent McDonald (West, Inc. statistician and programmer), Kellie Whitton (USFWS-Red Bluff biologist), Jim Earley (USFWS-Red Bluff biologist), David Colby (USFWS-Red Bluff biologist), Bill Poytress (USFWS-Red Bluff biologist), and Felipe Carrillo (USFWS-Red Bluff biologist). Field visits to the Red Bluff Diversion Dam, Battle Creek, and Clear Creek screw trap sites. PSMFC also examined databases provided by Cramer Fish Sciences and the USFWS Red Bluff office. These databases contained RST data collected on the Stanislaus River, Battle Creek, and Clear Creek. Recent Battle Creek and Clear Creek annual reports were examined to determine data analysis routines used by USFWS Red Bluff office. Analysis routines used by the USFWS Red Bluff office for RSTs located at the Red Bluff Diversion Dam were demonstrated by Felipe Carrillo.

In addition to corresponding with most of the people listed above, personnel at WEST Inc. reviewed the following documents relating to RST data and estimation techniques:

- Battle Creek RST report for Oct 2005 - Sep 2006;
- Red Bluff Diversion Dam RST reports for 2005 and 2006;
- Clear Creek RST report for Oct 2006 - Sep 2007;
- Mill and Deer Creek RST report for 1999;
- Butte and Big Chico Creek RST report for 2006-2007;
- Knights Landing RST report for Sep 1999 through Sep 2000;
- Feather River RST report for 2002 – 2004;
- Yuba River RST report for 2004 – 2005;
- Lower American RST River report for Oct 1998 – Sep 1999;
- Lower Mokelumne River RST report for Dec 2005 – Jul 2006;
- Lower Stanislaus River RST reports for 1999 and 2008;
- Lower Tuolumne River RST report for 2003;
- Lower Merced River RST report for 2008;
- the quantitative Appendix of the 2000 Red Bluff Diversion Dam RST report by Martin;
- "Determination of Salmonid Smolt Yield with Rotary-Screw Traps in the Situk River, Alaska, to Predict Effects of Glacial Flooding" by Thedinga et al (1994, *North American Journal of Fisheries Management,* p. 837-851);
- the 2000 review of Red Bluff Diversion Dam and Stanislaus River RST methods conducted by L. McDonald and S. Howlin;
- the 2000 review of Red Bluff Diversion Dam and Stanislaus River RST methods conducted by J. Skalski;
- response of D. Neeley to comments made by McDonald, Howlin, and Skalski during their review of the Red Bluff Diversion Dam and Stanislaus River RST methods; and
- the "Rotary Screw Traps and Inclined Plane Screen Traps" chapter of the American Fisheries Society protocol manual.

WEST Inc also reviewed the following databases provided by PSMFC:

- Cramer Fish Sciences databases containing RST data collected at Caswell State Park and Oakdale trapping sites on the Stanislaus River, and Hatfield State Park on the Merced River.
- USFWS databases containing RST data collected from the Lower Clear Creek, Upper Clear Creek, and Upper Battle Creek RST's.

Minimum Field Data

The list of variables in this sub-section represents a minimum set of field measurements to be collected at each site. From these data, other quantities (such as catch, efficiency, % water fished, etc.) can to be estimated and in turn used to estimate abundance. Additional field data pertinent to a site may be collected. Additional field data may be collected if they are useful for purposes other than abundance estimation, or if they pertain to a unique feature of the site and can explain variation in daily catch or trap efficiency. Additional data that might be pertinent to a site include staff gauge readings of water depth, stream width, fish weight, etc.

Field data are of three basic types: (1) trap placement data, (2) trap check data, and (3) efficiency trial data. Within trap check data, four classes of data exist: (a) trap operating characteristics, (b) physical environment measures, (c) fish counts (1 value per trap check), and (d) individual fish measures (multiple values per trap check). The minimum set of variables to be measured for each type of data is listed below.

Trap Placement Data

Any time a trap is turned on (e.g., after installation or after movement) or turned off (e.g., prior to removal or prior to movement), the following data should be recorded:

1. Trap ID
   Description: Manufacture's serial number or other unique code associated with the trap. This number should be used to identify the trap for the trap's entire lifetime. This number should not be changed or re-assigned to another trap.
2. Site code
   Description: Unique ID of the overall stream location, e.g., stream name and river mile. If trap was installed or turned on, this is the code for the trap's location after Date and Time (below). If the trap was pulled or turned off, this is the code for the trap's location prior to Date and Time.
3. Fishing location
   Description: Unique ID of fishing location *within* the site. For example, '01', '02', or '03' if there are 3 fishing locations at a site. If there is only one fishing location at the site, this number assigned should be '01'.
4. Date
   Description: Date of the change in trap status. Date that the trap began fishing, or date that trap stopped fishing.
5. Time
   Description: Time of the change in trap status. Time trap began fishing, or time trap stopped fishing.

6. Fishing?

   Description: A binary Yes/No variable. Yes = trap was fishing after the above Date and Time, No = trap was not fishing after the above Date and Time.

Every trap must have a [*site, fishing location, date, time*] quadruplet corresponding to when it began fishing, and a [*site, fishing location, date, time*] quadruplet when it stopped fishing, unless the trap remains fishing on the current date. *Trap ID* can be used to lookup cone diameter, max cone depth, and other characteristics of the trap. S*ite* can be used to lookup characteristics of the overall installation, such as river mile, latitude, longitude, etc. *Site* and *Fishing location* can be used to lookup characteristics of the trap's specific locations, such as channel location (thalweg, right bank, left bank, etc), bottom type, etc.

Trap Check Data

A trap check occurs when a RST is visited and captured fish are processed. The exact schedule of trap checks is left to the biologists in charge of each program, and can vary from RST to RST. Ideally, traps will be checked every day during the season when the species of interest is expected to be in the river. When traps are not checked on a day, data for that day will be treated as missing (see imputation method described in Abundance Estimation Methods).

At a minimum, the following data should be collected every time a trap is checked:

1. Trap operating characteristics:
   a. Site code (to match site code in Trap Placement data, e.g., stream name and river mile)
   b. Fishing location (to match fishing location in Trap Placement data)
   c. Date (of trap check)
   d. Time (of trap check)
   e. Cone rotation counter reading
   f. Cone rotation speed (rpm)
   g. Submerged cone depth (meters, measured from water surface to lowest part of submerged cone or read from gauge on trap), and
   h. Trap retention rate (intra-trap catch rate, depends on baffle configuration, usually 50% or 100%)
2. Physical environment variables
   a. Water velocity (m/s, near trap, preferably near front of cone)
   b. Water temperature (°C, in front of cone at depth)
   c. Debris occlusion (%, visual)
   d. Turbidity (at least Secci depth)
   e. Average flow between last check and current check (cubic meters per second, measured at most representative river gauge), and
   f. River gauge ID (Identifier of river gauge used to calculate the above average flow)
3. Fish counts
   a. Total number of unmarked fish caught (count or estimate)

    b. Estimate? (Yes/No, Yes = number of unmarked fish is an estimate, No = number of unmarked fish is a complete count)
    c. Total number of marked fish caught, and
    d. Total number of measured unmarked fish (number in subsample, if subsample was taken)
4. Individual fish data:
    a. For marked fish:
        i. Mark description code (e.g., AF-CL-BB = adipose fin – clipped – stained Bismarck brown, must be sufficient to identify the release group)
        ii. Fork length, and
        iii. Species
    b. For measured unmarked fish:
        i. Fork length
        ii. Species, and
        iii. Visual smolt index (0,1,2,3,4,5)

It is assumed that the trap has been fishing between date and time of the previous check until date and time of the current check. Cone rotations between previous check and current check, times rotation speed, will be used to compute amount of water sampled.

Efficiency trial data

Efficiency trials consist of releasing a known number of (uniquely or batch) marked fish upstream of a RST. By noting the number of marked fish from each release that were later captured in the RST, efficiency (probability of capture) can be estimated. Like trap checks, the exact schedule efficiency trials is left to the biologist in charge of each program. Ideally, it will be possible to release small batches of marked fish every day so that efficiency trials occur continuously throughout the season. However, large numbers of efficiency trials are not possible at many RSTs. In these cases, two to three efficiency trials per week are recommended. Less frequent efficiency trials are acceptable because probability of capture will be modeled after the season (see Abundance Estimation Methods below).

For every efficiency trial, the following should be recorded:
1. Date (of efficiency trial release)
2. Time (of efficiency trial release)
3. Dark? (Yes/No, Was sun down during release?)
4. Release location code (e.g., stream name and river mile)
5. Location of release in channel (e.g., LB, CC, RB, etc. for left bank, center current, right bank, etc.)
6. Nearest downstream RST site code
7. Distance from release location to nearest RST (river km)
8. Mark description code (to match fish data above)
9. Number of marked fish released.
10. Holding time (hours)
11. Fish source (wild, hatchery, etc.)
12. Species
13. Number of fish measured
14. Fork length for every measured fish

Abundance Estimation Methods

This section contains recommendations for statistical estimation of abundance from data collected by RSTs in the Central Valley. These estimation techniques are designed to utilize the minimum set of field data (previous section) and are intended to be applicable to all RST sites that collect this data. The recommended analysis is widely applicable because it applies to all sites that collect the minimum set of field data. The recommended analysis is stable in the sense that, when appropriate, sites and years can be combined to improve model estimation. Such combination of data would likely require judicious use of covariates (such as 'site' and 'year' variables), but can be done in some cases.

The analysis leaves open the exact protocol by which researchers measure variables contained in the minimum set of data. Ideally, each site can provide unbiased and precise (low variance) estimates of the basic variables listed in the previous section. This means, ideally, that each site could provide unbiased estimates of counts, velocity, cone depth, rotations, etc. If estimates of the basic variables are unbiased, the abundance estimates produced using methods in this section should also be unbiased. If unbiased estimates of the minimum dataset cannot be constructed, at least consistently measured estimates should be used. Readings from a poorly calibrated velocity or temperature meter is an example of a consistently measured, yet biased estimate. Consistently measured basic variables, when used in abundance estimation, will result in an index of juvenile abundance at the site that can at least be assessed for trends.

As called for in the cooperative agreement between CAMP and PSMFC, the methodological recommendations contained in this section were designed to estimate abundance of all life stages and runs of juvenile Chinook salmon. However, the methods outlined here are not specific to life stages or runs of a single species. The methods are applicable to all species, life stages, and runs provided similar and adequate data on these populations can be collected. The only caveat to wide-spread application of these methods is that the estimator's performance, while theoretically sound, may not perform well when samples sizes are low. A prudent amount of faith should be placed in abundance estimates produced by these methods for species other than Chinook.

General Estimation Approach

In his review of methods at Caswell State Park on the Stanislaus River, Skalski (2000) mentioned the virtues of a *design-based* estimation approach, and the vices of a *model-based* estimation approach. The general definition of a design-based approach is that the analysis relies on a few simple assumptions about the structure of the data and uses replication of measurements or samples as the basis for assessing variation. For example, if RST catch and trap efficiency could be assessed every day without error, a design-based approach would estimate abundance that day as catch divided by efficiency. Variation in abundance across days would be used to construct confidence intervals. Design-based approaches typically involve relatively simple estimators, like means, ratios, and products. On the other hand, model-based approaches make relatively weighty assumptions about the data structure, or what influences a particular variable, and uses these assumptions as the basis by which they assess variation. For example, a model-based approach could assume that the mean of a response variable follows a regression relationship, and that errors in the regression relationship follow a normal distribution. Model-based estimators can become quite complicated depending on the complexity of the situation and assumed model.

The virtues of a design-based approach include its simplicity and lack of assumptions (Skalski, 2000). It is hard to argue against properly designed and executed design-based estimates (Olsen and Smith, 1999). However, the two biggest vices of a design-based approach are its inability to include measurement error and a high data requirement that is generally required. Design-based approaches also have difficulty incorporating missing values into the analysis. The virtues of a model-based approach include its ability to incorporate measurement error, lower data requirements, and the ability to make estimates outside the data range (extrapolation) when necessary. However, the main vice of a model-based approach is the fact that its assumptions will always be violated to some extent and thus estimates are easy to question. Model-based approaches can use outputs of a model as substitutes for field data, thereby giving researchers the feeling that results are "far from" or "insulated from" the original data.

The abundance estimation procedure described here is neither fully design-based nor fully model-based. The approach advocated here uses raw data when it is appropriate, but assumes a flexible non-linear model for catches and efficiencies when raw counts or efficiencies do not apply to an entire interval between checks. The non-linear model allows estimation of daily abundance during times when the trap was not operating or when an efficiency trial has not been done for quite some time. In utilizing a model, the data collection requirements are reduced relative to a fully design-based approach because fewer checks and efficiency trials can be performed once the model is established and stable. If the models continue to be developed over time, accuracy and precision will increase through time. Utilizing a model for certain tasks also smoothes a portion of the random noise inherent in measurements, thus making estimate more stable.

The approach advocated here uses raw catch data when it is available, and model based estimates when raw catch is not available. On days when a RST check meets protocol, raw counts are inflated by a current estimate of trap efficiency without aid of a model for catch. A trap check 'meets protocol' if the interval between checks was 24 ± 2 hours (or, some other interval surrounding 24; in the remainder, 2 hours will be assumed) and the trap was in operation for that entire period. When counts are not available for a day (check does not 'meet protocol'), the approach employs a generalized additive model (GAM) (Hastie and Tibshirani, 1990) to estimate catch as a function of study covariates. To estimate trap efficiencies, the approach uses a second GAM estimated from past and current efficiency trials. Both of these GAMs can be functions of time (date of season) and other factors such as flow, percent flow sampled, turbidity, distance from trap to release site, etc.

Several RST operations in the Central Valley already employ models to infer various quantities when they are missing. For example, a 5-day moving average with a triangular weight function is used on data collected at Caswell State park to estimate catch on days when it is missing. Moving averages are special cases of a GAM model. Another model typically employed by RST operations is to assume trap efficiency remains constant between efficiency trials. On days when an efficiency trial has not been conducted, researchers typically use efficiency from the last trial to inflate raw counts.

Abundance Estimation

The basic quantities contained in this sub-section are estimable from the minimum set of field data collected at a site. At most sites, these quantities can be estimated from historical data and thus past estimates could be re-computed or updated using this methodology if necessary. In other cases, these estimates cannot be computed from historical data. At those sites, data collection procedures will need to change if these procedures are to be applied in the future.

The two basic quantities needed to estimate abundance at every site are:

- $\hat{c}_{ij}$ = either the enumerated or estimated catch of unmarked fish of a certain life stage in trapping location *i* at the site during the 24 hour period indexed by *j*. For example, $\hat{c}_{23}$ = estimated catch at the 2nd trapping location during day 3.
- $\hat{e}_{ij}$ = estimated trap efficiency at trapping location *i* of the site for a certain life stage during the 24 hour period indexed by *j*. For example, $\hat{e}_{23}$ = estimated efficiency at the 2nd trapping location during day 3.

Note that, for notational convenience, a subscript for site is not present in the above quantities. It is assumed that estimates will be computed separately for each site, thus eliminating the need for a site subscript.

Assuming the above quantities can be computed, an estimate of the number of fish passing the trap during the 24 hour period indexed by *j* is,

$$\hat{N}_{ij} = \frac{\hat{c}_{ij}}{\hat{e}_{ij}} \quad (1)$$

Estimation of $\hat{c}_{ij}$

The estimate of catch, $\hat{c}_{ij}$, will be computed in one of three ways. First, if the interval between check *j* and check $j - 1$was 24 ± 2 hours and the trap operated properly for the entire period, $\hat{c}_{ij}$ will be the total catch of unmarked fish in the trap at check *j*. Note that the amount of time the trap operated properly is estimated as the difference in rotation counter readings multiplied by cone rotation speed averaged over the two checks. When the check meets protocol, $\hat{c}_{ij}$ can either be a complete enumeration of captured fish, or an estimate based on random subsampling when too many fish are captured to enumerate.

The second method of computing $\hat{c}_{ij}$ will be used when the trap fishes for less than 22 hours. If the trap fished for less than 22 hours between check *j* and check $j - 1$, the fish count at time *j* will be adjusted using a diurnal logistic regression model. This diurnal logistic regression model will utilize efficiency trial data to estimate the proportion of a typical 24-hour fish count passing in a given period of time. To estimate this logistic regression, data from many efficiency trials and multiple checks will be used. Assuming m_i is the number of marked fish captured within 24 hours of release during the i^{th} efficiency trial, the logistic regression will estimate the proportion of m_i captured within *t* hours ($t < 22$) of release as a function of other covariates like day-night, flow, date, etc. To do this, the trap check time of the m_i marked fish must be known, and the interval between release and check must vary from 0 to 24 over multiple efficiency trials. When a trap is checked t hours ($t < 22$) after the previous check, $\hat{c}_{ij}$ will be computed as

$$\hat{c}_{ij} = \frac{c(t)}{p(t)}$$

where $c(t)$ is the catch of unmarked fish in the t hours since the last check and $p(t)$ is the estimated (via logistic regression) proportion of a typical 24-hour catch caught within t hours under similar conditions. Until sufficient data is available to adequately estimate the logistic regression model, $\hat{c}_{ij}$ will be treated as missing when a full 24 hours has not been sampled. In this case, $\hat{c}_{ij}$ will be estimated using the GAM (below).

The third method of computing $\hat{c}_{ij}$ will be employed ,when $\hat{c}_{ij}$ is missing for some reason (i.e., trap fished for >26 hours between checks). In this case, $\hat{c}_{ij}$ will be predicted after the season using a Poisson GAM model fitted to the $\hat{c}_{ij}$ that met protocol. The additive portion of this model will be of the general form,

$$log(E[\hat{c}_{ij}]) = s(j) + \beta_1 x_{ij1} + \cdots + \beta_p x_{ijp} \quad (2)$$

where $s(j)$ is a smooth (spline) function of the day index (i.e., smooth function of Julian date), the x_{ijk} are covariates associated with trap i during day j, and the β's are estimated coefficients. In other words, the GAM has a non-linear smoothing component, $s(j)$, as well as a linear component, symbolized by the $\beta_k x_{ijk}$. The smoothing component requires choice of the degree of smoothing that the function should do. Automatic and objective choice of the smoothing amount should be done by generalized cross-validation, or similar established technique.

Estimation of $\hat{e}_{ij}$

Efficiency estimates at the *i-th* trapping location on day j will be computed from a binomial GAM, unless sufficient efficiency trials (≥3 per week) have been performed. If sufficient efficiency trials have been conducted, and the assumption of constant efficiency between trials is justified, efficiency from the most recent trial will be used for $\hat{e}_{ij}$. When the most recent efficiency is not appropriate, a binomial GAM fitted to past and current efficiency trials will be estimated and used to compute $\hat{e}_{ij}$. The additive portion of this GAM model will be of the form,

$$log\left(\frac{E[\hat{e}_{ij}]}{1 - [E[\hat{e}]_{ij}]}\right) = s(j) + \gamma_1 z_{ij1} + \cdots + \gamma_p z_{ijp} \quad (3)$$

where $s(j)$ is again a smooth (spline) function of the day index (i.e., smooth function of Julian date), the z_{ijk} are covariates associated with the efficiency of trap i during day j, and the γ's are estimated coefficients. Again, automatic choice of the smoothing amount should be by generalized cross-validation, or similar established technique.

The current abundance estimation methods employed at Red Bluff Diversion Dam utilize a linear regression model containing the proportion of flow sampled between checks (i.e., %Q) to estimate trap efficiencies. The linear model used at Red Bluff Diversion Dam is a special case of the GAM proposed here (i.e., no $s(j)$ and only one z). The GAM proposed here allows for non-linear smoothing and inclusion of additional factors that may influence efficiency. For example, %Q, turbidity, and distance from release site could all be incorporated in the linear or non-linear parts of the model. Note that the absolute accuracy of covariates in the model (e.g., flow, %Q, etc.) is not paramount. It is only paramount that covariates in the model be consistently and objectively measured. Because the GAM model is invariant to linear transformations of the covariates, a proxy for any covariate can be used provided it is a linear transformation of the desired covariate.

Estimation of $\hat{\hat{N}}_{ij}$

Once $\hat{c}_{ij}$ and $\hat{e}_{ij}$ are estimated, and $\hat{N}_{ij}$ has been computed, abundance estimates for the site should be computed by summing over trap locations. The total number of fish passing a particular site on day *j* should be computed as

$$\hat{N}_j = \sum_{i=1}^{n_{ij}} \hat{N}_{ij}$$

where n_{ij} is the number of trapping locations fishing at site *i* during day *j*. Abundance on day *j* can then be summarized in a number of ways. The estimates $\hat{N}_j$ can be plotted against *j* to visually assess trends. $\hat{N}_j$ can be summed over a week, month, or year to produce weekly, monthly, or annual estimates of abundance. The time series of $\hat{N}_j$ estimates can be subjected to further analysis to detect and quantify trends.

Confidence Interval Estimates

The abundance estimator $\hat{N}_j$ is a mixture of measured and modeled fish counts, as well as modeled trap efficiency values. This mixture makes variance computation by traditional methods difficult because they rely on formulas and approximations. Here, confidence intervals for $\hat{N}_j$ will be computed by parametric bootstrap or Monte Carlo methods. This method has been successfully used at Battle and Clear Creek to compute confidence intervals for their abundance estimates.

Fish counts derived from trap checks are subject to measurement error. For instance, it is possible for technicians to miss-count fish, miss-classify species, or miss-classify life history stage. However, the measurement error inherent in raw counts is tiny compared to the day-to-day and seasonal fluctuation in fish passage. Day-to-day and seasonal fluctuation in fish passage is natural process variation, sometimes called sampling variation to distinguish it from measurement error. Because measurement error in $\hat{c}_{ij}$ is tiny compared to other sources of error, raw counts will be treated as known constants. Similarly, the measurement error in raw efficiency estimates is tiny compared to process variation.

Modeled values of $\hat{c}_{ij}$ and $\hat{e}_{ij}$ are not constants, and variation of these predicted values from their respective GAMs will be included by the parametric bootstrap procedure described below. Values of $\hat{c}_{ij}$ that have been corrected for less than 24-hour fishing periods are not constants; however, it is assumed that there are relatively few of these values and that it will take some time before sufficient data exists to estimate the logistic regression. If the logistic regression has been estimated, and numerous $\hat{c}_{ij}$ have been corrected for less than 24-hour fishing periods, the coefficients of the logistic regression should be included in the parametric bootstrap method outlined below. In this case, coefficients of the logistic regression would be treated the same as coefficients from the Poisson or binomial GAM.

Coefficients in both the Poisson GAM and binomial GAM are maximum likelihood estimates. A mathematical fact about maximum likelihood estimators is that their distribution converges to a multivariate normal distribution as sample size increases. Let $\hat{\boldsymbol{\beta}}$ represent the vector of smoothing and linear coefficients in the Poisson GAM model for missing fish counts, and let $\hat{\boldsymbol{\gamma}}$ represent the vector of smoothing and linear coefficients in the binomial GAM for trap efficiency. The parametric bootstrap procedure assumes both of these vectors are approximately multivariate normal random vectors, i.e.,

$$\hat{\boldsymbol{\beta}} \sim MVN\left(\boldsymbol{\beta}, \hat{V}\left(\hat{\boldsymbol{\beta}}\right)\right)$$
$$\hat{\boldsymbol{\gamma}} \sim MVN\left(\boldsymbol{\gamma}, \hat{V}(\hat{\boldsymbol{\gamma}})\right)$$

where *MVN* stands for the multivariate normal density function, and $\hat{V}\left(\hat{\boldsymbol{\beta}}\right)$ and $\hat{V}(\hat{\boldsymbol{\gamma}})$ are estimated variance-covariance matrices from the GAM model. $\hat{V}\left(\hat{\boldsymbol{\beta}}\right)$ and $\hat{V}(\hat{\boldsymbol{\gamma}})$ will be estimated using the 2nd derivative of the likelihood, or the observed Fisher information matrix.

Given these assumptions, the parametric bootstrap procedure proceeds as follows:

1. Generate realizations from the multivariate normal distribution. Specifically, generate the random vector $\hat{\boldsymbol{\beta}}^*$ from a $MVN(\hat{\boldsymbol{\beta}}, \hat{V}(\hat{\boldsymbol{\beta}}))$ distribution, and the random vector $\hat{\boldsymbol{\gamma}}^*$ from a $MVN(\hat{\boldsymbol{\gamma}}, \hat{V}(\hat{\boldsymbol{\gamma}}))$ distribution. If a logistic regression equation is in use to correct for less than 24-hours of fishing between checks, a random *MVN* vector representing its coefficients should also be generated.
2. Evaluate the Poisson GAM model in Equation (2) using $\hat{\boldsymbol{\beta}}^*$ for all days with missing fish counts. This will result in the random realizations $[E[\hat{c}]_{ij}]^*$ for all days with missing fish counts.
3. Evaluate the binomial GAM model in Equation (3) using $\hat{\boldsymbol{\gamma}}^*$ for all days. This will result in the random realizations $[E[\hat{e}]_{ij}]^*$ for all days.
4. For all days with missing fish counts, generate random Poisson variables $\hat{c}_{ij}^*$ from *Poisson*($[E[\hat{c}]_{ij}]^*$) distributions.
5. For all days, generate random binomial proportions $\hat{e}_{ij}^*$ from *binomial*($[E[\hat{e}]_{ij}]^*, \bar{r}_j$) distributions, where $\bar{r}_j$ is the (rounded) average number of released fish in the two efficiency trials on either side of day *j* temporally.
6. Recalculate $\hat{N}_{ij}$ for all days via Equation (1), substituting randomly generated values where appropriate. Specifically, use observed values of $\hat{c}_{ij}$ on days when counts are present, and substitute $\hat{c}_{ij}^*$ for $\hat{c}_{ij}$ on days when counts are missing. Substitute $\hat{e}_{ij}^*$ for $\hat{e}_{ij}$ on days when efficiency has been estimated from the binomial model. This results in a random series of abundance estimates for trap *i* of a particular site. Label these random estimates $\hat{N}_{ij}^*$.
7. Recalculate abundance for the site (i.e., $\hat{N}_j$) using the $\hat{N}_{ij}^*$. This results in a random time series of $\hat{N}_j^*$ values (for all *j*). Summarize these $\hat{N}_j^*$ values the same way they were summarized to compute the original estimates (i.e., sum over weeks, months, years, etc.).

8. Store the time series of $\hat{N}_j^*$ values and any derived summarizations.
9. Repeat the above steps 5000 times. This results in 5000 random realizations of $\hat{N}_j^*$ and subsequent summaries.
10. Finally, construct 90% confidence intervals as the 5th and 95th percentiles of the appropriate set of 5000 random abundance values. Specifically, the 90% confidence interval for $\hat{N}_j$ extends from the 5th percentile to the 95th percentile of the distribution of 5000 $\hat{N}_j^*$. Similarly for the confidence intervals on subsequent summarizations of $\hat{N}_j$. Error bands for visual displays of $\hat{N}_j$ can be computed by connecting the 5th and 95th percentile values in a graph of $\hat{N}_j$ through time.

A virtue of this parametric bootstrap technique is that it relies on only three parametric assumptions, and does not approximate any variances of derived estimators. The parametric assumptions this procedure makes are (1) missing fish counts follow a Poisson distribution, (2) efficiency values follow a binomial distribution, and (3) coefficients in both GAMs follow a multivariate normal distribution. A vice of this technique is that because it does not involve a mathematical formula, it must be computed using a conceptually simple but complex computer program. Note also that in order to carry out the computation, all covariate values must be available to evaluate the GAM models.

Trend Detection

There are two types of trend that can be detected from the time series of abundance estimates outlined above. The first type of trend is *abrupt change* that happens in a very short period of time (e.g., 1 or 2 years). The second type of trend is *long-term steady changes* in abundance that tend to move the mean in a single direction. Due to the high variability inherit in most juvenile production estimates, abrupt change is difficult to detect. Analyses to detect abrupt change can be run, but they will not be discussed here. It is assumed that long-term steady changes are of interest and an analysis designed to detect such trends will be discussed below. It should be kept in mind that the number of analyses that could be used to detect "trend" of some kind is large. The best analysis to detect trend is often a function of the specific objectives of the analysis and particulars of the data set being analyzed. In this section, a generic trend detection analysis (regression over time) will be described. It is hoped that this analysis will be applicable to a wide range of situations.

Detection of long-term trends can be divided into 2 inference scenarios. One inference scenario utilizes data from a single site and makes inference to parameters specific to that site. The other inference scenario assumes data from multiple sites within a region will be pooled to make inference about a parameter defined on the region. These latter inferences are called region-wide. Because multi-site region-wide trend detection analyses are generally extensions of single-site trend detection analyses, and because it is anticipated that single-site trend analyses will be more common, only single-site analyses will be discussed here. A qualified statistician should be consulted when multi-site trend detection analyses are to be performed.

It is assumed that trends in annual juvenile production are of interest. This assumption implies that total annual production will be the primary response of interest. It is assumed that an estimate of the standard error of annual production is available (see Confidence Interval Estimation above).

Long-term trends are estimated and detected using a mixed or fixed effect linear model and testing for the presence of non-zero slope coefficients. In matrix notation, a simple fixed effect model with no covariates (other than time) will be of the form,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}$$

where **Y** is the vector of annual production estimates,

$$\mathbf{X} = \begin{bmatrix} 1 & year_1 \\ 1 & year_2 \\ 1 & year_3 \\ \vdots & \vdots \\ 1 & year_n \end{bmatrix},$$

and

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix},$$

is a vector unknown coefficients to be estimated, and **E** is a vector of unknown random errors. The $year_i$ values in **X** are the actual years for each production estimate (e.g., 2006, 2007, 2010, etc.). If production was not estimated in a particular year, that year would not appear in **X**. Consequently, n is the number of data points, not the number of years that the overall monitoring program has been collecting data.

The above model assumes that the long term trend at a site is linear, but linearity is not necessary. Linearity of trend is not necessary because curvilinear or polynomial trends can be fitted and their coefficients tested for equality with zero. If auxiliary variables, such as mean temperature, flow, ocean conditions, etc. are correlated with annual production, these covariates can be incorporated into the model to explain variation and improve precision. If additional covariates are included, additional columns would be appended to **X**.

If production estimates are approximately normally distributed and residuals of the model are uncorrelated, standard least squares methods can be used to estimate and test whether the slope parameter in $\boldsymbol{\beta}$ is non-zero. If the slope is significantly different than zero, significant trend has been detected. If production estimates are not approximately normal, but residuals are uncorrelated, generalized linear model (GLM) estimation routines can be used to estimate and test whether the slope is zero. If production estimates are approximately normal, and residuals are correlated through time or space, mixed effect linear model estimation techniques can be used to estimate $\boldsymbol{\beta}$ and test for trend. Finally, if production estimates are not approximately normal, and residuals are correlated through time or space, generalized mixed linear model estimation techniques can be used. Alternatively, bootstrap methods can be used to test $\beta_1 = 0$ in the uncorrelated case, and block bootstrap methods (Lahiri, 2003) can be used in the correlated case. Bayesian analyses for each of the above cases are also available (consult a qualified statistician).

Conclusions

A unified data analysis procedure is feasible for the RST program in the Central Valley of California. Most RST sites are already collecting the minimum set of data required to carry out the estimation procedure set forth above. The data base, while complex, need only house the minimum set of variables to be useful for estimating abundance. The estimation procedure is flexible enough to allow missing counts, varying trap check intervals, variable efficiency trial schedules, and variable numbers of efficiency trials across sites.

If absolutely necessary, estimates of fish passage can be made during times of high flow by extrapolating the GAM models if the appropriate covariates are collected and if it can be assumed that the basic form of the model holds during high flows. If this assumption does not hold, estimates of fish passage during high flows cannot be made. As the GAM models are refined over time with more and more data, predictions should become more and more accurate and precise. For instance, it is not too much to hope that one day a RST will continue fishing during high flows. By using this information, however scant, to help estimate coefficients of the GAMs, researchers may one day be comfortable with abundance estimates during high flows.

Literature Cited

Hastie, T.J. and Tibshirani, R.J (1990) *Generalized Additive Models,* London: Chapman and Hall.

Lahiri, S. N. (2003) *Resampling methods for dependent data*. New York: Springer.

Olsen, A.R. and Smith, E.P. (1999) "Introduction to the Special Issue on Surveys Over Time", *Journal of Agricultural, Biological, and Environmental Statistics,* **4**(4), 328-330.

Skalski, J. (2000) "Technical Review of Demko et al. (2000): Outmigrant Trapping of Juvenile Salmonids in the Lower Stanislaus River, Caswell State Park Site, 1998" Report prepared for Scott Spaulding, USFWS Sacramento-San Joaquin Estuary Fishery Resource Office, Stockton, CA.