BIODIVERSITY
BUILDING
BLOCKS FOR
POLICY

D4.3 Report on the criteria for data quality and species characteristics for estimating species status and trends

27/02/2026

Author(s): Ward Langeraert, Katelyn Faulkner, Emma Cartuyvels, Quentin Groom, Toon Van Daele

Funded by
the European Union

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.

Prepared under contract from the European Commission
Grant agreement No. 101059592
EU Horizon Europe Research and Innovation Action

Project acronym:	**B3**
Project full title:	**Biodiversity Building Blocks for policy**
Project duration:	01.03.2023 – 31.08.2026 (42 months)
Project coordinator:	Dr. Quentin Groom, Agentschap Plantentuin Meise (MeiseBG)
Call:	HORIZON-CL6-2021-GOVERNANCE-01
Deliverable title:	Report on the criteria for data quality and species characteristics for estimating species status and trends
Deliverable n°:	D4.3
WP responsible:	WP4
Nature of the deliverable:	Report
Dissemination level:	Public
Licence of use:	Creative Commons Attribution 4.0 International
Lead partner:	INBO
Recommended citation:	Langeraert, W., Faulkner, K., Cartuyvels, E., Groom, Q. & Van Daele, T. (2026). ***Report on the criteria for data quality and species characteristics for estimating species status and trends***. B3 project deliverable D4.3.
Due date of deliverable:	Month n° 36
Actual submission date:	Month n° 36

Deliverable status:

Version	Status	Date	Author(s)
1.0	Final	27 February 2026	Ward Langeraert (INBO), Katelyn Faulkner (SANBI), Emma Cartuyvels (INBO), Quentin Groom (MeiseBG), Toon Van Daele (INBO)

Table of contents

Key takeaway messages 4
Executive summary 4
Non-technical summary 5
List of abbreviations 5
1. Introduction 6
1.1. Background and motivation 6
1.2. Data cubes and analytical framework 6
1.3. Comparative approach using structured monitoring data 7
1.4. Sampling bias and detectability 8
1.5. Software implementation and practical application 8
2. Comparative approach 9
2.1. Methods 9
2.2. Flanders case study 15
2.3. Western Cape case study 24
3. Technical challenges and data quality barriers 35
3.1. Data is out of bounds of selected area 35
3.2. Taxonomic inconsistencies and species name issues 38
3.3. Temporal coverage and publication delays in GBIF data 39
3.4. Disproportionate influence of component datasets 40
4. Sampling bias and survey completeness 50
4.1. Survey-effort, components and composite score 50
4.2. Probabilistic estimation of sample coverage 55
4.3. Conclusion 57
5. Species detectability and recording bias 59
5.1. Estimating detectability from opportunistic occurrence data 59
5.2. Taxonomic aggregation and outputs 60
5.3. Interpretation, bias, and limitations 62
5.4. Temporal and technological drivers of detectability 63
5.5. Conclusion 65
6. Software implementation 66
6.1. Simulating occurrence cubes in R 66
6.2. Explore occurrence cube quality and reliability in R 72
7. Guidelines for reliable indicator and trend calculations from occurrence cubes 74
7.1. Cube preparation and baseline data filtering 74
7.2. Exploratory assessment of data coverage and structure 74
7.3. Operational quality criteria and documentation 75
7.4. Examples 75
8. Acknowledgements 77
9. References 78
10. Annex 84
10.1. SQL queries Flanders 84
10.2. SQL queries Western Cape 86

Key takeaway messages

- Extensive quality control, data filtering and validation is required in order to produce robust and meaningful results when performing analysis based on GBIF data cubes.
- Comparison of structured monitoring data and unstructured cube data shows it is recommended to use coarse resolutions and longer time periods for spatial and trend analysis. In most cases a correction for survey effort or survey completeness is required.
- The use of data cubes based on GBIF datasets comes with some technical challenges related to differences in spatial uncertainty, taxonomic consistency between datasets, and the publication delay to the GBIF platform.
- The data cubes comprise highly heterogeneous datasets. Some datasets have a strong, dominant effect on the results of an analysis.
- A component-based survey effort score can be used to assess sampling bias and survey completeness. This index offers flexibility in tailoring to reflect regional knowledge of recording behaviour and expert understanding of how sampling occurs.

Executive summary

This deliverable report identifies and substantiates criteria for determining the reliability of species status and trend estimates derived from aggregated data cubes from the Global Biodiversity Information Facility (GBIF). To achieve this, the report adopts a comparative approach, contrasting unstructured GBIF cube data with structured monitoring data from bird surveys in Flanders (Belgium) and the Western Cape (South Africa).

Key findings from these case studies demonstrate that extensive quality control, data filtering, and validation are essential to producing robust results from unstructured data. Specific technical barriers identified include coordinate uncertainty, where large uncertainty radii can distort spatial signals; taxonomic inconsistencies, where unlinked accepted names artificially inflate species richness; and publication delays, which can create misleading temporal trends. Furthermore, the analysis reveals that data cubes are often dominated by a small number of highly influential component datasets, making indicators sensitive to the presence or absence of specific sources.

To address these biases, the report implements diagnostic frameworks to quantify survey effort and completeness. It introduces a survey-effort score, capturing record volume, temporal replication, and taxonomic coverage, and utilizes probabilistic estimators to assess survey completeness. Additionally, the report implementation examines species detectability, implementing survey-based detection probability metrics to distinguish between genuine ecological signals and reporting biases driven by technology or observer behaviour.

Finally, the report operationalizes these assessments through specialized software tools developed within the B3 project, specifically the **gcube** R package for simulating occurrence cubes and the **dubicube** R package for quality checks and quantifying indicator uncertainty. These insights are synthesized into a set of operational guidelines for reliable indicator and trend calculations, ensuring that biodiversity reporting based on aggregated occurrence data is transparent, reproducible, and robust.

Non-technical summary

As more people use mobile apps and new technologies to record nature, we are collecting massive amounts of "unstructured" biodiversity data. While this information is valuable for tracking the state of our environment, it is often messy and lacks a standardized plan, which can make it difficult to trust for official policy reporting.

This report explores how we can transform this large, inconsistent data into reliable indicators of how species are doing over time. By comparing public data from the Global Biodiversity Information Facility (GBIF) with professional bird monitoring programs in Belgium and South Africa, the study identifies several "traps" that can lead to wrong conclusions:

- **Location Errors:** If a recorded location is too vague, it can make it look like species are in places where they are not.
- **Naming Issues:** Different names used for the same species can make a region appear more diverse than it actually is.
- **Time Gaps:** Delays in uploading data can make it look like a species is declining, when the data just hasn't been shared yet.
- **Data Dominance:** Often, a single large dataset, like from a major citizen science initiative, can completely control the overall trend, making the results vulnerable if that one source stops sharing data.

To solve these problems, we investigated tools and quality checks for the data. These include a survey effort score to see if an area has been visited enough to be considered "well-sampled" and detectability metrics to account for species that are simply harder to see or record than others.

The report concludes with easy-to-follow guidelines and specialized software. These tools help scientists and policymakers filter out the "noise" in biodiversity data so they can produce accurate reports on the status and trends of species, ultimately leading to better-informed decisions for nature conservation.

List of abbreviations

ABV	Common Breeding Bird Survey Flanders
B3	Biodiversity Building Blocks for Policy
DOI	Digital Object Identifier
EEA	European Environment Agency
EU	European Union
GBIF	Global Biodiversity Information Facility
GIS	Geographic Information System
MGRS	Military Grid Reference System
R	R programming language
SABAP2	Southern African Bird Atlas Programme 2
UTM	Universal Transverse Mercator coordinate system

1. Introduction

1.1. Background and motivation

The increasing availability of large volumes of biodiversity occurrence data through global infrastructures such as the Global Biodiversity Information Facility (GBIF) has created new opportunities for assessing species status and trends at national and regional scales (https://www.gbif.org/). At the same time, the heterogeneous and largely unstructured nature of these data raises fundamental questions about the conditions under which reliable indicators can be derived. In this deliverable report, we address this challenge by examining the criteria for data quality and species characteristics that determine the reliability of species status and trend estimates derived from aggregated occurrence data cubes. This is particularly relevant given the growing reliance on GBIF data for policy-relevant biodiversity reporting (Groom et al., 2025).

Recent technological developments, including mobile applications, environmental DNA, automated sensors, computer vision and remote sensing, have accelerated the collection of so-called ‘big unstructured biodiversity data’. Such data are typically collected for broad descriptive purposes (e.g. documenting presence) without predefined sampling designs, rather than through statistically designed monitoring schemes (Bayraktarov et al., 2019). While these data hold considerable promise for large-scale biodiversity assessments, their suitability for deriving reliable indicators of species status and trends remains insufficiently understood.

The aim of this report is to identify and substantiate criteria for determining when models, trends and status estimates derived from aggregated GBIF data can be considered reliable. To achieve this, we examine key data-related factors, such as sampling effort, spatial and temporal coverage, taxonomic resolution and detection probability (Burgass et al., 2017; Isaac et al., 2014; Van Eupen et al., 2021). We also consider species-specific characteristics, including abundance, detectability and spatial and temporal dynamics (Callaghan et al., 2018; Kamp et al., 2016).

1.2. Data cubes and analytical framework

Within the B3 project (https://b-cubed.eu/), software has been developed to generate multidimensional species occurrence cubes that integrate taxonomic (what), spatial (where) and temporal (when) dimensions (Desmet et al., 2025). These data cubes provide a structured representation of biodiversity data that enables the consistent application of analytical workflows across regions, taxa and data sources. The cube concept has been operationalised through a public download service hosted by GBIF (Blisset et al., 2025), allowing users to generate occurrence cubes from all available GBIF data.

Despite these methodological advances, GBIF occurrence data are known to exhibit a range of biases and uncertainties that may compromise the reliability of derived indicators. These include:

- Taxonomic biases: Over-representation of well-known and easily identified or detected taxa, with under-sampling of taxonomically challenging groups and rare or cryptic species that are difficult to detect (Troudet et al., 2017). Detectability is the probability that a species will be recorded on a survey day given that it is truly present in a specific grid cell.

- Geographic biases: Non-random/uneven sampling effort in accessible or well-surveyed areas, leading to spatial clustering, such as in accessible areas (e.g. near roads) or popular sites (García-Roselló et al., 2023).
- Temporal biases: Opportunistic or non-random sampling over time and temporal changes in detectability confounds biological trends (Boakes et al., 2010). Detectability is influenced by biological visibility, reporting biases, and technological shifts such as the adoption of mobile applications and AI-assisted identification.
- Data quality issues: False positives due to misidentification or data processing errors.

As a result, indicators derived from aggregated (unstructured) occurrence cubes may exhibit misleading patterns or trends if these sources of uncertainty and bias are not adequately accounted for. A key objective of this report is therefore to determine the conditions under which these uncertainties are sufficiently constrained to allow robust estimation of species status and trends.

1.3. Comparative approach using structured monitoring data

To evaluate the reliability of indicators derived from unstructured occurrence data, this report adopts a comparative approach in which data cubes generated from GBIF are contrasted with cubes derived from structured monitoring and inventory schemes. These structured datasets are based on consistent sampling designs and standardized protocols, and they are therefore assumed to provide a largely accurate representation of underlying ecological patterns. The comparison focuses primarily on the accuracy of trend and status estimates derived from unstructured data, while statistical precision is addressed separately within the B3 project (see Langeraert et al., 2025).

To enable a robust evaluation across contrasting biogeographical contexts, we selected taxa and regions for which long-term, structured monitoring data are available (Cartuyvels et al., 2025; Langeraert et al., 2023). In Flanders (Belgium), the Common Breeding Bird Survey (ABV) provides structured data on breeding bird populations based on a fixed, stratified random sampling design (Onkelinx et al., 2023; Vermeersch, 2007; Vermeersch et al., 2018). For the Western Cape province of South Africa, the Southern African Bird Atlas Programme 2 (SABAP2) serves as the reference, complemented by the high-quality Hessequa Systematic Atlasing Subproject, which utilizes more intensive and systematic spatial coverage than the broader atlas dataset (Brooks et al., 2022).

For both regions, species occurrence cubes are constructed from GBIF data covering identical taxa, spatial extents, and temporal periods. By applying identical B3 workflows to both data sources, this task aims to determine which data quality thresholds and species-level attributes, such as rarity and abundance, are necessary for reliable national or regional status and trend estimates.

1.4. Sampling bias and detectability

To address spatial and temporal biases, the report implements diagnostic frameworks to evaluate where and when a dataset provides a sufficiently reliable basis for indicators. This involves quantifying survey effort through interpretable components, such as record volume, temporal replication, and seasonal coverage, which are combined into an aggregated survey-effort score to identify well-sampled areas. Furthermore, incidence-based sample coverage is estimated to provide a probabilistic measure of survey completeness.

The report also examines species detectability. It is influenced by biological visibility, reporting biases, and technological shifts such as the adoption of mobile applications and AI-assisted identification. By implementing survey-based detection probability metrics derived from repeated sampling events, the report aims to distinguish between genuine ecological signals and the institutional or human processes that govern the data collection lifecycle.

1.5. Software implementation and practical application

To operationalize the assessment of both technical data quality and underlying ecological biases, this report utilizes and describes specialized software tools developed within the B3 project framework. The **gcube** R package (Langeraert, 2026) provides a simulation environment that allows for the disentanglement of biological processes from observation processes by generating multi-species distributions and simulating diverse sampling biases. Complementing this, the **dubicube** R package (Langeraert & Van Daele, 2026) provides general measures for quantifying indicator uncertainty and data cube reliability, translating the "rules of thumb" identified in this study into practical diagnostics.

While the comparative case studies provide the empirical baseline for understanding data reliability, the software tools described here allow users to apply these insights to their own datasets.

2. Comparative approach

2.1. Methods

2.1.1. Selection of reference datasets

To ensure a robust comparison between structured and unstructured data, reference datasets were selected based on their ability to provide an accurate representation of ecological systems while minimizing inherent biases through standardized protocols. The selection process followed a common framework designed to balance sample size with practical applicability across taxonomic, spatial, and temporal dimensions (Langeraert et al., 2023).

The selection of monitoring projects was guided by specific criteria to ensure the reliability of species status and trend calculations. We prioritized datasets containing a sufficient number of different species with the inclusion of a spectrum of species ranging from rare to common. While very rare species often lack sufficient unstructured observations, very common species may suffer from underreporting by citizen scientists.

From a structural perspective, we favoured long-term monitoring programs, ideally spanning more than ten years, as these are better equipped to capture accurate trends and the increased observation effort characteristic of modern unstructured data. Furthermore, selected datasets required a large spatial extent to allow for comparisons across different data cube grid resolutions. It was also essential that the structured reference data covered the same temporal period as the unstructured GBIF data to facilitate direct comparisons.

2.1.1.1. Flanders case study

Common Breeding Bird Survey Flanders (ABV)

The Common Breeding Bird Survey Flanders (Algemene Broedvogelmonitoring Vlaanderen, ABV) is a long-term structured monitoring programme designed to assess population trends of common breeding bird species in Flanders, the northern region of Belgium (13,626 km²; ~7 million inhabitants) (Fig. 1). The programme has been running continuously since 2007 and targets approximately 100 widespread breeding bird species.

The ABV monitoring protocol is based on a random sample of 1 × 1 km Universal Transverse Mercator (UTM) grid cells, stratified according to the relative proportions of six land-use or ecosystem types: agricultural land, urban areas, forest, suburban areas, heathland and dunes, and marsh and open water (Vermeersch, 2007). In total, 1,200 grid cells are included in the sampling frame. These are divided into three groups of 300 grid cells, which are surveyed in a three-year rotation, such that each grid cell is visited once every three years.

Within each selected grid cell, six fixed monitoring locations are established where bird counts are conducted. Each grid cell is surveyed three times per breeding season, within fixed time windows and with a minimum interval of two weeks between visits. Counts are based on both visual and auditory detections and are carried out between sunrise and four hours after sunrise, following a strictly standardised protocol (Onkelinx et al., 2023; Piesschaert et al., 2022; Vermeersch et al., 2018, 2021).

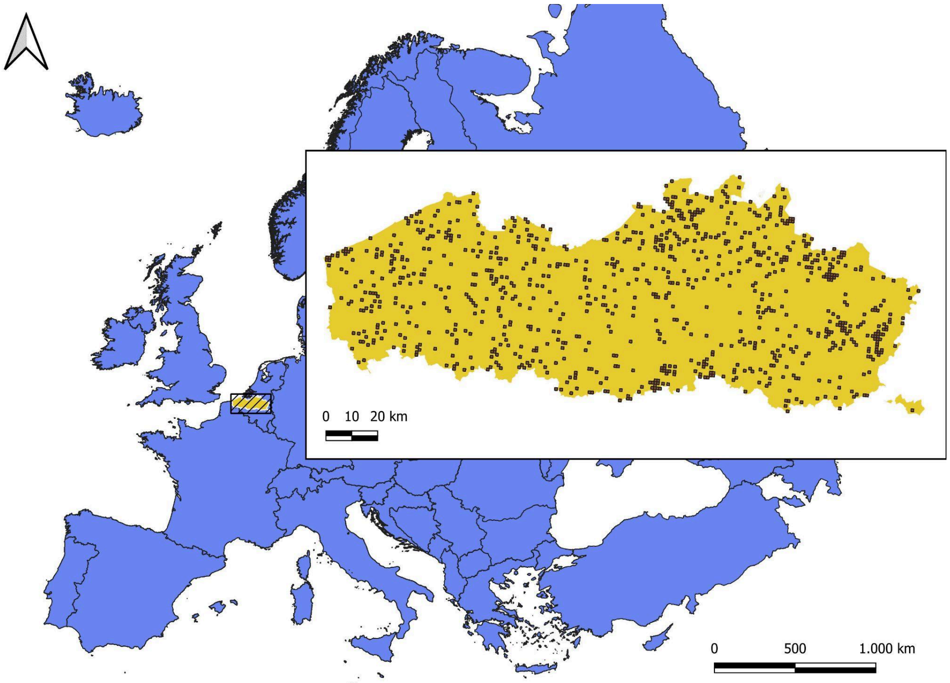


Figure 1: Location of Flanders within Europe and ABV monitoring locations within Flanders.

The ABV sampling design is optimised for detecting population trends rather than for estimating fine-scale spatial distributions. Certain ecosystem types that are relatively rare in Flanders are intentionally oversampled to ensure sufficient data for trend estimation within each ecosystem. As a result, although the survey covers approximately 6% of the total area of Flanders, the spatial distribution of sampling effort is not proportional to land cover, and the data do not directly provide unbiased estimates of species' spatial distributions.

Data access and representation

The ABV data are published on GBIF as two separate datasets covering the periods 2007–2016 (Vermeersch et al., 2021) and 2017–2021, following integration into the Meetnetten.be framework (Piesschaert et al., 2022). Occurrence records are spatially aggregated to the level of the 1 × 1 km UTM grid cell, with coordinates corresponding to the grid cell centroid. The associated coordinate uncertainty reflects the radius of a circle encompassing the grid cell (707 m). Temporal information is retained at full resolution, with exact observation dates available for each record.

The ABV dataset conforms well to the criteria required for use as a structured reference dataset in this study. It focuses on relatively common and easily detectable species, is based on a long-term and well-documented sampling protocol and targets a taxonomic group that is both popular and well-observed by citizen scientists. At the same time, variation among species in abundance, detectability and spatial extent provides a suitable basis for evaluating the conditions under which trends and status estimates derived from unstructured GBIF data are reliable.

2.1.1.2. Western Cape case study

Southern African Bird Atlas Programme 2 (SABAP2)

The Southern African Bird Atlas Programme 2 (SABAP2) is a large-scale bird monitoring and mapping project that began in 2007 (Brooks et al., 2022). The project aims to document the distribution and relative abundance of bird species across South Africa, Lesotho, Botswana, Namibia, Mozambique, Eswatini, Zimbabwe and Zambia. SABAP2 data from the Western Cape of South Africa (129 462 km², ~7.5 million inhabitants) and from the area where the Hessequa Systematic Atlasing Subproject is undertaken (~110 x 55 km, ~3500 inhabitants) are used as structured reference datasets for comparison with unstructured GBIF-derived occurrence data (Fig. 2).

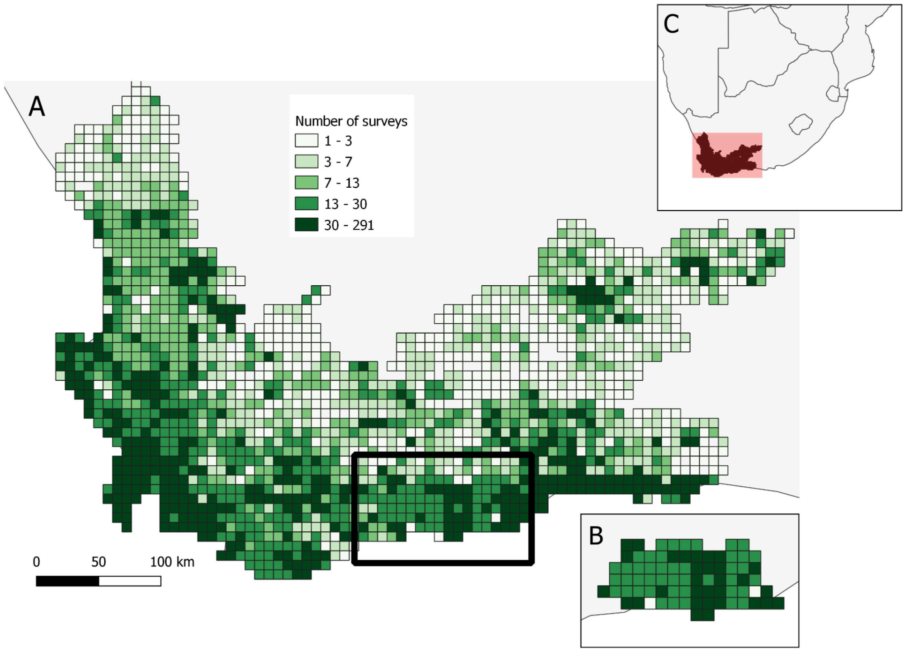


Figure 2: Number of SABAP2 standard protocol surveys per 5-minute² grid cell (pentad) in the Western Cape province of South Africa (panel A), and in the area of the Hessequa Systematic Atlasing Subproject (panel B). The inset in panel A shows the location of the Hessequa Systematic Atlasing Subproject in the Western Cape, and the inset in panel C shows the location of the Western Cape in South Africa.

SABAP2 data are collected primarily by citizen scientists following a standardised survey protocol, known as the BirdMap protocol (Brooks et al., 2022). Observations are recorded within fixed spatial grid cells (‘pentads’) with a resolution of 5 × 5 arc minutes, corresponding to approximately 8.2 × 8.2 km in southern Africa. For a standardised survey, observers spend a minimum of two hours and up to five days recording as many bird species as possible within a pentad. Species are recorded in the order in which they are seen or heard, and the hour of first detection is noted, providing approximate information on time-to-detection. Observers are expected to identify at least 90 % of species encountered and to survey all or most representative habitats within the grid cell. Survey duration, coverage of habitats and inclusion of nocturnal observations are self-reported.

In addition to standardised surveys, SABAP2 also accepts *ad hoc* records that do not meet the protocol requirements. These *ad hoc* records can be readily filtered from the dataset and were excluded from the analyses. All records submitted to SABAP2 undergo quality control procedures, including automated checks against known species' distributions and manual vetting by regional atlas committees for records flagged as unusual (Brooks et al., 2022). Further verification may involve follow-up communication with observers. From 2014 onwards, data quality and consistency improved substantially with the widespread adoption of BirdLasser, a mobile phone data collection application, which has reduced data capture errors (Brooks et al., 2022). Some grid cells have been surveyed more than once (Fig. 2), but for most of the area covered by SABAP2 repeated sampling over time is not standardised.

Hessequa Systematic Atlasing Subproject

Within SABAP2, the Hessequa Systematic Atlasing Subproject represents a high-quality, intensively monitored subset of data. The Hessequa Systematic Atlasing Subproject is undertaken near Stillbaai in the Western Cape of South Africa (Fig. 2) and covers 75 pentads (van Rooyen, 2018). Directed and systematic surveys following the SABAP2 standard protocol have been conducted in this area since October 2014 by a relatively consistent group of observers (Underhill & Van Rooyen, 2020; van Rooyen, 2018).

Although the survey intensity has evolved over time, every pentad has been surveyed at least once per year since the start of the project. Survey effort increased between 2017 and 2021, when each pentad was surveyed at least twice per year and surveys were distributed evenly across seasons. Since 2021, a protocol has been in place to ensure at least one survey per pentad per year, with seasonal coverage maintained.

Data access and representation

SABAP2 data are uploaded to GBIF at regular intervals (Brooks & Ryan, 2023). As GBIF standards do not capture all aspects of surveys, some information, such as survey effort (i.e., number of surveys per grid cell), is not directly accessible via GBIF and must be obtained through the SABAP2 data portal (https://sabap2.birdmap.africa/) or associated tools, such as the **rabm** R package (Clarance, 2019). Occurrence records are spatially generalised to the centroid of the pentad grid cell. Temporal information is retained at full resolution, with start and end dates available for each survey.

The SABAP2 dataset conforms well to the criteria required for use as a structured reference dataset. It includes records for over 500 species in the Western Cape, with over 300 species recorded through the Hessequa Systematic Atlasing Subproject. These encompass common and rare native species, and common alien species. Birds are generally easy to detect and identify, and the combination of standardised protocols, observer guidelines and record validation makes SABAP2 well suited as a reference dataset for evaluating the reliability of trends derived from aggregated occurrence data. While survey effort varies considerably across space and time for the Western Cape province, detailed information on survey effort per grid cell is available (Fig. 2) and has been used in previous assessments of sampling bias (Hugo & Altwegg, 2017).

Due to its repeated, directed sampling and the consistent application of the standardised protocol, the Hessequa Systematic Atlasing Subproject provides data of substantially higher quality than that available for most other areas covered by SABAP2. Although the spatial extent of the dataset is limited (Fig. 2), sufficient unstructured GBIF data are available for comparison: since 2015, over 400 bird species have been recorded in the area, with more than 100 records available for several dozen species. The Hessequa Systematic Atlasing Subproject dataset therefore serves as a valuable high-quality test case for evaluating the conditions under which reliable species trends and status estimates can be derived from aggregated occurrence data.

2.1.2. Data aggregation and analyses

All analyses were performed using the R programming language (R Core Team, 2025). Data aggregation and analysis procedures, including the choice of spatial resolution, were tailored to the structure and native spatial units of the reference monitoring data in each case study. In Flanders for example, the ABV scheme provides fully structured data with fixed grid cells, allowing direct cube generation via GBIF at 1 km resolution, with 10 km grids used to assess the effects of spatial aggregation. In contrast, SABAP2 is based on atlas-specific spatial units (pentads), requiring additional pre-processing and custom cube construction to ensure analyses are conducted at spatial scales consistent with the underlying data-generating process. For SABAP2, quarter degree grid cell grids were used to assess the effects of spatial aggregation.

2.1.2.1. Flanders case study

Structured monitoring data from ABV and unstructured occurrence cube data from GBIF were analysed over the same spatial and temporal extent to allow direct comparison. The ABV data are published as two datasets on GBIF and could therefore be downloaded from GBIF using the `occ_download_sql()` function from the **rgbif** R package (Chamberlain et al., 2025). The data were downloaded directly in data cube format; coordinate uncertainty was set to zero to prevent randomisation and preserve the original spatial resolution of the monitoring scheme. Unstructured cube data were downloaded using the same function, explicitly excluding ABV records to avoid data overlap.

Yearly occurrence cubes were generated at two spatial resolutions: 1 km and 10 km Military Grid Reference System (MGRS) grids. This aggregation was also done for the structured (ABV) data (Table 1). An additional (unstructured) data cube including a dataset dimension was created at 1 km resolution to support analyses described in Chapter 3.4. This means that this cube has, in addition to the classic three dimensions (year, 1 km grid cell, species), a fourth dimension which specifies the source dataset. This SQL query is provided in Annex 1 but can also be consulted from the DOI in Table 1.

All code used for the Flanders case study is publicly available on GitHub (Langeraert et al., 2026). The complete GitHub repository with all data files is available on Zenodo (v1.0.0, 10.5281/zenodo.18782579).

2.1.2.2. Western Cape case study

Structured data from the Southern African Bird Atlas Programme 2 (SABAP2), and unstructured GBIF occurrence cube data over the same spatial and temporal extents and resolutions were analysed and compared.

Table 1: Data cubes used for the Flanders case study. The DOI also shows the original SQL query used to obtain the data cube.

	Resolution	Structured	DOI	Description
1	1 km	yes	10.15468/dl.bjzbrv	Using zero meters of coordinate uncertainty. Identical to the original ABV data on grid cell level.
2	1 km	no	10.15468/dl.vvqewm	GBIF occurrence cube excluding ABV data.
3	10 km	yes	10.15468/dl.97pdjf	Using zero meters of coordinate uncertainty.
4	10 km	no	10.15468/dl.5grehw	GBIF occurrence cube excluding ABV data.
5	1 km	no	10.15468/dl.48vfzy	GBIF occurrence cube excluding ABV data with an extra dataset dimension (see Chapter 3.4).

The analyses were performed at two spatial extents: for the Western Cape province of South Africa and the area of the Western Cape covered by the Hessequa Systematic Atlasing Subproject (Fig. 2). The temporal extent of the analysis was 2015-2023. This period was selected by considering both the quality of the SABAP2 data (improved from 2014 onwards), the duration of the Hessequa Systematic Atlasing Subproject (commenced in late 2014), and delays in data submission to GBIF (see Chapter 3.3, the threshold of December 2023 was implemented following the recommendations made in Chapter 7). Analyses were performed at two spatial resolutions: pentad (5′ × 5′) and Extended Quarter Degree Grid Cell (QDGC).

Structured SABAP2 data for the Western Cape were downloaded from GBIF as a raw occurrence dataset (dataset 1, Table 2). Although the `occ_download_sql()` function from **rgbif** could have been used, this approach would have included *ad hoc* records that cannot be filtered out at the cube download stage. Therefore, the raw SABAP2 dataset was downloaded and pre-processed to exclude *ad hoc* records prior to cube construction. Yearly cubes for the Western Cape and the area covered by the Hessequa Systematic Atlasing Subproject (hereafter shortened to 'Hessequa area') were created using both pentad and QDGC grids. Details on the pentad of each record is available in GBIF (column 'verbatimLocality'), thus for pentads, cubes were created by aggregating the data based on this column and other relevant columns (e.g., 'year'). A function was written to obtain the QDGC for each record; with QDGC cubes created by aggregating the data based on this information and other relevant columns (e.g., 'year').

A yearly occurrence cube for the Western Cape at QDGC resolution was downloaded directly from GBIF using the `occ_download_sql()` function (dataset 2, Table 2) and subsequently subsetted to create a yearly cube at QDGC resolution for the Hessequa area. Pentad-year cubes could not be downloaded directly via **rgbif** and were therefore constructed using the `map_grid_designation()` function from the **gcube** R package (Langeraert, 2026), based on raw occurrence data (dataset 3, Table 2). Structured SABAP2 records were excluded from all unstructured occurrence cubes.

Other recommendations made in Chapter 7 were also followed (e.g., records with a minimum coordinate uncertainty of > 8 m for the pentad cubes, and > 27 km for the QDGC cubes were filtered out).

All code used for the Western Cape case study is publicly available on GitHub (Faulkner, 2026). The complete GitHub repository with all data files is available on Zenodo (v1.0.0, 10.5281/zenodo.18802553).

Table 2: Datasets used for the Western Cape case study. The DOI also shows the original SQL query or specifications used to obtain the data cube.

	Grid	Structured	DOI	Description
1	QDGC/pentad	yes	10.15468/dl.fa9thf	Raw structured occurrence data (SABAP2).
2	QDGC	no	10.15468/dl.nqf9x5	GBIF occurrence cube excluding SABAP2 data.
3	pentad	no	10.15468/dl.jxzwn4	Raw occurrence data excluding SABAP2 data, used as input for pentad cubes (see text).

2.2. Flanders case study

2.2.1. Exploration of data

The ABV data ranges from over 40,000 observations in 2007 to just over 20,000 observations in 2022 (Fig. 3A). 180 bird species were recorded, of which 34 were recorded less than 10 times (very rare), 29 were recorded between 10 and 100 times (rare), 59 were recorded between 100 and 1000 times (common), 43 were recorded between 1000 and 10000 times (very common) and 15 were recorded more than 10,000 times (extremely common) (Fig. 3B). This categorization is used throughout further analyses.

The unstructured occurrence cube contains information on 665 species, of which 157 were recorded less than 10 times (very rare), 126 were recorded between 10 and 100 times (rare), 135 were recorded between 100 and 1000 times (common), 84 were recorded between 1000 and 10000 times (very common) and 163 were recorded more than 10,000 times (extremely common) (Fig. 4B).

The data cube is made up of observations from several component datasets. With the largest dataset (*Waarnemingen.be*) containing most of the observations (74.7 %). For further analyses it is important to know that *Waarnemingen.be* data was last published in 2019 and currently runs only to 31-12-2018. Therefore, you can see a clear drop-off in the number of observations after this date (Fig. 4A). The influence of component datasets will be explored separately in Chapter 3.4.

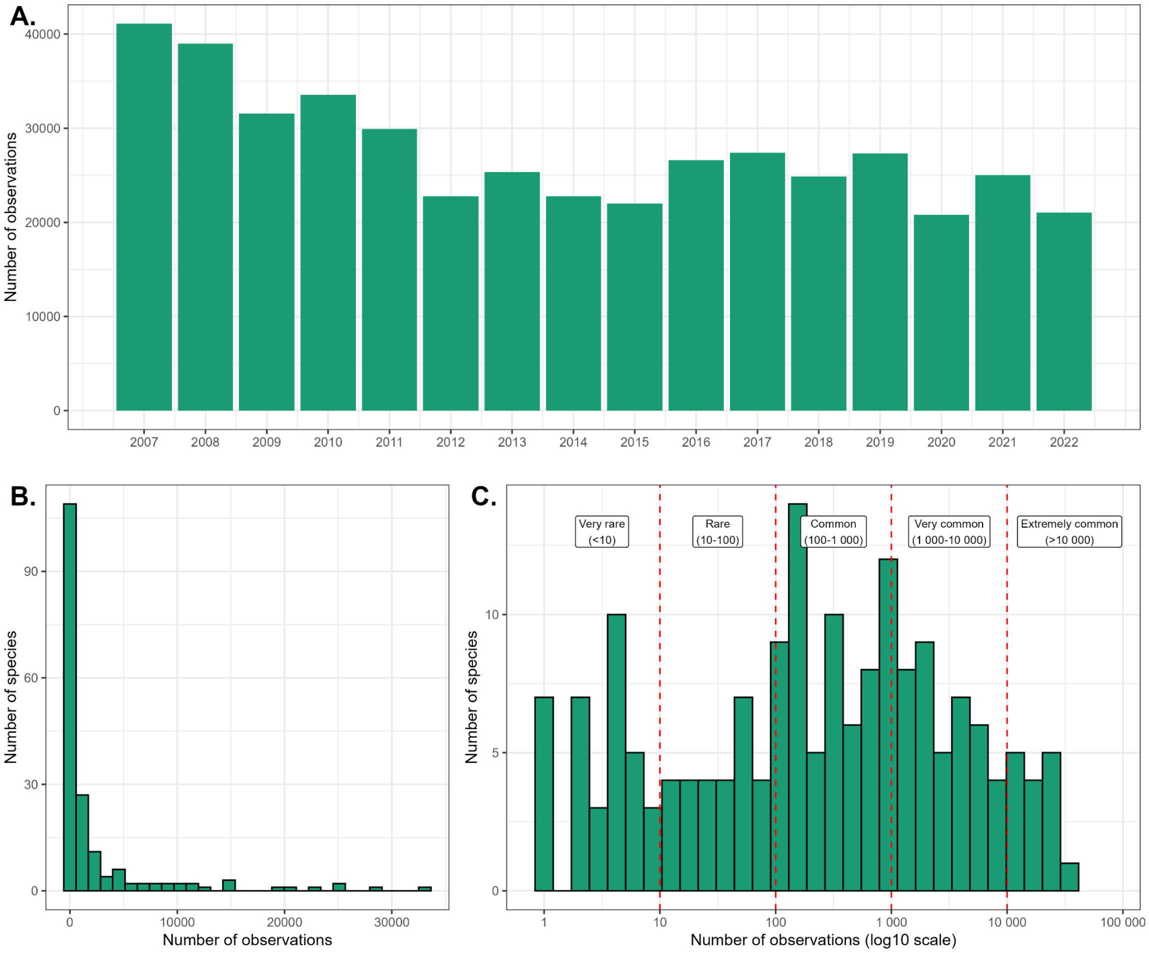


Figure 3: Overview of the structured dataset (ABV). A. Number of observations per year, showing temporal trends in sampling effort. B. Distribution of the number of species by number of observations, highlighting variation in species' abundance and detectability. C. Same species-level distribution on a log10 scale with rarity categories indicated (very rare, rare, common, very common, extremely common), illustrating how species occurrences are distributed across abundance classes.

There are very few observations with unknown coordinate uncertainty, most have an uncertainty of 3536 m (Fig. 5). This is due to the fact that the W*aarnemingen.be* data, making up 74.7 % of the data cube, are published based on an aggregation per 5 km UTM grid cell: $\sqrt{(5000/2)^2 + (5000/2)^2} = 3536$.

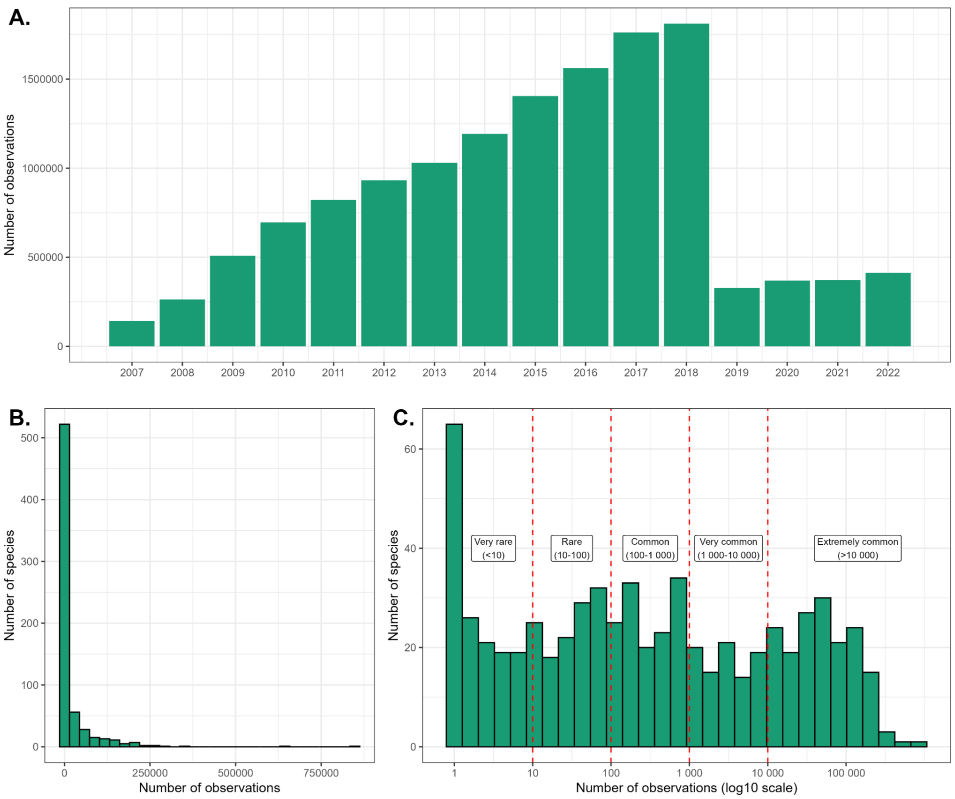


Figure 4: Overview of the unstructured occurrence cube. A. Number of observations per year, showing temporal trends in sampling effort. B. Distribution of the number of species by number of observations, highlighting variation in species' abundance and detectability. C. Same species-level distribution on a log10 scale with rarity categories indicated (very rare, rare, common, very common, extremely common), illustrating how species occurrences are distributed across abundance classes.

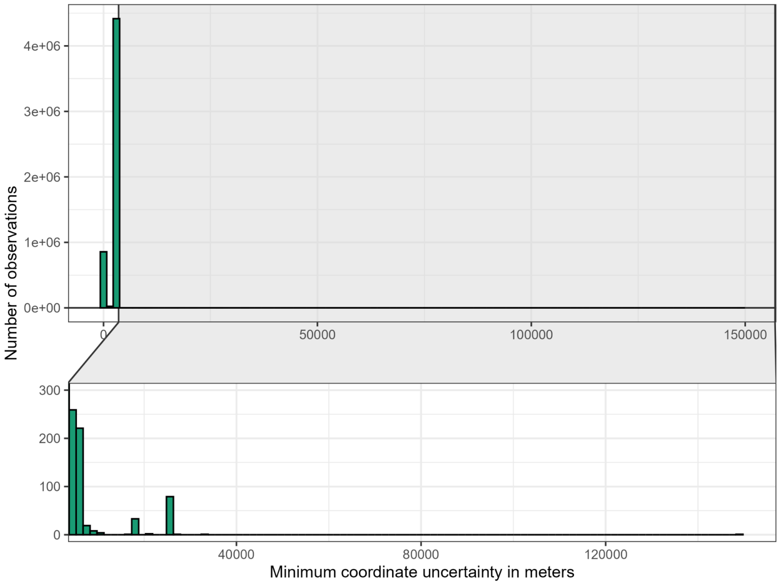


Figure 5: Distribution of minimum coordinate uncertainties in the cube.

2.2.2. Results

2.2.2.1. Range

The ABV monitoring is based on a (stratified) random sample, with a limited number of locations. The unstructured occurrence cube covers all grid cells in Flanders. Therefore, we do not compare the occupancy of each grid cell one-on-one but rather look if a species occurs in a similar percentage of the total number of 1 km grid cells.

There is a clear correlation (R = 0.77, p < 0.001) between the percentage of occupied grid cells in the structured dataset and unstructured dataset (Fig. 6A), but there is a big difference between the rarity classes. The range of rare and very rare species based on unstructured data appears to be overestimated compared to the ABV monitoring. The ABV monitoring is designed to track trends of common breeding birds. Rare species are not the scope of the monitoring and would require a much larger sample size. The statistical analysis of the ABV monitoring data is limited to species where the data fulfil some minimum requirements. We apply a similar filter to the unstructured occurrence cube. The species are only included when the following minimum requirements are fulfilled:

- A grid cell is only counted for a species if that species is observed more than once
- Observations in at least three different cells
- At least a hundred total observations

The result with the filtering is shown in Figure 6B. The very rare and rare species are not observed frequently enough to be included in the analysis and are filtered out.

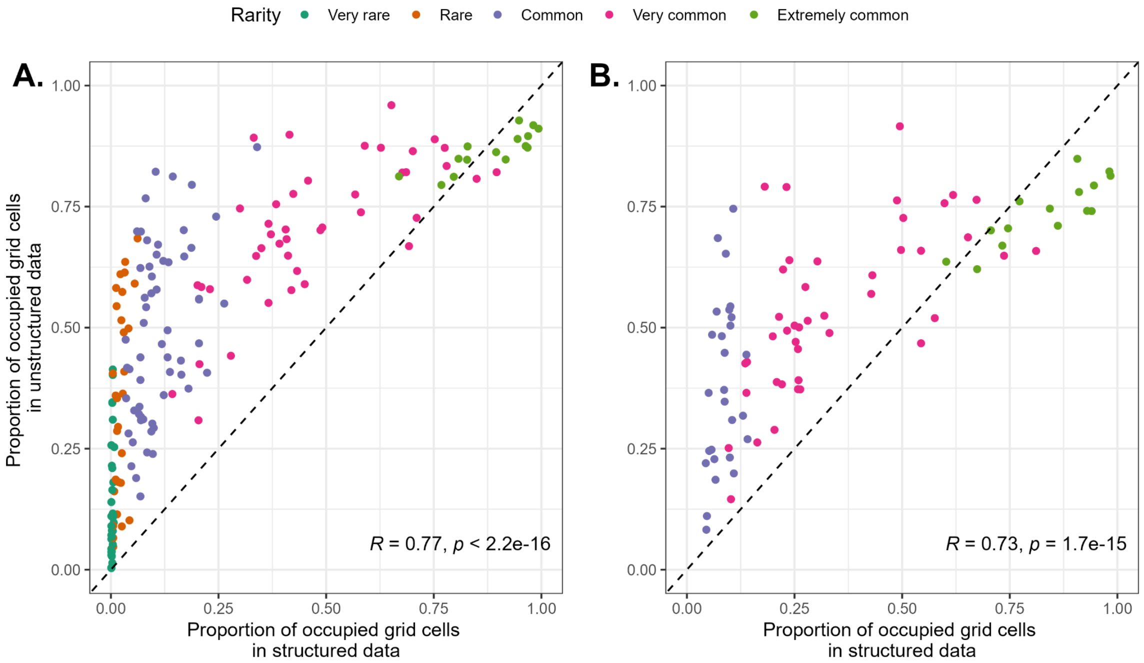


Figure 6: Correlation between the percentage of occupied grid cells between the structured (ABV) data and unstructured (GBIF) data per species. A. All records. B. Filter requirements (see text).

2.2.2.2.Trend comparison

To assess the similarity in temporal trends between the structured and unstructured data, we calculate the correlation coefficient for each species across the time series. For each species present in both datasets, we extract its occurrence trend over time and compute the Pearson correlation coefficient between the two time series. A high positive correlation indicates that both datasets show a similar pattern of change over time for that species, while a low or negative correlation suggests differing or opposing trends. This approach provides a quantitative measure of agreement between the datasets at the species level.

Due to the setup of the ABV data collection (with a subset of grid cells monitored in cycles of three years) it is more appropriate to look at the correlation between trends over the different cycles instead of the trend per year. The raw comparison of all species shows for most species an overall comparable trend, except for rare and very rare species, where the trend is just as often opposite as it is similar (Fig. 7A).

In order to have a more robust result we applied a filtering which is similar to the filtering used for the trend analysis of the ABV data (discussed above). With this filter applied, the very rare and rare species are eliminated, and trends are more often comparable (Fig. 7B).

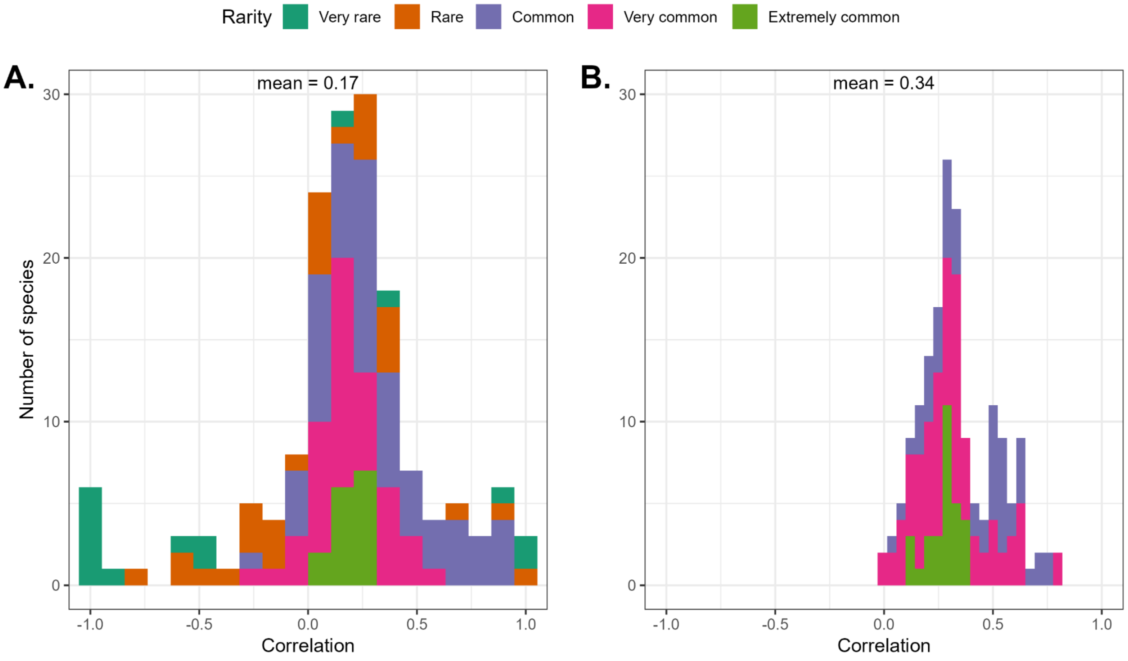


Figure 7: Per species correlation between the trendlines over different monitoring cycles for structured and unstructured data. A. All data. B. Filtered data excluding non-informative grid cells and data deficient species.

We look more in detail to the time series of occurrences for some individual species. We chose three species with different trends in the ABV dataset to compare with the unstructured data:

- **Cetti's warbler** (*Cettia cetti*): the species with a recent increase according to the ABV dataset:
- **Common nightingale** (*Luscinia megarhynchos*): a species with a rather stable trend in the ABV dataset:
- **Eurasian tree sparrow** (*Passer montanus*): the species with a large decrease according to the ABV dataset

As the majority of the unstructured data comes from the dataset 'waarnemingen.be' and this dataset has not been updated to GBIF since 2019, the time series is limited up until that year. Figure 8 shows the individual time series for the three species in the structured (ABV) and unstructured data. The trends in the unstructured data are clearly influenced by increasing survey effort and differ from those in the structured data (ABV). To correct for survey effort, we calculated the relative frequency of each of the three species in relation to the trend of all species within the same taxonomic order. The relative frequency shows patterns that match the structured ABV data quite well (see below). This suggests that relatively large changes and patterns can be detected, provided the species is relatively common and some correction is made for survey effort. Chapter 4 discusses sampling bias and survey effort in detail and proposes a method for combining multiple aspects of survey effort into a global survey effort score.

2.2.2.3.Diversity metrics

The **b3gbi** R package (Dove, 2026) provides functions to calculate several biodiversity indicators from GBIF occurrence cubes. In this section we calculate some biodiversity metrics and compare the results between the occurrence cubes based on unstructured data from GBIF vs. data occurrence cubes derived from the standardized ABV monitoring.

Although there are clear differences in species richness between individual 10 km grid cells, there is no clear spatial pattern for either the structured or unstructured data (Fig. 9) at this level of aggregation. Where the species richness for the structured data (ABV) does not show a significant trend over the years, there is a clear increase in species richness for the unstructured data till 2018 (Fig. 10). This is probably due to the gradual increase in survey effort which is reflected in an increase in occurrences (Fig. 4A).

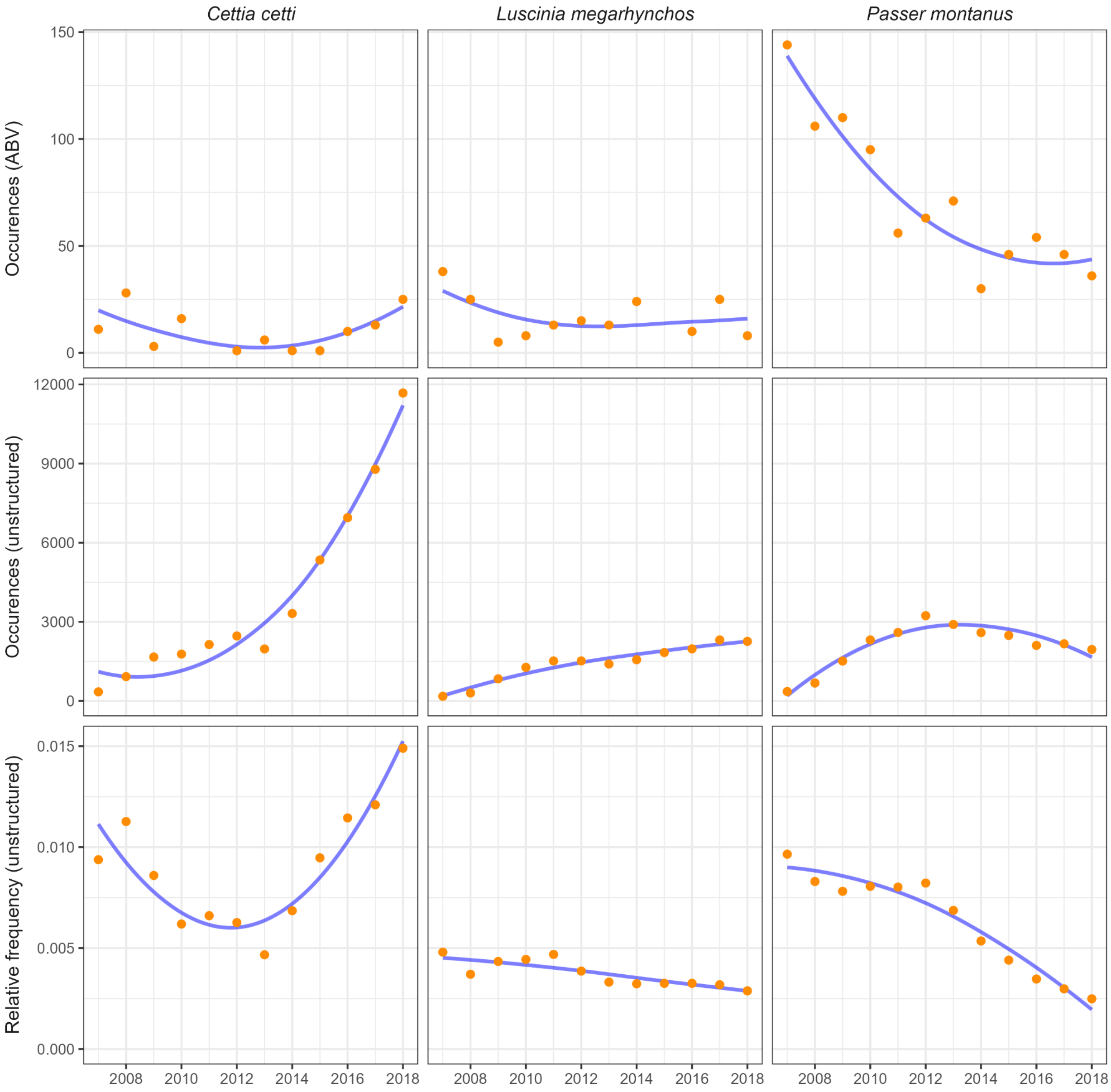


Figure 8: Occurrences from structured (ABV) data (upper row), from unstructured data (middle row) and relative occurrences from unstructured data (lower row) for three species: Cetti's warbler (*Cettia cetti*) (left column), Common nightingale (*Luscinia megarhynchos*) (middle column) and Eurasian tree sparrow (*Passer montanus*) (right column). The trendline is a LOESS smoother with a span of 2.

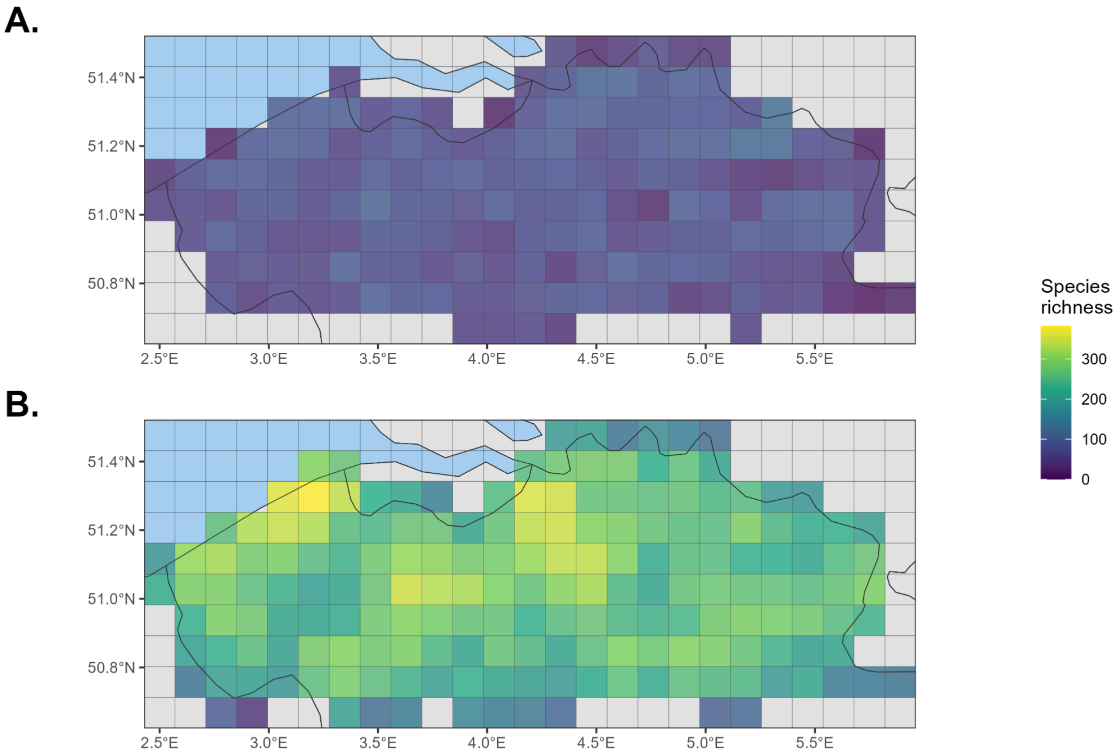


Figure 9: Observed species richness for structured (A) and unstructured (B) data.

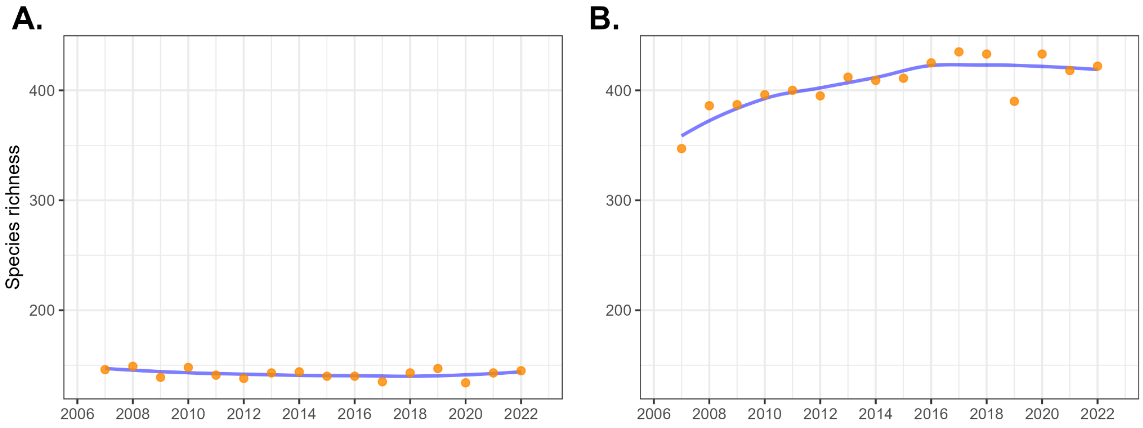


Figure 10: Observed species richness trend for structured (A) and unstructured (B) data. The trendline is a LOESS smoother.

The Pielou's Evenness index remains similar over the years for the structured (ABV) data but shows a strong influence of the datasets composing the unstructured data set (Fig. 11). After 2018 the waarnemingen.be data is no longer published on GBIF. This dataset contains a wide range of species from citizen science observations and once it is no longer included, the evenness drops due to the increased relative importance of datasets focused on a small subset of species from species targeted research projects. The observed change in the indicator therefore does not reflect an actual change in the evenness of birds in Flanders, but rather a change in the availability of specific data sources. Chapter 3.4 of this deliverable takes a closer look at the importance of this issue and analyses how individual datasets can affect biodiversity metrics.

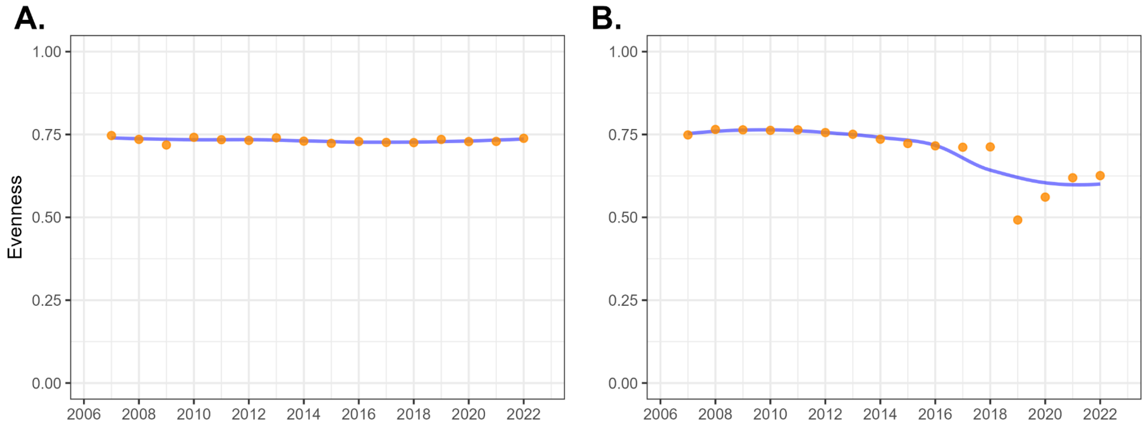


Figure 11: Trend of Pielou's Evenness over the years for the structured (A) and unstructured (B) data. The trendline is a LOESS smoother.

As expected, there is no trend in the estimated species richness for the structured (abv) dataset. The survey effort is constant by design of the monitoring scheme, and the observed species richness does not show any trend either. For the unstructured dataset the observed species richness gradually increases with time (Fig. 10). The increase in observed diversity is due to the increase in survey effort. The estimated diversity corrects for this using the rarefaction and extrapolation algorithm from the **iNEXT** package (Chao et al., 2014; Hsieh et al., 2016). The algorithm seems to eliminate the influence of survey effort well, even suggesting a slight decrease of estimated species richness in time (Fig. 12).

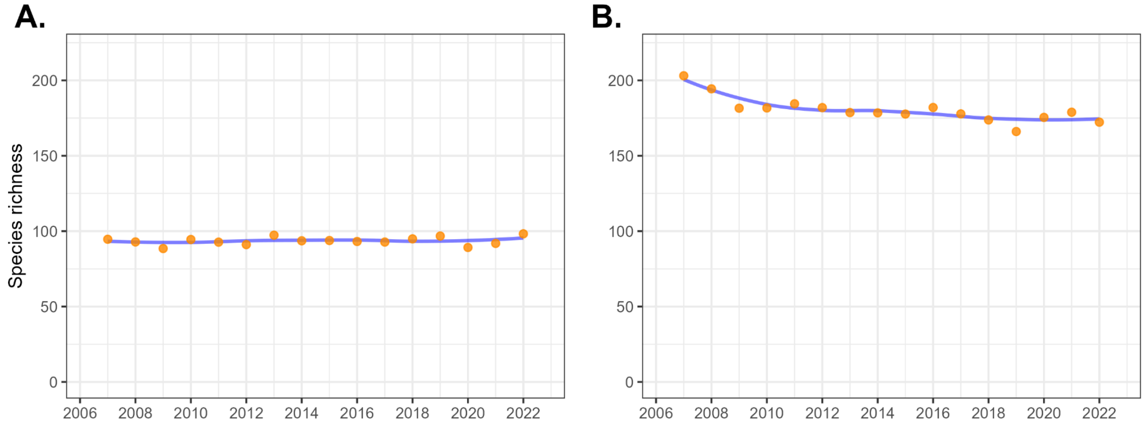


Figure 12: Trend of Hill0 estimate species richness over the years for the structured (A) and unstructured (B) data. The trendline is a LOESS smoother.

2.2.3.Conclusion

The comparison between structured (ABV) data and unstructured data shows that targeted filtering is required to obtain robust and meaningful results when analysing unstructured data. Thorough quality control is required, including checking for the loss of specific data sources and differences in publishing delay.

The ranges for rare species are much larger (and probably more complete) in the unstructured data than in the structured (ABV) data. This is to be expected because the structured monitoring was not designed for trend analysis of rare species. For the more common species, we see an equal range, and for the most common species, a slight underestimation of the range in the unstructured data.

The analysis of structured (ABV) data uses minimum criteria that relate to a minimum total number of observations, a minimum number of cells with observations, and observations in the same cell in consecutive years and this for each species. When these criteria are applied to the unstructured data, there’s some correlation of the trends between both datasets.

The patterns of the time series for the occurrences of individual species are different between the structured and unstructured datasets, mainly due to the increase in research effort. However, the relative frequency of the species in relation to the total number of occurrences within the same taxonomic order shows similar patterns. This suggests that it is possible to detect the most pronounced patterns, provided that corrections are made for survey effort.

The structured (ABV) data shows no overall trend in species richness. While there is an influence of increased survey effort on the observed species richness in the unstructured dataset, we do not find this trend in the estimated species richness (Hill0). The proxies for survey effort seem to work well. Metrics that take into account the extent to which species are evenly distributed within the species community (e.g., Pielou’s Evenness index) are very sensitive to the loss of certain datasets.

While these results demonstrate that survey effort proxies can mitigate bias, the high sensitivity of diversity metrics to the loss of specific datasets, such as the *Waarnemingen.be* drop-off after 2018, highlights a broader vulnerability. This suggests that the reliability of indicators is not just an ecological question, but also a technical one. Chapter 3 will therefore detail the specific data quality barriers, such as publication delays and taxonomic inconsistencies, which drive these observed patterns.

2.3. Western Cape case study

2.3.1. Exploration of data

Survey effort is not standardised in SABAP2, some grid cells have low survey effort and thus are unlikely to reflect true ecological patterns (Fig. 2). Survey effort data from SABAP2 (Fig. 2) and the `iNEXT()` function from the **iNEXT** package were used to get sample coverage estimates for each grid cell of the structured data cubes (Table 3). Sample coverage is a quantification of the sample completeness of the survey to assess the extent of undetected diversity. The function supports three types of data: abundance data (datatype="abundance"), raw incidence data (i.e., detection/non-detection data for each species; datatype = “incidence_raw”) and incidence-frequency data (i.e., total number of times each species was detected; datatype="incidence_freq"); with estimates for several diversity metrics available (i.e., Hill numbers). The structured data cubes provide information on the number of times each species was recorded in each grid cell, these are incidence-frequency data. The data were formatted to meet the requirements of **iNEXT** - lists of species incidence frequencies (one list per grid cell), with the first value of each list being the total number of surveys (i.e., number of sampling units). Sample coverage for q = 0 was calculated - i.e., the proportion of species that have been detected.

For the Western Cape sample coverage was generally high, but there were some grid cells with low coverage (Table 3). For the Hessequa area sample coverage was high for all grid cells. Grid cells with a sample coverage of < 0.9 were removed from the structured datasets before analysis. For alignment these grid cells were removed from the unstructured datasets.

Table 3: Mean sample coverage (with range in brackets) for the SABAP2 data from the Western Cape and Hessequa area at the two resolutions of analysis.

	Spatial extent	Resolution	Sample coverage
1	Western Cape	Pentad	0.91 (0.39 – 0.99)
2		QDGC	0.95 (0.004 – 0.999)
3	Hessequa area	Pentad	0.98 (0.96 – 0.99)
4		QDGC	0.99 (0.97 – 0.99)

For the Western Cape, 508 bird species and ~ 1 300 000 bird observations were recorded in SABAP2, with between 130 000 and 170 000 observations per year (Fig. 13). The unstructured data for the Western Cape included records for 562 bird species and ~ 1 200 000 bird observations, with the number of observations ranging between ~ 60 000 and 350 000 per year (Fig. 13).

In the Hessequa area, 323 bird species and ~ 110 000 bird observations were recorded in SABAP2, with between ~10 000 and 15 000 observations per year. For the same area, the unstructured data included records for 349 bird species and ~ 45 000 bird observations, with the number of observations ranging between ~ 2000 and 13 500 per year.

Overall, at both spatial scales there have been more observations in SABAP2 than in the unstructured data, but more species have been recorded in the unstructured data than in SABAP2 (partly due to issues highlighted in Chapter 3.2). The number of observations in the unstructured data has increased greatly over time, while the number has been more stable in the structured data (Fig. 13). The data were visually inspected to ensure that there was no drop off in observations in the downloaded data (Fig. 13).

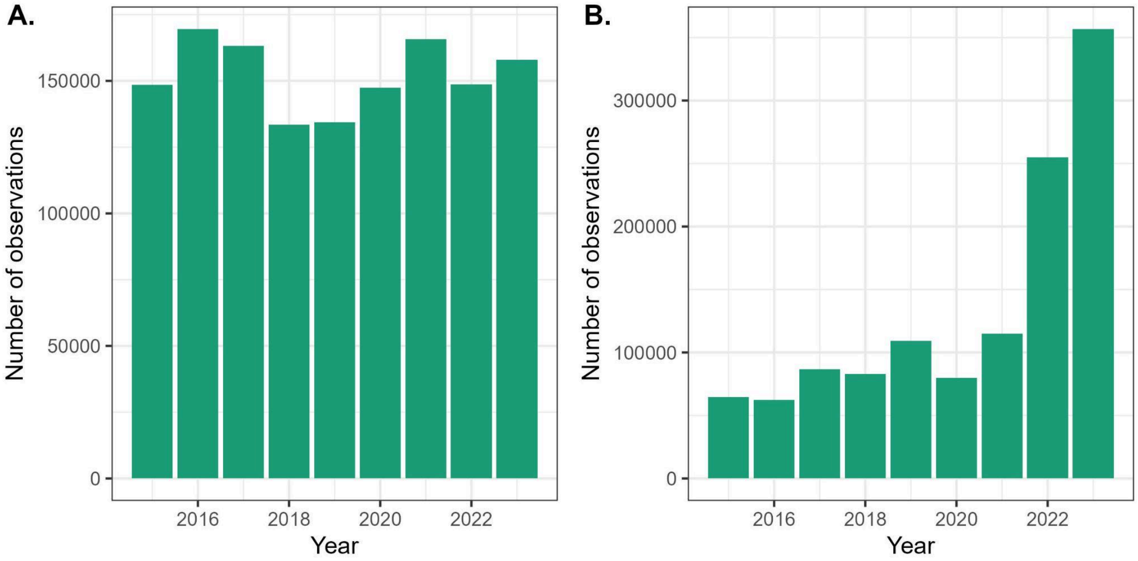


Figure 13: Number of observations per year in the structured (A) and unstructured (B) datasets for the Western Cape.

2.3.2. Results

Due to the nature of data collection, the data from the Western Cape were used in spatial comparisons.

2.3.2.1. Range

Range overlap

On average the ranges of the species based on SABAP2 data and unstructured data overlap by 22-69%, with this percentage being higher when analysed at larger spatial extents and coarser resolutions (Fig. 14). Most species for which there was 0 or 100% range overlap are very rare or rare species (Fig. 14).

Range extent

Range extent estimated using the unstructured data was positively and strongly correlated with range extent estimated using SABAP2 data (R^2 = 0.80-0.94; Fig. 15), with this correlation being stronger for analyses at larger spatial scales (Fig. 15). The resolution of the analysis had little effect on the results.

2.3.2.2. Diversity metrics

Observed species richness

The package **b3gbi** (Dove, 2026) was used for estimates of observed species richness. For the Hessequa area, observed species richness based on the unstructured data was not strongly, positively correlated with observed richness based on SABAP2 (QDGC: R^2 = -0.27, pentads: R^2 = 0.44), but the correlation was stronger, and positive for the Western Cape, particularly when the analysis was performed at coarser resolutions (QDGC: R^2 = 0.78 (Fig. 16A), pentad: R^2= 0.62).

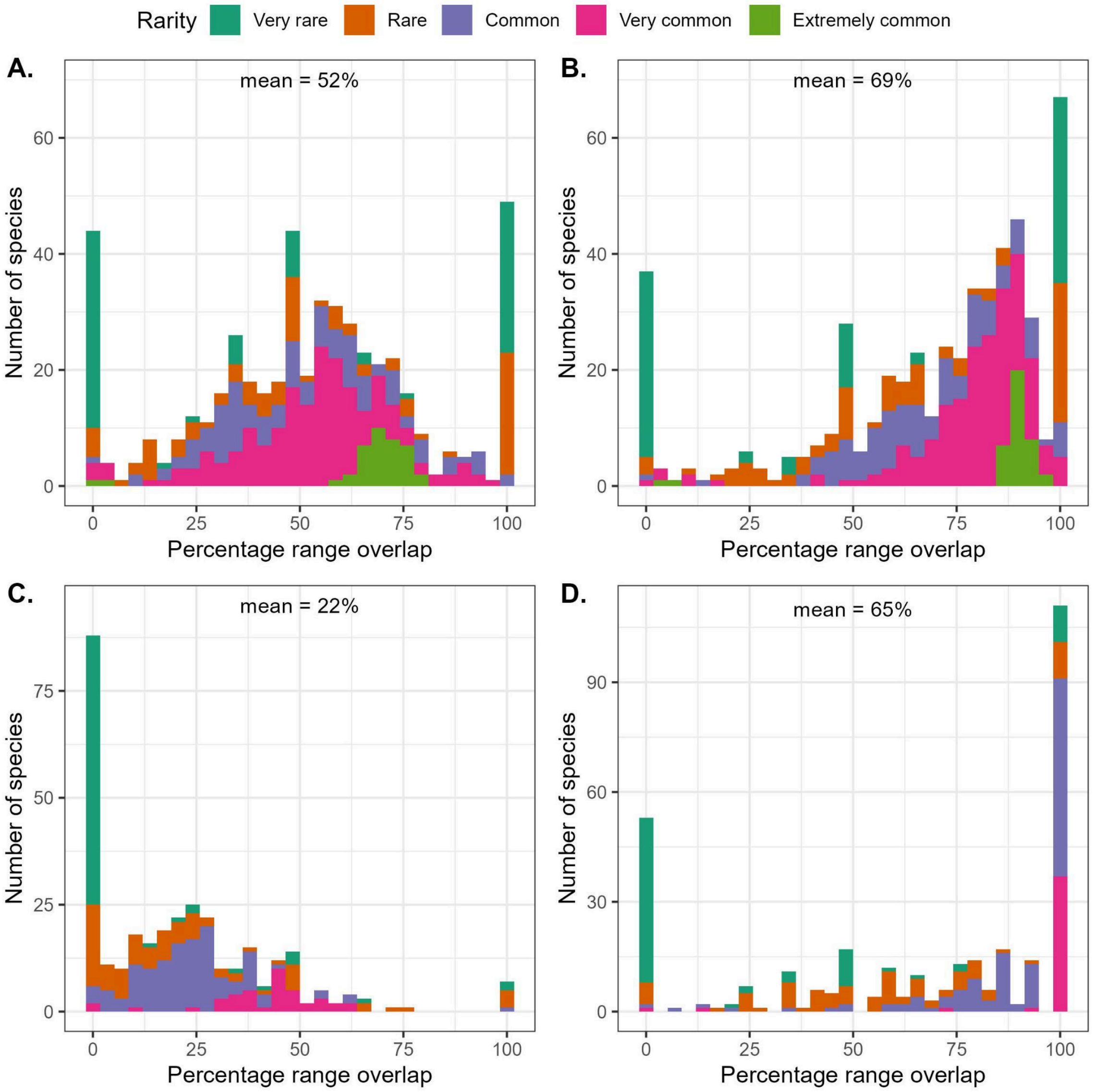


Figure 14: The percentage of range overlap - calculated as the percentage of SABAP2 grid cells in which a species has been recorded in both SABAP2 and the unstructured data - for A: Western Cape at pentad resolution and B: QDGC resolution, and C: area of Hessequa Systematic Atlasing Subproject at pentad resolution and D: QDGC resolution.

Hill numbers

Observed species richness based on the SABAP2 data is likely impacted by survey effort (Fig. 2) and thus may not reflect true trends. Therefore, we explored Hill number estimates (species richness, Shannon diversity and Simpson diversity) using the **iNEXT** package (which takes survey effort into account) and **b3gbi** (which uses survey effort proxies). As incidence data are recorded in SABAP2, the data type was set to "incidence_freq" for all estimates based on the SABAP2 data. Estimates using **iNEXT** for incidence data require the total number of sampling units for each assemblage (an estimate of survey effort), which for the SABAP2 data were the total number of surveys per grid cell.

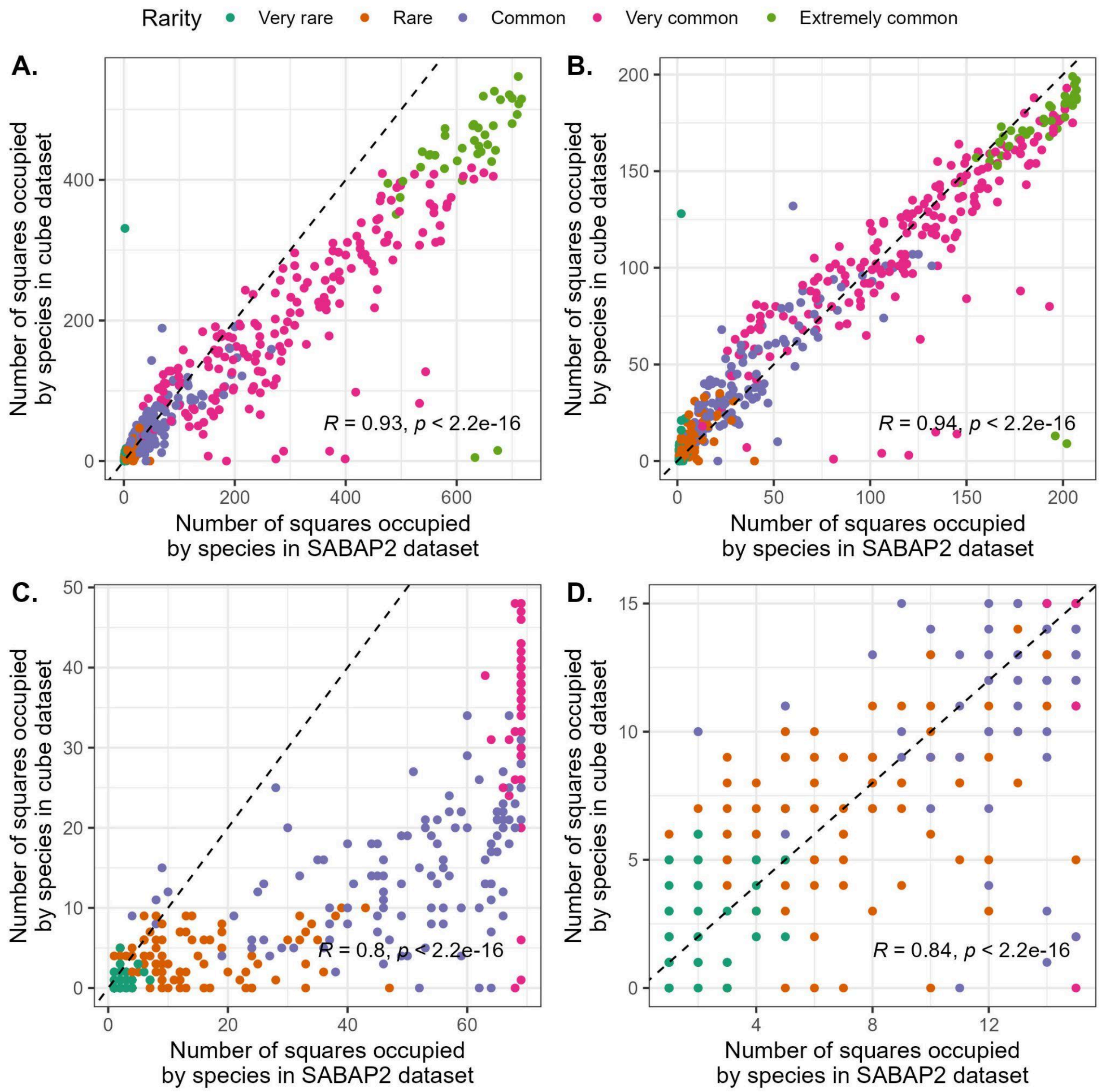


Figure 15: Correlation between range extent based on SABAP2 data and unstructured data - calculated as the number of SABAP2 grid cells in which a species is observed in SABAP2 and in which it is observed in the unstructured data - for A: Western Cape at pentad resolution and B: QDGC resolution, and C: area of Hessequa Systematic Atlasing Subproject at pentad resolution and D: QDGC resolution. The dashed lines are lines of equality (1:1).

When calculated using **b3gbi** there are several parameter options:

- select ‘data_type = incidence’ and ‘assume_freq = F’, which treats the data as presence/absence, and uses years as sampling units
- select ‘data_type = incidence’ and ‘assume_freq = T’, which assumes that the total number of observations of a species within a grid cell is equal to the number of sampling units in which that species was observed, and that the maximum number of observations for any species within a grid cell is equal to the total number of sampling units for that cell
- select ‘data_type = abundance’, which treats the species counts as an abundance proxy. The parameter ‘assume_freq’ is ignored if data_type is set to ‘abundance’

These analyses were only possible at a QDGC resolution, as estimates using a pentad grid are not possible through **b3gbi**.

The performance of **b3gbi** for estimates of Hill numbers was assessed by using **b3gbi** to calculate Hill numbers for the SABAP2 data, and comparing the estimates to those obtained for the same dataset using the **iNEXT** package and survey effort data from SABAP2. Generally, estimates calculated using **b3gbi** based on the SABAP2 data were strongly, positively correlated with estimates from **iNEXT** based on the same dataset (Table 4). These strong correlations indicate that the survey effort proxies implemented in **b3gbi** appear to work well. The strength of the correlations tended to be stronger for the Western Cape than the Hessequa area, but depended on the parameters used in **b3gbi** (Table 4).

Table 4: Correlation between Hill number estimates based on SABAP2 data for the Western Cape and Hessequa area when calculated using iNEXT (using survey effort data), and b3gbi (based on survey effort proxies). Note that various b3gbi parameter settings were tested (see text above for explanations of what these settings change) .

	Spatial extent	b3gbi parameters	Correlation for species richness	Correlation for Shannon index	Correlation for Simpson index
1	Western Cape	‘data_type = incidence’ AND ‘assume_freq = F’	$R^2 = 0.9$	$R^2 = 1$	$R^2 = 0.87$
2		‘data_type = incidence’ AND ‘assume_freq =T	$R^2 = 1$	$R^2 = 1$	$R^2 = 1$
3		‘data_type = abundance’	$R^2 = 1$	$R^2 = 1$	$R^2 = 1$
4	Hessequa area	‘data_type = incidence’ AND ‘assume_freq = F	$R^2 = 0.9$	$R^2 = 0.7$	$R^2 = 0.62$
5		‘data_type = incidence’ AND ‘assume_freq =T	$R^2 = 0.8$	$R^2 = 1$	$R^2 = 1$
6		‘data_type = abundance’	$R^2 = 0.8$	$R^2 = 1$	$R^2 = 0.99$

Estimated species richness

Observed and estimated species richness based on the unstructured data calculated using **b3gbi** were compared to estimated species richness based on SABAP2 using **iNEXT** (Fig. 16). **b3gbi** calculates diversity estimates for any specified sample coverage (a standardised level of sample completeness). Various parameters and sample coverage values were implemented in **b3gbi** to get estimates for the unstructured data, with the results based on the best sample coverage value (that which resulted in the strongest correlation) for each data type assessed (incidence and abundance) being discussed below and shown in Table 5.

Observed species richness based on the unstructured data was not strongly correlated with estimated species richness based on SABAP2, although the correlation was stronger at larger spatial scales (Table 5). This analysis, which corrected the structured data for survey effort, showed weaker correlations (Fig. 16B) than when the raw observed species richness values were analysed (Fig. 16A).

Table 5: Correlation between estimated species richness based on SABAP2 data (calculated using iNEXT) and observed and estimated species richness (calculated using b3gbi) based on unstructured data for the Western Cape and Hessequa area. Note that various b3gbi parameter settings were tested.

Spatial extent	Unstructured data	b3gbi parameters	Correlation
Western Cape	Observed species richness	NA	R^2 = 0.66
	Estimated species richness	‘data_type = incidence’ AND ‘assume_freq = FALSE’ AND ‘coverage = 0.05’	R^2 = 0.55
		‘data_type = abundance’ AND ‘coverage = 0.95’	R^2 = 0.70
Hessequa area	Observed species richness	NA	R^2 = -0.0018
	Estimated species richness	data_type = incidence’ AND ‘assume_freq = FALSE’ AND ‘coverage = 0.15’	R^2 = 0.13
		‘data_type = abundance’ AND ‘coverage = 0.55’	R^2 = 0.15

At the smaller spatial scale (i.e., the Hessequa area), estimated richness based on the unstructured data was not correlated with estimated species richness based on SABAP2, no matter the implemented parameters (Table 5). At the larger spatial scale (Western Cape) estimated species richness based on the unstructured data, with ‘data_type’ set to ‘abundance’ was positively and strongly correlated with estimated species richness based on SABAP2 (Table 5; Fig. 16D). Note that when ‘data_type’ was set to ‘incidence’ the sample coverage value that gave the best result was 0.05 (therefore had to be reduced significantly from the default of 0.95); but when set to ‘abundance’ implementing the default sample coverage value (0.95) gave the strongest correlation (Fig. 16C vs Fig. 16D). Although the strongest correlation was between observed species richness based on unstructured data and SABAP2 data (Fig. 16A), SABAP2 data are biased, and thus the correlations with estimated species richness that take survey effort into account are likely to give a truer picture.

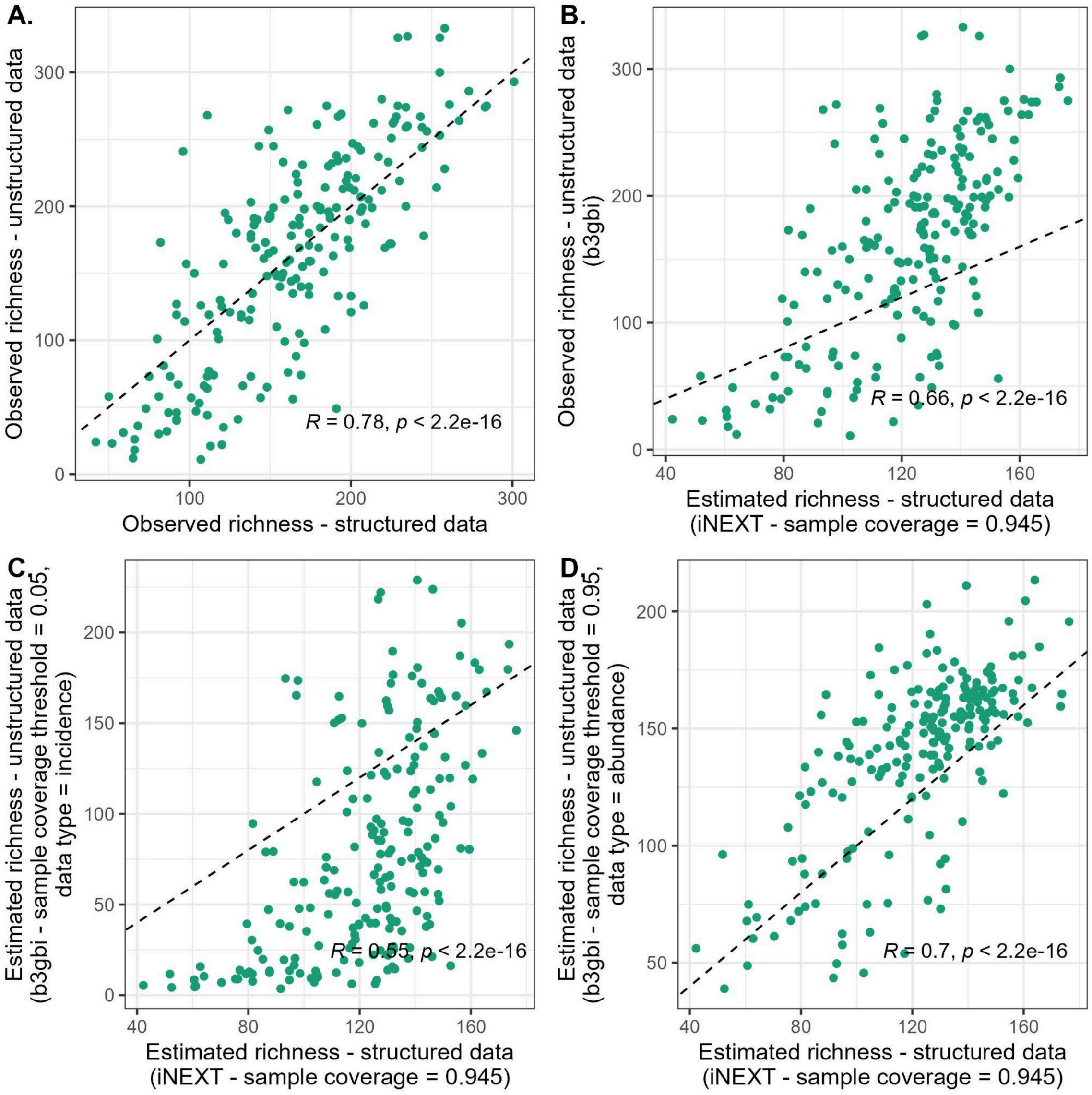


Figure 16: Correlations, for the Western Cape at a QDGC resolution, between A: observed species richness based on unstructured data and SABAP2; B: observed species richness based on unstructured data and estimated species richness based on SABAP2 (corrected for survey effort using iNEXT); C: estimated species richness based on unstructured data (using b3gbi with ‘data_type = incidence’) and estimated species richness based on SABAP2 (corrected for survey effort using iNEXT); and D: estimated species richness based on unstructured data (using b3gbi and ‘data_type = abundance’) vs estimated species richness from SABAP2 (i.e., corrected for survey effort using iNEXT). The dashed lines are lines of equality (1:1).

Shannon Index

As for species richness estimates, Shannon index estimates based on the unstructured data calculated using **b3gbi** were compared to estimates based on SABAP2 using **iNEXT** (Fig. 17). The same methods were followed as for estimated species richness - various parameters and coverage values were implemented in **b3gbi**.

For the Hessequa area, Shannon index estimated based on the unstructured data was not correlated with Shannon index estimated based on SABAP2 data, no matter the parameters (Table 6). For the Western Cape the correlation was moderate and, as for species richness estimates, was highest when ‘data_type = abundance’ and ‘coverage = 0.95’ (the default) (Table 6; Fig. 17B). Similar to species richness estimates, when ‘data_type’ was set to ‘incidence’ sample coverage had to be reduced significantly to improve performance (Table 6; Fig 17A).

Table 6: Correlation between Shannon index calculated based on SABAP2 data (using iNEXT and survey effort) and unstructured data (using b3gbi) for the Western Cape and Hessequa area. Note that various b3gbi parameter settings were tested.

Spatial extent	b3gbi parameters	Correlation
Western Cape	‘data_type = incidence’ AND ‘assume_freq = FALSE’ AND ‘coverage = 0.05’	R^2 = 0.56
	‘data_type = abundance’ AND ‘coverage = 0.95’	R^2 = 0.65
Hessequa area	data_type = incidence’ AND ‘assume_freq = FALSE’ AND ‘coverage = 0.45’	R^2 = 0.15
	‘data_type = abundance’ AND ‘coverage = 0.15’	R^2 = -0.012

Simpson Index

As for species richness estimates and Shannon index, Simpson index estimates based on the unstructured data calculated using **b3gbi** were compared to estimates based on SABAP2 using **iNEXT** (Fig. 18). The same methods were followed as for the other diversity metrics - various parameters and coverage values were implemented in **b3gbi**.

For the Hessequa area, Simpson index estimated based on the unstructured data was not correlated with Simpson index estimated based on SABAP2 data, no matter the parameters (Table 7). For the Western Cape the correlation was moderate, and was highest when ‘data_type = abundance’ and ‘coverage = 0.5’ (Table 7; Fig. 18B). In contrast to the other diversity metrics, for both data types (incidence and abundance) sample coverage had to be reduced significantly from the default (0.95) to improve performance (Table 7; Fig. 18). When data type was set to ‘abundance’ and coverage was set to the default the correlation was 0.52, in comparison to 0.59 when coverage was set to 0.5 (Table 7).

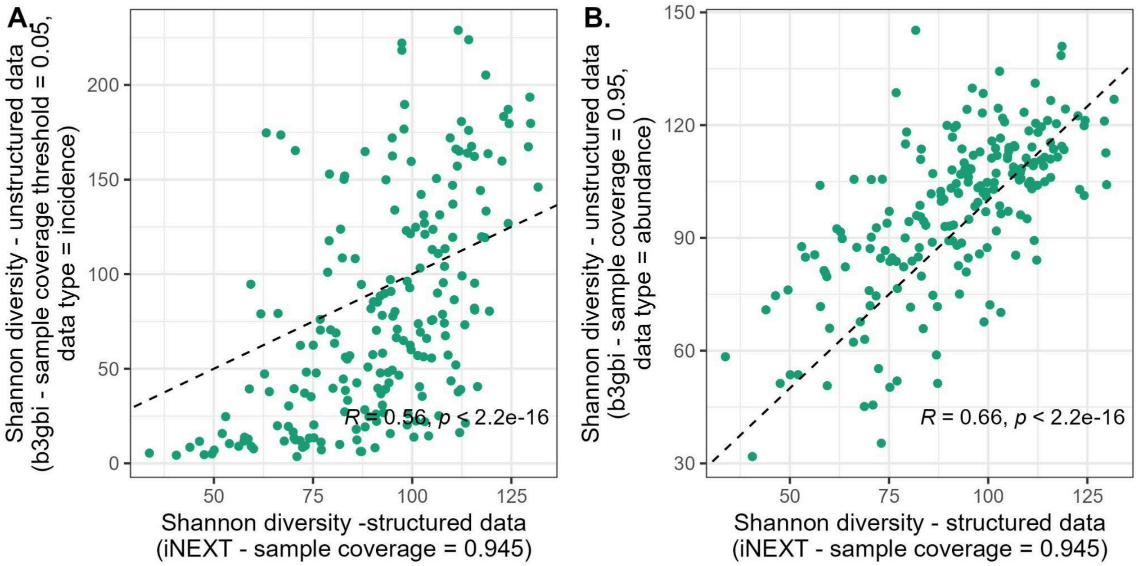


Figure 17: Correlations, for the Western Cape at a QDGC resolution, for Shannon index estimated based on A: unstructured data (using b3gbi and ‘data_type = incidence’) and SABAP2 (corrected for survey effort using iNEXT), and B: unstructured data (using b3gbi and ‘data_type = abundance’) and SABAP2 (corrected for survey effort using iNEXT). The dashed lines are lines of equality (1:1).

Table 7: Correlation between Simpson index calculated based on SABAP2 data (using iNEXT and survey effort) and unstructured data (using b3gbi) for the Western Cape and Hessequa area. Note that various b3gbi parameter settings were tested.

Spatial extent	b3gbi parameters	Correlation
Western Cape	‘data_type = incidence’ AND ‘assume_freq = FALSE’ AND ‘coverage = 0.05’	R^2 = 0.56
	‘data_type = abundance’ AND ‘coverage = 0.5’	R^2 = 0.59
Hessequa area	data_type = incidence’ AND ‘assume_freq = FALSE’ AND ‘coverage = 0.45’	R^2 = 0.20
	‘data_type = abundance’ AND ‘coverage = 0.05’	R^2 = 0.0057

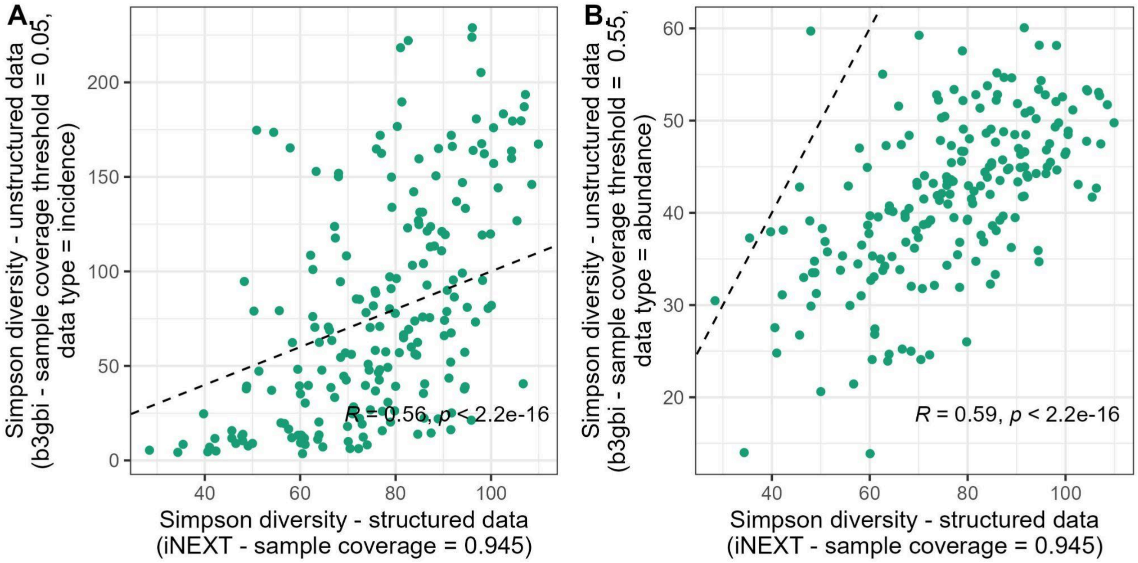


Figure 18: Correlations, for the Western Cape at a QDGC resolution, for Simpson index estimated based on A: unstructured data (using b3gbi and ‘data_type = incidence’) and SABAP2 (corrected for survey effort using iNEXT), and B: unstructured data (using b3gbi and ‘data_type = abundance’) and SABAP2 (corrected for survey effort using iNEXT). The dashed lines are lines of equality (1:1).

2.3.3. Conclusion

The comparative analysis for the Western Cape case study showed that range estimates based on unstructured data appear more reliable when analysed at larger spatial scales and coarser resolutions. However, range estimates based on unstructured data are likely to be less reliable for very rare and rare species than for species that are more common, even at large spatial scales and coarse resolutions. When using unstructured data, diversity metrics should be based on estimates that take survey effort (or survey effort proxies) into account, e.g., rather calculate estimated species richness than observed species richness. The diversity metric estimates from b3gbi and, therefore, the related survey effort proxies appear to perform well. Diversity metrics based on unstructured data appear more reliable when analysed at larger spatial scales and, when estimated using b3gbi, when data_type is set to 'abundance'.

The finding that range estimates are more reliable at coarser resolutions confirms that spatial 'noise' at fine scales can distort ecological signals. To address this, the next chapter establishes rigorous criteria for handling coordinate uncertainty to ensure that the spatial scales of analysis match the precision of the underlying data.

3. Technical challenges and data quality barriers

The transition from raw data to the use of aggregated GBIF occurrence cubes introduces several technical and qualitative challenges that must be addressed to ensure the reliability of biodiversity indicators. This chapter explores the primary barriers encountered before any analysis of the Flemish and Western Cape datasets.

3.1. Data is out of bounds of selected area

3.1.1. Handling coordinate uncertainty in occurrence cubes

Occurrences with large coordinate uncertainty (`coordinateUncertaintyInMeters`) can lead to observations being aggregated outside the intended study area, even when country-level filters are applied. This artefact arises because grid assignment randomly samples a location within the reported uncertainty radius, which may extend beyond the spatial extent used in the query. If not handled explicitly, this can result in misleading spatial patterns and biased indicators (Fig. 19).

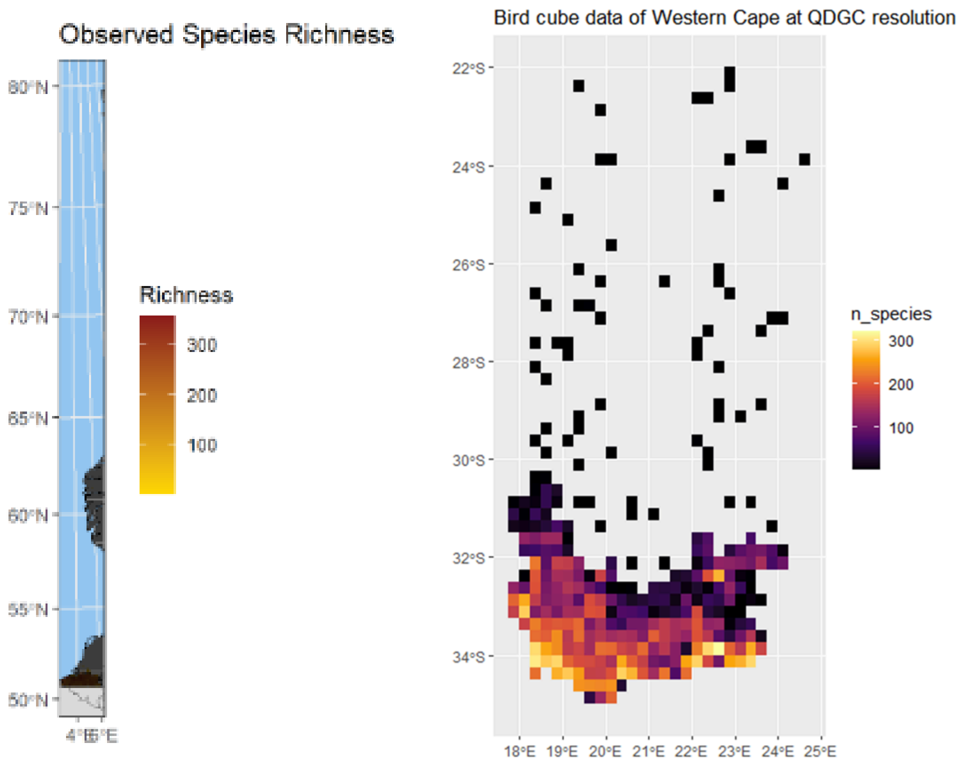


Figure 19: Preliminary analyses of biodiversity indicators gave strange patterns due to observations with large coordinate uncertainties. Left: Flanders. Right: Western Cape province of South Africa.

The appropriate treatment of coordinate uncertainty depends fundamentally on the spatial scale of the analysis and on whether the indicator is intended to describe spatial patterns or temporal trends. There is no single universally optimal solution; instead, the handling of coordinate uncertainty represents a trade-off between spatial precision and data retention that must be made explicit.

3.1.2. Implications for spatial and temporal indicators

For spatial indicators, where the objective is to quantify spatial variation, gradients, or hotspots, spatial precision is essential. Observations with coordinate uncertainties that exceed the grid resolution introduce uncertainty that is larger than the scale of the pattern under investigation. A practical and conservative rule of thumb is therefore to exclude records with a coordinate uncertainty larger than one edge length of the grid used for aggregation, for example greater than 1000 m when using a 1 km² grid (Fig. 20). This ensures that any randomisation of locations remains limited to neighbouring grid cells and does not overwhelm the spatial signal. Records with missing coordinate uncertainty should be treated cautiously in this context and are best excluded at fine spatial scales, as their true uncertainty is unknown and may be substantial. Although this approach can result in considerable data loss (~70% of the data for Flanders on 1 km² scale), it ensures that retained observations contribute meaningfully and consistently to spatial analyses.

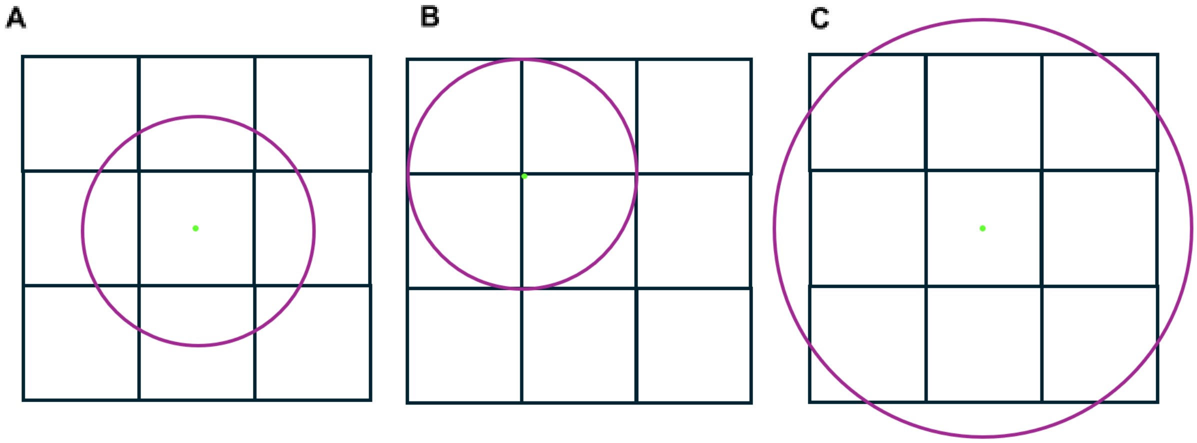


Figure 20: A. A point with coordinate uncertainty smaller than the edge of the grid will only ever result in that point being assigned to one of the neighbouring grid cells, even if the point is located at the very edge (B). C. Allowing points with coordinate uncertainties greater than the edge of the grid leads to high uncertainty compared to the scale of the pattern you want to analyse.

For temporal indicators, where the focus is on changes over time across the full spatial extent of the cube rather than on precise spatial allocation, the requirements differ. In this case, retaining observations is often preferable to avoid temporal bias caused by disproportionate data loss. Larger coordinate uncertainties, or even missing uncertainty values, may be acceptable provided that their inclusion does not distort the temporal signal. Methods that avoid random spatial assignment within uncertainty, or that tolerate greater uncertainty while aggregating temporally, are therefore more appropriate for this type of indicator.

3.1.3. Resolution-dependent recommendations and final guidance

The spatial resolution of the analysis, i.e., the grid resolution, further determines how both large and missing coordinate uncertainties should be handled (Tables 8, 9). At fine spatial resolutions, both large and missing uncertainties are generally unsuitable and should be excluded. At coarse spatial resolutions, where grid cells are much larger than typical coordinate uncertainties, it may be acceptable to retain such records, for example by assigning large and missing uncertainties a conservative value equal to the grid resolution. At intermediate resolutions, a hybrid approach may be warranted, in which extremely large uncertainties are filtered out while missing uncertainties are assigned a conservative, resolution-dependent value.

Table 8: Overview of recommended treatment of coordinate uncertainty by analysis resolution for spatial indicators. “Large coordinate uncertainty” refers to uncertainties exceeding one edge length of the grid used for aggregation. Assigning a “conservative value” to missing coordinate uncertainty implies setting it equal to the spatial resolution of the analysis.

Grid resolution	Large coord. uncertainty	Missing coord. uncertainty	Rationale
Fine	Exclude	Exclude	Spatial uncertainty exceeds the resolution of the pattern; inclusion leads to artefacts and unreliable spatial signals.
Intermediate	Exclude	Assign conservative value or exclude	Balance between spatial precision and data retention; avoid uncertainties larger than the analysis resolution.
Coarse	Assign conservative value	Assign conservative value	Grid cells are large relative to uncertainty; spatial signal remains robust.

Table 9: Overview of recommended treatment of coordinate uncertainty by analysis scale for temporal indicators.

Grid resolution	Large coord. uncertainty	Missing coord. uncertainty	Rationale
Fine	Exclude	Exclude	Spatial misallocation can bias temporal trends at fine resolution; conservative filtering is preferred.
Intermediate	Retain with caution or assign conservative value	Assign conservative value	Temporal signal dominates; limited spatial imprecision is acceptable if documented.
Coarse	Retain or assign conservative value	Assign conservative value	Maximising data retention is preferred; spatial precision is less critical

The resolution-dependent recommendations outlined above implicitly assume that the reported coordinates represent a meaningful estimate of the true location of the occurrence, such that the uncertainty radius reflects imprecision around an otherwise informative spatial reference. In practice, however, large coordinate uncertainties may arise from fundamentally different situations. In some cases, a large uncertainty still conveys useful spatial information at coarse resolutions, because the central coordinate remains a reasonable approximation of the occurrence location within the study area. In other cases, large uncertainties may reflect placeholder or non-informative coordinates, where the true location is effectively unknown within the spatial extent of the analysis.

The guidelines presented here are intended to address the former situation, in which uncertainty can be meaningfully related to analysis resolution. Records for which the reported coordinates do not provide a meaningful spatial reference, regardless of uncertainty radius, are not suitable for spatially explicit analyses at any resolution and should be excluded during data preparation. Distinguishing between these cases requires exploratory assessment of coordinate distributions and uncertainty patterns and cannot be resolved solely through threshold-based filtering.

In conclusion, coordinate uncertainty should never be handled implicitly. Its treatment must be explicitly aligned with the spatial resolution of the analysis and the intended use of the indicator, while recognising that not all large uncertainties are equally informative. Conservative filtering is recommended for spatial indicators, where spatial precision is critical, while more inclusive approaches are acceptable for temporal indicators, provided that the underlying assumptions and limitations of the data are clearly documented and justified.

3.2. Taxonomic inconsistencies and species name issues

3.2.1. Limitations of taxonomic harmonisation in GBIF

When comparing structured data with unstructured cube data for Flanders (Belgium), we identified several cases in which different scientific names referring to the same biological species were treated as distinct taxa. Examples include *Poecile montanus* and *Parus montanus*, (willow tit) as well as *Dendrocopus major* and *Dendrocopos major* (great spotted woodpecker). In these cases, both names are listed as accepted in the GBIF backbone taxonomy but are not linked as taxonomic equivalents. Consequently, occurrences are split across multiple names, leading to artificial inflation of species richness and inconsistencies between datasets.

Similar issues were observed in unstructured occurrence cubes for the Western Cape and the Hessequa area. For instance, *Tychaedon coryphoeus* (Karoo scrub robin) is represented by two accepted names in GBIF (*Tychaedon coryphoeus* (Lesson, 1831) and *Erythropygia coryphoeus* (Vieillot, 1817)). Because these names are not taxonomically linked, records are treated as belonging to different species in downstream analyses.

Resolving such cases requires submitting taxonomic issue reports to GBIF. However, identifying unlinked accepted names is often non-trivial and time-consuming, particularly in large datasets. Moreover, updates to the GBIF backbone taxonomy are implemented infrequently, typically on a semi-annual basis, and must accommodate a large volume of taxon-related issues.

3.2.2. Errors introduced during data publishing and taxonomic translation

Additional inconsistencies can arise during the data publishing process, when species names are incorrectly translated or mapped to the GBIF backbone. In the published ABV dataset, for example, occurrences attributed to *Saxicola torquatus* (African stonechat) are likely misassigned and should be corrected to *Saxicola rubicola* (European stonechat). In the published SABAP2 dataset, several species were missing entirely, while many records had missing species-level identification despite containing genus-level information.

Closer inspection revealed that these issues are likely caused by problems in the taxonomic mapping applied during publication to GBIF. For example, numerous SABAP2 records list the genus as "*Zosterope*", which GBIF treats as a synonym of *Zosterops* Vigors & Horsfield, 1827. Despite this, *Zosterops virens* (Cape white-eye) does not appear in the SABAP2 data downloaded from GBIF, even though the species is well represented in the Western Cape on the SABAP2 platform itself. Such inconsistencies result in incomplete or distorted species representations in derived data products.

Addressing these issues typically requires direct communication with the data publisher to correct taxonomic mappings at the source. As with backbone taxonomy issues, identifying affected taxa can be difficult and labour-intensive.

3.2.3. Species misidentifications in occurrence data

Finally, some discrepancies reflect genuine misidentifications rather than taxonomic or publishing artefacts. In the unstructured cube data for the Western Cape and the Hessequa area, several species were present that do not occur in the corresponding SABAP2 datasets for these regions. In some cases, this can be explained by likely field misidentifications.

A notable example is *Certhilauda curvirostris* (Cape lark), a range-restricted species that does not occur in the Hessequa area but appears in the unstructured occurrence cube. These records are plausibly misidentifications of *Certhilauda brevirostris* (Agulhas lark), a morphologically similar species that does occur in the area. Such errors can propagate into aggregated indicators if not detected, particularly when using opportunistic or unstructured data sources.

3.3. Temporal coverage and publication delays in GBIF data

Substantial delays between data collection and data publication in GBIF can affect the temporal completeness of both structured and unstructured data. These delays were observed consistently across the datasets analysed and can lead to mismatches between the intended temporal scope of an analysis and the actual coverage of the data retrieved from GBIF.

This issue was evident in both the structured and unstructured data used in this study. During preliminary analysis undertaken in 2025, it was found that the SABAP2 data downloaded from GBIF contained records only up to February 2023. This occurred despite SABAP2 being a dataset that is regularly published and updated in GBIF. Similarly, the unstructured occurrence cube contained relatively few observations from 2024, reflecting delays in the publication of recent records to GBIF.

Such publication lags can have important implications for analyses that rely on recent data, particularly temporal indicators and trend assessments. If not explicitly accounted for, these delays may lead to erroneous conclusions about recent changes, including apparent declines or plateaus that are in fact artefacts of incomplete data ingestion.

Users should therefore explicitly verify the effective temporal coverage of GBIF-derived data cubes before analysis and document any discrepancies between the intended and realised time periods. Where recent data are critical, complementary data sources or direct access to publisher datasets may be required to ensure adequate temporal completeness.

3.4. Disproportionate influence of component datasets

3.4.1. Introduction

GBIF hosts thousands of datasets that vary widely in spatial extent, taxonomic scope, and temporal coverage, ranging from single-species datasets to large, multi-taxon collections. Occurrence cubes are generated for a defined geographic range, taxonomic group, and time period by aggregating occurrence records from a selected subset of datasets published on GBIF. These cubes may fully overlap with some component datasets, while only partially covering others. This heterogeneity is a key reason why occurrence cubes are best understood as a standardised aggregation of heterogeneous data sources.

It is important to note that the technical issues mentioned in previous sections are often tied to specific component datasets. For instance, a delay in one major dataset can cause a sudden, artificial drop in both the number of records and the number of species detected in a cube. This makes it essential to analyse how much individual datasets influence the final indicators.

When an occurrence cube is generated through GBIF, users are provided with metadata on the contributing datasets, including their DOIs, citations, and the number of records included in the cube (see, for example, the links in Table 1). However, while this information supports transparency and reproducibility, it is not always directly informative for subsequent indicator calculations.

In this chapter, we investigate the influence of individual component datasets on indicator results through group-level sensitivity analysis.

3.4.2. Methods

3.4.2.1. Data

During the data exploration phase of the Flanders case study, substantial differences were observed in the number of occurrence records contributed by individual component datasets. To support the present analysis, we therefore constructed a dedicated occurrence cube that explicitly retains dataset-level information. This cube includes an additional dataset dimension, allowing occurrence records to be distinguished by their source dataset within each combination of year, grid cell, and species. As a result, the data are represented in a four-dimensional structure rather than the conventional three-dimensional cube (year × grid cell × species). The SQL query used to generate the cube is provided in Annex 1 and is also accessible via the DOI listed in Table 1.

All code is publicly available on GitHub (Langeraert et al., 2026). The complete GitHub repository with all data files is available on Zenodo (v1.0.0, 10.5281/zenodo.18782579).

3.4.2.2. Leave-one-dataset-out cross-validation

To assess the influence of individual component datasets on indicator outcomes, we performed a group-level sensitivity analysis for species prevalence (the proportion of occupied grid cells per species) using leave-one-dataset-out cross-validation as implemented in the **dubicube** R package (Langeraert & Van Daele, 2026). Cross-validation is a resampling technique that evaluates the robustness of indicators to the inclusion or exclusion of predefined groups, here individual component datasets contributing to the data cube.

Let $(\mathbf{X})$ denote the full biodiversity data cube, containing species occurrence information aggregated on a common spatial grid and grouped by a variable `dataset` identifying the component dataset of origin. For each species (i), prevalence is defined as the proportion of grid cells in which the species is recorded at least once across all component datasets. Formally, species prevalence is calculated as:

$$p_i = \frac{1}{C} \sum_{c=1}^{C} \mathbb{I}\left(\sum_{s=1}^{S} Z_{isc} > 0 \right)$$

where (C) is the total number of grid cells in the study area, (S) is the number of component datasets, and (Z_{isc}) is a binary variable indicating the presence (1) or absence (0) of species (i) in grid cell (c) as reported by dataset (s). The indicator function $(\mathbb{I}(\cdot))$ takes the value 1 if its argument is true and 0 otherwise. In this context, it evaluates whether the species has been observed at least once in a given grid cell across all datasets. As a result, each grid cell contributes either one (occupied) or zero (unoccupied) to the prevalence calculation, regardless of the number of observations or datasets reporting the species in that cell.

The prevalence estimate obtained from the full data cube is denoted by $(\hat{p}_i)$. To assess the sensitivity of this estimate to individual component datasets, we applied leave-one-dataset-out cross-validation. In each iteration, a reduced data cube $(\mathbf{X}_{-d_j})$ was constructed by removing all records associated with component dataset (j), while retaining the same spatial grid and species set. Species prevalence was then recalculated for each species using this reduced data cube, yielding an estimate $(\hat{p}_{i,-d_j})$.

To summarise the variability of prevalence estimates across cross-validation runs, we calculated, for each species, the root mean squared error (RMSE) and the mean relative error (MRE) between the leave-one-dataset-out estimates and the full-data estimate:

$$\mathrm{RMSE}_i = \sqrt{\frac{1}{S}\sum_{j=1}^{S}\left(\hat{p}_{i,-d_j} - \hat{p}_i\right)^2}$$

$$\mathrm{MRE}_i = \frac{1}{S}\sum_{j=1}^{S}\frac{|\hat{p}_{i,-d_j} - \hat{p}_i|}{\hat{p}_i}$$

RMSE captures the absolute magnitude of change in prevalence when datasets are excluded, while MRE expresses this change relative to the full-data prevalence, thereby facilitating comparisons across species with different baseline prevalence levels.

Finally, we examined whether the sensitivity of prevalence estimates was related to how evenly species occurrences were distributed across component datasets. For this purpose, we quantified dataset evenness for each species using Pielou's Evenness index, which is based on the normalised Shannon entropy of occurrences across datasets. Dataset evenness ranges from 0, indicating that occurrences are highly concentrated in a single dataset, to 1, indicating that occurrences are evenly distributed across all contributing datasets.

For species (i), Pielou's Evenness index (J_i) is calculated as:

$$J_i = \frac{-\sum_{s=1}^{S} p_{i,s}^{\mathrm{occ}} \ln(p_{i,s}^{\mathrm{occ}})}{\ln(S_i)}$$

where (S_i) is the number of component datasets in which species (i) occurs and ($p_{i,s}^{\mathrm{occ}}$) is the proportion of occurrences of species (i) contributed by dataset (s).

We then assessed the correlation between dataset evenness (J_i) and the cross-validation error metrics (RMSE and MRE) using linear regression, to evaluate whether species whose occurrences are unevenly distributed across datasets are more sensitive to the exclusion of individual component datasets.

3.4.2.3. Improvement score relative to ABV reference values

To evaluate whether leave-one-dataset-out cross-validation improves prevalence estimates relative to an external reference, we compared the CV-based estimates to ABV prevalence, which serves as the "true" benchmark. Let (p_i^{true}) denote the reference prevalence of species (i), ($\hat{p}_i$) the prevalence estimated from the full data cube, and ($\hat{p}_{i,-d_j}$) the prevalence estimated when component dataset (j) is omitted.

We first computed the absolute errors of the original and CV estimates relative to the reference:

$$e_i = \left|\hat{p}_i - p_i^{\text{true}}\right|, \quad e_{i,-d_j} = \left|\hat{p}_{i,-d_j} - p_i^{\text{true}}\right|$$

From these, an improvement score was defined for each species and dataset:

$$\Delta e_{i,-d_j} = e_i - e_{i,-d_j}$$

Positive values ($\Delta e_{i,-d_j} > 0$) indicate that omitting dataset (j) moved the prevalence estimate closer to the true value, zero indicates no change, and negative values ($\Delta e_{i,-d_j} < 0$) indicate that the CV estimate worsened relative to the reference. Across all species and datasets, we summarised improvement using the proportion of cases with $\Delta e_{i,-d_j} > 0$.

3.4.3. Results

3.4.3.1. Exploration

The occurrence cube comprises 40 component datasets (Fig. 21). Contribution levels are highly uneven: the largest dataset, *Waarnemingen.be*, accounts for 74.7 % of all occurrence records, and the six largest datasets together contribute 96.8 % of the total. The remaining 34 datasets each contribute relatively few records. When considering species richness rather than record counts (Fig. 21, right panel), a different pattern emerges. Dataset size in terms of number of observations is not necessarily correlated with the number of species represented. Large datasets do not always cover many species, and conversely, smaller datasets may include a comparatively high species richness.

The dominant contribution of *Waarnemingen.be* reflects its regional focus on Flanders and its broad taxonomic scope as a citizen science initiative, resulting in both a high number of records and a large number of species. In contrast, the iNaturalist dataset contributes relatively few records within the study area, reflecting its global scope, but still includes a large number of species due to its similarly broad, citizen science–based taxonomic coverage. Conversely, the second-largest dataset in terms of record count contributes many observations but is restricted to a single species (bird tracking data of *Larus argentatus*, European herring gull).

Over time, we see an increase of observations except for the steep drop-off after 2018 caused by the loss of the largest dataset (Fig. 22). Other datasets, such as the gull tracking dataset, seem to be temporary contributions limited to the duration of the project.

3.4.3.2. Leave-one-dataset-out cross-validation

When we compare the prevalence values between the unstructured cube data and the structured ABV data, we see a higher prevalence in general in the unstructured data (Fig. 23A). Only the most common species have a higher prevalence in the ABV data.

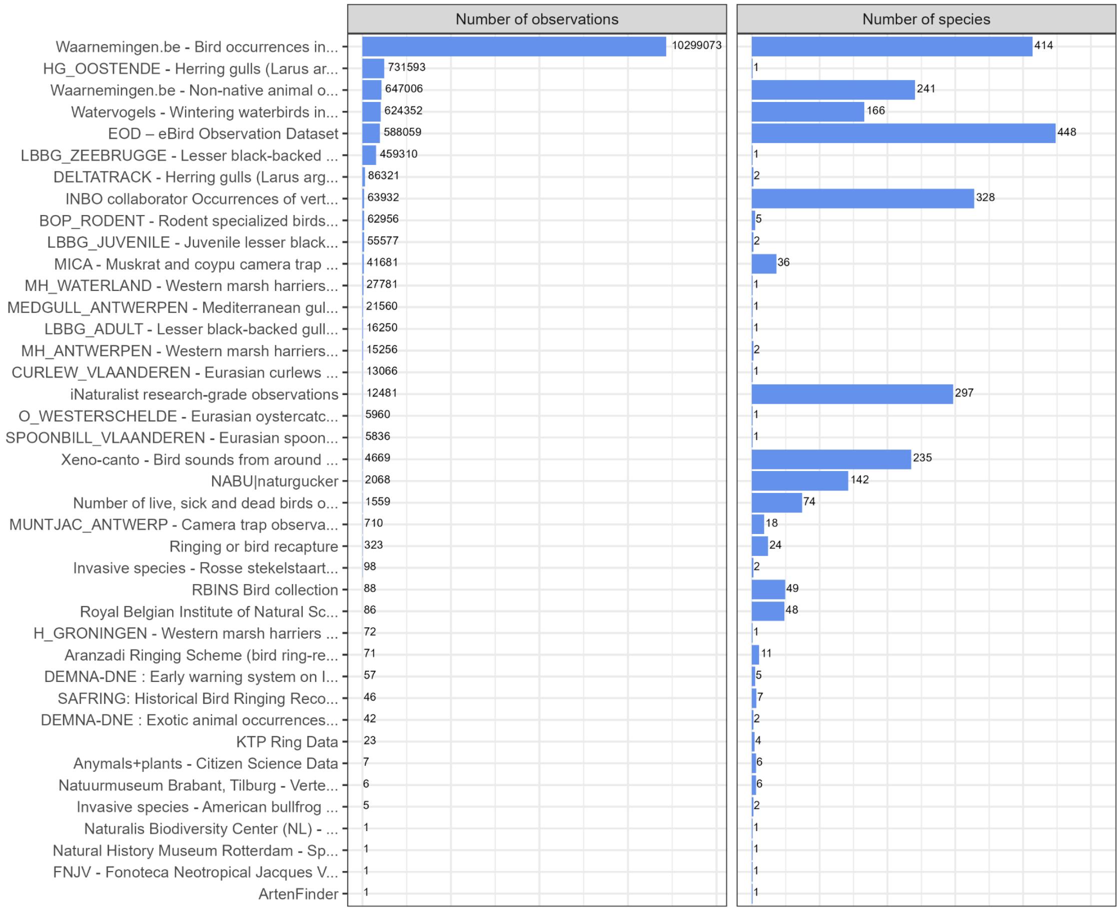


Figure 21: Number of observations and species per component dataset of the occurrence cube of the Flanders case study.

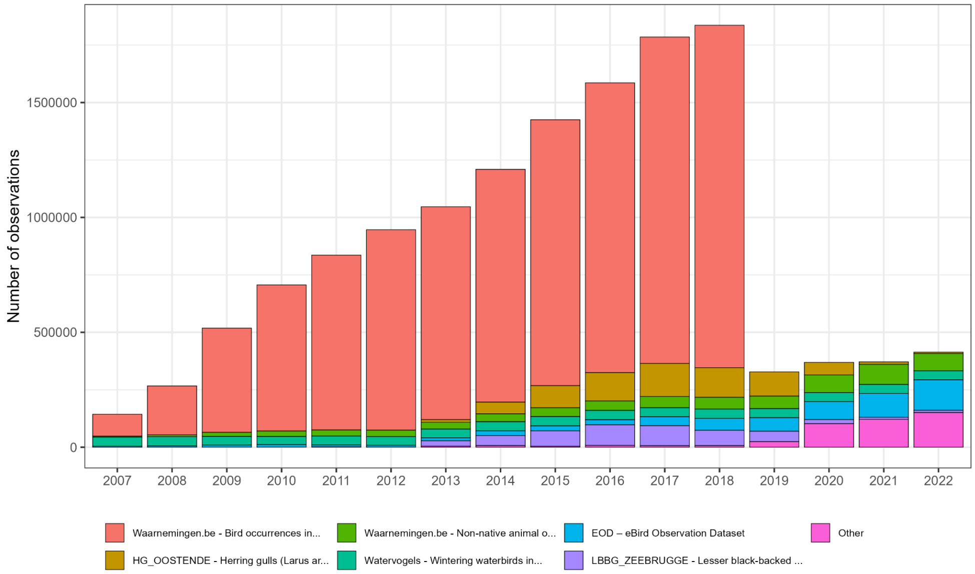


Figure 22: Number of observations and per year split per component dataset of the occurrence cube of the Flanders case study. The six largest contain 96.8 % of the cube's observations.

We calculated error measures for the indicator based on leave-one-dataset-out cross-validation (Fig. 23B-C). The RMSE is higher for more common species while the MRE is higher for rarer species (see also Fig. 24). This pattern is a direct consequence of the error measures used rather than an inherent property of the indicator. RMSE is an absolute error metric and therefore tends to be larger for common species, for which indicator values and absolute deviations are typically higher. In contrast, MRE is a relative (proportional) error measure, so for rare species even small absolute deviations can correspond to large relative errors, resulting in higher MRE values. The opposite trends observed for RMSE and MRE thus reflect the mathematical properties of absolute versus relative error measures and should not be interpreted as evidence of systematic differences in indicator performance between common and rare species.

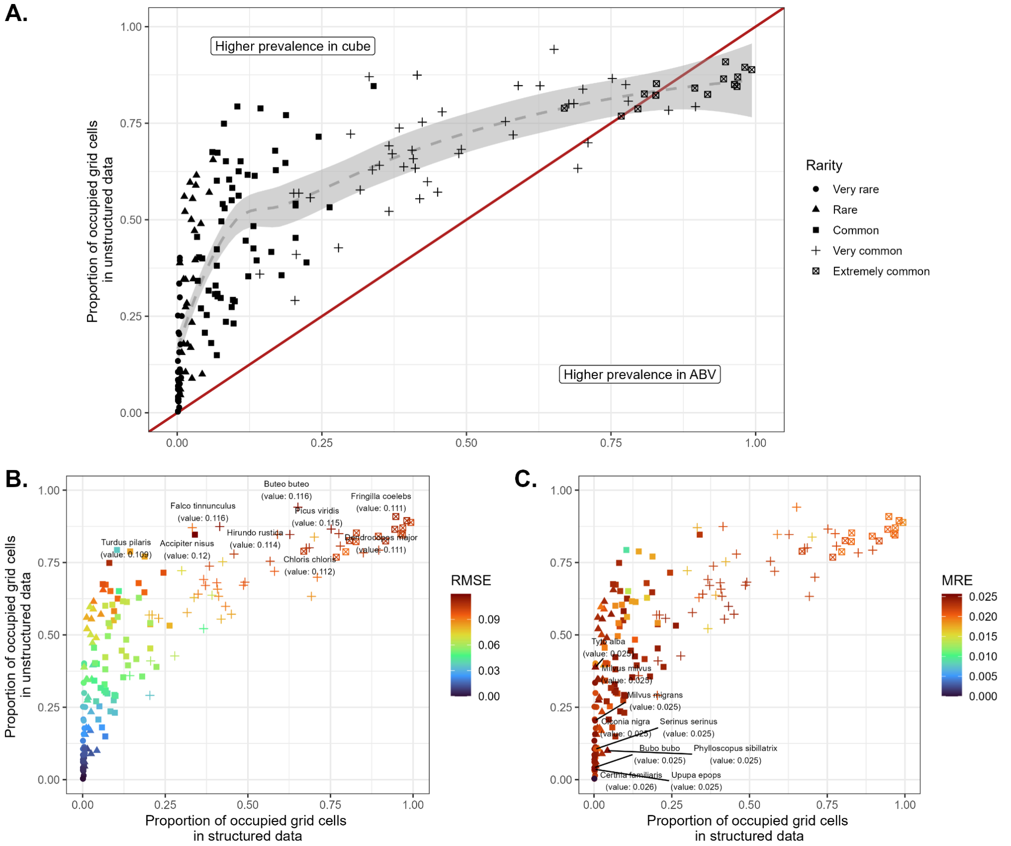


Figure 23: A. Comparison of prevalence between the structured (ABV) data and unstructured (GBIF) data per species. The red line shows a 1 on 1 relationship. The trendline is a LOESS smoother. B-C. Indicating RMSE (B) and MRE (C) on prevalence values. Labels are included for species with error values above the 95 % quantile (9 species).

Linear regressions were used to assess the relationship between dataset evenness and prediction error across species rarity categories (Fig. 24C-D). For RMSE, the relationship with dataset evenness is consistently negative across all rarity classes, indicating lower absolute errors if datasets are more equal in the number of observations (Table 10). However, this effect is only statistically significant at the 0.05 significance level for Rare, Very common, and Extremely common species. The strongest decline in RMSE with increasing evenness is observed for Extremely common species, followed by Rare species. For Very rare and Common species, confidence intervals overlap zero, suggesting no clear association between RMSE and dataset evenness in these groups.

For MRE, slopes are also negative across all rarity categories. This relationship is statistically significant for Very rare, Rare, Common, and Very common species, with the strongest effect observed for Common species. In contrast, Extremely common species show no significant relationship between MRE and dataset evenness, reflected by a wide confidence interval overlapping zero.

Overall, increasing dataset evenness is associated with reduced relative error for most rarity classes, while reductions in absolute error are more heterogeneous and depend on species rarity.

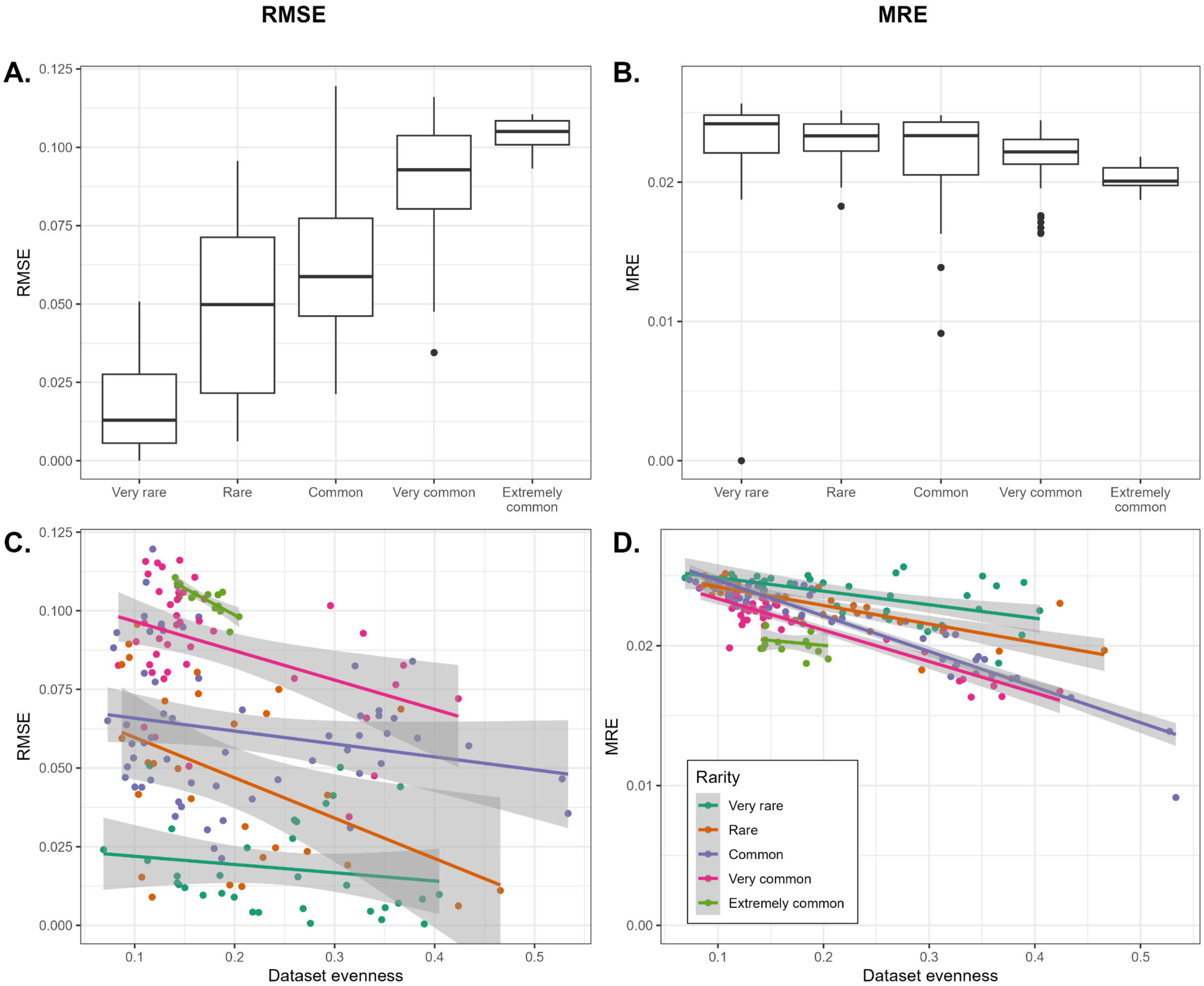


Figure 24: A-B. Boxplots of error measures per rarity category for RMSE (A) and MRE (B). C-D. Relationship between RMSE (C) and MRE (D) with dataset evenness across species rarity categories (linear regression).

Table 10: Results of linear regression between error measures and dataset evenness. Estimates of slopes are given with 95 % confidence limits (ll: lower limit, ul: upper limit).

Error	Rarity	Slope (ll, ul)	p-value
RMSE	Very rare	-0.026 (-0.084, 0.031)	0.35989
	Rare	-0.128 (-0.225, -0.032)	0.01127*
	Common	-0.041 (-0.090, 0.008)	0.10151
	Very common	-0.093 (-0.155, -0.031)	0.00425*
	Extremely common	-0.170 (-0.246, -0.094)	0.00033*
MRE	Very rare	-0.010 (-0.015, -0.004)	0.00123*
	Rare	-0.013 (-0.017, -0.009)	0.00000*
	Common	-0.025 (-0.028, -0.023)	0.00000*
	Very common	-0.022 (-0.026, -0.019)	0.00000*
	Extremely common	-0.007 (-0.031, 0.017)	0.53743

3.4.3.3. Improvement score relative to ABV reference values

Using leave-one-dataset-out cross-validation, we assessed the sensitivity of prevalence estimates derived from the bird data cube to the composition of the underlying datasets, using ABV prevalence as a reference benchmark. At the species level, median improvement scores are centred close to zero, indicating that for most species the omission of individual datasets had limited influence on prevalence estimates (Fig. 25). A smaller number of species exhibited consistently positive median improvements and only a few negative, showing higher sensitivity to data composition. Overall, cross-validation tends to move prevalence estimates closer to the true value.

Sensitivity patterns differed systematically across rarity classes (Fig. 26). Rare species showed more positive improvements than common species, which might suggest a stronger dependence on individual datasets. However, this pattern mainly reflects a structural constraint of the indicator rather than differences in data quality. When a dataset is removed, species prevalence cannot increase beyond its original value. For already common species with high prevalence, this leaves little room for improvement, and changes are therefore small or negative. In contrast, rare species have low baseline prevalence and are more sensitive to the influence of individual datasets, so removing a dataset can substantially alter their estimated prevalence and lead to larger apparent improvements. As a result, the observed differences among rarity classes arise primarily from the mathematical properties of the indicator and the underlying prevalence distributions in structured versus opportunistic data, rather than from intrinsic differences in dataset reliability.

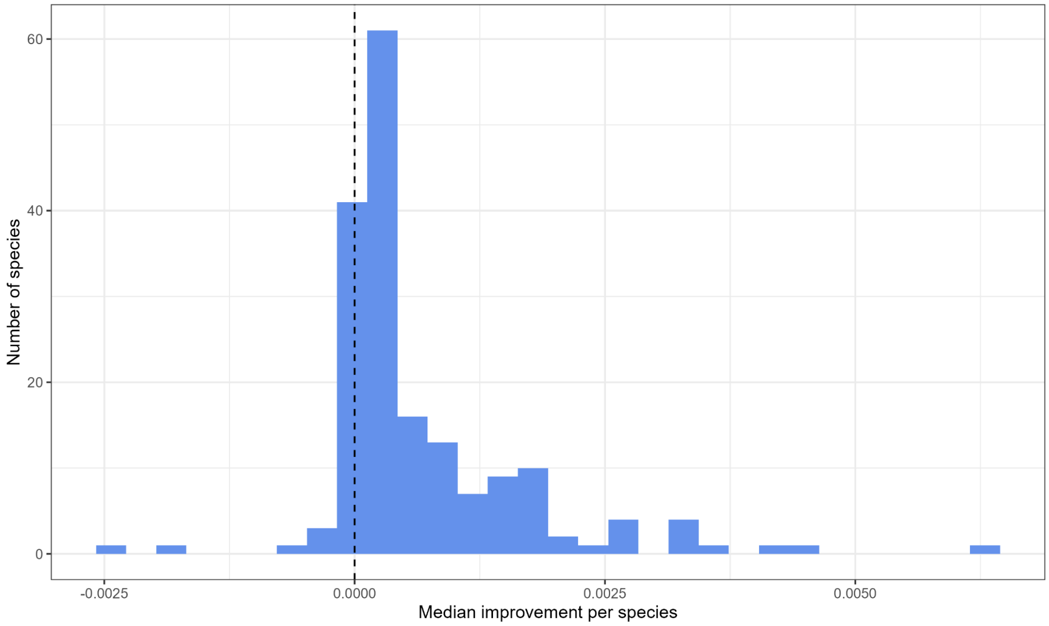


Figure 25: Distribution of median improvements per species.

Analysis at the dataset level showed that larger datasets exert the strongest influence on prevalence estimates (Fig. 26). The largest dataset, *Waarnemingen.be*, had a substantial impact, improving estimates for rare species while worsening them for common species.

3.4.4. Conclusion

This chapter examined the sensitivity of indicator results derived from occurrence cubes to the composition of their underlying component datasets, using species prevalence as a case study. The analysis shows that occurrence cubes are inherently shaped by strong heterogeneity among contributing datasets, with a small number of large datasets typically dominating the total number of records. As a result, indicator values should be interpreted as outcomes of a weighted aggregation of heterogeneous data sources rather than as direct summaries of uniform data quality.

Leave-one-dataset-out cross-validation proved to be a useful framework for quantifying this sensitivity at both the species and dataset level. Overall, prevalence estimates were relatively robust to the exclusion of individual datasets, with median changes close to zero for most species. However, sensitivity patterns differed systematically across species rarity classes and were closely linked to the mathematical properties of the prevalence indicator. Absolute and relative error measures exhibited contrasting behaviour for common versus rare species, and apparent improvements following dataset removal were largely constrained by baseline prevalence values. These patterns reflect structural characteristics of the indicator and underlying prevalence distributions rather than intrinsic differences in dataset reliability.

The relationship between dataset evenness and indicator uncertainty further highlights the importance of data structure. Species with unevenly distributed occurrences across datasets tended to show higher sensitivity to dataset omission, particularly in relative terms. Finally, it is important to note that all results presented here are based solely on species prevalence. Other indicators derived from occurrence cubes may exhibit different sensitivity patterns, and the conclusions of this chapter should therefore be understood as illustrative rather than universally applicable.

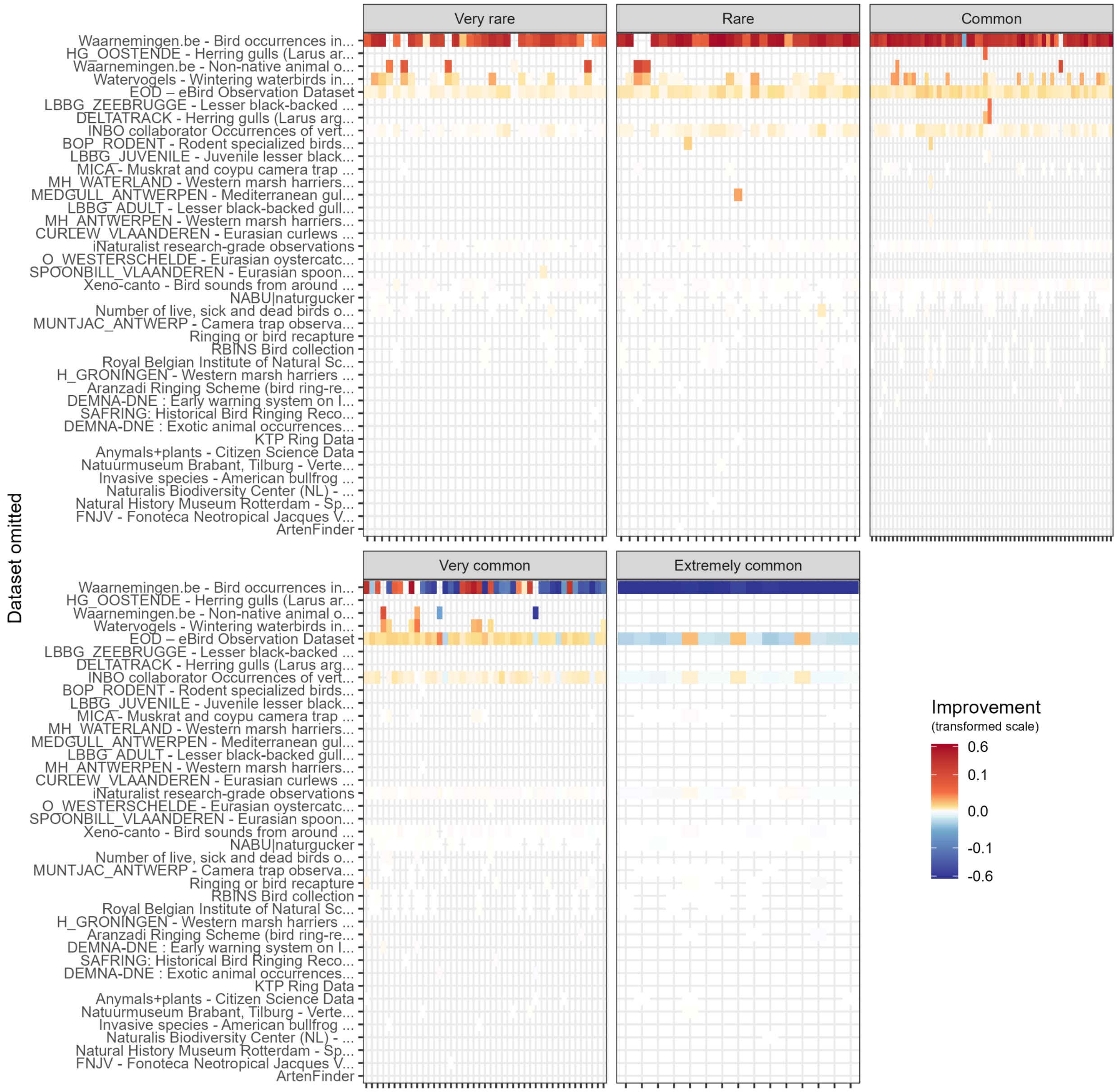


Figure 26: Improvement values of the indicator per species and per omitted dataset grouped by rarity class. Datasets are ordered from highest to lowest number of occurrences, as in Fig. 21. Colours indicate the magnitude and direction of improvement, with positive values reflecting improved indicator performance after dataset removal and negative values indicating deterioration. The colour scale is shown on a transformed (pseudo-log) scale to enhance contrast around zero.

4. Sampling bias and survey completeness

As established in the previous chapter, technical issues like publication delays and the dominance of specific datasets can create misleading patterns in biodiversity cubes. To address these challenges, this chapter introduces a survey-effort score and completeness metrics designed to identify well-sampled areas and provide a more reliable basis for indicator calculations.

Unstructured plant occurrence data are affected by strong spatial variation in observer effort, accessibility, seasonality, and taxon-specific detectability. To evaluate where the dataset (see 4.1) provides a sufficiently reliable basis for spatial indicators, we applied two complementary approaches at 1 × 1 km resolution. First, we quantified survey effort using a set of interpretable components that represent key dimensions of the recording process (record volume, temporal replication, seasonal coverage, and multi-year persistence) and combined these into an aggregated survey-effort score. Second, we estimated incidence-based sample coverage completeness following Chao et al. (2020), which provides a probabilistic measure of how completely the local flora has been detected and quantifies uncertainty through bootstrap confidence intervals. Together, these approaches provide both an effort-based diagnostic (how much and how consistently a cell has been sampled) and an outcome-based diagnostic (how complete the observed species set is likely to be), allowing robust identification of well-sampled grid cells and targeted prioritisation of under-sampled areas for future surveying.

4.1. Survey-effort, components and composite score

To characterise spatial variation in sampling effort and detectability in opportunistic flowering plant occurrence data in Flanders and Brussels over the past 50 years (1975 to 2025), we derived a set of complementary survey-effort components for each 1 × 1 km grid cell (GBIF.org (14 September 2025) GBIF Occurrence Download https://doi.org/10.15468/dl.m3nemg). The objective was to capture not only the volume of records but also the temporal replication, taxonomic breadth, and temporal continuity of sampling, which are all known to influence the probability of detecting species and the reliability of derived indicators. Because any single proxy (e.g. number of records) is susceptible to bias (e.g. repeated observations of common species, observer preference, or a few highly productive visits), the components were designed to represent multiple dimensions of effort and then combined into a composite index.

4.1.1. D — Number of observation days (within-cell temporal replication)

D represents the number of distinct sampling dates in a cell. It is used as an approximation of the number of independent visits or sampling occasions. Temporal replication is a key determinant of detectability and completeness: grid cells sampled on many different days are more likely to include seasonal species and detect species with intermittent visibility. Compared to raw record counts, **D** is less sensitive to multiple records from the same day and provides a more conservative proxy for sampling intensity.

4.1.2. N — Total number of occurrence records (sampling volume)

N is the total number of occurrence records within a cell. It captures the overall volume of information available and is useful as a broad indicator of effort. However, **N** can be inflated by repeated reporting of the same common taxa, targeted surveys of specific groups, or large datasets concentrated on particular sites. For this reason, **N** was retained as a component but not used in isolation.

4.1.3. Dn — Normalised temporal replication score

Dn is a rescaled version of **D** (bounded in [0, 1]) used to allow direct comparison among grid cells and facilitate aggregation with other components. Normalisation reduces the influence of extreme values of very heavily sampled cells and ensures that the composite index reflects relative differences in effort rather than being dominated by a small number of intensively sampled locations.

4.1.4. Rn — Normalised record volume score

Rn is a rescaled version of **N** (bounded in [0, 1]). As with **Dn**, normalisation provides comparability and prevents record-rich "hotspots" from disproportionately driving the composite index. Including both **Dn** and **Rn** allows the index to distinguish between cells with many records generated by few sampling occasions and cells with both high record volume and high temporal replication.

4.1.5. S — Seasonal coverage (phenological coverage)

To capture not only the volume and replication of sampling but also its phenological coverage, we included a seasonal score (**S**) based on the distribution of observation months for each grid cell. This component accounts for the fact that plant detectability and identification vary strongly through the year, and that opportunistic recording is often concentrated in a limited part of the growing season. Cells with high seasonal coverage are therefore more likely to detect a broader subset of the flora, including taxa with short phenological windows, whereas cells with low seasonal coverage may remain incomplete even when record counts are high.

4.1.6. Y — Number of years with records (temporal continuity)

Y quantifies the temporal continuity of recording in each grid cell by counting the number of years within a defined analysis window (2000–2025) in which the cell can be considered meaningfully surveyed. Rather than counting any year with a single record, a year is only counted when the grid cell has been recorded on at least distinct observation days. This threshold reduces sensitivity to incidental or opportunistic single-day events and provides a more robust indicator of sustained survey activity.

Formally, for each grid cell and year, the number of distinct observation dates is calculated, and a binary "surveyed year" indicator is assigned (1 if days , else 0). Y is then the sum of these indicators across all years in the analysis window, yielding the number of years in which the cell meets a minimum level of within-year replication. Cells with higher Y have been sampled repeatedly across years and are therefore less likely to reflect one-off campaigns or transient observer activity. This improves confidence that observed composition is representative and supports more reliable comparisons and indicator calculations.

4.1.7. M — Normalised measure of sampling consistency or diversity across sampling units and time

M is a temporal persistence multiplier that captures how consistently a grid cell has been surveyed across the full analysis window. It is derived from the proportion of years in which the cell meets the minimum survey threshold:

$$P = \min\left(\frac{Y}{k}, 1\right)$$

where:

- Y is the number of “surveyed years” (years with $\geq y_0$ observation days),
- k is the number of years in the analysis window (e.g. 2000–2025 $\rightarrow k = 26$).

Thus, P takes values between 0 and 1:

- $P = 0$: the cell is rarely surveyed across years (episodic)
- $P = 1$: the cell is surveyed every year (persistent)

Rather than using **P** directly as an effort component, it is transformed into a multiplier (**M**) ranging from 0.5 to 1.0:

$$M = 0.5 + 0.5P$$

This transformation ensures that temporal persistence acts as a moderating factor on overall effort, rather than as a stand-alone measure. Grid cells with high persistence ($P \approx 1$) receive **M** values close to 1.0 and are therefore not downweighted, while cells with very low persistence ($P \approx 0$) receive **M** values close to 0.5, resulting in moderate downweighting.

This design reflects the assumption that cells surveyed consistently across years provide more reliable information for comparative analyses and indicator estimation, particularly where interannual variation, phenological shifts, or long-term trends are of interest. At the same time, cells with irregular or short-lived sampling are not excluded entirely, as they may still contain valuable occurrence information. Instead, their contribution is reduced to account for the higher risk of temporal bias and incomplete detectability associated with episodic sampling.

Accordingly, **M** can be interpreted as an index of effort stability through time, with higher values indicating sustained monitoring and lower values indicating sporadic survey activity.

4.1.8. Core — Core survey-effort index

Core is a composite index summarising the “core” effort characteristics of a grid cell. It aggregates key normalised components that represent complementary aspects of effort (temporal replication, record volume, and taxonomic coverage and/or continuity), producing a single value in [0, 1]. Core provides a compact representation of multi-dimensional survey effort and is useful for ranking grid cells or identifying areas requiring increased sampling.

4.1.9. Score — Final survey-effort score

Score is the final aggregated survey-effort score, expressed on a convenient scale (e.g. 0–100) (Figs. 27, 28). This score is designed for communication and mapping, and it can be used as a quality layer to:

1. mask cells with insufficient effort,
2. weight analyses by survey effort,
3. prioritise areas for targeted recording or survey campaigns, and
4. support transparency in biodiversity indicator reporting by explicitly accounting for sampling intensity.

A Jupyter notebook used to calculate survey effort is archived in a public GitHub repository: https://github.com/AgentschapPlantentuinMeise/surveyeffort.

4.1.10. Why use multiple components?

Opportunistic biodiversity data are affected by sampling bias at multiple levels: spatial accessibility, observer effort, taxon popularity, and seasonality. A multi-component approach improves robustness by ensuring that no single bias dominates the assessment. For example, high **N** but low **D** may reflect many records from a single day or event, whereas high **D** and high **Y** indicates broad temporal sampling likely to improve detectability. Combining these components supports more defensible decisions about which grid cells are sufficiently sampled for indicator calculations, and pinpoints where additional data collection is most valuable. The panels of Figure 28 show the empirical distributions of the individual components used to quantify survey effort and data availability per grid cell. Distributions highlight strong right-skew in raw effort measures and increasing regularisation after normalisation and aggregation.

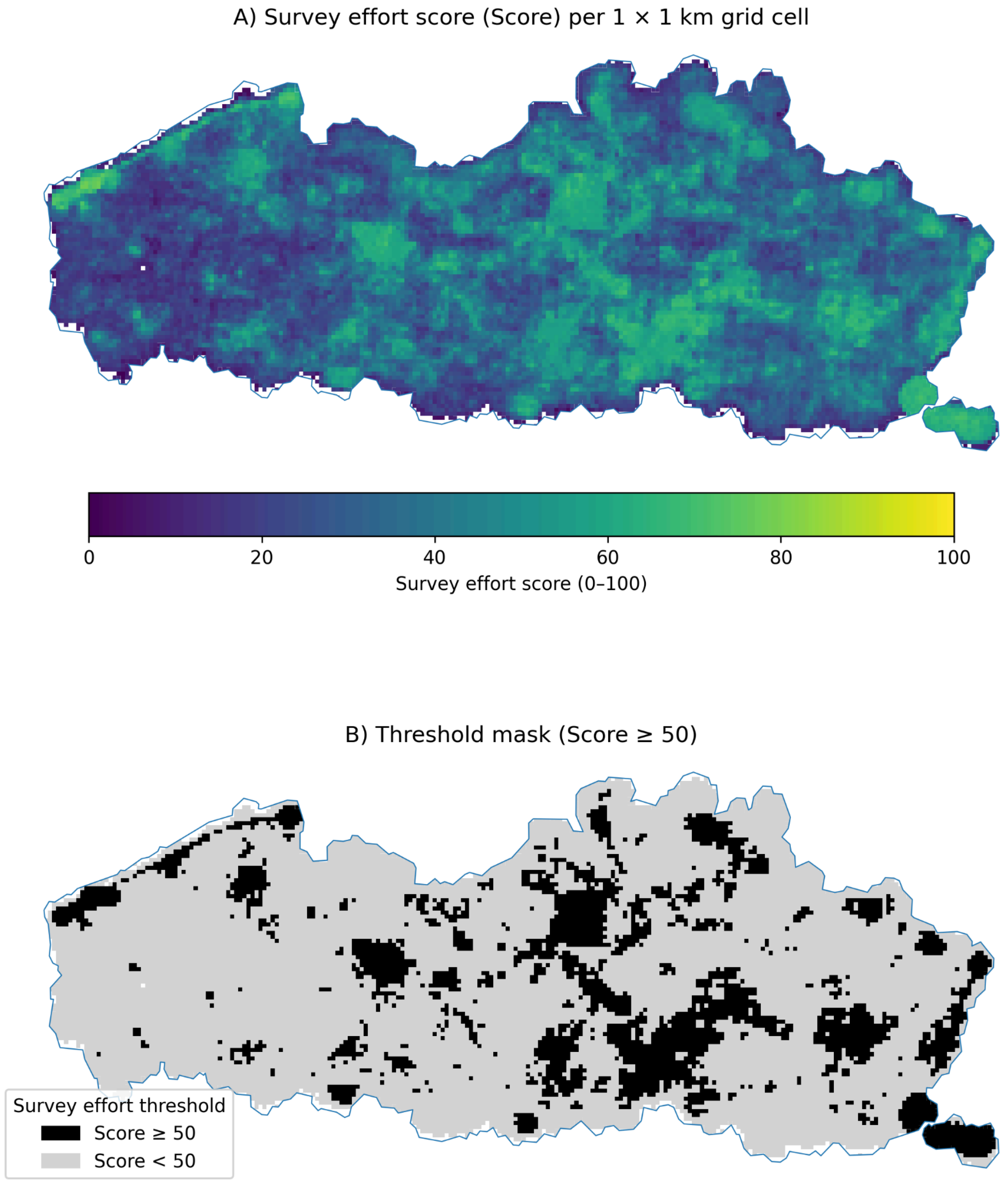

A) Survey effort score (Score) per 1 × 1 km grid cell
0
20
40
60
80
100
Survey effort score (0–100)
B) Threshold mask (Score ≥ 50)
Survey effort threshold
Score ≥ 50
Score < 50

Figure 27: Survey-effort score and threshold mask for 1 × 1 km grid cells in Flanders. A. Survey-effort score (0–100) derived from multiple components capturing record volume, temporal replication, seasonal coverage, and multi-year persistence. B. Binary threshold mask identifying grid cells with Score ≥ 50 (black), used to highlight areas considered sufficiently surveyed for downstream analyses; cells below the threshold are shown in grey. The grid is cropped to the Flanders boundary.

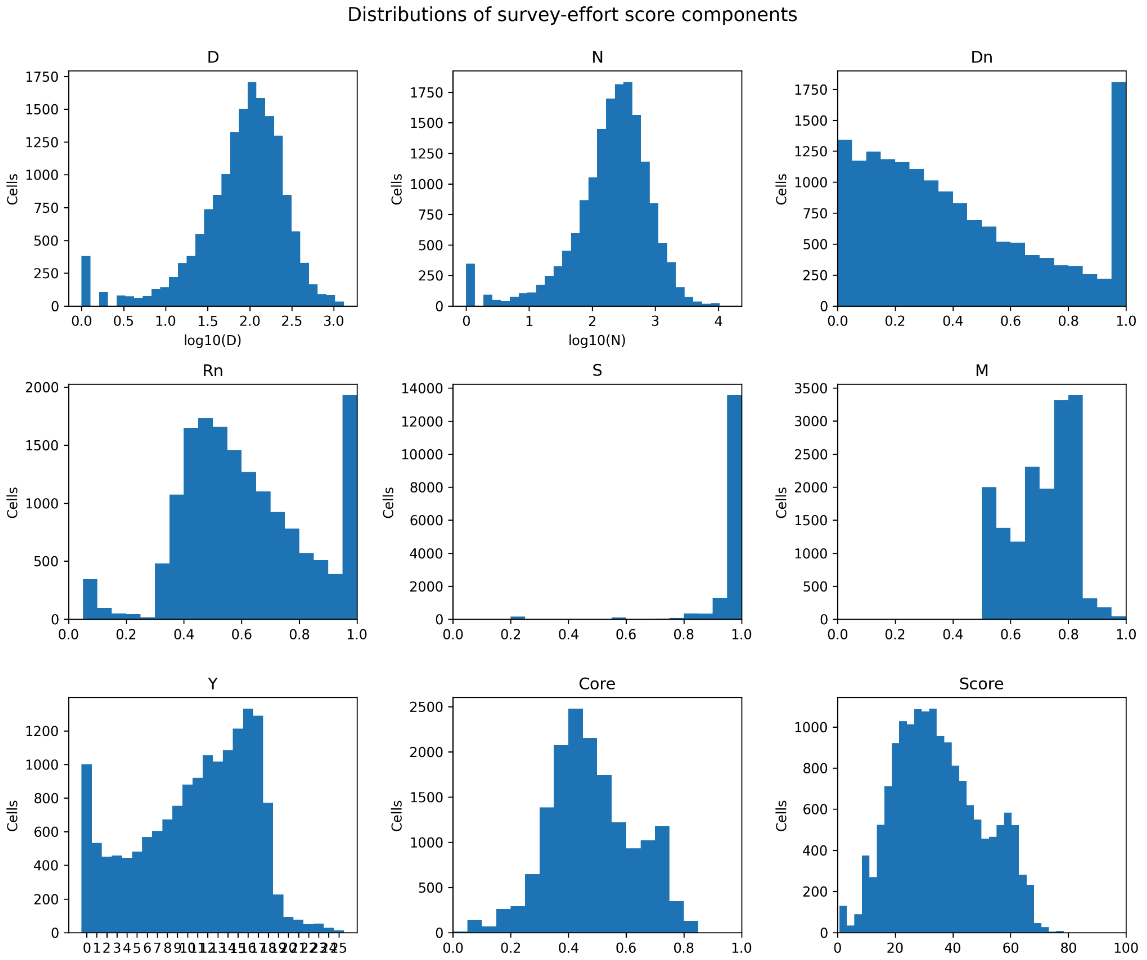


Figure 28: Distributions of survey-effort components across 1 × 1 km grid cells. D represents the total number of observation days, and N the total number of occurrence records. Dn and Rn are normalised measures of temporal and recording intensity, scaled to the interval [0, 1]. S represents taxonomic coverage, expressed as a normalised measure of species representation. M captures spatial or methodological consistency in recording effort, while Y denotes the number of distinct years with observations. Core is a weighted combination of the normalised components. Core = 0.4D + 0.2R + 0.2S. Score is a composite index of how well a grid cell has been surveyed, combining the number of survey days, recording intensity, seasonal timing and spacing of visits, and consistency of coverage across years. Score= 100 × Core × M.

4.2. Probabilistic estimation of sample coverage

Unstructured biodiversity occurrence data are inherently affected by spatially and temporally heterogeneous sampling effort, observer preferences, and variation in species detectability. As a result, spatial patterns derived from aggregated occurrence cubes may reflect differences in recording intensity rather than true ecological patterns (Boakes et al., 2010; Isaac et al., 2014). Assessing and accounting for these sources of bias is therefore a prerequisite for reliable indicator use.

4.2.1. Sample coverage as a measure of survey completeness

To quantify the degree to which plant occurrence data are affected by uneven sampling effort across Flanders, we estimated sample coverage, *sensu* Chao et al. (2020), for each 1 × 1 km grid cell. Sample coverage represents the proportion of the total incidence probability attributable to species already observed in the sample and can be interpreted as the probability that a newly added record belongs to a species that has already been detected. Unlike richness-based completeness metrics, coverage is comparatively robust to the presence of undetected rare species and stabilises at lower sampling intensities.

Coverage was estimated using the incidence-based estimator $\hat{C}_1$, which depends on the number of species recorded in exactly one sampling unit (uniques, *Q1*), the total number of incidences (U), and a small-sample bias correction. Following Chao et al. (2020), this metric was preferred over richness-based completeness ($\hat{C}_0$) for fine-grained spatial analyses, as the latter is highly sensitive to rare species and unstable when replication is low.

4.2.2. Defining sampling units and incidence data

For the Flemish plant data, occurrences were aggregated into 1 × 1 km EEA reference grid cells. Within each grid cell, sampling units were defined as unique observation dates. Presence–absence incidence data were constructed by deduplicating records within each cell–species–date combination. This approach approximates repeated surveys within a grid cell and allows incidence-based estimators to be applied to opportunistic data in a consistent manner.

4.2.3. Quantifying uncertainty using bootstrap resampling

While point estimates of sample coverage provide a useful first indication of survey completeness, they do not convey the uncertainty associated with uneven and often sparse sampling. To address this, we implemented a non-parametric bootstrap procedure, following the recommendations of Chao et al. (2020), to derive confidence intervals for coverage estimates.

For each grid cell, sampling units (i.e. observation dates) were resampled with replacement, preserving the number of sampling units per cell. For each bootstrap replicate, incidence frequencies were recalculated and sample coverage was re-estimated. Repeating this procedure multiple times yielded an empirical distribution of $\hat{C}_1$, from which 95 % confidence intervals were derived. This approach captures uncertainty arising from limited replication and heterogeneous detection across sampling units, without imposing parametric assumptions about species detectability.

4.2.4. Spatial representation of completeness and uncertainty

Coverage estimates and their associated confidence intervals were mapped by generating grid-cell polygons directly from the EEA grid codes, ensuring exact spatial correspondence between occurrence data and geometry.

Three complementary spatial diagnostics were produced:

1. Point estimates of sample coverage ($\hat{C}_1$), indicating the overall level of survey completeness per grid cell.
2. Lower confidence bounds of coverage, representing a conservative estimate of completeness under uncertainty.
3. Confidence-interval width, highlighting areas where coverage estimates are particularly uncertain due to limited or uneven sampling.

Together, these maps provide a transparent spatial assessment of sampling bias and detectability constraints across Flanders. Areas with low coverage or wide confidence intervals can be identified as insufficiently sampled and may be excluded from downstream indicator calculations or prioritised for targeted data collection (Fig. 29).

All scripts used to generate completeness estimates and uncertainty maps were archived in a public GitHub repository: https://github.com/AgentschapPlantentuinMeise/completeness. This repository enables full reproducibility of the analysis and includes GIS-ready outputs and figure-generation code.

4.2.5. Implications for indicator reliability

By explicitly quantifying both survey completeness and its uncertainty, this approach allows spatial variation in data quality to be incorporated into indicator interpretation. Rather than assuming uniform reliability across space, coverage-based diagnostics provide an evidence-based basis for masking poorly sampled grid cells, comparing regions at equivalent levels of completeness, and communicating uncertainty to end users. This is particularly important for fine-resolution analyses based on unstructured occurrence data, where sampling bias and detectability are major limiting factors.

4.3. Conclusion

The survey-effort score and the incidence-based completeness metric (Chao et al., 2020) yielded highly similar spatial patterns, indicating that both capture the dominant gradients in sampling intensity and detectability across Flanders. The Chao-based approach is attractive for routine application because it is comparatively simple, theoretically grounded, and provides an interpretable measure of completeness with confidence intervals derived via bootstrap resampling. The component-based index, however, offers greater methodological flexibility: individual components can be tailored to reflect regional knowledge of recording behaviour (e.g. strong seasonal bias in botanical recording, uneven multi-year persistence, or minimum thresholds for defining a surveyed year). This flexibility allows the score to be aligned with expert understanding of how sampling occurs, while still providing a transparent and mappable data-quality layer.

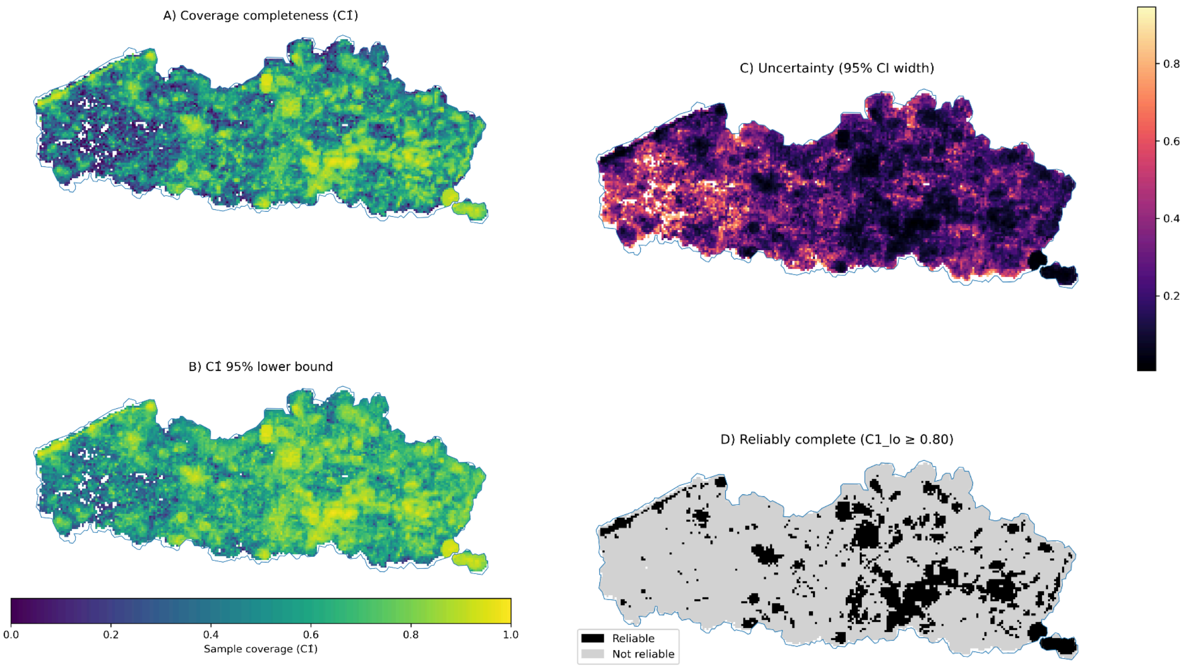

A) Coverage completeness (C1̂)
C) Uncertainty (95% CI width)
0.8
0.6
0.4
0.2
B) C1̂ 95% lower bound
D) Reliably complete (C1_lo ≥ 0.80)
0.0
0.2
0.4
0.6
0.8
1.0
Sample coverage (C1̂)
Reliable
Not reliable

Figure 29: Spatial survey completeness and uncertainty of plant occurrence records in Flanders (1 × 1 km). A. Incidence-based sample coverage completeness ($\hat{C}_1$) estimated for each 1 × 1 km EEA reference grid cell, representing the probability that an additional record from the cell would belong to a species already detected. B. Lower 95 % confidence bound of $\hat{C}_1$, derived from non-parametric bootstrap resampling of sampling units (observation dates) within each cell ($\mathrm{n} = 100$ bootstrap replicates), providing a conservative estimate of minimum plausible completeness. C. Width of the 95% confidence interval ($\hat{C}_{1high} - \hat{C}_{1low}$), indicating spatial variation in uncertainty driven by heterogeneous sampling effort and detectability. D. Binary reliability mask identifying grid cells considered reliably complete (black) according to the criterion $\hat{C}_{1low} \geq 0.80$; cells not meeting this criterion are shown in grey. Grid cells with fewer than 10 sampling units were excluded from analysis.

5. Species detectability and recording bias

Where the previous chapter establishes whether a location has been surveyed enough to be 'reliable,' this chapter addresses whether the species themselves are likely to be detected even when effort is present.

5.1. Estimating detectability from opportunistic occurrence data

To quantify variation in species detectability in opportunistic occurrence data, we implemented a simple, survey-based detection probability metric derived directly from repeated sampling events within EEA reference grid cells (Van Strien et al., 2013). The aim was to estimate, for each species, how likely it is to be observed on a day when a grid cell is surveyed, conditional on the species being truly present in that cell. This provides a pragmatic detectability proxy that can be used to interpret spatial patterns in occurrence indicators, identify taxa that are typically under-recorded, and support aggregation of detectability statistics at higher taxonomic levels (e.g. family and order).

5.1.1. Survey definition and effort quantification

Occurrences were aggregated into EEA reference grid cells (here applied to a 5 × 5 km grid) and sampling effort within each cell was represented by the number of unique observation dates ("survey days"). A survey day was defined as any calendar date on which at least one occurrence was recorded in that cell. This allows opportunistic occurrence records to be treated as a repeated-visit dataset, where each unique date is interpreted as an independent sampling occasion (or an approximation thereof). Duplicate records were removed at the level of species × cell × day to avoid inflated detection counts due to multiple reports of the same species during the same visit. Source data can be found at: GBIF.org (31 December 2025) GBIF Occurrence Download https://doi.org/10.15468/dl.zz5c5u.

Formally, for each grid cell c, total survey effort was calculated as:

$$\text{survey\_days} = \#\left\{d : \exists \text{ any occurrence in cell } c \text{ on date } d\right\}$$

5.1.2. Confirmed presence filtering

Because opportunistic datasets can include sporadic or erroneous records, detection probabilities were only calculated for species-cell combinations where the species is likely to be truly established and repeatedly recorded over time. A "confirmed presence" rule was used to select occupied cells for each species. A species was considered present in a grid cell if it had:

1. records in at least three distinct years, and
2. a minimum span of at least ten years between the earliest and latest record in that cell.

Let $y_{s,c}$ be the set of years in which species s was recorded in cell c. Then s is treated as present in c if:

$$|y_{s,c}| \geq 3, \text{and } (max(y_{s,c}) - min(y_{s,c})) \geq 10$$

This criterion excludes one-off observations and short-term occurrences and focuses detection estimates on persistent populations.

5.1.3. Detection probability and shrinkage estimation

For each confirmed species–cell combination, detectability was estimated as the proportion of survey days in that cell on which the species was recorded:

$$p_{s,c} = \frac{\mathrm{detection\_days}_{s,c}}{\mathrm{survey\_days}_{c}}$$

where:

1. $\mathrm{detection\_days}_{s,c}$ is the number of unique dates on which species s was observed in cell c,
2. $\mathrm{survey\_days}_{c}$ is the total number of unique survey dates (any plant occurrence) in cell c

This metric can be interpreted as: the probability that species s will be recorded on a randomly selected survey day in cell c, given that the species is confirmed present in that cell.

However, detection probability estimates can be unstable when survey effort is low, i.e. when there are few survey days, especially for rare or cryptic taxa. To stabilise estimates for low-effort cells, an optional Bayesian shrinkage estimator was computed using a Beta prior:

$$p^{*}_{s,c} = \frac{\mathrm{detection\_days}_{s,c} + \alpha}{\mathrm{survey\_days}_{c} + \alpha + \beta}$$

with $\alpha = 1$ and $\beta = 9$, corresponding to a prior mean of 0.10 (i.e. a weak prior expectation that a species will be detected on ~10% of survey occasions). This shrinkage estimator pulls extremely high or low values toward the prior mean when effort is low, while having negligible effect when survey effort is high.

5.2. Taxonomic aggregation and outputs

5.2.1. Aggregation across taxonomic levels

To allow detectability to be summarised at broader taxonomic levels, each species was linked to its higher classification (family, order, class, phylum, kingdom) by querying the GBIF Species API using speciesKey identifiers. Results were cached locally to ensure reproducibility and efficiency. Detection probabilities were then summarised across all confirmed occupied cells and across species:

- species-level summaries: mean and median detectability across occupied cells,
- family-level summaries: average detectability across species within families,
- order-level summaries: average detectability across species within orders.

These summaries support interpretation of detectability bias in taxonomic groups and can be used as inputs for downstream analyses, including taxonomic tree visualisations of detectability. Figure 30 is a taxonomic tree showing variation in estimated detectability across higher taxa, based on opportunistic occurrence records. Node colour represents the mean detectability of descendant orders, calculated as the average probability that a species is recorded on a given survey day within grid cells where it is confirmed present. Detectability values were computed for species within grid cells meeting a confirmed-presence criterion (≥ 3 distinct recording years spanning ≥ 10 years) and then summarised across species to obtain mean order-level detectability estimates.

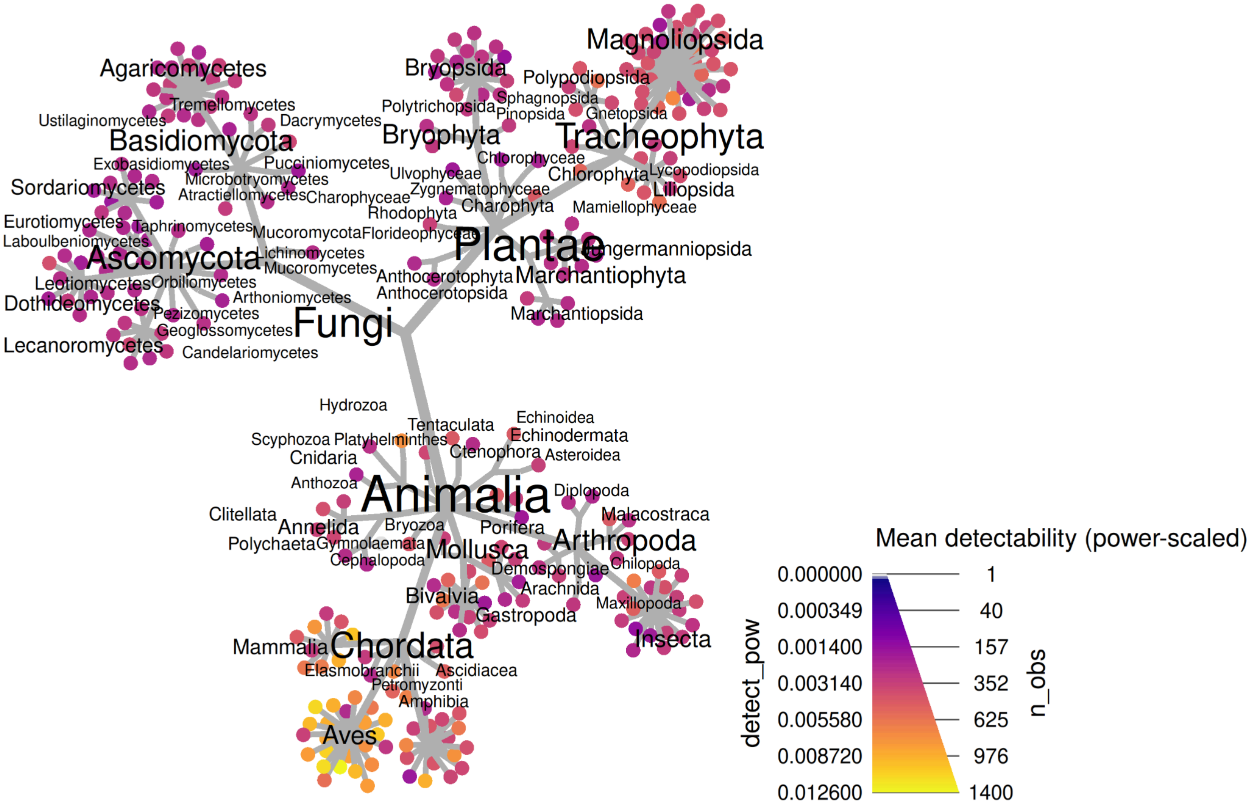


Figure 30: Taxonomic distribution of detectability across major kingdoms, based on opportunistic occurrence records. Colours are power-transformed to enhance visual contrast among detectability values. Only higher taxonomic ranks are labelled for clarity.

5.2.2. Data products and script outputs

The scripts used to generate detection probabilities were archived in a public GitHub repository: https://github.com/AgentschapPlantentuinMeise/detectability.

The script produces:

1. **species × cell detectability table** (`species_cell_detectability.tsv`) containing:
    - confirmed presence statistics (years present, span),
    - survey effort (`survey_days`),
    - detection counts (`detection_days`),
    - detection probability $p_{s,c}$ and shrinkage $p^*_{s,c}$,
    - higher taxonomy fields (kingdom → family).

2. **summary tables** for species, family, and order detectability.

3. **taxonomy and feature tables** compatible with **metacoder** workflows, with `orderKey` used as the OTU_ID (Foster et al., 2017).

5.3. Interpretation, bias, and limitations

This approach quantifies detectability as a function of recorded co-occurrence with any survey effort in a cell. It is intentionally simple and relies only on the internal structure of opportunistic data. However, several caveats apply:

- Survey days are inferred from observation dates rather than planned sampling effort and thus reflect both recorder activity and reporting behaviour.
- Detection probability is conditional on confirmed presence, but presence is still inferred from repeated records rather than independent validation.

Detectability estimated from opportunistic occurrence data reflects not only biological visibility or abundance of species, but also the human and institutional processes that determine whether an observation is made, identified, and reported. In particular, detection probabilities may be shaped by observer preferences, uneven taxonomic expertise, and reporting biases that vary across space, time, and taxonomic groups. For example, conspicuous, charismatic, or easily identified species may be disproportionately recorded even when they are not particularly abundant, while small, cryptic, ephemeral, or taxonomically challenging species can remain under-recorded despite being locally common. Similarly, species may be systematically overlooked if they require specialised identification (e.g. microscopic characters, critical groups), are only identifiable during brief phenological windows, or are likely to be misidentified and therefore avoided by non-specialists. As a result, the detectability metrics derived here should be interpreted as an empirical measure of recording detectability—the probability of being recorded given the local survey process—rather than a direct proxy for ecological detectability alone. Importantly, this aligns with the intended purpose of the metric: it captures taxon-specific and observer-driven biases that influence the reliability of opportunistic biodiversity indicators.

5.4. Temporal and technological drivers of detectability

5.4.1. Temporal trends in recording detectability

Detectability estimates derived from opportunistic occurrence data show pronounced temporal change over the period analysed (Fig. 31). This indicates that the probability of recording a species during a survey day is not constant through time but is strongly influenced by changing data-collection practices, observer behaviour, and technological developments.

A plausible explanation for the observed decline in detectability for many taxa is that recording effort has shifted towards more frequent but less complete recording events. As citizen-science platforms expanded and mobile applications made it easier to submit individual observations, observers may increasingly record on more days but submit fewer complete lists of all species encountered during each visit. Such behaviour reduces the probability that any particular species is recorded on a given survey day, especially for common or less “interesting” taxa that are easily overlooked or under-reported in opportunistic submissions. At the same time, increased participation and survey activity can substantially increase the number of survey days, causing the denominator (survey days) to rise faster than the numerator (species detection days), thereby lowering detectability estimates.

These shifts coincide with the expansion of major citizen-science reporting infrastructures. In Belgium, waarnemingen.be was launched by Natuurpunt on 14 May 2008, after which reporting volumes increased sharply. Globally, iNaturalist launched in 2008, but remained web-based until its first mobile application was released in 2011. These developments likely contributed to step changes in the volume and style of recording, which can be observed in detectability trends and should be considered when interpreting temporal patterns in occurrence-based indicators.

Overall, the detectability trends provide strong evidence that temporal variation in observation practices introduces a systematic bias into opportunistic biodiversity data. Consequently, changes in detectability through time should be interpreted cautiously and should be treated as a data-quality signal, reflecting evolving recording behaviour and sampling heterogeneity, rather than as a direct ecological signal of changing species abundance.

5.4.2. Drivers of temporal change: technology and citizen science

It is also important to recognise that biological recording has undergone several major technological shifts in recent years, which have altered both the volume and composition of records and the probability of detecting particular taxa. The widespread adoption of smartphone-based citizen science applications, increasingly supported by AI-assisted identification, has lowered barriers to participation and increased the rate at which individual observations can be submitted, potentially favouring conspicuous or photographable organisms while reducing the frequency of complete lists. In parallel, new monitoring technologies have expanded detectability for groups that were historically under-recorded in opportunistic datasets. Camera traps have greatly increased detection of many mammals and other elusive vertebrates, while bat detectors and automated acoustic recorders have transformed survey capacity for bats and vocalising taxa. Similarly, animal tracking devices provide presence data independent of observer encounters, improving detectability for mobile species with low encounter rates. More recently, molecular approaches such as eDNA sampling and metabarcoding have enabled detection of organisms that are cryptic, rare, nocturnal, or difficult to identify visually, and can dramatically shift the apparent detectability of entire groups relative to traditional observational recording. These innovations mean that temporal changes in detectability and taxonomic coverage may reflect evolving tools and recording workflows as much as ecological change, reinforcing the need to interpret detectability trends as a combined signal of biological visibility, observer behaviour, and technological capability.

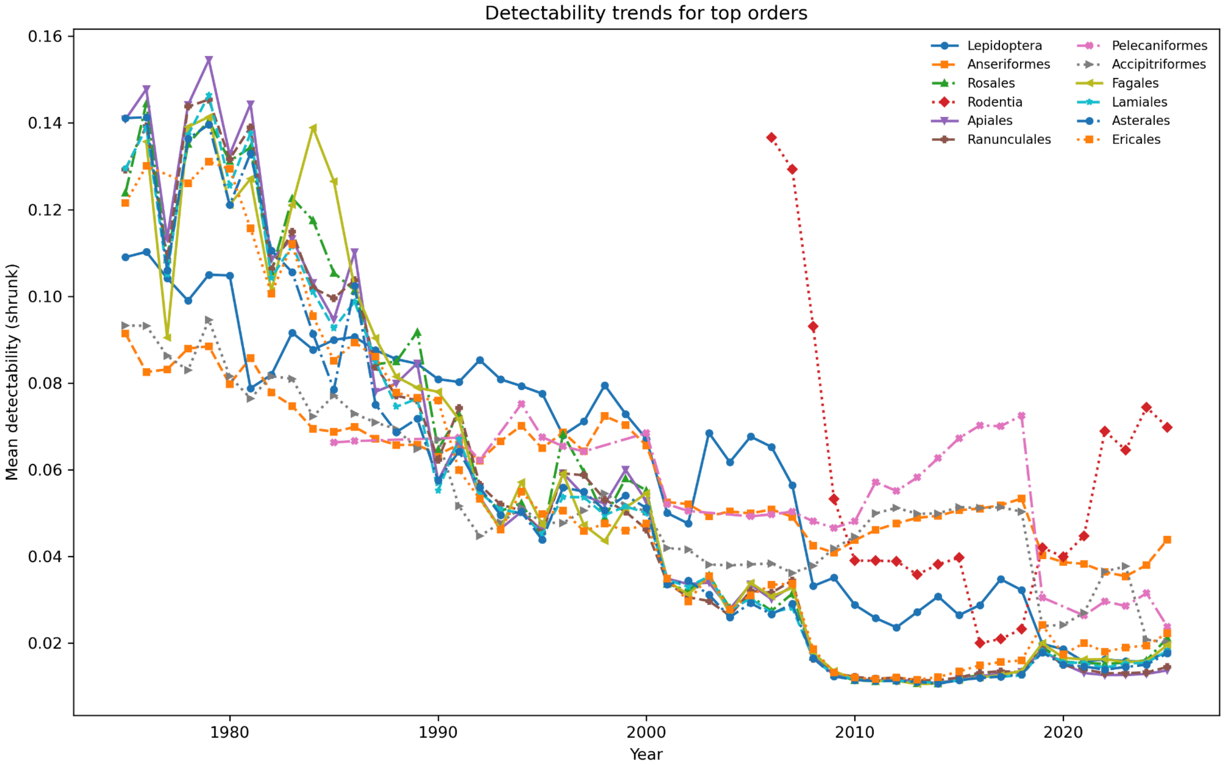


Figure 31: The annual detectability for the most detectable taxonomic orders, calculated from opportunistic occurrence records aggregated into EEA reference grid cells in Flanders. Within each cell and year, survey effort was quantified as the number of unique observation dates with any occurrence record. For each order, detection days were the number of survey days in which at least one species from the order was recorded. Annual detectability was estimated as the mean of species × cell detectability values, conditional on confirmed presence (species recorded in ≥3 distinct years spanning ≥10 years within the cell). Shrinkage estimates were applied to reduce instability in cells with low survey effort.

5.5. Conclusion

Based on the order-level detectability estimates, there are clear systematic differences among major taxonomic groups in how frequently taxa are recorded on days when a grid cell is surveyed. Animals show the highest overall detectability, with order means substantially exceeding those of plants and fungi (mean of order means ≈ 0.0039 for Animalia, compared with ≈ 0.0017 for Plantae and ≈ 0.0010 for Fungi). This pattern is driven largely by highly conspicuous and well-surveyed vertebrate groups, particularly birds and mammals: several avian orders (e.g. *Columbiformes*, *Suliformes*, *Podicipediformes*, *Piciformes*) rank among the most detectable, consistent with the fact that birds are visually conspicuous and actively targeted by observers. By contrast, fungal orders generally show the lowest detectability, reflecting strong constraints from seasonality (fruiting bodies), identification difficulty, and chronic under-reporting of fungi in opportunistic datasets. Plant detectability is intermediate, but still relatively low overall, with many orders clustered at low values, consistent with heterogeneous observer focus and the fact that many plants are overlooked unless flowering, distinctive, or of particular interest. Across all kingdoms, the lowest detectability values are often associated with orders represented by few species or few confirmed occupied cells, indicating that limited sample size also contributes to unstable or low detectability estimates.

Importantly, these detectability differences are not temporally stable. Detectability has changed substantially through time for many orders, indicating that recording detectability reflects evolving observation practices and data mobilisation, not only biological detectability. In particular, trends suggest a shift toward more frequent recording on more survey days but less complete recording events, consistent with the increasing use of opportunistic submission platforms and mobile applications that facilitate rapid reporting of selected observations rather than complete lists. Step changes in detectability coincide with major developments in citizen-science infrastructure (e.g. the launch of waarnemingen.be in 2008), and more broadly with technological advances such as AI-assisted identification, camera traps, acoustic detectors, tracking devices, and eDNA/metabarcoding. Together, these results indicate that detectability in opportunistic occurrence cubes is structured by taxon-specific visibility and phenology but is also strongly shaped by recorder behaviour and technological change. Consequently, temporal variation in detectability should be treated as an explicit data-quality signal and accounted for when interpreting spatial and temporal patterns in occurrence-based biodiversity indicators.

6. Software implementation

We were able to implement useful tools into two R packages: (1) a simulation tool that was initiated and accelerated during the B3 hackathon 'Hacking Biodiversity Data Cubes for Policy' (Abraham et al., 2025), and (2) a package that focuses on developing general measures of data cube quality and reliability developed within Task 5.4 of the B3 project.

6.1. Simulating occurrence cubes in R

6.1.1. The **gcube** package

Simulation-based approaches play a crucial role in biodiversity research because they allow complex ecological and sampling processes to be explored in a controlled and flexible way. By mimicking real-world systems under predefined assumptions, simulations make it possible to study mechanisms and sensitivities that are difficult or impossible to isolate using empirical data alone (Christie et al., 2019; Zurell et al., 2010). Biodiversity data are shaped by a combination of biological processes, such as species responses to environmental gradients and temporal change, and observation processes, including variation in sampling effort, detection probability, spatial uncertainty, and observer bias. Simulation frameworks enable these processes to be modelled separately or jointly, allowing their individual effects on observed patterns to be disentangled (Münkemüller et al., 2012; Sokol et al., 2017; Zurell et al., 2010).

The **gcube** package provides an R-based framework for simulating occurrence data cubes (Langeraert, 2026). It supports the generation of multi-species distributions across space and time, followed by the simulation of diverse observation and sampling processes that yield realistic occurrence records. These simulated occurrences can subsequently be aggregated to a spatial grid, taking positional uncertainty into account, to construct occurrence cubes. A general example is provided in the next paragraph.

In addition to fully simulated data, **gcube** also facilitates the creation of data cubes from virtual species generated with the **virtualspecies** package (Leroy et al., 2016), as well as from user-supplied datasets. Comprehensive tutorials and examples are available through the package website (https://b-cubed-eu.github.io/gcube/) and the broader B3 documentation platform (https://docs.b-cubed.eu/).

The name **gcube**, short for "generate cube", reflects the package's primary goal: enabling the construction of occurrence cubes from minimal and transparent inputs. The package was originally conceived during the 'Hacking Biodiversity Data Cubes for Policy' hackathon, where it received first prize in the 'Visualisation and Training category' (Langeraert et al., 2024).

6.1.2. Simulation example

This is a basic reproducible example which shows the workflow for simulating a biodiversity data cube. It is divided in three steps or processes:

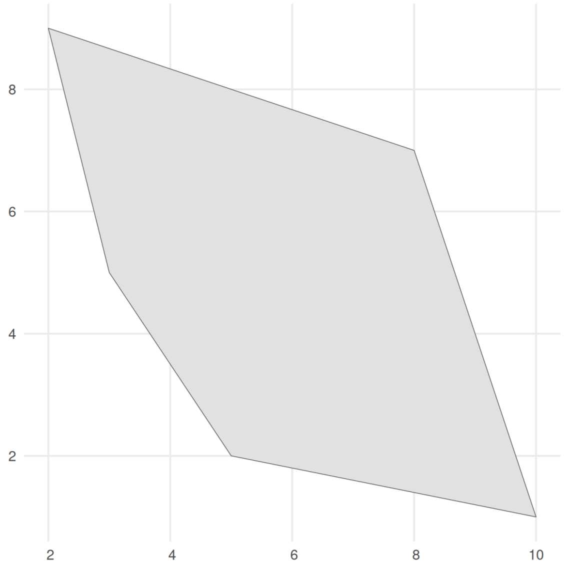
1. Occurrence process
2. Detection process
3. Grid designation process

The functions are set up such that a single polygon as input is enough to go through this workflow using default arguments. The user can change these arguments to allow for more flexibility.

```
# Load packages
library(gcube)

library(sf)       # working with spatial objects
library(dplyr)    # data wrangling
library(ggplot2) # visualisation with ggplot
```

We create a polygon as input. It represents the spatial extent of the species.

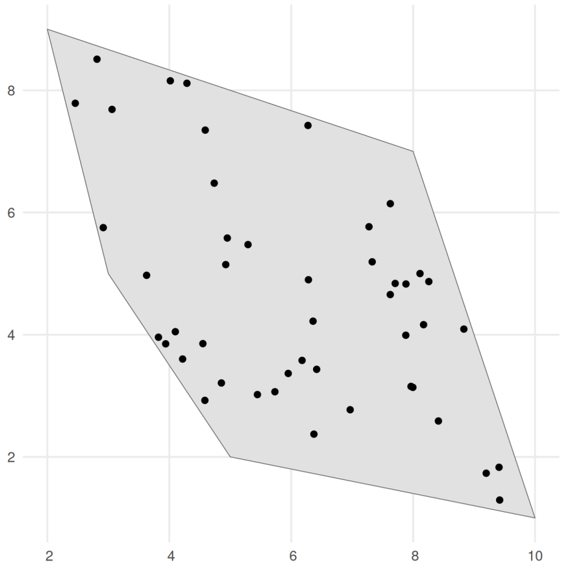
```
# Create a polygon to simulate occurrences within
polygon <- st_polygon(list(cbind(c(5, 10, 8, 2, 3, 5), c(2, 1, 7,9, 5,
2))))

# Visualise
ggplot() +
  geom_sf(data = polygon) +
  theme_minimal()
```



1. Occurrence process

We generate occurrence points within the polygon using the `simulate_occurrences()` function. In this function, the user can specify different levels of spatial clustering and define the trend of number of occurrences over time. The default is a random spatial pattern and a single time point with `rpois(1, 50)` occurrences.

```
# Simulate occurrences within polygon
occurrences_df <- simulate_occurrences(
  species_range = polygon,
  initial_average_occurrences = 50,
  spatial_pattern = c("random", "clustered"),
  n_time_points = 1,
  seed = 123)

# Visualise
ggplot() +
  geom_sf(data = polygon) +
  geom_sf(data = occurrences_df) +
  theme_minimal()
```

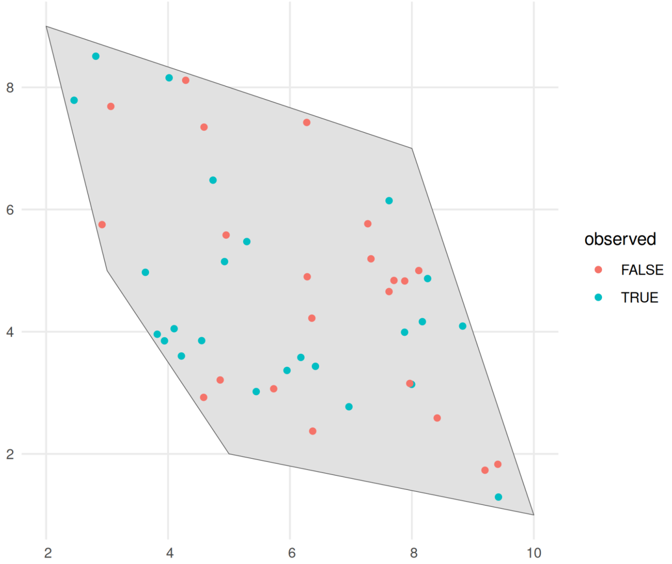


2. Detection process

In the second step we define the sampling process, based on the detection probability of the species and the sampling bias. This is done using the `sample_observations()` function. The default sampling bias is `"no_bias"`, but bias can be added using a polygon or a grid as well.

```
# Detect occurrences
detections_df_raw <- sample_observations(
  occurrences = occurrences_df,
  detection_probability = 0.5,
  sampling_bias = c("no_bias", "polygon", "manual"),
  seed = 123)
```

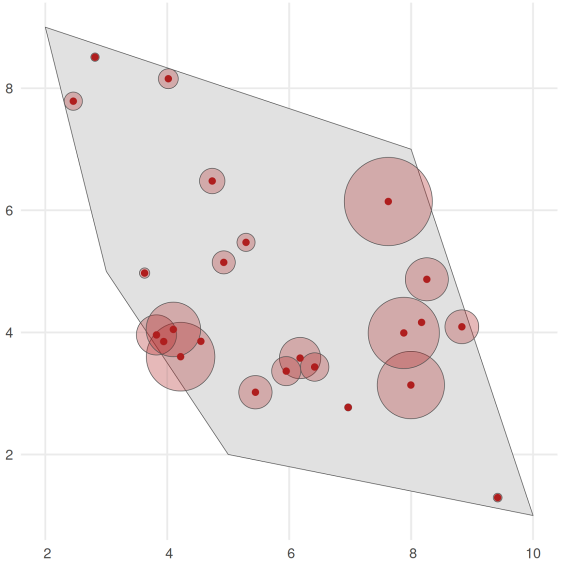
```
# Visualise
ggplot() +
  geom_sf(data = polygon) +
  geom_sf(data = detections_df_raw,
          aes(colour = observed)) +
  theme_minimal()
```



We select the detected occurrences and add an uncertainty to these observations. This can be done using the `filter_observations()` and `add_coordinate_uncertainty()` functions, respectively.

```
# Select detected occurrences only
detections_df <- filter_observations(
  observations_total = detections_df_raw)

# Add coordinate uncertainty
set.seed(123)
coord_uncertainty_vec <- rgamma(nrow(detections_df), shape = 2, rate = 6)
observations_df <- add_coordinate_uncertainty(
  observations = detections_df,
  coords_uncertainty_meters = coord_uncertainty_vec)

# Created and sf object with uncertainty circles to visualise
buffered_observations <- st_buffer(
  observations_df,
  observations_df$coordinateUncertaintyInMeters)
```

```
# Visualise
ggplot() +
  geom_sf(data = polygon) +
  geom_sf(data = buffered_observations,
          fill = alpha("firebrick", 0.3)) +
  geom_sf(data = observations_df, colour = "firebrick") +
  theme_minimal()
```



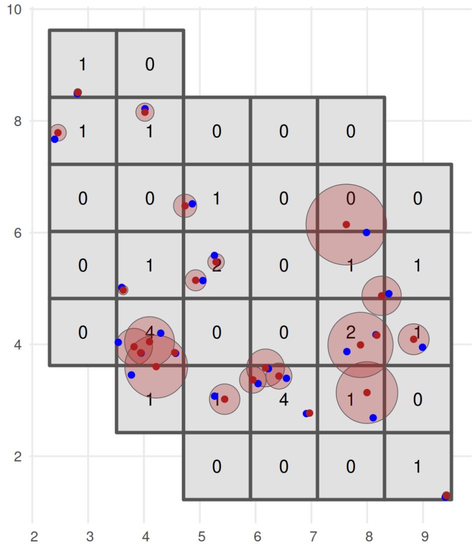
3. Grid designation process

Finally, observations are designated to a grid with `grid_designation()` to create an occurrence cube. We create a grid over the spatial extent using `sf::st_make_grid()`.

```
# Define a grid over spatial extend
grid_df <- st_make_grid(
    buffered_observations,
    square = TRUE,
    cellsize = c(1.2, 1.2)
  ) %>%
  st_sf() %>%
  mutate(intersect = as.vector(st_intersects(geometry, polygon,
                                             sparse = FALSE))) %>%
  dplyr::filter(intersect == TRUE) %>%
  dplyr::select(-"intersect")
```

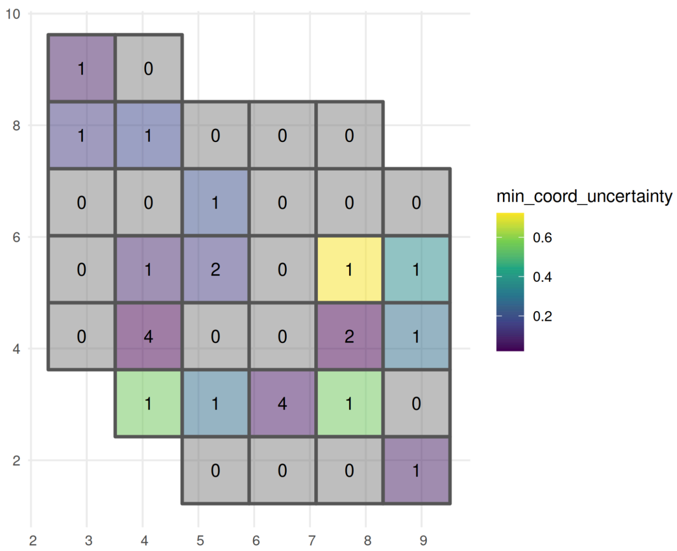
To create an occurrence cube, `grid_designation()` will randomly take a point within the uncertainty circle around the observations. These points can be extracted by setting the argument `aggregate = FALSE`.

```
# Create occurrence cube
occurrence_cube_df <- grid_designation(
  observations = observations_df,
  grid = grid_df,
  seed = 123)

# Get sampled points within uncertainty circle
sampled_points <- grid_designation(
  observations = observations_df,
  grid = grid_df,
  aggregate = FALSE,
  seed = 123)

# Visualise grid designation
ggplot() +
  geom_sf(data = occurrence_cube_df, linewidth = 1) +
  geom_sf_text(data = occurrence_cube_df, aes(label = n)) +
  geom_sf(data = buffered_observations,
          fill = alpha("firebrick", 0.3)) +
  geom_sf(data = sampled_points, colour = "blue") +
  geom_sf(data = observations_df, colour = "firebrick") +
  labs(x = "", y = "", fill = "n") +
  theme_minimal()
```



The output gives the number of observations per grid cell and minimal coordinate uncertainty per grid cell.

```
# Visualise minimal coordinate uncertainty
ggplot() +
  geom_sf(data = occurrence_cube_df, aes(fill = min_coord_uncertainty),
          alpha = 0.5, linewidth = 1) +
  geom_sf_text(data = occurrence_cube_df, aes(label = n)) +
  scale_fill_continuous(type = "viridis") +
  labs(x = "", y = "") +
  theme_minimal()
```



6.1.3. Creating cubes for multiple species

Each cube simulation function mentioned in the previous paragraph has a corresponding mapping function (Table 11). These mapping functions are designed to handle operations for multiple species simultaneously by using the `purrr::pmap()` function. Consult the documentation for detailed information on how these mapping functions are implemented.

6.2. Explore occurrence cube quality and reliability in R

Task 5.4 of the B3 project focuses on the development of general measures for assessing the quality and reliability of biodiversity data cubes. This task also provides a framework for quantifying and interpreting indicator uncertainty, which is essential for robust estimation of species status and trends (Langeraert et al., 2025). A central outcome of this work is the development of the **dubicube** R package (Langeraert & Van Daele, 2026). Key insights from this deliverable (see next chapter) are being translated into practical data quality diagnostics (e.g. automated warnings) to support transparent and informed occurrence cube use and indicator calculation in R.

Table 11: Overview of gcube core simulation functions.

Single species	Multiple species	Description
simulate_occurrences()	map_simulate_occurrences()	Simulate species occurrences within a spatiotemporal scope
sample_observations()	map_sample_observations()	Sample observations from a larger occurrence dataset
filter_observations()	map_filter_observations()	Filter detected occurrences
add_coordinate_uncertainty()	map_add_coordinate_uncertainty()	Add coordinate uncertainty to observations
grid_designation()	map_grid_designation()	Observations to grid designation to create a data cube

7. Guidelines for reliable indicator and trend calculations from occurrence cubes

7.1. Cube preparation and baseline data filtering

This first step defines the analytical boundaries and removes technically invalid records. It establishes the spatial, temporal, and biological scope within which indicators will be calculated.

1. Explicitly **define the spatial and temporal scope of the analysis** before generating occurrence cubes, including grid resolution, geographic extent, and intended indicator type (spatial vs. temporal).

2. Apply **baseline data quality filters** in the SQL query to remove obvious technical errors and unsuitable records:

```
occurrenceStatus = 'PRESENT'
AND NOT occurrence.basisofrecord IN ('FOSSIL_SPECIMEN', 'LIVING_SPECIMEN')
AND NOT ARRAY_CONTAINS(issue, 'ZERO_COORDINATE')
AND NOT ARRAY_CONTAINS(issue, 'COORDINATE_OUT_OF_RANGE')
AND NOT ARRAY_CONTAINS(issue, 'COORDINATE_INVALID')
AND NOT ARRAY_CONTAINS(issue, 'COUNTRY_COORDINATE_MISMATCH')
AND speciesKey IS NOT NULL
AND decimalLatitude IS NOT NULL
AND decimalLongitude IS NOT NULL
```

7.2. Exploratory assessment of data coverage and structure

This step is descriptive and diagnostic. Its purpose is to understand how the data are distributed and structured before defining analytical constraints or exclusion rules.

1. Examine the **spatial distribution** of observations to ensure the study area is well covered and that data density is sufficient for the chosen grid resolution.

2. Summarise **coordinate uncertainty** to quantify the proportion of missing or large uncertainties.

3. Inspect the **temporal distribution** of observations to identify gaps, irregularities, or strong drop-offs in recent years.

4. Perform **basic taxonomic checks**, including counts of species, identification of unlinked accepted names, and detection of unexpected or implausible occurrences that may indicate misidentifications or taxonomic inconsistencies.

5. Inspect the **contribution of individual datasets.** Assess how many datasets contribute to the occurrence cube and how uneven their contributions are in terms of observations

and species coverage. Strong dominance by a few datasets is common and should be expected.

6. **Group-level sensitivity analyses** can help to explore how indicator values respond to the inclusion or exclusion of predefined groups (e.g. datasets, species, spatial units, or time periods). Such analyses are useful for identifying influential components, understanding structural properties of indicators, and distinguishing methodological effects from patterns that may otherwise be interpreted as ecological signals. Sensitivity analysis supports more informed interpretation of indicator results when working with heterogeneous occurrence data.

7.3. Operational quality criteria and documentation

This step translates the study design and diagnostic findings into explicit, reproducible analytical rules. It defines what data are considered suitable for indicator calculation at the chosen resolution and how decisions are documented.

1. **Define coordinate uncertainty thresholds consistent with the spatial scale and indicator purpose**. For spatial indicators, exclude records with coordinate uncertainty exceeding the grid resolution and treat records with missing uncertainty conservatively. For temporal indicators covering broad areas, prioritise data completeness while documenting potential spatial imprecision.

2. **Define temporal inclusion criteria, including cut-off years where necessary**, particularly for recent years. Publication delays can lead to incomplete data for the most recent periods, and apparent declines or anomalies in temporal trends should not be interpreted without first confirming data completeness.

3. **Apply taxonomic harmonisation procedures where required**, including unlinked accepted names, incorrect taxonomic mappings during data publication, and potential species misidentifications. Such issues can result in artificial species duplication, missing species, or implausible occurrences outside known distributions, and may require local harmonisation or reporting issues to GBIF.

4. Where exploratory **sensitivity analyses reveal strong dependence** on specific datasets, taxa, spatial units, or time periods, explicitly define and implement alternative subsets (e.g. excluding dominant datasets or specific taxonomic groups), recalculate the indicator under these conditions, and quantify how results differ from the main analysis.

5. **Document all filtering, harmonisation, and data-quality decisions transparently**, including uncertainty thresholds, temporal cut-off dates, taxonomic corrections, and any dataset-level decisions to ensure reproducibility and transparency.

7.4. Examples

Below are examples from the SQL query to obtain the unstructured data used in the spatial analysis of bird data for the Western Cape of South Africa. The query followed the recommendations above. The full query can be found in Annex 2.

1. Data used covers 2015-2023 for an analysis performed in 2025.

```
AND \"year\" >= 2015
AND \"year\" <= 2023
```

2. The analysis was performed at the quarter-degree grid cell resolution, and thus occurrences with a coordinateUncertainty > 27 000 m (27 km) have been filtered out.

```
AND (coordinateUncertaintyInMeters <= 27000 OR
coordinateUncertaintyInMeters IS NULL)
```

3. Unknown coordinate uncertainties have been set to 27 000 m (the length of the edge of the grid cells).

```
COALESCE(coordinateUncertaintyInMeters, 27000))
```

4. Coordinates have been randomly assigned to grid cells within their uncertainty.

```
MIN(COALESCE(coordinateUncertaintyInMeters, 27000)) AS
minCoordinateUncertaintyInMeters
```

5. Suggested data quality filters have been implemented.

```
AND occurrenceStatus = 'PRESENT'
AND NOT occurrence.basisofrecord IN ('FOSSIL_SPECIMEN', 'LIVING_SPECIMEN')
AND NOT ARRAY_CONTAINS(issue, 'ZERO_COORDINATE')
AND NOT ARRAY_CONTAINS(issue, 'COORDINATE_OUT_OF_RANGE')
AND NOT ARRAY_CONTAINS(issue, 'COORDINATE_INVALID')
AND NOT ARRAY_CONTAINS(issue, 'COUNTRY_COORDINATE_MISMATCH')
AND speciesKey IS NOT NULL
AND decimalLatitude IS NOT NULL
AND decimalLongitude IS NOT NULL
```

8. Acknowledgements

The authors would like to thank Lissa Breugelmans and Sandra MacFadyen for their thorough review of this document. We also thank Shawn Dove for his technical assistance with the **b3gbi** R package, and Johan van Rooyen for providing essential information regarding the Hessequa Systematic Atlasing Subproject. Finally, the authors acknowledge the use of Google Gemini, a large language model, which was used following the completion of the main draft to refine the structure, continuity, and clarity of the text.

9. References

Abraham, L., Hendrickx, L., Groom, Q., Yovcheva, N., Rocchini, D., & Dove, S. (2025). *Hackathon results* (Deliverables B-Cubed D1.7). https://b-cubed.eu/library

Bayraktarov, E., Ehmke, G., O'Connor, J., Burns, E. L., Nguyen, H. A., McRae, L., Possingham, H. P., & Lindenmayer, D. B. (2019). Do Big Unstructured Biodiversity Data Mean More Knowledge? *Frontiers in Ecology and Evolution*, *6*, 239. https://doi.org/10.3389/fevo.2018.00239

Blisset, M., Høfft, M., Waller, J., Rodrigues, A., Noesgaard, D., Robertson, T., & Desmet, P. (2025). *Occurrence cube service* (Deliverables B-Cubed D2.3). https://b-cubed.eu/library

Boakes, E. H., McGowan, P. J. K., Fuller, R. A., Chang-qing, D., Clark, N. E., O'Connor, K., & Mace, G. M. (2010). Distorted Views of Biodiversity: Spatial and Temporal Bias in Species Occurrence Data. *PLoS Biology*, *8*(6), e1000385. https://doi.org/10.1371/journal.pbio.1000385

Brooks, M., Rose, S., Altwegg, R., Lee, A. T., Nel, H., Ottosson, U., Retief, E., Reynolds, C., Ryan, P. G., Shema, S., Tende, T., Underhill, L. G., & Thomson, R. L. (2022). The African Bird Atlas Project: A description of the project and BirdMap data-collection protocol. *Ostrich*, *93*(4), 223–232. https://doi.org/10.2989/00306525.2022.2125097

Brooks, M., & Ryan, P. (2023). *Southern African Bird Atlas Project 2* [Data set]. FitzPatrick Institute of African Ornithology. https://www.gbif.org/dataset/906e6978-e292-4a8b-9c39-adf6bb0f3323

Burgass, M. J., Halpern, B. S., Nicholson, E., & Milner-Gulland, E. J. (2017). Navigating uncertainty in environmental composite indicators. *Ecological Indicators*, *75*, 268–278. https://doi.org/10.1016/j.ecolind.2016.12.034

Callaghan, C. T., Martin, J. M., Major, R. E., Kingsford, R. T., Callaghan, C. T., Martin, J. M., Major, R. E., & Kingsford, R. T. (2018). Avian monitoring – comparing structured and unstructured citizen science. *Wildlife Research*, *45*(2), 176–184. https://doi.org/10.1071/WR17141

Cartuyvels, E., Faulkner, K., Langeraert, W., & Van Daele, T. (2025). *Preliminary criteria for data quality and species characteristics for estimating species status and trends* (Milestones B-Cubed No. MS19). https://b-cubed.eu/library

Chamberlain, S., Barve, V., McGlinn, D., Oldoni, D., Desmet, P., Geffert, L., & Ram, K. (2025). *rgbif: Interface to the Global Biodiversity Information Facility API* [Computer software]. https://CRAN.R-project.org/package=rgbif

Chao, A., Gotelli, N. J., Hsieh, T. C., Sander, E. L., Ma, K. H., Colwell, R. K., & Ellison, A. M. (2014). Rarefaction and extrapolation with Hill numbers: A framework for sampling and estimation in species diversity studies. *Ecological Monographs*, *84*(1), 45–67. https://doi.org/10.1890/13-0133.1

Chao, A., Kubota, Y., Zelený, D., Chiu, C., Li, C., Kusumoto, B., Yasuhara, M., Thorn, S., Wei, C., Costello, M. J., & Colwell, R. K. (2020). Quantifying sample completeness and comparing diversities among assemblages. *Ecological Research*, *35*(2), 292–314. https://doi.org/10.1111/1440-1703.12102

Christie, A. P., Amano, T., Martin, P. A., Shackelford, G. E., Simmons, B. I., & Sutherland, W. J. (2019). Simple study designs in ecology produce inaccurate estimates of biodiversity responses. *Journal of Applied Ecology*, *56*(12), 2742–2754. https://doi.org/10.1111/1365-2664.13499

Clarance, D. (2019). *rabm: An interface to the Africa Bird Atlas API* [Computer software]. https://github.com/davidclarance/rabm

Desmet, P., Oldoni, D., Blisset, M., & Robertson, T. (2025). *Specification for species occurrence cubes and their production* (Deliverables B-Cubed D2.1v1.1). https://b-cubed.eu/library

Dove, S. (2026). *b3gbi: General Biodiversity Indicators for Biodiversity Data Cubes* [Computer software]. https://github.com/b-cubed-eu/b3gbi

Faulkner, K. (2026). *Rsa-unstructured-data-comp* (Version 1.0.0) [Computer software]. https://github.com/b-cubed-eu/rsa-unstructured-data-comp

Foster, Z. S. L., Sharpton, T. J., & Grünwald, N. J. (2017). Metacoder: An R package for visualization and manipulation of community taxonomic diversity data. *PLOS Computational Biology*, *13*(2), e1005404. https://doi.org/10.1371/journal.pcbi.1005404

García-Roselló, E., González-Dacosta, J., & Lobo, J. M. (2023). The biased distribution of existing information on biodiversity hinders its use in conservation, and we need an integrative approach to act urgently. *Biological Conservation*, *283*, 110118. https://doi.org/10.1016/j.biocon.2023.110118

Groom, Q., Adriaens, T., Desmet, P., Vanderhoeven, S., & Yovcheva, N. (2025). *Why countries need the Global Biodiversity Information Facility: Lessons from Belgium [policy brief]* (Version 4). https://doi.org/10.5281/ZENODO.16890979

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

Hugo, S., & Altwegg, R. (2017). The second Southern African Bird Atlas Project: Causes and consequences of geographical sampling bias. *Ecology and Evolution*, *7*(17), 6839–6849. https://doi.org/10.1002/ece3.3228

Isaac, N. J. B., van Strien, A. J., August, T. A., de Zeeuw, M. P., & Roy, D. B. (2014). Statistics for citizen science: Extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution*, *5*(10), 1052–1060. https://doi.org/10.1111/2041-210X.12254

Kamp, J., Oppel, S., Heldbjerg, H., Nyegaard, T., & Donald, P. F. (2016). Unstructured citizen science data fail to detect long-term population declines of common birds in Denmark. *Diversity and Distributions*, *22*(10), 1024–1035. https://doi.org/10.1111/ddi.12463

Langeraert, W. (2026). *gcube: Simulating Biodiversity Data Cubes* [Computer software]. https://github.com/b-cubed-eu/gcube

Langeraert, W., Barhdadi, W., Brosens, D., Cortès, R., Desmet, P., Di Musciano, M., Earl, C., Govaert, S., Huybrechts, P., Martini, M., Rodrigues, A. V., Veeken, A., Yahaya, M. M., & Van Daele, T. (2024). *Unveiling ecological dynamics through simulation and visualization of biodiversity data cubes*. BioHackrXiv. https://doi.org/10.37044/osf.io/vcyr7

Langeraert, W., Cartuyvels, E., & Van Daele, T. (2026). *Compare unstructured data* (Version 1.0.0) [Computer software]. https://github.com/b-cubed-eu/comp-unstructured-data

Langeraert, W., Desmet, P., & Van Daele, T. (2025). *Design of R packages for indicator calculation* (Milestones B-Cubed No. MS28). https://b-cubed.eu/library

Langeraert, W., Faulkner, K., Zengeya, T., Martini, M., Rocchini, D., Breugelmans, L., Cortès Lobos, R. B., & Van Daele, T. (2023). *Selection of the monitoring and inventory projects: Selection of species (groups), spatial and temporal extent* (Milestones B-Cubed No. MS18). https://b-cubed.eu/library

Langeraert, W., & Van Daele, T. (2026). *dubicube: Calculation and Interpretation of Data Cube Indicator Uncertainty* [Computer software]. https://github.com/b-cubed-eu/dubicube

Leroy, B., Meynard, C. N., Bellard, C., & Courchamp, F. (2016). Virtualspecies, an R package to generate virtual species distributions. *Ecography*, *39*(6), 599–607. https://doi.org/10.1111/ecog.01388

Münkemüller, T., De Bello, F., Meynard, C. N., Gravel, D., Lavergne, S., Mouillot, D., Mouquet, N., & Thuiller, W. (2012). From diversity indices to community assembly processes: A test with simulated data. *Ecography*, *35*(5), 468–480. https://doi.org/10.1111/j.1600-0587.2011.07259.x

Onkelinx, T., Vermeersch, G., & Devos, K. (2023). *Trends op basis van de Algemene Broedvogelmonitoring Vlaanderen (ABV): Technisch achtergrondrapport voor de periode 2007-2022* (No. 1; Rapporten van Het Instituut Voor Natuur- En Bosonderzoek). Instituut voor Natuur- en Bosonderzoek. https://doi.org/10.21436/inbor.89419879

Piesschaert, F., Vermeersch, G., Brosens, D., Westra, T., Desmet, P., Feys, S., Van De Poel, S., & Pollet, M. (2022). *ABV - Common breeding birds in Flanders, Belgium (post 2016)* [Data set]. Research Institute for Nature and Forest (INBO). https://doi.org/10.15468/PJ2V6H

R Core Team. (2025). *R: A Language and Environment for Statistical Computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Sokol, E. R., Brown, B. L., & Barrett, J. E. (2017). A simulation-based approach to understand how metacommunity characteristics influence emergent biodiversity patterns. *Oikos*, *126*(5), 723–737. https://doi.org/10.1111/oik.03690

Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R., & Legendre, F. (2017). Taxonomic bias in biodiversity data and societal preferences. *Scientific Reports*, *7*(1), 9132. https://doi.org/10.1038/s41598-017-09084-6

Underhill, L., & Van Rooyen, J. (2020). Systematic atlasing in Hessequa – Report on the first cycle of seasonal monitoring. *Biodiversity Observations*, *11*. https://doi.org/10.15641/bo.933

Van Eupen, C., Maes, D., Herremans, M., Swinnen, K. R. R., Somers, B., & Luca, S. (2021). The impact of data quality filtering of opportunistic citizen science data on species distribution model performance. *Ecological Modelling*, *444*, 109453. https://doi.org/10.1016/j.ecolmodel.2021.109453 van Rooyen, J. A. (2018). Systematic atlasing in Hessequa—Moving from mapping to monitoring. *Biodiversity Observations*, *9*(10), 1–13. https://doi.org/10.15641/bo.v9i0.508

Van Strien, A. J., Termaat, T., Kalkman, V., Prins, M., De Knijf, G., Gourmand, A.-L., Houard, X., Nelson, B., Plate, C., Prentice, S., Regan, E., Smallshire, D., Vanappelghem, C., & Vanreusel, W. (2013). Occupancy modelling as a new approach to assess supranational trends using opportunistic data: A pilot study for the damselfly Calopteryx splendens. *Biodiversity and Conservation*, *22*(3), 673–686. https://doi.org/10.1007/s10531-013-0436-1

Vermeersch, G. (2007). *Sampling framework for the common breeding bird survey in Flanders, Belgium* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.10103472

Vermeersch, G., Anselin, A., Herremans, M., & Brosens, D. (2021). *ABV - Common breeding birds in Flanders, Belgium* [Data set]. Research Institute for Nature and Forest (INBO). https://doi.org/10.15468/XJ0IKB

Vermeersch, G., Ledegen, H., & Feys, S. (2018). *Methodehandleiding bij het project Algemene Broedvogelmonitoring Vlaanderen (ABV)* (Rapporten van het Instituut voor Natuur- en Bosonderzoek No. 93; Rapporten van het Instituut voor Natuur- en Bosonderzoek). Instituut voor Natuur en Bosonderzoek. https://doi.org/10.21436/inbor.15674942

Zurell, D., Berger, U., Cabral, J. S., Jeltsch, F., Meynard, C. N., Münkemüller, T., Nehrbass, N., Pagel, J., Reineking, B., Schröder, B., & Grimm, V. (2010). The virtual ecologist approach: Simulating data and observers. *Oikos*, *119*(4), 622–635. https://doi.org/10.1111/j.1600-0706.2009.18284.x

10. Annex

10.1. SQL queries Flanders

SQL query with extra component dataset dimension used in Chapter 3.4.

```
"SELECT
  \"year\",
  GBIF_MGRSCode(1000, decimalLatitude, decimalLongitude,
  COALESCE(coordinateUncertaintyInMeters, 1000)) AS mgrsCode,
  speciesKey,
  species,
  family,
  datasetName,
  datasetKey,
  COUNT(*) AS n,
  MIN(COALESCE(coordinateUncertaintyInMeters, 1000)) AS
minCoordinateUncertaintyInMeters,
  IF(ISNULL(family), NULL, SUM(COUNT(*)) OVER (PARTITION BY family)) AS
familyCount
  FROM
  occurrence
  WHERE
  occurrenceStatus = 'PRESENT'
  AND NOT ARRAY_CONTAINS(issue, 'ZERO_COORDINATE')
  AND NOT ARRAY_CONTAINS(issue, 'COORDINATE_OUT_OF_RANGE')
  AND NOT ARRAY_CONTAINS(issue, 'COORDINATE_INVALID')
  AND NOT ARRAY_CONTAINS(issue, 'COUNTRY_COORDINATE_MISMATCH')
  AND level1gid = 'BEL.2_1'
  AND \"year\" >= 2007
  AND \"year\" <= 2022
  AND speciesKey IS NOT NULL
  AND decimalLatitude IS NOT NULL
  AND decimalLongitude IS NOT NULL
  AND class = 'Aves'
  AND collectionCode != 'ABV'
  GROUP BY
  \"year\",
  mgrsCode,
  speciesKey,
  species,
  family,
  datasetName,
datasetKey
ORDER BY
\"year\" ASC,
mgrsCode ASC,
speciesKey ASC,
datasetName ASC"
```

10.2. SQL queries Western Cape

The SQL query to obtain the unstructured data used in the spatial analysis of bird data for the Western Cape of South Africa.

```
occ_download_sql("SELECT \"year\", GBIF_EQDGCode(2, decimalLatitude,
decimalLongitude, COALESCE(coordinateUncertaintyInMeters, 27000))
AS qdgcCode,speciesKey,species, \"order\", family, genus,
COUNT(*) AS n,
MIN(COALESCE(coordinateUncertaintyInMeters, 27000)) AS
minCoordinateUncertaintyInMeters,
IF(ISNULL(\"order\"), NULL, SUM(COUNT(*)) OVER (PARTITION BY \"order\")) AS
orderCount, IF(ISNULL(family), NULL, SUM(COUNT(*)) OVER (PARTITION BY
family)) AS familyCount, IF(ISNULL(genus), NULL, SUM(COUNT(*)) OVER
(PARTITION BY genus)) AS genusCount FROM occurrence
WHERE class = 'Aves'
AND occurrenceStatus = 'PRESENT'
AND (coordinateUncertaintyInMeters <= 27000 OR
coordinateUncertaintyInMeters IS NULL) AND NOT occurrence.basisofrecord IN
('FOSSIL_SPECIMEN', 'LIVING_SPECIMEN')
AND NOT ARRAY_CONTAINS(issue, 'ZERO_COORDINATE')
AND NOT ARRAY_CONTAINS(issue, 'COORDINATE_OUT_OF_RANGE')
AND NOT ARRAY_CONTAINS(issue, 'COORDINATE_INVALID')
AND NOT ARRAY_CONTAINS(issue, 'COUNTRY_COORDINATE_MISMATCH')
AND level1Gid = 'ZAF.9_1'
AND \"year\" >= 2015
AND \"year\" <= 2023
AND speciesKey IS NOT NULL
AND decimalLatitude IS NOT NULL AND decimalLongitude IS NOT NULL
AND collectionCode != 'SABAP2'
GROUP BY
\"year\",
qdgcCode,
speciesKey,
\"order\",
family,
genus,
species
ORDER BY
"year\" ASC,
qdgcCode ASC,
speciesKey ASC")
```