1. Data integration

Data integration was performed in the statistical programming language R (R Core Team 2024) using the ‘tidyverse’ family of packages (Wickham et al. 2019), and packaged into the R package LTMRdata v2.0.0 (Bashevkin et al. 2022). The latest version of the R package is available at https://github.com/sbashevkin/LTMRdata. Source datasets were obtained either in flat format from data publications or directly from the Access databases. When pulled from Access databases, the individual relational data tables were joined together following the relationships diagram in their respective Access databases.

To ensure consistency across datasets, columns and categorical data (e.g., species names) were renamed and units converted. All time data were converted to Pacific Time (PST/PDT). In the Salvage dataset, daily sampling occurs throughout the 24-hour period and due to the manual recording of sampling times, there are instances where time was recorded as 2 AM on the Spring daylight savings time day, a nonexistent time since 2 AM is skipped on these days. These entries were recoded to 3 AM, but that time may be incorrect and off by one hour.

Data from the fish sampling gear(s) or sampling times (i.e., normal counts or secondary flushes in the Salvage database) of each survey were retained. Data that did not include a fish count and only a water quality sample were removed. In the Salvage dataset, these entries were identified using the unstandardized comments along with a zero-catch count.

To allow users to standardize catch across samples, we calculated metrics of effort for each sampling program and method. Sampling effort was quantified following the methods of the component surveys, as either tow area (for the otter trawls) or tow volume (for all other gear types). The tow volume values for Salvage represent the daily export volume for the sampled facility. Thus, although multiple salvage samples are collected each day, the tow volume value represents the daily effort. Some tow volumes values are 0 for Salvage and these represent days when no pumping was conducted. For the Enhanced Delta Smelt Monitoring Program (EDSM) and Delta Juvenile Fish Monitoring Program (DJFMP) datasets, tow volume was set to NA when debris was detected in the flowmeter, and samples were excluded with gear condition codes 3 (poor sampling), 4 (no sample attempted), or 9 (fish captured outside of live box or cod end and could not be assigned to a specific tow).

Catch count in the Salvage dataset was calculated as salvage:

*S* = *N*­(*Tp/Tc*),

where *S* = Salvage, *N* = the observed count of each species, *Tp* = the total time pumped (minutes), and *Tc* = the total time sampled (minutes; CDFW 2018).

Generally, surveys measure lengths for only a subset of the total catch of a given species from each sample in a process called subsampling. To estimate the frequency of each recorded length across all individuals caught in a sample, we calculated the adjusted length frequencies for each species in each sample following the methods of FMWT (James White, personal communication):

A𝑙= *C*(𝐹𝑙*/M*),

where A𝑙 = adjusted frequency of each recorded length, C = total catch, 𝐹𝑙 = measured frequency of each recorded length, and *M*= total number of fish measured. For instances when subsampling did not occur (e.g., species listed on the Endangered Species Act or the California Endangered Species Act, or a low catch that did not require subsampling), the calculated adjusted length frequency is equivalent to the recorded length frequency. Some surveys occasionally subsampled for measurement separately on different groups within a species (generally based on size classes, age, or Salmon races). In these cases, we calculated the adjusted frequencies separately for each group. In the Suisun Study database, these groups were identified with unstandardized comments describing the size or age class. When possible, these comments were converted into a fish size range that could then be assigned a group. However, not all comments could be translated into a fish size range, leaving some fish sizes unknown in the Suisun Study data. Fish sizes are occasionally unknown in many of the surveys when no individuals of a species were measured in a sample.

The datasets were then combined by simply binding them together by their now-standardized column names. The last step in data integration was to fill in zeros. The surveys did not commonly record when a species was not present in a trawl, but these absences are critically important for statistical analysis. For any samples in which a species was not recorded, we added a record of that species with a count of 0. Note that this resulted in 0s being added even for years in which a particular taxon was not sampled for (e.g., some surveys only started counting jellyfish in more recent years, but this will fill in 0s even before they were counted). Thus, users should consult the documentation of the component datasets for taxa that have not always been counted. However, zeros were filled separately for each survey, so zeros were only filled for taxa that had been recorded historically by that particular survey. This should ensure that zeros are only filled based on the taxa list (and historical observations) of each individual survey.

2. Fish lengths

While 9 of the surveys recorded fish lengths as fork length (or total length if no fork), the Suisun Study recorded standard length. We thus compiled conversion equations for 20 species to convert Suisun Study lengths (for those 20 species) to fork length (or total length if no fork). The equations are of the form:

FL = a + b × SL

where *FL* = fork length (or total length for species with no fork) in mm, *a* = intercept, *b* = slope, and *SL* = standard length in mm. The Splittail equation was calculated via a median quantile regression, while the other equations were obtained from the California Department of Fish and Wildlife Length-Weight Study (unpublished).

While the equations are provided, lengths are not converted in this dataset, so Suisun Study lengths are not comparable to the other surveys unless converted. Sampling programs also often discard (not count nor measure) fish below a size cutoff, but those cutoffs differ among surveys and over time. The integrated dataset does not account for these different cutoffs, so users should consult the documentation of their dataset of interest for these details.

3. Final format

Before publication, the joined dataset was split into separate tables for the fish-level data (taxa, length, count, etc.) and the sample-level data (environmental variables, sampling effort, etc.). The fish length conversion equations are published in a separate table.

The dataset can be downloaded and accessed via the R package deltafish (https://github.com/Delta-Stewardship-Council/deltafish). The package downloads the data from this EDI publication, processes it into a database, and provides tools to access that database within R.

4. Literature cited

Bashevkin, S. M., J. W. Gaeta, T. X. Nguyen, L. Mitchell, and S. Khana. 2022. LTMRdata: An R package to integrate data from 9 fish surveys in the San Francisco Estuary. v2.2.0. Zenodo. doi:10.5281/zenodo.6048977

[CDFW] California Department of Fish and Wildlife (2018). Chinook salmon loss estimation for Skinner Delta Fish Protective Facility and Tracy Fish Collection Facility. California Department of Fish and Wildlife.

R Core Team. 2024. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Wickham, H., M. Averick, J. Bryan, and others. 2019. Welcome to the Tidyverse. Journal of Open Source Software 4: 1686. doi:10.21105/joss.01686