Package: neonDivData (via r-universe)

October 18, 2024

Type Package

Title Standardized NEON Organismal Data for Biodiversity Research

Version 0.1.1

Description Cleaned, simplified, and standardized NEON organismal data for biodiversity research. The following taxonomic groups are included so far: algae, beetles, birds, fish, herptiles, macroinvertebrates, mosquitoes, plants, small_mammals, ticks, tick_pathogens, and zooplankton. NEON input data (<https://data.neonscience.org>) were processed and standardized using R package `ecocomDP` (<https://github.com/EDIorg/ecocomDP>).

License CC0

Encoding UTF-8

LazyData true

Depends R (>= 2.10)

RoxygenNote 7.3.1

Roxygen list(markdown = TRUE)

Imports tibble

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

Config/testthat/edition 3

Repository https://daijiang.r-universe.dev

RemoteUrl https://github.com/daijiang/neonDivData

RemoteRef HEAD

RemoteSha 232350a1af754a6faafeb586f1f6799cb1513b41

Contents

`data_algae` *Periphyton, seston, and phytoplankton collection*

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, 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.

- `sampleCondition`: Condition of samples.
- `perBottleSampleVolume`: Per bottle sample volumn (milliliter).
- `release`: Version of data release by NEON.
- `habitatType`: Habitat type sampled.
- `algalSampleType`: Type of algal sample collected.
- `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.
- `samplerType`: Type of sampler used to collect the sample.
- `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

Here, we:

- Only records from the 'alg_biomass' with the 'analysis_type' taxonomy were used for the observation table.
- Combined the preservative sample volume and lab's recorded volumes from the 'alg_biomass' file and 'alg_tax_long' for total sample volume. If missing 'perBottleSampleVolume', they were created using NEON domain lab volumes. This new volume column was labeled 'perBSVol'.
- Combined 'Alg_field_data' with 'alg_biomass' via 'parentSampleID' to add in field conditions and 'benthicArea'.
- Corrected sample units from 'cellsperBottle' to density by dividing 'algalParameterValue' updated sample volume 'perBSVol. Benthic samples units were then corrected to density by multiplying 'algalParameterValue' by 'fieldSampleVolume' and dividing by 'benthicArea' sampled.

Note

Details of locations (e.g. latitude/longitude coordinates can be found in neon_location).

Author(s)

Lara Jansen

`data_beetle` *Ground beetles sampled from pitfall traps*

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, 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.
- `boutID`: Identifier for bout.
- `nativeStatusCode`: The process by which the taxon became established in the location. '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.
- `remarks`: Remarks (techinical notes) of record.
- `samplingProtocolVersion`: The NEON document number and version where detailed information regarding the sampling method used is available; format 'NEON.DOC.######vX'.
- `trappingDays`: Decimal days between trap setting and collecting events.
- `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 process data we:

1. Remove all non-carabid bycatch samples.
2. Use the most expert taxonomy when available.
3. Update abundances based on new taxonomy.
4. Create a boutID that identifies all trap collection events at a site in the same bout (replacing eventID).
5. Update collectDate to reference the most common collection day in a bout, maintaining one collectDate per bout.
6. Create a new `trappingDays` column for the number of days a trap was set before being collected.
7. Correct trap days to account for entries where the trap set date was not updated based on a previous collection.

Note

Details of locations (e.g. latitude/longitude coordinates can be found in neon_location).

Author(s)

Kari Norman

`data_bird` *Breeding landbird point counts 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 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`. `NA` represents no bird observed.
- `unit`: Unit of the values in the `value` column.
- `pointCountMinute`: The minute of sampling within the point count period.
- `targetTaxaPresent`: Indicator of whether the sample contained individuals of the target taxa.
- `nativeStatusCode`: The process by which the taxon became established in the location. 'A': Presumed absent, due to lack of data indicating a taxon's presence in a given location; 'N': Native; 'I': Introduced; 'UNK': Status unknown.
- `observerDistance`: Radial distance between the observer and the individual(s) being observed (unit: 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 can not 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 of record.

- `clusterCode`: Alphabetic code (A-Z) linked to clusters (groups of individuals of the same species) spanning multiple records. It is only used to link clusters that take up multiple lines on the data sheet.
- `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

The bird data provided by NEON is already well organized. We only removed some columns that likely won't be used in biodiversity studies. These columns include: 'identifiedBy', 'measuredBy', 'laboratoryName', 'samplingImpractical', 'samplingImpracticalRemarks', 'publicationDate', 'technicianID', 'observerInstitutionName', 'evaluationMethod', and 'evaluationScore'. If any of these information is important for the specific question asked by users, they should modify our code accordingly or download the raw data from NEON data portal directly. We also removed records without 'taxon_id'.

Note

Details of locations (e.g. latitude/longitude coordinates can be found in neon_location). The sampling protocol has evolved over time, so users are advised to check whether the `samplingProtocolVersion` fits their data requirements and subset as necessary.

Author(s)

Daijiang Li, Eric Sokol

`data_fish` *Fish survey data collected by NEON*

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 #' @referencesJensen, 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` *Vertebrate Herpetofauna Bycatch sampled from pitfall traps*

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_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 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 a 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 *Macroinvertebrate data*

Description

This dataset was derived from NEON data portal 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, 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.
- estimatedTotalCount: Estimated total count.
- individualCount: Individual count.
- subsamplePercent: Percent of the total sample contained in the subsample.
- release: Version of data release by NEON.
- 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

Here, we:

1. Removed field samples that had logistical issues (samplingImpractical!="OK")
2. Removed duplicate records in the field data table
3. Used count data that were already corrected for subsampling at the external taxonomy lab (estimatedTotalCount)
4. Summed estimatedTotalCount by scientificName per sampleID in inv_taxonomyProcessed to collapse sizeClass and lifeStage
5. Calculated density by dividing estimatedTotalCount from the inv_taxonomyProcessed table by benthicArea from the inv_fieldData table to standardize density by area sampled

Note

Details of locations (e.g. latitude/longitude coordinates can be found in neon_location).

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, usually it has location and date information.
- `sampleID`: Identifier for sample.
- `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.
- `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.
- `nativeStatusCode`: The process by which the taxon became established in the location. '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.
- `remarks_sorting`: Technician notes; free text comments accompanying the 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.
- `subsampleWeight`: Weight of subsample in gram.
- `totalWeight`: Weight of entire sample in gram.
- `trapHours`: Number of hours between trap setting and collecting events.
- `weightBelowDetection`: Notes regarding the weight relative to scale detection limit.
- `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

Here, we:

1. Removed trapping records that had no `collectDate`, `eventID`, `namedLocation`, and/or `sampleID`
2. Removed sorting records that had no `collectDate`, `sampleID`, `namedLocation`, and/or `subsampleID`
3. Removed archive records that had no `startCollectDate`, `archiveID`, and/or `namedLocation`
4. Removed expert taxonomy records that had no `collectDate`, `subsampleID`, and/or `namedLocation`
5. Verified there were no duplicated trapping, sorting, or archive records
6. Verified there were no trapping `sampleID` values missing from the sorting table
7. Verified there were no sorting `subsampleID` values missing from the expert taxonomist table
8. Converted blanks to NA throughout the datasets
9. Convert 0 to NA for unidentified samples - where `individualCount` in the expert taxonomist table was 0 and there was no listed `taxonID`
10. Left-joined the trapping table to the sorting table and verified barcode consistency between tables
11. Left-joined the expert taxonomy table to the trapping/sorting table and verified barcode and lab name consistency between tables
12. Renamed columns and added estimated total individuals as: number individuals iD'ed * (total subsample weight/ subsample weight)
13. Left-joined archive table and verified barcode consistency between tables

Note

Details of locations (e.g. latitude/longitude coordinates can be found in neon_location).

Author(s)

Natalie Robinson

`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 unique samples, usually it has location and date information.
- `subplotID`: This is the NEON provided subplot ID in the format of `subplot_id`, then `subsubplot_id` and 1 or 10 (m^2); if the sampling unit is 100 m^2, the values are 31, 32, 40, and 41.
- `subplot_id`: Subplot ID; each plot normally has four 100 m^2 subplots (31, 32, 40, 41).
- `subsubplot_id`: Subsubplot ID (1, 2, 3, 4) for sampling units at 1 m^2 or 10 m^2.
- `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. 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`: Value of the variable(s) specified by `variable_name`. If the individual was observed out of 1 square meter subplots, the value will be `NA` (i.e., *presence only*).
- `unit`: Unit of the values in the `value` column.
- `presence_absence`: All 1s since every record represent a species.
- `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.
- `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).
- `release`: Version of data release by NEON.
- `sample_area_m2`: The area of the sampling unit that the observed plant was located in. Potential values are 1, 10, or 100.
- `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

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

1. Removed 1 m^2 data with `targetTaxaPresent = N`
2. Removed rows missing values for plotID, subplotID, boutNumber, endDate, and/or taxonID
3. Removed duplicate taxa between nested subplots (each taxon should be represented once for the bout/plotID/year). For example, if a taxon/date/bout/plot combo is present in 1 m^2 data, remove from 10 m^2 and above
4. Stacked species occurrence from different scales into a long data frame. Therefore,
   - to get the species list at 1 m^2 scale, we need all the data with `sample_area_m2 == 1` (e.g. `dplyr::filter(plants, sample_area_m2 == 1)`); the unique sample unit id can then be generated with `paste(plants$plotID, plants$subplot_id, plants$subsubplot_id, sep = "_")`
   - to get the species list at 10 m^2 scale, we need all the data with `sample_area_m2 <= 10` (e.g. `dplyr::filter(plants, sample_area_m2 <= 10)`); the unique sample unit id can then be generated with `paste(plants$plotID, plants$subplot_id, plants$subsubplot_id, sep = "_")`
   - to get the species list at 100 m^2 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 unique sample unit id can then be generated with `paste(plants$plotID, plants$subplot_id, sep = "_")`
   - to get the species list at 400 m^2 scale (i.e. one plot with four subplots), we need aggregate the data at `plotID` level (the sample unit is the plot now).

Note

Details of locations (e.g. latitude/longitude coordinates can be found in neon_location).

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` *Small mammal box trap data collected by NEON*

Description

This dataset was derived from NEON data portal 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, 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.
- `year`: Observation year.
- `month`: Observation month.
- `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. Remove all records that are designated as "dead", "escaped", or "nontarget".
2. Remove all records designated as recaptures (i.e., only first captures are retained)

Note

Details of locations (e.g. latitude/longitude coordinates can be found in neon_location).

Author(s)

Marta Jarzyna

Source

https://data.neonscience.org

---

data_summary *Data product last modification time*

Description

This data frame records the last modify time for each data product in this package.

Usage

```
data_summary
```

Format

A data frame with the following columns:

- `taxon_group`: The taxa group that the location information can be used for. Note that some taxa groups may have the same 'plotID' but their latitude/longitude may differ slightly, which justifies the need of this column.
- `data_package_id`: Identifier for the data generated by `{ecocomDP}`.
- `n_taxa`: Number of species included. Note that we did not clean species anmes though we did remove records with species identified above genus level for well studied groups such as plant and fish. However, we keep all records regardless of their species identification levels for taxonomic groups that are hard to identifiy (e.g. macroinvertebrate, beetle).
- `n_sites`: Number of NEON sites included.
- `sites`: All site codes that have data, separated by `|`.
- `start_date`: The earliest date that have records.
- `end_date`: The latest date that have records.

- data_package_title: Title of the dataset.
- neon_ecocomdp_mapping_method: This is the ID that ecocomDP::read_data used to download and process the NEON data product with NEON's orignal ID specified in original_neon_data_product_id.
- original_neon_data_product_id: The NEON data product ID. See neonUtilities:::table_types for all available data types and their data product IDs provided by NEON.
- original_neon_data_version: NEON data release version.
- original_neon_data_doi: Original NEON data doi.
- r_object: The name of R objects for each taxonomic group. By calling the R objects, we can get the processed and standardized NEON organismal data for downstream diversity analysis.
- variable_names: The variable names that represent "abundance" information.
- units: The units of the variable_names.

---

data_tick *Ticks sampled using drag cloths*

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, 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.

- `LifeStage`: Life stage of the sample ('Adult', 'Larva', or 'Nymph').
- `release`: Version of data release by NEON.
- `remarks_field`: Technician notes; free text comments accompanying the record.
- `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

Here, we:

1. Removed field samples that had logistical issues (`samplingImpractical!="OK"`)
2. Removed field samples that had were not counted yet (e.g. 2019 data) e.g. require non NA values in `adultCount`, `nymphCount`, and `larvaCount`
3. Removed samples that had field counts but no associated taxonomy record, e.g. ticks were present in the field but there was no associated `sampleID` in the taxonomy (mostly 2019 records that haven't been ID'd yet). Also remove records in the taxonomy table that have no associated field data (`sampleID` present in taxonomy table but not field table). Records removed are mostly legacy.
4. Removed any sample in the taxonomy dataset that did not have a `sampleCondition` of "OK", or has an NA in `identifiedDate` or `acceptedTaxonID`
5. Removed samples in the taxonomy dataset that had `remarks` indicating a mis-identification. This included the following remarks: ("insect", "mite", "not a tick", "NOT A TICK", "arachnid", "spider").
6. Created a `lifeStage` column: this was either Nymph, Larvae, or Adult. Any tick that was sexed (e.g. `sexOrAge` was Male or Female) was assigned to the Adult lifestage. The Nymph or Larvae assignments were taken from the `sexOrAge` column.
7. Widened the taxonomy data so that each column is a unique taxonomicID_lifestage and each row is a sample. In doing so, we dropped the Sex information and summed M and F counts into a total adult count (i.e. we grouped by `sampleID`, `acceptedTaxonID`, and `lifeStage` and summed counts).
8. Left-joined the tick field data to tick taxonomy data using the `sampleID`. This retains 0 counts from the field data which will not have a record in the taxonomy table.
9. For any record where no ticks were found in the field (`targetTaxaPresent=="N"`), assigned 0s to all taxonomicID_lifestage columns.

10. For records where ticks were mis-id'd (see step 5) and there are no tick counts, adjusted the field count to 0.
11. Fixed count discrepancies between field and lab (note that as of 2019 counts may ONLY come from lab so this may not be necessary in future). Note that the lab count was used as the source of final counts per species-lifestage. When discrepancies were minor we trusted the lab counts. When discrepancies were larger, we used the following decisions: 11a. In cases where the field count total was greater than the taxonomic count total AND the discrepancy was in the larval stages, we assumed the lab stopped ID'ing after a certain count. We added the additional, un-ID'd "field" larvae to `IXOSP2_Larva` (the highest taxonomic ID) 11b. In other cases where field count was greater than the taxonomic count and there were `remarks_field` we assumed the remarks were about lost ticks (a common remark). We assigned any un-identified ticks to IXOSP2. 11c. Some counts did not match because there were too many samples sent to the lab and the invoice limit was reached. If the taxonomy table contained remarks about "invoice limit" or "billing limit" for a given sample ID, we trusted the field counts and added any difference between field and lab counts to the IXOSP2 column. *Note this was a conservative decision but one could reasonably assume that counts assigned to IXOSP2 would really belong to whatever lower order taxonomy was commonly ID'd for the remaining samples.*. 11d. Removed any remaining samples where field and lab counts were off by >30% and there was no obvious explanation.

Note

Details of locations (e.g. latitude/longitude coordinates can be found in neon_location).

Author(s)

Wynne Moss, Melissa Chen, Brendan Hobart, Matt Bitters

`data_tick_pathogen` *Tick-borne pathogen status*

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.
- `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.
- 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.
- lifeStage: Life stage of the host (all Nymph).
- 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 samples with flagged quality checks
2. Removed samples with missing metadata (domain/site/plotID; Lat/Long; plotType; nlcdClass, elevation, collectDate, subsampleID, batchID, testingID, testPathogenName
3. Removed samples where sampleCondition!="OK"
4. Removed samples where testResult==NA
5. Removed samples where HardTick DNA Quality (under testPathogenName) are not 'Positive', and then removed hardtack DNA and Ixodes pacificus tests.
6. Combined B. burgdeferi and B burgdeferi sensu lato into a single test pathogen type (B. burgedferi sensu lato)

Note

Details of locations (e.g. latitude/longitude coordinates can be found in neon_location).

Author(s)

Melissa Chen, Wynne Moss, Brendan Hobart, Matt Bitters

`data_zooplankton` *Zooplankton density data*

Description

This dataset was derived from NEON data portal 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.
- `siteID`: NEON site code.
- `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.
- `release`: Version of data release by NEON.
- `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

Here, we:

- Only keep those with `sampleCondition` to be "condition OK" for the `zoo_taxonomyProcessed` table.
- Combined the `zoo_fieldData` and the `zoo_taxonomyProcessed`.
- Calculated zooplankton density as the ratio of `adjCountPerBottle` and `towsTrapsVolume`.

Note

Details of locations (e.g. latitude/longitude coordinates can be found in neon_location).

Author(s)

Lara Jansen; Stephanie Parker

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

Description

We extracted location information from all taxanomic 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 *Site information*

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`: Latitide 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 *Taxanomic names of all groups*

Description

This data frame was put together from each data product.

Usage

```
neon_taxa
```

Format

A data frame with the following columns:

- `taxon_id`: Species code, based on one or more sources. For algae, macroinvertebrate, and tick, this is from the `acceptedTaxonID` column (which was removed here) so that all taxonomic groups have the same variable name. In another word, algae, macroinvertebrate, and tick only have `acceptedTaxonID` and we just renamed it to `taxon_id` for these groups following other groups.
- `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.
- `taxon_group`: The taxonomic group that the location information can be used for. Note that some taxa groups may have the same 'plotID' but their latitude/longitude may differ slightly, which justifies the need of this column.

Note

Some taxonomic groups used `taxonID` (renamed as `taxon_id` here) in the data product while other groups used `acceptedTaxonID`. In addition, all data were from NEON and we did not do extra clean up.

Index