Delta Smelt Life Cycle Model Data Files

Lara Mitchell and Ken Newman

June 6, 2014

DRAFT

Many relevant data sets have been assembled in support of the Delta Smelt Life Cycle Model (DSLCM) in recent years. These data sets (referred to here as "raw" data sets) have been used to create "clean" data sets designed specifically for model fitting purposes. The clean data sets are produced using a series of R scripts that take the raw data files as input, carry out various data cleaning procedures, and save the resulting clean data sets as csv files and R objects. This document provides descriptions of the raw data sets, clean data sets, and R scripts, organized by data type.

Version History

May 28, 2014 – Version history section added. Tables 13 and 14 also added.
May 29, 2014 – Table 15 added.

1 Fish Survey Data

Four separate data sets, categorized as station, catch, length, and tide data, have been compiled for each of five fish surveys conducted by the California Department of Fish and Wildlife: Spring Kodiak Trawl (SKT), Fall Midwater Trawl (FMWT), Bay Study Midwater Trawl (Bay), 20mm Survey (Twentymm), and Summer Townet (STN). Each data set is saved as an Excel file with one worksheet containing a flat file with one or more lines of heading. Some Excel files have additional worksheets containing metadata or data summary calculations. A copy of each flat file has been saved in csv format (with the same name as the parent Excel file) so that the data can easily be read into R. All csv files are formatted to have only one header line, regardless of how many were used in the corresponding Excel file. Data from the Chipps Midwater Trawl survey (conducted by the U.S. Fish and Wildlife Service) have been divided into only two categories, catch data and tide data, and saved in Excel and csv files similar to those of the other surveys.

1.1 Raw Data

1.1.1 Station

Each station data set provides the latitude, longitude, region, subregion, and three digit station number of the sampling sites used in the given survey. All five station data sets (corresponding to STN, FMWT, SKT, Bay, and Twentymm) have the format shown in Table 1, with latitude (LatDD) and longitude (LonDD) represented in decimal degrees.

1.1.2 Catch

Each catch data set describes the fish species composition of the survey on a per-tow basis. In the case of STN, FMWT, SKT, Bay, and Twentymm, each record in the data set corresponds to a single tow and

Table 1: An example of a station data set.

Station	LatDD	LonDD	Region	SubRegion
323	38.05	-122.28	Far West	East San Pablo Bay
328	38.06	-122.35	Far West	Mid San Pablo Bay
329	38.06	-122.30	Far West	East San Pablo Bay
334	38.08	-122.34	Far West	Mid San Pablo Bay
335	38.07	-122.32	Far West	East San Pablo Bay
336	38.06	-122.28	Far West	East San Pablo Bay
⋮				

contains information on when and where the tow took place, what species were caught, how many individuals of each species were caught, and what the physical conditions were like at the time of sampling. Table 2 describes a set of data fields common to many of the catch files. This table reflects some of the fields of primary interest for the purposes of this modeling effort, but we note that many of the catch data sets include additional fields not described here. Further details on how the data are collected or calculated are available elsewhere (`www.dfg.ca.gov/delta/`).

The Chipps Excel file contains two worksheets, labeled "Chipps Island Trawls" and "Chipps Island Larval DSM remove". The first worksheet consists of a data set containing count and length data for multiple fish species. Each record describes the number of individuals of a given species and size caught in a given tow on a given date. Descriptions of the fields of interest are given in Table 3, with further details available elsewhere (`www.fws.gov/stockton/jfmp`). If no organisms were caught at a given date-time, the record appears in the data set with a blank value in the Organism field. It should be noted that Chipps is carried out at one location and hence does not sample from a range of stations like the other surveys. Between 1976 and roughly 1996, larval delta smelt (defined as delta smelt less than 25 mm in forklength) were counted and recorded as part of the Chipps survey. The second worksheet consists of the same data as in the first worksheet except with pre-1996 records identified as "larval delta smelt" removed. Some uncertainty remains about whether any records in this data set, in particular those without length information, still include larval delta smelt. The Chipps csv catch file is a copy of the second worksheet.

1.1.3 Length

Each length data set contains delta smelt age and length information. For FMWT, Bay, Twentymm, and STN the data set has a single record for each unique combination of date-station that was sampled, and contains, at a minimum, the nine fields described in Table 4. Some data sets also contain physical variables and measures of CPUE. The last six fields in the table represent length statistics for age-0 and age-1 delta smelt, many of which are missing because delta smelt were not caught, length measurements were not collected, or only one delta smelt was measured (in which case standard deviation cannot be calculated). The SKT data set has a separate record for every delta smelt that includes the individual's date-of-catch, forklength, sex, and reproductive stage.

The date columns across all raw catch and length files have been given identical m/d/yyyy formatting (e.g., 1/30/1999 represents January 30, 1999). However, to avoid potential confusion, separate Year, Month, and Day columns have been added to most of the data sets.

Table 2: A partial summary of the data fields in the SKT, FMWT, Bay, Twentymm, and STN raw catch files. ✓ indicates that the field is present, X that it is absent.

		Survey				
Field Name	Description	SKT	FMWT	Bay	Twenty-mm	STN
Date	Date of tow.	✓	✓	✓	✓	✓
Year	Year of tow.	✓	✓	✓	✓	✓
Month	Month of tow.	✓	✓	✓	✓	✓
Day	Day of the month of tow.	✓	✓	✓	✓	✓
TimeStart	Time at start of tow.	✓	✓	✓	✓	✓
Survey	A number describing the progression of the survey on a biweekly or monthly basis.	✓	✓	✓	✓	✓
Station	Station number.	✓	✓	✓	✓	✓
Tow	The unique tow number at a given station, on a given date.	X	X	✓	✓	✓
TowDirection	Tow direction code: 1 = with current, 2 = against current, 3 = unknown (during slack).	X	✓	✓	X	X
Secchi	Secchi depth (cm).	✓	✓	✓	✓	✓
CondSurf	Specific conductivity of the first foot of water from the surface (µS).	✓	✓	X	✓	✓
CondBott	Specific conductivity of the first foot of water from the bottom (µS).	X	✓	X	✓	✓
TempSurf	Water temperature (°C).	✓	✓	✓	✓	✓
Tide	Tide code: 1 = high slack, 2 = ebb, 3 = low slack, 4 = flood.	✓	✓	✓	✓	✓
Depth	Depth of water (Bay: m, all others: ft).	✓	✓	✓	✓	✓
Volume	Estimate of water volume sampled (m^3).	✓	✓	✓	✓	✓
SalinSurf	Salinity (ppt) for first meter of water column.	X	X	✓	X	X
delta.smelt	Number of delta smelt in the tow.	✓	✓	✓	✓	✓
delta.smelt.age0	Number of age-0 delta smelt in the tow.	✓	✓	✓	✓	✓
delta.smelt.age1	Number of age-1 delta smelt in the tow.	✓	✓	✓	✓	✓

1.1.4 Tide

The tide data sets (prepared by CM2H-Hill) contain data related to the tidal stages during fish survey sampling. The primary data fields are summarized in Table 5. There exists one record per date-station in the case of SKT, FMWT, Bay, Twentymm, and STN, and one record per date-time in the case of Chipps.

1.2 Clean Data

The R script *Data Cleaner - Fish Surveys.r* allows the user to specify a survey (SKT, FMWT, Bay, Twentymm, or STN) and merge that survey's raw data sets to produce a standardized data set containing delta

Table 3: A partial summary of the data fields in the Chipps raw data set.

Field Name	Description
SampleDate	Date of tow.
TimeStart	Time at start of tow.
TowNumber	The unique tow number on a given date.
TowDirection	Tow direction code: U = upstream, D = downstream.
Secchi	Secchi depth (m).
WaterTemp.	Water temperature (°C).
Volume	Estimate of volume sampled (m^3).
Organism	Organism code (DSM = delta smelt).
ForkLength	Fork length (mm).
Count	Number of fish in the given tow that have the given organism code and forklength.

Table 4: A partial summary of the data fields in the FMWT, Bay, Twentymm, and STN raw length data sets.

Field Name	Description
Date	Date of data collection.
Station	Station number.
Age0_n_L	Number of age-0 delta smelt measured for length.
Age0_L_bar	Mean length of the measured age-0 delta smelt.
Age0_s_L	Standard deviation of the measured age-0 lengths.
Age1_n_L	Number of age-1 delta smelt measured for length.
Age1_L_bar	Mean length of the measured age-1 delta smelt.
Age1_s_L	Standard deviation of the measured age-1 lengths.

smelt age and length statistics, select physical variables, and tide information. The clean data sets are "updated" versions of the raw catch data sets, designed to have one record per unique date-station and 40 standardized data fields, shown in Table 6. These refer to the same fields, by name, as those described in Tables 1, 2, 4, and 5, with unit conversions carried out as necessary so that all clean data sets have the set of units shown. One exception is the Month field, which reflects the month to which the record has been designated for modeling purposes, which may differ from the month in which the data were actually collected.

Part of the cleaning code is generic, meaning that the same changes are made to all of the surveys. For example, region, subregion, latitude, and longitude data are transferred from each station data set to the corresponding catch data set. Other tasks, such as removing special (non-routine) tows, are specific to individual surveys. Because of the structural differences between the Chipps catch file and the other catch files, the Chipps data are processed using a separate script, *Data Cleaner - Chipps.r*. The Chipps clean data set is similar to those of the other surveys except that each record represents a unique date-time as opposed to a unique date-station. Both scripts reference a series of functions saved in a third script, *Utility Functions.r*. A summary of the changes made by the R scripts is provided here. Distinctions between procedures used by the two scripts are made as necessary.

Field insertion:

For SKT, a Tow column is created that is identical to TimeStart; this is used to aggregate over unique tows later in the code. For Bay, the SalinSurf field is used to calculate CondSurf using the following conversion

Table 5: A partial summary of the data fields in the SKT, FMWT, Bay, Twentymm, STN, and Chipps raw tide data sets.

Field Name	Description
Date	Date of data collection.
TimeStart	Time at start of data collection. (If there are multiple tows at a date-station, the earliest tow time is used.)
Region	Region of data collection.
Station	Station number.
TideStage	Tide level (in feet) relative to NGVD29.
HighType	Closest peak high tide: HH = High High, LH = Low High.
Time-to-High-Min	Difference between sampling time and the closest peak high tide time (min).
LowType	Closest peak low tide: HL = High Low, LL = Low Low.
Time-to-Low-Min	Difference between sampling time and the closest peak low tide time (min).
TideVelocity	Instantaneous Velocity (ft/s).
Ebb-Type	Closest peak ebb velocity: HE = High Ebb, LE = Low Ebb.
Time-to-Ebb-Min	Difference between sampling time and the closest peak ebb velocity time (min).
FloodType	Closest peak flood velocity: HF = High Flood, LF = Low Flood.
Time-to-Flood-Min	Difference between sampling time and the closest peak flood velocity time (min).
Time-to-Slack-Min	Difference between sampling time and the closest slack velocity time (min).

equation: $Conductivity = 178500\left(1 - e^{-0.01 * Salinity}\right)$. This equation may need correcting (Wim Kimmerer, personal communication).

For Chipps, the TowNumber field is simply renamed Tow, and WaterTemp is renamed TempSurf. In addition, Region is added with the value "West", Station is added with the value "Chipps", Lat is added with the value 38.055, and Lon is added with the value -121.9109.

For all six surveys, additional fields are added as necessary so that every clean data set has all of the fields shown in Table 6. These added fields generally contain all NA's.

Station removal:

Data Cleaner - Fish Surveys: Records corresponding to sampling stations outside of the four main regions (Far West, West, North, and South) are not included in the clean data sets. Nor are records with station numbers that do not have matches in the corresponding station data set.

Records corresponding to stations 347, 348, and 349 (all of which are located in the Upper Napa River subregion of the Far West region) are also removed from the Twentymm data set. These stations were sampled relatively infrequently over time but have large catches that make these records outliers. See Table 7 for a summary of the sampling frequency and catch data for these stations.

Unit conversion:

The Depth field in the Bay data set is converted from meters to feet. For Chipps, the Secchi field is

Table 6: Data field names and units used in the clean data files. Units, indicated in parentheses, are not part of the field names. The field ordering in the clean data files is shown column-wise; e.g., Year is the first field, Date is the second field. An exception is made in cases where there are multiple tows, see "Tow Aggregation".

Year	Tow	delta.smelt	HighType
Date	TowDirection	delta.smelt.age0	Time-to-High-Min (min)
Month	Secchi (cm)	delta.smelt.age1	LowType
TimeStart	CondSurf (µS)	Age0_n_L	Time-to-Low-Min (min)
Survey	TempSurf (°C)	Age0_L_bar	TideVelocity (m/s)
Station	CondBott (µS)	Age0_s_L	EbbType
Region	Tide	Age1_n_L	Time-to-Ebb-Min (min)
SubRegion	Depth (ft)	Age1_L_bar	FloodType
Lat (dd)	Volume (m^3)	Age1_s_L	Time-to-Flood-Min (min)
Lon (dd)	SalinSurf (ppt)	TideStage	Time-to-Slack-Min (min)

Table 7: Daily tow frequency and total catch for stations 347, 348, and 349 in the Twentymm data set.

	Station					
	347		348		349	
Date	# tows	total catch	# tows	total catch	# tows	total catch
1996-07-26	1	0	0	0	0	0
2001-03-24	3	3	3	40	3	16
2001-04-07	3	787	3	1742	3	746
2001-05-05	3	58	3	254	3	250
2001-06-04	3	0	3	0	3	1
2002-03-22	1	0	0	0	0	0
2002-04-06	1	0	0	0	0	0
2002-04-19	1	0	0	0	0	0

converted from meters to centimeters. For all six surveys, TideVelocity is converted from feet-per-second to meters-per-second.

Factor levels:

In cases where the Tide field has numerical factor levels (1, 2, 3, 4), these values are changed to descriptive strings ("High Slack", "Ebb", "Low Slack", "Flood") for clarity. Similarly, in cases where the TowDirection field has numerical levels (1, 2, 3), these values are changed to descriptive strings ("With", "Against", "Neither").

Special survey/tow removal:

Special (non-routine) surveys and tows in the raw data files were not included in the clean data sets. For SKT, special surveys are indicated by a Survey number greater than 5; for Twentymm, by a Survey number greater than or equal to 10. For STN, special tows are indicated by a Tow number of 4. In the case of SKT, these records are removed per advice from Julio Adib-Samii (personal communication).

Additional record removal:

Any records more recent than 2010 are removed from all six surveys.

Data Cleaner - Fish Surveys: For STN, records prior to 1962 are also removed, while for Twentymm, records from July 2001 are removed.

Data Cleaner - Chipps: Records prior to 1978 are removed. Delta smelt records with lengths less than 25mm or greater than 100mm are reclassified as "Other Smelt" and hence not used in constructing the Chipps clean data set. The decision to remove fish less than 25mm is based on a meeting with Matt Dekar, Joseph Kirsch, Jonathan Speegle, and Pat Brandes on September 16, 2013. The decision to remove fish greater than 100mm is based on the hypothesis that larger delta smelt may have been misidentified in the past (William Bennett, personal communications). See Table 8 for a summary of the records removed based on length.

Table 8: Frequency of Chipps delta smelt records removed with length > 100mm or < 25mm, by year and month.

Month	Year																	Total
	1979	1981	1982	1983	1985	1989	1990	1991	1992	1994	1995	1996	1998	1999	2002	2003	2004	
January	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
May	0	2	0	2	2	1	1	0	1	0	0	1	1	0	0	0	0	11
June	1	0	2	0	0	2	0	1	0	0	1	0	0	0	0	0	0	7
July	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	2
November	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
December	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	4
Total	1	2	2	2	2	3	1	1	1	1	3	1	2	1	1	1	1	26

(a) Length > 100mm.

Month	Year						Total
	1979	1983	1984	1993	1994	1995	
May	0	0	0	1	0	20	21
June	4	2	7	3	4	157	177
July	0	0	0	0	0	26	26
Total	4	2	7	4	4	203	224

(b) Length < 25mm.

Month designation:

Some changes, summarized below, are made to Month fields. These changes are made when the station is part of a survey set where the majority of stations in that survey were sampled in an adjacent month. For example, an SKT station in the fifth set of the year (in May typically) was sampled at the end of April, and its month designation was changed to May. Note that these changes are made to the Month field only, and not the Date field.

SKT: Records from April 2007 with a Survey value of 5 are changed to May.

FMWT: Month values are assigned according to Survey number, with Surveys 1 through 12 corresponding to July through June, respectively.

Bay: Records from 1980 with sampling dates near the end of the month (the 17^{th} or later) are changed to the subsequent month (per advice from Kathy Hieb, personal communication). In this case the Survey number is also increased by one so that the Survey number is consistent with the Month. For example, records from January 31, 1980 (Survey 1) are given a new Month value of February and a new Survey value of 2.

Twentymm: Records from March 1997 are changed to April. Records from August with a Survey number of 8 or 9 are changed to July. Records from June with a Survey number of 9 are also changed to July.

STN: Records from July 2008 with a Survey value of 3 are changed to July. Records from July 2008 with a Survey value of 5 are changed to August. Records from September 2005 are changed to August.

Volume imputation:

Data Cleaner - Fish Surveys: First, an attempt is made to impute missing volumes using median volumes calculated by date-station. If no data are available to calculate the median on this scale, an attempt is made to impute the missing volumes using median volumes calculated by year-region. In the case of FMWT, no volume data are available prior to 1985. Volumes from 1985 are therefore used to fill in the pre-1985 missing values on a per-region basis.

Data Cleaner - Chipps: First, an attempt is made to impute missing volumes using mean volumes by year-month. In cases where no data are available from the same year-month as the missing value, volumes from the two adjacent months are used to calculate the mean.

Tow aggregation:

Data Cleaner - Fish Surveys: The Twentymm and STN surveys typically take three replicate tows at a given station (on a given date). Some limited tow replication also takes place during SKT sampling. In the clean data sets, these replicate tows are aggregated to form one unique record per date-station. The value of delta.smelt for the aggregated record is found by summing the values of this field across the individual tows; similarly for delta.smelt.age0, delta.smelt.age1, and Volume. Volume imputation is carried out before tow aggregation; age and length imputation (described below) are carried out after tow aggregation.

In cases where tows are aggregated, three additional fields are added to the clean data set for each of the n unique tow numbers. For $i = 1, \ldots, n$, the fields ds.tow$i$, ds.age0.tow$i$, and ds.age1.tow$i$ indicate the total number of delta smelt, the number of age-0 delta smelt, and the number of age-1 delta smelt caught in tow i, respectively. These fields are inserted between delta.smelt.age1 and Age0_n_L.

Length statistics:

Data Cleaner - Fish Surveys: As described previously, the SKT raw length file has a separate record for each delta smelt. As part of the cleaning process, an age assignment key is used to assign an age (0 or 1) to each of these fish based on length and month-of-catch. The key, summarized below, was developed by CDFW (Steve Slater, personal communication). The length statistics Age0_n_L, Age0_L_bar, Age0_s_L, Age1_n_L, Age1_L_bar, and Age1_s_L are then calculated on a date-station basis, and merged with the clean SKT data set.

Data Cleaner - Chipps: As described previously, the Chipps raw catch file has a separate record for each date-time-species-length combination. As part of the data cleaning process, length values of zero are first changed to NA's. Ages are then assigned to each delta smelt record using the CDFW age-assignment key, and length statistics (Age0_n_L, Age0_L_bar, etc.) are calculated on a date-time basis and merged with the clean catch data set.

Table 9: Delta smelt age assignment key. The numbers indicate the cut-off length (in millimeters) used to distinguish between age-0 and age-1 fish in the given month. Individuals below the cut-off length are taken to be age-0; individuals at or above this length are taken to be age-1.

Jan.	Feb.	Mar.	Apr.	May	June	July	Aug	Sep.	Oct.	Nov.	Dec.
40	50	50	50	50	60	65	70	75	80	80	80

Age fields:

Data Cleaner - Chipps: Values of delta.smelt.age0 and delta.smelt.age1 in the clean Chipps data set are filled in as follows: In cases where the value of delta.smelt is zero, delta.smelt.age0, delta.smelt.age1, Age0_n_L, and Age1_n_L are set to zero. In cases where Age1_n_L is equal to delta.smelt, delta.smelt.age1 is set equal to delta.smelt, while delta.smelt.age0 and Age0_n_L are set to zero. In cases where Age0_n_L is equal to delta.smelt, delta.smelt.age0 is set equal to delta.smelt, while delta.smelt.age1 and Age1_n_L are set to zero. If Age0_n_L + Age1_n_L is equal to delta.smelt, delta.smelt.age0 and delta.smelt.age1 are set equal to Age0_n_L and Age1_n_L, respectively.

Length imputation:

Data Cleaner - Fish Surveys: First, an attempt is made to impute missing values of Age0_L_bar (the mean length of age-0 fish), using sample size-weighted averages of the available age-0 mean lengths by year-month. Next, an attempt is made impute those values that are still missing using sample size-weighted monthly (across year) averages of the available age-0 mean lengths. Missing values of Age1_L_bar are imputed analogously. No attempt is made to impute sample sizes (Age0_n_L, Age1_n_L) or standard deviations (Age0_s_L, Age1_s_L).

Age imputation:

Data Cleaner - Fish Surveys: To impute the number of age-0 delta smelt (delta.smelt.age0) and age-1 delta smelt (delta.smelt.age1) represented in a given record, we first calculate the fraction of age-0 delta smelt caught on a year-month basis using the available values of delta.smelt.age0 and delta.smelt.age1. For each record with missing age counts, the imputed value of delta.smelt.age0 is given by the product of the total number of delta smelt (delta.smelt) and the estimated fraction of age-0's for that year-month. The imputed value of delta.smelt.age1 is calculated accordingly as the product of delta.smelt and one minus the age-0 fraction. For FMWT, month-specific fraction estimates (provided by Dave Contreras, personal communication) are used whenever calculated values are not available (including years prior to 1975). These fractions are as follows: January: 0, February: 0, March: 0, May: 0, September: 0.9, October: 1, November: 1, December: 1.

Data Cleaner - Chipps: The age imputation process for Chipps is the same as for the other five surveys except that the median fraction of age-0 delta smelt for a given month (across all years) is used whenever a calculated year-month estimate is not available.

Unavailable length statistics:

In cases where the value of delta.smelt.age0 is zero, the values of Age0_n_L, Age0_L_bar, and Age0_s_L are set to 0, NA, and NA, respectively. Similarly for cases where the value of delta.smelt.age1 is zero.

Clean data files:

A copy of each clean data set is saved as (1) a data frame in an R object file, and (2) a csv file, both of which have the same abbreviated name as the fish survey. All fields other than those shown in Table 6, and those that may be added during tow aggregation, are removed before saving.

Table 10 gives a brief summary of each clean data set, including the number of records for which volume, age, and/or length imputation was carried out.

Table 10: Summary of the clean catch data sets. The last three columns indicate the number of records for which missing values of Volume, delta.smelt.age0 (delta.smelt.age1), and Age0_L_bar (Age1_L_bar) were imputed. Note that for SKT – STN each record corresponds to a unique date-station, whereas Chipps has a separate record for each date-time.

Survey	Year range	Number of records	Number of fish	Multiple tows combined	Imputation		
					Volume	Age	Length
SKT	2002 – 2010	1611	5010	yes	0	0 (0)	0 (0)
FMWT	1967 – 2010	21370	20518	no	8683	1189 (1189)	828 (635)
Bay	1980 – 2010	7886	2857	no	0	0 (0)	0 (0)
Twentymm	1995 – 2010	5525	22162	yes	21	0 (0)	0 (0)
STN	1962 – 2010	5959	48915	yes	0	0 (0)	0 (6)
Chipps	1978 – 2010	41912	93675	–	70	936 (936)	0 (0)

2 Prey Data

Delta smelt prey are currently divided into four categories: zooplankton, mysids, amphipods, and fish larvae.

2.1 Zooplankton and Mysids

2.1.1 Raw Data

Zooplankton and mysid data are available through the Environmental Monitoring Program's Zooplankton Study, which started in 1972 and uses three gear types: (1) a pump targeting microzooplankton less than 1 mm in length, (2) a modified Clarke-Bumpus (CB) net targeting mesozooplankton 0.5 – 3.0 mm in length, and (3) a macrozooplankton net targeting zooplankton 1 – 20 mm in length, including mysid shrimp. Further information is available at `http://www.water.ca.gov/bdma/meta/zooplankton.cfm` and `http://www.dfg.ca.gov/delta/projects.asp?ProjectID=ZOOPLANKTON`.

Two separate raw EMP data sets have been compiled, one containing data on zooplankton species including copepods, clacoderans, and rotifers (referred to as the zooplankton data set), and one containing data on mysids (referred to as the mysid data set). The zooplankton data set presents catch-per-unit-volume (number caught/m^3), CPUV, for a variety of taxa, with each record corresponding to a unique combination of sample date and sampling station. The mysid data set contains biomass-per-unit-volume (micrograms of Carbon/m^3), BPUV, with each record again corresponding to a unique combination of sample date and sampling station. The sampling stations used in the EMP surveys are different from those used in the fish surveys described in the previous section; see Figure 1 for a map of the EMP stations. A separate EMP station file has been created that indicates to which of the four DSLCM regions (Far West, West, North, or South) each EMP station belongs.

Table 11 shows select fields from the raw zooplankton and mysid data sets that are used to calculate biomass metrics in the clean zooplankton and mysid data sets (see next section). Note that zooplankton are restricted to calanoid copepods, cyclopoid copepods, cladocerans. Different taxa have been collected at different times throughout the history of the survey, as indicated by the "Sampling Period" column in Table 11. Differences in collection periods are due, in part, to the fact that many of the species are non-indigenous to the bay-delta.

The Weight value associated with each Field in Table 11 indicates the estimated Carbon biomass of an individual in that Field, and serves as an indicator of how "nutritious" the individual is: the higher the weight, the more nutritious (Wim Kimmerer, personal communication). Mysid weights are highly dependent upon individual size (Wim Kimmerer, personal communication). The following formulas give the Carbon

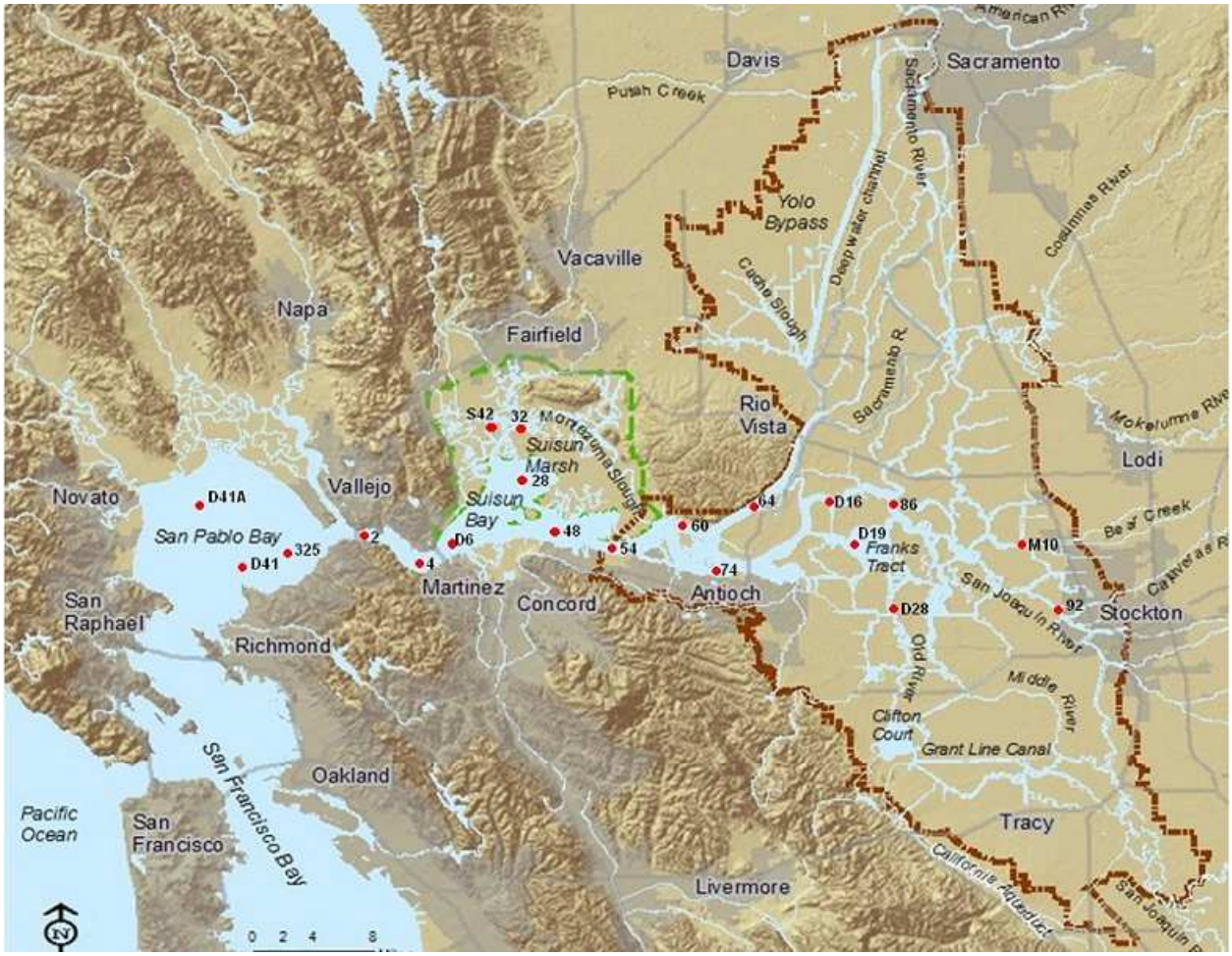

Figure 1: EMP Zooplankton Study sampling locations, shown as red dots (`http://www.dfg.ca.gov/`).

biomass (μgC) for individual mysids based on species and length (mm), and were used to calculate the BPUV values in the raw mysid data set (April Hennessy, personal communication):

Species	Carbon biomass equation
Hyperacanthomysis longirostris	$0.4 * \left(0.0103 * \text{Length}^{2.2593}\right)$
Neomysis mercedis	$0.4 * \left(0.0012 * \text{Length}^{3.2533}\right)$
All other species	$0.4 * \left(0.0012 * \text{Length}^{3.2533}\right)$

It has been hypothesized that organisms sampled by the pump component of the EMP Zooplankton study may be too small for juvenile and adult delta smelt to actively target as prey (Matt Nobriga, personal communication). For this reason, the pump data are not being used at this time. Zooplankton data collected as part of the Twentymm fish survey are also not being used because they are temporally limited relative to the EMP study, and because there is tentative evidence for correlation with the EMP data (Steve Slater, personal communication).

2.1.2 Clean Data

The R script *Data Cleaner - Zooplankton.r* uses the raw zooplankton data set to create a new data set containing measures of zooplankton biomass calculated by year-month-region for the years 1972 – 2010. We first remove any records from the raw zooplankton data set that fall outside of the four main regions, and replace any 0 CPUV values outside of each field's sampling period with NA's.

For each combination of year, month, and region, we calculate three measures of prey biomass using the fields shown in Table 11: β_1, which consists of the copepod nauplii and juveniles, β_2, which consists of the copepod juveniles and adults, and β_3, which consists of the copepod juveniles, copepod adults, and cladocerans. Let ymr represent a unique year-month-region, S_{ymr} represent the set of n_{ymr} records corresponding to ymr (where each record represents a unique date-station), and K_i represent the set of data fields corresponding to β_i. Then for $i = 1$, 2, or 3,

$$\beta_{i,ymr} = \frac{1}{n_{ymr}} \sum_{s \in S_{ymr}} \ln\left(\sum_{k \in K_i} W_k * \text{CPUV}_{ymr,s,k} + \epsilon\right), \tag{1}$$

where W_k is the biomass "weight" for field k (from Table 11) and ϵ is a small (< 1) adjustment factor to ensure that the natural log is always defined. Note that the quantity $W_k * \text{CPUV}_{ymr,s,k}$ is an estimate of BPUV for field k from record ymr, s. Each value of $\beta_{i,ymr}$ is therefore a log geometric mean of the "total" biomass per unit volume, per year-month-region.

The clean zooplankton data set contains values of $\beta_{i,ymr}$, and is structured as shown in Table 12, where the columns naup_juvBPUV, juv_adultBPUV, and juv_adult_cladBPUV, correspond to β_1, β_2, and β_3, respectively. Missing values of naup_juvBPUV are imputed by linearly interpolating across the year-month time series in a given region. Missing values of juv_adultBPUV and juv_adult_cladBPUV are calculated similarly. See Table 13 for a summary of when and where values of naup_juvBPUV, juv_adultBPUV, and juv_adult_cladBPUV are currently imputed.

The mysid data cleaning process, carried out by the script *Data Cleaner - Mysid.r*, is similar to that of the zooplankton data except for two primary differences: (1) the values in the raw mysid data set are measures of BPUV that have already been weighted according to mysid length, and (2) there exists only one prey metric, mysidBPUV, consisting of both mysid species listed in Table 11. Missing values are again imputed by linearly interpolating across the year-month time series in a given region. See Table 12 for an example of a clean mysid data set, and Table 14 for a summary of the imputed data.

2.2 Amphipod

2.3 Fish larvae

References

[1] Orsi, J.J., Bowman, T.E., Marelli, D.C., Hutchinson, A. 1983. Recent introduction of the planktonic calanoid copepod *Sinocalanus doerrii* (Centropagidae) from mainland China to the Sacramento-San Joaquin Estuary of California. Journal of Plankton Research, 5(3): 357 – 375.

[2] Orsi, J.J. 1999. Long-term trends in mysid shrimp and zooplankton. Interagency Ecological Program Newsletter, 12(2): 13 – 15.

[3] Kimmerer, W., Penalva, C., Bollens, S., Avent, S., Cordell, J. 1999. Interagency Ecological Program Newsletter, 12(2): 16 – 21.

Mysid MAST file has the weighted values in the "Mysid BPUE Matrix 1972 – 2011" tab ... that's the one I use ... make that clear

Table 11: A summary of select fields from the raw zooplankton data set (above the double line) and the raw mysid data set (below the double line), organized by taxon and, in some cases, life stage. *Field* gives the field name used in the raw data set, *Description* describes the species or group of species represented by the field, and *Sampling Period* shows the year range during which the fields have been used. Asterisks indicate "catch all" categories that exclude species that were explicitly being counted at the time. *Weight* represents the estimated carbon biomass of an individual organism in the corresponding field (Wim Kimmerer, personal communication). *Status* indicates whether a field represents native or introduced species. In the latter case, the last column gives the year the species are hypothesized to have been introduced, or the year in which they first became abundant [1, 2, 3].

Taxon	Life Stage	Field	Description	Sampling Period	Weight (μgC)	Status	Intro Year
Copepod (Calanoid)	nauplius	COPNAUP	Copepod nauplii*	1972 – 1988	0.1		
		OTHCOPNAUP	Other copepod nauplii*	1989 – present	0.1		
		EURYNAUP	*Eurytemora affinis* nauplii	1989 – present	0.1	Introduced?	?
		SINONAUP	*Sinocalanus doerrii* nauplii	1989 – present	0.07	Introduced	1979
		PDIAPNAUP	*Pseudodiaptomus* spp. nauplii	2000 – present	0.1		
Copepod (Calanoid)	juvenile	CALJUV	Calanoid copepodids*	1972 – 1988	1.5		
		OTHCALJUV	Other calanoid copepodids*	1989 – present	1.5		
		EURYJUV	*Eurytemora affinis* copepodids	1989 – present	1.4	Introduced?	?
		SINOCALJUV	*Sinocalanus doerrii* copepodids	1989 – present	1.81	Introduced	1979
		PDIAPJUV	*Pseudodiaptomus* spp. copepodids	1990 – present	1.25		
		ASINEJUV	*Acartiella sinensis* copepodids	2006 – present	1.16	Introduced	...
		ACARJUV	*Acartia* spp. copepodids	2006 – present	1.3	Native	NA
		DIAPTJUV	*Diaptomidae* copepodids (includes several genera)	2006 – present	2		
		TORTJUV	*Tortanus* spp. copepodids	2006 – present	7.95		
Copepod (Calanoid)	adult	EURYTEM	*Eurytemora affinis*	1972 – present	3.07	Introduced?	?
		OTHCALAD	Other Calanoid adults*	1972 – present	3		
		SINOCAL	*Sinocalanus doerrii*	1978 – present	3.4	Introduced	1979
		PDIAPFOR	*Pseudodiaptomus forbesi*	1988 – present	2.66	Introduced	1988
Copepod (Cyclopoid)	adult	AVERNAL	*Acanthocyclops vernalis*	1972 – present	3.36		
		LIMNOSPP	*Limnoithona* spp.	1979 – present	0.13		
		LIMNOSINE	*Limnoithona sinensis*	2007 – present	0.13	Introduced	1993
		LIMNOTET	*Limnoithona tetraspina*	2007 – present	0.13	Introduced	1994
Caldoceran		BOSMINA	*Bosmina longirostris*	1972 – present	0.6		
		DAPHNIA	*Daphnia* spp.	1972 – present	4		
		DIAPHAN	*Diaphanosoma* spp.	1972 – present	1		
		OTHCLADO	Other cladocera*	1972 – present	1		
Mysid		H_longirostris	*Hyperacanthomysis longirostris* (formerly *Acanthomysis bowmani*)	1993 – present	size-dependent	Introduced	1993
		N_mercedis	*Neomysis mercedis*	1972 – present	size-dependent	Native	NA

Table 12: An example of a clean zooplankton data set (top) and a clean mysid data set (bottom). Both contain prey metrics calculated by year-month-region, as in Equation 1.

Year	Month	Region	naup_juvBPUV	juv_adultBPUV	juv_adult_cladBPUV
1972	January	Far West	6.189	6.971	6.971
1972	February	Far West	6.037	7.019	7.030
1972	March	Far West	6.192	6.915	6.916
1972	April	Far West	5.228	5.122	5.138
1972	May	Far West	7.377	7.462	7.462
⋮					

Year	Month	Region	mysidBPUV
1972	January	Far West	4.147
1972	February	Far West	5.190
1972	March	Far West	6.426
1972	April	Far West	7.134
1972	May	Far West	6.422
⋮			

Table 13: A summary of the year-month-region combinations for which zooplankton prey metrics were imputed. Values in the table represent region (FW = Far West; W = West; N = North; S = South; All = FW, W, N, S).

	Month						
Year	January	February	March	July	October	November	December
1973	All	All	0	0	0	0	All
1974	All	All	0	0	0	0	All
1975	All	All	0	0	0	0	All
1976	All	All	0	0	0	0	0
1977	0	0	0	0	0	FW	0
1978	0	N	0	0	0	0	N, S
1979	N	0	0	0	0	0	S
1980	N, S	N, S	0	0	0	0	0
1981	0	0	0	0	0	0	N, S
1982	N, S	N, S	0	0	0	0	N, S
1983	N, S	N, S	0	0	0	0	All
1984	All	All	0	0	0	0	All
1985	All	All	0	0	0	0	All
1986	All	All	N	0	0	0	All
1987	All	All	0	0	0	0	All
1988	All	All	0	All	0	0	All
1989	All	All	0	0	0	0	All
1990	All	All	0	0	0	0	All
1991	All	All	FW, N, S	0	0	0	All
1992	All	All	0	0	0	0	All
1993	All	All	0	0	0	0	All
1994	All	All	0	0	0	0	0
1995	0	0	0	0	FW	0	0
1998	FW	0	0	0	0	0	0
2002	0	0	0	0	N	0	0

Table 14: A summary of the year-month-region combinations for which mysid prey metrics were imputed. (FW = Far West; W = West; N = North; S = South; All = FW, W, N, S).

	Month					
Year	January	February	March	July	October	December
1973	All	All	N	0	0	All
1974	All	All	0	0	0	All
1975	All	All	0	0	0	All
1976	All	All	0	0	0	0
1978	0	N	0	0	0	N, S
1979	N	0	0	0	0	S
1980	N, S	N, S	0	0	0	0
1981	0	0	0	0	0	N, S
1982	N, S	N, S	0	0	0	N, S
1983	N, S	N, S	0	0	0	All
1984	All	All	0	0	0	All
1985	All	All	0	0	0	All
1986	All	All	N	0	0	All
1987	All	All	0	0	0	All
1988	All	All	0	All	0	All
1989	All	All	0	0	0	All
1990	All	All	0	0	0	All
1991	All	All	FW, N, S	0	0	All
1992	All	All	0	0	0	All
1993	All	All	0	0	0	All
1994	All	All	0	0	0	0
2002	0	0	0	0	N	0

Table 15: Each R cleaning file name has the prefix "Data Cleaner -". For example, the full name of the Benthos cleaning file is "Data Cleaner - Benthos.R".

R cleaning file (.R)	**Data Type**	**"Raw" data files (.csv)**	**"Clean" data files (.csv and .R)**
Benthos	Benthos	Allbenthic1	Benthos
		Allbenthic2.csv	
		benthos.station.coordinates	
Fish Surveys	Catch	Bay_Catch_1980_2011	Bay
		Chipps_Catch_1976_2011	Chipps
		FMWT_Catch_1967_2010	FMWT
		SKT_Catch_2002_2010	SKT
		STN_Catch_1959_2010	STN
		Twentymm_Catch_1995_2010	Twentymm
	Length	Bay_DSM_Lengths	
		FMWT_DSM_Lengths	
		SKT_DSM_Lengths	
		STN_DSM_Lengths	
		Twentymm_DSM_Lengths	
	Station	Bay_Stations_coords	
		FMWT_Stations_coords	
		SKT_Stations_coords	
		STN_Stations_coords	
		Twentymm_Stations_coords	
	Tide	Bay_Tide_Vars	
		Chipps_Tow_Tide_Vars	
		FMWT_Tide_Vars	
		SKT_Tide_Vars	
		STN_Tide_Vars	
		Twentymm_Tide_Vars	
Mysid	Mysid	1972-2010 Mysid Matrix	Mysid
		EMPMysidMatrixMAST	Mysid_cohort
		Mysids.taxon.cutoffs	
		ZPStations	
Salvage	Salvage	Salvage.Physical	
Zooplankton	Zooplankton	CB.taxon.cutoffs	Zooplankton
		zooplankton 1972-2010 CB Matrix	Zooplankton_cohort
		ZPStations	