Title: neonPlantEcology: an R package for preparing NEON plant data for use in ecological research.

Authors: Adam L. Mahood[1-3,*], Ranjan Muthukrishnan[4], Jacob A. Macdonald[1], David T. Barnett[5], Eric R. Sokol[5], Samuel M. Simkin[5]

Affiliations: 1. USDA-ARS; 2. Earth Lab, CU Boulder; 3. CU Boulder Geography; 4. Boston University Dept. of Biology; 5. National Ecological Observatory Network, Battelle, Boulder, Colorado, USA

*corresponding author: admahood@gmail.com

ORCID IDs: DTB 0000-0002-0485-3567; SMS 0000-0003-2418-4265; ERS 0000-0001-5923-0917; RM 0000-0002-7001-6249; JAM 0009-0009-3093-0667; ALM 0000-0003-3791-9654

Author Contributions: Conceptualization: ALM, RM; Data curation: ALM; Formal Analysis: ALM; Investigation: ALM, RM; Methodology: ALM; DTB; RM, ES, SS; Project administration: ALM Software: ALM, ES, DTB, SS; Visualization: ALM, JAM; Writing – original draft: ALM, JAM, RM; Writing – review & editing: ALM, RM, JAM

Acknowledgments: This project originated from collaborative partnerships formed at the 2019 NEON Science Summit, which was funded by NSF Award #DBI 1906144. The National Ecological Observatory Network is a program sponsored by the National Science Foundation and operated under cooperative agreement by Battelle. This material is based in part upon work supported by the National Science Foundation through the NEON Program.

Data Availability Statement: The code for the *neonPlantEcology* package is hosted on GitHub at https://github.com/admahood/neonPlantEcology and can be installed via remotes::install_github("admahood/neonPlantEcology").

Abstract

The National Ecological Observatory Network (NEON) is a continental-scale endeavor of ecological data collection for 30 years. We created a software package, *neonPlantEcology* that automatically arranges the raw data from the plant presence and percent cover (DP1.10058.001) data product from NEON into tables familiar to plant ecologists. Because of the broad scale of the observatory, it is necessary to tailor the data collection to the idiosyncrasies of each of 47 different ecosystems. Furthermore, data collection practices are occasionally modified for various reasons. These complexities, along with the volume and multiscalar nature of the data, need to be understood and accounted for in order to correctly process the data. This is particularly true for the plant diversity data product. We present three case studies using the package, centered around the three primary functions of *neonPlantEcology*. By automating the process of preparing NEON's plant diversity data, *neonPlantEcology* makes it more accessible to a wide range of users.

Keywords
National Ecological Observatory Network, NEON, Plant Data, Plant Ecology, R package

1. Introduction

In most terrestrial ecosystems, plants provide both the energetic foundation and the physical structure for ecological communities. Plants also lie at the interface between biogeochemical fluxes in the soil and in the atmosphere. Plant communities can be thought of as an expression of these fluxes, and thus tracking changes in those communities is critically important in the understanding of ecosystem dynamics. But collecting plant community data is time-consuming, requires deep local expertise, and often must be done at particular times of the year. These challenges make sampling at broad scales difficult. There are collections of plot data from disparate sources (e.g. vegBank citation), but these are often collected by different protocols, which are vulnerable to different types of observer error, making data harmonization an exercise in caution. The need for broad-scale, consistently collected data was one of the reasons for the formation of The National Ecological Observatory Network (NEON) (Keller *et al.*, 2008). NEON began collecting data on a myriad of ecosystem components using consistent protocols and observers at a handful of sites in 2014, eventually coming into full operation at 47 terrestrial sites across the United States in 2019. NEON data have great potential for use in plant ecology studies (Gill *et al.*, 2021; Muthukrishnan *et al.*, 2022), and are just now reaching a point in their lifespan where they have the potential to reveal groundbreaking insights, particularly when joined to an unprecedented array of *in situ* and remotely-sensed ancillary data collected at each NEON site (Meier, Thibault and Barnett, 2023). NEON technicians collect plant presence, cover and height annually or sub-annually around peak productivity, in a multiscalar framework (Barnett *et al.*, 2019; NEON, 2023). The nuance in the timing and frequency of data collection, combined with the standardization across sites is the main strength of NEON data products, but if these details are not understood and accounted for it can lead to errors in data preparation and interpretation.

The full set of NEON sites represent the breadth of natural systems that exist across the United States, but because these sites can have vastly different ecological circumstances, there is no one size fits all solution for collecting data. NEON divides the US into 20 domains, and each domain has one to three terrestrial sites where sampling is conducted, for a total of 47 terrestrial sites (Keller *et al.*, 2008). Sampling designs are aligned across all sites but at each site sampling approaches are adapted to reflect local considerations (Thorpe *et al.*, 2016; Barnett *et al.*, 2019). For example, higher latitude areas typically have their growing season peak in the mid to late summer, and so plant sampling is conducted in one bout during the peak of the growing season in order to capture peak productivity. But for hot deserts in the southwest productivity peaks in the spring and fall, and different plant species are abundant in these respective seasons. In these systems, plant sampling is conducted in two bouts corresponding

4

to the bimodal peaks in productivity. The NEON protocols also occasionally see small changes that are implemented due to logistical challenges, and these must be accounted for as well.

Here, we present an R package called *neonPlantEcology* that facilitates the retrieval and initial processing of NEON plant diversity data. The *neonPlantEcology* package processes raw Plant Presence and Percent Cover data (DP1.10058.001)(NEON, 2023), retrieved using NEON's API, into structures familiar to plant ecologists that are compatible with commonly-used packages like *vegan* (Oksanen *et al.*, 2022). *neonPlantEcology* converts the raw data into either a long data frame where each row is an observation of the cover of one plant species at one location, or a wide community matrix, where each row is a site, and each column is a species, at any spatial or temporal level of aggregation. It contains functions for obtaining height and groundcover data, as well as calculating biodiversity metrics from those same data objects (**Table 1**). It also has site- and plot-level geographic coordinates and polygons as included datasets. To support broad usage and easy modification, *neonPlantEcology* is coded in *tidyverse* syntax, which is easy to interpret and modify by end users in the community, and it is fast through use of a *data.table* backend via *dtplyr* (Dowle and Srinivasan, 2023; Wickham *et al.*, 2023). The package is currently focused on plant diversity data. Planned future updates will incorporate more functionality to seamlessly integrate ancillary data including vegetation structure, herbaceous biomass, remotely-sensed products and flux tower measurements to link to the plant community data outputs.

 5

Table 1. Functions in neonPlantEcology

Name: npe_<**name**>	purpose	input	output
1. download	Download data	Site abbreviation(s)	List object of raw data
2. site_info	Get site metadata	No input	Shapefile of site coordinates with metadata
3. longform	Turn raw data into longform cover data frame	Raw diversity data (1)	Longform data frame with each row as a cover value for each species at each subplot or plot
4. community_matrix	Create community matrix from raw data	Raw diversity data (1)	Data frame with each row as a site, each column as a species, and each cell value is either cover or occurrence
5. diversity_info	Summary diversity info at a chosen scale	Raw diversity data (1)	Summary diversity data
6. plot_centroids	Get plot centroids	Output from 3-5	Data frame of plot centroids
7. cm_metadata	Get metadata from community matrix	Community matrix from (4)	Data frame with rownames from 2, translated to metadata for each plot
8. change_native_status	Post-hoc change native status	Longform output (3)	Altered longform output
10. groundcover	Get ground cover estimates	Raw Diversity Data (1)	Ground cover for each spatial unit in a longform data frame
11. heights	Get height estimates	RawDiversity Data (1)	Height for each species at each spatial unit in a longform data frame
12. site_ids	Get 4-letter site codes	none	Vector of 4-letter site codes

2. Package Description

2.1 Plant diversity sampling structure. The aim of *neonPlantEcology* is to help ecologists acquire and process NEON Plant Presence and Percent Cover data (NEON, 2023) into familiar formats for ecological analyses. There are several facets to the NEON plant diversity data to which attention needs to be paid to properly format it at different scales. First two different types of data are collected within each plot at 4 spatial scales (Barnett *et al.*, 2019). In each 400m^2 (20m x 20m) plot, technicians estimate percent cover and height for all species within 6-8, 1m^2 subplots. Then, they record the occurrence of all species that did not occur in the 1m^2 subplots

 6

in $10m^2$ (3.16m x 3.16m) subplots that surround each $1m^2$ subplot. Occurrence is recorded again in two $100m^2$ (10m x 10m) subplots, each of which surrounds one of the 10m2 subplots.

The raw data are packaged as a list containing one data frame for cover and heights recorded in the $1m^2$ subplots, and one data frame with occurrences recorded in the $10m^2$ and $100m^2$ subplots. After 2019, the two central $1m^2$ and $10m^2$ subplots were excluded from sampling to minimize trampling at the plot centroid. Within a given year, the data are collected in bouts. Most sites have only one bout, but some have multiple bouts to account for multiple peaks in greenness, which may correspond to different species being active and abundant. The site, bout and year are recorded in the data in the “eventID” column in the raw data. *neonPlantEcology* allows the user to select a temporal resolution of subannual, annual or the whole time series, and this will be reflected in the eventID column. Subannual preserves bout-level information, and the eventID column will be formatted “site.bout.year”. Annual uses the maximum cover if a species is observed in the same subplot in both bouts, and the eventID column will be the year. If the entire time series is used, the eventID will be the range of years.

 7

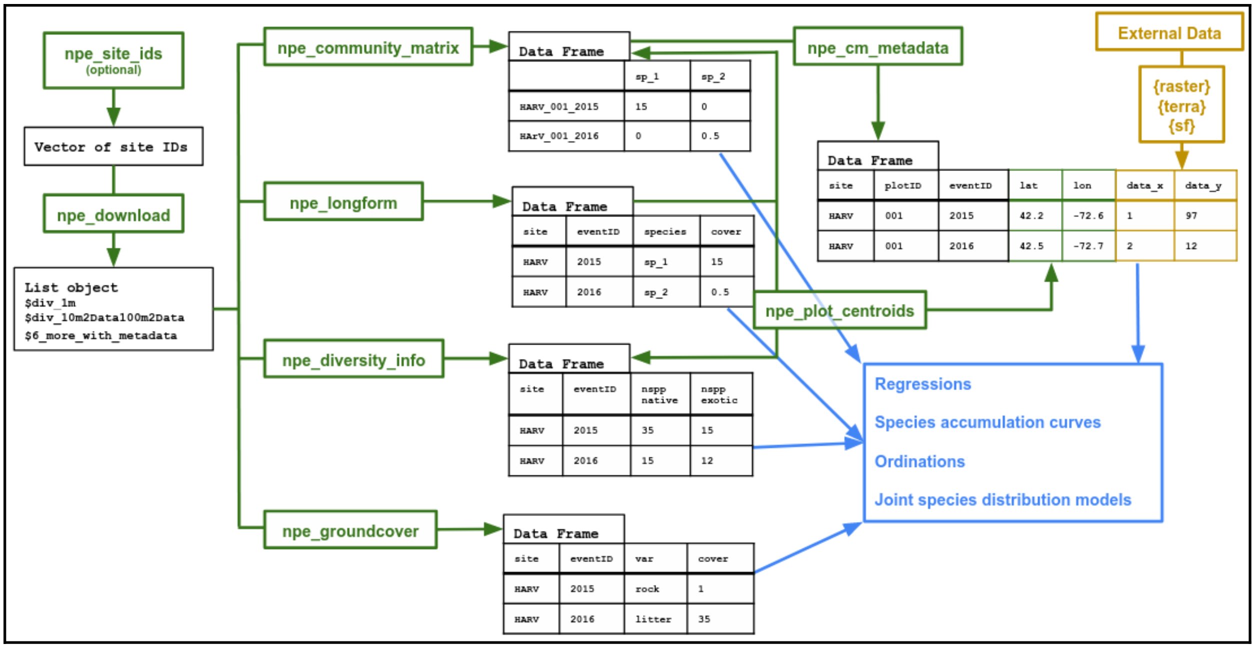

Figure 1. The main functions. `**npe_site_ids**` can be used to assist in site selection. Raw data is acquired through `**npe_download**`. It is then formatted to the desired structure and spatial and temporal scales of interest via `**npe_longform**` for long format, or `**npe_community_matrix**` for wide format. Metadata can be obtained from `**npe_cm_metadata**`, which can be used in concert with `**npe_plot_centroids**` to join community data with external ancillary data. `**npe_groundcover**` extracts ancillary data collected on site, and `**npe_diversity_info**` calculates higher-level information from the community matrix.

2.2 Functions. The neonPlantEcology package is based on a set of functions that pull community data from the NEON API and process it into more easily usable formats or that provide other useful data about specific NEON sites or plots, providing the components of a workflow starting at the raw data and ending with analysis-ready data (**Figure 1**).

`**npe_download**` uses `**loadByProduct**` from *neonUtilities* to download any data product from the NEON API. It downloads the Plant Presence and Percent Cover product (NEON, 2023) by default. `**npe_site_ids**` creates a vector of four letter site codes based on the domain, Koppen-Geiger climate classification, or aridity index that one can feed into `**npe_download**.` There are four data sets included in the package that can be loaded with the data function. **"site_polygons"** is polygons of each terrestrial NEON site location, along with some basic metadata in the row names (site, plot number, subplot ID and event ID). **"sites"** is a

 8

data frame of all terrestrial sites with additional metadata of aridity index and Koppen-Geiger climate classifications. **“Plot_centroids”** is point locations of each individual plot for the entire network. **“D14”** is the raw Plant Presence and Percent Cover data for domain 14 (NEON, 2023), which includes the Jornada and Santa Rita Experimental Ranges.

`npe_longform` creates a *long* data frame where each row has one cover value for one species, and there are columns defining the plot, subplot, site, eventID and so on. This function processes two list objects from the raw data, one which contains the 1 m^2 cover data and one which contains 10 and 100 m^2 occurrence data. It aggregates to the spatial scale (1m^2, 10m^2, 100m^2, the whole 400m^2 plot, or the site) and at a temporal scale (annual, sub-annual, or the full time series) chosen by the user. If the scale is 1m^2, the 10 and 100 m^2 subplots are discarded. If the scale is greater, the 10 and 100 m^2 subplots are given a trace value (default is 0.05%), then cover is calculated at the scales specified by the user.

`npe_community_matrix` creates a *wide* data frame, where each row is a site, plot or subplot, and each column is a species. The user can specify whether to return an abundance matrix (values 0-100) or an occurrence matrix (values 0 or 1).

`npe_diversity_info` calculates biodiversity and cover indices at the scales specified by the user. It returns a data frame with Shannon-Weaver diversity index (Shannon and Weaver, 1949) number of species, percent cover, and relative percent cover at each site. Each index is calculated for native, introduced, unknown, and all species together (**Table S1**). It optionally calculates all of these metrics for families specified by the user (**see Example 3**). The user also has the option of getting the beta diversity indexes of turnover and nestedness (Baselga, 2012) among the 1m^2 subplots for each plot, or among plots at each site.

`npe_cm_metadata` extracts the metadata from the data frame created by **`npe_community_matrix`** and puts it into a data frame. **`npe_plot_centroids`** downloads the spatial coordinates for each plot. NEON technicians also estimate other ground cover variables

 9

(rock, bare ground, etc), which are obtained with `**npe_groundcover**`, and the height of each species in each 1m2 subplot, obtained with `**npe_heights**`.

3. Package Installation and Examples

The package can be installed via `remotes::install_github("admahood/neonPlantEcology")`.

Example 1: summary data for all sites. For the first example, we downloaded the plant diversity data for all terrestrial sites and used `**npe_diversity_info**` to get site-level information on species richness, relative cover, and Shannon-Weaver alpha diversity, grouped by biogeographic origin (**Figure 2**).

```
library(neonPlantEcology)
all_sites <- npe_site_ids(all = TRUE) |>
  npe_download(sites = _)
di <- npe_diversity_info(all_sites, scale = "site", timescale = "all")
```

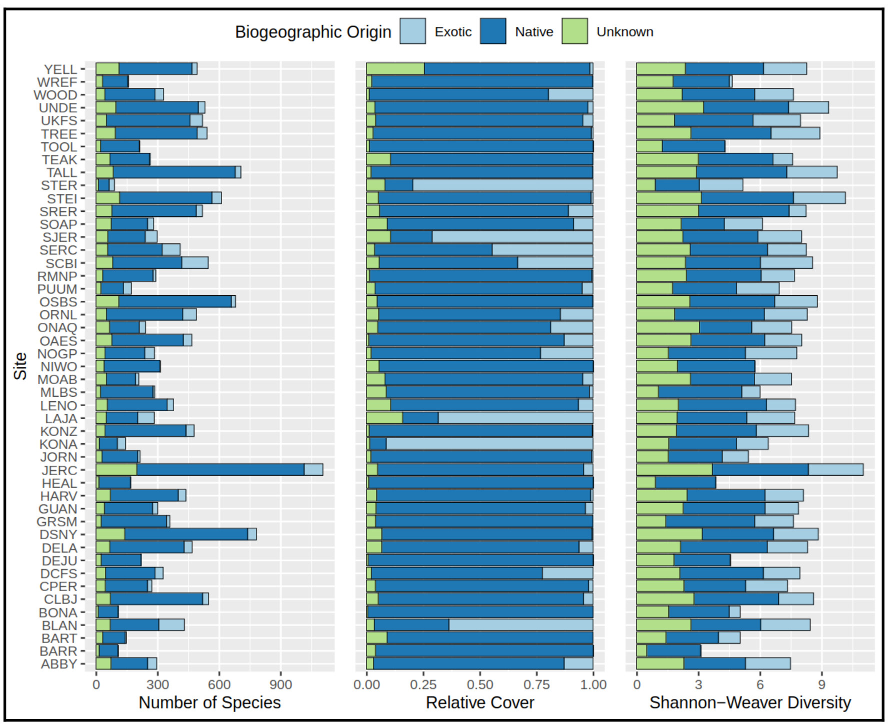

Figure 2. Number of species, relative cover, and Shannon-Weaver diversity grouped by biogeographic origin for all 47 sites. Plotting code for Figures 2-4 is in the package vignette.

 10

Example 2: community analysis for Domain 14. Next, we used the included Domain 14 diversity data object via `**data("D14")**`, which was fed to the `**npe_community_matrix**` function to produce a community matrix at an annual time scale and plot-level spatial scale. We then used `**npe_cm_metadata**` to get the plot ID numbers, site, and eventID for each row in the community matrix and joined that with additional plot-level metadata using the location information contained in `**data("plot_centroids")**`. Using these data we conducted a non-metric multidimensional scaling analysis (Minchin, 1987), from which we see separation in plant community composition between the two sites and within-year clustering (**Figure 3**).

```
data("D14"); data("plot_centroids")
library(tidyverse); library(sf); library(neonPlantEcology)
comm <- npe_community_matrix(D14)
metadata <- npe_plot_info(comm) |>
  left_join(plot_centroids |> st_set_geometry(NULL))
nmds <- metaMDS(comm)
nmds_sites <- nmds$points |>
  as_tibble(rownames = "rowname") |>
  left_join(metadata)
```

 11

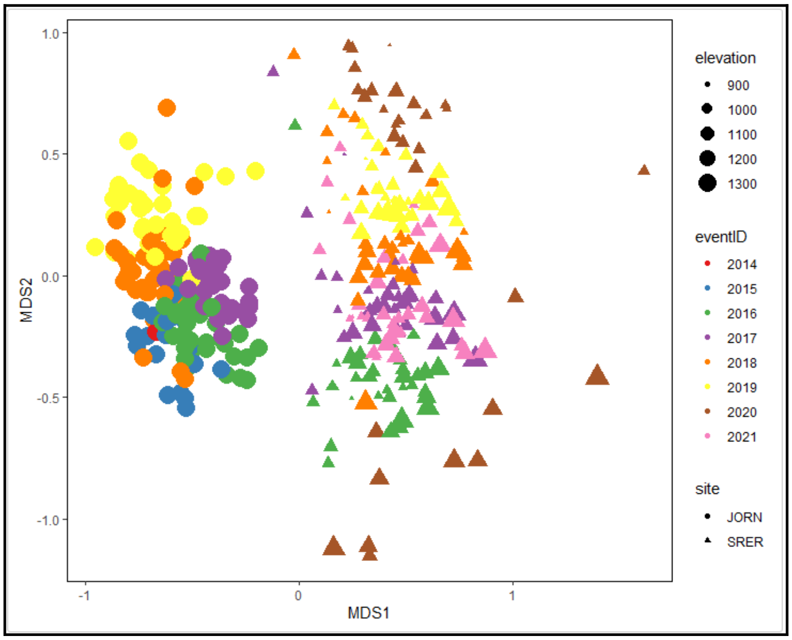

Figure 3. A non-metric, multidimensional scaling analysis for the plant communities at the two sites in Domain 14, the Santa Rita Experimental Range (SRER, triangles) and the Jornada Experimental Range (JORN, circles).

Example 3: cover by family through time. For the third example we use `**npe_longform**` to collect species cover by year at the plot scale, then aggregate by two families of interest: Fabaceae and Poaceae (Figure 4).

```
data("D14"); library(tidyverse); library(neonPlantEcology)
lf <- npe_longform(D14, scale = "plot", timescale = "annual")
lf_f <- lf |>
  mutate(family = ifelse(!family %in% c("Poaceae", "Fabaceae"),
                         "Other", family)) |>
  group_by(site, plotID, eventID, family) |>
  summarise(cover = sum(cover, na.rm=T)) |>
  ungroup()
```

 12

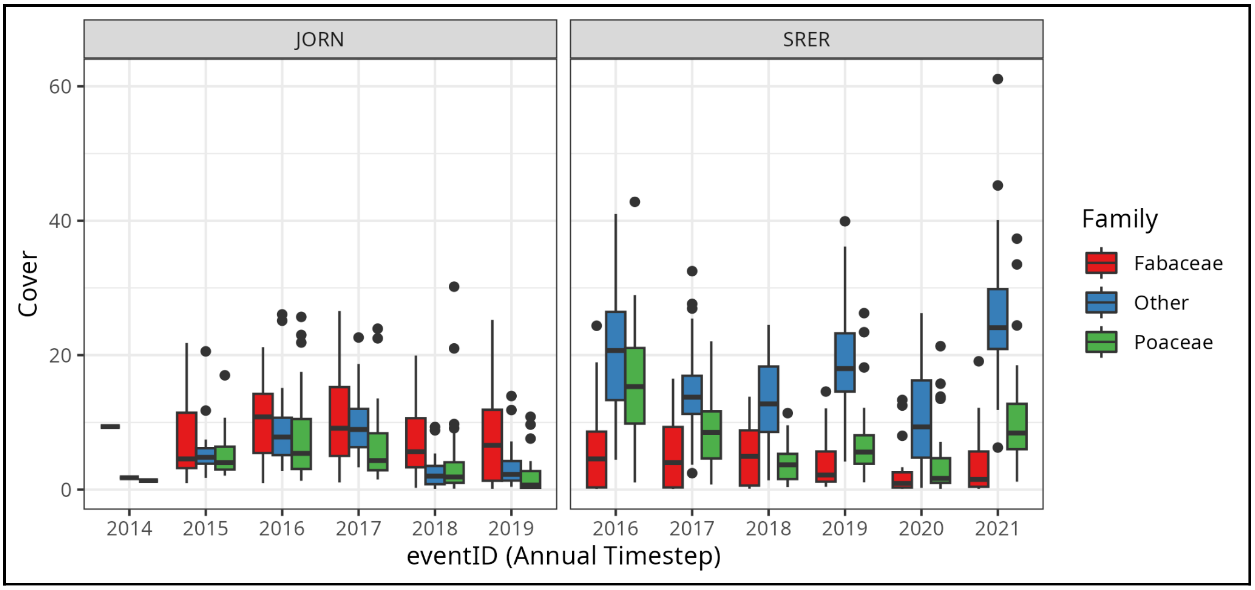

Figure 4. The sum of the annual percent cover of plants in the Fabaceae, Poaceae and all other families for the Jornada Experimental Range (JORN) and the Santa Rita Experimental Range (SRER).

4. Conclusion

neonPlantEcology complements the existing software ecosystem for working with NEON data by providing the basic service of conducting all of the steps of processing the diversity data from its raw form, accounting for spatial and temporal scale, sampling effort and changes in sampling design, to formats that are readable by programs and packages such as PC-ORD or R vegan which are familiar to ecologists. We aimed to create a package that contains sensible defaults at each decision point, but provides the flexibility for the end user to modify those decisions if it makes sense for their analysis. Wider adoption of this package will simplify the process of acquiring and processing of NEON data and facilitate broader usage by community ecologists, and assist and encourage researchers to conduct more cross site comparisons. Scaling up to multi-site or whole network analyses will be critical for achieving the broadest goals of NEON to understand the robustness or context dependence of ecological theory (Nagy *et al.*, 2021; Record *et al.*, 2021).

 13

References

Barnett, D.T. *et al.* (2019) ‘The plant diversity sampling design for The National Ecological Observatory Network’, *Ecosphere*, 10(2). Available at: https://doi.org/10.1002/ecs2.2603.

Baselga, A. (2012) ‘The relationship between species replacement, dissimilarity derived from nestedness, and nestedness’, *Global Ecology and Biogeography*, 21(12), pp. 1223–1232. Available at: https://doi.org/10.1111/j.1466-8238.2011.00756.x.

Dowle, M. and Srinivasan, A. (2023) *data.table: Extension of `data.frame`*. Available at: https://CRAN.R-project.org/package=data.table.

Gill, N.S. *et al.* (2021) ‘Six central questions about biological invasions to which NEON data science is poised to contribute’, *Ecosphere*, 12(9). Available at: https://doi.org/10.1002/ecs2.3728.

Keller, M. *et al.* (2008) ‘A continental strategy for the National Ecological Observatory Network’, *Frontiers in Ecology and the Environment*, 6(5), pp. 282–284. Available at: https://doi.org/10.1890/1540-9295(2008)6[282:ACSFTN]2.0.CO;2.

Meier, C.L., Thibault, K.M. and Barnett, D.T. (2023) ‘Spatial and temporal sampling strategy connecting NEON Terrestrial Observation System protocols’, *Ecosphere*, 14(3), p. e4455. Available at: https://doi.org/10.1002/ecs2.4455.

Minchin, P.R. (1987) ‘An evaluation of the relative robustness of techniques for ecological ordination’, *Vegetatio*, 69, pp. 89–107.

Muthukrishnan, R. *et al.* (2022) ‘Harnessing NEON to evaluate ecological tipping points: Opportunities, challenges, and approaches’, *Ecosphere*, 13(3). Available at: https://doi.org/10.1002/ecs2.3989.

Nagy, R.C. *et al.* (2021) ‘Harnessing the NEON data revolution to advance open environmental science with a diverse and data-capable community’, *Ecosphere*, 12(12), p. e03833. Available at: https://doi.org/10.1002/ecs2.3833.

NEON, (National Ecological Observatory Network) (2023) ‘Plant presence and percent cover

 14

(DPI.10058.001), RELEASE-2023.’ Available at: https://doi.org/10.48443/9579-a253.

Oksanen, J. *et al.* (2022) *vegan: Community Ecology Package*. Available at: https://CRAN.R-project.org/package=vegan.

Record, S. *et al.* (2021) ‘Novel Insights to Be Gained From Applying Metacommunity Theory to Long-Term, Spatially Replicated Biodiversity Data’, *Frontiers in Ecology and Evolution*, 8, p. 612794. Available at: https://doi.org/10.3389/fevo.2020.612794.

Shannon, C.E. and Weaver, W. (1949) ‘A mathematical model of communication’, *Urbana, IL: University of Illinois Press*, 11, pp. 11–20.

Thorpe, A.S. *et al.* (2016) ‘Introduction to the sampling designs of the National Ecological Observatory Network Terrestrial Observation System’, *Ecosphere*, 7(12), p. e01627. Available at: https://doi.org/10.1002/ecs2.1627.

Wickham, H. *et al.* (2023) *dtplyr: Data Table Back-End for ‘dplyr’*. Available at: https://CRAN.R-project.org/package=dtplyr.

Supplement

Table S1. Variables created by **`npe_diversity_info`**.

Variable	Description
shannon_<exotic/native/unknown/total>	Shannon-Weaver diversity of exotic, native, unknown or all species
evennness_<exotic/native/unknown/total>	Pielou's evenness of exotic, native, unknown or all species
nspp_<exotic/native/unknown/total>	number of species of exotic, native, unknown or all species
cover_<exotic/native/unknown/total>	Absolute cover as measured by technicians of exotic, native, unknown or all species
rel_cover_<exotic/native/unknown/total>	Relative cover - the absolute cover divided by the total cover of all species of exotic, native, unknown or all species

15

nfamilies	number of families
shannon_family	Shannon-Weaver diversity, but aggregated by family instead of species
evenness_family	Pielou's evenness, but aggregated by family instead of species
scale	The scale of aggregation (1m, 10m, 100m, plot or site)
invaded	Is there at least one exotic species present?
turnover	Species turnover (Baselga 2012) (if betadiversity = TRUE)
nestedness	Nestedness (Baselga 2012) (if betadiversity = TRUE)