

# Package: neonDivData (via r-universe)

October 23, 2026

**Type** Package

**Title** Standardized NEON Organismal Data for Biodiversity Research

**Version** 0.3.0

**Description** Cleaned, simplified, and standardized NEON organismal data for biodiversity research. The following taxonomic groups are included so far: algae, beetles, birds, macroinvertebrates, mosquitoes, plants, small\_mammals, ticks, tick\_pathogens, and zooplankton. NEON input data (<<https://data.neonscience.org>>) were processed and standardized.

**License** CC0

**Encoding** UTF-8

**LazyData** true

**LazyDataCompression** xz

**Depends** R (>= 3.5)

**Roxygen** list(markdown = TRUE)

**Imports** tibble

**Suggests** tidyverse, neonUtilities, lubridate, testthat (>= 3.0.0)

**Config/testthat/edition** 3

**Config/roxygen2/version** 8.0.0

**Repository** <https://daijiang.r-universe.dev>

**Date/Publication** 2026-10-23 19:59:01 UTC

**RemoteUrl** <https://github.com/daijiang/neonDivData>

**RemoteRef** HEAD

**RemoteSha** 6f6c514c0390607c2cb402c6c39ab6fa93967b3c

## Contents

|                       |   |
|-----------------------|---|
| data_algae . . . . .  | 2 |
| data_beetle . . . . . | 4 |
| data_bird . . . . .   | 6 |

|                                  |    |
|----------------------------------|----|
| data_fish . . . . .              | 9  |
| data_herp_bycatch . . . . .      | 11 |
| data_macroinvertebrate . . . . . | 13 |
| data_mosquito . . . . .          | 15 |
| data_plant . . . . .             | 17 |
| data_small_mammal . . . . .      | 21 |
| data_summary . . . . .           | 23 |
| data_tick . . . . .              | 24 |
| data_tick_pathogen . . . . .     | 26 |
| data_zooplankton . . . . .       | 28 |
| neon_location . . . . .          | 29 |
| neon_sites . . . . .             | 30 |
| neon_taxa . . . . .              | 31 |

|              |           |
|--------------|-----------|
| <b>Index</b> | <b>32</b> |
|--------------|-----------|

---

|            |                                                         |
|------------|---------------------------------------------------------|
| data_algae | <i>Periphyton, seston, and phytoplankton collection</i> |
|------------|---------------------------------------------------------|

---

## Description

This dataset was derived from **NEON data portal** with data product ID 'DP1.20166.001'. Details about this data product can be found at <https://data.neonscience.org/data-products/DP1.20166.001>.

## Usage

data\_algae

## Format

A data frame (also a tibble) with the following columns:

- location\_id: Location id.
- siteID: NEON site code.
- unique\_sample\_id: Identity of unique samples (equals sampleID).
- observation\_datetime: Observation date and time.
- taxon\_id: Accepted taxon code.
- taxon\_name: Scientific name associated with the taxon ID.
- taxon\_rank: The lowest level taxonomic rank that can be determined for the individual or specimen.
- variable\_name: The variable name(s) represented by the value column.
- value: Cell density value.
- unit: Either cells/mL (water column) or cells/cm2 (benthic).
- count\_method: How the count was obtained: "soft algae count", "diatom slide proportion x soft algae diatom count", or "diatom slide count".

- `redundant_genus`: TRUE for genus-level records that co-occur with a finer identification of the same genus in the same sample.
- `sampleCondition`: Non-OK condition of the taxonomy sample; NA if OK.
- `perBottleSampleVolume`: Sample volume per bottle (milliliter); fallback-filled where originally missing.
- `release`: Version of data release by NEON.
- `habitatType`: Habitat type sampled.
- `algalSampleType`: Type of algal sample collected.
- `samplerType`: Type of sampler used to collect the sample.
- `benthicArea`: Area of the benthos sampled (square meter).
- `samplingProtocolVersion`: The NEON document number and version where detailed information regarding the sampling method used is available; format NEON.DOC.#####vX.
- `substratumSizeClass`: Size class of the substratum sampled.
- `phytoDepth1`: First phytoplankton sample depth (meter) at sampling location.
- `phytoDepth2`: Second phytoplankton sample depth (meter) at sampling location.
- `phytoDepth3`: Third phytoplankton sample depth (meter) at sampling location.
- `latitude`: The geographic latitude (in decimal degrees, WGS84) of the geographic center of the reference area.
- `longitude`: The geographic longitude (in decimal degrees, WGS84) of the geographic center of the reference area.
- `elevation`: Elevation (in meters) above sea level.

## Details

To clean the data, we:

1. Filtered `alg_biomass` to `analysisType == "taxonomy"` records only.
2. Computed a fallback bottle volume (`preservativeVolume + labSampleVolume`, user guide equation 1) used when `perBottleSampleVolume` is NA or 0 (2014 samples).
3. Kept `alg_taxonomyProcessed` records with `algalParameterUnit == "cellsPerBottle"` and a taxon ID.
4. Reconciled diatom counts following user guide section 3.9. Soft algae + diatom counts are quantitative, while diatom slide counts give species-level relative abundance only. For samples with both, the total diatom count from the soft algae analysis was split among slide taxa in proportion to their slide counts, and replaced the soft algae diatom records. When a sample had slide data but no soft algae diatoms, slide counts were used as reported. Adding both analyses would double count diatoms. The method used for each record is given in `count_method`.
5. Joined taxonomy to `alg_biomass` and `alg_fieldData` via `sampleID` and `parentSampleID`.
6. Computed cell density: water column samples (seston/phytoplankton) as cells/mL; benthic samples as cells/cm<sup>2</sup> (cells per bottle / bottle volume x field sample volume / benthic area).
7. Resolved within-sample duplicates (separate morphotype or size records of one taxon) by summing density across records sharing the same `sampleID` and `acceptedTaxonID`.

**Note**

Details of locations (e.g. latitude/longitude coordinates can be found in [neon\\_location](#)).

**Temporal coverage.** NEON did not analyze algal taxonomy between November 2020 and December 2023, and the data product was suspended in 2024 (user guide section 3.3). Samples collected in late 2020 were counted only as natural units or valves, not cells, so cell density in this dataset ends in July 2020.

**Zero rows.** Algal samples always contain cells, so this dataset has no zero rows (the zero-row convention of the other datasets does not apply).

**Flags.** sampleCondition gives non-OK conditions of the taxonomy sample; NA if OK. Seston taxonomy exists only for 2014-2017 and has very low counts. redundant\_genus marks genus-level records that co-occur with a finer identification of the same genus in the same sample.

**Author(s)**

Lara Jansen

---

|             |                                     |
|-------------|-------------------------------------|
| data_beetle | <i>Ground beetle abundance data</i> |
|-------------|-------------------------------------|

---

**Description**

This dataset was derived from [NEON data portal](#) with data product ID 'DP1.10022.001'. Details about this data product can be found at <https://data.neonscience.org/data-products/DP1.10022.001>.

**Usage**

data\_beetle

**Format**

A data frame (also a tibble) with the following columns:

- location\_id: Location id.
- siteID: NEON site code.
- plotID: Plot identifier (NEON site code\_XXX).
- unique\_sample\_id: Identity of unique samples (equals sampleID).
- trapID: Identifier for the trap.
- observation\_datetime: Collection date of the trap.
- taxon\_id: Accepted species code, based on one or more sources. NA for zero rows.
- taxon\_name: Scientific name, associated with the taxonID. This is the name of the lowest level taxonomic rank that can be determined.
- taxon\_rank: The lowest level taxonomic rank that can be determined for the individual or specimen.

- `variable_name`: The variable name(s) represented by the value column.
- `value`: Abundance (count per trap day); 0 for zero-catch samples.
- `unit`: Unit of the values in the value column.
- `boutID`: Sampling bout identifier (`siteID_boutDate`, where the bout date is the most common collection date of the bout; collection can span two days).
- `nativeStatusCode`: The process by which the taxon became established in the location. 'A': Presumed absent; 'N': Native; 'I': Introduced; 'UNK': Status unknown.
- `redundant_genus`: TRUE for genus-level records that co-occur with a finer identification of the same genus in the same sample.
- `release`: Version of data release by NEON.
- `remarks`: Remarks (technical notes) of record.
- `samplingProtocolVersion`: The NEON document number and version where detailed information regarding the sampling method used is available; format 'NEON.DOC.#####vX'.
- `samplingImpractical`: Reason sampling was impractical; NA if sampling was OK.
- `sampleCondition`: Non-OK condition of the field sample; NA if OK.
- `trapConditionFlag`: Cup, lid, and fluid level problems; NA if none.
- `trappingDays`: Number of days the trap was set before this collection (numeric).
- `latitude`: The geographic latitude (in decimal degrees, WGS84) of the geographic center of the reference area.
- `longitude`: The geographic longitude (in decimal degrees, WGS84) of the geographic center of the reference area.
- `elevation`: Elevation (in meters) above sea level.
- `nlcdClass`: National Land Cover Database Vegetation Type Name.

## Details

To clean the data, we:

1. Filtered `bet_fielddata` to `sampleCollected == "Y"`. `trappingDays` is the number of days since the trap was set, or since its previous collection when a trap was collected more than once from the same set date. Placeholder eventIDs in a few 2014-2015 records were rebuilt from site, year, and week.
2. Built a taxonomy resolution hierarchy: Expert ID overrides Parataxonomist ID; both override Sorting taxonomy. For the rare individuals with more than one expert record, the most recent expert ID was used. Unpinned other carabid individuals are excluded because their sorting-level taxonomy is too coarse (2013-2016 protocol).
3. Computed unpinned counts per subsample (sorting total minus pinned individuals); clamped to zero where anomalous.
4. Dropped collected samples with no record at all in `bet_sorting` (mostly 2013-2014); these were never processed, so their catch is unknown.
5. Left-joined downward from field effort to taxonomy so that sorted samples with no carabids are kept as zero rows.
6.  $\text{Abundance} = \text{count} / \text{trappingDays}$  (count per trap day).

**Note**

Details of locations (e.g. latitude/longitude coordinates can be found in [neon\\_location](#)).

**Zero rows.** A trap sample that was collected and sorted but contained no carabids is kept as one row with `taxon_id` = NA and `value` = 0. This zero-row convention is shared by all `neonDivData` observation datasets.

**Flags.** Trap and sample problems are flagged rather than removed: `trapConditionFlag` lists cup, lid, and fluid level problems (e.g. "lid: Disturbed; fluid: High"; user guide section 5.4), `sampleCondition` gives non-OK sample conditions (e.g. "sample incomplete"), and `samplingImpractical` is carried through. All are NA when no problem was recorded. `redundant_genus` marks genus-level records that co-occur with a finer identification of the same genus in the same sample.

**Author(s)**

Kari Norman

---

data\_bird

*Breeding landbird point count data*

---

**Description**

This dataset was derived from [NEON data portal](#) with data product ID 'DP1.10003.001'. Details about this data product can be found at <https://data.neonscience.org/data-products/DP1.10003.001>.

**Usage**

data\_bird

**Format**

A data frame (also a tibble) with the following columns:

- `location_id`: Location id.
- `siteID`: NEON site code.
- `plotID`: Plot identifier (NEON site code\_XXX).
- `pointID`: Identifier for a point location.
- `unique_sample_id`: Identity of the sampling bout at a site (equals `eventID`, e.g. BRD.HARV.2019.1); a single point count is identified by `unique_sample_id`, `plotID`, and `pointID` together (point numbers repeat across plots).
- `observation_datetime`: Observation date and time.
- `taxon_id`: Accepted species code, based on one or more sources. NA for zero rows.
- `taxon_name`: Scientific name, associated with the `taxonID`. This is the name of the lowest level taxonomic rank that can be determined.

- **taxon\_rank**: The lowest level taxonomic rank that can be determined for the individual or specimen.
- **variable\_name**: The variable name(s) represented by the value column.
- **value**: Cluster size (count of individuals observed); 0 for zero-bird points; NA when a bird was detected but its cluster size was not recorded.
- **unit**: Unit of the values in the value column.
- **boutNumber**: Sampling bout within the season (most points are surveyed twice per year).
- **pointCountMinute**: The minute of sampling within the point count period; 88 = observed outside the count; NA for zero rows.
- **incidental\_observation**: TRUE when pointCountMinute == 88.
- **targetTaxaPresent**: Indicator of whether the sample contained individuals of the target taxa.
- **redundant\_genus**: TRUE for genus-level records that co-occur with a finer identification of the same genus in the same point count.
- **nativeStatusCode**: The process by which the taxon became established in the location. 'A': Presumed absent; 'N': Native; 'I': Introduced; 'UNK': Status unknown.
- **observerDistance**: Radial distance between the observer and the individual(s) being observed (meter).
- **detectionMethod**: How the individual(s) was (were) first detected by the observer.
- **visualConfirmation**: Whether the individual(s) was (were) seen after the initial detection.
- **sexOrAge**: Sex of individual if detectable, age of individual if individual cannot be sexed.
- **release**: Version of data release by NEON.
- **startCloudCoverPercentage**: Observer estimate of percent cloud cover at start of sampling.
- **endCloudCoverPercentage**: Observer estimate of percent cloud cover at end of sampling.
- **startRH**: Relative humidity as measured by handheld weather meter at the start of sampling.
- **endRH**: Relative humidity as measured by handheld weather meter at the end of sampling.
- **observedHabitat**: Observer assessment of dominant habitat at the sampling point at sampling time.
- **observedAirTemp**: The air temperature (Celsius) measured with a handheld weather meter.
- **kmPerHourObservedWindSpeed**: The average wind speed measured with a handheld weather meter, in kilometers per hour.
- **samplingProtocolVersion**: The NEON document number and version where detailed information regarding the sampling method used is available; format 'NEON.DOC.#####vX'.
- **remarks**: Remarks recorded for the point count (from brd\_perpoint).
- **clusterCode**: Alphabetic code (A-Z) linking clusters of individuals of the same species spanning multiple records.
- **latitude**: The geographic latitude (in decimal degrees, WGS84) of the geographic center of the reference area.
- **longitude**: The geographic longitude (in decimal degrees, WGS84) of the geographic center of the reference area.
- **elevation**: Elevation (in meters) above sea level.
- **nlcdClass**: National Land Cover Database Vegetation Type Name.
- **plotType**: NEON plot type in which sampling occurred: tower, distributed or gradient.

## Details

To clean the data, we:

1. Filtered `brd_perpoint` to `samplingImpractical == "OK"` to retain only surveyed points.
2. Left-joined downward from point survey records to `brd_countdata` by `eventID`, `plotID`, and `pointID` to preserve zero-bird survey points. Points without any count record were dropped.
3. Kept every detection record. Rows that look identical (same species, minute, distance, and cluster size) are separate detections and are not collapsed.
4. Collapsed each point count with no birds detected (all minutes `targetTaxaPresent == "N"`) into one zero row. Minute-level `targetTaxaPresent == "N"` rows at points where birds were detected in other minutes carry no information and were dropped.
5. `value` = cluster size (count of individuals); `unique_sample_id` = `eventID`.

## Note

Details of locations (e.g. latitude/longitude coordinates can be found in [neon\\_location](#)); coordinates are for the plot (grid) center, not the point. The sampling protocol has evolved over time, so users are advised to check whether the `samplingProtocolVersion` fits their data requirements and subset as necessary.

**Zero rows.** A point count with no birds detected is kept as one row with `taxon_id` = NA and `value` = 0. This zero-row convention is shared by all `neonDivData` observation datasets.

**Records to consider removing.** Each row is one detection (one cluster in one minute), so counts for a species at a point must be summed. Before analysis, consider:

- `incidental_observation == TRUE` (`pointCountMinute == 88`): birds observed outside the timed count, with unknown location (user guide section 3.5).
- Flyovers (`detectionMethod`), which have no distance measured.
- Records not identified to species (`taxon_rank` of class, family, subfamily, genus, or species-Group, e.g. "Aves sp." or slash taxa); `redundant_genus` marks genus-level records that co-occur with a species of the same genus in the same point count.
- Detections with an unknown cluster size (`value` = NA).
- Most points are surveyed twice per season (`boutNumber`).

## Author(s)

Daijiang Li, Eric Sokol

data\_fish

*Fish survey data collected by NEON (Discontinued for now)***Description**

This dataset was derived from **NEON data portal** with data product ID 'DP1.20107.001'. Details about this data product can be found at <https://data.neonscience.org/data-products/DP1.20107.001>. Sampling methods and the design are detailed here: <https://www.neonscience.org/data-collection/fish> and <https://www.neonscience.org/observatory/observatory-blog/one-fish-two-fish-learn-how-neon-samples-fish>

**Usage**

data\_fish

**Format**

A data frame (also a tibble) with the following columns:

- location\_id: Location id.
- siteID: NEON site code.
- pointID: NEON sampling point identifier.
- unique\_sample\_id: Identity of unique samples, usually it has location and date information.
- observation\_datetime: Observation date and time.
- taxon\_id: Accepted species code, based on one or more sources.
- taxon\_name: Scientific name, associated with the taxonID. This is the name of the lowest level taxonomic rank that can be determined.
- taxon\_rank: The lowest level taxonomic rank that can be determined for the individual or specimen.
- variable\_name: The variable name(s) represented by the value column.
- value: Value of the variable(s) specified by variable\_name.
- unit: Unit of the values in the value column.
- reachID: An identifier for the set of information associated with the reach.
- samplerType: Type of sampler used to collect the sample.
- fixedRandomReach: An indication of whether the reach is fixed or random.
- measuredReachLength: The length of the reach as measured by the technicians when the reach was established (meters).
- efTime: Operational time of the electrofisher (second).
- passStartTime: The start time of the pass.
- passEndTime: The end time of the pass.
- mean\_efishtime: Average efish time (in second).

- `release`: Version of data release by NEON.
- `netSetTime`: Time the net was set.
- `netEndTime`: Time the net was pulled.
- `netDeploymentTime`: Total time of deployment of the net (hours).
- `netLength`: Length of the net (meter).
- `netDepth`: Deployment depth of the net (meter).
- `efTime2`: Operational time of the electrofisher for the second electrofisher (second).
- `latitude`: The geographic latitude (in decimal degrees, WGS84) of the geographic center of the reference area.
- `longitude`: The geographic longitude (in decimal degrees, WGS84) of the geographic center of the reference area.
- `elevation`: Elevation (in meters) above sea level.

### Details

- We downloaded all fish data (i.e., `fsh_perPass`, `fsh_fieldData`, `fsh_bulkCount`, `fsh_perFish`), including the complete taxon table for fish, for both stream and lake sites surveyed via the NEON API.
- We joined the `'fsh_perPass'`, `'fsh_fieldData'`, and `'fsh_bulkCount'` datasets to produce a table with bulk-processed data that merged `'fsh_perPass'`, `'fsh_fieldData'`, and `'fsh_perFish'` to concatenate individual-level data.
- Finally both individual-level and bulk-processed datasets were appended into a single table. If `'fsh_bulkCount'` dataset does not have a `'taxonRank'` column, we added that information based on data stored in `'scientificName'` column particularly to separate species level identifications. For each finer-resolution taxon in the individual-level dataset, we considered the relative abundance as one since each row represented a single individual fish.
- Whenever possible, we substituted missing data by cross-referencing other data columns, omitted completely redundant data columns, and retained records with species-level taxonomic resolution. For the appended dataset, we also calculated the relative abundance for each species per sampling reach or segment at a given site.
- To calculate species-specific catch per unit effort (`'catch_per_effort'`), we normalized the relative abundance by either average electrofishing time (i.e., `'efTime'`, `'efTime2'`) or trap deployment time (i.e., the difference between `'netEndTime'` and `'netSetTime'`). In this case, we assumed that size of the traps used, water depths, number of netters used, and the reach lengths (a significant proportion of bouts had reach lengths missing) to be comparable across different sampling reaches and segments.

### Note

Details of locations (e.g. latitude/longitude coordinates can be found in [neon\\_location](#)).

### Author(s)

Stephanie Parker, Thilina Surasinghe

**Source**

<https://data.neonscience.org>; <https://data.neonscience.org/data-products/DP1.20107.001#collectionAndProcessing>

**References**

Jensen, B., S. Parker, and C. Scott. 2017. Neon user guide to fish electrofishing, gill netting, and fyke netting counts (NEON.DP1.20107). NEON, National Ecological Observatory Network, Boulder, CO, USA.

---

|                   |                                                                                          |
|-------------------|------------------------------------------------------------------------------------------|
| data_herp_bycatch | <i>Vertebrate Herpetofauna Bycatch sampled from pitfall traps (Discontinued for now)</i> |
|-------------------|------------------------------------------------------------------------------------------|

---

**Description**

This dataset was derived from [NEON data portal](https://data.neonscience.org) with data product ID 'DP1.10022.001'. Details about this data product can be found at <https://data.neonscience.org/data-products/DP1.10022.001>.

**Usage**

data\_herp\_bycatch

**Format**

A data frame (tibble) with the following columns:

- location\_id: Location id.
- siteID: NEON site code.
- plotID: Plot identifier (NEON site code\_XXX).
- unique\_sample\_id: Identity of unique samples, usually it has location and date information.
- trapID: Identifier for trap.
- observation\_datetime: Observation date and time.
- taxon\_id: Accepted species code, based on one or more sources.
- taxon\_name: Scientific name, associated with the taxonID. This is the name of the lowest level taxonomic rank that can be determined.
- taxon\_rank: The lowest level taxonomic rank that can be determined for the individual or specimen.
- variable\_name: The variable name(s) represented by the value column.
- value: Value of the variable(s) specified by variable\_name.
- unit: Unit of the values in the value column.
- trappingDays: Cleaned up decimal days between trap setting and collecting events
- release: Version of data release by NEON.

- nativeStatusCode: The process by which the taxon became established in the location only provided for vert bycatch herp
- remarksSorting: Technician notes; free text comments accompanying the record from sorting table
- remarksFielddata: Technician notes; free text comments accompanying the record from fielddata table
- latitude: The geographic latitude (in decimal degrees, WGS84) of the geographic center of the reference area.
- longitude: The geographic longitude (in decimal degrees, WGS84) of the geographic center of the reference area.
- elevation: Elevation (in meters) above sea level.
- plotType: NEON plot type in which sampling occurred: tower, distributed or gradient.
- nlcdClass: National Land Cover Database Vegetation Type Name.

## Details

To process data we:

1. Cleaned trappingDays.
  - So that is the number of days a trap was set before being collected.
  - Correct trap days to account for entries where the trap set date was not updated based on a previous collection.
2. Create a boutID that identifies all trap collection events at a site in the same bout essentially replacing eventID.
3. Update collectDate to reference the most common collection day in a bout, maintaining one collectDate per bout.
4. sampleType provides the group that was caught in the pit fall trap. This was changed to have three levels
  - "vert bycatch herp" - these are the samples
  - "no data collected" - these samples in fielddata not in sorting
  - "not herp" - this is an aggregate of all the other types "other carabid", "invert bycatch", "carabid", "vert bycatch mam" And we only kept "vert bycatch herp" in the final data product.

## Note

This script was derived from the script written by Kari Norman to process the pit fall traps of beetles. Additional variables were added and missing samples were retained in herp\_bycatch.

## Author(s)

Matt Helmus and Kari Norman

## Source

<https://data.neonscience.org>

## References

Hoekman, David, Katherine E. LeVan, Cara Gibson, George E. Ball, Robert A. Browne, Robert L. Davidson, Terry L. Erwin, et al. "Design for Ground Beetle Abundance and Diversity Sampling within the National Ecological Observatory Network." *Ecosphere* 8, no. 4 (2017): e01744.

---

data\_macroinvertebrate

*Aquatic macroinvertebrate density data*

---

## Description

This dataset was derived from [NEON data portal](https://data.neonscience.org/data-products/DP1.20120.001) with data product ID 'DP1.20120.001'. Details about this data product can be found at <https://data.neonscience.org/data-products/DP1.20120.001>.

## Usage

data\_macroinvertebrate

## Format

A data frame (also a tibble) with the following columns:

- `location_id`: Location id.
- `siteID`: NEON site code.
- `unique_sample_id`: Identity of unique samples (equals `sampleID`).
- `observation_datetime`: Observation date and time.
- `taxon_id`: Accepted species code, based on one or more sources. NA for zero rows.
- `taxon_name`: Scientific name, associated with the `taxonID`. This is the name of the lowest level taxonomic rank that can be determined.
- `taxon_rank`: The lowest level taxonomic rank that can be determined for the individual or specimen.
- `variable_name`: The variable name(s) represented by the value column.
- `value`: Density (count per square meter); 0 for zero rows.
- `unit`: Unit of the values in the value column.
- `redundant_genus`: TRUE for genus-level records that co-occur with a finer identification of the same genus in the same sample.
- `estimatedTotalCount`: Estimated total count (summed across size classes).
- `individualCount`: Raw individual count (summed across size classes).
- `subsamplePercent`: Percent of the total sample contained in the subsample (character; multiple values are separated by "|").
- `release`: Version of data release by NEON.

- **sampleCondition**: Non-OK conditions of the taxonomy records (separated by "|"); NA if all OK.
- **dataQF**: NEON data quality flag for the field sample; NA if none.
- **benthicArea**: Area sampled (square meter).
- **habitatType**: Habitat type sampled.
- **samplerType**: Type of sampler used to collect the sample.
- **substratumSizeClass**: Size class of the substratum sampled.
- **remarks**: Remarks of record.
- **ponarDepth**: Depth (meter) of petite ponar sample.
- **snagLength**: Length (centimeter) of snag sampled.
- **snagDiameter**: Diameter (centimeter) of snag sampled.
- **latitude**: The geographic latitude (in decimal degrees, WGS84) of the geographic center of the reference area.
- **longitude**: The geographic longitude (in decimal degrees, WGS84) of the geographic center of the reference area.
- **elevation**: Elevation (in meters) above sea level.

## Details

To clean the data, we:

1. Deduplicated `inv_fieldData` by `sampleID` using `slice(1)` as a guard against duplicate field records (none occur in the current release).
2. Filtered `inv_taxonomyProcessed` to `targetTaxaPresent == "Y"` and non-missing `estimatedTotalCount`; summed `estimatedTotalCount` and `individualCount` across size-class records sharing the same `sampleID` and `acceptedTaxonID` (user guide section 3.10).
3. Kept processed samples with no target taxa (`targetTaxaPresent == "N"`) as zero rows.
4. Inner-joined taxonomy to field data (inner join required because `benthicArea` is needed to compute density; records without field metadata are dropped).
5.  $\text{Density} = \text{estimatedTotalCount} / \text{benthicArea}$  (count per square meter, user guide equation 1).

## Note

Details of locations (e.g. latitude/longitude coordinates can be found in [neon\\_location](#)).

**Zero rows.** A processed sample with no target taxa is kept as one row with `taxon_id = NA` and `value = 0`. This zero-row convention is shared by all `neonDivData` observation datasets.

**Flags.** `sampleCondition` lists non-OK conditions of the taxonomy records (e.g. "damaged, affecting taxonomy"), and `dataQF` gives NEON field quality flags (e.g. "kicknet with incorrect mouth size used", which underestimates density); both are NA when no problem was recorded. `redundant_genus` marks genus-level records that co-occur with a finer identification of the same genus in the same sample; many macroinvertebrates are identified only to family or higher, so check `taxon_rank` before estimating richness.

**Author(s)**

Stephanie Parker, Eric Sokol

data\_mosquito

*Mosquitoes sampled from CO2 traps***Description**

This dataset was derived from **NEON data portal** with data product ID 'DP1.10043.001'. Details about this data product can be found at <https://data.neonscience.org/data-products/DP1.10043.001>.

**Usage**

data\_mosquito

**Format**

A data frame (also a tibble) with the following columns:

- `location_id`: Location id.
- `siteID`: NEON site code.
- `unique_sample_id`: Identity of unique samples (equals `sampleID`; `plotID.YYYYMMDDTHHMMZ.empty` for empty traps).
- `subsampleID`: Unique identifier associated with each subsample per `sampleID`.
- `observation_datetime`: Observation date and time.
- `taxon_id`: Accepted species code, based on one or more sources. NA for zero rows.
- `taxon_name`: Scientific name, associated with the `taxonID`. This is the name of the lowest level taxonomic rank that can be determined.
- `taxon_rank`: The lowest level taxonomic rank that can be determined for the individual or specimen.
- `variable_name`: The variable name(s) represented by the value column.
- `value`: Abundance (count per trap hour); 0 for zero-catch traps.
- `unit`: Unit of the values in the value column ('count per trap hour').
- `nativeStatusCode`: The process by which the taxon became established in the location. 'A': Presumed absent; 'N': Native; 'I': Introduced; 'UNK': Status unknown.
- `redundant_genus`: TRUE for genus-level records that co-occur with a finer identification of the same genus in the same sample.
- `proportionIdentified`: Proportion of the total catch that was subsampled and identified.
- `release`: Version of data release by NEON.
- `remarks_sorting`: Technician notes; free text comments accompanying the sorting record.

- `samplingProtocolVersion`: The NEON document number and version where detailed information regarding the sampling method used is available; format 'NEON.DOC.#####vX'.
- `sex`: M for male, F for female, U for unknown.
- `sortDate`: Date sample was sorted.
- `trapHours`: Number of hours between trap setting and collecting events.
- `nightOrDay`: Whether the trap was deployed during the night or the day.
- `sampleCondition`: Non-OK sample conditions from trapping, sorting, and expert identification; NA if none.
- `trapConditionFlag`: Trap operation problems (fan, catch cup, CO2); NA if none.
- `latitude`: The geographic latitude (in decimal degrees, WGS84) of the geographic center of the reference area.
- `longitude`: The geographic longitude (in decimal degrees, WGS84) of the geographic center of the reference area.
- `elevation`: Elevation (in meters) above sea level.
- `nlcdClass`: National Land Cover Database Vegetation Type Name.
- `plotType`: NEON plot type in which sampling occurred: tower, distributed or gradient.

## Details

To clean the data, we:

1. Kept deployed trap-nights from `mos_trapping` (`trapHours > 0`; `samplingImpractical` of OK or missing; `targetTaxaPresent` of Y, N, or U).
2. Left-joined `mos_trapping` to `mos_sorting` (by `sampleID`) and to `mos_expertTaxonomistIDProcessed` (by `subsampleID`), keeping traps that caught nothing.
3. Summed duplicate expert records of the same taxon and sex within a sample.
4. Kept expert identifications below family rank, and zero rows for traps (or identified subsamples) with no target taxa. Uncertain traps (`targetTaxaPresent == "U"`) without expert identifications were dropped.
5. Estimated total individuals per sample = `individualCount / proportionIdentified` (equivalent to the weight-based correction in the NEON user guide).
6. Abundance = estimated total individuals / `trapHours` (count per trap hour).

## Note

Details of locations (e.g. latitude/longitude coordinates can be found in [neon\\_location](#)).

**Zero rows.** A trap-night that caught no mosquitoes is kept as one row with `taxon_id = NA` and `value = 0`. NEON does not assign a `sampleID` to empty traps, so their `unique_sample_id` is built from the plot and the collection time in UTC (`plotID.YYYYMMDDTHMMZ.empty`); this format differs from NEON sampleIDs, which use local time, so the two cannot collide. This zero-row convention is shared by all `neonDivData` observation datasets.

**Flags.** Compromised samples and trap problems are flagged rather than removed: `sampleCondition` collects non-OK conditions from trapping, sorting, and expert identification (e.g. "trapping: Cold chain broken"), and `trapConditionFlag` collects fan, catch cup, and CO2 problems (e.g. "fan:

Off; CO2: Absent"). Both are NA when no problem was recorded. Genus-level identifications are kept; redundant\_genus marks those that co-occur with a finer identification of the same genus in the same sample. Day and night traps are distinguished by nightOrDay.

### Author(s)

Natalie Robinson, Daijiang Li

---

data\_plant

*Plant survey data collected by NEON*

---

### Description

This dataset was derived from **NEON data portal** with data product ID 'DP1.10058.001'. Details about this data product can be found at <https://data.neonscience.org/data-products/DP1.10058.001>

### Usage

data\_plant

### Format

A data frame (also a tibble) with the following columns:

- location\_id: Location id.
- siteID: NEON site code.
- plotID: Plot identifier (NEON site code\_XXX).
- unique\_sample\_id: Identity of the 100 m<sup>2</sup> subplot in one sampling event, in the format location\_id\_subplot\_id\_year-boutNumber.
- subplotID: This is the NEON provided subplot ID in the format of subplot\_id, then 1 or 10 (m<sup>2</sup>), then subsubplot\_id (e.g. 31\_1\_1, 31\_10\_1); if the sampling unit is 100 m<sup>2</sup>, the format is subplot\_id\_100 (e.g. 31\_100).
- subplot\_id: Subplot ID; each plot normally has four 100 m<sup>2</sup> subplots (31, 32, 40, 41).
- subsubplot\_id: Subsubplot ID (1, 2, 3, 4) for sampling units at 1 m<sup>2</sup> or 10 m<sup>2</sup>; NA for 100 m<sup>2</sup> records.
- observation\_datetime: Observation date and time.
- taxon\_id: Accepted species code, based on one or more sources. NA for zero rows.
- taxon\_name: Scientific name, associated with the taxonID. This is the name of the lowest level taxonomic rank that can be determined.
- taxon\_rank: The lowest level taxonomic rank that can be determined for the individual or specimen. Values are 'genus', 'species', 'speciesGroup', 'subspecies', or 'variety'. Species accounts for the majority of the entries. *Higher ranks have already been filtered out* because we think they are too vague for biodiversity research.

- `variable_name`: The variable name(s) represented by the value column.
- `value`: Percent cover for 1 m<sup>2</sup> records (0 for zero rows). For 10 m<sup>2</sup> and 100 m<sup>2</sup> records the value is NA (i.e., *presence only*); a small number of 1 m<sup>2</sup> records (mostly 2014-2018) also lack a cover estimate.
- `unit`: Unit of the values in the value column.
- `presence_absence`: 1 for records of a taxon; 0 for zero rows.
- `boutNumber`: Number of sampling bout, most sites sample only 1 bout.
- `nativeStatusCode`: Whether the species is a native or non-native species. 'A': Presumed absent, due to lack of data indicating a taxon's presence in a given location; 'N': Native; 'N?': Probably Native; 'I': Introduced; 'I?': Probably Introduced; 'NI': Native and Introduced, some infrataxa are native and others are introduced; 'NI?': Probably Native and Introduced, some infrataxa are native and others are introduced; 'UNK': Status unknown.
- `identificationQualifier`: NEON identification qualifier (e.g. "cf. species", "aff. species"), indicating uncertain identifications.
- `morphospeciesID`: Field morphospecies identifier, if the record was first identified as a morphospecies.
- `redundant_genus`: TRUE for genus-level records that co-occur with a finer identification of the same genus in the same `unique_sample_id`.
- `heightPlantOver300cm`: Indicator of whether individuals of the species in the sample are taller than 300 cm.
- `heightPlantSpecies`: Ocular estimate of the height (centimeter) of the plant species (if height is < 300 cm).
- `sample_area_m2`: The area of the sampling unit that the observed plant was located in. Potential values are 1, 10, or 100.
- `sampled_scales_m2`: Nested scales (m<sup>2</sup>) that were actually surveyed in this 100 m<sup>2</sup> subplot during this sampling event; the same value is repeated on every row of the subplot-event. Each 100 m<sup>2</sup> subplot contains two 1 m<sup>2</sup> quadrats and two 10 m<sup>2</sup> subplots. Since 2019, NEON surveys the 1 m<sup>2</sup> quadrats every year but the 10 m<sup>2</sup> and 100 m<sup>2</sup> scales only every other year (and in half of the plots in 2018). Values:
  - "1,10,100": full survey; the records give the complete species list of the 100 m<sup>2</sup> subplot.
  - "1": only the 1 m<sup>2</sup> quadrats were surveyed; the records cover about 2 m<sup>2</sup> of the subplot, so its species list is far from complete (e.g. subplot 31 of HARV\_001 has 21-26 species in fully surveyed years but 4-10 in 1 m<sup>2</sup>-only years).
  - Other combinations (e.g. "1,10", "10,100"): partial surveys with one scale missing; rare (<1% of subplot-events).

For species lists at 10 m<sup>2</sup>, 100 m<sup>2</sup>, or 400 m<sup>2</sup> (plot) scale, keep only full surveys, e.g. `dplyr::filter(data_plant, sampled_scales_m2 == "1,10,100")`; otherwise richness appears to drop every other year as a sampling artifact. Analyses at the 1 m<sup>2</sup> scale (percent cover, quadrat richness) need no filter.
- `latitude`: The geographic latitude (in decimal degrees, WGS84) of the geographic center of the reference area.
- `longitude`: The geographic longitude (in decimal degrees, WGS84) of the geographic center of the reference area.
- `elevation`: Elevation (in meters) above sea level.

- `plotType`: NEON plot type in which sampling occurred: tower, distributed or gradient.
- `nlcdClass`: National Land Cover Database Vegetation Type Name.
- `release`: Version of data release by NEON.
- `accepted_wcvp_name`: WCVP-accepted scientific name (binomial or trinomial, without authors) for species-, subspecies-, and variety-level taxa. For names matched to WCVP synonyms, this is the accepted name; for names with no accepted WCVP name, this falls back to the original NEON `taxon_name` (with authors). Genus-level taxa retain their original NEON names; speciesGroup taxa are assigned to one member species (see `species_group_assigned`).
- `accepted_wcvp_name_binomial`: Binomial (genus + specific epithet, without authors) extracted from `accepted_wcvp_name`. Useful for joining with other datasets that use binomial names only.
- `species_group_assigned`: TRUE for NEON speciesGroup records whose WCVP name columns give one assigned member species rather than a confirmed identification.

## Details

The detailed design of NEON plant survey can be found in [Barnett et al. 2019](#). Here, we:

1. Kept surveyed subplots only (`targetTaxaPresent` of Y or N; records flagged `samplingImpractical` have no taxa) and removed rows missing `plotID`, `subplotID`, `eventID`, or `endDate`.
2. Applied morphospecies resolutions from `div_morphospecies` (joined by `siteID` and `morphospeciesID`, NEON user guide sections 3.7 and 3.9), so field morphospecies later identified to genus or species carry the resolved `taxon_id`, `taxon_name`, and `taxon_rank`. Only unambiguous (one-to-one) resolutions were applied. Unresolved morphospecies that share a coarse code (e.g. 2PLANT) are kept as distinct taxa and can be told apart by `morphospeciesID`.
3. Where a subplot was surveyed on two dates within one sampling event (repeat surveys), kept the first survey. Where the same taxon was recorded twice in one subplot survey, kept the record with the highest percent cover.
4. Removed duplicate taxa between nested subplots (each taxon should be represented once per subplot per event). If a taxon is present in a 1 m<sup>2</sup> subplot, it is removed from the 10 m<sup>2</sup> subplot containing it; if present in any nested 1 or 10 m<sup>2</sup> subplot, it is removed from the 100 m<sup>2</sup> subplot.
5. Kept taxa identified to genus, speciesGroup, species, subspecies, or variety.
6. Standardized taxon names against the World Checklist of Vascular Plants (WCVP) by joining species-, subspecies-, and variety-level NEON names against a shared plant name base built with the `rWCVP` package (WCVP version recorded in `data-raw/plant_name_lookup.csv`). Specifically:
  - The join uses the full NEON name including authorship, because the same name can refer to different taxa depending on its author (homonyms).
  - In the name base, raw names are split into canonical name and authorship, matched exactly first and fuzzily only when needed, and synonyms are followed to the WCVP-accepted name. Fuzzy matches whose author contradicts the matched WCVP record, and names whose candidate matches disagree on the accepted taxon, are left unresolved rather than guessed.
  - The accepted names are stored in `accepted_wcvp_name` and `accepted_wcvp_name_binomial`; the original NEON name is retained in `taxon_name`.

- Names with no accepted WCV name (not in the name base, no match, or unresolved) and genus-level taxa retain their original NEON name.
  - speciesGroup taxa (groups of species that observers could not tell apart, e.g. "Fraxinus americana/pennsylvanica") were assigned to one member species in accepted\_wcvp\_name and accepted\_wcvp\_name\_binomial: the first-listed member by default or, if it is not recorded as a species at the same site, the first later-listed member that is. Parts that name only a genus (e.g. "Andropogon L.") are never assigned. These records are flagged with species\_group\_assigned = TRUE; the original NEON name is kept in taxon\_name.
7. Stacked species occurrence from different scales into a long data frame. A sampling event is identified by plotID, boutNumber, and the year of observation\_datetime (or unique\_sample\_id, which combines them with the 100 m<sup>2</sup> subplot). Always include the event when building sample units, otherwise all years are pooled. Therefore,
- to get the species list at 1 m<sup>2</sup> scale, we need all the data with sample\_area\_m2 == 1 (e.g. `dplyr::filter(plants, sample_area_m2 == 1)`); the sample unit is unique\_sample\_id plus subsubplot\_id (or unique\_sample\_id plus subplotID).
  - to get the species list at 10 m<sup>2</sup> scale, we need all the data with sample\_area\_m2 <= 10; the sample unit is unique\_sample\_id plus subsubplot\_id.
  - to get the species list at 100 m<sup>2</sup> scale, we need use the whole data set since the maximum value of sample\_area\_m2 is 100 (i.e. a 10 m by 10 m subplot); the sample unit is unique\_sample\_id.
  - to get the species list at 400 m<sup>2</sup> scale (i.e. one plot with four subplots), aggregate the data by plotID, boutNumber, and year.

## Note

Details of locations (e.g. latitude/longitude coordinates can be found in [neon\\_location](#)).

**Zero rows.** A 1 m<sup>2</sup> subplot that was surveyed but contained no plant species is kept as one row with taxon\_id = NA, value = 0, and presence\_absence = 0. This zero-row convention is shared by all neonDivData observation datasets.

**Sampling effort.** Since 2019, the 10 m<sup>2</sup> and 100 m<sup>2</sup> scales are surveyed only every other year (and in half of the plots in 2018), while 1 m<sup>2</sup> subplots are surveyed every year. Use sampled\_scales\_m2 to restrict analyses to subplots where all scales were surveyed (sampled\_scales\_m2 == "1,10,100") before comparing 100 m<sup>2</sup> or 400 m<sup>2</sup> species lists among plots or years.

**speciesGroup records.** About 0.7% of records are NEON speciesGroup taxa that were assigned to one member species in the WCV name columns (species\_group\_assigned = TRUE). For most of these, the assigned species is the only member recorded at that site, or both members occur there, so the assignment does not add a spurious species. Use the assigned binomial for community-level analyses (e.g. richness, composition); drop flagged records for species-specific analyses of either member; or collapse them to genus for a conservative richness estimate.

**Flags.** redundant\_genus marks genus-level records that co-occur with a finer identification of the same genus in the same unique\_sample\_id; drop them (!redundant\_genus) to avoid inflating species richness. identificationQualifier (e.g. "cf. species") marks uncertain identifications.

## Author(s)

Daijiang Li, Michael Just

**Source**

<https://data.neonscience.org>

**References**

Barnett, D.T., Adler, P.B., Chemel, B.R., Duffy, P.A., Enquist, B.J., Grace, J.B., Harrison, S., Peet, R.K., Schimel, D.S., Stohlgren, T.J. and Vellend, M., 2019. The plant diversity sampling design for the national ecological observatory network. *Ecosphere*, 10(2), p.e02603.

---

|                   |                                                     |
|-------------------|-----------------------------------------------------|
| data_small_mammal | <i>Small mammal box trap data collected by NEON</i> |
|-------------------|-----------------------------------------------------|

---

**Description**

This dataset was derived from [NEON data portal](https://data.neonscience.org) with data product ID 'DP1.10072.001'. Details about this data product can be found at <https://data.neonscience.org/data-products/DP1.10072.001>.

**Usage**

data\_small\_mammal

**Format**

A data frame (also a tibble) with the following columns:

- location\_id: Location id.
- siteID: NEON site code.
- plotID: Plot identifier (NEON site code\_XXX).
- unique\_sample\_id: Identity of unique samples (location\_id\_eventID, one plot-bout).
- observation\_datetime: Date of the last trapping night of the bout.
- taxon\_id: Accepted species code, based on one or more sources. NA for zero rows.
- taxon\_name: Scientific name, associated with the taxonID. This is the name of the lowest level taxonomic rank that can be determined.
- taxon\_rank: The lowest level taxonomic rank that can be determined for the individual or specimen.
- variable\_name: The variable name(s) represented by the value column.
- value: Unique individuals per 100 trap nights per plot per bout; 0 for zero rows.
- unit: Unit of the values in the value column.
- n\_individuals: Number of unique individuals captured in the plot-bout.
- target\_taxon: FALSE for taxa that are not NEON small mammal targets; NA for zero rows.
- redundant\_genus: TRUE for genus-level records that co-occur with a finer identification of the same genus in the same plot-bout.

- year: Year of the first trapping night of the bout.
- month: Month of the first trapping night of the bout.
- n\_trap\_nights\_per\_bout\_per\_plot: Number of trap nights per bout per plot.
- n\_nights\_per\_bout: Number of nights per bout.
- nativeStatusCode: Whether the species is a native or non-native species. 'A': Presumed absent, due to lack of data indicating a taxon's presence in a given location; 'N': Native; 'I': Introduced; 'UNK': Status unknown.
- release: Version of data release by NEON.
- latitude: The geographic latitude (in decimal degrees, WGS84) of the geographic center of the reference area.
- longitude: The geographic longitude (in decimal degrees, WGS84) of the geographic center of the reference area.
- elevation: Elevation (in meters) above sea level.
- plotType: NEON plot type in which sampling occurred: tower, distributed or gradient.
- nlcdClass: National Land Cover Database Vegetation Type Name.

## Details

To process data we:

1. Filter nights from mam\_perplotnight where samplingImpractical is not "OK" (field added 2020; pre-2020 records without this field are retained).
2. Identify bouts using eventID from mam\_perplotnight (site x year x week).
3. Compute trapping effort as the number of unique trap coordinates actually set per night (traps with status "1 - trap not set" and placeholder coordinates such as "XX" are excluded), summed across all valid nights within a bout per plot (n\_trap\_nights\_per\_bout\_per\_plot).
4. Retain only capture records with trapStatus "5 - capture" or "4 - more than 1 capture in one trap", restricting to taxa identified to genus or finer rank.
5. Resolve within-bout recaptures: tagged individuals (non-missing tagID) are counted once per plot per bout; untagged individuals (missing tagID) are each treated as unique using their row identifier. When recaptures of one individual were assigned different taxa, the most frequent identification is used.
6. Compute value as  $100 * n\_individuals / n\_trap\_nights\_per\_bout\_per\_plot$  (unique individuals per 100 trap nights per plot per bout). Effort and captures are both aggregated by plot and bout, including bouts that span two months.

## Note

Details of locations (e.g. latitude/longitude coordinates can be found in [neon\\_location](#)).

**Zero rows.** A plot-bout in which no target small mammal was captured is kept as one row with taxon\_id = NA, value = 0, and n\_individuals = 0. This zero-row convention is shared by all neonDivData observation datasets.

**Non-target taxa.** NEON targets cricetid, heteromyid, and murid rodents and small sciurids; shrews, moles, rabbits and pikas, weasels, opossums, large tree squirrels, and pocket gophers are captured only incidentally (user guide section 3). These captures are kept with target\_taxon =

FALSE. To analyze target taxa only, use `dplyr::filter(data_small_mammal, target_taxon | is.na(taxon_id))`; plot-bouts with only non-target captures already have a zero row.

**Comparability.** Diversity grids are trapped one night per bout and pathogen grids three nights per bout (`n_nights_per_bout`), so unique individuals per 100 trap nights is lower on three-night grids as recaptures accumulate. Use `n_individuals` with `n_nights_per_bout` (or mark-recapture models on the NEON data) when comparing grid types. `redundant_genus` marks genus-level records that co-occur with a finer identification of the same genus in the same plot-bout.

### Author(s)

Marta Jarzyna

### Source

<https://data.neonscience.org>

---

|              |                                                               |
|--------------|---------------------------------------------------------------|
| data_summary | <i>Summary of NEON data products included in this package</i> |
|--------------|---------------------------------------------------------------|

---

### Description

One row per taxonomic group. Computed directly from the cleaned datasets by `data-raw/02_clean_save_data.R`.

### Usage

`data_summary`

### Format

A data frame with the following columns:

- `taxon_group`: Taxonomic group name (e.g. "ALGAE", "BEETLES").
- `neon_product_id`: NEON data product ID (e.g. "DP1.20166.001").
- `r_object`: Name of the R data object for this group (e.g. "data\_algae").
- `n_taxa`: Number of unique taxa included.
- `n_sites`: Number of NEON sites with data.
- `sites`: All site codes that have data, separated by |.
- `start_date`: The earliest observation date in the dataset.
- `end_date`: The latest observation date in the dataset.
- `variable_names`: The variable name(s) in the value column.
- `units`: The unit(s) of the value column.

---

|           |                                        |
|-----------|----------------------------------------|
| data_tick | <i>Ticks sampled using drag cloths</i> |
|-----------|----------------------------------------|

---

### Description

This dataset was derived from **NEON data portal** with data product ID 'DP1.10093.001'. Details about this data product can be found at <https://data.neonscience.org/data-products/DP1.10093.001>.

### Usage

data\_tick

### Format

A data frame (also a tibble) with the following columns:

- location\_id: Location id.
- siteID: NEON site code.
- plotID: Plot identifier (NEON site code\_XXX).
- unique\_sample\_id: Identity of unique samples (plotID\_collectDate, where collectDate includes the time).
- observation\_datetime: Observation date and time.
- taxon\_id: Accepted species code, based on one or more sources. NA for zero rows.
- taxon\_name: Scientific name, associated with the taxonID. This is the name of the lowest level taxonomic rank that can be determined.
- taxon\_rank: The lowest level taxonomic rank that can be determined for the individual or specimen.
- variable\_name: The variable name(s) represented by the value column.
- value: Abundance (count per square meter); 0 for zero rows.
- unit: Unit of the values in the value column.
- LifeStage: Life stage of the tick (Larva, Nymph, or Adult).
- count\_source: Source of the count: "lab", "field count excess", or "field count excess + lab"; NA for zero rows.
- redundant\_genus: TRUE for genus-level records that co-occur with a finer identification of the same genus and life stage in the same sample.
- release: Version of data release by NEON.
- remarks\_field: Technician notes; free text comments accompanying the record.
- sampleCondition: Non-OK condition of the field sample; NA if no known compromise.
- samplingMethod: Name or code for the method used to collect or test a sample.
- targetTaxaPresent: Indicator of whether the sample contained individuals of the target taxa ('Y' or 'N').

- `totalSampledArea`: Total area sampled (square meter).
- `latitude`: The geographic latitude (in decimal degrees, WGS84) of the geographic center of the reference area.
- `longitude`: The geographic longitude (in decimal degrees, WGS84) of the geographic center of the reference area.
- `elevation`: Elevation (in meters) above sea level.
- `nlcdClass`: National Land Cover Database Vegetation Type Name.
- `plotType`: NEON plot type in which sampling occurred: tower, distributed or gradient.

## Details

To clean the data, we:

1. Kept field samples with `samplingImpractical` of OK or missing. Drags coded as having no ticks (`targetTaxaPresent == "N"`) that nonetheless have lab-identified ticks were recoded to "Y".
2. Mapped life stage from `tck_taxonomyProcessed` sex/age codes to Larva, Nymph, or Adult; rows with unmappable values are dropped. Lab records are not filtered on `sampleCondition`, because damaged ticks are still identified and counted.
3. Aggregated taxonomy counts by `sampleID`, `acceptedTaxonID`, and `LifeStage`; concatenated release values.
4. Reconciled field and lab counts (2014-2018 only; see note): when field counts exceed lab counts, the unidentified remainder is added as `Ixodidae sp.` (IXOSP, family rank), the code NEON itself uses for unidentified ticks. The remainder is capped so the sample total equals the larger of the field and lab totals, avoiding double counts when field staff assigned ticks to the wrong life stage.
5. Left-joined downward from field effort to lab taxonomy to preserve empty drag events (`targetTaxaPresent == "N"`).
6. `Abundance = individualCount / totalSampledArea` (count per square meter). Samples with ticks but no usable counts (e.g. lost samples after 2018 without lab data) are dropped.

## Note

Details of locations (e.g. latitude/longitude coordinates can be found in [neon\\_location](#)).

**Zero rows.** A drag with no ticks is kept as one row with `taxon_id = NA` and `value = 0`. This zero-row convention is shared by all `neonDivData` observation datasets.

**Protocol change in 2019.** From 2014 to 2018, field staff counted ticks by life stage, and the lab did not always identify every tick (invoice limits, subsampling of large larval samples). Field counts stopped in 2019; since then the lab counts every tick. `count_source` shows whether a count comes from the lab, from the field-count remainder, or both. Abundance of IXOSP (especially larvae) is therefore much higher before 2019.

**Flags.** `sampleCondition` gives non-OK field sample conditions (e.g. "Entire sample lost", "Tick count estimated"); NA if no compromise. `samplingMethod` distinguishes drag and flag sampling, which differ in efficiency. `redundant_genus` marks genus-level records that co-occur with a finer identification of the same genus and life stage in the same sample.

**Author(s)**

Wynne Moss, Melissa Chen, Brendan Hobart, Matt Bitters

---

|                    |                                   |
|--------------------|-----------------------------------|
| data_tick_pathogen | <i>Tick-borne pathogen status</i> |
|--------------------|-----------------------------------|

---

**Description**

This dataset was derived from **NEON data portal** with data product ID 'DP1.10092.001'. Details about this data product can be found at <https://data.neonscience.org/data-products/DP1.10092.001>.

**Usage**

data\_tick\_pathogen

**Format**

A data frame (also a tibble) with the following columns:

- location\_id: Location id (named location).
- siteID: NEON site code.
- plotID: Plot identifier.
- unique\_sample\_id: Identity of unique samples (namedLocation\_collectDate).
- observation\_datetime: Observation date and time.
- taxon\_id: Pathogen name (standardized).
- taxon\_name: Pathogen name (same as taxon\_id).
- taxon\_rank: Taxonomic rank inferred from name ("genus" for sp./spp. names, otherwise "species").
- variable\_name: The variable name(s) represented by the value column.
- value: Positivity rate (n\_positive\_test / n\_tests).
- unit: Unit of the values in the value column ("positive tests per pathogen per sampling event").
- vector\_taxon\_id: NEON taxon code of the tested tick (e.g. IXOSCA = Ixodes scapularis, AMBAME = Amblyomma americanum, HAELON = Haemaphysalis longicornis).
- lifeStage: Life stage of the host tick (N = nymph, L = larva; extracted from subsampleID).
- testProtocolVersion: The protocol version used to test the sample.
- release: Version of data release by NEON.
- n\_tests: Number of tests conducted.
- n\_positive\_test: Number of tests that were positive.
- latitude: The geographic latitude (in decimal degrees, WGS84) of the geographic center of the reference area.

- longitude: The geographic longitude (in decimal degrees, WGS84) of the geographic center of the reference area.
- elevation: Elevation (in meters) above sea level.
- nlcdClass: National Land Cover Database Vegetation Type Name.
- plotType: NEON plot type in which sampling occurred: tower, distributed or gradient.

## Details

To clean the data, we:

1. Removed tests from batches that failed quality criteria (`criteriaMet != "Y"` in `tck_pathogenqa`). QA records are linked by batch only from 2021 onward, and all batches in the current release met the criteria.
2. Removed samples where `sampleCondition != "OK"` or `testResult` is NA.
3. Applied the DNA quality fix: identified ticks (`testingID`) whose HardTick DNA Quality test was not "Positive" and dropped all test rows for those ticks (not just the DNA quality row itself), because pathogen results from ticks with degraded DNA are unreliable.
4. Removed assays that are not pathogens: HardTick DNA Quality and the tick species confirmation tests (*Ixodes pacificus*, *Ixodes scapularis*, *Amblyomma americanum*, *Dermacentor variabilis*, *Haemaphysalis longicornis*; added 2021-2022). Unified "*Borrelia burgdorferi*" into "*Borrelia burgdorferi sensu lato*".
5. Extracted the vector tick taxon (`vector_taxon_id`) and lifeStage from `subsampleID(plotID.date.vectorTaxonID)`.
6. Aggregated to one row per location x date x vector taxon x pathogen x life stage: `value = n_positive_test / n_tests`.

## Note

Details of locations (e.g. latitude/longitude coordinates can be found in [neon\\_location](#)).

**Vector species.** Each tick genus is tested for a different pathogen panel (e.g. *Borrelia burgdorferi* s.l., *Babesia*, and *Anaplasma* in *Ixodes*; *Ehrlichia* and *Francisella* in *Amblyomma*), and prevalence differs strongly among tick species. Always analyze positivity rates within `vector_taxon_id`. The panel also determines which pathogen rows exist, so the absence of a pathogen row means it was not tested, not that it was absent.

**Interpretation.** *Borrelia* sp. is a genus-level screen; ticks positive for any *Borrelia* species are also positive for *Borrelia* sp., so do not sum positives across *Borrelia* rows. Some pathogens (e.g. *Rickettsia parkeri* and *Rickettsia philipii*) were added to the panel only in 2022. Many rows are based on few tests (`n_tests`); weight by `n_tests` rather than averaging value. `taxon_id` is the pathogen name, not a NEON taxon code. This dataset has no zero rows: a rate of 0 means the pathogen was tested and not detected.

## Author(s)

Melissa Chen, Wynne Moss, Brendan Hobart, Matt Bitters

---

|                  |                                 |
|------------------|---------------------------------|
| data_zooplankton | <i>Zooplankton density data</i> |
|------------------|---------------------------------|

---

## Description

This dataset was derived from [NEON data portal](https://data.neonscience.org/data-products/DP1.20219.001) with data product ID 'DP1.20219.001'. Details about this data product can be found at <https://data.neonscience.org/data-products/DP1.20219.001>. Zooplankton are collected from the water column of lakes near NEON sensor infrastructure. The type of sampler used depends on the depth of water at the sampling location. Multiple tows or traps are collected at each location and composited into a single sample. Zooplankton sampling is quantitative and based on the volume of water collected during sampling.

## Usage

data\_zooplankton

## Format

A data frame (also a tibble) with the following columns:

- `location_id`: Location id (named location).
- `siteID`: NEON site code.
- `unique_sample_id`: Identity of unique samples (equals `sampleID`).
- `observation_datetime`: Observation date and time.
- `taxon_id`: Accepted taxon code.
- `taxon_name`: Scientific name associated with the taxon ID.
- `taxon_rank`: The lowest level taxonomic rank that can be determined for the individual or specimen.
- `variable_name`: The variable name(s) represented by the value column.
- `value`: Density (count per liter).
- `unit`: Unit of the values in the value column ("count per liter").
- `redundant_genus`: TRUE for genus-level records that co-occur with a finer identification of the same genus in the same sample.
- `release`: Version of data release by NEON.
- `sampleCondition`: Non-OK conditions of the taxonomy records (separated by "|"); NA if all OK.
- `samplerType`: Type of sampler used to collect the sample.
- `towsTrapsVolume`: Sample volume (liter) collected for zooplankton.
- `latitude`: The geographic latitude (in decimal degrees, WGS84) of the geographic center of the reference area.
- `longitude`: The geographic longitude (in decimal degrees, WGS84) of the geographic center of the reference area.
- `elevation`: Elevation (in meters) above sea level.

## Details

To clean the data, we:

1. Deduplicated zoo\_fieldData by sampleID using slice(1) to guard against NEON's known aquatic duplicate metadata issue.
2. Summed adjCountPerBottle across size-class and life-stage records sharing the same sampleID and taxonID. adjCountPerBottle is already corrected for subsampling by the laboratory. Records are not filtered on sampleCondition: most non-OK values are life-stage notes ("other (specified in remarks)", e.g. "Copepodids") or affect body length measurements only.
3. Inner-joined taxonomy to field data (inner join required because towsTrapsVolume is needed to compute density; records without field metadata are dropped).
4.  $\text{Density} = \text{sum}(\text{adjCountPerBottle}) / \text{towsTrapsVolume}$  (count per liter), per user guide equation 2.

## Note

Details of locations (e.g. latitude/longitude coordinates can be found in [neon\\_location](#)).

**Zero rows.** NEON reports no zooplankton samples without target taxa, so this dataset has no zero rows (the zero-row convention of the other datasets does not apply).

**Flags.** sampleCondition lists non-OK conditions of the taxonomy records (separated by "|"); NA if all OK. Consider excluding "damaged, affecting analysis". Tow net (whole water column) and Schindler trap samples differ in volume and depth sampled (samplerType), so check comparability before pooling. redundant\_genus marks genus-level records that co-occur with a finer identification of the same genus in the same sample.

## Author(s)

Lara Jansen, Stephanie Parker

---

neon\_location

*Information of all locations included in this data package.*

---

## Description

We extracted location information from all taxonomic groups and saved as one file here. *Note that some aquatic sites do not have lat/long information though.*

## Usage

neon\_location

**Format**

A data frame with the following columns:

- `location_id`: Location id.
- `siteID`: NEON site code.
- `plotID`: Plot identifier (NEON site code\_XXX).
- `latitude`: The geographic latitude (in decimal degrees, WGS84) of the geographic center of the reference area.
- `longitude`: The geographic longitude (in decimal degrees, WGS84) of the geographic center of the reference area.
- `elevation`: Elevation (in meters) above sea level.
- `nlcdClass`: National Land Cover Database Vegetation Type Name for terrestrial sites.
- `aquaticSiteType`: Type of aquatic systems ('lake', 'river', 'stream').

---

|            |                         |
|------------|-------------------------|
| neon_sites | <i>Site information</i> |
|------------|-------------------------|

---

**Description**

Full names, types, coordinates of all 81 NEON sites.

**Usage**

```
neon_sites
```

**Format**

A data frame with the following columns:

- `Site Name`: Full site name.
- `siteID`: NEON site code.
- `Domain Name`: Full domain name.
- `domainID`: Unique identifier of the NEON domain.
- `State`: The state name of the site locates in.
- `Latitude`: Latitude of the site (in decimal degrees, WGS84).
- `Longitude`: Longitude of the site (in decimal degrees, WGS84).
- `Site Type`: The type of the site (e.g. Core Terrestrial).
- `Site Subtype`: Second level site type, for aquatic sites only (e.g. Lake, Wadeable Stream, Non-wadeable River).
- `Site Host`: Host organization of the site.

---

|           |                                      |
|-----------|--------------------------------------|
| neon_taxa | <i>Taxonomic names of all groups</i> |
|-----------|--------------------------------------|

---

**Description**

This data frame was assembled from all taxonomic data products in the package. It is updated each release by `data-raw/02_clean_save_data.R` and preserves taxa from previous releases so that names are not lost between NEON data versions.

**Usage**

```
neon_taxa
```

**Format**

A data frame with the following columns:

- `taxon_id`: Accepted species code, based on one or more sources.
- `taxon_name`: Scientific name associated with the taxon ID. This is the name of the lowest level taxonomic rank that can be determined.
- `taxon_rank`: The lowest level taxonomic rank that can be determined for the individual or specimen.
- `taxon_group`: The taxonomic group the taxon belongs to (e.g. "ALGAE", "BEETLES").

**Note**

NEON source tables use either `taxonID` or `acceptedTaxonID` depending on the data product. Both are standardized to `taxon_id` here.

# Index

## \* datasets

- data\_algae, [2](#)
- data\_beetle, [4](#)
- data\_bird, [6](#)
- data\_fish, [9](#)
- data\_herp\_bycatch, [11](#)
- data\_macroinvertebrate, [13](#)
- data\_mosquito, [15](#)
- data\_plant, [17](#)
- data\_small\_mammal, [21](#)
- data\_summary, [23](#)
- data\_tick, [24](#)
- data\_tick\_pathogen, [26](#)
- data\_zooplankton, [28](#)
- neon\_location, [29](#)
- neon\_sites, [30](#)
- neon\_taxa, [31](#)

- data\_algae, [2](#)
- data\_beetle, [4](#)
- data\_bird, [6](#)
- data\_fish, [9](#)
- data\_herp\_bycatch, [11](#)
- data\_macroinvertebrate, [13](#)
- data\_mosquito, [15](#)
- data\_plant, [17](#)
- data\_small\_mammal, [21](#)
- data\_summary, [23](#)
- data\_tick, [24](#)
- data\_tick\_pathogen, [26](#)
- data\_zooplankton, [28](#)

- neon\_location, [4](#), [6](#), [8](#), [10](#), [14](#), [16](#), [20](#), [22](#), [25](#),  
[27](#), [29](#), [29](#)
- neon\_sites, [30](#)
- neon\_taxa, [31](#)