Data preparation for analyses

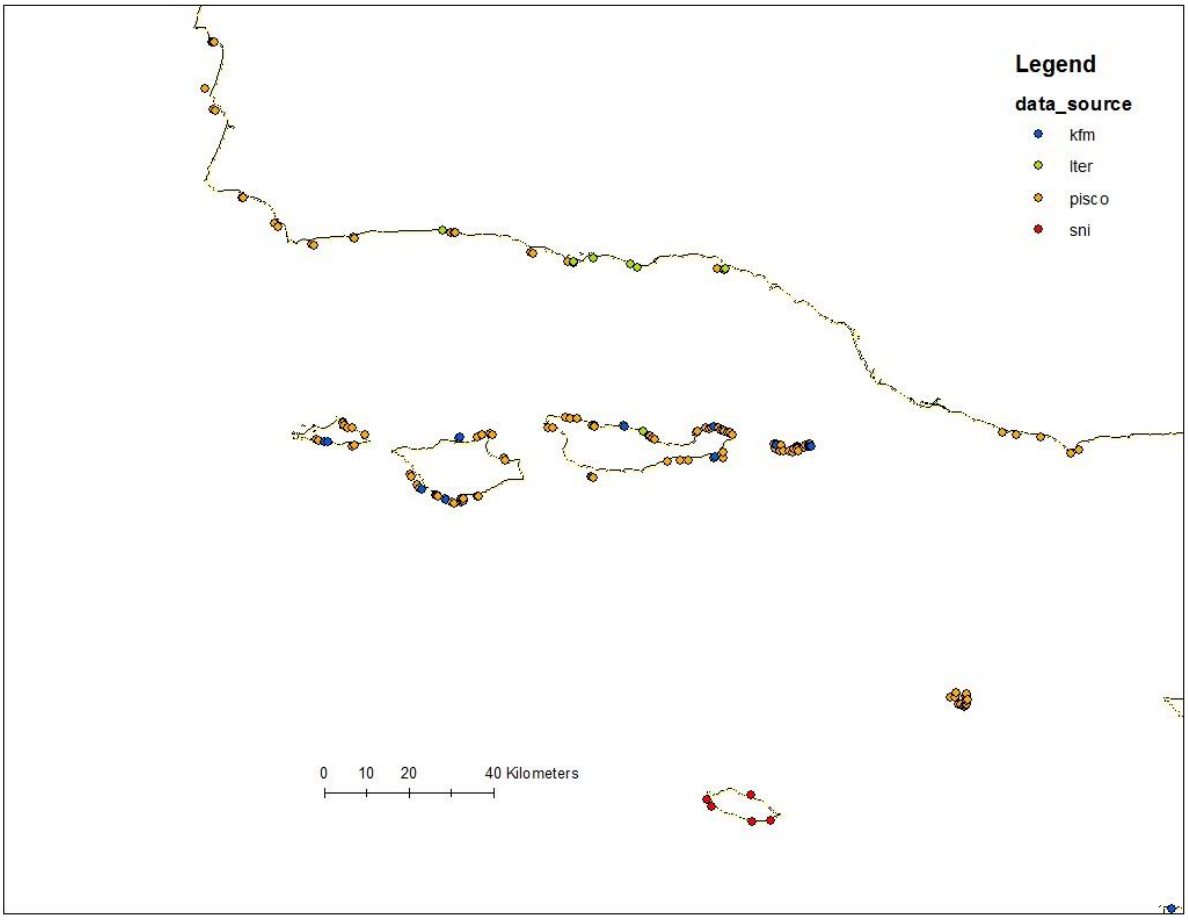
The four Santa Barbara Channel MBON kelp forest data sets required considerable processing to integrate due to their different sampling methodologies and locations. Analyses using integrated data from these different data sets should consider the notes listed below. The notes will be incorporated into the integrated datasets posted on the MBON website as part of the methods documentation.

The surveying methods, raw data, and metadata files for each project can be found online:

- Santa Barbara Coastal LTER (LTER): http://sbc.lternet.edu
- Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO): http://www.piscoweb.org/access-data
- Kelp forest monitoring (KFM): http://www.esapubs.org/archive/ecol/E094/245/
- San Nicolas Island monitoring (SNI): http://www.esapubs.org/archive/ecol/E094/244/

Overall:

1. The four programs survey different regions of the Santa Barbara Channel and outer waters: KFM is mainly located at the five Northern Channel Islands, PISCO has sites at both islands and mainland shore, LTER sites primarily occur along the mainland, and SNI occurs only at San Nicolas Island.



2. The survey transects were fixed in the KFM, SNI, and LTER. The PISCO transects are random.

3. The integrated dataset includes the same spatial-level identifiers for all four monitoring programs: site_id, subsite_id, transect_id, and replicate_id. However, the ids indicate different experimental units among programs because of the programs' different sampling designs. See Table 1 below:

Table 1. The spatial-level identifiers for each of the programs. The detail description for each of the surveys can refer to the links shown at the beginning of this document.

Survey	**Site_id**	**subsite_id**	**transect_id**	**replicate_id**
KFM fish (RDFC)	37 sites	Same as site	1 transect each site	1-8 replicate (divers); 2000 m^2 each for 30 mins

KFM fish (visual fish)	37 sites	Same as site	1 transect each site	4 replicates; 2(W) × 50 (L) m^2 each
KFM Swath/band	37 sites	Same as site	1 transect each site (same as fish)	12 replicates; 60 m^2 each
KFM quad	37 sites	Same as site	1 transect each site (same as fish)	6 (2 m^2) or 20 (10 m^2) replicates
KFM rpc	37 sites	Same as site	1 transect each site (same as fish)	15 replicates; 40 points
LTER fish	11 sites	Same as site	2-8 transects each site	1 replicate: 2 (W)× 40 (L) m^2.
LTER quad	11 sites	Same as site	2-8 transects each site (same as fish)	6 replicates: 1 m^2 each.
LTER swath	11 sites	Same as site	2-8 transects each site (same as fish)	4 replicates; 20 m^2 each
LTER upc	11 sites	Same as site	2-8 transects each site (same as fish)	1 replicate with 80 points in each transect
PISCO fish	68 sites	2-3 subsites (“side” in original datasheet)	2-4 transects within each subsite (“zone” in original datasheet)	3-4 replicates for each transect (“transect” in original, treated as replicates, random locations): 2 (W)× 30 (L) m^2.
PISCO quad	68 sites	2 subsites (“side” in original datasheet)	3 transects in each subsite (“zone” in original datasheet)	2 replicates for each transect (“transect” in original, treated as replicates, random locations, different from fish transects); 1 m^2
PISCO swath	68 sites	2 subsites (“side” in original datasheet)	3 transects in each subsite (“zone” in original datasheet)	2 replicates for each transect (“transect” in original, treated as replicates, random locations, different from fish transects); 60 m^2
PISCO upc	68 sites	2 subsites (“side” in	3 transects in each subsite (“zone” in	2 replicates for each transect (“transect” in

		original datasheet)	original datasheet)	original, treated as replicates, random locations, different from fish transects); 30-90 pts total
SNI fish	5 sites (super station in original datashe et)	1-2 subsites (station in original datasheet)	1 transect in each subsite	1 replicate: 2 (W)× 50 (L) m^2 for the bottom and 5 (W)× 50 (L) m^2 for the mid water.
SNI quad	5 sites	Same as site (stop survey in 2014)	1 transect in each subsite	5 replicates with 1 m^2.
SNI swath	5 sites	1-2 subsites	1 transect in each subsite	5 replicates with 20 m^2.
SNI upc	5 sites	1-2 subsites	1 transect in each subsite	10 replicates in 1m^2 area. 20 points

4. The number of transects per site differs among the different monitoring programs: KFM has 33 sites and 1 transect per site, PISCO 68 sites and ~24 transects for fish survey and ~16 for benthic survey, LTER 11 sites and 2-8 transects per site, and SNI 5 sites with 1-2 transects per site.

5. For all programs, the number of sites/subsites/transects changes over the course of the time series. *Recommendation*: This needs to be considered for analyses that requires a balanced sampling effort over time and space.

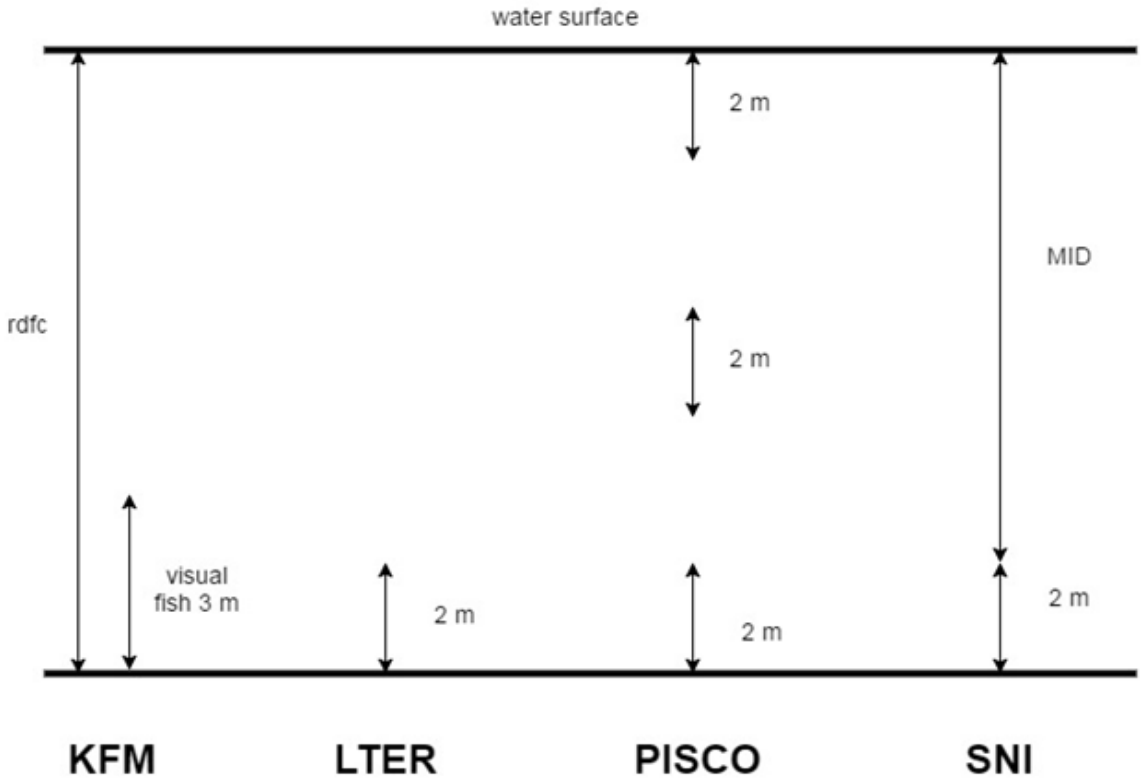
6. Survey timing differs among programs. PISCO's benthic surveys are early in the summer and fish later in the fall. LTER has all their surveys conducted between late July and early August. KFM surveys mainly in summer; SNI surveys in spring and a second survey later in the fall.

7. Transects were sampled multiple times a year in some surveys. For example, KFM surveyed several times a year in some years, and SNI surveyed twice a year in most years. In some cases a single PISCO survey was conducted over multiple days. *Recommendation*: Consider using only the survey that is closest in time to the mean sampling date of all surveys (mean sampling day of year = Day 275).

8. Organisms were recorded at different taxonomic levels (e.g. class, genus, or species) by the different monitoring programs, e.g., for quad and swath surveys, LTER has finer taxonomic resolution than PISCO. Thus, even though the same species may be observed, programs document them differently. *Recommendation*: Consider looking through the taxa report to identify taxa that are common to all programs. R code is included with the data package to filter taxa

Fish:

1. Fish are surveyed in different parts of the water column by the different monitoring programs (see graph below). This needs to be taken into account when fish data among the different monitoring programs.



2. PISCO fish transects can be located in a very shallow zone (< 3 m in depth) or deep zone (> 16 m in depth). Consider removing those extreme depth transects when making comparisons with other projects.

3. For some taxa, projects counted individuals according to population characteristics, e.g., age class, sex, life-history stage, size. In the integrated dataset, these distinctions within a species

were maintained by providing distinct local taxon ids, but all population subsets have the same authoritative ID (as auth_taxon_id, e.g., WORMs ID).

4. KFM uses multiple divers to collect data for the RDFC survey, and the replicate_id for the RDFC survey indicates a unique diver code. The number of divers varies, but for analyses that require balanced sampling can be standardized by selecting a fixed number of random divers (replicate_ids). KFM RDFC surveys are timed, covering 30 minute periods within a 2000 square meter area, and do not necessarily cover the entire area.

5. The visual fish survey in KFM has a closed list of 13 species, but has longer time series than LTER or PISCO (starting in early 1980's). The RDFC also began in the early 1980's, but only coded relative abundance (single, many, etc) was recorded until 2002, when counts began, allowing accurate density computation. For the species which abundance codes were used in the early years, average fish density for the species was estimated from recent data and the abundance code replaced. If the density from recent years are not available for the species, NA replaces the abundance code.

Quad and Swath:

1. Giant Kelp abundance is categorized in all programs, but the categories and definitions used (<1 m, > 1 m, holdfast, dead holdfast, w/o stipe count, or all of above) varied among the monitoring programs. The integrated dataset has one count for giant kelp, either stipe/frond or entire plant, depending on the survey methods. LTER counted the number of stipes/fronds per plant (the stipe total were included in the integrated data set). In the quad/swath survey, PISCO counted the *Macrocystis* plants which is included as "plant count" in the integrated dataset. PISCO can provide additional data for the stipe counts but not included in the integrated dataset. KFM counted plants but had additional stipe count survey (not included in the integrated dataset) could be requested. SNI counted plants.

2. LTER, KFM, and PISCO have sea urchin size frequency as a separate table (which was not retained in this integrated dataset).

5. KFM changes their quad/swath survey area over time. In addition, KFM moved their survey species from band transect to quad when the densities are high, e.g. *Crassedoma giganteus*, *Lytechinus anamesus*. Refer to the KFM survey protocols for more detail.

Percent cover:

1. The total number of points used to estimate percent cover varies over time for PISCO monitoring: 1999 was ~90 points, prior to 2006 was 60 points, and after 2007 was 30 points.

Recommendation: Consider dropping the 1999 PISCO point count survey. Similarly, KFM shifted total point counts over time: 20 points in 1982, 10 points in 1983, 50 points in 1984, and 40 points after 1984 and has been constant since.

2. In the UPC count, PISCO has superlayer points that are only recorded in the case that the point landed on *Laminaria farlowii*, brittle stars, *Pachythyone rubra*, or drift algae. They are marked as “percent_cover_superlayer” in the “unit_type” column in the integrated taxa csv file. Similarly, there are other layers called “relief” and “substrate” in the point count.
Recommendation: Researchers who are not interested in this information can remove the corresponding “taxon” from the dataset.

3. PISCO has its substrate survey in the quad survey. Other projects record substrate in the UPC/RPC survey.

4. The areas that form the basis of a UPC/RPC plot are different across projects, for example, PISCO uses 30-meter transect to set up 30 points and SNI uses 1 m^2 plot to set up 20 points.