popler: An R package for synthesis of population time series from long-term ecological research

Aldo Compagnoni*[a,], Andrew J. Bibian[a], Brad M. Ochocki[a], Sam Levin[b], Margaret O'Brien[c], Kai Zhu[d] and Tom E.X. Miller[a]

[a]Department of BioSciences, Program in Ecology and Evolutionary Biology, Rice University, Houston, TX USA

[b]Sam's affiliation

[c]Margaret's affiliation

[d]Kai's affiliation

Running headline: The popler database and R package

[Tom's comments appear in red italics.]

*email@aldo.edu

Abstract

[Very rough draft, will need to be improved.]

1. Population dynamics has a central role in the historical and current development of both fundamental and applied ecological science. The nascent culture of open data promises to increase the value of population dynamics studies to the field of ecology. However, users interested in data synthesis using population data are stymied by the heterogeneity of the available datasets, and by how difficult it is to identify relevant datasets.

2. To obviate these issues, we built popler, a relational database that can accommodate the vast majority of population data under a common structure, without the need for aggregating raw observations. Such common structure allows the user to more easily identify, compare, and manipulate datasets. The availability of raw data confers the largest possible freedom when devising scientific analyses.

3. We implemented popler as a PostgreSQL instance, where we stored population data originated by the United Stated Long Term Ecological Research (LTER) Network. Our focus on the US LTER data aims to leverage the untapped potential of this vast open data resource. However, we plan to extend this database to population data outside of the US LTER network.

4. To facilitate data accessibility we created an R package designed to explore and query the database. Our package allows users unfamiliar with the SQL language and with the structure of this database to easily identify, download, and analyze the datasets salient to their needs.

5. Finally, we developed an R package to explore and query this resource.

...

Keywords

Introduction

Population dynamics – changes in species' abundance and composition through time and space – are central to ecology for both applied and fundamental reasons. Population studies are a standard tool in conservation, particularly to perform risk assessment and viability analysis (Morris and Doak, 2002; Beissinger and McCullough, 2002). Population studies are also an integral part of fundamental ecological research. Populations are the building blocks of ecological dynamics at higher scales of organization, and examples abound showing how the study of population ecology improves understanding in evolution (Metcalf and Pavard, 2007), community ecology (Levine and HilleRisLambers, 2009), and ecosystem ecology (Medvigy et al., 2009; Fisher et al., 2018).

Given their central role in ecology, studies of population dynamics will be an essential component in the advances allowed by the flourishing culture of open access and data synthesis. The increase in freely available ecological data sets is poised to change ecological science (Laurance et al., 2016). The rising focus on open data is clear in changing publishing standards, in the design of observational networks like the National Ecological Observatory Network (Schimel et al., 2007), and in the online availability of previously proprietary data (Kratz et al., 2003; Bechtold et al., 2005). This deluge of open data holds tremendous promise to facilitate comparative analyses and to test the generality of ecological hypotheses. For population dynamics in particular, it is the increasing availability of long-term data that will likely yield the most substantial scientific advances, as long time series are required to detect trends in abundance (Lindenmayer et al., 2012), quantify temporal variance (Compagnoni et al., 2016), and identify endogenous (Knape and de Valpine, 2012) or exogenous (Knape and de Valpine, 2011; Hampton et al., 2013) drivers of population fluctuations.

There is ample evidence demonstrating the importance of long-term data for population ecology to facilitate comparative and synthetic studies (Lindenmayer et al., 2012; Giron-Nava et al., 2017). To our knowledge, there is currently just one publicly accessible database focused on long-term population dynamics: the Global Population Dynamics Database (GPDD, Inchausti and Halley, 2001). The GPDD provides over 5000 time series of population size longer than 10 years for over 1800 animal species. This database has been powerfully leveraged for comparative analyses and syntheses (e.g., Knape and de Valpine, 2012) but it has some important limitations. GPDD time series are not spatially replicated – there is one observation of population size or density temporal replicate, with no estimate of uncertainty – making it difficult or impossible to isolate different sources of variability and requiring the assumption that population size is known without error. Retaining information on spatial replication would expand the variety of research questions that can be asked and improve estimation of uncertainty in the answers. Additionally, the GPDD focuses exclusively on single species dynamics, making it difficult or impossible to link the dynamics of multiple fluctuating populations within ecological communities (e.g., Ushio et al., 2018).

One of the best sources of publicly available long-term observations of population abundance of diverse plants and animals is the Long-Term Ecological Research (LTER) network supported by the U.S. National Science Foundation. The LTER was founded in 1980 and grew from the original six sites to the current 28 sites throughout North America plus one each in Puerto Rico and Antarctica. Synthetic and comparative studies from the LTER network have made valuable contributions to ecological understanding (Knapp et al., 2012). However, the majority of LTER synthesis research has focused on ecological dynamics at the community (e.g. Wilcox et al. (2017)) and ecosystem (e.g. Knapp and Smith (2001)) scales. Nevertheless, every LTER site collects population abundance data as one of its five core areas of continuous observations (Callahan, 1984). These population time series include both single- and multi-species studies. In our opinion, these data, which have been accumulating since 1980, are under-used. One issue that may limit the use of LTER population data in synthetic, comparative studies is their heterogeneity. Indeed, several authors have noticed that progress in ecology is hindered by difficulties in the way data is accessed, by heterogeneities in the way data is structured, and by failure to cite data originators *[Jones et al. 2004, Reichman et al. 2011]*. The structure of LTER data sets may be widely different, employing a variety of data types (counts of individuals, biomass estimates, percent cover, etc.), experimental designs driven by the priorities of particular PIs, and diverse replication schemes – idiosyncrasies that may be difficult to accommodate in a one-size-fits-all database. However, these challenges also present valuable opportunities. For example, the hierarchical replication structure of many LTER studies (e.g., subplots within plots within transects) can facilitate more sophisticated statistical investigation than would be possible with simpler, aggregated, or unreplicated data.

To overcome the issues posed by heterogeneous data structures, we developed popler (POPulation dynamics in Long-term Ecological Research), an online database of LTER population studies. We also developed a companion R package to aid in discovery, querying, and synthesis. The popler database defines a common data structure to accommodate nearly every long-term population data set provided by the LTER network as of this writing. The common data structure facilitates the identification, access, and manipulation of raw population data through a user-friendly R package. Our goals here are to provide introductions to the popler database and R package, which we believe will be useful resources in population, community, and macro ecology. Our focus here is on LTER time series but our database schema can in principle, accommodate any population level dataset and expanding popler beyond the LTER network is a priority for future development.

The **popler** database

popler aims to organize population data from the LTER network using a common structure. To achieve this, we identified a set of variables relevant to population studies (Table 1) and organized them into a relational database (Fig. 1). In popler, we store "raw" data, warts and all; we have not modified or aggregated the original observations. Our goal was to organize data under a common structure. If needed, this structure facilitates downstream aggregation and subsetting of the original data.

Population data

Time series of population size are the heart of the popler database. We defined 'population data' as time-series of observations on the size or density of a population of a species or other taxonomic unit. Observations of population size are stored in a variable called abundance_observation and can be measured as a count, biomass, density, or cover. Counts and biomass are absolute measures, density is a measure of population size taken with respect to a one-, two-, or three-dimensional unit of observation, and cover is a relative (e.g., percentage) or absolute (e.g., area) measure. These four types of population data are stored in the homonymous tables of the database (Fig. 1A).

The population datasets contained in popler are always replicated temporally. Temporal replicates are identified with up to three variables: `year`, `month`, and `day`. In theory, population data can be collected at a smaller temporal resolution (e.g. hourly). However, we did not encounter LTER datasets whose sampling was more frequent than daily.

Population data are also almost always spatially replicated. Hence, they provide measures of population abundance at multiple locations. Furthermore, spatial replicates are often nested, where for example a study might be replicated at separate sites (e.g. a watershed, a lake, an island, etc.), each of which contains intermediate spatial replicates (e.g. a transect, a block), which in turn contain the smallest spatial replicate at which observations are made (e.g. a plot, a quadrat). The hypothetical study described above would have three nested levels of spatial replication, identified by three numbered `spatial replication` variables. In `popler`, we accommodate data sets with up to five spatial replication levels (Table 1). For simplicity, we call the first, and therefore largest spatial replicate "study site" (Fig. 1C) – which is not the LTER site, which is identified elsewhere in the study's metadata (Fig. 1D), but rather a location within it.

`popler` contains both observational and experimental studies. Experimental datasets contained information on one or more experimental treatments. Popler accommodates information on up to three experimental treatments, identified by three numbered `treatment type` variables (Table 1).

Most datasets contain one or more variables in addition to the ones described above which we store in a list of variable called `covariates`. Covariates might report, for example, the hour of day or the temperature when population data was collected, the Cartesian coordinates of spatial replicates, and so on. Covariates can be useful for time series that contain information on population structure *[Would this be indicated in the metadata? How would someone search for such studies or know if they are working with one?]*. In these datasets, population size refers to subdivisions of the entire population. These subdivisions can be, for example, males and females, large and small individuals, developmental different stages, and so forth. However, we found a handful of examples of such dataset.

Finally, in addition to time series of abundance, `popler` contains datasets that provide information on the states or attributes of the individuals, or a subset thereof, that make up a population. We store this information in a dedicated table ("Individual'"", Fig. 1A). As individual attributes we consider variables that describe identity, size, sex, life stage (e.g. instar versus adult in insects) or status (e.g. reproductive or non-reproductive, live or dead, etc.). We refer to these individual attributes with the term "structure": `popler` accommodates data sets that measure up to four types of structure simultaneously. We store these data in up to four numbered `structure_type` variables. While these data are not population time series, we chose to include them in `popler` because they provide information on demographic transitions that can be used to derive estimates of population growth. Moreover, in the cases of datasets that sample all of the individuals in a population, individuals can be aggregated (i.e. summed) as a measure of population size. *[Are 'structure' and 'individual' operationally the same thing? Would there ever be individual data that is not somehow 'structured'? Seems like no, especially if identity is treated as structure. I am wondering if we can simplify this by just calling this individual-level data, and saying that this will always be accompanied by one or more individual attributes. We introduce the concept of structure above, as a possible covariates for time series. Here 'structure' has a related but different interpretation. I would like to prevent confusion.]*

Taxonomic information

Each observation of abundance or individual state corresponds to a taxonomic unit (Fig. 1B). Population data typically refer to a species or a genus, but we also include data that refer to a higher taxonomic rank, such as family, or order. Popler provides 15 taxonomic ranks, and two additional variables that refer to how taxonomic information is recorded in the original datasets. The additional variables are `sppcode`, which are taxon-specific alphanumeric codes, and `common_name`, the common name of each taxonomic unit (Table S1).

Popler stores the taxonomic information linked to each study in two tables: one containing the original taxonomic information reported by the PI, the other containing the accepted taxonomic information derived from the former (Fig. 1B). Raw taxonomic data typically contains ambiguities derived by the dynamic changes in species classifications (Chamberlain and Szöcs, 2013). The raw data also typically fail to include higher-level taxonomic information above the genus level. To provide as much taxonomic information as possible, popler provides a second table linking taxonomic units provided by the authors to accepted taxonomic units according to the algorithms provided by the R package taxize (Chamberlain and Szöcs, 2013). This package links taxon names to the unambiguous entry of an online taxonomic database.

Study site

We stored the locations of datasets by recording the latitude (lat_study_site) and longitude (lng_study_site) of study sites (Fig. 1C). Storing this information in a separate table allows for explicit connections between independent data sets collected at the same locations within LTER sites. We emphasize that "study sites" refer to the first and therefore largest level of nested spatial replication. These study sites are at a smaller scale than "LTER sites" (the 28 NSF-supported locations; Table A1).

Metadata

The metadata table describes 48 variables (Table S2) with general information about the dataset, its temporal and spatial replication, and its study design (Fig. 1D).*[It would be helpful to clarify how these 48 variables relate to the variables in Table 1.]* The general information includes a title designated by the original authors, link to online metadata, contact information for data originators, and the type of data provided by the dataset (i.e., which of the five tables in Fig. 1A the data is stored in).

The metadata also contains variables on temporal and spatial replication. These are derived variables that aim to facilitate the user in the search of salient datasets. To quantify temporal replication we provide duration_years, the years elapsed between the first and last observation, and the sampling frequency. We report the most prevalent sampling frequency, because in many datasets the sampling frequency is not prefectly constant. For example, in yearly sampling designs, years of sampling are often skipped. We quantify spatial replication with the number of levels of nested spatial replicates, and with the number of replicates for each spatially nested level.

Finally, we provide basic information on the design of the study that originated the dataset.

The variable `community` shows whether studies focus on a single species or on multiple species. Studies can also be observational or experimental; if a study is experimental, we provide information on the type of treatments imposed by the study. Finally, while the objective of popler is to store raw population data, some publicly available datasets provide aggregated data. In this case, we consider these data as derived.

Creation and contents of the `popler` database

We implemented `popler` as an instance of PostgreSQL 9.2.18. We selected the candidate datasets online, pre-processed the associated files, and uploaded them in the database through a graphic user interface we developed in Python. We selected the candidate datasets by scraping information from the internet sites associated with each one of the 26 LTER sites (Table A1). We scraped this data using libraries scrapy in Python, and rvest and RSelenium in R.

We included or excluded studies based on several criteria. First, we only included studies that reported repeated observations of populations or individuals through time. Second, we included studies with at least five years of data. We performed this selection in early 2017, so datasets that are not currently stored in `popler` might now meet this criterion. We did not require population censuses to be consecutive (some studies have an irregular sampling frequency). Third, we selected only datasets for which the observations had corresponding taxonomic information, which implies that we ignored time series data on species richness or on the abundance of functional types.

Before uploading to the online database, we preprocessed datasets in python and R. We often had to combine the separate files associated with the same dataset, transform datasets from "wide" to "long" form, convert non-ASCII characters, and handling ambiguous study sites. We provide the details of this pre-processing in Appendix S1.

The `popler` database resulting from this process contains data from 265 studies (118 of which are experimental) representing 3436 cumulative years of observations with a mean study duration of 12.97 years. `popler` contains data from 691 plant species, 349 animal species, and 1 fungal species.

The **popler** package

The `popler` R package consists of three core functions that allow users to browse and retrieve data from the database (Fig. 2). In order of intended use, these functions are: `dictionary()`, `browse()`, and `get_data()`. This order of use revolves around a summary table made up of the variables contained in the metadata (Fig. Fig. 1D) and taxonomic (Fig. Fig. 1B) tables of the online database. Each row of the summary table corresponds to a dataset and each column corresponds to one of the metadata variables that describe it. The `dictionary()` function explains what these variables (or columns) are, and it shows their content. This content provides the information needed to subset and visualize metadata through the `browse()` function, which returns studies (or rows) that satify the subset criteria. Finally, following exploration of the metadata, raw data can be downloaded with the `get_data()` function. Below we describe these functions and illustrate their use in greater detail.

The **`dictionary()`** function

The dictionary function is a good place for new users to begin working with `popler` (Fig. 2). With no arguments provided, this function returns a subset of the most useful metadata variables associated with each dataset (Fig. 1):

```
##              variable                                          description
## 1              title                             title of project
## 2  proj_metadata_key                            unique project id
## 3             lterid                                    lter name
## 4           datatype type of abundance data (e.g. count,biomass)
## 5          studytype         experimental or observational study?
## 6     duration_years                 duration of project in years
## 7          community         does data set contain multiple taxa?
## 8          structure                types of indidivual structure
## 9          treatment                           types of treatment
## 10          lat_lter                           lter site latitude
## 11          lng_lter                          lter site longitude
## 12           species         specific epithet of a taxonomic unit
## 13           kingdom                                      kingdom
## 14            phylum                                       phylum
## 15             class                                        class
## 16             order                                        order
## 17            family                                       family
## 18             genus                                        genus
```

Setting argument `full_tbl` to `TRUE` returns all 76 metadata variables. Each one of these variables name can be provided as an argument to `dictionary()`, which will then return the possible unique values of the variable. For example, `dictionary(lterid)` returns the three letter codes for all sites in the LTER network, `dictionary(genus)` returns all the genera represented in the database, etc. This output provides the starting point for developing criteria with which to identify relevant datasets. *[Note for Aldo: I thought we decided that dictionary(variable) would return the unique levels of factors and quantiles of continuous variables. This is not how it is working. dictionary(duration_years) returns unique levels of years. Also, the output shows studies less than 5 years, including zero years and one NA. This is not consistent with the criteria stated in the text. This will need some attention. Lastly, we should probably cut 'Chromalveolata'* *from the kingdom data.]*

The **browse()** function

Once the user is familiar with the meaning and content of the variables that define `popler` datasets, they are ready to dig deeper using `browse()` (Fig. 2). Running `browse()` without arguments provides the metadata from all of the datasets currently stored in popler. This will be a $265 * 19$ data frame, with each row corresponding to a study and each column corresponding to a variable defined by `dictionary()`.

The full strength of `browse()` is achieved by subsetting studies according to desired criteria using standard logical expressions. For example, the user might want to consider only studies whose duration is 30 years or greater, which can be subsetted with:

```
LTER_30 <- browse( duration_years > 29)
```

This operation will create the object `LTER_30`, which provides metadata for the data sets that satisfy the specified criterion. Multiple criteria may be combined. For example, 30+ year studies of plants can be browsed with

```
LTER_30_plants <- browse( duration_years > 29 & kingdom == "Plantae")
```

It is at the `browse()` stage that users should vet the data sets that meet their criteria. To facilitate data exploration, `browse()` output can be printed in a more readable settings by providing `report = TRUE` as an argument, which opens up a formatted html document (Fig. 4). The metadata provided by `browse()` not only contains information on the characteristics of a study but also information on how to cite the study, its unique identifiers, including doi, Knowledge Network for Biocomplexity (`http://knb.ecoinformatics.org`) catalog system identifier, the contact information of study PIs, and a hyperlink to the url providing the original data.

The **get_data()** function

[Random observation: the study titles in popler are sometimes in quotes, sometimes not. This should be cleaned up. Also I would like to edit the default message of getdata() before we publish.]

Once data sets of interest have been identified, `get_data()` downloads the data from a server that hosts the database. This function can take as its first argument a `browse` object, a logical expression, either, or both. For example,

```
LTER_30_dat <- get_data(LTER_30)
```

downloads the raw data from the studies described in the browse object `LTER_30` and

```
BNZ_30_dat <- get_data(LTER_30,lterid == "BNZ")
```

downloads the subset of data sets in `LTER_30` from the Bonanza Creek LTER site. The three letter code "BNZ" could have been located running `dictionary(lterid)` (we show the meaning of three letters codes in Table S3). The data downloaded from `popler` are in "long" form, meaning that each row of data reports a single measure of population size, and separate variables indicate the temporal and spatial replicate, taxonomic information, etc. This format makes it easy to further subset downloaded datasets with the aim of visualization and analysis. For example, to visualize the yearly population size of a single species across time and its spatial replicates, we can subset the study, the species, and flag its spatial replicates with a different color. We provide such an example using a 33-year long data set collected at the Bonanza Creek LTER. We select a dataset that measures the population size of plants as counts of observed individuals. We highlight one example species, *Viburnum edulis* *[Aldo wrote: ???]* and color code the highest level of spatial replication (the study sites of Fig. 1B). The following code draws the plot shown in Fig. ??.

```
plot(abundance_observation ~ year,
     col = as.factor(spatial_replication_level_1),
     data = subset(BNZ_30_dat, proj_metadata_key == 195 & sppcode == 'VIBUEDUL') )
```

This plot shows three idiosynchrasies of this particular dataset. First, because this study includes three nested spatial replication levels, each site contains several spatial replicates and therefore multiple points with the same color. Second, the temporal resolution of this dataset is not consistent. Observations were collected at unequal intervals and, while the first and last observation are 33 years apart, the sampling was concentrated between 1983 and 2003, with only one or two sites sampled before and after. *[It would be nice to take this example a little further, drawing on the online study description - the type of digging one would have to do to really understand the data. For example, the abundance observation is a count, but per what sampling unit? Are these counts in a plot, counts on a transect, etc? We can't tell just from popler but we can probably get that information by reading the study design. It would be good to add something to this effect.]*

Ancillary functions

`popler` also provides three additional functions to open the url of the original dataset, unpack covariates, and provide a citation for each dataset. First, we encourage users to consult the online study description associated with each dataset before starting their analyses. As described above, we have not modified the original datasets ; rather, we have rearranged their structure and added accepted taxonomies when possible. The function `metadata_url()` takes as its argument data objects produced by either the `browse()` or `get_data()` functions and launches a web browser. Second, as described above, `popler` stores all covariates associated with each observation but does not deliver these by default. To access covariates, the `cov_unpack()` function takes an object created by `get_data()` and extracts an R data frame with rows corresponding to abundance observaions and columns corresponding to any covariates provided by the original PI. To interpret these covariates, users should read the documentation of the respective dataset using `metadata_url()`. Third we strongly encourage users of `popler` data to cite data originators and we provide the function `popler_citation()` to do so. The argument of this function is, again, a data object produced by either `browse()` or `get_data()`.

Limitations and opportunities for development

[I did not do too much to this section. I think it needs to be stronger but we can continue to polish this. Probably OK for sending to coauthors.] Working with raw, spatially replicated, and non-aggregated data provides key advantages in quantitative analyses of population dynamics, and these advantages were a driving force behind the development of `popler`. However, because we did not aggregate data, the user needs examine individual datasets and the associated online

study descriptions to understand their peculiarities. This is important for several reasons. First, some studies do not define the difference between a missing value and an abundance observation of zero. Second, many datasets have gaps or changes in the design during the length of the study. Third, the covariates variable can hold key information which is best understood when consulting the online documentation of the original dataset.

In the future, there are opportunities to increase the size of `popler` and expand its scope. First, many of the studies included in `popler` are ongoing, so there will be opportunities to update these entries in the future with new observations. Second, because our schema (Fig. 1) is very general, the database could be expanded to include population datasets outside of the LTER network. Third, it would be valuable to explicitly associate `popler`'s population-level data with environmental drivers, especially climate. It is our intention and hope that the resources provided by `popler` will advance ecological understanding of population dynamics within the LTER network, and more generally.

Acknowledgements

Support for database and package development was provided by the National Science Foundation (DEB-1543651). *[We should probably acknowledge NSF support of LTER, as well as coauthor-specific acknowledgements.]*

References

W. A. Bechtold, P. L. Patterson, et al. *The enhanced forest inventory and analysis program: national sampling design and estimation procedures*, volume 80. US Department of Agriculture Forest Service, Southern Research Station Asheville, North Carolina, 2005.

S. R. Beissinger and D. R. McCullough. *Population viability analysis.* University of Chicago Press, 2002.

J. T. Callahan. Long-term ecological research. *BioScience*, 34(6):363–367, 1984.

S. A. Chamberlain and E. Szöcs. taxize: taxonomic search and retrieval in r. *F1000Research*, 2, 2013.

A. Compagnoni, A. J. Bibian, B. M. Ochocki, H. S. Rogers, E. L. Schultz, M. E. Sneck, B. D. Elderd, A. M. Iler, D. W. Inouye, H. Jacquemyn, and T. E. X. Miller. The effect of demographic correlations on the stochastic population dynamics of perennial plants. 86:480–494, 2016. ISSN 0012-9615. doi: 10.1002/ecm.1228.

R. A. Fisher, C. D. Koven, W. R. Anderegg, B. O. Christoffersen, M. C. Dietze, C. E. Farrior, J. A. Holm, G. C. Hurtt, R. G. Knox, P. J. Lawrence, et al. Vegetation demographics in earth system models: A review of progress and priorities. *Global change biology*, 24(1):35–54, 2018.

A. Giron-Nava, C. C. James, A. F. Johnson, D. Dannecker, B. Kolody, A. Lee, M. Nagarkar, G. M. Pao, H. Ye, D. G. Johns, et al. Quantitative argument for long-term ecological monitoring. *Marine Ecology Progress Series*, 572:269–274, 2017.

S. E. Hampton, E. E. Holmes, L. P. Scheef, M. D. Scheuerell, S. L. Katz, D. E. Pendleton, and E. J. Ward. Quantifying effects of abiotic and biotic drivers on community dynamics with multivariate autoregressive (mar) models. *Ecology*, 94(12):2663–2669, 2013.

P. Inchausti and J. Halley. Investigating long-term ecological variability using the global population dynamics database. *Science*, 293(5530):655–657, 2001.

J. Knape and P. de Valpine. Effects of weather and climate on the dynamics of animal population time series. *Proceedings of the Royal Society of London B: Biological Sciences*, 278(1708):985–992, 2011.

J. Knape and P. de Valpine. Are patterns of density dependence in the global population dynamics database driven by uncertainty about population abundance? *Ecology letters*, 15(1): 17–23, 2012.

A. K. Knapp and M. D. Smith. Variation among biomes in temporal dynamics of aboveground primary production. *Science*, 291(5503):481–484, 2001.

A. K. Knapp, M. D. Smith, S. E. Hobbie, S. L. Collins, T. J. Fahey, G. J. Hansen, D. A. Landis, K. J. La Pierre, J. M. Melillo, T. R. Seastedt, et al. Past, present, and future roles of long-term experiments in the lter network. *BioScience*, 62(4):377–389, 2012.

T. K. Kratz, L. A. Deegan, M. E. Harmon, and W. K. Lauenroth. Ecological variability in space and time: Insights gained from the us lter program. *AIBS Bulletin*, 53(1):57–67, 2003.

W. F. Laurance, F. Achard, S. Peedell, and S. Schmitt. Big data, big opportunities. *Frontiers in Ecology and the Environment*, 14(7):347–347, 2016.

J. M. Levine and J. HilleRisLambers. The importance of niches for the maintenance of species diversity. *Nature*, 461(7261):254, 2009.

D. B. Lindenmayer, G. E. Likens, A. Andersen, D. Bowman, C. M. Bull, E. Burns, C. R. Dickman, A. A. Hoffmann, D. A. Keith, M. J. Liddell, et al. Value of long-term ecological studies. *Austral Ecology*, 37(7):745–757, 2012.

D. Medvigy, S. Wofsy, J. Munger, D. Hollinger, and P. Moorcroft. Mechanistic scaling of ecosystem function and dynamics in space and time: Ecosystem demography model version 2. *Journal of Geophysical Research: Biogeosciences*, 114(G1), 2009.

C. J. E. Metcalf and S. Pavard. Why evolutionary biologists should be demographers. *Trends in Ecology & Evolution*, 22(4):205–212, 2007.

W. Morris and D. Doak. Quantitative conservation biology; theory and practice in conservation biology. *Sinauer, Sunderland, Massachusetts, USA*, 2002.

D. Schimel, W. Hargrove, F. Hoffman, and J. MacMahon. Neon: A hierarchically designed national ecological network. *Frontiers in Ecology and the Environment*, 5(2):59–59, 2007.

M. Ushio, C.-H. Hsieh, R. Masuda, E. R. Deyle, H. Ye, C.-W. Chang, G. Sugihara, and M. Kondoh. Fluctuating interaction network and time-varying stability of a natural fish community. *Nature*, 554:360–363, Feb. 2018. ISSN 1476-4687. doi: 10.1038/nature25504.

K. R. Wilcox, A. T. Tredennick, S. E. Koerner, E. Grman, L. M. Hallett, M. L. Avolio, K. J. La Pierre, G. R. Houseman, F. Isbell, D. S. Johnson, J. M. Alatalo, A. H. Baldwin, E. W.

Bork, E. H. Boughton, W. D. Bowman, A. J. Britton, J. F. Cahill, S. L. Collins, G. Du, A. Eskelinen, L. Gough, A. Jentsch, C. Kern, K. Klanderud, A. K. Knapp, J. Kreyling, Y. Luo, J. R. McLaren, P. Megonigal, V. Onipchenko, J. Prevéy, J. N. Price, C. H. Robinson, O. E. Sala, M. D. Smith, N. A. Soudzilovskaia, L. Souza, D. Tilman, S. R. White, Z. Xu, L. Yahdjian, Q. Yu, P. Zhang, and Y. Zhang. Asynchrony among local communities stabilises ecosystem function of metacommunities. *Ecology letters*, 20:1534–1545, Dec. 2017. ISSN 1461-0248. doi: 10.1111/ele.12861.

Table 1: Variables used to store population or individual data in `popler`.

Variable	Description
`abundance_observation`	Measure of population abundance at a specific time and location. This variable measures abundance as a count, biomass, density, or cover. For individual data sets this variable is always equal to 1, because each attribute or set of attributes refer to a single individual.
`day`	Day of observation
`month`	Month of observation
`year`	Year of observation
`spatial_replicate_n`	The n^{th} level of spatial replication, where `spatial_replicate_1` is the study site. `popler` accommodates up to five levels of spatial replication.
`treatment_type_n`	For datasets originating from an experimental study, the n^{th} treatment. `popler` popler accommodates up to three treatments.
`covariates`	Ancillary observations that do not fall into the standard schema of `popler`.
`structure_type_n`	For individual data, these variables measure the n^{th} attribute of individuals (identity, size, sex, status, stage). `popler` accommodates up to four structure types per dataset.

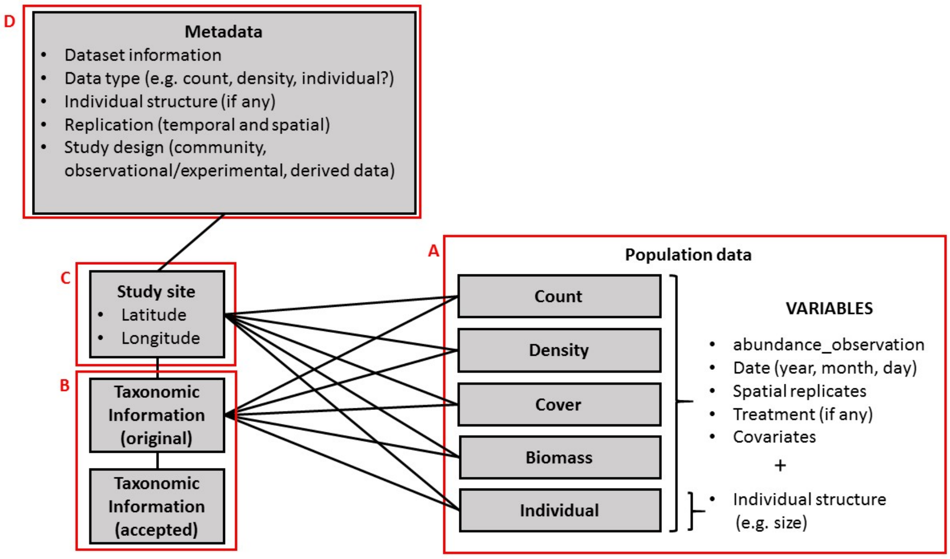

D
Metadata
Dataset information
Data type (e.g. count, density, individual?)
Individual structure (if any)
Replication (temporal and spatial)
Study design (community, observational/experimental, derived data)
C
Study site
Latitude
Longitude
B
Taxonomic Information (original)
Taxonomic Information (accepted)
A
Population data
Count
Density
Cover
Biomass
Individual
VARIABLES
abundance_observation
Date (year, month, day)
Spatial replicates
Treatment (if any)
Covariates
+
Individual structure (e.g. size)

Figure 1: Schematic representation of the entity relationship diagram of the `popler` database. `popler` provides metadata on the studies that originated abundance data points (D). This metadata contains information on the unique identifiers of each study, on its design (observational or experimental), temporal and spatial replication. Popler stores the latitude and longitude of the study site (C). Each abundance data point corresponds to a specific taxonomic unit (B). Finally, the time series population data collected in a study can be of five different types: count, density, biomass, cover, and individual data (A).

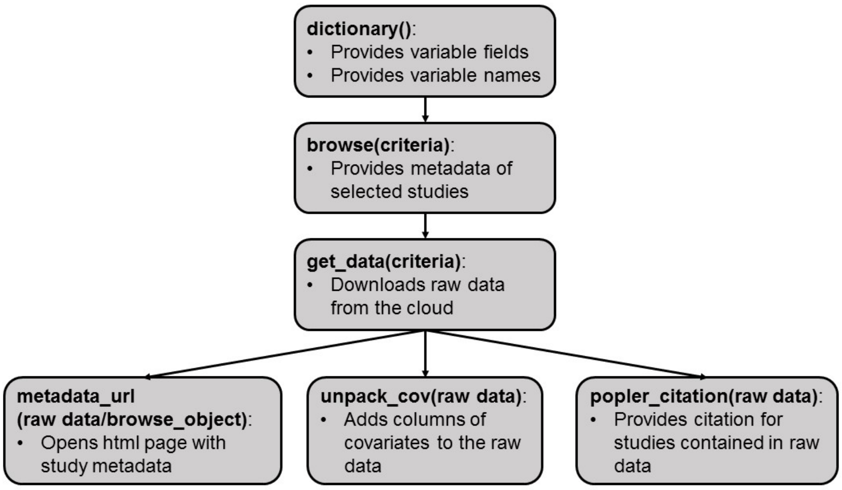

dictionary():
• Provides variable fields
• Provides variable names
browse(criteria):
• Provides metadata of selected studies
get_data(criteria):
• Downloads raw data from the cloud
metadata_url (raw data/browse_object):
• Opens html page with study metadata
unpack_cov(raw data):
• Adds columns of covariates to the raw data
popler_citation(raw data):
• Provides citation for studies contained in raw data

Figure 2: Suggested workflow when using the `popler` R package to interface with the homonymous online database. The function `dictionary()` refers to the variables of the metadata that describe the data sets contained in `popler`. `dictionary()` describes these variables and returns their possible values. This information advises which criteria to use when subsetting `popler` The user can provide a criterion (that is, a logical statement) to subset the metadata, using `browse()`, and download dataset from the cloud using `get_data()`. Moreover, the output of `get_data()` (a data frame) can be the argument of three ancillary functions: `metadata_url()` opens the webpage containing the original dataset and their associated online metadata. `unpack_cov()` can be used to format the covariates contained in a raw data object into separate columns of a data frame. Finally, `popler_citation()` provides a citation for the downloaded data set(s).

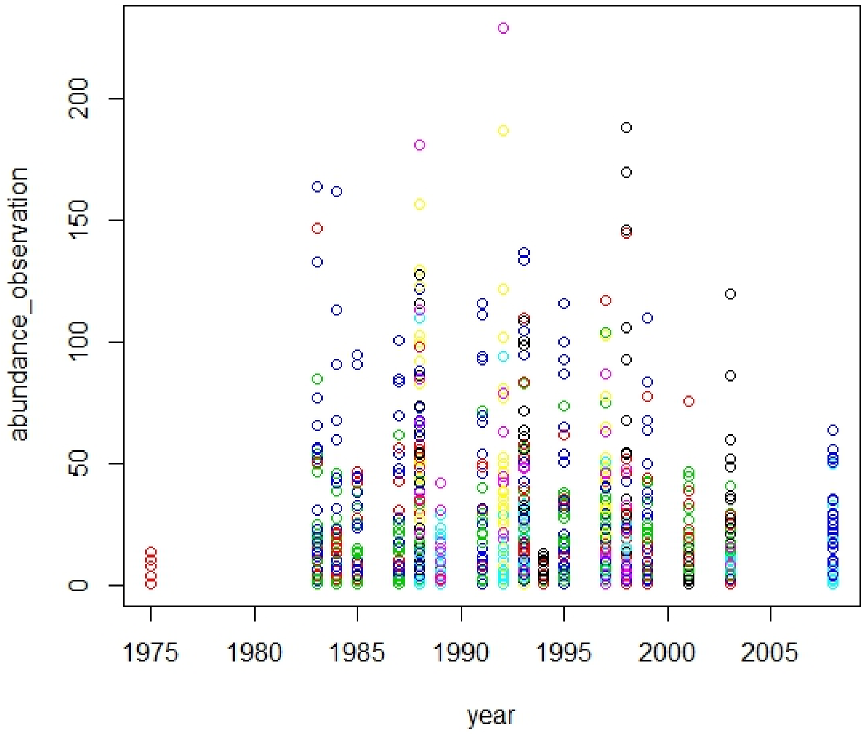
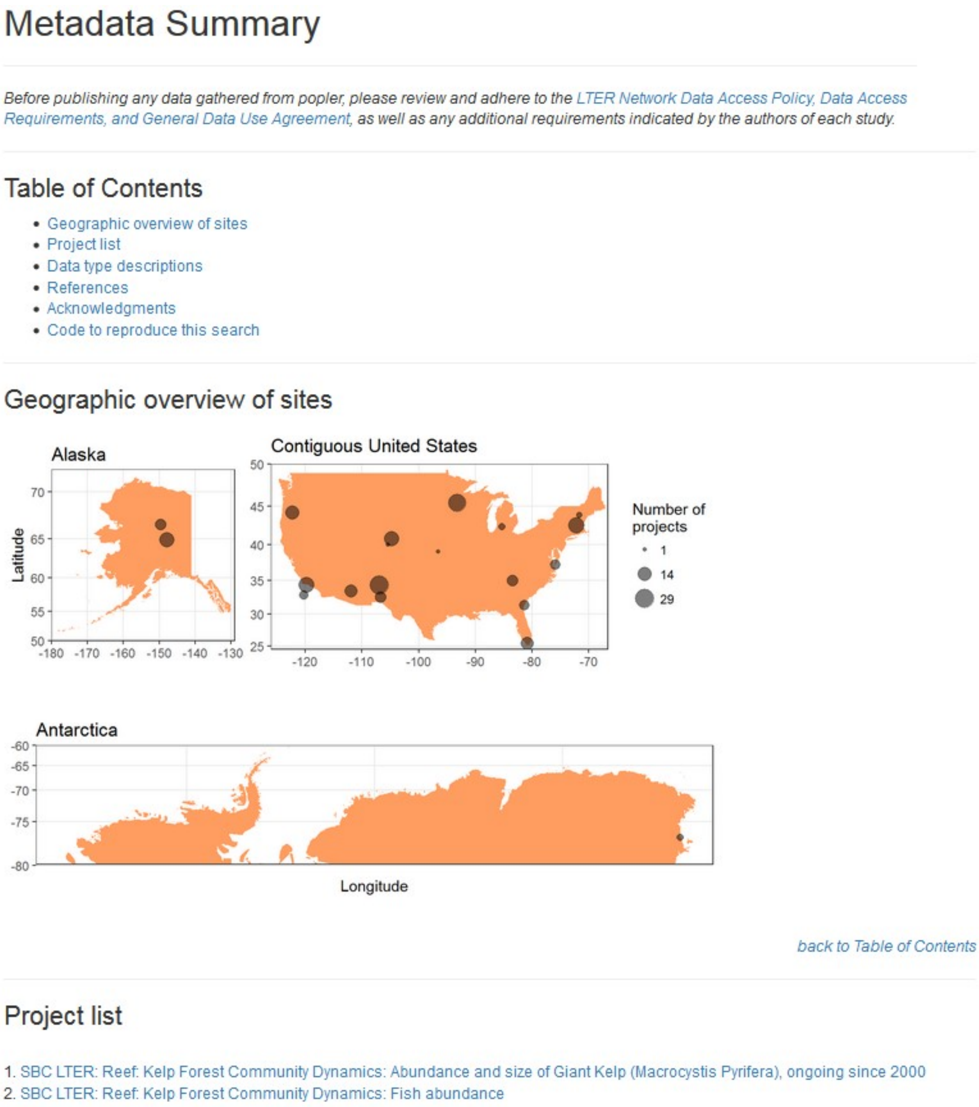


Figure 3: Time series of *Viburnum edulis* abundance counts at Bonanza Creek LTER. The abundance observation is a count and colors correspond to multiple study sites, the largest scale of spatial replication in `popler`. There are additional, smaller scales of spatial replication in this study, and hence multiple points per site per year.



Metadata Summary

Before publishing any data gathered from popler, please review and adhere to the LTER Network Data Access Policy, Data Access Requirements, and General Data Use Agreement, as well as any additional requirements indicated by the authors of each study.

Table of Contents

Geographic overview of sites

back to Table of Contents

Project list

1. SBC LTER: Reef: Kelp Forest Community Dynamics: Abundance and size of Giant Kelp (Macrocystis Pyrifera), ongoing since 2000
2. SBC LTER: Reef: Kelp Forest Community Dynamics: Fish abundance

Figure 4: The html output of function `browse()` when argument report = TRUE.

Appendix S1: Pre-processing **popler** data

Before uploading datasets into the online popler database, we combined datasets, transformed datasets from wide to long form, converted non-ASCII characters, and modified ambiguous study site names.

The variables of many datasets were contained in two or more separate files, which we combined in a single file. When the original dataset provided data in wide form, we transformed it into long form. In wide form datasets, abundance data associated with different species was stored in separate columns. popler stores these datasets in long form, whereby each row of abundance data is related to a specific taxonomic unit in the table containing taxonomic information (Fig. 1B). We converted all data in ASCII format, because the encoding of the database is the UTF-8. We often re-defined study site names to unambiguously associate them with one of the 26 LTER sites. Many site names are alphanumeric codes (e.g. “U1”) which can overlap across several LTER sites. Hence, we changed site names following a standard formula (namely, from “U1” to “site_sbc_U1”, where “sbc” refers to the Santa Barbara coastal LTER site).

In a handful of cases, we removed single data rows from the original dataset. These data rows were associated with two types of typos in the original dataset. First, some abundance observations were not associated with a time of observation. We removed this data because popler can only accommodate population information associated with a time of observation. Second, a handful of abundance data points were clear typos (e.g. the letter “l” instead of a numeric value). We substituted these data points with a missing value (NULL in the database). We uploaded these pre-processed datasets in the popler database through a Graphic User Interface developed in Python using libraries panda and pyqt5.

Table S1: Taxonomic variables contained in the popler table on original taxonomic information.

Variable
sppcode
kingdom
subkingdom
infrakingdom
superdivision
division
subdivision
superphylum
phylum
subphylum
class
subclass
order
family
genus
species
common_name

Table S2: Metadata variables used to describe the datasets stored in popler.

Variable	Description
proj_metadat_key	Unique ID
lter_project_key	ID of LTER site
lter_project_key	ID of LTER site
title	Title of study
samplingunits	Unit of measure (if any) referred to population data.
datatype	Data type: count, biomass, cover, density, and individual. These correspond to the tables in Fig. 1A.
structured_type_n	If individual data, this shows what type of structure is stored. A study can contain up to $n = 4$ types of structure.
structured_type_n_units	Unit of measure (if any) referred to structure data.
studystartyr	Start year of the study
studyendyr	End year of the study
duration_years	Duration of the study in years
samplefreq	Frequency of population census
studytype	Whether study is observational or experimental
community	Whether study includes single taxon (community = F) or multiple taxa (community = T)
spatial_replication_level_n_extent	Extent of spatial replication level number n. A dataset can have up to to 5 replication levels.
spatial_replication_level_n_unit_extent	Unit of spatial extent of the n spatial replication level.
spatial_replication_level_n_label	Label of the spatial replication level (e.g. transect, plot, quadrat, ect.). The label of spatial replication level 1 is “site”.
spatial_replication_level_n_number_of_unique_reps	The number of unique replicates for the nth level of spatial replication.
treatment_type_n	The type of treatment. *[This is vague and does not correspond to level n. PROBLEM.]*
control_group	If study is experimental, this shows the field(s) that identify the control replicate.
derived	Is population size data raw, or is it derived (e.g. it is aggregated)?
authors	Author(s) of the original dataset
authors_contact	Email address(es) of the author(s) associated with the original dataset.
metalink	url of the original dataset
knbid	Knowledge Network for Biocomplexity identifier.

Table S3: LTER identification acronyms and their meaning as used in the popler database.

Variable	LTER name
SBC	Santa Barbara Coastal LTER
SEV	Sevilleta LTER
SGS	Shortgrass Steppe
VCR	Virginia Coastal Reserve LTER
AND	Andrew Forest LTER
NWT	Niwot Ridge LTER
BNZ	Bonanaza Creek LTER
CDR	Cedar Creek Ecosystem Science Reserve
GCE	Georgia Coastal Ecosystems LTER
ARC	Arctic LTER
CAP	Central Arizon - Phoneix LTER
FCE	Florida Coastal Everglades LTER
HFR	Harvard Forest LTER
KBS	Kellogg Biological Station LTER
CWT	Coweeta LTER
HBR	Hubbard Brook LTER
MCM	McMurdo Dry Valleys LTER
JRN	Jornada Basin LTER
CCE	California Current Ecosystem LTER
KNZ	Konza Prairie LTER