|  |  |
| --- | --- |
| CSCI {CSCI} | R Documentation |

Score samples using the CSCI tool

**Description**

A function that aggregates many of the steps involved in the scoring of the California Stream Condition Index (CSCI) into a single function. These steps include data quality flagging, conversions of taxonomic names, iterative subsampling (20 iterations), metric calculations, prediction of expected taxa and metric values, scoring, and aggregation into a final index. Input data includes sample-wise raw, unprocessed taxonomy in a flat format, and station-wise predictor data in a crosstab format. See demo data for examples. A complete description of the index is provided in Mazor et al. (in review).

**Usage**

CSCI(bugs, stations, rand = sample.int(10000, 1))

**Arguments**

|  |  |
| --- | --- |
| bugs | A data frame with BMI data (see details) |
| stations | A data frame with environmental data (see details), one row per station. |
| rand | An integer to control the random number generator (RNG) seed for the subsampling. By default set to sample.int(10000, 1) |

**Details**

A valid "bugs" data frame consists of the following columns: StationCode, SampleID, FinalID (i.e., taxa names), LifeStageCode ("A", "L", "P", or "X"), BAResult (i.e., taxa counts), and Distinct (a positive integer where the taxonomist has indicated distinctiveness, else left blank or 0). Values for FinalID and LifeStageCode must conform to values from SWAMP lookup tables (<http://swamp.mpsl.mlml.calstate.edu/>). See CSCI guidance document for details on these fields.

A valid "stations" data frame consists of the following columns: StationCode (must match with same column in the "bugs" data frame), New\_Lat, New\_Long, AREA\_SQKM , ELEV\_RANGE, BDH\_AVE, PPT\_00\_09, LPREM\_mean, KFCT\_AVE, TEMP\_00\_09, P\_MEAN, N\_MEAN, PRMH\_AVE, SITE\_ELEV, MgO\_Mean, S\_Mean, SumAve\_P, CaO\_Mean. See CSCI guidance document for details on these fields.

The data frames are also subject to the following constraints: no missing blank cells in any field in either data frame (except for the Distinct column); all values under StationCode in the "bugs" data frame must be represented under StationCode in the "stations" data frame; every SampleID must be associated with only a single StationCode; no duplicated data in either data frame (e.g., every combination of the SampleID, FinalID, LifeStageCode, and Distinct should be unique in the “bugs” data frame).

In order to produce replicable results, the RNG seed can be controlled using the rand argument. Enter any integer.

**Value**

A list of data frames that serve as reports in varying detail:

|  |  |
| --- | --- |
| core | A summary of the CSCI results, and data quality flags, averaged across 20 iterations. |
| Suppl1\_mmi | A detailed breakdown of the results of the MMI component of the CSCI, averaged across 20 iterations. |
| Suppl1\_grps | Probability of biotic group membership in a SampleID by Group format |
| Suppl1\_OE | A detailed breakdown of the results of the O/E component of the CSCI, averaged across 20 iterations. Capture probabilities and mean abundances of each OTU are provided. |
| Suppl2\_mmi | Similar to Suppl1\_mmi, except broken down by iteration. |
| Suppl2\_OE | Similar to Suppl1\_OE, except brown down by iteration. Iteration-wise O/E scores are also provided. |

**Examples**

data(bugs\_stations) #A list of two data frames: bugs and stations

results <- CSCI(bugs = bugs\_stations[[1]], stations = bugs\_stations[[2]])

ls(results) #see all the components of the report.

results$core #see the core report.