**Data sources**

Data for each CDFW survey are currently made available through a publicly accessible FTP site. Processed data files that report catch-per-tow or catch-per-station are available for several of the surveys, but these files exclude important information including lengths. Comprehensive survey data are stored as Microsoft Access Databases which typically include separate files for tow, catch and length data in addition to a variety of necessary lookup tables. Previous survey review efforts have identified several drawbacks to this data management approach including the propriety nature of the file type (i.e. users must pay for Microsoft Office in order to open Access files) and the substantial amount of preprocessing that is necessary to prepare the data for typical analysis and modeling applications (LTMR Chapter 3). We echo the recommendations of previous reviewers in suggesting that data be stored in a non-proprietary flat format. For each survey we downloaded the Access Database file and any associated metadata.

**Single survey data preparation**

Each database was converted into a comprehensive, long-format flat file by first exporting individual database tables in .csv format and then using each survey’s respective relationship diagram to join the tables using the ‘tidyverse’ family of packages (Wickham et al. 2019) in the programming language R (R Core Team 2020). At this initial stage all variables were retained. Many variables required renaming to create consistency between surveys and facilitate later data integration. Each survey has a unique protocol for which taxa, and how many of each are measured, but in all cases when large catches of a taxon occur, only a subsample is measured. A such, it was necessary to adjust length-frequencies to account for unmeasured individuals and the typical CDFW protocol was used, calculating Fa,l, the adjusted frequency of each recorded length as: 𝐹𝑎,𝑙 = 𝑇𝑐(𝐹𝑚,𝑙/𝑇𝑚) where Tc is the total catch of a taxon, Fm,l  is the measured frequency of each recorded length, and Tm is the total number of fish measured. In early years of the FMWT and STN surveys it was also common to only measure a subsample of target taxa. In cases where no individuals of a species were measured, an adjusted length-frequency could not be calculated, and so an NA was entered for the length of each counted individual. Such unmeasured individuals were retained at this initial stage. Adjusted length frequencies were rounded to the nearest integer, and the data tables were then expanded so that a unique row exists for each individual captured.

**Survey integration**

**External survey overview**

Although these five CDFW surveys are the focus of the current review, they exist within a broader system of fish surveys that occur in the Bay-Delta; several of which overlap substantially in space, time and methodology with individual CDFW surveys. In order to gain a more complete understanding of the frequency and intensity of fish sampling throughout a broader set of habitats, we obtained data from four additional long-term fish monitoring studies: USFWS’s Delta Juvenile Fish Monitoring Program (DJFMP) and the associated Enhanced Delta Smelt Monitoring (EDSM), CDFW’s San Francisco Bay Study (Bay Study) and UC Davis’ Suisun Marsh Study (Suisun Study). In addition, we obtained data from a recent study conducted by ICF that has deployed trawl gears in habitats not presently captured by long-term monitoring efforts. The data from these external surveys present many of the same challenges for integration as described for the core surveys (e.g. XXX). These data present additional challenges because they are collected and maintained by several different agencies and use a variety of methods in addition to pelagic trawling (i.e. benthic trawling, beach seining).

Given these data integration challenges and the

**Species selection**

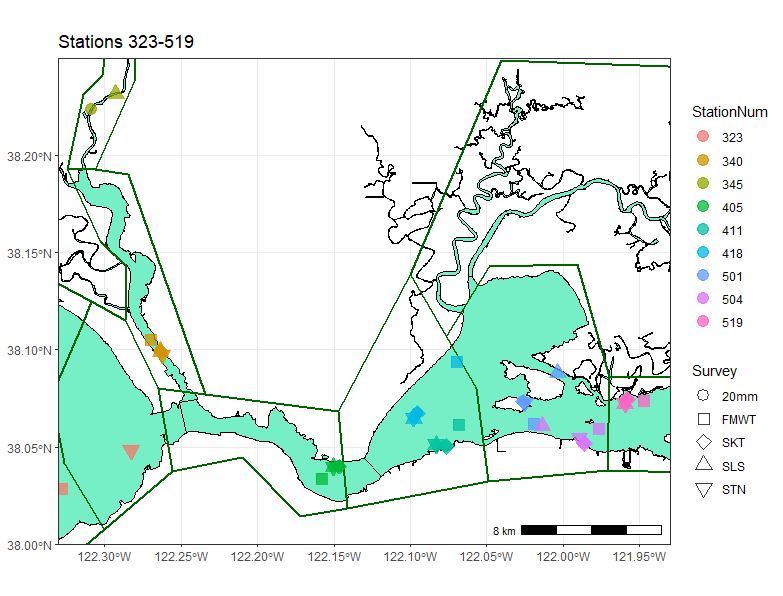
Across all CDFW surveys, more than 150 species or higher taxonomic groupings are included in the data.

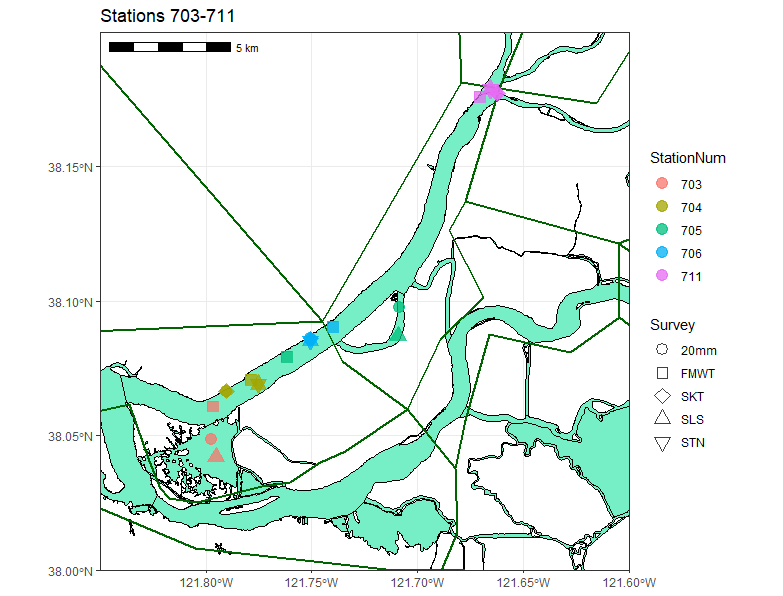
**Age distinction**

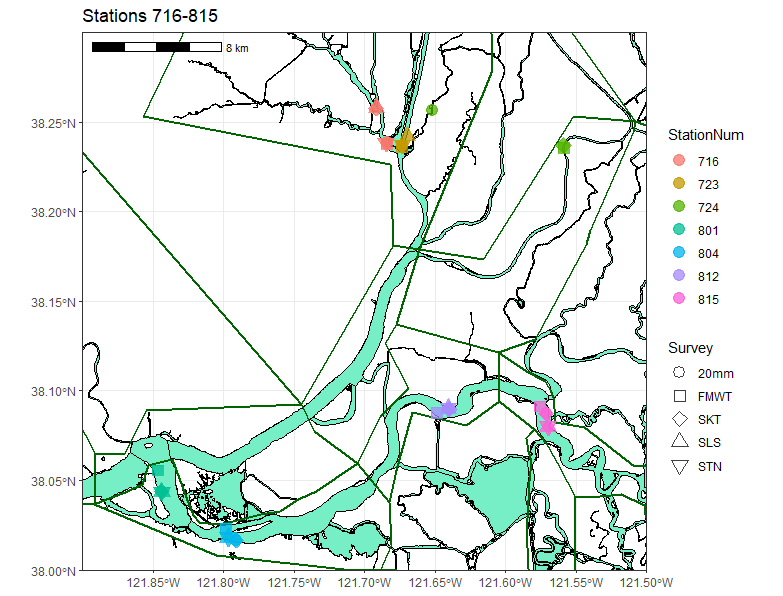
**Reconciling geographic locations of survey stations:**

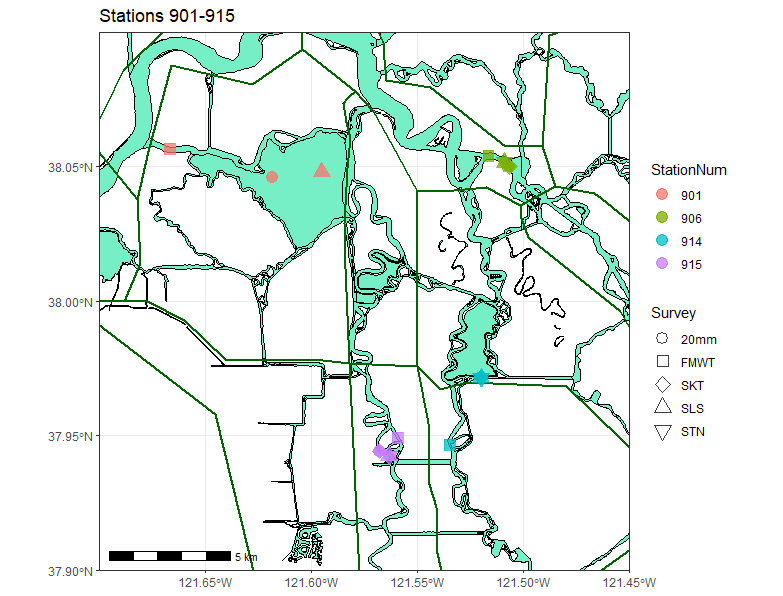
The five pelagic surveys under review were not designed in a simultaneous or coordinated effort, but rather were initiated between the 1950s and 200s to address questions specific to the ecological and management context of their respective eras. As such, in effect the surveys are independent entities, but many efforts to coordinate between surveys have occurred through the decades resulting in a challenging situation in which some aspects of the surveys are logically related, but others are not. One area that poses is challenge is the naming convention for survey stations. Each survey uses a three-digit identifier for each station, and 60 station identifiers are shared by at least two surveys (2 surveys: 21 stations, 3 surveys: 5 stations, 4 surveys: 5 stations, 5 surveys: 29 stations). In most cases, when the same identifier is used by multiple surveys the geographic location is shared, but there are numerous exceptions to this pattern. Integrated analysis of catch data requires reconciling which station identifiers differ in location between surveys. However, this process is inherently somewhat subjective, since the dynamic conditions of the Delta and practical challenge of towing nets through this environment mean that the recorded locations of sampling stations must be understood as approximate. Moreover, the practical relevance of a given geographic discrepancy will vary across regions, with minor variation in location unlikely to be of much importance in open water regions, but potentially more impactful in channel and slough areas. As such, there is some scope for minor variation in the station coordinates between surveys, but such variation must be considered on a case-by-case basis.

In total, 190 station identifiers have been used across the five surveys. We first selected only stations for which the identifier was used in at least two surveys (n=60). We next calculated the minimum and maximum decimal longitude and latitude recorded for each of these station locations, calculated the absolute value of the differences between each pair of values, and summed the results to derive a rough metric of total geographic variability for a station identifier. We next removed from further examination all stations for which the total discrepancy was less than 0.01 degrees, determining through visual inspection that differences smaller than this were within the likely range of variation typical for these sampling methods. After these exclusions, 25 stations remained with potentially consequential distances between survey locations. To allow for visual comparison at an appropriate resolution, we divided these stations into four groups and plotted the station coordinates on a map of the Delta.









**Calculations**

𝑉 = (𝐹𝑒 − 𝐹𝑠)𝐶𝑀

**Data filtering**

**Final review data formats**

**Defining spatial strata**

General approach:

Summarize per-survey table to station or subregion level via mean and in some cases estimates of within and between-year variability

For catch, remove 0 columns, standardize to 0-mean unit SD, calculate Euclidean distance matrix, apply agglomerative hierarchical clustering.

Potential subsets and/or combinations

Catch

Age-0 only vs Age-0 and Age-1

Initial conclusions from examining maps:

* 340, 519, 405, 711, 704, 706, 804, 812, 723, 815, 906 can **probably be grouped** as one location.
* 716, 801, 345, 411, 504 have somewhat greater geographic spread but could **potentially** be combined.

**Remaining stations will need to be split into at least two locations:**

* 724(20mm in separate channel),
* 901 (Potentially all unique locations),
* 915 (FMWT in different channel?),
* 914(FMWT very far off),
* 705(FMWT in main channel, others inside channel),
* 703 (FMWT not in Sherman Lake, others are), 323 (STN and FMWT 5+km apart),
* 418 (FMWT several km separate),
* 501 (surveys spread on different sides of islands)