DSM TN 1. Delta Smelt Life Cycle Model Data Files

Lara Mitchell and Ken Newman

U.S. Fish and Wildlife Service

May 9, 2017

DRAFT

Many data sets have been assembled in support of the Delta Smelt Life Cycle Model. 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 both CSV files and R objects with the same root file names. For brevity, only the CSV file names are presented here. This document provides descriptions of the raw data sets, clean data sets, and R cleaning scripts, organized by data type. All files and procedures described here are subject to change.

Acronyms

DSLCM - Delta Smelt Life Cycle Model

CDFW - California Department of Fish and Wildlife
EMP - Environmental Monitoring Program
FWS - U.S. Fish and Wildlife Service
IEP - Interagency Ecological Program for the San Francisco Estuary

Bay - Bay Study Midwater Trawl, conducted by CDFW
Chipps - Chipps Survey, conducted by FWS
FMWT - Fall Midwater Trawl Survey, conducted by CDFW
SKT - Spring Kodiak Trawl Survey, conducted by CDFW
STN - Summer Townet Survey, conducted by CDFW
Twentymm - 20mm Survey, conducted by CDFW

CCFB - Clifton Court Forebay
CVP - Central Valley Project water management project
SWP - State Water Project water management project

1 Fish Survey Data

The R script `DataCleaner_FishSurveys.r` creates separate clean catch data sets for the Bay, FMWT, SKT, STN, and Twentymm fish surveys. For each survey, the script merges four raw data sets, categorized as station, catch, length, and tide data, to produce a standardized data set containing delta smelt catch, age, and fork length information, select environmental variables, and tide information. The clean data sets are "updated" versions of the raw catch data sets, designed to have one record per unique combination of sampling date and sampling station.

The script `DataCleaner_Chipps.r` creates a clean catch data set for the Chipps survey. 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.

1.1 Raw Data

1.1.1 Station Data

The raw station data sets are stored in the Excel files listed below, with the corresponding survey name or abbreviation included in the file name. Copies of these files were saved in CSV format for reading in to R. Within a data set, each row represents a different sampling station for the corresponding survey. The columns provide the three digit station code (defined by CDFW), the latitude and longitude of the station in decimal degrees, and the region and subregion in which the station falls (see documentation on the DSLCM for details on how region and subregion are defined). Part of the STN station file is shown in Table 1. The other station files are similar. There is no station file for the Chipps survey because it takes place at only one location: Chipps Island.

Raw Data Files

`Bay_Stations_coords.xlsx`
`FMWT_Stations_coords.xlsx`
`SKT_Stations_coords.xlsx`
`STN_Stations_coords.xlsx`
`Twentymm_Stations_coords.xlsx`

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

1.1.2 Fish Survey Catch Data

The raw fish survey catch data sets are stored in the Excel files listed below, with the corresponding survey name included in the file name. Copies of these files were saved in CSV format for reading in to R.

Raw Data Files

`BayStudy_MWT_1980-2014_FishMatrix.xlsx`
`Chipps_Catch_1976_2011.xlsx`
`Chipps_Catch_2011_2016.xlsx`
`FMWT 1967-2015 Catch Matrix_updated.xlsx`
`Mitchell_SKT_2016Update.xlsx`
`LMithcell_DatReq_STN_2016.xls`
`Mitchell_20_mm_CatchMatrix_1995_2016.xlsx`

The raw Bay catch file was downloaded from the CDFW ftp site in February 2016.

The Chipps catch files were provided by Jonathan Speegle (FWS) on February 19, 2016.

The FMWT file was provided by Sarah Finstad (CDFW) on February 12, 2016.

The SKT file was provided by Lauren Damon on October 27, 2016.

The STN file was provided by Felipe La Luz (CDFW) in October 2016. Note that the STN data file appears to only contain core index stations (plus station 340, which has been sampled since 1978), while other data files contain both index and non-index stations.

The Twentymm catch file was provided by Lauren Damon (CDFW) on September 9, 2016.

Each catch data set describes the fish species composition of the survey on a per-tow basis. In the case of Bay, FMWT, SKT, STN, and Twentymm, each record in the data set corresponds to a single tow and 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; some of these field names vary between surveys (for example, Time vs. TimeStart). Further details on how the data are collected or calculated are available through the CDFW website.

The first 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 through the Lodi FWS website. 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 fork length) 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 second Chipps Excel file contains data from later years not included in the first file.

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.	✓	✓	✓	✓	✓
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	For Bay: an indication of tow "quality". For the other surveys: the unique tow number at a given station, on a given date.	X	X	✓	✓	✓
Volume	Estimate of water volume sampled (m^3).	✓	✓	✓	✓	✓
TowDirection	Tow direction code: 1=with current, 2=against current, 3=unknown (during slack).	✓	✓	✓	X	X
Secchi	Secchi depth (For Chipps: m; other surveys: 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 codes (Bay: 1=flood, 2=ebb, 3=low slack, 4=high slack; other surveys: 1=high slack, 2=ebb, 3=low slack, 4=flood).	✓	✓	✓	✓	✓
Depth	Depth of water at the station (Bay: m; other surveys: ft).	✓	✓	✓	✓	✓
SalinSurf	Salinity (ppt) for first meter of water column.	X	X	✓	X	X
delta.smelt	Number of delta smelt in the tow.	✓	✓	✓	✓	✓

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 fork length.

1.1.3 Delta Smelt Length Data

The raw length data files contain fork length measurements on delta smelt caught in the fish surveys. For each tow, the number of smelt measured for length is usually (but not always) equal to the total number caught. The Excel files containing the raw length data are listed below, with the corresponding survey name included in the file name. Copies of these files were saved in CSV format for reading in to R. All delta smelt fork lengths are in millimeters.

Raw Data Files

`Bay_DSM_Lengths_1980_2014.xlsx`
`FMWT_DSM_Lengths_1967_2015.xlsx`
`Mitchell_SKT_2016Update_DSM_Lengths.xlsx`
`STN_DSM_Lengths_1959_2016.xlsx`
`Mitchell_20-mm_DS_Lengths_1995_2016.xlsx`

The Bay length file was generated by Lara Mitchell using a copy of the Bay Study Access data base obtained from the CDFW ftp site in February 2016. This file contains fork length measurements for delta smelt caught by the Bay midwater trawl. Each record represents a unique tow and fork length combination, with the field `Frequency` giving the total number of delta smelt represented by that record.

The FMWT length file was provided by Sarah Finstad on January 15, 2016. It has a structure that is similar to the Bay file, except that the frequency column is labelled `LengthFrequency`.

The STN length file was provided by Felipe La Luz by in October 2016. It has the same structure as the FMWT length file, and also uses the field name `LengthFrequency`. Fork length measurements are not available prior to 1973.

The Twentymm length file was provided by Lauren Damon on September 9, 2016. It has the same structure as the FMWT file, but uses the frequency field name `CountOfLength`.

The SKT length file was provided by Lauren Damon on October 27, 2016. It has a separate record for each individual delta smelt caught in the survey. In addition to fork length data, it contains sex and reproductive information.

Chipps delta smelt fork length information is represented in the `ForkLength` field of the Chipps raw catch data file, rather than in a separate length file.

1.1.4 Tide Data

The tide data sets listed below were created by Chandra Chilmakuri (CM2H-Hill), and contain tidal information from the times and locations where fish surveys took place. The primary data fields are summarized in Table 4. There exists one record per date-station in the case of Bay, FMWT, SKT, STN, and Twentymm, and one record per date-time in the case of Chipps. Versions of these files were saved in CSV format for reading in to R.

Raw Data Files

`Bay_Tide_Vars.xlsx`
`Chipps_Tow_Tide_Vars.xlsx`
`FMWT_Tide_Vars.xlsx`
`SKT_Tide_Vars.xlsx`
`STN_Tide_Vars.xlsx`
`Twentymm_Tide_Vars.xlsx`

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

Field Name	Description
Date	Fish survey sample date.
TimeStart	Fish survey sample time.
Region	Fish survey region designation.
Station	Fish survye 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).

1.2 Clean Data

The six clean catch data files are listed below. The field names and units used in the clean catch files are described below that. Additional columns are added when a survey conducts replicate tows; see the section on aggregating replicate tows.

Clean Data Files

`Bay_80_14.csv`
`Chipps_78_15.csv`
`FMWT_67_15.csv`
`SKT_02_15.csv`
`STN_59_16.csv`
`Twentymm_95_16.csv`

Clean Data Field Names

`Date` - Sample date.
`Year` - Sample year.
`Month` - Sample month.
`Survey` - Survey number.
`Station` - Station code.
`TimeStart` - Sample time.
`Volume` - Volume of water sampled (m^3).
`TowDirection` - Tow direction string (With_Current, Against_Current, or Neither).
`Region` - Sampling region, as defined in the DSLCM.
`SubRegion` - Sampling subregion.
`Lat` - Latitude (degree decimal).
`Lon` - Longitude (degree decimal).
`Secchi` - Secchi depth (cm).
`CondSurf` - Surface conductivity (µS).
`TempSurf` - Surface temperature (°C).
`SalinSurf` - Surface salinity (ppt).
`CondBott` - Bottom conductivity (µS).
`Tide` - Tide string (High_Slack, Ebb, Low_Slack, or Flood).
`Depth` - Depth to bottom (ft).
`Inland_silverside` - Number of inland silverside caught.
`Striped_bass_age0` - Number of age 0 striped bass caught.
`Striped_bass_age1_plus` - Number of age 1+ striped bass caught.
`Striped_bass_all` - Total number of striped bass caught.
`Longfin_Smelt` - Total number of longfin smelt caught.
`Threadfin_Shad` - Total number of threadfin shad caught.
`Tridentiger_spp` - Total number of Tridentiger gobies caught.
`delta.smelt` - Number of delta smelt caught.
`delta.smelt.age0` - Number of age 0 delta smelt caught.
`delta.smelt.age1` - Number of age 1 delta smelt caught.
`Age0_n_L` - Number of age 0 delta smelt measured for fork length.
`Age0_L_bar` - Age 0 delta smelt mean fork length (mm).
`Age0_s_L` - Age 0 delta smelt fork length standard deviation (mm).
`Age1_n_L` - Number of age 1 delta smelt measured for fork length.
`Age1_L_bar` - Age 1 delta smelt mean fork length (mm).
`Age1_s_L` - Age 1 delta smelt mean fork length standard deviation (mm).

Age0_L_min - Minimum age 0 delta smelt fork length (mm).
Age0_L_max - Maximum age 0 delta smelt fork length (mm).
Age1_L_min - Minimum age 1 delta smelt fork length (mm).
Age1_L_max - Maximum age 0 delta smelt fork length (mm).
TideStage - Tide level (converted from ft to m).
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 (converted from ft/s to m/s).
EbbType - 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).
Cable.Out - Length of cable let out during tow (ft). Used to calculate EstimatedTowDepth_ft.
EstimatedTowDepth_ft - Estimated maximum depth that the trawl reached during a tow (ft).
Age0_age_in_days - Pseudo age (in days) of an age 0 delta smelt given its catch date and an assumed "cohort-wide" hatch date of March 1st.
Age0_pgt - Estimated probability of the trawl catching an age 0 delta smelt on the given sample date, given an assumed population length distribution.
Age1_age_in_days - Pseudo age (in days) of an age 1 delta smelt given its catch date and an assumed "cohort-wide" hatch date of March 1st.
Age1_pgt - Estimated probability of the trawl catching an age 1 delta smelt on the given sample date, given an assumed population length distribution.

General Procedure:

The general procedure for producing a clean catch data set is described here. Details specific to a given fish survey are included as necessary.

1. Merge Station Data

First, the station and catch data set are merged by station code. Records with stations that are not included in the station data set, and records with stations that are located outside of the four DSLCM regions (Far West, West, North, and South), are removed from the merged data. Next, fields are renamed as necessary so that the merged file has the standardized field names shown above. Fields not originally represented in the catch file are added and filled in with the value NA.

2. Make Survey-Specific Changes:

Bay

The field SalinSurf is used to calculate CondSurf using the following conversion equation: $Conductivity = 178500\left(1 - e^{-0.01 * Salinity}\right)$. This equation may need correcting (Wim Kimmerer, personal communication). The Depth field is converted from meters to feet. The numerical levels of Tide and the numerical levels of TowDirection are changed to descriptive strings for ease of interpretation (see below). The mapping for Tide is: 1 = "Flood", 2 = "Ebb", 3 = "Low_Slack", 4 = "High_Slack". The mapping for TowDirection the mapping is: 1 = "With_Current", 2 = "Against_Current", 3 = "Neither".

FMWT

The numerical levels of Tide and the numerical levels of TowDirection are changed to descriptive strings for ease of interpretation. The mapping for Tide is: 1 = "High_Slack", 2 = "Ebb", 3 = "Low_Slack", 4 =

"Flood". The mapping for TowDirection is: 1 = "With_Current", 2 = "Against_Current", 3 = "Neither".

SKT

Records with survey numbers greater than or equal to 6 are removed from the clean data set. These are special, non-routine surveys. The record from 3/9/2004, survey 3, station 610, at 13:30 has indeterminate Tide and TowDirection values (indicated by 0's). The previous 7 tows from that date have numerical values of 2 for both fields, so the 0's are replaced with 2's. After this, the numerical levels of Tide and the numerical levels of TowDirection are changed to descriptive strings for ease of interpretation. The mapping for Tide is 1 = "High_Slack", 2 = "Ebb", 3 = "Low_Slack", 4 = "Flood". The mapping for TowDirection is: 1 = "With_Current", 2 = "Against_Current", 3 = "Neither".

STN

Tow volumes prior to 2003 are unavailable, but an average volume of 735 m^3 has been provided by CDFW in the raw catch file, and this value is also used in the clean data file. Some values of TimeStart are missing; these are left as blank strings. Infrequent 4th tows, indicated by a Tow value of 4, are removed per advice from Julio Adib-Samii (personal communication). The numerical levels of Tide are changed to descriptive strings for ease of interpretation. The mapping for Tide is: 1 = "High_Slack", 2 = "Ebb", 3 = "Low_Slack", 4 = "Flood". Note that many environmental field variables are imputed in the STN data set (see Table 6).

Twentymm

Some values of TimeStart are missing; these are left as blank strings. Records with survey numbers greater than or equal to 10 are removed from the clean data set. These are special, non-routine surveys. Missing values of the field Depth are filled in with the value 32 (provided by Trishelle Morris) rather than an average value. The numerical levels of Tide are changed to descriptive strings for ease of interpretation. The mapping for Tide is: 1 = "High_Slack", 2 = "Ebb", 3 = "Low_Slack", 4 = "Flood".

Chipps

The clean Chipps catch data set is structured to have one record for every unique date-startTime combination. The field Secchi is converted from meters to centimeters. The field Region is filled in with the value "West", SubRegion is filled in with "Honker Bay". Station is filled in "Chipps", Lat is filled in with 38.055, and Lon is filled in with -121.9109. Some records have a ForkLength value of 0. These are changed to NA before length and age statistics are calculated. Delta smelt records with fork lengths less than 25 mm or greater than 100 mm are reclassified as "Other Smelt" and hence not used in constructing the Chipps clean data set. The decision to remove fish less than 25 mm 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 100 mm is based on the hypothesis that larger delta smelt may have been misidentified in the past (William Bennett, personal communications). See Table 5 for a summary of the records removed based on fork length.

Table 5: Frequency of Chipps delta smelt records removed with fork length > 100 mm or < 25 mm, by year and month.

Month	Year																		Total
	1979	1981	1982	1983	1985	1989	1990	1991	1992	1994	1995	1996	1998	1999	2002	2003	2004	2013	
January	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
March	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
April	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
May	0	2	0	2	2	1	1	0	1	0	0	1	1	0	0	0	0	0	11
June	1	0	2	0	0	2	0	1	0	0	1	0	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	0	2
November	0	0	0	0	0	0	0	0	0	0	1	0	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	0	4
Total	1	2	2	2	2	3	1	1	1	1	3	1	2	1	1	1	1	2	28

(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 < 25 mm.

3. Include Predator and Competitor Fields

Fields containing counts of striped bass (Morone saxatilis), inland silverside (Menidia beryllina), longfin smelt, threadfin shad, and Tridentiger goby are included in the clean data set. Some raw catch files have separate fields for age 0, 1, 2, and/or 3 striped bass. In the clean data set, there are separate fields for age 0 striped bass, age 1+ striped bass, and total striped bass.

4. Impute Volume and Environmental Fields

At this point, attempts are made to fill in values of `Volume`, `Secchi`, `CondSurf`, `TempSurf`, `SalinSurf`, `CondBott`, and `Depth` that are either missing or physically unrealistic. For depth, values of 0 are considered physically unrealistic and replaced. Similarly, secchi values of 0 are replaced. For FMWT, volumes less than 3000 are also replaced (Sarah Finstad, personal communication).

The general procedure is to try substituting with mean values calculated by date-station, then by date-subregion, then date-region, year-month-subregion, year-month-region, month-region, and finally by year-region. Averages calculated by date-station are tried first because this method uses data from records that are close to the missing record in both time and space. This method can be used when multiple tows were carried out at a single date-station. For records that are still missing or physically unrealistic after this, an attempt is made to substitute mean values calculated per date-subregion. In this case, the available data are still close to the missing record in time, but cover a wider geographic range. Next, mean values calculated per date-region are substituted, when available. If at this point values are still missing, the time frame is expanded to the same month as the missing record (within the same year), and substitutions are carried out using means calculated per year-month-subregion, then per year-month-region. Finally, means calculated per month-region (across years) and year-region (across months) are tried.

Missing CondBott values are first filled in using predicted values from a linear model for bottom conductivity as a function of the corresponding surface conductivity. After this, any missing values are filled in using the procedure described above with the seven alternative average values.

For Chipps, missing volumes are filled in with mean volumes calculated 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.

Table 6: Number of missing or physically unrealistic values imputed by fish survey (row) and imputation method (column). Fields not shown either did not need to have values imputed, or did not exist in the raw catch file to begin with. Dashes indicate that all missing or invalid values were successfully replaced.

	Date-Station	Date-SubRegion	Date-Region	Year-Month-SubRegion	Year-Month-Region	Month-Region	Year-Region	Year Month	Linear Model
Bay									
CondSurf	0	51	17	16	26	183	-	-	-
TempSurf	0	53	22	17	26	171	-	-	-
SalinSurf	0	51	17	16	26	183	-	-	-
Chipps									
Volume	0	0	0	0	0	0	0	79	-
Secchi	788	0	0	381	-	-	-	-	-
TempSurf	162	0	0	159	-	-	-	-	-
FMWT									
Volume	300	513	70	56	40	6701	22	-	-
Secchi	1	191	54	33	28	16	-	-	-
CondSurf	1	58	43	15	69	16	-	-	-
TempSurf	0	41	23	39	57	7	-	-	-
CondBott	0	0	0	0	0	0	0	0	15870
Depth	0	11	9	4	24	4583	22	-	-
SKT									
Secchi	2	1	7	-	-	-	-	-	-
CondSurf	2	-	-	-	-	-	-	-	-
TempSurf	2	-	-	-	-	-	-	-	-
Depth	0	7	4	-	-	-	-	-	-
STN									
Volume	1	-	-	-	-	-	-	-	-
Secchi	0	55	61	239	7	3267	-	-	-
CondSurf	0	11	25	84	41	3327	-	-	-
TempSurf	0	36	39	255	47	4494	-	-	-
CondBott	0	0	0	0	0	0	0	0	11026
Depth	0	42	44	162	60	4290	-	-	-
Twentymm									
Volume	10	-	-	-	-	-	-	-	-
Secchi	0	15	69	101	21	-	-	-	-
CondSurf	0	9	37	108	-	-	-	-	-
TempSurf	0	15	22	60	-	-	-	-	-
CondBott	373	-	-	-	-	-	-	-	-

5. Create Selectivity and Catchability Fields

The last five fields of the clean catch data set are intended to be used for analyzing gear selectivity for delta smelt. They are based on preliminary analyses and are subject to change.

The field `EstimatedTowDepth_ft` gives an estimate of the maximum depth in feet reached by the trawl during a tow. For Twentymm, this is calculated as $(3.937/25) * \texttt{Cable.Out} - 8.3$, where `Cable.Out` is the number of feet of cable let out (from the raw catch data set) and it is estimated that for every 25 feet of cable let out, the trawl drops 3.937 feet. 8.3 is the average distance, in feet, from the block to the water surface for the Twentymm boats (Trishelle Morris, personal communication). For FMWT, the formula is $(3.937/25) * \texttt{Cable.Out} - 6.67$ (Sarah Finstad, personal communication). Bay is operated such that the net descends to close to the station depth, but not so deep that the net plows the substrate (Kathy Hieb, personal communication). For example, the tow depth at a 20 foot-deep station is roughly 18 to 19 feet. The net does not fish below 40 feet though, so if a station depth is greater than 40 feet, the tow depth will remain 40 feet (Kathy Hieb, personal communication). In the Bay clean data set, `EstimatedTowDepth_ft` is set equal to `Depth` except in cases where `Depth` is greater than 40 feet, in which case 40 feet is used instead.

`Cable.Out` values that are missing or 0 are imputed using the process described in step 2. For FMWT, one cable value of 15 is also replaced because it leads to a negative tow depth. For Twentymm, cable values of less than 53 feet are replaced because they lead to negative tow depths. See Table 7 for a summary of imputed `Cable.Out` values. In cases where `EstimatedTowDepth_ft` ends up being greater than `Depth`, the value of `EstimatedTowDepth_ft` is replaced by the value of `Depth`; see Table 8.

The fields `Age0_age_in_days` and `Age1_age_in_days` give the pseudo ages (in days) of age-0 and age-1 delta smelt based on its catch date and assuming a "cohort-wide" hatch date of March 1st. The fields `Age0_pgt` and `Age1_pgt` give estimates of the probabilities of catching age-0 or age-1 delta smelt on that date given an assumed population length distribution. These values are currently coming from the file `prob.catch.bygear.dayD.df_3_7_2016.csv`. Details on how these values are calculated will be coming soon.

Table 7: Number of missing values of `Cable.Out` imputed by fish survey (row) and imputation method (column). Note that when `Cable.Out` is imputed, `EstimatedTowDepth_ft` is also necessarily imputed.

	Date-Station	Date-SubRegion	Date-Region	Year-Month-SubRegion	Year-Month-Region	Month-Region	Year-Region	Year Month
FMWT								
Cable.Out	1544	1522	327	35	22	13671	482	-
STN								
Cable.Out	0	67	158	173	66	2995	-	-
Twentymm								
Cable.Out	3	66	44	-	-	-	-	-

Table 8: Number of EstimatedTowDepth_ft values replaced with the corresponding value of Depth.

Survey	Number Replaced
FMWT	3434
STN	794
Twentymm	145

6. Merge Length Data

The raw length data set is used to calculate age and length-related fields in the clean data set. For SKT, which has a separate record for every fish, an age assignment key is used to assign an age (0 or 1) to each fish based on fork length and month-of-catch. The key, shown in Table 9, was developed by CDFW (Steve Slater, personal communication). The records are then aggregated by unique tow, and the fields `Age0_n_L` through `Age1_L_max` are calculated for each tow. This aggregated length data set is then merged with the clean data set. The fields `delta.smelt.age0` and `delta.smelt.age1` represent the total number of age 0 and age 1 delta smelt caught, and are calculated by multiplying the total delta smelt catch in the tow by the proportion of age 0 and age 1 individuals represented in the length data for that tow (e.g., `Age0_n_L`/(`Age0_n_L` + `Age1_n_L`) would be the proportion of age 0 smelt). The same process is used for Bay, FMWT, STN, and Twentymm, except first each record in the length data set is duplicated according to the frequency column in order to produce a data set with the same structure as the SKT length data set.

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`. Ages are then assigned to each delta smelt record using the CDFW age-assignment key, and length statistics are calculated on a date-time basis and merged with the clean data set.

Table 9: Delta smelt age assignment key. The numbers indicate the cut-off length (in mm) 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

7. Impute Ages

When length information is not available to calculate the age fields `delta.smelt.age0` and `delta.smelt.age1`, values are imputed using the following procedure. A mean proportion of age 0 smelt is calculated using the same procedure described in step 2. That is, first an attempt is made to calculate a mean by date-station, then if that is not possible, an attempt is made to calculate a mean by date-subregion, and so on. The imputed value of `delta.smelt.age0` is given by the product of the mean age 0 proportion and `delta.smelt`, rounded to the nearest integer. The imputed value of `delta.smelt.age1` is given by `delta.smelt` - `delta.smelt.age0`. See Table 10 for a summary of imputed age information.

For FMWT, the following age 0 proportions are used whenever calculated values are not available, including years prior to 1975: January: 0, February: 0, March: 0, May: 0, September: 0.9, October: 1, November: 1, December: 1. Additionally, if the calculated values in September, October, or November are less than 0.9, the value 0.9 is used instead. All of these values were provided by Dave Contreras (personal communication).

8. Aggregate Replicate Tows

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 given by summing the values of this field across the replicate tows; similarly for `delta.smelt.age0`, `delta.smelt.age1`, `Volume`, `Inland_silverside`, `Striped_bass_age0`, `Striped_bass_age1_plus`, `Striped_bass_all`, `Longfin_Smelt`, `Threadfin_Shad`, and `Tridentiger_spp`.

Table 10: Number of missing values of delta.smelt.age0 and delta.smelt.age1 imputed by fish survey (row) and imputation method (column). Note that when delta.smelt.age0 is imputed, delta.smelt.age1 is also imputed.

	Date-Station	Date-SubRegion	Date-Region	Year-Month-SubRegion	Year-Month-Region	Month-Region	Year-Region	Year Month
Chipps								
delta.smelt.age0/1	461	0	0	394	0	268	-	-
FMWT								
delta.smelt.age0/1	0	56	19	23	20	1153	6	-
STN								
delta.smelt.age0/1	16	15	14	57	4	1726	-	-
Twentymm								
delta.smelt.age0/1	6	-	-	-	-	-	-	-

The value of EstimatedTowDepth_ft for the aggregated record is given by the mean of the replicate tow depths. The fields Age0_n_L through Age1_L_max are recalculated at the date-station level using the length data set. All other fields, including TimeStart and TowDirection, are taken from the *first* tow record.

When tows are aggregated, additional fields are added to the clean data set in order to preserve catch information from the replicate tows. Let n be the maximum number of replicates conducted for any date-station combination. These additional fields have the same names as the aggregated fields except with .tow$i$ appended to the end, where $i = 1, \ldots, n$. For example, the fields delta.smelt.tow$i$, delta.smelt.age0.tow$i$, and delta.smelt.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.

9. Impute Lengths

After any replicate tows are aggregated, attempts are made to imput missing values of Age0_L_bar and Age1_L_bar. The process is the same as that used in step 2. See Table 11 for a summary of imputed length information. If values are unable to be imputed, they are left as NA. No attempts are made to impute sample sizes (Age0_n_L, Age1_n_L) or standard deviations (Age0_s_L, Age1_s_L).

10. Merge Tide Data

At this point, the tide data set is merged with the clean data set. The field TideStage is converted from feet to meters, and TideVelocity is converted from ft/s to m/s.

Table 11: Number of missing values of `Age0_L_bar` and `Age1_L_bar`) imputed by fish survey (row) and imputation method (column).

	Date-Station	Date-SubRegion	Date-Region	Year-Month-SubRegion	Year-Month-Region	Month-Region	Year-Region	Year Month
Chipps								
Age0_L_bar	131	0	0	306	0	263	-	-
Age1_L_bar	312	0	0	183	0	164	-	-
FMWT								
Age0_L_bar	0	50	15	13	18	815	5	-
Age1_L_bar	0	9	3	6	2	590	2	-
STN								
Age0_L_bar	0	7	5	23	2	826	-	-
Age1_L_bar	0	0	2	11	0	121	-	-

2 Average Temperature and Secchi Data

The clean Bay, Chipps, FMWT, SKT, STN, and Twentymm data sets were used to calculate mean water temperature and secchi values for every combination of Year-Month-Region between January 1980 and December 2015. The Year, Month, Region, TempSurf, and Secchi fields of the six clean fish survey data sets were combined into one data frame which was then used to calculate the average temperature and secchi values. Hence, these averages are calculated across survey type, sampling date, and sampling location within a given Year-Month-Region. Missing mean values were imputed by averaging over averages with the same Month and Region (i.e., by averaging across years). 145 mean temperatures and 145 mean secchis were imputed. The resulting data set contains the fields Region, Month, Year, MeanTemperature, and MeanSecchi, and has a separate record for each Year-Month-Region combination. This data set is created with the script Create_FishSurvey_TempSecchi.r, and saved in the file Mean_Temp_Secchi.csv. The mean temperature and secchi values are also stored in individual 3D arrays in the files Mean_Temp_3Darray.R and Mean_Secchi_3Darray.R.

Data collected in the Mid San Pablo Bay subregion *were* included in these calculations, which is at odds with the next two sections. I want to discuss this with everyone before making any changes, though.

3 Predator and/or Competitor Indexes

The clean FMWT, SKT, STN, and Twentymm data sets were used to calculate indexes of abundance for age 0 striped bass, age 1+ striped bass, inland silverside, threadfin shad, and Tridentiger goby by year-month-region using a stratified ratio-expansion procedure. See the technical note "TN2_Design_Based_Estimates_of_Delta_Smelt_Abundance" for a general description of the procedure. Note that the resulting values are indexes of abundance because we did not try to account for gear selectivity or fish availability/catchability. The calculations are done by the script Create_FishSurvey_PredCompetitor.r, and the resulting data set is saved in the file FishSurvey_PredCompetitor_long.csv, which has fields Calendar_Year, Month, Region, Gear, Species, and Index. The abbreviated name SBAge0 is used for age 0 striped bass, SBAge1Plus is used for age 1+ striped bass, ISS is used for inland silverside, TFS is used for threadfin shad, and TriGoby is used for Tridentiger goby. The file FishSurvey_PredCompetitor_wide.csv contains a wide-formatted version of the data set with field names Calendar_Year, Month, Region, Index_SBAge0_TMM, Index_ISS_TMM, etc.

Data collected in the Mid San Pablo Bay subregion were not included in these calculations.

4 Mean Length Data

The script Create_FishSurvey_MeanLength.r uses fish survey length data to calculate mean fish lengths in a given year-month (calculated over the stations sampled in that year-month). It produces the data files FishSurvey_MeanLength.csv and FishSurvey_MeanLength_cohort.csv. Some of the average lengths are adjusted for gear selectivity.

Mean lengths are calculated separately for age-0 and age-1 delta smelt. An typical field in the data file looks like this: MeanLength_TMM_DSM_Apr0_adj, where TMM indicates that the lengths came from the 20mm Survey, DSM indicates that the species is delta smelt, Apr0 indicates that it is the mean length of age 0 delta smelt in April, and adj means that 20mm gear selectivity estimates were used to try to adjust for gear selectivity when calculating the mean. MeanLength_TMM_DSM_Apr0_unadj would be the version without gear selectivity adjustments.

Threadfin shad (TFS) and Tridentiger goby (TriGoby) mean lengths are now also included in the clean length files. A typical field name looks like this: `MeanLength_STN_TFS_Jul`. No attempts are made to account for gear selectivity or separate out juvenile and adults. In the data file organized by cohort year, mean TFS and TriGoby lengths from year-month `y-m` are assigned to cohort year `y` if `m` is in March to December and cohort year `y - 1` otherwise.

Data collected in the Mid San Pablo Bay subregion were not included in these calculations.

5 Entrainment-Related Physical Variables

The R script `DataCleaner_EntrainPhysicalVar.r` creates a clean data set containing physical variable measurements, including delta flows and turbidity.

5.1 Raw Data

5.1.1 Dayflow Data

All of the files listed below, with the exception of `Daily Outflow and X2 1930-2011.xlsx`, were downloaded from the Dayflow home page by Lara Mitchell; the date on which each file was retrieved is shown in parentheses. These files contain, among other fields, daily values of Sacramento River flows, San Joaquin River flows, SWP and CVP exports, delta outflow, and QWest flows. X2 values are only present in the files for water year 1997 and later. Flow values are in thousands of acre-feet per day and X2 is in km. These raw data files were combined into a single clean data set spanning water years 1969 - 2016.

The file `Daily Outflow and X2 1930-2011.xlsx` was created by Fred Feyrer and contains X2 values from October 1, 1929 to December 31, 2011, with values prior to water year 1997 calculated according to the X2 equation provided on the Dayflow documentation website. A copy of this file was provided by Ken Newman on March 15, 2016. A copy of the worksheet named "Daily Outflow and X2 1930-2011" was saved in csv format for reading in to R. All values of X2 in the clean data set prior to water year 1997 come from this file.

Raw Data Files

`wy1970-1983.csv` (downloaded on February 24, 2016)
`wy1984-1996.csv` (downloaded on February 24, 2016)
`dayflowCalculations1997.csv` (downloaded on February 24, 2016)
`dayflowCalculations1998.csv` (downloaded on February 24, 2016)
`dayflowCalculations1999.csv` (downloaded on February 24, 2016)
`dayflowCalculations2000.csv` (downloaded on February 24, 2016)
`dayflowCalculations2001.csv` (downloaded on February 24, 2016)
`dayflowCalculations2002.csv` (downloaded on February 24, 2016)
`dayflowCalculations2003.csv` (downloaded on February 24, 2016)
`dayflowCalculations2004.csv` (downloaded on February 24, 2016)
`dayflowCalculations2005.csv` (downloaded on February 24, 2016)
`dayflowCalculations2006.csv` (downloaded on February 24, 2016)
`dayflowCalculations2007.csv` (downloaded on February 24, 2016)
`dayflowCalculations2008.csv` (downloaded on February 24, 2016)
`dayflowCalculations2009.csv` (downloaded on February 24, 2016)
`dayflowCalculations2010.csv` (downloaded on February 24, 2016)
`dayflowCalculations2011.csv` (downloaded on February 24, 2016)
`dayflowCalculations2012x.csv` (downloaded on February 24, 2016)
`dayflowCalculations2013x.csv` (downloaded on February 24, 2016)
`dayflowCalculations2014a.csv` (downloaded on February 24, 2016)
`dayflowCalculations2015.csv` (downloaded on February 24, 2016)
`dayflowCalculations2016.csv` (downloaded on April 21, 2017)
`Daily Outflow and X2 1930-2011.xlsx`

5.1.2 OMR Data

The file `OMR_Q_wy1980-cy2014.csv` contains combined daily Old River and Middle River flows in cfs. Values from water years 1987 - 2014 are based on data obtained online from USGS by Pete Smith (USGS). Values from water years 1980 - 1986 were imputed by Pete Smith from a regression of combined flows on exports and San Joaquin River flows at Vernalis.

The file `omr-2010-2017.csv` contains combined daily Old River and Middle River flows in cfs. The data were obtained online by Lara Mitchell on April 21, 2017 from the same USGS website cited above. For this file, missing values were imputed using simple linear interpolation. The code for creating this file is located in the R script `DataCleaner_EntrainPhysicalVar.r`, below where clean physical variable files are saved in csv and RData format.

In cases where both files contained the same date, values from the first file were used.

5.1.3 Turbidity Data

The file `daily_CCFB_turbidity_Mar88-Aug12.csv` contains daily turbidity measurements (in ntu) from CCFB for the years 1988 - 2010. The data were obtained from the CDEC website by Pete Smith.

The file `CCFB_Turbidity_Daily_2012_2017.txt` contains daily CCFB turbidity data from 2012 to 2017. These data also come from CDEC, and were retrieved by Lara Mitchell on April 21, 2017.

Missing values in both files were imputed using a simple moving average. In cases where both files contained the same date, values from the first file were used.

5.2 Clean Data

5.2.1 Daily Physical Variable Data Set

The Dayflow, OMR, and turbidity data described above were used to construct the clean file `Entrain_Physical_Daily_69_16.csv`, the fields of which are described below. Each row of the file represents a unique date, with all dates between October 1, 1969 and September 30, 2016 represented.

Clean Data Field Names

`Date` - Unique date.
`Year` - Calendar year corresponding to Date.
`Month` - Month corresponding to Date.
`Inflow` - Total Delta inflow (converted to cfs). Source: Dayflow Data.
`SacFlow` - Sacramento River flow (converted to cfs). Source: Dayflow Data.
`SJRFlow` - San Joaquin River flow (converted to cfs). Source: Dayflow Data.
`Outflow` - Total Delta outflow at Chipps Island (converted to cfs). Source: Dayflow Data.
`QWEST` - San Joaquin River flow past Jersey Point (converted to cfs). Source: Dayflow Data.
`SWP.Exports` - State Water Project (SWP) exports (converted to cfs). Source: Dayflow Data.
`CVP.Exports` - Central Valley Project (CVP) exports (converted to cfs). Source: Dayflow Data.
`Total.Exports` - Total exports, including SWP, CVP, and others (converted to cfs). Source: Dayflow Data.
`X2` - Distance from Golden Gate to 2ppt Salinity (km). Source: Dayflow Data.
`OMR` - Sum of Old River and Middle River flow (cfs). Source: OMR Data.
`OMR.scale` - OMR divided by the standard deviation of all daily OMR values.
`CCFB.Turbidity` - Clifton Court Forebay turbidity (ntu). Source: Turbidity Data.

CCFB.Turbidity.scale - CCFB.Turbidity divided by the standard deviation of all daily CCFB.Turbidity values.

5.2.2 Monthly Physical Variable Data Set

The file Entrain_Physical_Monthly_69_16.csv is a version of the daily file that is aggregated by year-month. Within a year-month, we sum over the fields Inflow, SacFlow, SJRFlow, Outflow, QWEST, SWP.Exports, CVP.Exports, and Total.Exports, and take the mean of the fields X2, OMR, OMR.scale, CCFB.Turbidity, and CCFB.Turbidity.scale. Field names remain the same between the daily and monthly files. The following values were substituted for the calculated values because of suspected errors in the daily data (Pete Smith, personal communication). The substituted values came from a regression analysis carried out by Pete Smith. Currently, changes are *not* made to the problematic daily data.

January 1991: 10.4 ntu
January 1994: 4.0 ntu
February 1994: 6.0 ntu
February 1995: 16.6 ntu
January 1996: 14.8 ntu
February 1996: 26.8 ntu
February 1997: 39.2 ntu
March 1997: 30.0 ntu

6 Prey Data

The R script `DataCleaner_ZooMysid_median_vX.r` creates a clean data set containing information on delta smelt prey items, including zooplankton and mysids.

6.1 Raw Data

6.1.1 Zooplankton and Mysid Data

Zooplankton and mysid data are collected through the CDFW Zooplankton Study, which is part of IEP's Environmental Monitoring Program. The study 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.

The raw data files listed below contain data collected by the Zooplankton Study, and were used to create a clean delta smelt prey data set.

Raw Data Files

`EMPCBMatricesMASTMay2017.xlsx`
`EMPMysidMatricesMASTMay2017.xlsx`
`CB.taxon.cutoffs.csv`
`Mysids.taxon.cutoffs.csv`
`ZPStations.csv`

The first file contains data on zooplankton species including copepods, cladocerans, and rotifers, and was provided by April Hennessy on May 4, 2017. The worksheet named "1972-2014 CB BPUE Matrix" contains Carbon biomass-per-unit-volume (BPUV, micrograms of Carbon/m^3) estimates for a variety of taxa with each record corresponding to a unique combination of sample date and sampling station. This worksheet was used as the raw zooplankton file. The second file contains data on mysids, and was provided by April Hennessy on May 3, 2017. The worksheet named "MysidBPUEMatrix1972-2016" contains BPUV estimates for different taxa with each record corresponding to a unique combination of sample date and sampling station. This worksheet was used as the raw mysid file. The carbon biomass of an organism 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 zooplankton and mysid Excel files describe how the BPUV estimates were calculated.

The third and fourth files contain information on the years during which each species in the zooplankton and mysid files have been monitored. This information is used to distinguish cases of 0 catches from cases in which data were not collected. The fifth file indicates to which of the DSLCM regions each EMP sampling station belongs; see Figure 1 for a map of the EMP stations. The last three files were provided by Ken Newman in 2014.

Table 12 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 12. Differences in collection periods are due, in part, to the fact that many of the species are non-indigenous to the bay-delta.

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

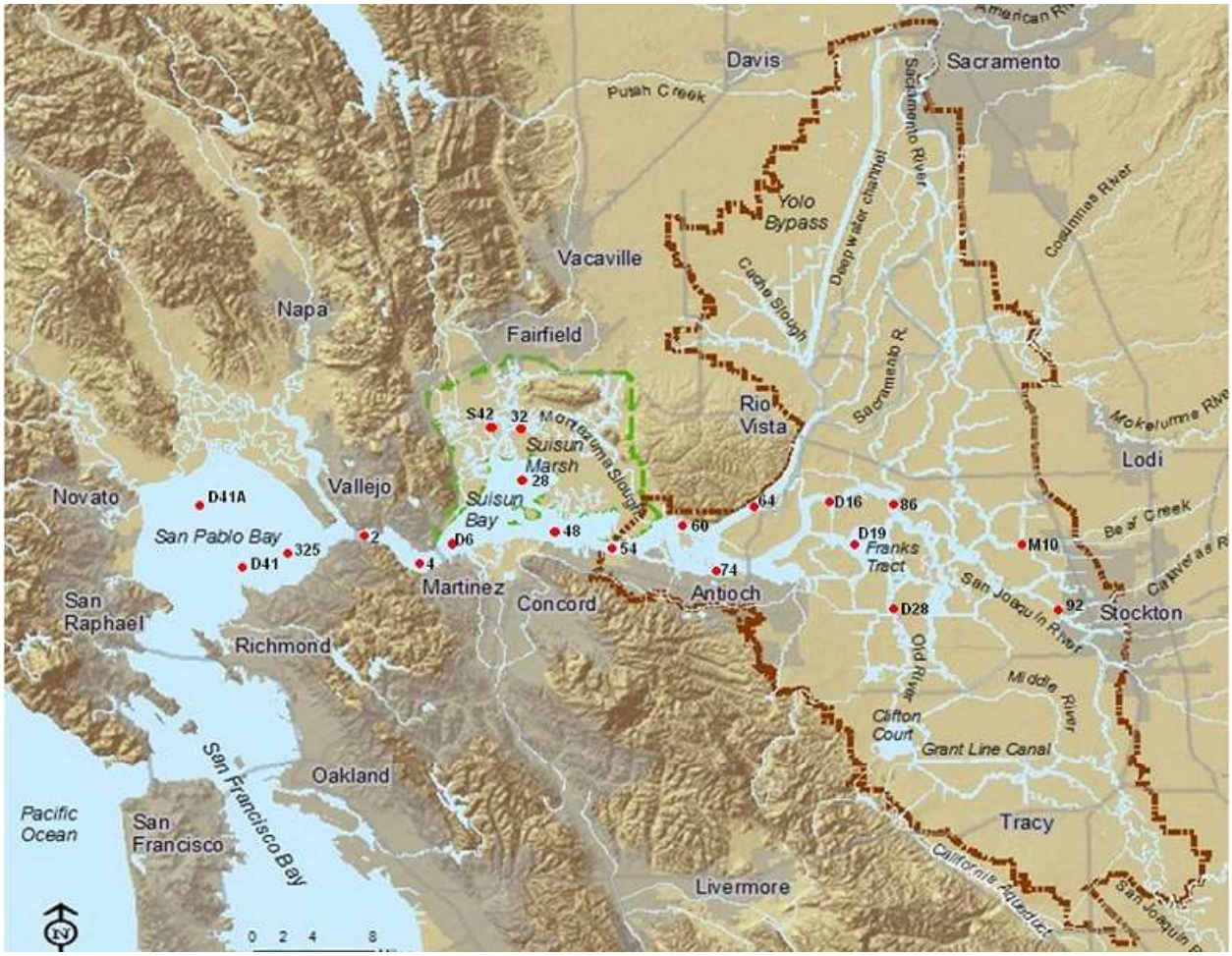

Figure 1: EMP Zooplankton Study sampling locations, shown as red dots (`http://www.dfg.ca.gov/`).

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).

6.2 Clean Data

The R script `DataCleaner_ZooMysid_median_vX.r` uses the raw zooplankton and mysid data sets to create a clean data file, called `ZooMysid_74_16_df_median.csv`, containing measures of zooplankton and mysid biomass calculated by year-month-region for the years 1974 – 2016. The field names in the clean file are listed below.

Clean Data Field Names

`Year` - Sample year.
`Month` - Sample month.
`Region` - Sampling region, as defined in the DSLCM.
`NJ_BPUV` - Prey metric composed of copepod nauplii and juveniles.
`JA_BPUV` - Prey metric composed of copepod juveniles and adults.
`JAC_BPUV` - Prey metric composed of copepod juveniles and adults, and cladocerans.
`NJAC_BPUV` - Prey metric composed of copepod nauplii, juveniles, and adults, and cladocerans.
`M_BPUV` - Prey metric composed of mysids.
`JACM_BPUV` - Prey metric composed of copepod juveniles, copepod adults, cladocerans, and mysids.
`NJACM_BPUV` - Prey metric composed of copepod nauplii, juveniles, and adults, cladocerans, and mysids.
`ACM_BPUV` - Prey metric composed of copepod adults, cladocerans, and mysids.

The first step in creating the clean file is to remove any records from the raw zooplankton data set that fall outside of the four main regions and replace any 0 BPUV values outside of each field's sampling period with `NA`'s. Next, any records that fall outside of the core sampling stations 1 and 2, surveys 3 – 11, and the years 1974+ are removed per a recommendation in the document "ReadMeZooplanktonStudyMatricesJune2015.doc." Then, eight measures of aggregated prey biomass are calculated from different combinations of zooplankton and mysid species at different life stages. Separate biomass estimates are calculated for each combination of year, month, and region, with the median BPUV being calculated across all sampling stations within the given region. The fields in the clean prey data set are described below, with the eight aggregated biomass field names ending in "BPUV." Details on the specific species used to construct each field are available in Table 12. Missing values were imputed by linearly interpolating across the year-month time series in a given region when data were available to do these calculations. Table 13 summarizes the imputation scheme for the `NJ_BPUV` field. The other fields were handled similarly.

Table 12: 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. *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 (Orsi et al. 1983; Orsi 1999; Kimmerer et al. 1999).

Taxon	Life Stage	Field	Description	Sampling Period	Status	Intro Year
Copepod (Calanoid)	nauplius	COPNAUP	Copepod nauplii*	1972 – 1988		
		OTHCOPNAUP	Other copepod nauplii*	1989 – present		
		EURYNAUP	*Eurytemora affinis* nauplii	1989 – present	Introduced?	?
		SINONAUP	*Sinocalanus doerrii* nauplii	1989 – present	Introduced	1979
		PDIAPNAUP	*Pseudodiaptomus* spp. nauplii	2000 – present		
Copepod (Calanoid)	juvenile	CALJUV	Calanoid copepodids*	1972 – 1988		
		OTHCALJUV	Other calanoid copepodids*	1989 – present		
		EURYJUV	*Eurytemora affinis* copepodids	1989 – present	Introduced?	?
		SINOCALJUV	*Sinocalanus doerrii* copepodids	1989 – present	Introduced	1979
		PDIAPJUV	*Pseudodiaptomus* spp. copepodids	1990 – present		
		ASINEJUV	*Acartiella sinensis* copepodids	2006 – present	Introduced	...
		ACARJUV	*Acartia* spp. copepodids	2006 – present	Native	NA
		DIAPTJUV	*Diaptomidae* copepodids (includes several genera)	2006 – present		
		TORTJUV	*Tortanus* spp. copepodids	2006 – present		
Copepod (Calanoid)	adult	EURYTEM	*Eurytemora affinis*	1972 – present	Introduced?	?
		OTHCALAD	Other Calanoid adults*	1972 – present		
		SINOCAL	*Sinocalanus doerrii*	1978 – present	Introduced	1979
		PDIAPFOR	*Pseudodiaptomus forbesi*	1988 – present	Introduced	1988
Copepod (Cyclopoid)	adult	AVERNAL	*Acanthocyclops vernalis*	1972 – present		
		LIMNOSPP	*Limnoithona* spp.	1979 – present		
		LIMNOSINE	*Limnoithona sinensis*	2007 – present	Introduced	1993
		LIMNOTET	*Limnoithona tetraspina*	2007 – present	Introduced	1994
Cladoceran		BOSMINA	*Bosmina longirostris*	1972 – present		
		DAPHNIA	*Daphnia* spp.	1972 – present		
		DIAPHAN	*Diaphanosoma* spp.	1972 – present		
		OTHCLADO	Other cladocera*	1972 – present		
Mysid		H_longirostris	*Hyperacanthomysis longirostris* (formerly *Acanthomysis bowmani*)	1993 – present	Introduced	1993
		N_mercedis	*Neomysis mercedis*	1972 – present	Native	NA

Table 13: A summary of the year-month-region combinations for which the field `NJ_BPUV` was imputed. Values in the table represent region (FW = Far West; W = West; N = North; S = South; All = Far West, West, North, and South). A value of 0 means that no imputation was necessary.

	Month								
Year	January	February	March	May	July	August	October	November	December
1974	0	0	0	0	0	0	0	0	All
1975	All	All	FW	0	0	0	0	0	All
1976	All	All	0	0	0	0	0	0	All
1977	All	All	0	0	0	0	0	FW,N	All
1978	All	All	0	0	0	0	0	0	All
1979	All	All	0	0	0	0	0	0	All
1980	All	All	0	0	0	0	0	0	All
1981	All	All	FW	0	0	0	0	0	All
1982	All	All	0	0	0	0	0	0	All
1983	All	All	0	0	0	0	0	0	All
1984	All	All	0	0	0	0	0	0	All
1985	All	All	0	0	0	0	0	0	All
1986	All	All	N	0	0	0	0	0	All
1987	All	All	0	0	0	0	0	0	All
1988	All	All	FW	0	All	0	0	0	All
1989	All	All	0	0	0	0	0	FW	All
1990	All	All	0	0	0	0	0	0	All
1991	All	All	FW,N,S	0	0	0	0	0	All
1992	All	All	0	0	0	0	0	0	All
1993	All	All	0	0	0	0	0	0	All
1994	All	All	0	0	0	0	0	0	All
1995	All	All	0	0	0	0	FW	0	All
1996	All	All	0	0	0	0	0	0	All
1997	All	All	0	0	0	0	0	0	All
1998	All	All	0	0	0	0	0	0	All
1999	All	All	0	0	0	0	0	0	All
2000	All	All	0	0	0	0	0	0	All
2001	All	All	0	0	0	0	0	0	All
2002	All	All	0	0	0	0	N	0	All
2003	All	All	0	0	0	0	0	0	All
2004	All	All	0	0	0	0	0	0	All
2005	All	All	0	0	0	0	0	0	All
2006	All	All	0	0	0	0	0	0	All
2007	All	All	0	0	0	0	0	0	All
2008	All	All	0	0	0	0	0	0	All
2009	All	All	0	0	0	0	0	0	All
2010	All	All	0	FW	0	0	0	0	All
2011	All	All	0	0	0	0	N	0	All
2012	All	All	0	0	0	0	0	0	All
2013	All	All	0	0	0	0	0	0	All
2014	All	All	0	0	0	0	0	0	All
2015	All	All	0	0	FW	N	0	0	All
2016	All	All	0	0	0	N	0	0	0

7 Salvage Data

The R script `DataCleaner_Salvage.r` creates a clean data set containing data on delta smelt salvaged at the State Water Project or the Central Valley Project.

7.1 Raw Data

The files listed below were provided by Geir Aasen (CDFW) and contain count and fork length (mm) information on salvaged delta smelt. Further information on data collection is available on the salvage section of the CDFW ftp site.

Raw Data Files

`ForkLengths-1979-1992.csv`
`ForkLengths-1993-2014.csv`
`Salvage-1979-1992.csv`
`Salvage-1993-2014.csv`

7.2 Clean Data

The clean salvage files are listed below, and cover the years 1979 to 2014. The first file contains the total number of age 0 delta smelt salvaged at the SWP and CVP combined. The data are grouped by year (row) and month (column). The second file is structured similarly, and contains total age 1 salvage. The third file contains daily smelt salvage counts and mean fork lengths, partitioned by age group (0 or 1) and facility (SWP or CVP).

Clean Data Files

`Salvage.Age0.Year.by.Month.csv`
`Salvage.Age1.Year.by.Month.csv`
`Salvage.Daily.csv`
`Salvage.Monthly.csv`

8 Spawning Water Quality Index

The R script `DataCleaner_WaterQuality.r` creates an index reflecting the quality of water temperature for delta smelt spawning on a year-month basis.

8.1 Raw Data

The files listed below contain hourly water temperature measurements (°F) from five data collection stations in the Bay-Delta. The data were downloaded from the California Data Exchange Center (CDEC) website on April 26, 2016. According to the CDEC station metadata site, the Antioch (ANC), Pittsburg (PTS), Rio Vista (RIV), and San Andreas Landing (SAL) stations are operated by the U.S. Bureau of Reclamation, and the Martinez (MRZ) station is operated by the California Department of Water Resources. The rows in each file represent unique date-hour combinations. Some changes were made to these files immediately after they were downloaded. Namely, HTML formatting statements were removed, and missing temperatures,

originally indicated by two dashes (- -), were replaced with the text NA. We note that there are other data collection stations represented on the CDEC website that could be considered beyond those considered here.

Raw Data Files

`CDEC_Temp_ANC_3-1-99_to_4-20-16.txt`
`CDEC_Temp_MRZ_7-1-94_to_4-20-16.txt`
`CDEC_Temp_PTS_4-1-99_to_4-20-16.txt`
`CDEC_Temp_RIV_2-22-99_to_4-20-16.txt`
`CDEC_Temp_SAL_2-23-99_to_4-20-16.txt`

8.2 Clean Data

The clean spawning water quality index data file, `SpawningWaterQualityIndex.csv`, contains a water quality index value for each combination of year-month between January 1995 and March 2016, with each row corresponding to a unique year-month.

An outline of the procedure for producing spawning water quality index values is as follows: clean the hourly temperature data; use the clean hourly data to calculate mean daily temperatures; calculate the water quality index for a given month as a weighted sum of mean daily temperatures within that month, where higher weights are given to temperatures that are more favorable for delta smelt spawning.

The following procedure was carried out for each of the five data sets in order to produce clean hourly and mean daily temperature values. A data frame containing every hour of every date was created with temperature values of `NA`, then temperature values from the raw data set were filled in. This was done to ensure that every date-hour combination was represented in the clean data set even if any combinations were missing from the raw data set. A visual inspection of all five time series indicated that temperatures outside of the interval [40°F, 85°F] were probably not realistic, so any values outside of this range were replaced with `NA`s. Most of these invalid temperatures were rather extreme, e.g., 2000, and appeared to be the result equipment malfunction. Next, empirical lower and upper bounds were calculated for a given date-time, and any temperatures falling outside the range defined by the bounds were replaced with `NA`s. This was done to detect and remove potentially problematic points that fell outside of the overall visual pattern of the data. Bounds were constructed by splitting the temperature data by day (within a year) and hour, e.g., January 1 at 12:00 pm, and calculating $\hat{\theta} \pm 2\hat{\sigma}_\theta$, where $\hat{\theta}$ and $\hat{\sigma}_\theta$ are the calculated mean and standard deviation (ignoring any missing temperature values). For a given day-time combination, this gives a rough 95% confidence interval calculated across years. This method is simple and systematic, but is also very crude and, based on a visual inspection, probably overestimated the number of problematic temperatures. We alternatively considered comparing individual temperature measurements with a moving average, but this method was not always able to detect points that we thought should have been detected and removed. The reason for this was that some potentially problematic temperatures occurred in sequences, leading to problematic moving average values. At this point, day-time specific mean temperatures were recalculated (to exclude problematic values) and these means used to impute all missing (`NA`) temperature values. Finally, the cleaned hourly data set was used to calculate mean daily temperatures in both °F and °C.

We found that the mean daily temperatures across the five sites were highly correlated (see Figure 2). Martinez is further west than spawning is likely to occur, but because the Martinez data go back further in time than the other four data sets, and because water temperatures at the five stations were so correlated, we chose to use Martinez temperatures for calculating the spawning water quality index. The index for a given year-month combination is calculated as a weighted sum of the mean daily temperatures in °C, where the weights range from 0 to 1 and reflect how favorable a temperature is for delta smelt spawning with higher weight indicating higher favorability. The weighting function, shown in the top panel of Figure 3, is

based on work by Wang (1986) suggesting that delta smelt spawn between 7 and 15°C, and on observations of aquaculture delta smelt spawning between 12 and 22°C (Lindberg et al. 1997; Bennett 2005). Spawning water quality indices were not calculated for incomplete year-months, i.e., year-month combinations that had missing days. In this case, the year-month was simply excluded from the final data set.

9 Utility Files and Functions

The R scripts described in this document use functions defined in the file `DataCleaner_Utility.r`.

References

J J Orsi, T E Bowman, D C Marelli, and A Hutchinson. 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, 1983.

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

W Kimmerer, C Penalva, S Bollens, S Avent, and J Cordell. *Interagency Ecological Program Newsletter*, 12 (2):16 – 21, 1999. URL `http://www.water.ca.gov/iep/products/newsletter.cfm`.

J C S Wang. Fishes of the sacramento-san joaquin estuary and adjacent waters, california: a guide to the early life histories. Technical report, IEP, California Dept. of Water Resources, Sacramento, CA, 1986.

J Lindberg, R Mage, B Bridges, and S Dorshov. Status of delta smelt culture. *Interagency Ecological Program Newsletter*, 10(Summer):21 – 22, 1997. URL `http://www.water.ca.gov/iep/products/newsletter.cfm`.

W A Bennett. Critical assessment of the delta smelt population in the San Francisco estuary, California. *San Francisco Estuary and Watershed Science*, 3(2), 2005. URL `http://www.escholarship.org/uc/item/0725n5vk`.

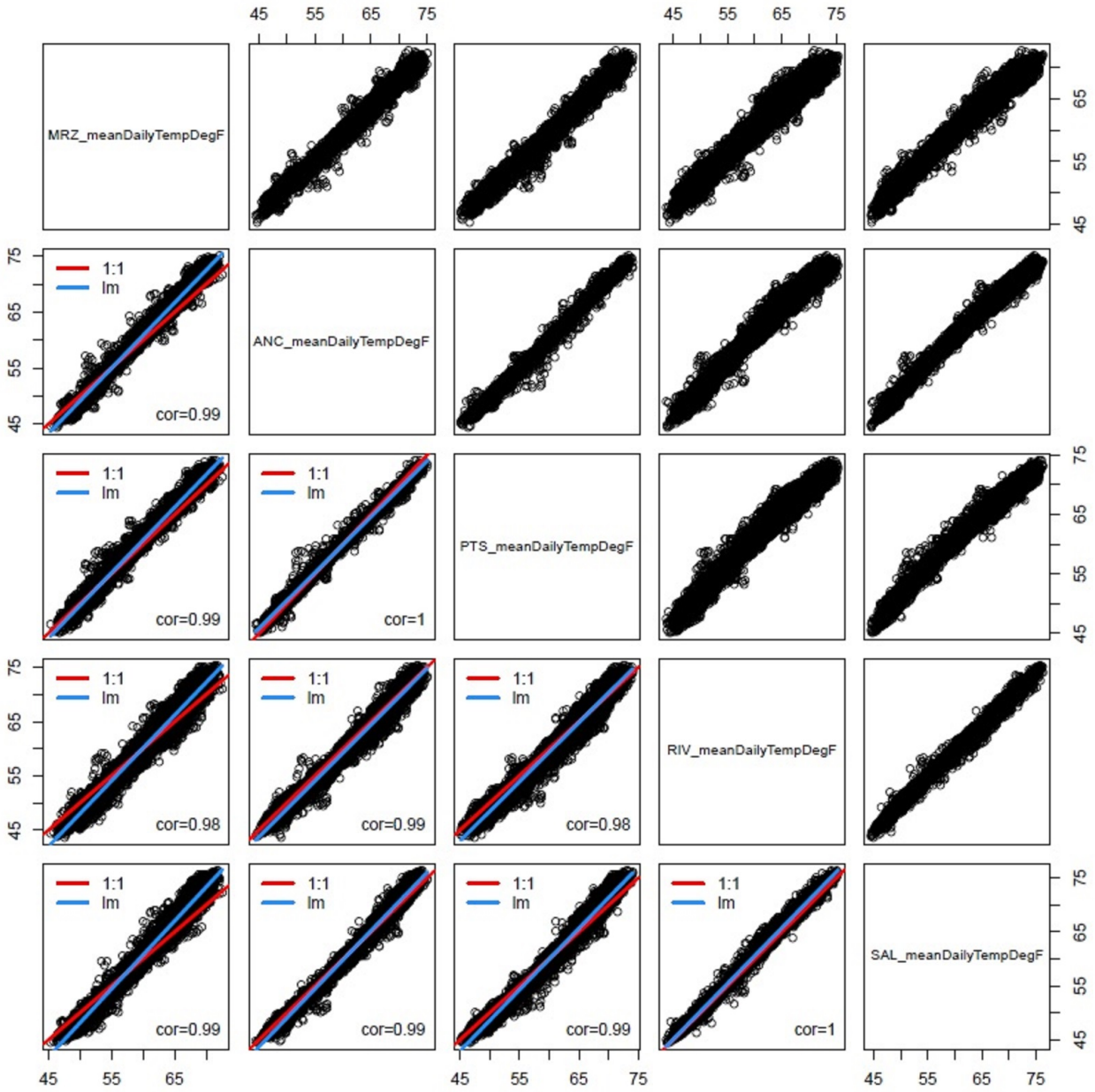

Figure 2: Correlation between average daily temperatures (°F) from five monitoring stations: MRZ, ANC, PTS, RIV, and SAL. "lm" stands for fitted linear model.

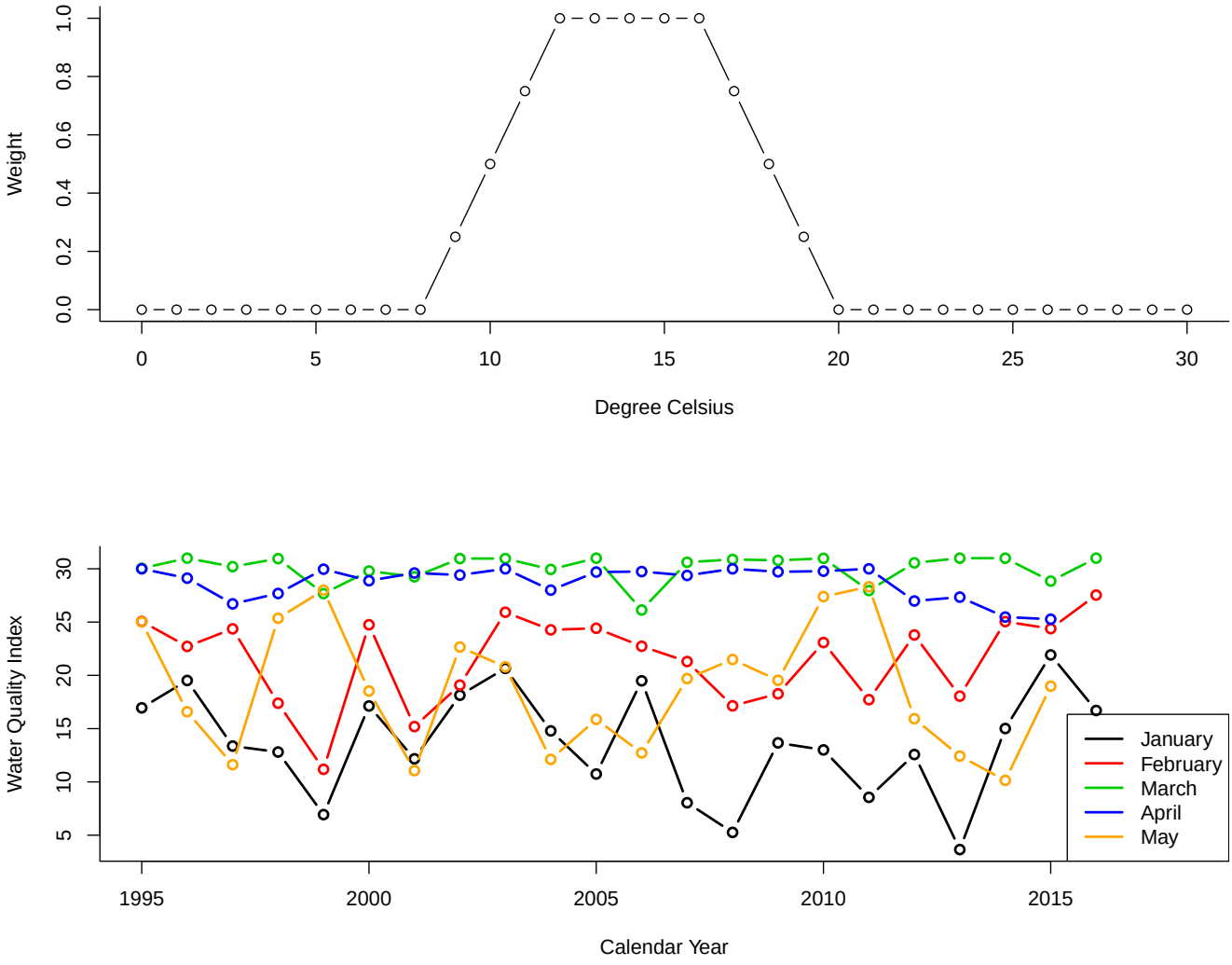

Figure 3: Spawning water quality index weighting function (top panel) and calculated index values by year and month (bottom panel).