1. Data integration

Data integration was performed in the statistical programming language R (R Core Team 2021) 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 either pulled in flat format from data publications, or ingested from the raw Access databases. When pulled from Access databases, the individual data tables were exported to csv format to facilitate import into R. The data tables from each survey were then joined together following the relationships diagram in their respective Access databases. Data from the primary sampling gear(s) of each survey were retained. Columns and categorical data (e.g., species names) were then renamed and units converted for consistency across surveys.

To standardize catch across trawls, we calculated metrics of effort for each trawl. Sampling effort was quantified following the methods of the component surveys, as either a tow area (for the otter trawls) or a tow volume (for all other gear types). For EDSM and DJFMP data, 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).

Generally, a subset of the total catch of a given species is measured from each trawl. To estimate the frequency of each size class of fish, we calculated the adjusted length frequencies for each species in each trawl after the methods of FMWT (James White, personal communication):

where Fa,l = adjusted frequency of each recorded length, Tc = total catch, Fm,l = measured frequency of each recorded length, and Tm = total number of fish measured. 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 trawls in which a species was not recorded, we added a record of that species with an adjusted frequency of 0. Note that this will result in 0s being added even for years in which a particular taxa was not counted in samples (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 will need to peruse documentation of the component datasets for taxa that have not always been counted.

2. Fish lengths

While 8 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 (**Table 3.1**) to convert Suisun Study lengths (for those 20 species) to fork length (or total length if no fork). The equations are of the form:

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-Wight Study (unpublished).

Lengths are especially important for the data integration because each survey would discard (not count nor measure) fish below a size cutoff, but those cutoffs have differed among surveys and over time. 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.

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). This 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

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