M26 Design of data and indicator robustness measures

28/02/2025

Author(s): Langeraert Ward, Van Daele Toon

Funded by
the European Union

This project receives funding from the European Union's Horizon Europe Research and Innovation Programme (ID No 101059592). Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the EU nor the EC can be held responsible for them.

Table of Contents

Summary

Biodiversity indicators derived from occurrence cubes must be assessed for reliability and meaningfulness. Key aspects include robustness measures, uncertainty quantification, and interpretation frameworks. Robustness measures evaluate adequacy and representativeness of the data. Furthermore, uncertainty quantification, using bootstrapping, ensures correct indicator interpretation, supporting informed decision-making. Best practices are explored based on existing techniques and preliminary analyses in R.

In light of data variability, measures for data cube and species robustness are proposed. Data cube robustness metrics assess data quality across spatial, temporal, and taxonomical dimensions. These metrics can serve as early warning systems during data exploration, e.g., indicating when not enough data is present or when strong data clustering is present along one or more dimensions. For species robustness, a cross-validation technique is proposed where species are systematically excluded and the indicator is recalculated: leave-one-species-out cross-validation. The method is a tool for data exploration that quantifies the influence of a single species on indicator calculation.

In light of indicator variability, methods for uncertainty quantification and effect classification are discussed. Indicator uncertainty can be calculated using the bootstrap resampling technique, from which confidence intervals can be generated. Four interval types are compared: (1) normal: assumes normal distribution, (2) basic: centers interval using percentiles, (3) percentile: uses bootstrap distribution percentiles, and (4) bias-corrected and accelerated (BCa): percentile that adjusts for bias and skewness. Based on literature and preliminary analysis, the BCa interval is recommended over the percentile interval as it accounts for bias and skewness in the bootstrap distribution. The normal and basic intervals are included for the sake of simplicity, but rarely recommended in practice. Finally, effect classification helps interpret trends by comparing confidence limits with reference values and thresholds.

The proposed methods will be bundled in an R package called **dubicube**. The functions in this package can be used for exploratory analyses of occurrence cubes, as well as uncertainty calculation and interpretation of derived indicators.

1 Introduction

Although we can always calculate indicators from an occurrence cube, it is essential to evaluate whether those indicators are reliable and meaningful. The key questions are as follows: Can the calculated indicator values be interpreted with confidence? Do we have sufficient and representative data to support valid conclusions? To address this, we focus on three complementary aspects: robustness measures, indicator uncertainty, and interpretation frameworks.

In this report, we introduce measures of robustness as a way to evaluate the applicability and reliability of biodiversity indicators. These measures assess whether the underlying data are adequate and representative for drawing meaningful conclusions about a given species, time period, or area. Note that this is different from the statistical definition of robustness which entails the resistance of a statistical method, estimator, or model to outliers or violations of assumptions. Our robustness measures are properties of the data, calculated alongside indicator values. They are not a property of the indicator itself.

Indicator values are sometimes calculated without a corresponding measure of uncertainty, such as confidence intervals (Rowland et al., 2021), although its importance is paramount for correct interpretation, and consequently, for making adequate management and policy decisions (Fischhoff & Davis, 2014; Milner-Gulland & Shea, 2017). For indicators from occurrence cubes, we propose a framework for uncertainty calculation using bootstrapping, a flexible, non-parametric method. Furthermore, to aid in interpretation, we introduce an effect classification method to categorise indicator effects where confidence intervals are compared with reference values.

By combining measures of robustness, uncertainty quantification, and a clear interpretation framework, we ensure that biodiversity indicators are not only calculated but also contextualised in a way that supports confident and meaningful decision-making.

2 Example Analyses and Dataset

Example analyses were performed with R v4.4.2 (R Core Team, 2024) using RStudio (Posit team, 2024). The code can be consulted in the repository of Langeraert et al. (2025, v1.4.0). Data wrangling and visualisation were done using the **tidyverse** package (Wickham et al., 2019).

As an example dataset, we used the ‘europe_insect_cube’ occurrence cube of the **b3gbi** R package v0.2.2 (Dove, 2024). This cube is derived from European insect data from the ‘Natural History Collections of the Faculty of Biology AMU’ dataset published on GBIF (GBIF.org, 2024). We selected the data from 2011-2020. This data cube is structured as a single-resolution 1 km² grid within the EEA grid reference system. The cube contains 149 grid cells covering coordinates from (4557000, 1695000) to (5284000, 3479000), with a total of 141,168 observations, representing 482 species.

3 Data Variability and Robustness

3.1 Measures for Data Cube Robustness

We aim to develop general robustness measures for data cubes to assess their applicability, either in general or in relation to specific biodiversity indicators. These measures evaluate data quality across different dimensions of the cube and support exploratory data analysis:

- Spatial
  - Data Distribution: Is the data clustered in specific areas or evenly spread across the region?
  - Geographical Coverage: Are species confined to small areas, or are they widespread throughout the region?
- Temporal
  - Temporal Variation: How do species occurrences change over time? Are there sufficient observations over different time periods?
  - Comparative Stability: How do these occurrences compare to a higher taxonomic level over time?
- Taxonomical
  - Prevalence or Abundance: How do species prevalence or abundance affect multispecies indices? Are certain species overrepresented?

To implement this, we propose two functionalities. Section 5 provides a more general description of software development and implementation.

1. **Measuring Data Cube Robustness**
   - Function(s) that measure the applicability of a dataset in general, or related to a certain biodiversity indicator
     i. **Input:** occurrence cube of class ‘processed_cube’, indicator function when using leave-one-out cross-validation (see further)
     ii. **Output:** summary with warning/flag system (e.g. with colours) and a short description

These measures support data exploration and will be categorized into different robustness levels:

category	
Robustness summary	
Note	
Important note	
Very important note	

Preliminary rules are provided in Annex 8.1. Further refinements will be made in parallel with the results from data quality assessments within the B3 project (T4.5).

2. **Filtering Data Cube Observations**
    - Function(s) that filter a dataset based on default or custom rules
        i. **Input:** occurrence cube of class 'processed_cube'
        ii. **Output:** filtered occurrence cube

To enhance data exploration, this filtering function allows users to refine datacubes using robustness criteria. Default settings will be provided—for instance, trend analysis over multiple years is unlikely to be meaningful for a bird observed only once in a single year. Such rare sightings may be anomalies rather than a true trend for that species in that region.

3.2 Measures for Species Robustness

For indicators developed across multiple species, we developed a cross-validation (CV) technique to assess how each species impacts the overall calculation of an indicator. It is a leave-one-out method, where each species is systematically excluded from the dataset and the indicator is recalculated without that species. We call this method leave-one-species-out cross-validation (LOSO-CV), which provides insights into the robustness of the indicator and helps guide users in interpreting results, particularly when specific species might disproportionately affect the outcomes.

1. Original Sample Data: $\mathbf{X} = \{X_{11}, X_{12}, X_{13}, \ldots, X_{sn}\}$

- The initial set of data points, where there are s different species and n total sample size across all species. n corresponds to the number of cells in a data cube or the number of rows in tabular format.

2. Statistic of Interest: θ

- The parameter or statistic being estimated, such as the mean $\bar{X}$, variance σ^2, or a biodiversity indicator. Let $\hat{\theta}$ denote the estimated value of θ calculated from the complete dataset $\mathbf{X}$.

3. Cross-Validation (CV) Sample: $\mathbf{X}_{-s_j}$

- The full dataset $\mathbf{X}$ excluding all occurrences belonging to species j. This subset is used to investigate the influence of species j on the estimated statistic $\hat{\theta}$.

4. CV Estimate for Species j: $\hat{\theta}_{-s_j}$

- The value of the statistic of interest calculated from $\mathbf{X}_{-s_j}$, which excludes species j. For example, if θ is the sample mean, $\hat{\theta}_{-s_j} = \bar{X}_{-s_j}$.

5. Error Measures:
Error: Error_{s_j}

- The difference between the statistic estimated without species j ($\hat{\theta}_{-s_j}$) and the statistic calculated on the complete dataset ($\hat{\theta}$).

$$\text{Error}_{s_j} = \hat{\theta}_{-s_j} - \hat{\theta}$$

Relative Error: Rel. Error_{s_j}

- The absolute error, normalised by the true estimate $\hat{\theta}$ and a small error term $\epsilon = 10^{-8}$ to avoid division by zero.

$$\text{Rel. Error}_{s_j} = \frac{|\hat{\theta}_{-s_j} - \hat{\theta}|}{\hat{\theta} + \epsilon}$$

7. Summary Measures:
Mean Relative Error: MRE

- The average of the relative errors over all species.

$$\text{MRE} = \frac{1}{s}\sum_{j=1}^{s} \text{Rel. Error}_{s_j}$$

Mean Squared Error: MSE

- The average of the squared errors.

$$\text{MSE} = \frac{1}{s}\sum_{j=1}^{s} (\text{Error}_{s_j})^2$$

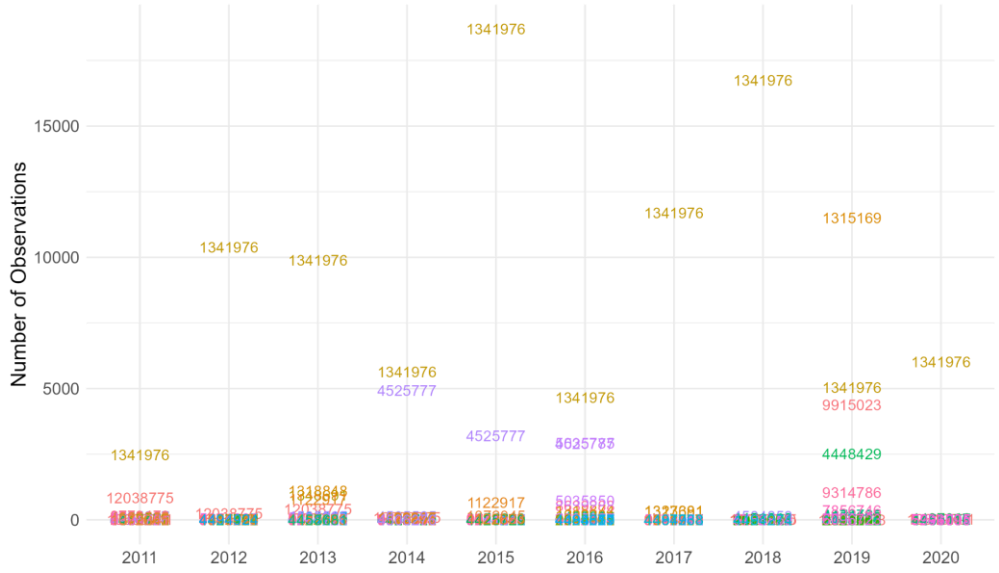
Root Mean Squared Error: RMSE

- The square root of the MSE:

$$\text{RMSE} = \sqrt{\text{MSE}}$$

As an example, we calculated Pielou's Evenness using the **b3gbi** package v0.2.2 (function `b3gbi::pielou_evenness_ts()`). Evenness is a useful indicator because the influence of a single species can make a big difference to the indicator value. Higher evenness values indicate a more balanced community (a value of 1 means that all species are equally abundant), while low values indicate a more unbalanced community (a value of 0 means that one species dominates completely). The influence of a species on this indicator is therefore related to the number of observations for that species compared to the number of observations for each of the other species (Fig. 1A). Figure 1B shows the LOSO-CV error for each species each year. An error value of 0.2 for species 1341976 (= GBIF key for European honey bee, *Apis mellifera* Linnaeus, 1758), for example, means that the indicator value is 0.2 higher without that species than when the species is included.

A.



B.

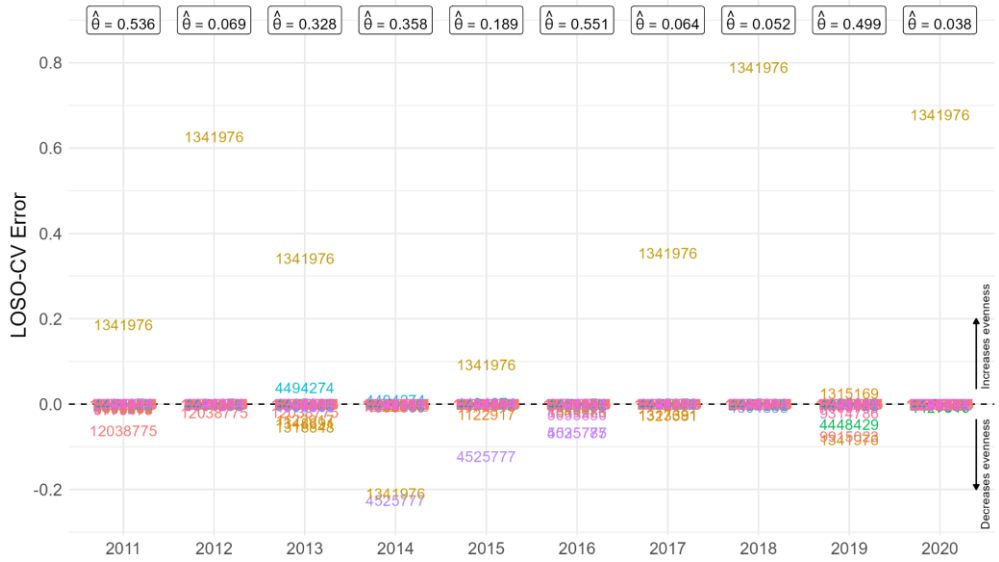


Figure 1: Visualisation of the influence of each species on the calculation of the indicator. The numbers indicate the GBIF species keys. A: Number of occurrences for each species in the dataset per year. B: LOSO-CV error for each species each year. Note that the species key in B means that the species was not included (see the report text for explanation).

When one dominant species is present, removal of that species increases the evenness value (positive error value for that species), e.g., species 1341976 in 2011, 2012, 2013, 2015, 2017, 2018 and 2020. This is because when the dominant species is removed, the community becomes more balanced. When two equally dominant species are present, removing each of the species decreases the evenness value (negative error value for that species), e.g., species 1341976 and 4525777 (red-belted clearwing, *Synanthedon myopaeformis* (Borkhausen, 1789)) in 2014. This is because when one of the two species is removed, the other becomes the only dominant species

and evenness decreases. When there are multiple dominant species present, removing each of the species has only a small effect on the evenness value, e.g., in 2016 and 2019. This is because when one of the species is removed, there are still other dominant species and evenness does not change much.

This example demonstrates that the LOSO-CV method functions as expected. It serves as an exploratory tool for assessing the influence of individual species on indicator calculations. In 2012, we observe a notably low evenness of 0.069, indicating the dominance of a single species. But which species is responsible? And is this expected? One possible explanation is dataset bias—some datasets included in the cube may disproportionately focus on certain species (see further). However, this does not imply that species with large errors should be excluded. In this case, species 1341976 had significantly more observations in 2012 than others, making the low evenness value expected. Whether this is problematic depends on the research question.

For context, the dominant species, 1341976, is the European honey bee (*Apis mellifera*). One might expect honeybees to be less dominant compared to smaller, less conspicuous insects like ants or aphids. However, *A. mellifera* appears to dominate the dataset, likely because it is easier to spot, identify, and study. This highlights a clear bias and a lack of robustness in the data cube.

While LOSO-CV is a valuable approach, other leave-one-out methods could also provide important insights. For example, “leave-one-dataset-out cross-validation” could help assess how individual datasets influence indicator calculations. This would be particularly useful for identifying biases introduced by certain data sources (Ferro & Flick, 2015). Combining LOSO-CV with similar techniques can provide a more comprehensive understanding of data reliability and robustness in biodiversity assessments.

4 Indicator Variability and Robustness

4.1 Measures for Indicator Uncertainty

To quantify indicator uncertainty, we followed the guidelines of (Rowland et al., 2021) that advise the selection of an appropriate approach to present uncertainty in biodiversity indicators (Fig. 2). On the basis of these, we propose the calculation of confidence intervals for indicators calculated from occurrence cubes based on the bootstrap resampling method (hereafter, bootstrapping).

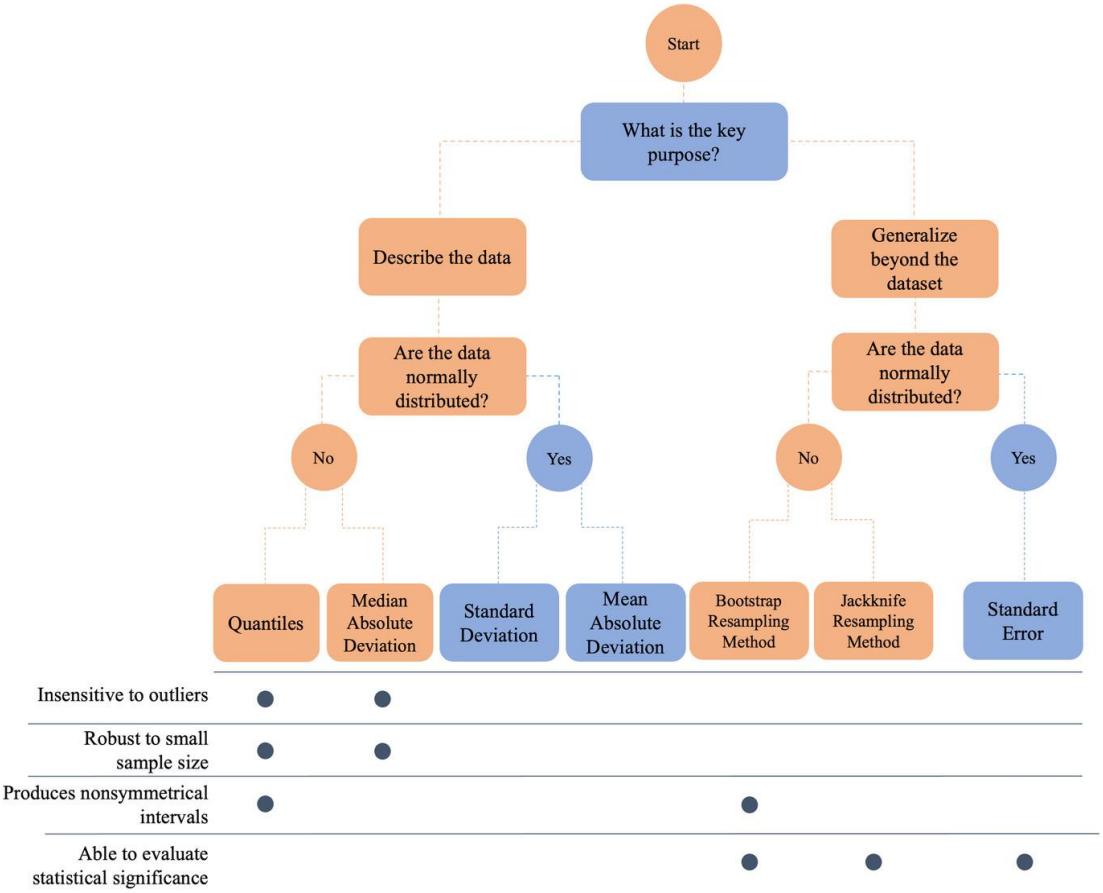


Figure 2: Figure 1 from Rowland et al. (2021): "Decision tree for presenting data variability or uncertainty in biodiversity indicators with interval methods. This list of approaches is not exhaustive."

Bootstrapping provides a flexible, non-parametric approach to estimate the variability of indicators without relying on strong assumptions about the underlying data distribution (Davison & Hinkley, 1997; Efron & Tibshirani, 1994). This flexibility is meaningful for this research, as both occurrence cube datasets and the derived indicators can be very distinct in nature (Dixon, 2001). Each occurrence cube has a unique spatial, temporal, taxonomic, … scope, and spatial and temporal indicators are developed related to prevalence, abundance, phylogenetic diversity, impact of alien species, etc. (Breugelmans et al., 2024; Dove, 2024; Yahaya & Kumschick, 2025). In bootstrapping, the dataset is resampled multiple times with replacement. For each resampled dataset, the statistic of interest, e.g. a biodiversity indicator, is calculated. Based on the

distribution of these indicator values, we get an idea about indicator uncertainty, e.g. via calculation of confidence intervals (Fig. 3).

1. Original Sample Data: $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$

- The initial set of data points. Here, n is the sample size. This corresponds to the number of cells in a data cube or the number of rows in tabular format.

2. Statistic of Interest: θ

- The parameter or statistic being estimated, such as the mean $\bar{X}$, variance σ^2, or a biodiversity indicator. Let $\hat{\theta}$ denote the estimated value of θ calculated from the dataset $\mathbf{X}$.

3. Bootstrap Sample: $\mathbf{X_b}^* = \{X_1^*, X_2^*, \ldots, X_n^*\}$

- A sample of size n drawn with replacement from the original sample $\mathbf{X}$. Each X_i^* is drawn independently from $\mathbf{X}$.
- A total of B bootstrap samples are drawn from the original data. Common choices for B are 1000 or 10,000 to ensure a good approximation of the distribution of the bootstrap replications (see further).

4. Bootstrap Replication: $\hat{\theta}_b^*$

- The value of the statistic of interest calculated from the b-th bootstrap sample $\mathbf{X}_b^*$. For example, if θ is the sample mean, $\hat{\theta}_b^* = \bar{X}_b^*$.

5. Bootstrap statistics

Bootstrap Estimate of the Statistic: $\hat{\theta}_{\text{boot}}$

- The average of the bootstrap replications:

$$\hat{\theta}_{\text{boot}} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b^*$$

Bootstrap Bias: $\text{Bias}_{\text{boot}}$

- This bias indicates how much the bootstrap estimate deviates from the original sample estimate. It is calculated as the difference between the average bootstrap estimate and the original estimate:

$$\text{Bias}_{\text{boot}} = \frac{1}{B} \sum_{b=1}^{B} (\hat{\theta}_b^* - \hat{\theta}) = \hat{\theta}_{\text{boot}} - \hat{\theta}$$

Bootstrap Standard Error: SE_{boot}

- The standard deviation of the bootstrap replications, which estimates the variability of the statistic.

Bootstrap Confidence Interval: CI_{boot}

- Confidence intervals (CIs) for the statistic of interest can be constructed using the bootstrap distribution of $\hat{\theta}^*$. Several methods are explained in more detail below.

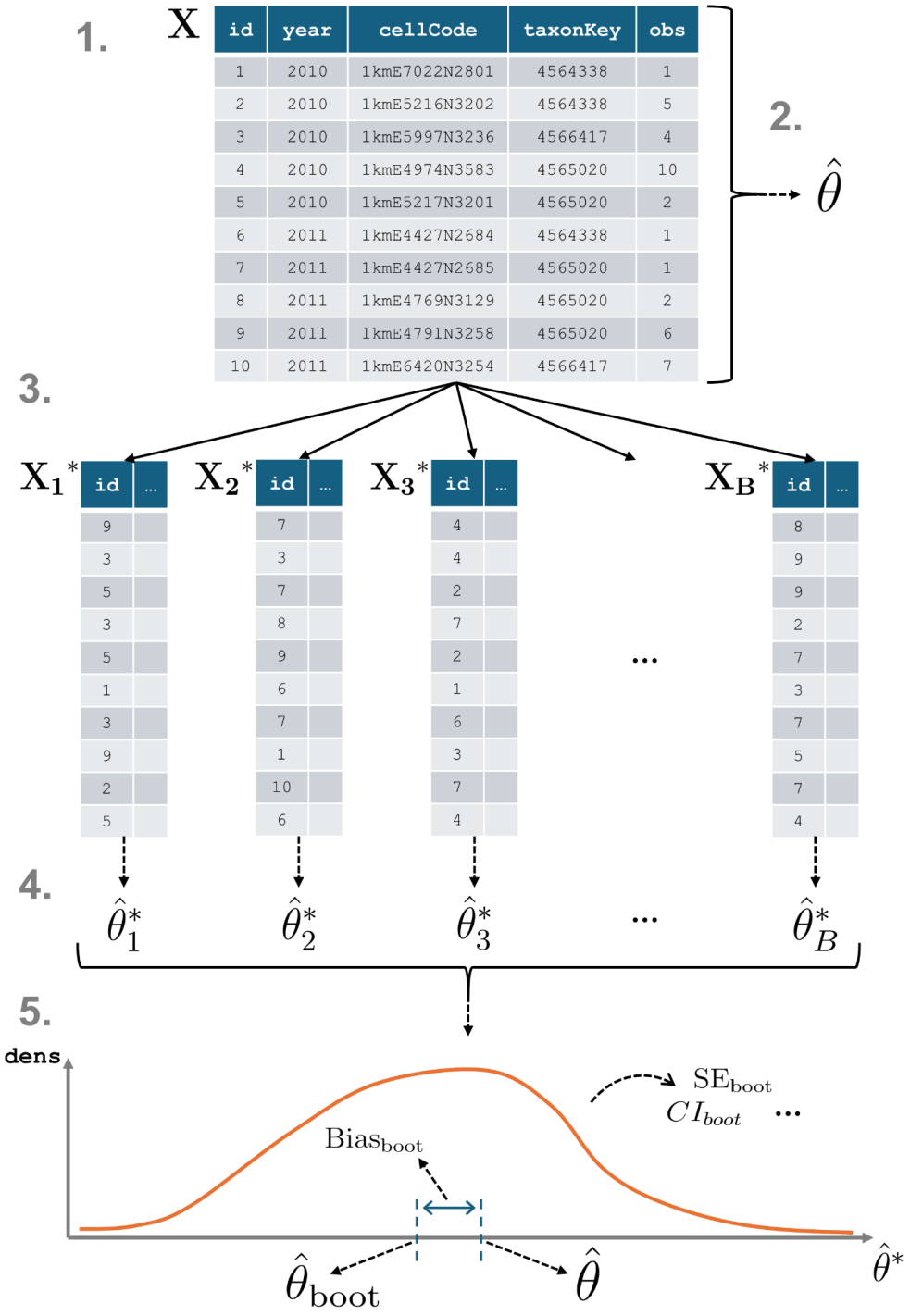


Figure 3: Use of bootstrapping for the quantification of indicator uncertainty. See the report text for an explanation of the mathematical notation.

Bootstrapping can be used to assess the uncertainty about an indicator estimate via confidence intervals (Davison & Hinkley, 1997; Dixon, 2001; Rowland et al., 2021). We consider four different types of intervals (with confidence level α). The choice for confidence interval types and their calculation is in line with the **boot** package in R (Canty & Ripley, 1999) to ensure smooth implementation in software (see further). They are based on the definitions provided by Davison & Hinkley (1997, Chapter 5) (see also DiCiccio & Efron, 1996; Efron, 1987).

1. Normal: Assumes the bootstrap distribution of the statistic is approximately normal

$$CI_{norm} = \left[\hat{\theta} - \text{Bias}_{\text{boot}} - \text{SE}_{\text{boot}} \times z_{1-\alpha/2}, \hat{\theta} - \text{Bias}_{\text{boot}} + \text{SE}_{\text{boot}} \times z_{1-\alpha/2}\right]$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

2. Basic: Centers the interval using percentiles

$$CI_{basic} = \left[2\hat{\theta} - \hat{\theta}^*_{(1-\alpha/2)}, 2\hat{\theta} - \hat{\theta}^*_{(\alpha/2)}\right]$$

where $\hat{\theta}^*_{(\alpha/2)}$ and $\hat{\theta}^*_{(1-\alpha/2)}$ are the $\alpha/2$ and $1-\alpha/2$ percentiles of the bootstrap distribution, respectively.

3. Percentile: Uses the percentiles of the bootstrap distribution.

$$CI_{perc} = \left[\hat{\theta}^*_{(\alpha/2)}, \hat{\theta}^*_{(1-\alpha/2)}\right]$$

where $\hat{\theta}^*_{(\alpha/2)}$ and $\hat{\theta}^*_{(1-\alpha/2)}$ are the $\alpha/2$ and $1-\alpha/2$ percentiles of the bootstrap distribution, respectively.

4. Bias-Corrected and Accelerated (BCa): Adjusts for bias and acceleration

Bias refers to the systematic difference between the observed statistic from the original dataset and the center of the bootstrap distribution of the statistic as introduced above. The bias correction term is calculated as follows:

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#(\hat{\theta}^*_b < \hat{\theta})}{B}\right)$$

where $\#$ is the counting operator and Φ^{-1} the inverse cumulative density function of the standard normal distribution.

Acceleration quantifies how sensitive the variability of the statistic is to changes in the data.

- $a = 0$: The statistic's variability does not depend on the data (e.g., symmetric distribution)
- $a > 0$: Small changes in the data have a large effect on the statistic's variability (e.g., positive skew)
- $a < 0$: Small changes in the data have a smaller effect on the statistic's variability (e.g., negative skew).

The acceleration term is calculated as follows:

$$\hat{a} = \frac{1}{6} \frac{\sum_{i=1}^{n} (I_i^3)}{(\sum_{i=1}^{n} (I_i^2))^{3/2}}$$

where I_i denotes the influence of data point x_i on the estimation of θ. I_i can be estimated using jackknifing. Examples are (1) the negative jackknife: $I_i = (n-1)(\hat{\theta} - \hat{\theta}_{-i})$, and (2) the positive jackknife $I_i = (n+1)(\hat{\theta}_{-i} - \hat{\theta})$ (Frangos & Schucany, 1990). Here, $\hat{\theta}_{-i}$ is the estimated value leaving out the i'th data point x_i. The **boot** package also offers infinitesimal jackknife and regression estimation (Canty & Ripley, 1999). The different jackknife options can be explored in the future.

The bias and acceleration estimates are then used to calculate adjusted percentiles.

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{\alpha/2}}{1 - \hat{a}(\hat{z}_0 + z_{\alpha/2})}\right), \quad \alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha/2}}{1 - \hat{a}(\hat{z}_0 + z_{1-\alpha/2})}\right)$$

So, we get

$$CI_{bca} = \left[\hat{\theta}^*_{(\alpha_1)}, \hat{\theta}^*_{(\alpha_2)}\right]$$

4.2 Comparison of Interval Types

We compared all four interval types mentioned above in the case of three different general biodiversity indicators ($\alpha = 0.05$). The following indicators were selected:

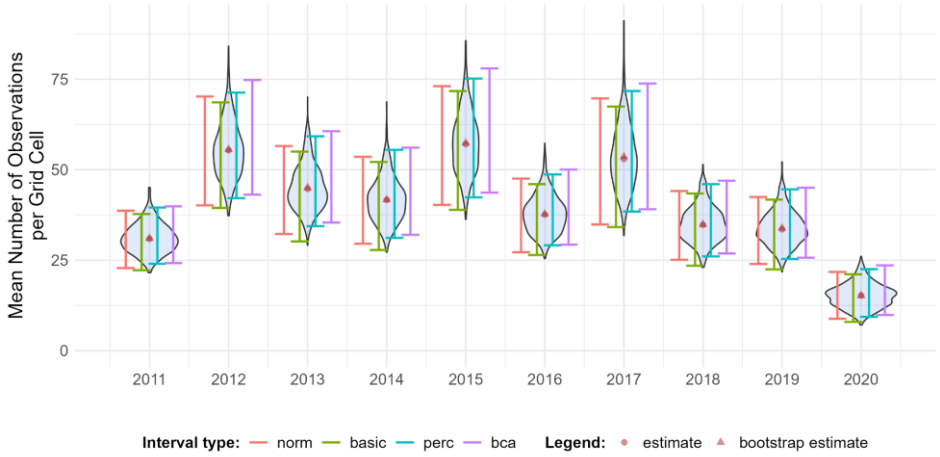
1. Mean Number of Observations per Grid Cell
    - Custom function
    - Calculates the mean number of observations per grid cell per year
    - Positive real number
2. Pielou's Evenness
    - As calculated by the **b3gbi** package v0.2.2: `b3gbi::pielou_evenness_ts()`
    - Calculates the evenness per year
    - Real number between 0 and 1

3. Observed Species Richness
    - As calculated by the **b3gbi** package v0.2.2: `b3gbi::obs_richness_ts()`
    - Calculates the number of species per year
    - Positive integer

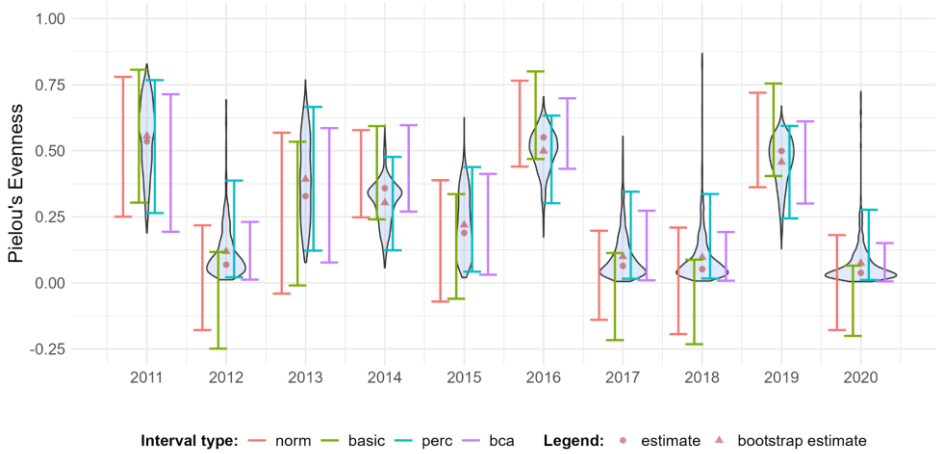
(1) The Mean Number of Observations per Grid Cell show very low bootstrap bias and more or less symmetrical bootstrap distributions (Fig. 4A). Therefore, the four interval types are similar. (2) Pielou's Evenness shows moderate bias and a strong skewness in the bootstrap distributions unless values are around 0.5 (Fig. 4B). Therefore, the symmetric intervals, normal and basic, are not recommended. Truncation or the use of transformation functions (in this case logit and expit) could allow these intervals to be limited to 0-1 range of evenness. These options will be considered in the future. The percentile interval can be used for asymmetrical distributions, but does not take bias into account. Therefore, the BCa interval is recommended in this case. (3) The Observed Species Richness shows high bias and more or less symmetrical bootstrap distributions (Fig. 4C). The BCa interval could not be calculated when all bootstrap replications are lower than the original estimate (see formula $\hat{z}_0$). None of the calculated intervals cover the original estimate. This is the case because bootstrap resampling will never introduce new species (Dixon, 2001, p. 287). Alternative approaches should be used in this case. For example, the **vegan** (Oskanen et al., 2024) and **iNEXT** (Hsieh et al., 2016) packages in R offer alternative species richness indices with uncertainty calculation. Some of those are already implemented for occurrence cubes in a newer version of **b3gbi** (Dove, 2024, v0.4.0).

Because of the wide range of indicator types, we recommend the use of percentile or BCa intervals, because they have no strong assumptions regarding the bootstrap distribution. The BCa interval is recommended as it accounts for bias and skewness. However, due to the jackknife estimation of the acceleration parameter, the calculation time is significantly longer. The use of the normal and basic confidence intervals is not recommended, but could be used in combination with truncations or transformations. The assumption of normality can be checked by making a Q-Q plot of the bootstrap replications (Davison & Hinkley, 1997). An overview of the recommendations is provided in Table 1. This is not an exhaustive review of the topic, but based on existing literature and our preliminary results, these recommendations provide a useful starting point for selecting appropriate interval types.

A.



B.



C.

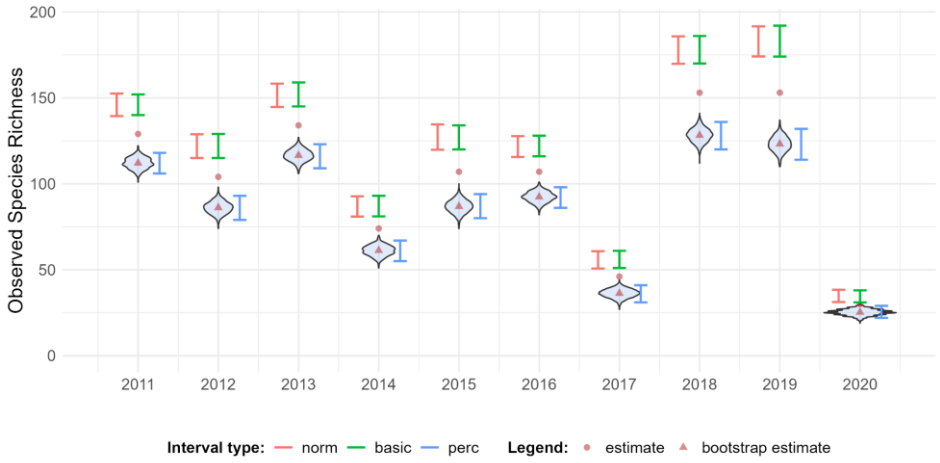


Figure 4: Comparison of different interval types for three different indicators. The violin plots show the bootstrap distributions. The estimate refers to the original estimate $\hat{\theta}$. A: Mean Number of Observations per Grid Cell. B: Pielou's Evenness. C: Observed Species Richness.

Table 1: Overview of the advantages and disadvantages between the considered confidence interval types.

Interval type	Advantages	Disadvantages	References
Normal	- Simplicity - Understanding of bootstrap and CI theory	- Assumes bootstrap distribution is normal - Often, transformation needed for more accurate results - Erratic coverage error in practise	(Davison & Hinkley, 1997, Chapter 5; Efron & Tibshirani, 1994, Chapter 13; Hesterberg, 2015)
Basic	- Simplicity - Understanding of bootstrap and CI theory	- Assumes symmetric bootstrap distribution - Typically substantial coverage error	(Carpenter & Bithell, 2000; Davison & Hinkley, 1997, Chapter 5; Hesterberg, 2015)
Percentile	- Simplicity - No assumptions about bootstrap distribution - Implicitly uses the existence of a good transformation	- Does not take bias into account - Substantial coverage error if the distribution is not nearly symmetric	(Carpenter & Bithell, 2000; Davison & Hinkley, 1997, Chapter 5; Efron & Tibshirani, 1994, Chapter 13)
BCa	- No assumptions about bootstrap distribution - Implicitly uses the existence of a good transformation - Adjusts for bias - Adjusts for skewness - Smaller coverage error than the other methods	- Involved calculation of acceleration parameter a - Unstable coverage when a is small ($a < 0.025$) and for small sample sizes	(Carpenter & Bithell, 2000; Davison & Hinkley, 1997, Chapter 5; Dixon, 2001; Efron & Tibshirani, 1994, Chapter 14)

4.3 Interpretation of Indicator Uncertainty

In interpreting indicators, too much emphasis is sometimes placed on relatively small differences in indicator values. Small differences are to be expected due to natural variability and limited sample size. To avoid over-interpretation of changes in indicator values, it is important to include the confidence intervals around the calculated values in the final evaluation. This section presents a general approach to interpreting effects. The aim is to provide a general framework for interpreting changes (increases or decreases) for the very different indicators developed and implemented under the B3 project.

The approach is presented below using Pielou's evenness index with the BCa intervals from the previous subsection. To create a more continuous representation of change over time, we can apply LOESS (Locally Estimated Scatterplot Smoothing) to the estimates and confidence limits. This smoothing technique fits local regressions across subsets of the data, producing a flexible trend line that helps visualize broader patterns while retaining important details. However, for the example of Pielou's Evenness, this figure does not provide a clear handle for making a statement

about whether the indicator value is increasing or decreasing (Fig. 5). An alternative option is using a fan plot based on the standard error of the estimates. This is for example implemented in the **effectclass** package (Onkelinx, 2023) and will be further developed in the future.

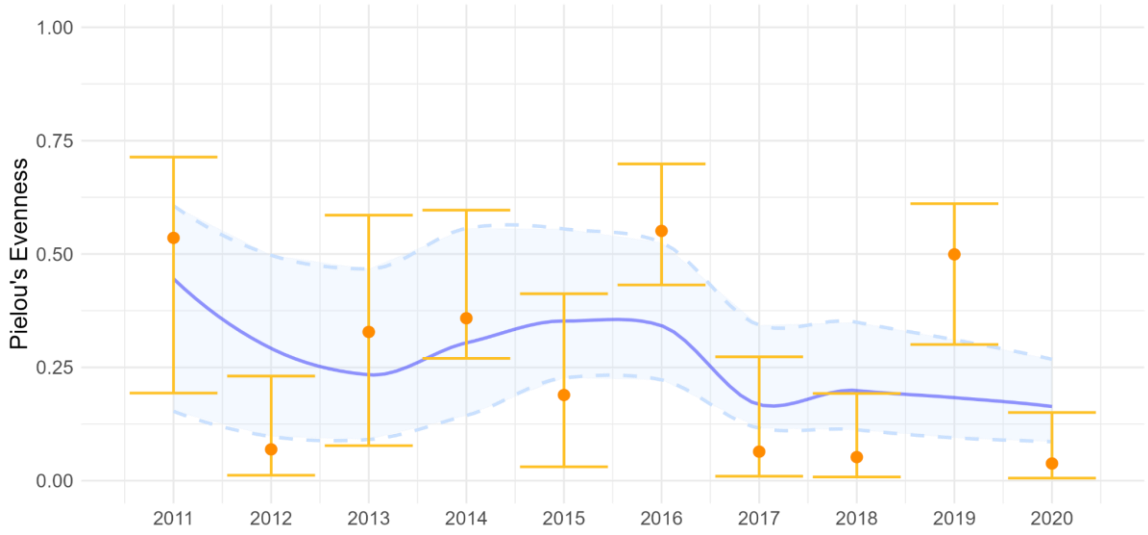


Figure 5: Trend visualisation using LOESS smoothers based on the estimates and the confidence limits.

Interpretation of the results can be done using effect classification based on the **effectclass** package. If the confidence interval is above/under the reference line, we call it an increase/decrease. Threshold values can be provided to make a distinction between stable (confidence interval covering the reference line) and uncertain trends (confidence interval covering the reference line and at least one of the threshold lines). The thresholds can also be used to classify even further (Table 2, Fig. 6). Threshold values should be manually selected around the reference line at a level deemed negligible, allowing trends within this range to be classified as 'no effect'.

Table 2: Rules for effect classification to aid the interpretation of indicator effects/trends (Onkelinx, 2023). Continues on the next page.

Symbol	Fine effect/trend	Coarse effect/trend	Rule
++	strong positive effect/strong increase	positive effect/ increase	confidence interval above the upper threshold
+	positive effect/ increase	positive effect/ increase	confidence interval above reference and contains the upper threshold
+~	moderate positive effect/ moderate increase	positive effect/ increase	confidence interval between reference and the upper threshold

Table 2 continued

Symbol	Fine effect/trend	Coarse effect/trend	Rule
~	no effect/stable	no effect/stable	confidence interval between thresholds and contains reference
-~	moderate negative effect/moderate decrease	negative effect/ decrease	confidence interval between reference and the lower threshold
-	negative effect/ decrease	negative effect/ decrease	confidence interval below reference and contains the lower threshold
--	strong negative effect/strong decrease	negative effect/ decrease	confidence interval below the lower threshold
?+	potential positive effect/potential increase	unknown effect/ unknown	confidence interval contains reference and the upper threshold
?-	potential negative effect/potential decrease	unknown effect/ unknown	confidence interval contains reference and the lower threshold
?	unknown effect/ unknown	unknown effect/ unknown	confidence interval contains the lower and upper threshold

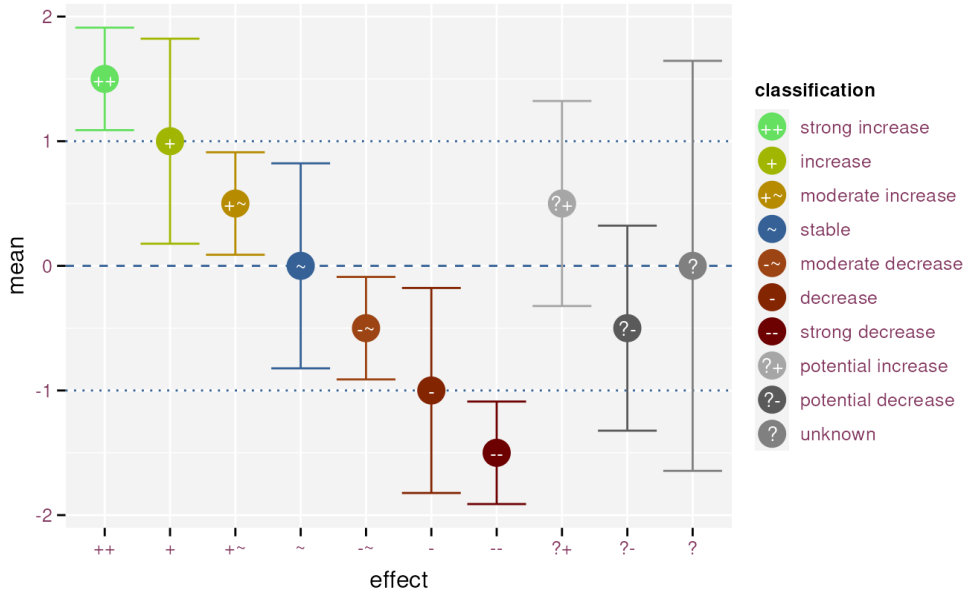
We can compare these to a certain constant. For example, we are interested to see whether evenness is significantly different from 0.5, where we choose arbitrary thresholds of 0.1 (Fig. 7A). We see that evenness is significantly lower than 0.5 in 2012, 2015, 2017, 2018 and 2020. In 2012, 2017, 2018 and 2020, this is a strong difference (under the lower threshold). In all other cases, there is an uncertain trend.

Perhaps more interesting is the comparison with a reference group. In the case of a time series, we might want to compare indicator values with the value for the last year (or any reference period, see further). For this, we need to bootstrap again. This time over the difference between indicator functions:

1. Resample the dataset with replacement
2. Calculate the indicator for each period (e.g., year)
3. Take the difference between the indicator values for each non-reference period with the indicator value for the reference period
4. Repeat steps 1-3 B times

Steps 1, 2 and 4 are the same as before. The difference is that we now get bootstrap replicate distributions for a difference between indicator values (Step 3). From these distributions, we can again calculate the confidence intervals as defined previously. The reference value for effect classification is now typically equal to 0 (no difference with the reference year 2020). We choose arbitrary thresholds of 0.2 (Fig. 7B). We see that evenness was significantly higher in 2014, 2016, and 2019 compared to 2020. In all other cases, there is an uncertain trend.

A.



B.

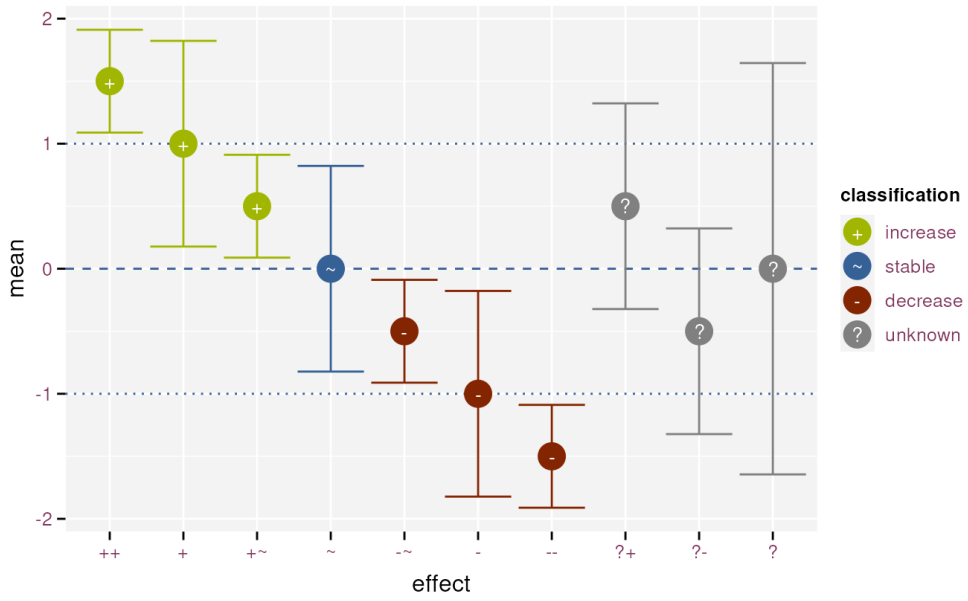
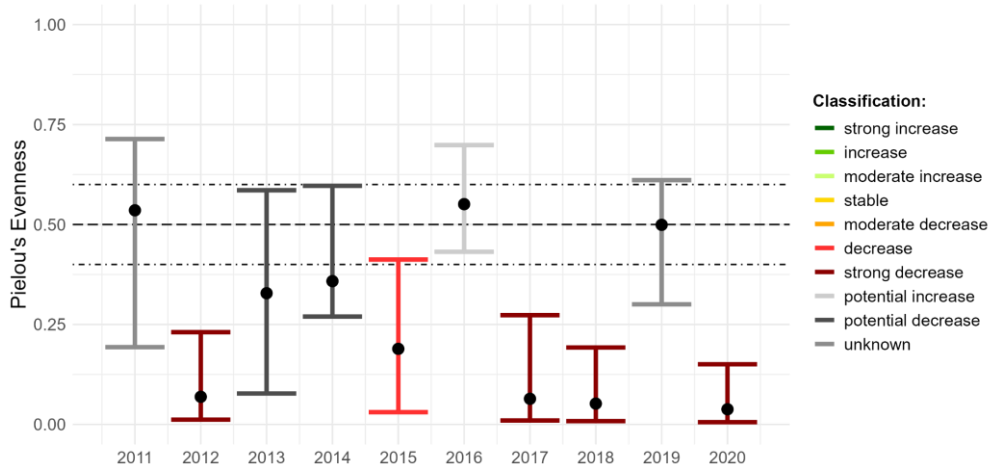


Figure 6: Visualisation of effect classification. A: Fine classification. B: Coarse classification. Figures from Onkelinx (2023): https://inbo.github.io/effectclass/articles/visualisation.html.

A.



B.

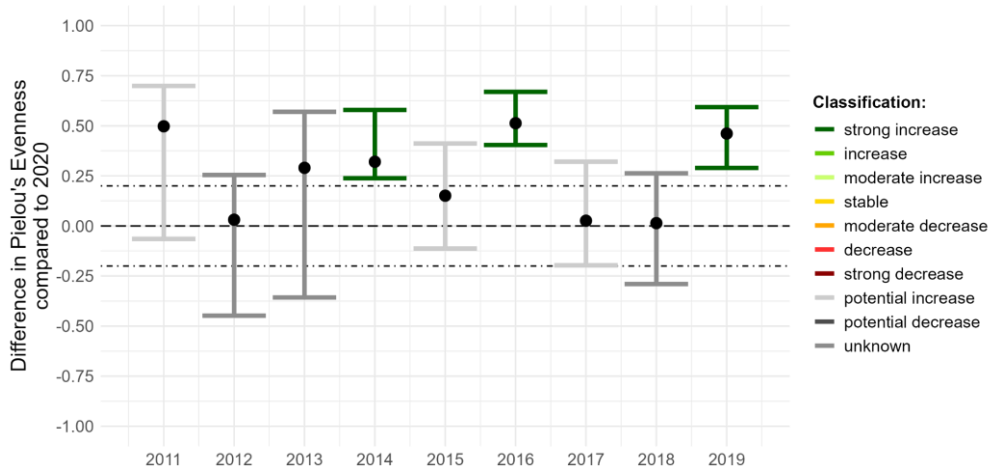


Figure 7: Visualisation and interpretation for the percentile intervals of Pielou's Evenness. A: Comparison with a constant (0.5 ± 0.1). B: Comparison with a reference period (2020: 0 ± 0.2).

Note that the choice of the reference year should be well considered. Keep in mind which comparisons should be made, and what the motivation is behind the reference period. A high or low value in the reference period relative to other periods, e.g. an exceptional bad or good year, can affect the magnitude and direction of the calculated differences. Whether this should be avoided or not, depends on the motivation behind the choice and the research question. A reference period can be determined by legislation, or by the start of a monitoring campaign. A specific research question can determine the periods that need to be compared. Furthermore, the variability of the estimate of reference period affects the width of confidence intervals for the differences. A more variable reference period will propagate greater uncertainty. In the case of GBIF data, more data will be available in recent years than in earlier years. If this is the case, it could make sense to select the last period as a reference period as done in our example above. In a way, this also avoids the arbitrariness of choice for the reference period. You compare previous situations with the current situation (last year), where you could repeat this comparison

annually, for example. Note that it will require some caution/adaptation for the interpretation of the trend classification in time-series where 'increase' suggests an increase with time and not backwards. Finally, when comparing multiple indicators, we recommend using a consistent reference period to maintain comparability.

In the case of the BCa interval, if we compare with a reference group, we need to estimate the acceleration using jackknifing in a different way than before. Consider $\hat{\theta} = \hat{\theta}_1 - \hat{\theta}_2$, where $\hat{\theta}_1$ is the estimate for the indicator value of a non-reference period (sample size is n_1) and $\hat{\theta}_2$ is the estimate for the reference period (sample size is n_2). The acceleration term is now calculated as follows:

$$\hat{a} = \frac{1}{6} \frac{\sum_{i=1}^{n_1+n_2} (I_i^3)}{(\sum_{i=1}^{n} (I_i^2))^{3/2}}$$

Remember I_i can be calculated using the negative jackknife: $I_i = (n-1)(\hat{\theta} - \hat{\theta}_{-i})$, or the positive jackknife $I_i = (n+1)(\hat{\theta}_{-i} - \hat{\theta})$. Such that

$\hat{\theta}_{-i} = \hat{\theta}_{1,-i} - \hat{\theta}_2$ for $i = 1, \ldots, n_1$, and
$\hat{\theta}_{-i} = \hat{\theta}_1 - \hat{\theta}_{2,-i}$ for $i = n_1 + 1, \ldots, n_1 + n_2$

4.4 Calculation and Interpretation of Uncertainty for Spatial Indicators

So far we only focussed on temporal indicators, e.g. per year, but indicators are also calculated over a spatial extent (Fig. 8). The methods discussed above have not been tested on spatial indicators. However, bootstrapping should be analogous to the examples above. Instead of grouping by year, we can group by grid cell code (or by species and year, or species and grid cell code for species-specific indicators). The confidence intervals can then be derived as before. The main difference will be visualisation of uncertainty and the way we can use effect classification.

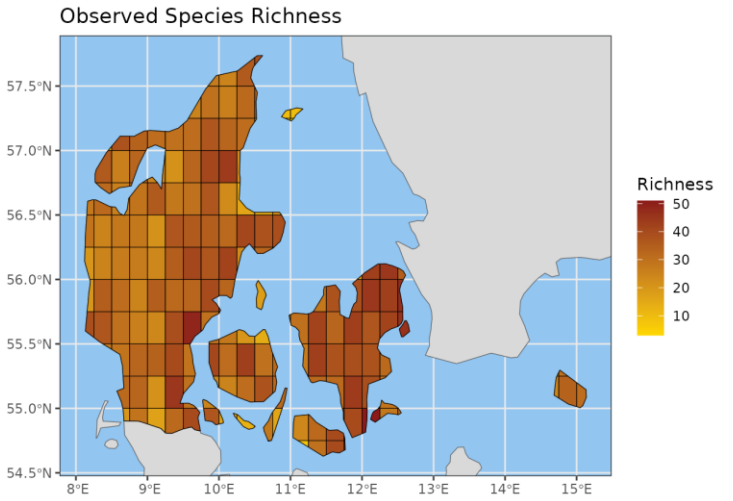


Figure 8: Spatial indicator (Observed Species Richness) for the mammals of Denmark. Figure from Dove (2024): https://b-cubed-eu.github.io/b3gbi/articles/b3gbi.html.

To visualise the uncertainty from confidence intervals, we can map the CI width, the bootstrap standard error or a relative measure like the CI width divided by two times the estimate, where larger values indicate greater uncertainty. Alternatively, we can create separate maps for the lower and upper CI bounds. For visualising both the estimate and uncertainty in a single map, we can use circles within the grid cells that vary in blur (Figs 9A-B, best w.r.t. user intuitiveness), or transparency (Figs 9C-D, best w.r.t. user performance ~ accuracy, speed) (Kinkeldey et al., 2014; MacEachren et al., 2005, 2012).

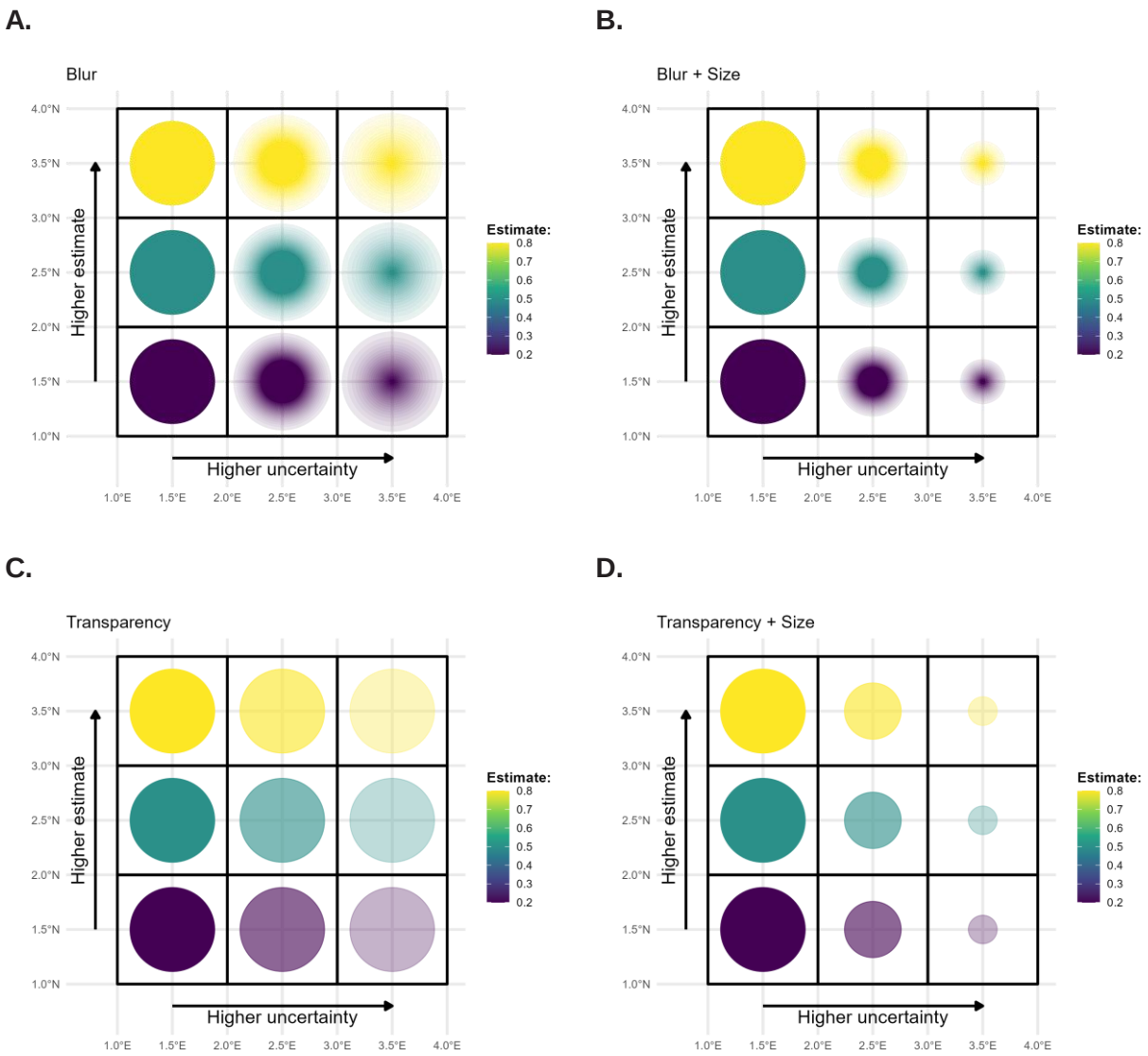


Figure 9: Visualisation of indicator estimate and uncertainty within a spatial grid. A: Blur. B: Blur and size. C: Transparency. D: Transparency and size. Created using R packages ggplot2 (Wickham, 2016), dplyr (Wickham et al., 2023), sf (Pebesma, 2018) and ggblur (mikefc, 2025) (code provided in Langeraert et al., 2025, v1.4.0).

For effect classification, we can use the same classification technique as before, where we can compare with a constant value (mean or reference value). For visualisation, the grid cells can be coloured according to the effect. Comparison with a reference group (in this case a reference grid cell) may be less useful than for temporal indicators, but is possible in a similar way.

5 Software Implementation

The code for calculating the robustness measures, indicator uncertainty via bootstrapping, and effect classification will be bundled in an R package called **dubicube** (Fig. 10) (Langeraert & Van Daele, 2025). The functions in this package can be used for exploratory analyses (Section 3) as well as uncertainty calculation and interpretation (Section 4). It can also serve as a dependency for packages calculating indicators from occurrence cubes, e.g. to retrieve confidence intervals.

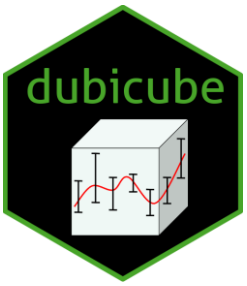


Figure 10: The logo of the dubicube R package.

We plan to explore the use of the **boot** package (Canty & Ripley, 1999) for implementing bootstrapping functionalities in the **dubicube** package. In this way, we will ensure correct programming of bootstrap resampling and confidence level calculation. However, preliminary implementation tests have already shown that modifications to the original **boot** code will be required to make it compatible with the structure and specific requirements of occurrence cubes. This will likely involve adapting or extending the functionality to handle cube-based data effectively.

Additionally, since bootstrapping and jackknife methods can be computationally intensive, we will need to focus on optimising the code for performance. This includes investigating the use of parallel processing options, which could significantly reduce computational time and make the implementation more efficient for large datasets. Some options are also provided by the **boot** package.

6 Acknowledgements

We would like to thank Peter Desmet for coming up with the R package name for **dubicube**. Emma Cartuyvels, Thierry Onkelinx, Shawn Dove, and Lissa Breugelmans provided valuable comments that helped improve the text. Additionally, we appreciate the input of Emma Cartuyvels, Toon Westra, Floris Vanderhaeghe, and Shawn Dove regarding the visualisation of spatial uncertainty.

7 References

Breugelmans, L., Trekels, M., & Hendrickx, L. (2024). *pdindicatoR: Calculate and visualize phylogenetic diversity indicators based on species occurrence data cubes* [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.14237551

Canty, A., & Ripley, B. (1999). *boot: Bootstrap Functions (Originally by Angelo Canty for S)* [Computer software]. https://CRAN.R-project.org/package=boot

Carpenter, J., & Bithell, J. (2000). Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, *19*(9), 1141–1164. https://doi.org/10.1002/(SICI)1097-0258(20000515)19:9<1141::AID-SIM479>3.0.CO;2-F

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap Methods and their Application* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511802843

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, *11*(3). https://doi.org/10.1214/ss/1032280214

Dixon, P. M. (2001). The Bootstrap and the Jackknife: Describing the Precision of Ecological Indices. In S. M. Scheiner & J. Gurevitch (Eds.), *Design and Analysis of Ecological Experiments* (Second Edition, pp. 267–288). Oxford University PressNew York, NY. https://doi.org/10.1093/oso/9780195131871.003.0014

Dove, S. (2024). *b3gbi: General Biodiversity Indicators for Biodiversity Data Cubes* [Computer software]. https://github.com/b-cubed-eu/b3gbi

Efron, B. (1987). Better Bootstrap Confidence Intervals. *Journal of the American Statistical Association*, *82*(397), 171–185. https://doi.org/10.1080/01621459.1987.10478410

Efron, B., & Tibshirani, R. J. (1994). *An Introduction to the Bootstrap* (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429246593

Ferro, M. L., & Flick, A. J. (2015). “Collection Bias” and the Importance of Natural History Collections in Species Habitat Modeling: A Case Study Using Thoracophorus costalis

Erichson (Coleoptera: Staphylinidae: Osoriinae), with a Critique of GBIF.org. *The Coleopterists Bulletin*, *69*(3), 415–425. https://doi.org/10.1649/0010-065X-69.3.415

Fischhoff, B., & Davis, A. L. (2014). Communicating scientific uncertainty. *Proceedings of the National Academy of Sciences*, *111*(supplement_4), 13664–13671. https://doi.org/10.1073/pnas.1317504111

Frangos, C. C., & Schucany, W. R. (1990). Jackknife estimation of the bootstrap acceleration constant. *Computational Statistics & Data Analysis*, *9*(3), 271–281. https://doi.org/10.1016/0167-9473(90)90109-U

GBIF.org. (2024). *GBIF Occurrence Download* [Dataset]. The Global Biodiversity Information Facility. https://doi.org/10.15468/DL.QK4F2Z

Hesterberg, T. C. (2015). What Teachers Should Know About the Bootstrap: Resampling in the Undergraduate Statistics Curriculum. *The American Statistician*, *69*(4), 371–386. https://doi.org/10.1080/00031305.2015.1089789

Hsieh, T. C., Ma, K. H., & Chao, A. (2016). iNEXT: An R package for rarefaction and extrapolation of species diversity Hill numbers). *Methods in Ecology and Evolution*, *7*(12), 1451–1456. https://doi.org/10.1111/2041-210X.12613

Kinkeldey, C., MacEachren, A. M., & Schiewe, J. (2014). How to Assess Visual Communication of Uncertainty? A Systematic Review of Geospatial Uncertainty Visualisation User Studies. *The Cartographic Journal*, *51*(4), 372–386. https://doi.org/10.1179/1743277414Y.0000000099

Langeraert, W., Dove, S., & Van Daele, T. (2025). *Investigate indicator uncertainty* [Computer software]. https://doi.org/10.5281/zenodo.14754768

Langeraert, W., & Van Daele, T. (2025). *dubicube: Calculation and Interpretation of Data Cube Indicator Uncertainty* [Computer software]. https://github.com/b-cubed-eu/dubicube

MacEachren, A. M., Robinson, A., Hopper, S., Gardner, S., Murray, R., Gahegan, M., & Hetzler,

E. (2005). Visualizing Geospatial Information Uncertainty: What We Know and What We Need to Know. *Cartography and Geographic Information Science*, *32*(3), 139–160. https://doi.org/10.1559/1523040054738936

MacEachren, A. M., Roth, R. E., O'Brien, J., Li, B., Swingley, D., & Gahegan, M. (2012). Visual Semiotics & Uncertainty Visualization: An Empirical Study. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2496–2505. https://doi.org/10.1109/TVCG.2012.279

mikefc. (2025). *ggblur: Blurry Geoms for Ggplot2* [Computer software]. https://github.com/coolbutuseless/ggblur

Milner-Gulland, E. J., & Shea, K. (2017). Embracing uncertainty in applied ecology. *The Journal of Applied Ecology*, *54*(6), 2063–2068. https://doi.org/10.1111/1365-2664.12887

Onkelinx, T. (2023). *effectclass: Classification and visualisation of effects* [Computer software]. https://inbo.github.io/effectclass/

Oskanen, J., Simpson, G., Blanchet, F., Kindt, R., Legendre, P., Minchin, P., O'Hara, R., Solymos, P., Stevens, M., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Carvalho, G., Chirico, M., De Caceres, M., Durand, S., … Weedon, J. (2024). *vegan: Community Ecology Package* [Computer software]. https://CRAN.R-project.org/package=vegan

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, *10*(1), 439. https://doi.org/10.32614/RJ-2018-009

Posit team. (2024). *RStudio: Integrated Development Environment for R* [Computer software]. Posit Software, PBC. http://www.posit.co/

R Core Team. (2024). *R: A Language and Environment for Statistical Computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rowland, J. A., Bland, L. M., James, S., & Nicholson, E. (2021). A guide to representing

variability and uncertainty in biodiversity indicators. *Conservation Biology*, *35*(5), 1669–1682. https://doi.org/10.1111/cobi.13699

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A Grammar of Data Manipulation* [Computer software]. https://CRAN.R-project.org/package=dplyr

Yahaya, M. M., & Kumschick, S. (2025). *impIndicator: Impact Indicators of Alien Taxa* [Computer software]. https://github.com/b-cubed-eu/impIndicator

8 Annex

8.1 Preliminary Rules for Data Cube Robustness

8.1.1 Robustness measures along basic dimensions

8.1.1.1 Temporal

Minimal number of time points

The absolute number of time points is largely the responsibility of the user during the occurrence cube specification. The number of time points is not very important for indicator calculation but could be flagged as a note if only a single time point is available, since this might be a mistake from the user.

category	
> 1 time points	
1 time point	
-	
-	

Missing time points

Percentage of missing time points over the whole time period (first - last) of the data cube for each species.

category	
< 10%	
> 10%	
> 25%	
> 50%	

8.1.1.2 Spatial

Spatial range

Number of grid cells with occurrences as a percentage of the total number of grid cells for all species and each individual species.

category	
> 50%	
-	
-	
< 5%	

Overall spatial clustering

While ecological data are expected to have a considerable degree of structure, extreme values for clustering or dispersion may indicate underlying problems with the dataset. The degree of spatial clustering of the data in the datacube will be investigated using spatial autocorrelation analysis. High spatial autocorrelation means that neighbouring cells tend to have similar values, whereas low spatial autocorrelation means that values are very different when moving from one cell to another.

Moran's I will be used to summarise spatial autocorrelation at varying lag distances for a geographical area. It indicates whether the data are dispersed, random or clustered with values between -1 and 1.

category	
-0.5 < x < 0.8	
x < -0.5 or x > 0.8	
x < -0.7 or x > 0.9	
x < -0.8 or x > 0.95	

Local indicators for spatial association

To gain more insight into where local spatial patterns occur, Local Indicators of Spatial Association (LISA) will be used. They can be used to identify areas with statistically significant spatial patterns, unusual high or low values, detect spatial outliers with values significantly different from their surroundings and local variations.

Several indicators are possible, but at least the most common Local Moran’s I and Geary ratio will be implemented.

8.1.1.3 Taxonomical

Minimal number of taxa

The user specifies a species group from which to make an occurrence cube. It is possible to get a cube with less taxa (e.g. species) than expected because data for only a few taxa in this group are available on GBIF. The number of taxa may or may not be important for indicator calculation, depending on the indicator function.

category	
> 5 taxa	
3-5 taxa	
2-3 taxa	
1 taxon	

8.1.1.4 Observations

Minimal number of observations

Flag when not enough observations.

category	
> 40 observations	
30-40 observations	
20-30 observations	
< 20 observations	

8.1.1.5 Minimal coordinate uncertainty

Minimal coordinate uncertainty

The minimal coordinate uncertainty in meters should not be (much?) larger than the resolution of the grid cell.

category	
All minimal coordinate uncertainty smaller than resolution	
1-5 rows with minimal coordinate uncertainty larger than resolution	
5-10 rows with minimal coordinate uncertainty larger than resolution	
> 20 rows with minimal coordinate uncertainty larger than resolution	

8.1.2 Robustness measures along interactive dimensions

8.1.2.1 Temporal + Spatial

Spatial similarity for consecutive time steps

The spatial distribution of occurrences naturally fluctuates, but abrupt changes may indicate underlying issues in the dataset. To assess spatial changes over time, we calculate the similarity of presence/absence maps (where presence is defined as occurrences > 0, and absence as no occurrences) between successive time points. This is measured using the Jaccard similarity index:

Jaccard similarity = (number of cells with occurrences in both time points) / (number of cells with occurrences in both sets)

The index ranges from 0 (no similarity) to 1 (complete similarity). It can be computed for the entire data cube or for individual species. Results can be visualized as a time series with (time period - 1) steps or summarized as the mean similarity over the full time period.

category	
> 0.8	
0.5 > x < 0.8	
> 0.2 x < 0.5	
< 0.2	

8.1.2.2 Temporal + Taxonomical

Minimal number of taxa per time point

The user specifies a species group from which to make an occurrence cube. It is possible to get a cube with less taxa (e.g. species) than expected because data for only a few taxa in this group is available on GBIF. The number of taxa may or may not be important for indicator calculation depending on the indicator function.

category	
> 5 taxa	
3-5 taxa	
2-3 taxa	
1 taxon	

Relative difference of indicator value per time point

Via leave-one-species-out cross-validation, we can calculate the relative difference between the indicator value calculated with and without each taxon. If the relative difference is too large for one or more taxa, they can be flagged. We can take the median (or mean?) per time point. If the median relative difference is too large, this time point can be flagged.

category	
< 10 %	
10-50 %	
50-100 %	
> 100 %	

8.1.2.3 Temporal + Observations

Minimal number of observations per time point
Flag when not enough observations per time point.

category	
> 30 observations	
20-30 observations	
10-20 observations	
< 10 observations	

Relative difference of number of observations
We can calculate the relative difference between the number of observations per time point and the median number of observations over all time points. If the relative difference is too large for one or more time points, they can be flagged.

category	
< 10 %	
10-50 %	
50-100 %	
> 100 %	

8.1.2.4 Taxonomical + Observations

Minimal number of observations per taxon
Flag when not enough observations per taxon.

category	
> 30 observations	
20-30 observations	
10-20 observations	
< 10 observations	

Relative difference of number of observations

We can calculate the relative difference between the number of observations per taxon and the median number of observations over all taxa. If the relative difference is too large for one or more taxa, they can be flagged.

category	
< 10 %	
10-50 %	
50-100 %	
> 100 %	