

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

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

^aDepartment of BioSciences, Program in Ecology and Evolutionary Biology, Rice
University, 6100 Main St, MS-170, Houston, TX 77005

^bInstitute of Biology, Martin Luther University Halle-Wittenberg, Am Kirchtor
1, 06108 Halle (Saale), Germany

^cGerman Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig,
Deutscher Platz 5e, 04103 Leipzig, Germany

^dMarine Science Institute, University of California, Santa Barbara, CA 93016,
United States

^eDepartment of Environmental Studies, University of California, Santa Cruz, CA
95064, USA

Running headline: The popler database and R package

[Tom's comments appear in red italics.] [Aldo's comments appear in blue italics.]

^{*}aldo.compagnoni@gmail.com

Abstract

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 difficulty in identifying relevant datasets, and by the heterogeneity of the available data.
2. To obviate these issues, we built a relational database, `popler`, and its R client library (`"popler"`) package. `popler` accommodates the vast majority of population data under a common structure, and without the need for aggregating raw observations. `library("popler")` is designed for users unfamiliar with the SQL language and with the structure of the database. This R library allows users to identify, download, explore, and cite datasets salient to their needs.
3. We implemented `popler` as a PostgreSQL instance, where we stored population data originated by the United States 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. The database currently contains 274 datasets from 25 LTER sites.
4. The combination of the online database and the R library (`"popler"`) is a resource to data synthesis efforts in population ecology. The common structure of `popler` simplifies comparative analyses, and the availability of raw data confers the largest freedom when exploiting datasets. `library("popler")` maximizes these opportunities by providing a user-friendly interface to the online database.

Keywords

- 1 open long-term population data, US Long Term Ecological Research Network data, online
- 2 database, database structure, PostgreSQL, R package, data synthesis, comparative analysis

3 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., Knappe 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

62 with simpler, aggregated, or unreplicated data.

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

74 **The `popler` database**

75 `popler` aims to organize population data from the LTER network using a common structure.
76 To achieve this, we identified a set of variables relevant to population studies (Table 1) and
77 organized them into a relational database (Fig. 1). In `popler`, we store “raw” data, meaning
78 that we have not modified or aggregated the original observations. Our goal was to organize data
79 under a common structure which, if needed, facilitates downstream aggregation and subsetting
80 of the original data.

81 **Population data**

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

89 database (Fig. 1A).

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

94 Population data are also almost always spatially replicated. Hence, they provide measures
95 of population abundance at multiple locations. Furthermore, spatial replicates are often nested,
96 where for example a study might be replicated at separate sites (e.g. a watershed, a lake, an
97 island, etc.), each of which contains intermediate spatial replicates (e.g. a transect, a block),
98 which in turn contain the smallest spatial replicate at which observations are made (e.g. a plot,
99 a quadrat). The hypothetical study described above would have three nested levels of spatial
100 replication, identified by three numbered `spatial_replication` variables. In `popler`, we
101 accommodate data sets with up to five spatial replication levels (Table 1). For simplicity, we call
102 the first, and therefore largest spatial replicate “study site” (Fig. 1C). Note that this does not
103 refer to the LTER site, one of the 28 NSF-supported locations (Table S3).

104 `popler` contains both observational and experimental studies. Experimental datasets con-
105 tained information on one or more experimental treatments. `Popler` accommodates information
106 on up to three experimental treatments, identified by three numbered `treatment_type` vari-
107 ables (Table 1).

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

116 Finally, in addition to time series of abundance, `popler` contains individual-level data. This
117 data provide information on the attributes of the individuals, or a subset thereof, that make
118 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). 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.

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 entries of online taxonomic databases.

144 Study site

145 We stored the locations of datasets by recording the latitude (`lat_study_site`) and longitude
146 (`lng_study_site`) of study sites (Fig. 1C). Storing this information in a separate table allows
147 for explicit connections between independent data sets collected at the same locations within
148 LTER sites. We emphasize that “study sites” refer to the first and therefore largest level of
149 nested spatial replication. We reiterate that these study sites are at a smaller scale than “LTER
150 sites” (the 28 NSF-supported locations; Table S3).

151 Metadata

152 The metadata table describes 48 variables (Table S2) with general information about the dataset,
153 its temporal and spatial replication, and its study design (Fig. 1D). The general information
154 includes a title designated by the original authors, link to online metadata, contact information
155 for data originators, and the type of data provided by the dataset (i.e., which of the five tables
156 in Fig. 1A the data is stored in).

157 All remaining metadata is related to the variables stored in the tables of 1A and 1B. These
158 metadata provide information on temporal replication, spatial replication, and study design.
159 First, we provide the years elapsed between the first and last observation (`duration_years`),
160 and the sampling frequency (`samplefreq`). We report the prevalent sampling frequency, be-
161 cause in most datasets sampling changes during the length of the study. These metadata relate
162 to the variables `day`, `month`, `year` of the original datasets (Table 1). Second, we quantify spa-
163 tial replication providing the number of levels of nested spatial replicates, and with the number
164 of replicates for each spatially nested level. These metadata relate to `spatial_replicate_n`
165 of the original datasets. Third, we provide basic information on the design of the study that
166 originated each dataset. We show whether studies focus on a single species or on multiple
167 species through `community`. This variable indicates whether the taxonomic tables (1B) con-
168 tain more than one taxa. Studies can also be observational or experimental (`studytype`). If a
169 study is experimental, we provide information on the type of treatments imposed by the study
170 (`treatment_type_n`) and, when available, which is the control treatment (`control_group`).
171 The (`studytype`) variable refers to variables `treatment_type_n` in the original datasets. Fi-

nally, when abundance data stored in the `abundance_observation` variable is aggregated across space or time, rather than raw, we consider these data as derived (`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 28 LTER sites (Table S3). 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 274 studies (118 of which are experimental) representing 3613 cumulative years of observations with a mean study duration of 13.19 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: `pplr_dictionary()`,

199 `pplr_browse()`, and `pplr_get_data()`. This order of use revolves around a summary ta-
200 ble made up of the variables contained in the metadata (Fig. Fig. 1D) and taxonomic (Fig.
201 Fig. 1B) tables of the online database. Each row of the summary table corresponds to a
202 dataset and each column corresponds to one of the metadata variables that describe it. The
203 `pplr_dictionary()` function explains what these variables (or columns) are, and it shows
204 their content. This content provides the information needed to subset and visualize metadata
205 through the `pplr_browse()` function, which returns studies (or rows) that satisfy the sub-
206 set criteria. Finally, following exploration of the metadata, raw data can be downloaded with
207 the `pplr_get_data()` function. Below we describe these functions and illustrate their use in
208 greater detail.

209 **The `pplr_dictionary()` function**

210 The dictionary function is a good place for new users to begin working with popler (Fig. 2).
211 With no arguments provided, this function returns a subset of the most useful metadata variables
212 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 individual 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

213 Providing argument `full_tbl = TRUE` returns all 76 metadata variables. Each one of these
 214 variables name can be provided as an argument to `pplr_dictionary()`, which then returns the
 215 possible unique values of the variable. For example, `pplr_dictionary(lterid)` returns the
 216 three letter codes of the LTER network sites included in `popler`. For numeric variables such as
 217 `duration_years`, `pplr_dictionary()` returns a summary including quantiles, mean, and
 218 median. This output provides the starting point for developing criteria with which to identify
 219 relevant datasets.

220 The `pplr_browse()` function

221 Once the user is familiar with the meaning and content of the variables that define popler
222 datasets, they are ready to dig deeper using `pplr_browse()` (Fig. 2). Running `pplr_browse()`
223 without arguments provides the metadata from all of the datasets currently stored in popler.
224 This will be a 274×19 data frame, with each row corresponding to a study and each column
225 corresponding to a variable defined by `pplr_dictionary()`.

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

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

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

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

232 It is at the `pplr_browse()` stage that users should vet the data sets that meet their
233 criteria. To facilitate data exploration, `pplr_browse()` output can be printed in a more
234 readable settings by providing `report = TRUE` as an argument, which opens up a formatted
235 html document (Fig. 3). The metadata provided by `pplr_browse()` not only contains in-
236 formation on the characteristics of a study but also information on how to cite the study, its
237 unique identifiers, including digital object identifier (DOI), Knowledge Network for Biocomplex-
238 ity (<http://knb.ecoinformatics.org>) catalog system identifier, and the contact informa-
239 tion of study PIs. In a few instances when we could not obtain the DOI of a dataset, we provided
240 the universal resource locator (URL) associated with the original data.

241 The `pplr_get_data()` function

242 Once data sets of interest have been identified, `pplr_get_data()` downloads the data from a
243 server that hosts the database. This function can take as its first argument a browse object, a

244 logical expression, either or both. For example,

```
LTER_30_dat <- pplr_get_data(LTER_30)
```

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

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

246 downloads the subset of data sets in LTER_30 from the Bonanza Creek LTER site. The
247 three letter code “BNZ” could have been located running `dictionary(lterid)`. The data
248 downloaded from `pplr` are in “long” form, meaning that each row of data reports a single
249 measure of population size, and separate variables indicate the temporal and spatial replicate,
250 taxonomic information, etc. This format makes it easy to further subset downloaded datasets
251 with the aim of visualization and analysis. For example, to visualize the yearly population size
252 of a single species across time and its spatial replicates, we can subset the study, the species, and
253 flag its spatial replicates with a different color. We provide such an example using a 33-year long
254 data set collected at the Bonanza Creek LTER. We select a dataset that measures the population
255 size of plants as counts of observed individuals. We highlight one example species, *Viburnum*
256 *edule* and color code the highest level of spatial replication (the study sites of Fig. 1B). The
257 following code draws the plot shown in Fig. 4.

```
plot(abundance_observation ~ jitter(year), pch = 16,  
     xlab = "Year (jittered)",  
     col = as.factor(spatial_replication_level_1),  
     data = subset(BNZ_30_dat, proj_metadata_key == 195 &  
                   sppcode == 'VIBUEDUL') )
```

258 This plot shows three idiosyncrasies of this particular dataset. First, because this study
259 includes three nested spatial replication levels, each site contains several spatial replicates and
260 therefore multiple points with the same color. Second, the temporal resolution of this dataset
261 is not consistent. Observations were collected at unequal intervals and, while the first and last
262 observation are 33 years apart, the sampling was concentrated between 1983 and 2003, with only
263 one or two sites sampled before and after. This visualization invites a closer look to understand

264 the temporal replication and spatial replication this dataset. We can directly visualize the
265 temporal replication of this study plotting how many years of data are available per site:

```
ggplot(data = BNZ_30_dat,  
       aes(x = year, y = spatial_replication_level_1) ) +  
geom_point(size = 5) +  
theme_bw() + xlab("Year") + ylab("Site")
```

266 The resulting plot (Fig. 5) shows that just a few sites were sampled annually. To further
267 understand the sampling design, exploring the BNZ_30_dat object shows that there are two
268 replication levels: SITE and plot. Accessing the online metadata using pplr_metadata_url
269 shows that each site contains 20 plots of 4 square meters of surface.

270 Ancillary functions

271 popler also provides three additional functions to open the url of the original dataset, unpack
272 covariates, and provide a citation for each dataset. First, we encourage users to consult the on-
273 line study description associated with each dataset before starting their analyses. As described
274 above, we have not modified the original datasets ; rather, we have rearranged their structure
275 and added accepted taxonomies when possible. The function metadata_url() takes as its
276 argument data objects produced by either the pplr_browse() or pplr_get_data() func-
277 tions and launches a web browser. Second, as described above, popler stores all covariates
278 associated with each observation but does not deliver these by default. To access covariates,
279 the cov_unpack() function takes an object created by pplr_get_data() and extracts an R
280 data frame with rows corresponding to abundance observations and columns corresponding to any
281 covariates provided by the original PI. To interpret these covariates, users should read the docu-
282 mentation of the respective dataset using metadata_url(). Third we strongly encourage users
283 of popler data to cite data originators and we provide the function pplr_citation() to do
284 so. The argument of this function is, again, a data object produced by either pplr_browse()
285 or pplr_get_data().

286 Limitations and opportunities for development

287 Working with raw, spatially replicated, and non-aggregated data provides key advantages in
288 quantitative analyses of population dynamics, and these advantages were a driving force behind
289 the development of `popler`. However, because we did not aggregate data, the user needs examine
290 individual datasets and the associated online study descriptions to understand their peculiarities.
291 This is important for several reasons. First, some studies do not define the difference between
292 a missing value and an abundance observation of zero. Second, many datasets have gaps or
293 changes in the design during the length of the study. Third, the covariates variable can hold key
294 information which is best understood when consulting the online documentation of the original
295 dataset.

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

304 Acknowledgements

305 Support for database and package development was provided by the National Science Foundation
306 (DEB-1543651). This research was supported by a Julian Huxley Faculty Fellowship from Rice
307 University and a Faculty Research Grant awarded by the Committee on Research from the
308 University of California, Santa Cruz (KZ). *[We should probably acknowledge NSF support of*
309 *LTER, as well as coauthor-specific acknowledgements.]*

310 Authors' contributions

311 AC, AB, KZ, MO, TEXTM designed and built the database. AC AB, KZ, BD, SM, and TEXTM
312 designed and built the R package. AC and TEXTM led the writing of the manuscript. All authors
313 contributed critically to the drafts and gave final approval for publication.

314 References

- 315 W. A. Bechtold, P. L. Patterson, et al. *The enhanced forest inventory and analysis program:*
316 *national sampling design and estimation procedures*, volume 80. US Department of Agriculture
317 Forest Service, Southern Research Station Asheville, North Carolina, 2005.
- 318 S. R. Beissinger and D. R. McCullough. *Population viability analysis*. University of Chicago
319 Press, 2002.
- 320 J. T. Callahan. Long-term ecological research. *BioScience*, 34(6):363–367, 1984.
- 321 S. A. Chamberlain and E. Szöcs. taxize: taxonomic search and retrieval in r. *F1000Research*, 2,
322 2013.
- 323 A. Compagnoni, A. J. Bibian, B. M. Ochocki, H. S. Rogers, E. L. Schultz, M. E. Sneek, B. D.
324 Elder, A. M. Iler, D. W. Inouye, H. Jacquemyn, and T. E. X. Miller. The effect of demographic
325 correlations on the stochastic population dynamics of perennial plants. 86:480–494, 2016. ISSN
326 0012-9615. doi: 10.1002/ecm.1228.
- 327 R. A. Fisher, C. D. Koven, W. R. Anderegg, B. O. Christoffersen, M. C. Dietze, C. E. Farrior,
328 J. A. Holm, G. C. Hurtt, R. G. Knox, P. J. Lawrence, et al. Vegetation demographics in earth
329 system models: A review of progress and priorities. *Global change biology*, 24(1):35–54, 2018.
- 330 A. Giron-Nava, C. C. James, A. F. Johnson, D. Dannecker, B. Kolody, A. Lee, M. Nagarkar, G. M.
331 Pao, H. Ye, D. G. Johns, et al. Quantitative argument for long-term ecological monitoring.
332 *Marine Ecology Progress Series*, 572:269–274, 2017.
- 333 S. E. Hampton, E. E. Holmes, L. P. Scheef, M. D. Scheuerell, S. L. Katz, D. E. Pendleton, and

334 E. J. Ward. Quantifying effects of abiotic and biotic drivers on community dynamics with
335 multivariate autoregressive (mar) models. *Ecology*, 94(12):2663–2669, 2013.

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

338 J. Knappe and P. de Valpine. Effects of weather and climate on the dynamics of animal population
339 time series. *Proceedings of the Royal Society of London B: Biological Sciences*, 278(1708):985–
340 992, 2011.

341 J. Knappe and P. de Valpine. Are patterns of density dependence in the global population dy-
342 namics database driven by uncertainty about population abundance? *Ecology letters*, 15(1):
343 17–23, 2012.

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

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

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

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

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

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

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

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

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

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

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

370 K. R. Wilcox, A. T. Tredennick, S. E. Koerner, E. Grman, L. M. Hallett, M. L. Avolio, K. J.
371 La Pierre, G. R. Houseman, F. Isbell, D. S. Johnson, J. M. Alatalo, A. H. Baldwin, E. W.
372 Bork, E. H. Boughton, W. D. Bowman, A. J. Britton, J. F. Cahill, S. L. Collins, G. Du,
373 A. Eskelinen, L. Gough, A. Jentsch, C. Kern, K. Klanderud, A. K. Knapp, J. Kreyling, Y. Luo,
374 J. R. McLaren, P. Megonigal, V. Onipchenko, J. Prevéy, J. N. Price, C. H. Robinson, O. E.
375 Sala, M. D. Smith, N. A. Soudzilovskaia, L. Souza, D. Tilman, S. R. White, Z. Xu, L. Yahdjian,
376 Q. Yu, P. Zhang, and Y. Zhang. Asynchrony among local communities stabilises ecosystem
377 function of metacommunities. *Ecology letters*, 20:1534–1545, Dec. 2017. ISSN 1461-0248. doi:
378 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.

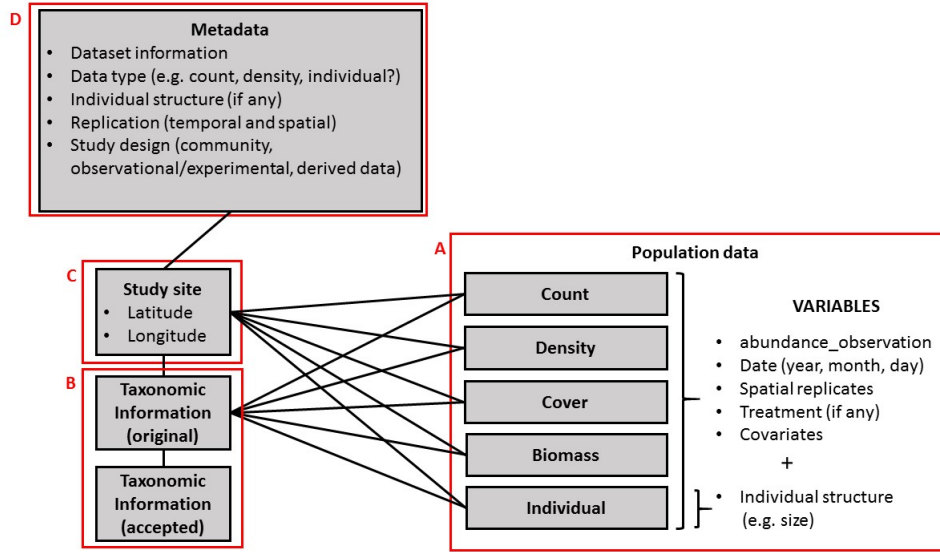


Figure 1: Schematic representation of the entity relationship diagram of the `popler` database. `popler` provides metadata on the studies that originated abundance data points (D). This metadata contains information on the unique identifiers of each study, on its design (observational or experimental), 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).

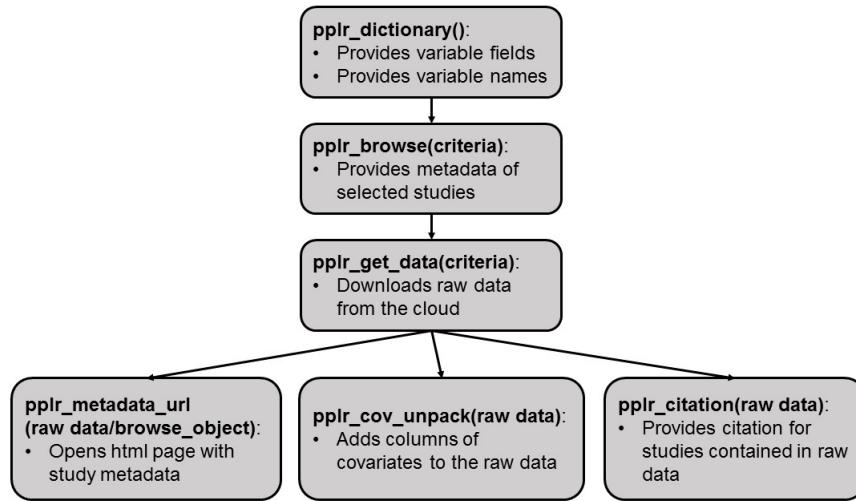


Figure 2: Suggested workflow when using the `popler` R package to interface with the homonymous online database. The function `pplr_dictionary()` refers to the variables of the metadata that describe the data sets contained in `popler`. `pplr_dictionary()` describes these variables and returns their 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 `pplr_browse()`, and download dataset from the cloud using `pplr_get_data()`. Moreover, the output of `pplr_get_data()` (a data frame) can be the argument of three ancillary functions: `pplr_metadata_url()` opens the webpage containing the original dataset and their associated online metadata. `pplr_cov_unpack()` can be used to format the covariates contained in a raw data object into separate columns of a data frame. Finally, `pplr_citation()` provides a citation for the downloaded data set(s).

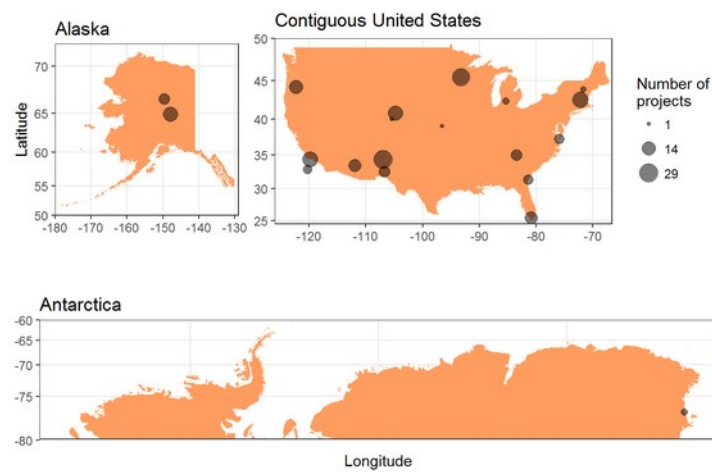
Metadata Summary

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](#)
- [Project list](#)
- [Data type descriptions](#)
- [References](#)
- [Acknowledgments](#)
- [Code to reproduce this search](#)

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 3: The html output of function `pplr_browse()` when argument `report = TRUE`.

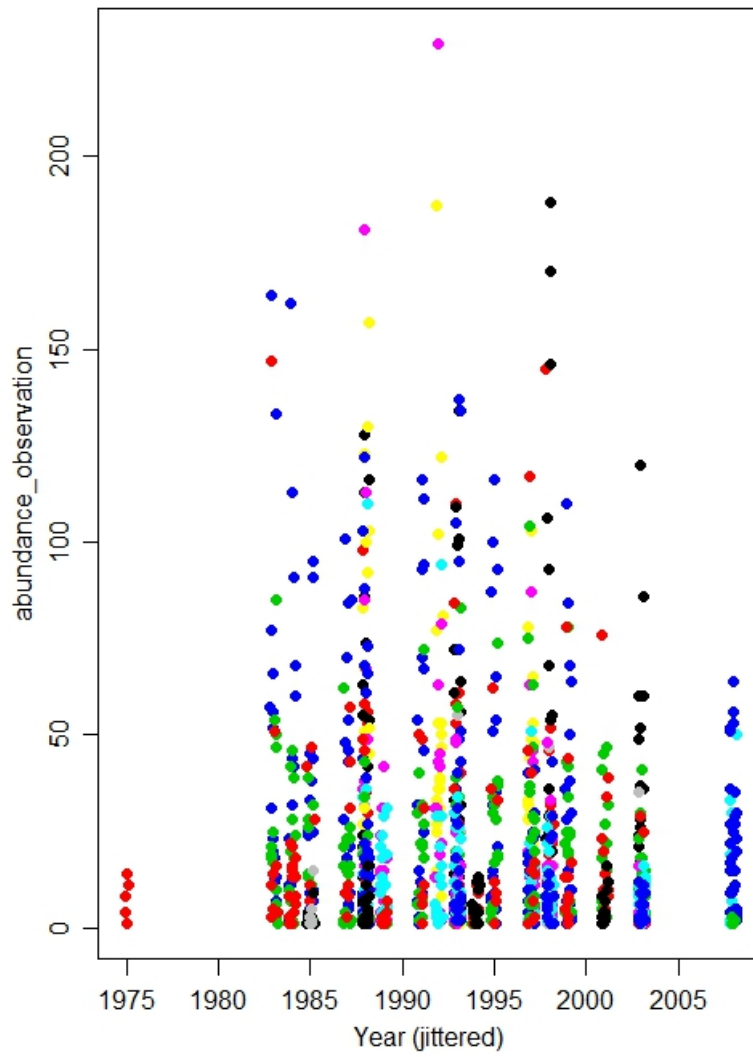


Figure 4: Time series of *Viburnum edule* abundance counts at Bonanza Creek LTER. The abundance observation is a count and colors correspond to multiple study sites, the largest scale of spatial replication in poplar. There are additional, smaller scales of spatial replication in this study, and hence multiple points per site per year.

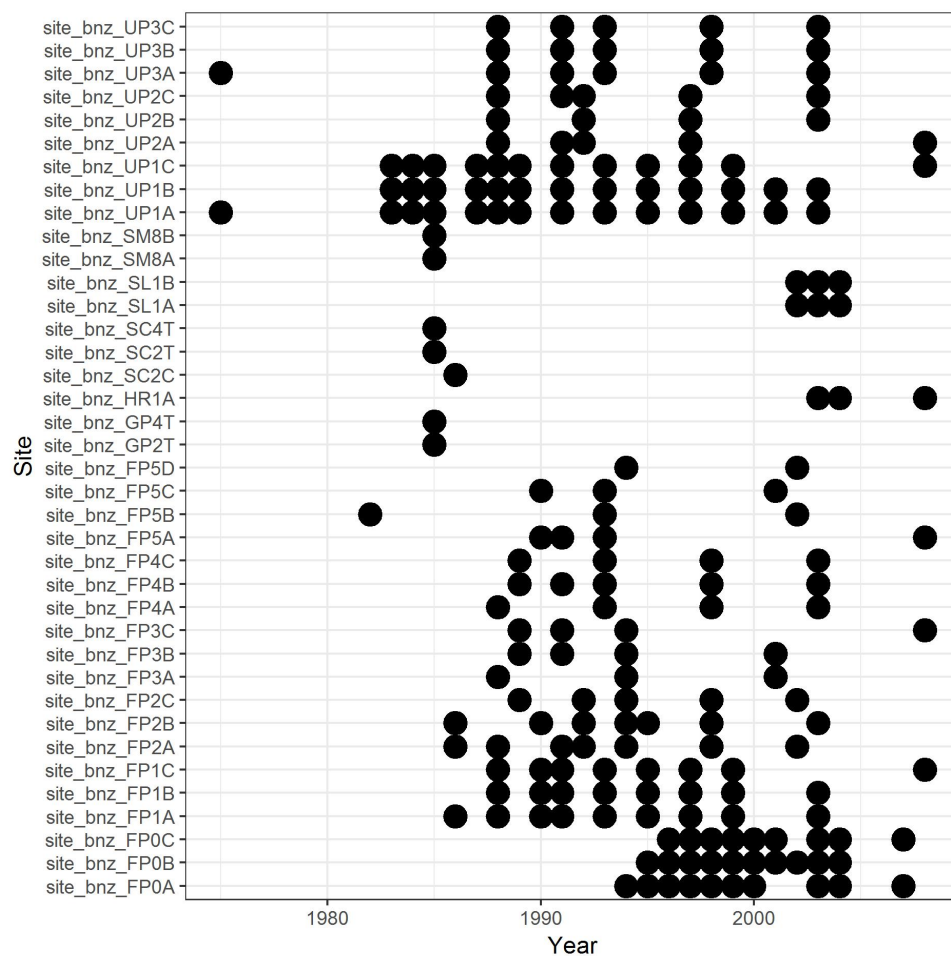


Figure 5: The temporal replication of the Bonanza Creek dataset.

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

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

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

393 In a handful of cases, we removed single data rows from the original dataset. These data
394 rows were associated with two types of typos in the original dataset. First, some abundance
395 observations were not associated with a time of observation. We removed this data because
396 **popler** can only accommodate population information associated with a time of observation.
397 Second, a handful of abundance data points were clear typos (e.g. the letter “l” instead of a
398 numeric value). We substituted these data points with a missing value (NULL in the database).
399 We uploaded these pre-processed datasets in the **popler** database through a Graphic User
400 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 (<code>community = F</code>) or multiple taxa (<code>community = T</code>)
spatial_replication_level	Extent of spatial replication level number n . A dataset can have up to to 5 replication levels.
spatial_replication_level	Unit of spatial extent of the n spatial replication level.
spatial_replication_level	Label of the spatial replication level (e.g. transect, plot, quadrat, ect.). The label of spatial replication level 1 is "site".
spatial_replication_level	The number of unique replicates for the n th level of spatial replication.
treatment_type_n	The type of treatment. <i>[This is vague and does not correspond to level n. PROBLEM.]</i>
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