BIODIVERSITY BUILDING BLOCKS FOR POLICY

MS19 Preliminary criteria for data quality and species characteristics for estimating species status and trends.

30/04/2025

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

Funded by the European Union

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

Table of contents

Summary

Datasets on GBIF are known to have several inherent biases, which can lead to misleading patterns or trends in derived indicators. To establish the necessary conditions for ensuring the accuracy of these indicators, this study compared unstructured cube data from GBIF with data from structured monitoring programs, which are designed to minimize biases and provide a largely accurate representation of the ecological systems.

The initial challenges encountered were related to technical aspects and data quality. These included observations falling outside the selected region due to substantial coordinate uncertainty and subsequent randomization, necessitating filtering methods that could potentially result in data loss. Inconsistencies in species names, such as non-synonymous names for the same species, translation errors during publication, and misidentifications, also required rigorous quality control. Furthermore, delays in data publication to GBIF hampered the assessment of recent temporal trends. Lastly, technical issues arose when attempting to map indicators using the b3gbi R package, particularly with non-standard grid resolutions.

The analysis revealed that unstructured data are less reliable for capturing the ranges and trends of rare or very rare species, often exhibiting poor range overlap and trend correlation with structured data. This discrepancy can be attributed to factors like underreporting, misidentification, and the absence of standardized sampling effort in unstructured data, a problem likely to be amplified for less popular species groups.

Despite these limitations, the study also identified positive developments. Range correlations between structured and unstructured data generally improved over time, indicating enhanced data coverage and quality within GBIF. This encouraging trend underscores the potential of unstructured data to serve as a valuable complement to structured data for biodiversity monitoring, especially with continued improvements in data availability and quality. Nevertheless, the inherent heterogeneity and potential biases in unstructured data necessitate careful filtering, validation, and interpretation to ensure the generation of robust and meaningful results.

List of abbreviations

EU	European Union
ABV	Common Breeding Bird Survey Flanders (Algemene Broedvogelmonitoring Vlaanderen)
UTM	Universal Transverse Mercator
GBIF	Global Biodiversity Information Facility
SABAP2	Southern African Bird Atlas Programme 2
QDGC	Quarter-degree Grid Cell

1 Introduction

A data cube is a multidimensional representation of data that allows for efficient storage, retrieval, and analysis of information along multiple dimensions. In the context of biodiversity data, a data cube can integrate various dimensions such as taxonomic (what), temporal (when) and spatial (where), enabling researchers to explore complex ecological patterns and relationships more effectively.

In the context of the B3 project, software was created to generate species occurrence cubes (Desmet et al. 2023). This software has been implemented as a public download service by the Global Biodiversity Information Facility (GBIF) (Blissett et al. 2025). GBIF is an international network and data infrastructure that hosts over 100,000 biodiversity datasets. Multiple workflows developed in the project start from occurrence cubes made from all available GBIF data. However, it is known that the available data on GBIF has several issues:

1. Taxonomic biases (Troudet et al. 2017)
2. Geographic biases (García-Roselló, González-Dacosta, and Lobo 2023)
3. Species detectability issues that lead to gaps in data for taxa that are not easily detected
4. Non-random distribution of sampling effort. Some areas are surveyed more intensively than others (e.g., near roads, in accessible locations (Reddy and Dávalos 2003), popular tourist sites, biodiverse places (Trimble and van Aarde 2012))
5. False positives in species records are common (misidentification, data entry errors, contamination when working with (e)DNA)
6. Changes in sampling effort over time (Boakes et al. 2010)
7. Changes in detection probability over time

Therefore, indicators developed with these data can demonstrate misleading patterns or trends. It will, therefore, be necessary to determine what conditions need to be met to ensure that the resulting indicators are accurate.

To determine the preliminary criteria for data quality and species characteristics, unstructured cube data from GBIF was compared with data collected through structured monitoring programs. The structured monitoring programs used are:

- Common Breeding Bird Survey Flanders (Algemene Broedvogelmonitoring Vlaanderen, ABV)
- Southern African Bird Atlas Programme 2 (SABAP2)

This comparison allows us to assess the accuracy of the unstructured data, rather than its precision, which will be addressed separately under work package 5.4 using the dubicube R package. We expect the structured monitoring data to provide a largely accurate representation of the system, as potential biases have been minimized through careful design and standardized protocols. By evaluating how well the unstructured data aligns with this reference, we aim to gain insights into its reliability and potential for broader biodiversity monitoring applications.

1.1 ABV

The ABV monitoring aims to assess the trend of common breeding birds in Flanders. Flanders is the northern region of Belgium, covering an area of 13.626 km² with a population of almost 7 million.

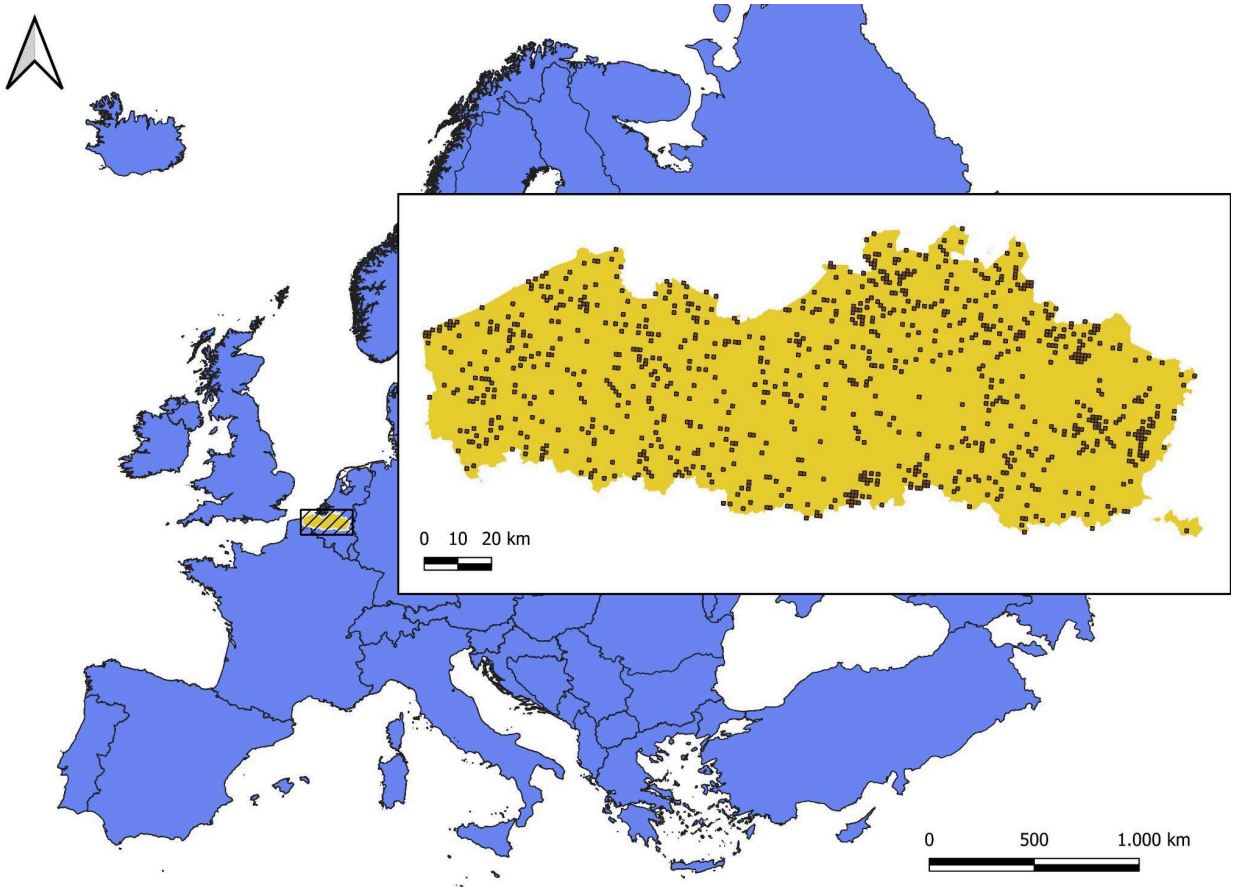

Figure 1: Location of Flanders within Europe and ABV monitoring locations within Flanders.

The ABV monitoring started in 2007 and is still ongoing. The protocol involves selecting a random sample of 1200 UTM 1x1 km grid cells, stratified by land use. These cells are divided into groups of 300, and 300 grid cells are visited each year on a three-year rotation. Each grid cell contains six monitoring locations where bird counts are conducted. The data collection is standardized, with each grid cell being visited three times a year at fixed intervals (at least two weeks apart). Birds are counted based on sound and sight, and counts are conducted between sunrise and 4 hours after sunrise.

This protocol is optimized for detecting trends and is stratified according to six ecosystem types (i.e., agricultural, urban, forest, suburban, heathland and dunes, marsh and water). Ecosystems that are not common in Flanders are oversampled to ensure enough data is collected for trend detection per ecosystem. Due to this and the fact that only ~6% of the total area is monitored the ABV data does not inherently provide accurate spatial distribution patterns.

1.2 Southern African Bird Atlas Programme 2

SABAP2 is a bird monitoring project that began in 2007 (Brooks et al. 2022). The project maps the distribution and relative abundance of birds in South Africa, Lesotho, Botswana, Namibia, Mozambique, Eswatini, Zimbabwe, and Zambia (Brooks et al., 2022). The data are collected by citizen scientists following a standardised protocol (i.e., BirdMap protocol) (Brooks et al., 2022). At least two hours, but a maximum of five days, are spent recording as many different bird species in a grid cell as possible. Observers record birds in the order that they are seen or heard, the hour each species was first detected, and the end of each hour of the survey. The observer is expected to identify at least 90% of the species encountered, and all or most of the representative habitats within the grid cell should be surveyed. All records submitted to SABAP2 undergo a quality check (for details see Brooks et al., 2022). Ad hoc records (those not following the standardised protocol) are also submitted, but these were excluded from our analysis. The grid cells are pentads, i.e., they are at a 5-minute^2 spatial resolution (~ 8.2 km^2 in southern Africa) (Brooks et al., 2022).

Data on when and how many times each grid cell has been surveyed are available. However, repeated surveys over time are not standardised, except for in the Hessequa Systematic Atlasing Subproject (https://sabap2.birdmap.africa/coverage/group/Hssq1).

The Hessequa Systematic Atlassing Subproject is undertaken near Stillbaai in the Western Cape province of South Africa (Van Rooyen 2018). In this subproject, ongoing directed, systematic bird surveys following the SABAP2 standard protocol have been conducted since October 2014 (Van Rooyen 2018; van Rooyen and Underhill 2020). The systematic atlasing evolved over time but every grid cell has been surveyed at least once a year since the start of the project, and since 2021 a system has been in place where each pentad is surveyed twice a year distributed evenly over the seasons (see van Rooyen, 2018; van Rooyen & Underhill, 2020; and https://sabap2.birdmap.africa/coverage/group/Hssq1).

The SABAP2 data from the Western Cape, and from the Hessequa Systematic Atlassing Subproject were used (Fig.2). The Western Cape is 129 462 km² and has a population of over almost 7.5 million; while the Hessequa Systematic Atlassing Subproject is being undertaken in an area that is ~110 x 55 km (75 grid cells), and has a population of ~3500. The analyses were done at two resolutions pentad resolution and quarter-degree grid cell resolution (QDGC). Nine pentads fit into one QDGC (15 min^2).

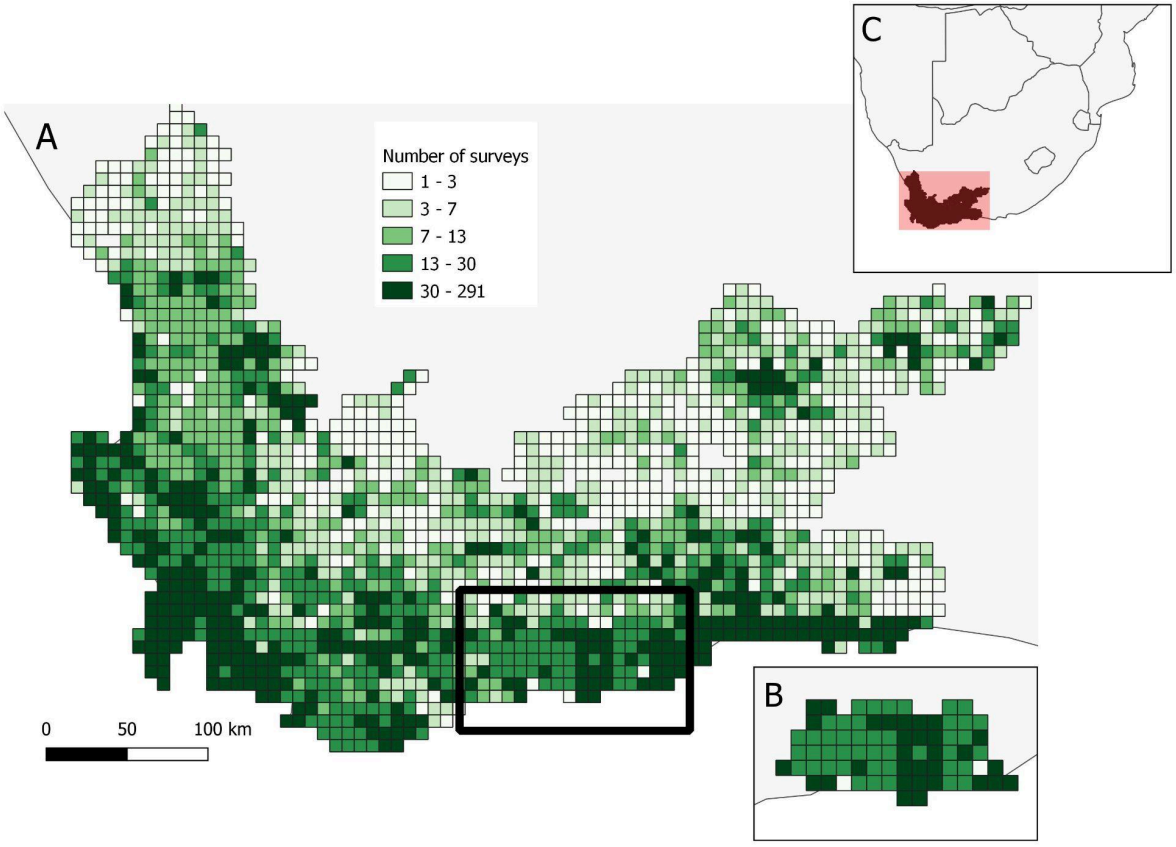

Figure 2: Number of SABAP2 standard protocol surveys per 5-minute² grid cell (pentad) in the Western Cape of South Africa (panel A), and in the area of the Hessequa Systematic Atlassing Subproject (panel B). The inset in panel A shows the location of the Hessequa Systematic Atlassing Subproject in the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

2 Exploration of the data

2.1 Flanders

The ABV data are published on GBIF and were downloaded using the `occ_download_sql()` function from the rgbif R package (Chamberlain, Oldoni, and Waller 2022). Note that GBIF now offers the functionality to do this directly through the GBIF portal, see blogpost. We downloaded the ABV data in the cube format to allow for easy comparisons (GBIF.Org User 2025a). The cube data was downloaded using the same functionality (GBIF.Org User 2025b). All the code used for the Flemish data can be found here.

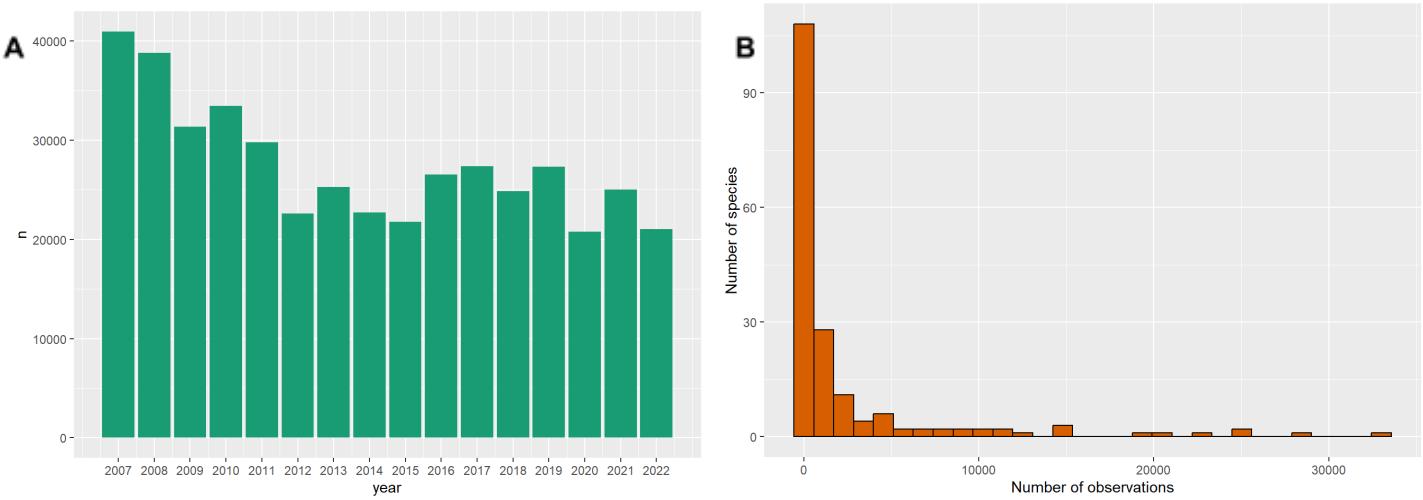

Figure 3: A: Number of observations per year for the ABV dataset. B: Number of observations per species.

The ABV data ranges from over 40,000 observations in 2007 to just over 20,000 observations in 2022 (Fig.3 A). 180 bird species were recorded, of which 34 were recorded < 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.3 B). This categorization is used throughout further analyses.

The cube data was also downloaded directly from GBIF. Observations from ABV monitoring published on GBIF were excluded.

The cube is made up of several datasets, in descending order of number of occurrences:

- Waarnemingen.be - Bird occurrences in Flanders and the Brussels Capital Region, Belgium
- Watervogels - Wintering waterbirds in Flanders, Belgium
- HG_OOSTENDE - Herring gulls (Larus argentatus, Laridae) breeding at the southern North Sea coast (Belgium)
- EOD – eBird Observation Dataset
- Waarnemingen.be - Non-native animal occurrences in Flanders and the Brussels Capital Region, Belgium
- LBBG_ZEEBRUGGE - Lesser black-backed gulls (Larus fuscus, Laridae) breeding at the southern North Sea coast (Belgium and the Netherlands)
- Broedvogels - Atlas of the breeding birds in Flanders 2000-2002
- European Seabirds At Sea (ESAS)
- And 80+ smaller datasets

With the first dataset (waarnemingen.be) containing most of the observations (67%). 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 (Fig.4 A).

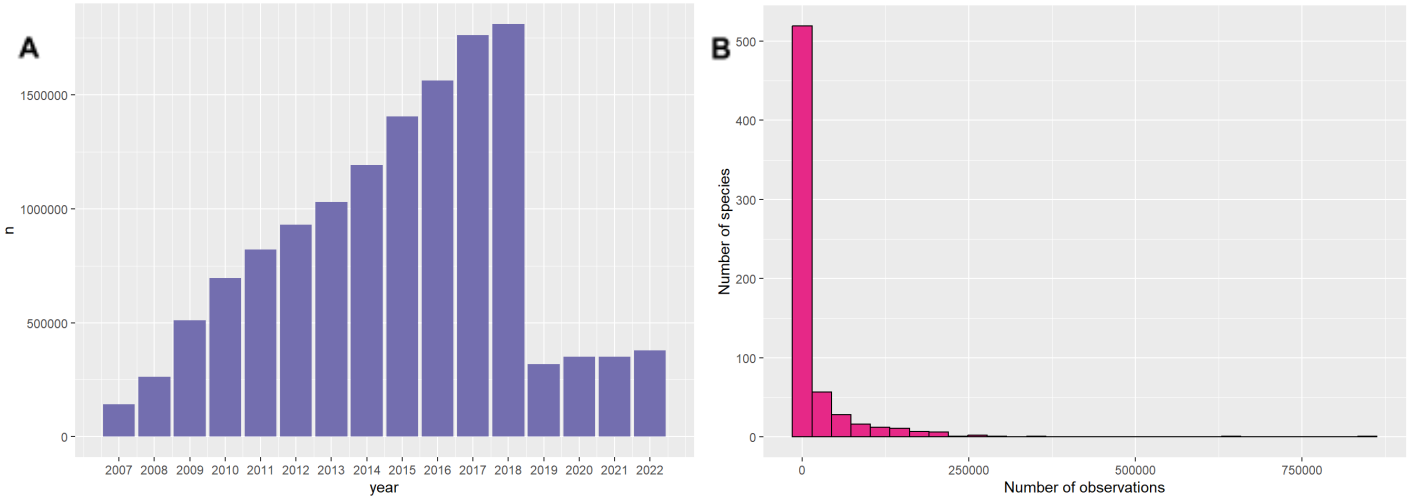

Figure 4: A: Number of observations per year for the cube dataset. B: number of observations per species.

The cube contains information on 663 species. 25 of these are hybrids. 281 of these were observed less than a 100 times, 247 were observed more than a 1000 times (Fig.4 B).

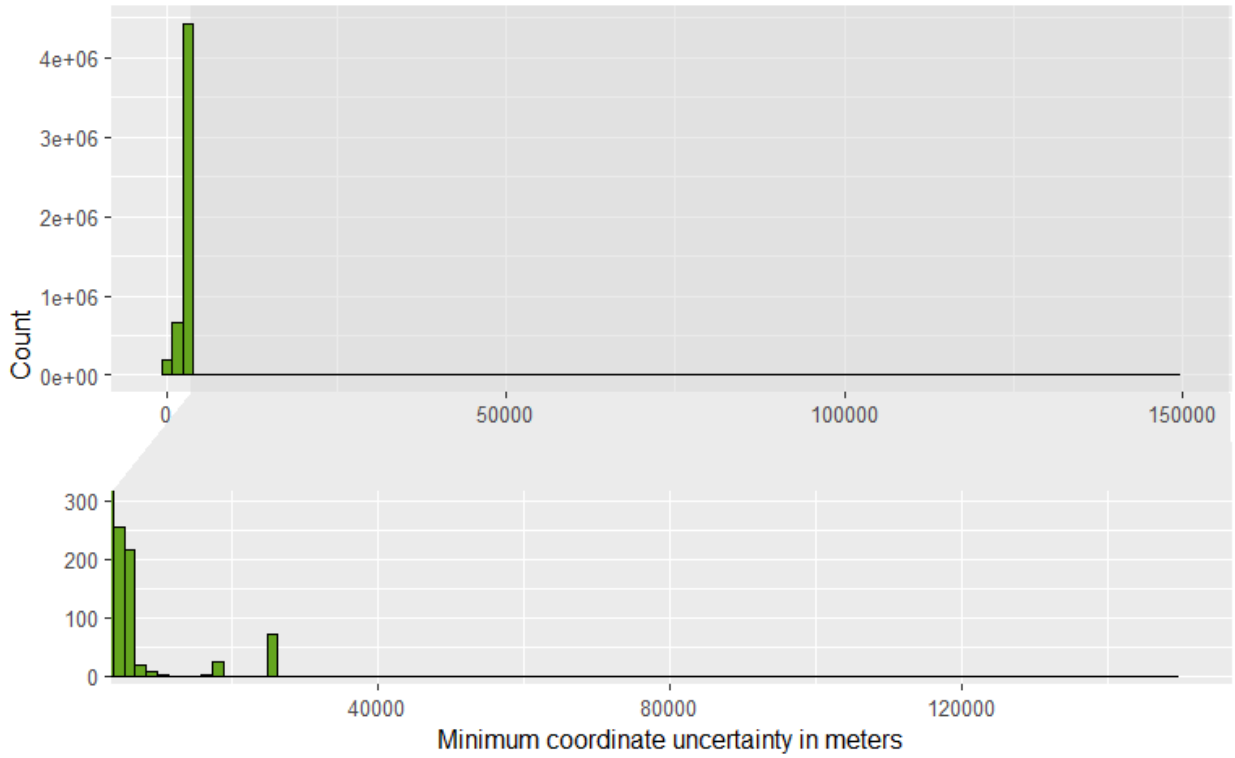

Figure 5: Count of different minimum coordinate uncertainties in the cube.

There are very few observations with zero coordinate uncertainty, most have an uncertainty of 3536 m. This is due to the fact that the waarnemingen.be data (making up 67% of the data) are published based on an aggregation per 5 km.

2.2 South Africa

The SABAP2 data for the Western Cape were downloaded as an occurrence dataset from GBIF (GBIF.Org User 2025c). Although the `occ_download_sql()` function from the rgbif R package (Chamberlain, Oldoni, and Waller 2022) could have been used, the resultant data cubes would have included ad hoc records, and there is no filter available to easily remove them. Therefore, the SABAP2 dataset was downloaded and manipulated to remove ad hoc records. Yearly pentad and QDGC cubes were created for the Western Cape of South Africa and the area of the Hessequa Systematic Atlasing Subproject (hereafter shortened to 'Hessequa'). The cubes for the Western Cape spanned the length of the SABAP2 project (2008-2024), while those for Hessequa spanned the period over which systematic repeated surveys have been performed in that area (2015-2024). All the code used for the South African data can be found here.

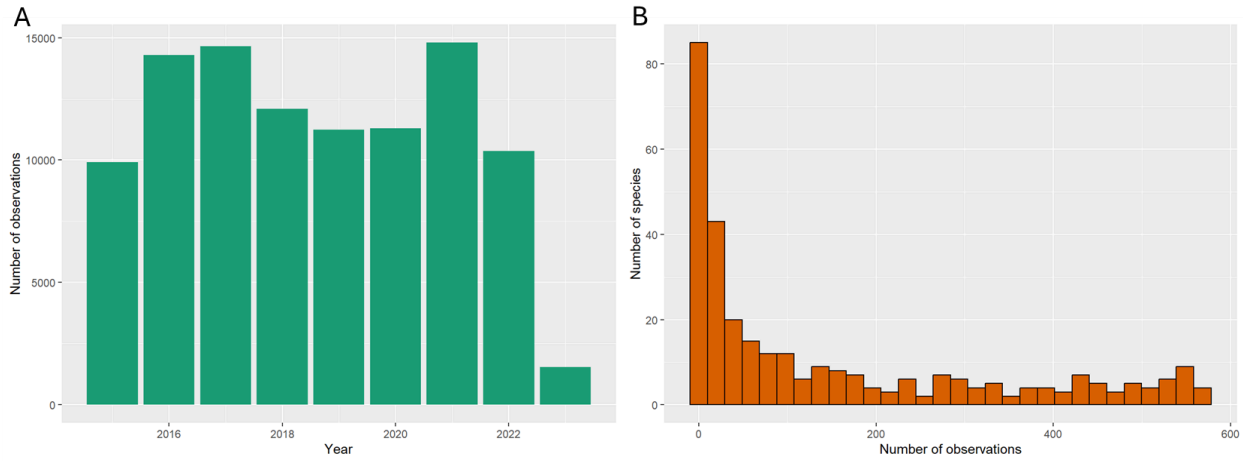

Figure 6: A: Number of observations per year for the SABAP2 dataset for the Hessequa Systematic Atlasing Subproject. B: number of observations per species.

In the SABAP2 data for Hessequa there have been over 100,000 bird observations, with the number of observations ranging between 10,000 - 15,000 per year, excluding 2023 and 2024 (Fig.6 A). 310 bird species were recorded in the Hessequa Systematic Atlasing Subproject between 2015 and 2024 (Fig.6 B), most of which (123 species) were recorded between 100 and 1000 times (common).

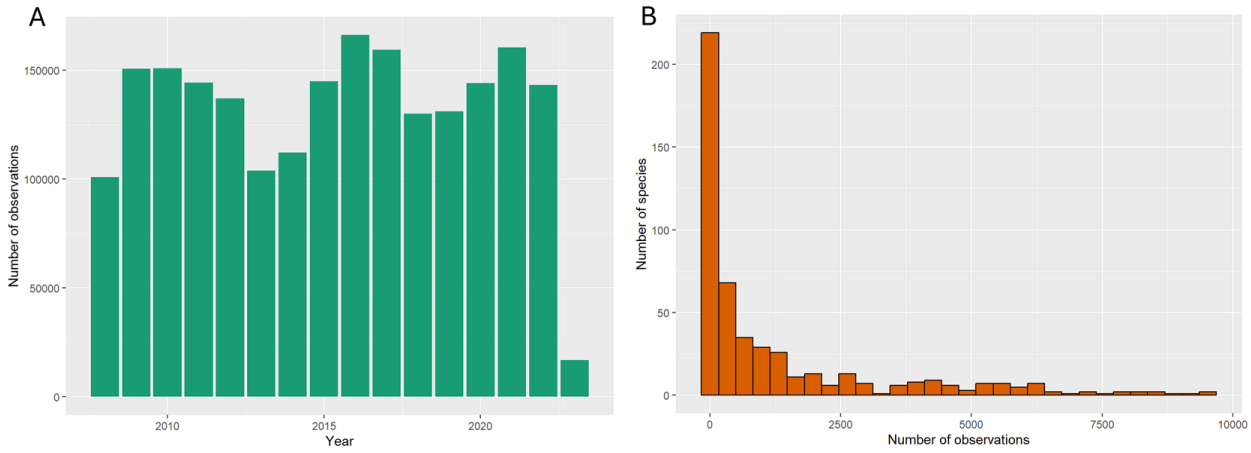

Figure 7: A: Number of observations per year for the SABAP2 dataset for the Western Cape. B: number of observations per species.

In the SABAP2 data for the Western Cape there have been over 2 million bird observations, with the number of observations ranging between 100,000 - 175,000 per year, excluding 2023 and 2024 (Fig.7 A). 502 bird species were recorded in the Western Cape between 2008 and 2024 (Fig.7 B), most of which (162 species) were recorded between 1000 and 10,000 times (very common).

Importantly, there are few observations in SABAP2 for 2023 and none for 2024, as only data up until February 2023 have been submitted to GBIF (see Section 3.3). These delays need to be considered when assessing temporal trends below.

An unstructured QDGC, year data cube for the Western Cape (2008-2024) was downloaded from GBIF (GBIF.Org User 2025d) using the `occ_download_sql()` function from rgbif (Chamberlain, Oldoni, and Waller 2022). These data were subsetted to create a QDGC data cube for Hessequa (2015-2024). It is not possible to download pentad data cubes using rgbif. Therefore, the `map_grid_designation()` function from the gcube R package (Langeraert 2025) was used to create the pentad cube for Hessequa (GBIF.Org User 2025e). It is not yet possible to create a cube using gcube for a large area such as the Western Cape, therefore, the pentad cube for the Western Cape (2008-2024) was created using an occurrence dataset from GBIF (GBIF.Org User 2025f). Observations from SABAP2 were excluded. Note that the period of time the cubes spanned was the same as the SABAP2 data. After the creation of the cubes, records with a minimum coordinate uncertainty of > 8 m for the pentad cubes, and > 27 km for the QDGC cubes were filtered out (see Section 3.1).

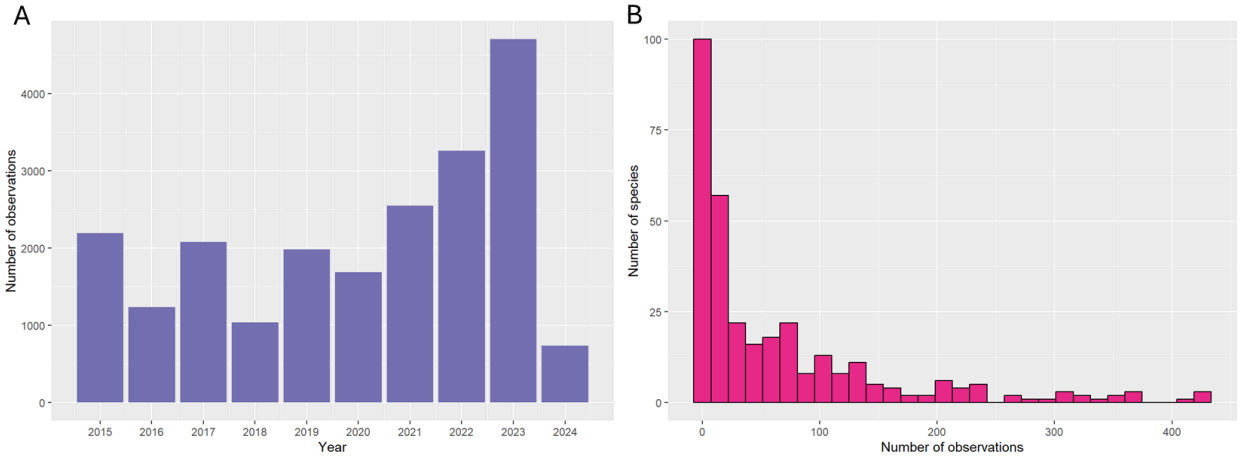

Figure 8: A: Number of observations per year for the unstructured pentad dataset for the area in which the Hessequa Systematic Atlasing Subproject is performed. B: number of observations per species. The unstructured QDGC dataset showed the similar trends.

There were less observations in the unstructured data cubes for Hessequa (21459 for the pentad cube and 38678 for the QDGC cube), than in the SABAP2 data, but the number of observations increased over time (Fig.8). There were more bird species recorded in the unstructured data cubes (322 species in the pentad cube and 356 species in the QDGC cube) for Hessequa than in SABAP2. Most of the species (137) in the unstructured pentad data for Hessequa were recorded between 10 and 100 times, whereas in the QDGC data, most (126) were recorded between 100 and 1000 times. For Hessequa, most of the species in the SABAP2 data were in the unstructured data cubes (88% for the pentad data and 94% for the QDGC data), with the missing species being very rare or rare species.

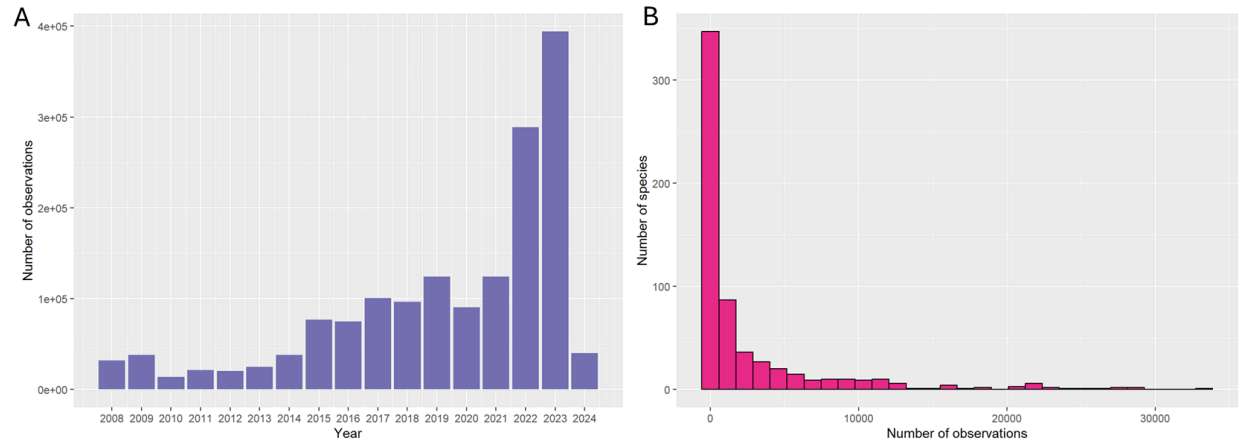

Figure 9: A: Number of observations per year for the unstructured QDGC dataset for the Western Cape. B: number of observations per species. The unstructured pentad dataset showed the same trends.

There were less observations in the unstructured data cubes for the Western Cape (1 645 858 for the pentad cube and 1 599 485 for the QDGC cube), than in the SABAP2 data, but the number of observations increased over time (Fig.9). There were more bird species recorded in the unstructured data cubes (636 species in the pentad cube and 614 species in the QDGC cube) for the Western Cape than in SABAP2 (see Section 3.2). Most of the species (171 in the pentad cube and 168 in the QDGC cube) in the unstructured data for the Western Cape were recorded between 1000 and 10 000 times. For the Western Cape, most of the species in the SABAP2 data were in the unstructured data cubes (98% for the pentad data and 97% for the QDGC data), with the missing species being mostly very rare species.

There were species missing from the SABAP2 data cubes that appeared in the unstructured data cubes, the reasons are discussed in Section 3.2. The unstructured data cubes included few observations for 2024, this was due to delays in data feeding into GBIF (see Section 3.3). For example, the unstructured data in the pentad cube for Hessequa came from nine publishers, but most observations were from Cornell Lab of Ornithology (i.e., Ebird), with the second most from iNaturalist. Ebird data dominated in all years, and appears to have not been submitted to GBIF for 2024, which explains the major decrease in records in 2024 (see Fig.7 A). These delays in submission to GBIF need to be considered when assessing temporal trends below.

3 General issues

3.1 Data is out of bounds of area selected

When generating an occurrence cube for a certain region, e.g. Belgium, it is still possible to get a cube with observations outside of this region. This is the case for observations with large `coordinateUncertaintyInMeters`.

For grid designation, a location is randomly chosen within the radius of coordinateUncertaintyInMeters. If this uncertainty is very large, it is possible that this random point is outside the region that is used in the SQL query.

After the aggregation function, in your SQL, you filter on region/country, but these observations with large uncertainty aggregated outside Belgium, still have "Belgium" in the dataset column country. Therefore it will not be excluded even if you filter on Belgium in your SQL.

There are a few possible ways to deal with this:

1. Filtering out occurrences with a high `coordinateUncertainty` before aggregating
2. Not randomly assigning coordinates to grid within their uncertainty
3. Filtering out records with high `minCoordinateUncertaintyInMeters` after aggregating

Methods 1 and 3 will result in some to a lot of data loss, depending on what cut-off is used for the filtering. Method 2 will not result in any data loss but might lead to data being aggregated in specific points. Method 1 seems preferable for making spatial indicators whereas method 2 seems preferable for making temporal indicators spanning the whole area of the cube.

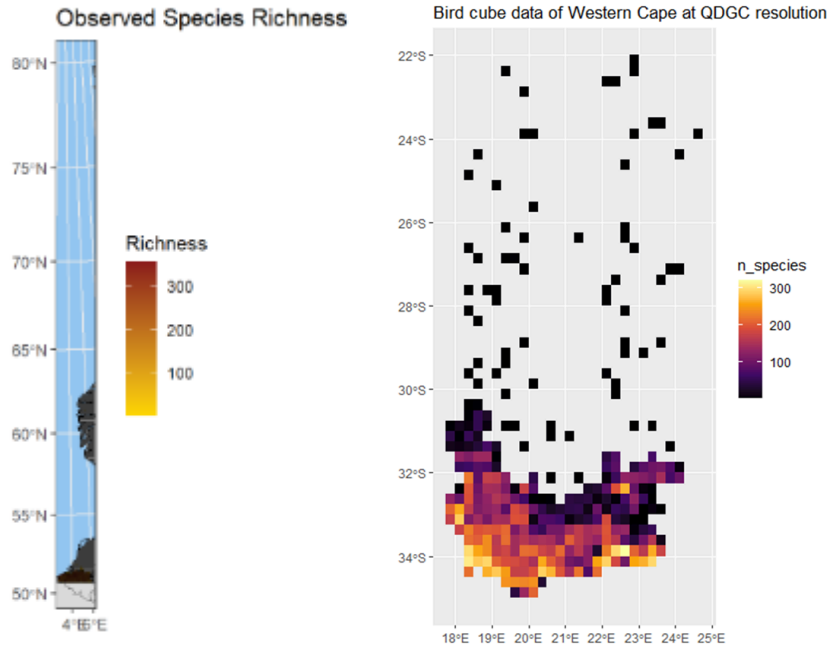

Figure 8: preliminary analyses of biodiversity indicators gave strange patterns due to observations with large coordinate uncertainties.

How to filter and what to do with coordinateUncertainty = NA

A good rule of thumb is to filter data with a coordinate uncertainty smaller than one edge of the grid used. For example, if you are using a grid of 10 km², all coordinate uncertainties should be smaller than 10 000 m. This will ensure that the randomization only ever selects a cell adjoining the one containing the coordinates (Fig.9). Using a higher cut-off will result in a big uncertainty of the actual location of the occurrence vs. the pattern one wishes to determine.

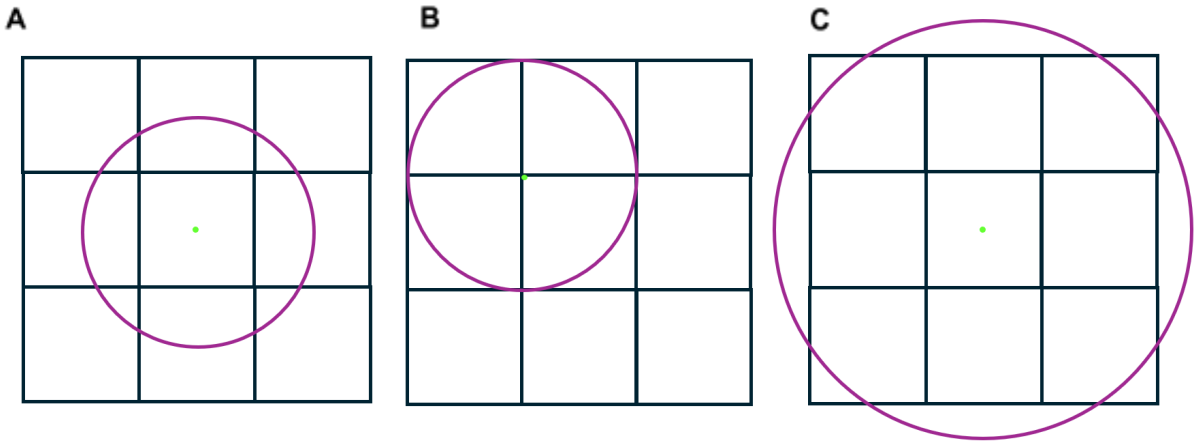

Figure 9: A: A point with coordinate uncertainty smaller than the edge of your grid will only ever result in that point being assigned to one of the neighboring grid cells, even if the point is located at the very edge (B). C: Allowing points with coordinate uncertainties greater than the edge of your grid leads to high uncertainty compared to the scale of the pattern you are looking at.

Compared to the data in the Flemish cube, the standard 1000 m of coordinate uncertainty for observations without coordinate uncertainty provided seems rather low. Following the rule of thumb stated above, if we analyse the cube on the 1 km² scale we would discard ~70% of the data but keep all data with unknown coordinate uncertainty. Our recommendation would be to first determine the scale at which to run the analysis (taken into account our rule of thumb), then set the coordinate uncertainty at the same scale of the analysis for the data with missing coordinate uncertainty.

3.2 Issues with species names

Names for the same species are not synonyms in GBIF

When comparing the structured data with the unstructured data cube for Flanders, Belgium we found that *Poecile montanus* and *Parus montanus*, *Dendrocopus major* and *Dendrocopos major* both refer to the same species. Both species names are accepted names in GBIF but they are not linked. The same issue was found for several species in the unstructured data cubes for the Western Cape and the area of the Hessequa Systematic Atlasing Subproject, South Africa - for example, there are two accepted species names in GBIF for *Tychaedon coryphoeus* (*Tychaedon coryphoeus* (Lesson, 1831) and *Erythropygia coryphoeus* (Vieillot, 1817)) and for

Dessonornis caffer (*Cossypha caffra* (Linnaeus, 1771) and *Dessonornis caffer* (Linnaeus, 1771))
Solution: an issue should be created through GBIF (our example). However, identifying such species can be difficult and time-consuming in many instances. It's important to note that GBIF's backbone is updated infrequently (usually every six months), and they manage a significant number of taxon-related issues.

Species names are wrongly translated in the publishing process
For the published ABV data we found that *Saxicola torquatus* is most likely a wrong name and needs to be replaced with *Saxicola rubicola*. For the published SABAP2 data there were species missing, and many records for which `species` was NA. Further investigation indicated that for these species there appears to be an issue when the data is published to GBIF. For example, there are many records where `species` is NA, but `genus` is “Zosterope”: a synonym according to GBIF of Zosterops Vigors & Horsfield, 1827. Yet, *Zosterops virens* (Cape white-eye) does not appear in the SABAP2 data downloaded from GBIF, despite there being many records in the Western Cape for this species on the SABAP2 website.
Solution: contact the data publisher and ask them to rectify this, although in many cases it will be difficult and time consuming to identify such species.

Species misidentifications
Some species recorded in the unstructured data cubes for the Western Cape and the area of the Hessequa Systematic Atlasing Subproject, did not occur in the SABAP2 data for these areas. In some cases, this was likely due to the misidentification of the species. For example, *Certhilauda curvirostris* (Cape Lark) is a range restricted species, which does not occur in the area of the Hessequa Systematic Atlasing Subproject, but is included in the unstructured data cubes for the area. These records could be misidentifications, as the species could be confused with *Certhilauda brevirostris* (Agulhas lark), which does occur in this area.

3.3 Delays in data being published to GBIF

There are substantial delays in data being published to GBIF. This was evident in both the structured and unstructured data cubes. For example, the period covered by the SABAP2 and unstructured data cubes was up to 31 December 2024. However, the SABAP2 data cube included data only up until February 2023, despite this dataset being regularly published to GBIF. Similarly, due to delays in data being published to GBIF, the unstructured data cube included few observations from 2024.

3.4 Mapping indicators with b3gbi

For the South African data cubes the indicators could not be mapped using the b3gbi R package (Dove 2025). Data at pentad resolution could not be mapped as this is not one of the standard grids, and there were also issues mapping the data at QDGC resolution, although this is one of the standard grids.
Solution: We will engage with the developer of the package to rectify this issue (https://github.com/b-cubed-eu/b3gbi/issues/54).

4 Range comparisons

4.1 Flanders

Given the limitations of the ABV monitoring setup we won't compare the occupancy of each square one-on-one, but rather look if a species occurs in a similar percentage of the total number of squares.

Furthermore, we explore if certain filters can help improve performance of range comparisons between the cube data and the structured ABV data. The following filters are used:

- No filter
- Filter 1: only species that were reported on in the ABV framework (Onkelinx et al. 2024)
- Filter 2: a specific set of rules to exclude non-informative squares and data deficient species (loosely based on the rules set in the ABV methodology)
    - A square is only relevant for a species if that species is observed in this square for more than one time period
    - A minimum of three relevant squares are needed to include the species
    - A minimum of a hundred observations are needed to include the species

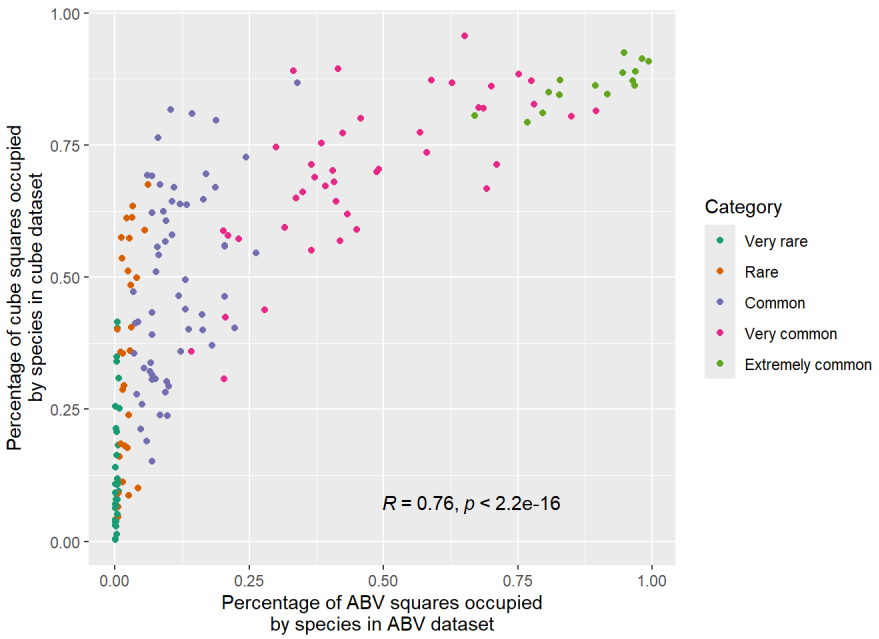

Figure 10: Correlation between the percentage of occupied ABV squares and the percentage of occupied cube squares per species.

There is a significant correlation (R = 0.76, p < 0.001) between the percentage of occupied ABV squares and the percentage of occupied cube squares (Fig.10). This correlation improves very little (R = 0.77) when using filter 1 and worsens slightly (R = 0.73) when using filter 2 (Fig.11 A and B).

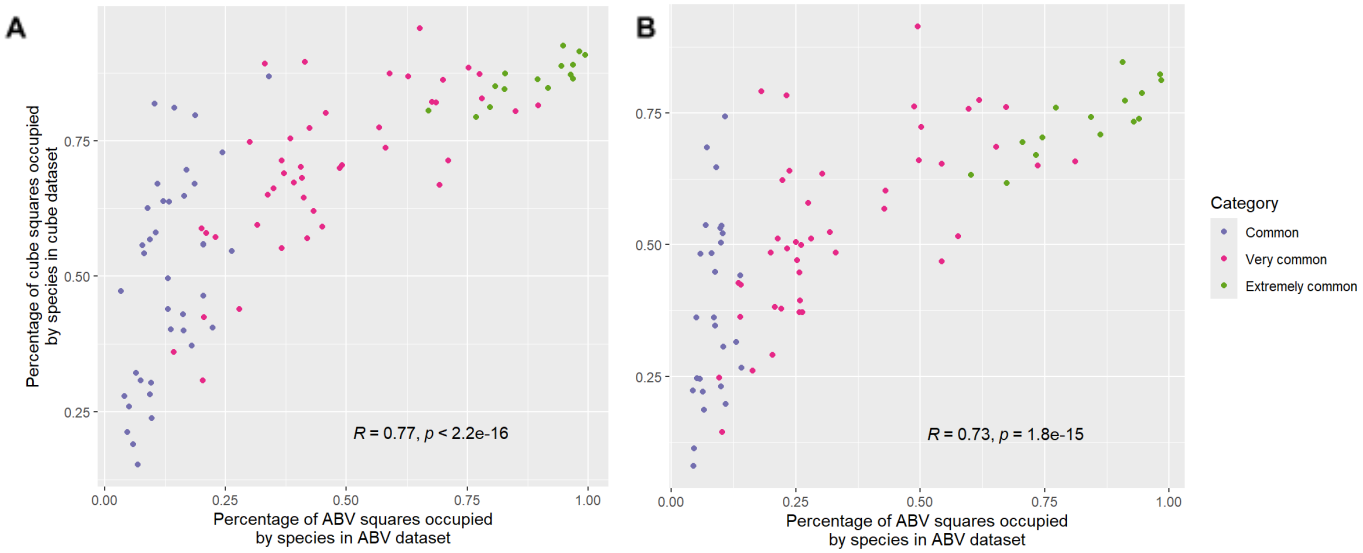

Figure 11: Correlation between the percentage of occupied ABV squares and the percentage of occupied cube squares per species. Filtered for A: only species that were analysed in the ABV framework, and B: a specific set of rules to exclude non-informative squares and data deficient species.

4.2 South Africa

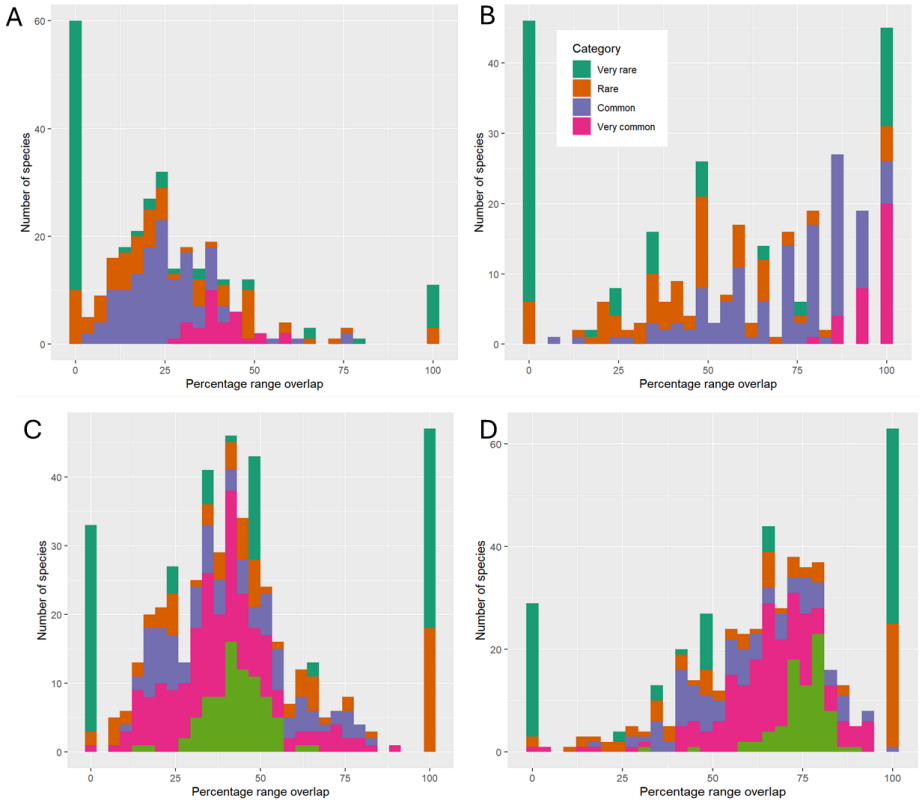

Figure 12: The percentage of range overlap - calculated as the percentage of SABAP2 grid cells in which a species has been recorded in both SABAP2 and the unstructured data - for A: Area of Hessequa Systematic Atlasing Subproject at pentad resolution, and B: QDGC resolution, and C: Western Cape at pentad resolution, and D: QDGC resolution.

On average the ranges of the species based on SABAP2 data and unstructured data overlap by 25-64%, with this percentage being higher when analysed at larger spatial extents and coarser resolutions (Fig.12). Most species for which there was 0 or 100% range overlap are very rare or rare species (Fig.12). These correlations are also calculated for two-year periods for Hessequa and three-year periods for the Western Cape - the correlations got stronger over time for both study areas.

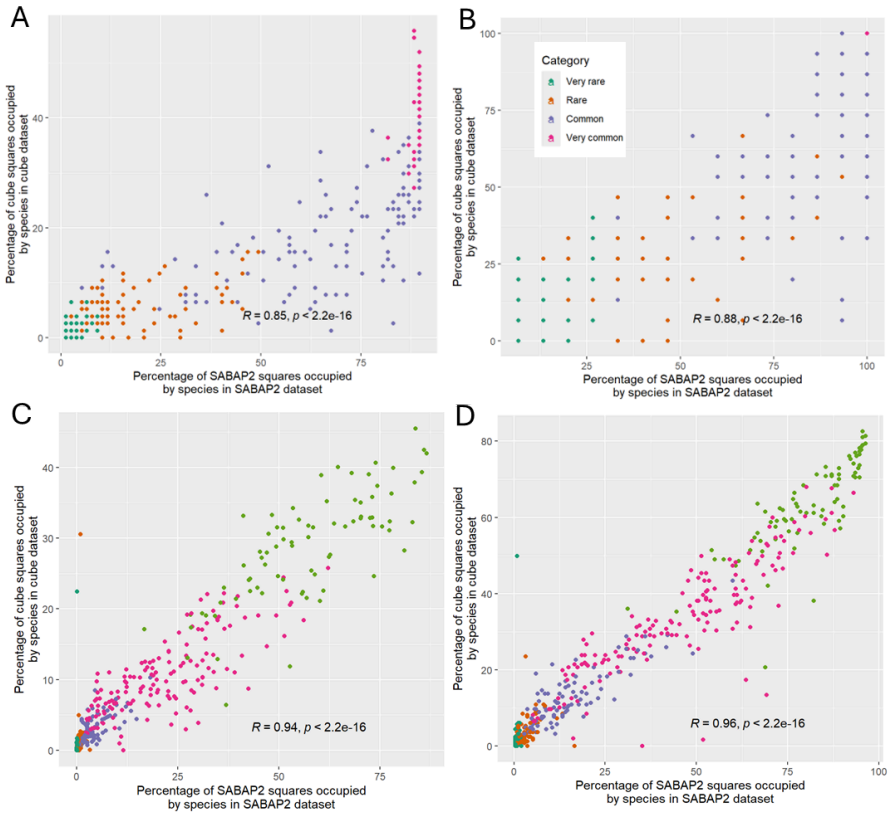

Figure 13: Correlation between the percentage of SABAP2 grid cells in which a species is observed in SABAP2 and in which it is observed in the unstructured data - for A: Area of Hessequa Systematic Atlasing Subproject at pentad resolution, and B: QDGC resolution, and C: Western Cape at pentad resolution, and D: QDGC resolution.

The percentage of grid cells in which species are observed in SABAP2 is significantly and positively correlated with the percentage of grid cells in which the species are observed in the unstructured data (Fig.13). The strength of this correlation is higher when the analysis is performed for larger spatial extents and coarser resolutions (Fig.13). This analysis was also performed by excluding very rare and rare species - for the Western Cape the correlations remained strong (R^2 = ~ 0.96), but for Hessequa they declined (R^2 = ~ 0.65). These correlations are also calculated for two-year periods for Hessequa and three-year periods for the Western Cape - the correlations got stronger over time for the Western Cape (e.g., R^2 = 0.81 for 2008-2010 to R^2 = 0.96 for 2020-2022), but not for Hessequa (e.g., e.g., R^2 ~ 0.86 or every time period).

5 Trend comparisons

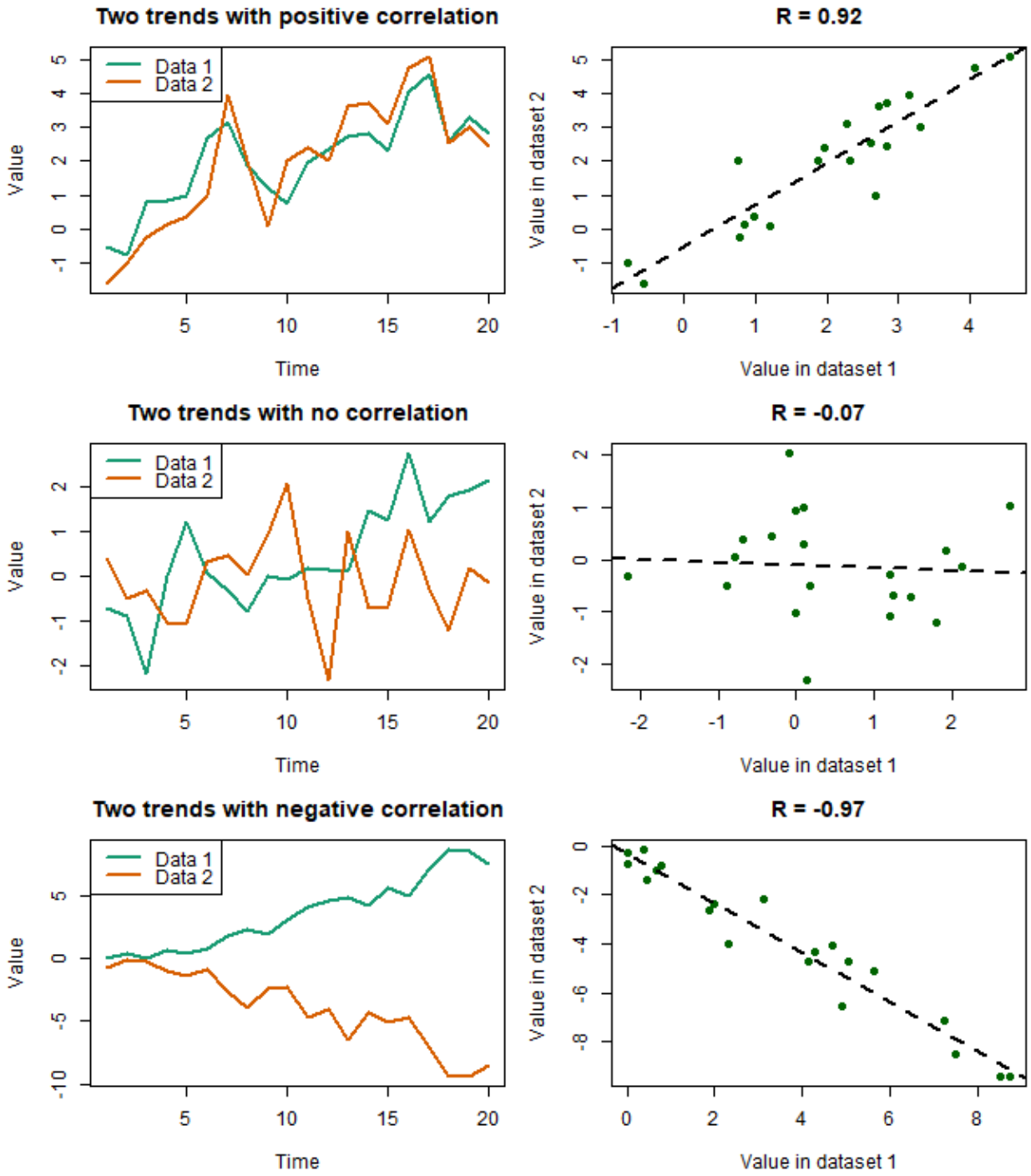

Figure 14: Comparison of time series and scatterplots for three pairs of datasets with different levels of correlation. The top row shows a strong positive correlation (R ≈ 1), the middle row shows no correlation (R ≈ 0), and the bottom row shows a strong negative correlation (R ≈ -1). Time series plots (left) illustrate the trends over time, while scatterplots (right) show the relationship between paired data points.

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 annual occurrence trend 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 (Fig.14). This approach provides a quantitative measure of agreement between the datasets at the species level.

5.1 Flanders

Due to the setup of the ABV data collection (with a subset of squares monitored in cycles of three years) it makes little sense to look at the trend per year (but the figures are included in the appendix).

So, we look at the correlation between trends over the different cycles. Here we also explore if certain filters can help improve performance of trend comparisons between the cube data and the structured ABV data. The following filters are used:

- No filter
- Filter 1: only species that were reported on in the ABV framework (Onkelinx et al. 2024)
- Filter 2: a specific set of rules to exclude non-informative squares and data deficient species (loosely based on the rules set in the ABV methodology)
    - A square is only relevant for a species if that species is observed in this square for more than one time period
    - A minimum of three relevant squares are needed to include the species
    - A minimum of a hundred observations are needed to include the species
- Filter 3: number of observations per species and time period is weighed by total number of observations for that time period
- Filter 4: a combination of filter 2 and filter 3.

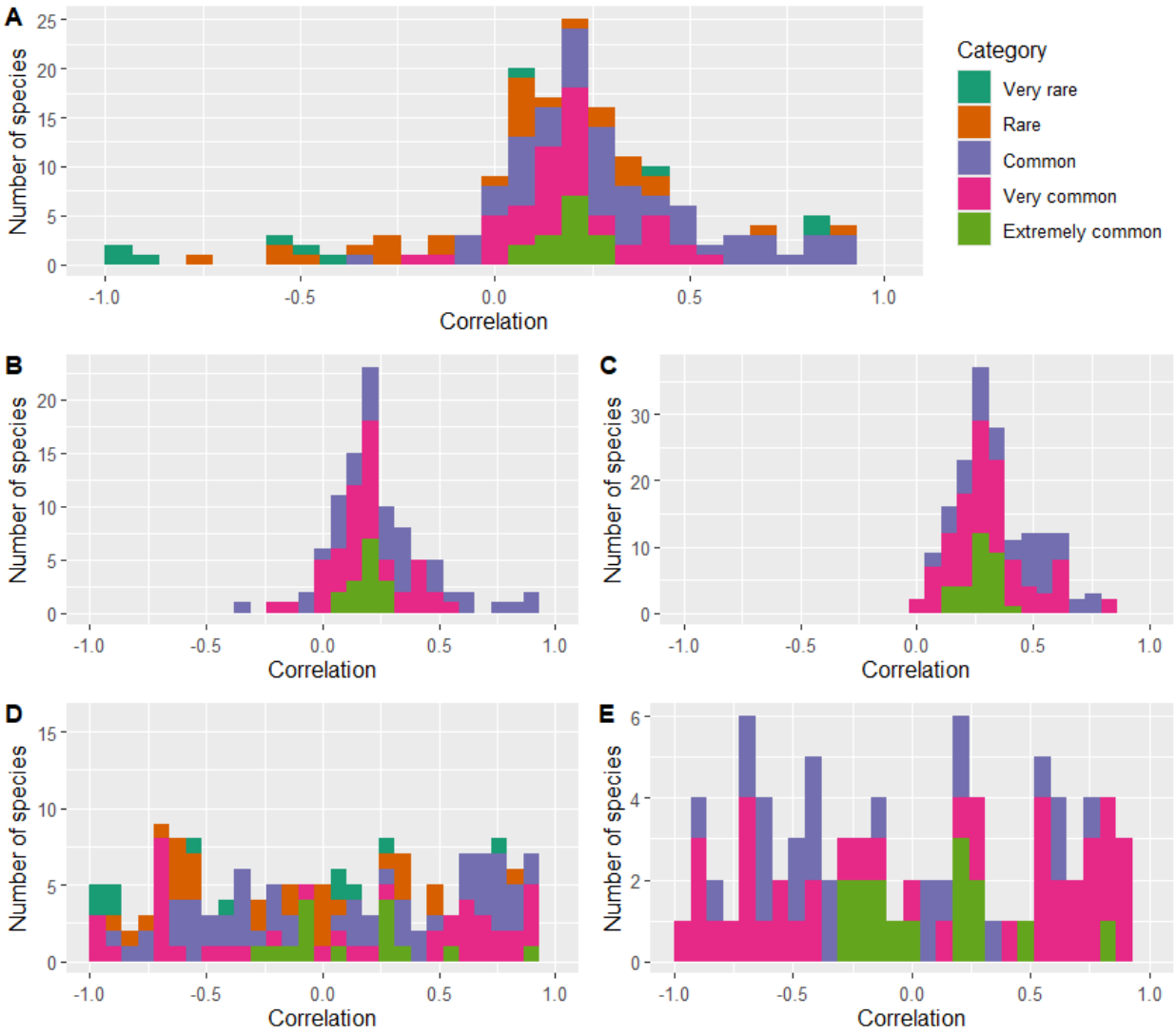

Figure 15: Correlation per species of the trend over the different cycles between the structured and unstructured data. A: Not filtered. B:Filtered for only species that were analysed in the ABV framework. C: Filtered for a specific set of rules to exclude non-informative squares and data deficient species. D: Weighed by the total number of occurrences per time period E: A combination of the specific set of rules and the weighing.

Overall we find a poor correlation between the trends in the structured and the unstructured data (Fig.15). The correlations for rare and very rare species range from -1 to almost 1 (Fig.15 A). Only the specific set of rules filter seems to increase the correlations slightly (Fig.15 C).

5.2 South Africa

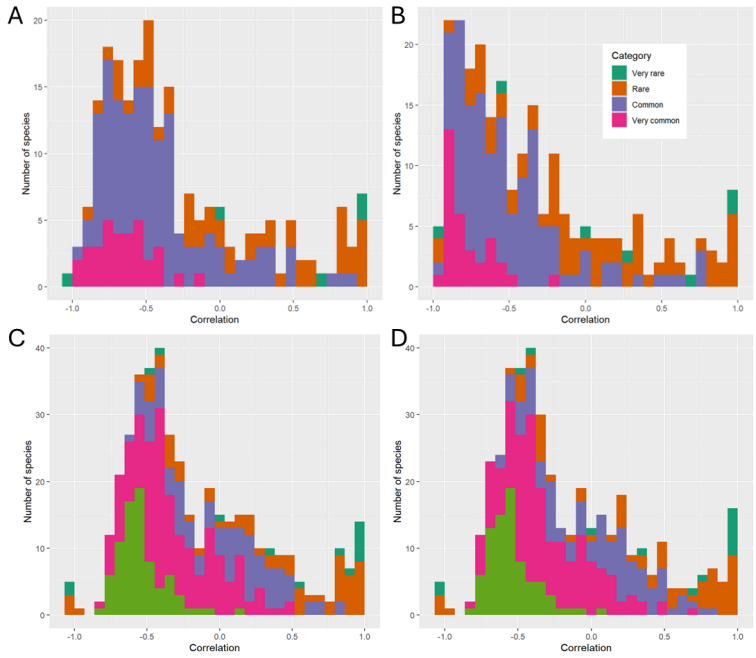

Figure 16: Correlations between yearly trends in species observations in SABAP2 and unstructured data - for A: area of Hessequa Systematic Atlasing Subproject at pentad resolution, and B: QDGC resolution, C: Western Cape at pentad resolution, and D: QDGC resolution.

There are poor, mostly negative correlations when yearly trends are compared, no matter the spatial extent or resolution of the analysis (Fig.16). The correlations improve if the number of observations for each species per year is divided by the total number of observations for that year, with this improvement evident no matter the spatial extent or resolution of analysis (Fig.17). The correlations also improve if years are grouped into time-periods, with greater improvements when the analysis is performed at larger spatial extents (Fig.18). Most species for which there are high correlations (positive and negative) are rare and very rare species. This was the case for all analyses, except when performed at a large spatial extent and coarse resolution, and years are grouped into time periods (Fig.18 D).

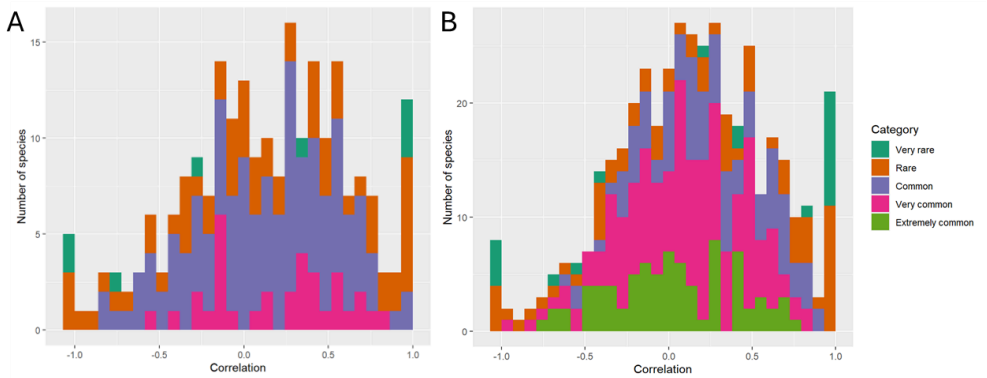

Figure 17: Correlations between yearly trends in species observations in SABAP2 and unstructured data, when the observations per species per year are divided by the total number of observations per year - for A: Area of Hessequa Systematic Atlasing Subproject at pentad resolution, and B: Western Cape at QDGC resolution. The trends are similar for the other extent-resolution combinations.

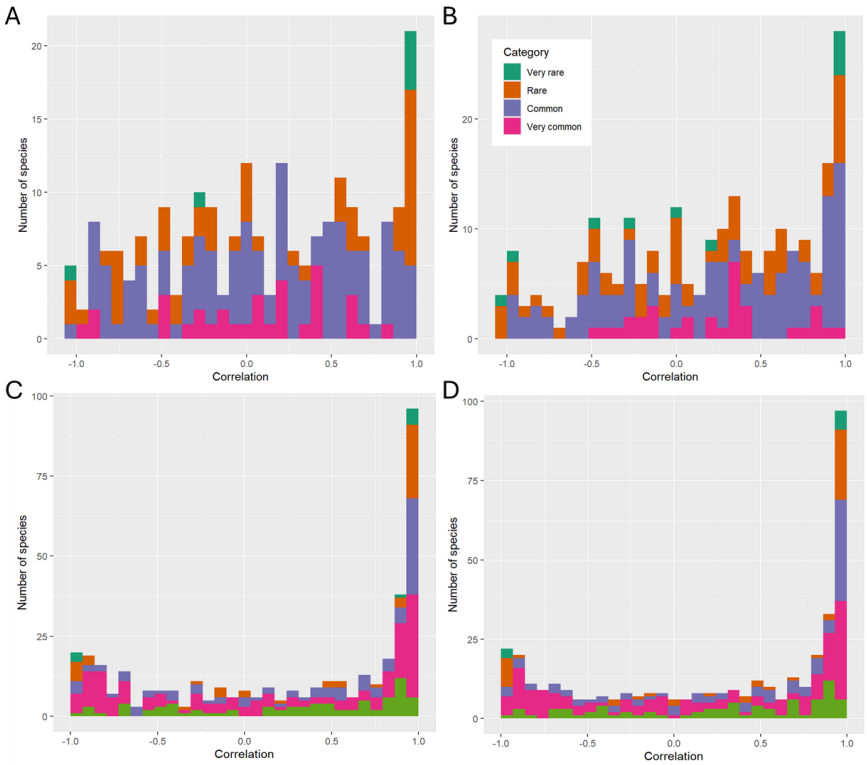

Figure 18: Correlations between trends over time periods in species observations in SABAP2 and unstructured data - for A: area of Hessequa Systematic Atlasing Subproject at pentad resolution, and B: QDGC resolution, C: Western Cape at pentad resolution, and D: QDGC resolution. Two-year time periods were used for the area of the Hessequa Systematic Atlasing Subproject and three-year periods for the Western Cape.

6 Comparing biodiversity indicators

One aspect of the B3 project focusses on delivering standardized workflows that simplify the process of calculating common biodiversity indicators from GBIF data cubes. For this the b3gbi R package is being developed (Dove 2025) which provides a one stop shop for taking occurrence cubes from GBIF and calculating several biodiversity indicators. In this section we look at these biodiversity indicators and compare the results when these are calculated from unstructured occurrence cubes from GBIF vs. structured occurrence cubes derived from standardized monitoring.

6.1 Flanders

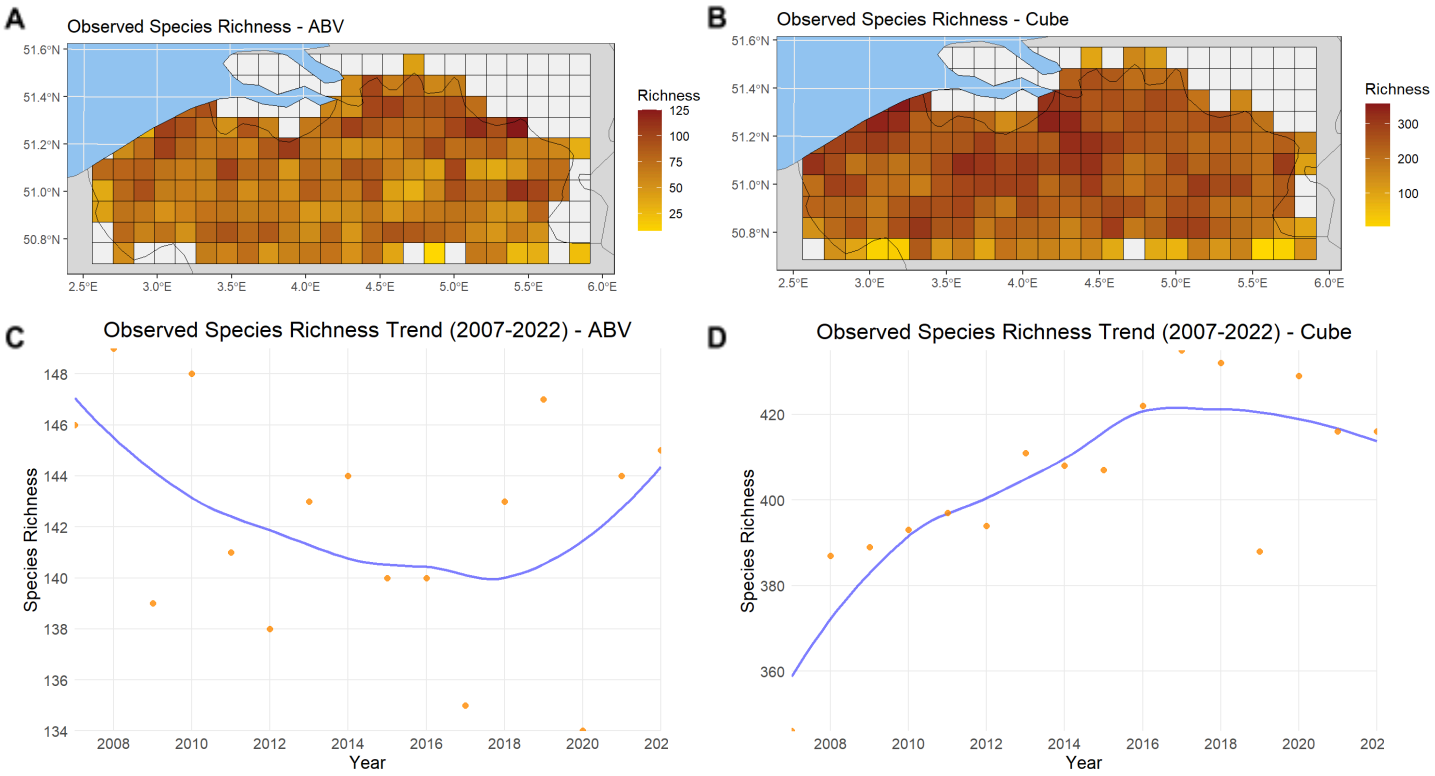

Figure 19: observed species richness per 10 km² for the structured (A) and unstructured data (B). observed species richness per year for the structured (C) and unstructured data (D).

There is no clear spatial pattern in species richness for either the structured or unstructured data (Fig.19). This is in line with expectations for Flanders as it is a small, densely populated area. There is no clear temporal trend for the observed species richness for the structured data, while there is a clearly increasing species richness for the unstructured data (Fig.19 D). This is most likely due to the increase in total number of occurrences over the years (Fig.4 A).

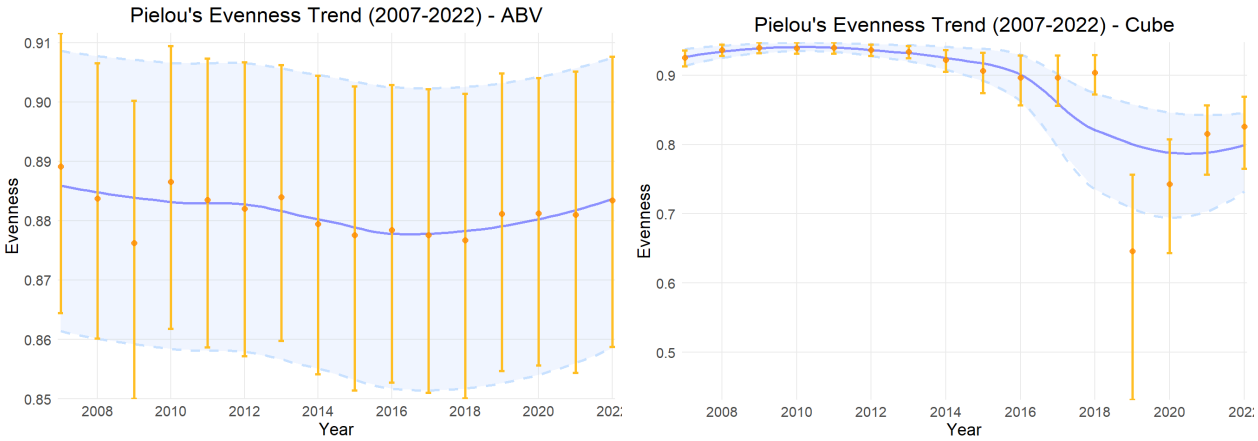

Figure 20: Pielou's evenness over the years for the structured (ABV) and unstructured (cube) data.

When looking at another biodiversity indicator, i.e. Pielou's Evenness, we see a very strong influence of the datasets used in constructing the cube (Fig.20). 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 diminishes due to the increased relative importance of datasets focused on a small subset of species from research projects and the like. 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 types of data.

When looking at species specific indicators for three different species we see that the unstructured data yield highly similar trend shapes for all three species, regardless of their true population dynamics (Fig.21), suggesting a systematic bias due to uncorrected sampling effort bias. This highlights the necessity of applying correction methods to account for variation in sampling effort when using unstructured data sources in biodiversity monitoring.

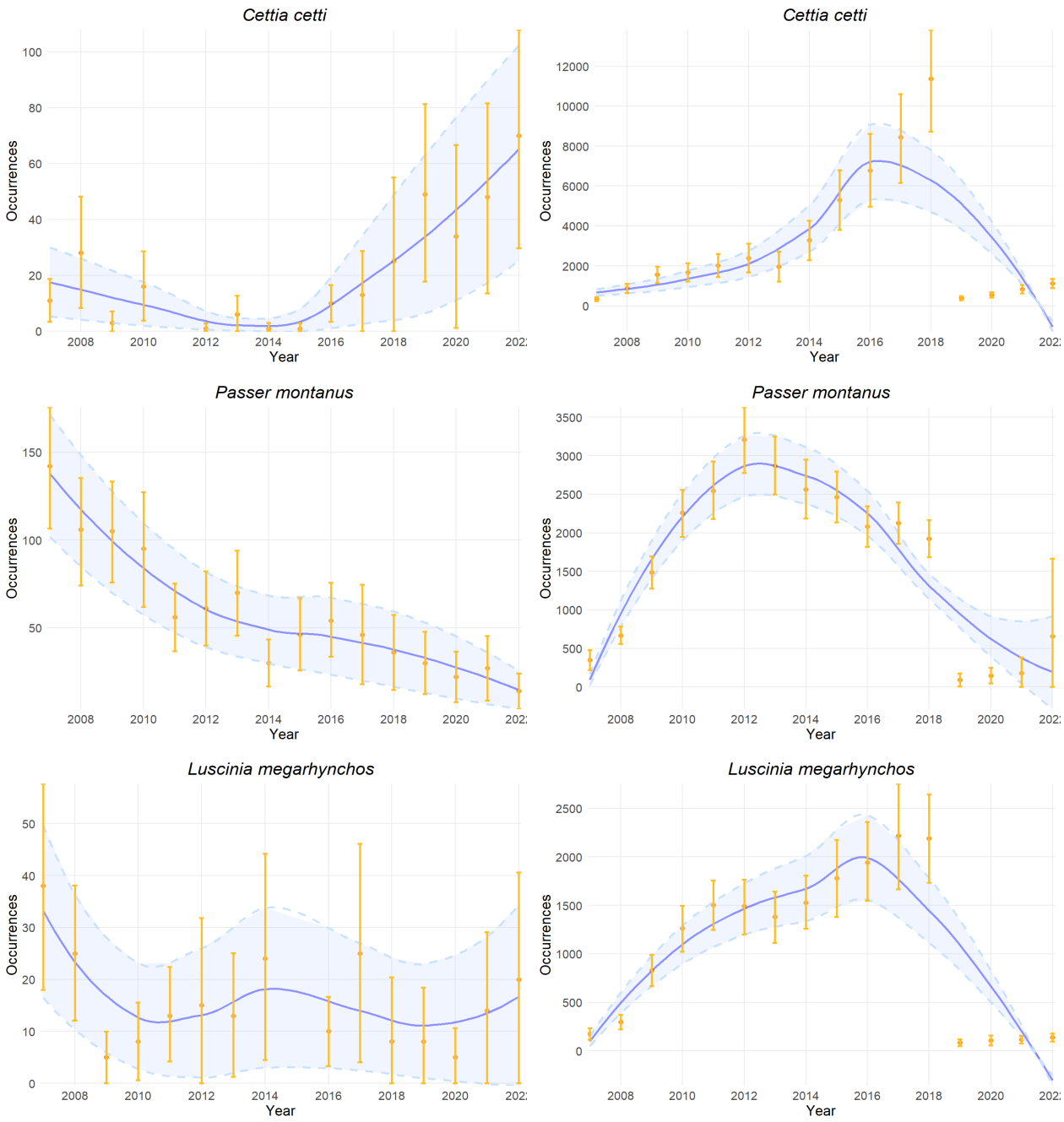

Figure 21: The species specific indicator on the number of occurrences for: the species with the biggest increase according to the ABV: Cetti's warbler (*Cettia cetti*), the species with the biggest decrease according to the ABV: Eurasian tree sparrow (*Passer montanus*) and a species with a very stable trend according to the ABV: Common nightingale (*Luscinia megarhynchos*). Species specific indicator calculated for the structured data (left) and for the unstructured data (right).

6.2 South Africa

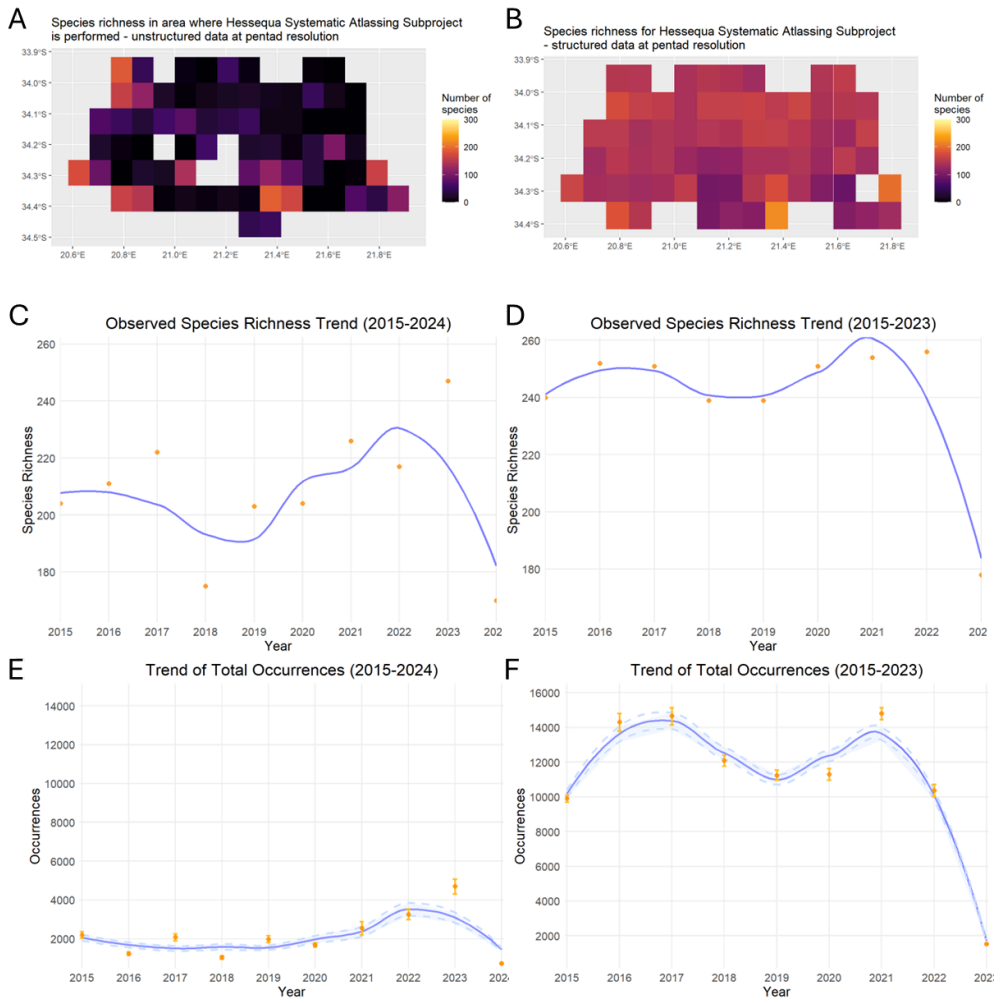

Figure 22: Spatial and temporal trends in species richness and number of observations for the area of the Hessequa Systematic Atlasing Subproject at pentad resolution, based on unstructured data (A, C and E): and structured data (B, D and F). Similar trends were evident for this area at QDGC resolution.

The spatial and temporal trends in species richness for Hessequa differ between the unstructured and structured datasets (Fig.22). There is less spatial and temporal variation in species richness based on the structured data (spatial: range of 89-221 species with standard deviation of 22.2; temporal: range ~240-260), than the unstructured data (spatial: range of 1-193 species with standard deviation of 48.3; temporal: range ~175-250), and species richness is generally higher in the structured data (Fig.22). These trends are evident no matter the resolution of the data (pentad and QDGC). Species richness increases over time based on the unstructured data, this reflects an increase over time in the unstructured data in the number of observations (Fig.22). In contrast, species richness does not increase over time in the structured data, but the trend reflects temporal changes to the number of observations (Figure 22), which is aligned with temporal trends in the number of surveys in the area.

Figure 23: Spatial trends in species richness for the Western Cape at pentad resolution for A: unstructured data and B: structured data; and at QDGC resolution for C: unstructured data and D: structured data.

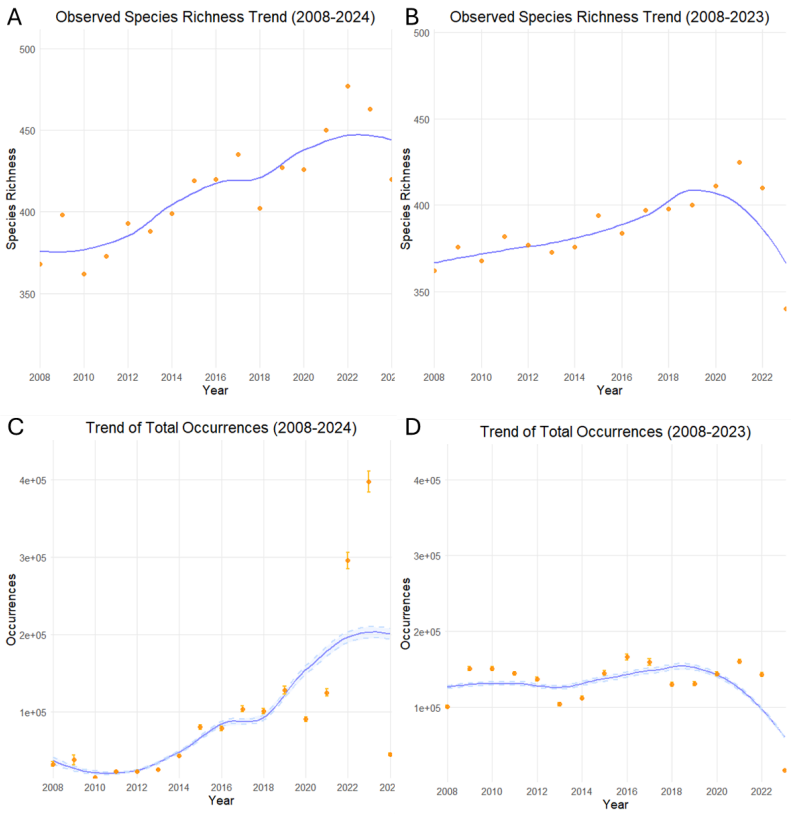

Figure 24: Temporal trends in species richness and number of observations for the Western Cape for unstructured data (A and C) and structured data (B and D).

For the Western Cape, there is less spatial variation in species richness based on the structured data than the unstructured data, but the patterns are more similar when analysed at a coarser resolution (Fig.23). No matter the resolution of analysis (pentad or QDGC) in the unstructured data species richness increases over time, with this increase reflecting an increase in the number of observations (Fig.24). In the structured data there is also a slight increase in species richness, and while the number of observations is relatively stable (Fig.24) there was an increase in the number of surveys over time (not shown).

At all resolutions and spatial extents of the analysis and in both the structured and unstructured data there are a few grid cells that had a high number of observations, and these were generally where many people are found. Similar to species richness, there is more spatial variation in number of observations in the unstructured data than the structured data, with patterns becoming more similar at larger extents and resolutions of analysis.

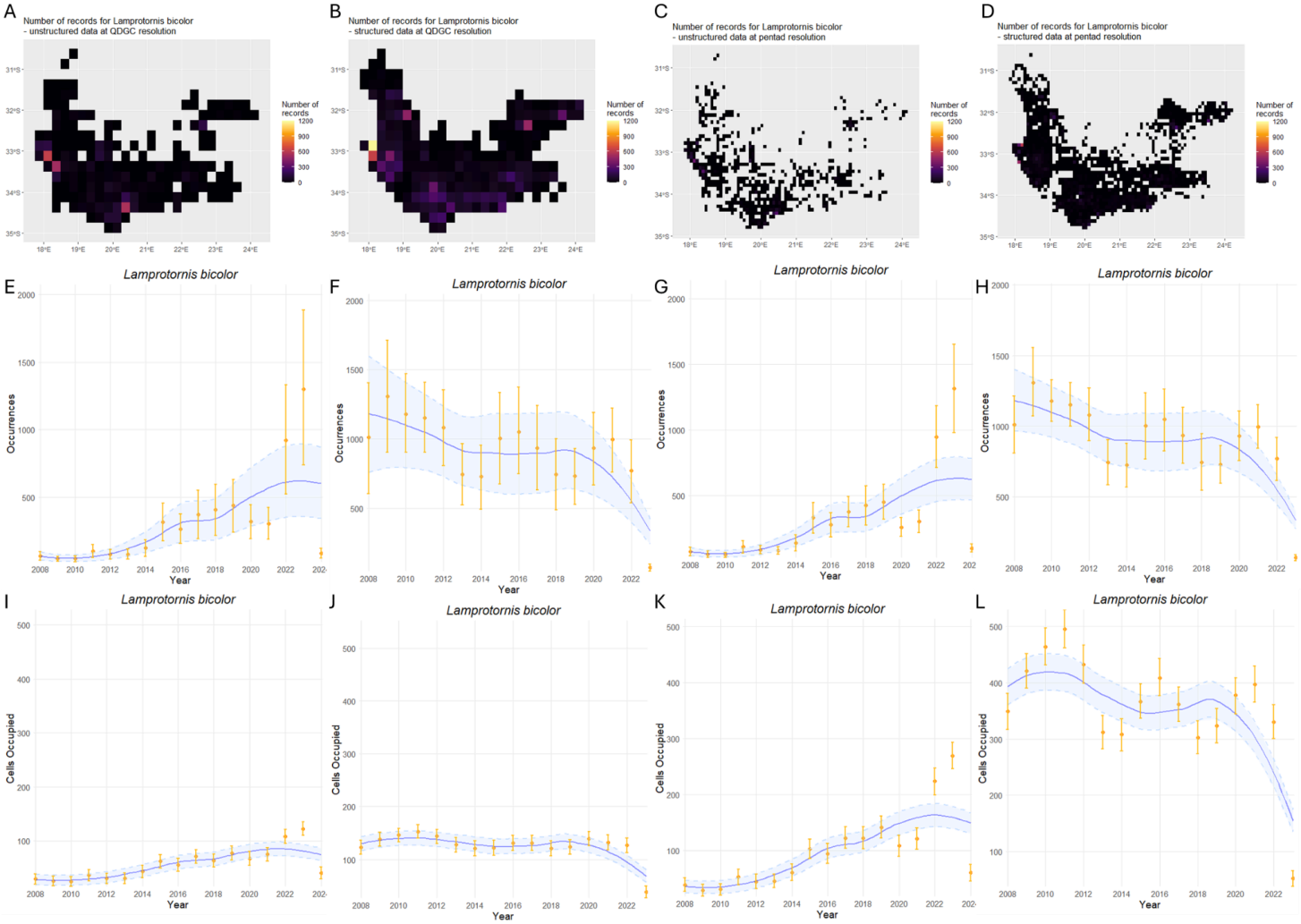

Figure 25: Temporal and spatial variation in number of records and cells occupied in the Western Cape for the species *Lamprotornis bicolor* (African Pied Starling) based on: unstructured data at QDGC resolution (A, E, I), structured data at QDGC resolution (B, F, J), unstructured data at pentad resolution (C,G,K), and structured data at pentad resolution (D,H,L).

The species specific indicators show that when unstructured data are analysed at high resolutions, the spatial distribution and number of cells that are occupied are likely to be underestimated (Fig.25). Analysing the data at coarser resolutions makes the spatial distribution and number of cells occupied more similar to what is seen in the structured data. In terms of number of occurrences, the spatial and temporal patterns are different between the structured and unstructured data, no matter the extent or resolution of the analysis. These trends are evident in a number of the species for which these analyses were performed.

7 Discussion

Unstructured data, while valuable, is prone to taxonomic and geographic biases, detectability issues, non-random sampling effort, false positives, and changes in sampling effort over time, all of which can impact the accuracy and reliability of species status and trend estimations. Our investigation into the use of unstructured biodiversity data from the Global Biodiversity Information Facility (GBIF) for estimating species status and trends reveals significant challenges.

A first layer of issues are related to technical issues and data quality. Observations are sometimes found outside the selected region due to large coordinate uncertainty and subsequent randomization, necessitating the implementation of filtering methods, though these might lead to data loss. Species names can present inconsistencies, including non-synonymous names for the same species, translation errors during publication, and misidentifications, all requiring careful quality control. Furthermore, delays in data publication to GBIF are evident, affecting the assessment of recent temporal trends. Finally, technical issues are encountered with mapping indicators using the b3gbi R package (Dove 2025), particularly for non-standard grid resolutions.

In this study we use two structured datasets, both the results of bird monitoring programs. Birds are overrepresented on GBIF compared to other taxonomic groups (Troudet et al. 2017), potentially skewing results and limiting the generalizability of the findings. By comparing the structured data with the unstructured data, we find that unstructured data are less reliable for capturing ranges and trends of rare or very rare species, often leading to poor range overlap and trend correlation with structured data. This could be attributed to several factors such as underreporting, misidentification, and lack of standardized sampling effort in unstructured data. For less popular species groups this problem will likely be even bigger.

Despite these limitations, the study also reveals some positive aspects of unstructured data. Range correlations between structured and unstructured data generally increase over time, suggesting improved data coverage and quality in GBIF. This trend is encouraging and highlights the potential of unstructured data to complement structured data for biodiversity monitoring, especially as data availability and quality continue to improve. However, it is crucial to acknowledge that the heterogeneity and potential biases in unstructured data necessitate careful data filtering, validation, and interpretation to ensure robust and meaningful results.

In preparation of the final deliverable it would be useful to try and correct SABAP2 data based on the number of surveys, and then look at trends and indicator comparisons. We would also like to explore bringing in estimates of sampling effort such as family counts or indicators on sample completeness per year or per grid cell. Another thing to explore would be a way to weigh the number of occurrences based on the dataset they originate from, this way not only looking at the total number of occurrences but also the specificity of certain datasets.

8 Preliminary recommendations

- For determining ranges, analyses at larger spatial scales and coarser resolutions are recommended.
- For assessing trends, it is advised to either assess the trends across time-periods, rather than years, or weighing the number of observations per species, e.g. by the total number of observations.
- For spatial trends in species richness, analyses at larger scales and coarser resolutions are recommended.
- Filter data with a coordinate uncertainty smaller than the edge of the grid used. For example, if using a grid of 10km x 10km, all coordinate uncertainties should be smaller than 10,000 m.
- For data with missing coordinate uncertainty, determine the analysis scale, taking into account the rule of thumb regarding coordinate uncertainty, then set the coordinate uncertainty at the same scale of the analysis.
- Always perform quality control before starting analysis, looking for unforeseen species names, specific drop-offs in data availability and other unexpected patterns.

9 Acknowledgements

We would like to thank all members of the B3 consortium for the interesting feedback on our presentation at the annual consortium meeting and for answering our questions.
We would like to thank Thierry Onkelinx, Dimitri Brosens and Stijn Cooleman at INBO for answering our questions about the ABV data and making slight corrections when we found issues in the published data.
We would like to thank Damiano Oldoni for answering many of our technical questions regarding the species occurrence cubes and for reviewing this document.

AI tools were utilized to enhance the writing and assist with portions of the data analysis presented in this document.

10 References

Blissett, Matthew, Morten Høfft, John Waller, Andrew Rodrigues, Daniel Noesgaard, Robertson, Tim, and Peter Desmet. 2025. 'Occurrence Cube Service. B3 Project Deliverable D2.3.'

Boakes, Elizabeth H., Philip JK McGowan, Richard A. Fuller, Ding Chang-qing, Natalie E. Clark, Kim O'Connor, and Georgina M. Mace. 2010. 'Distorted Views of Biodiversity: Spatial and Temporal Bias in Species Occurrence Data'. *PLoS Biology* 8 (6): e1000385.

Brooks, Michael, Sanjo Rose, Res Altwegg, Alan TK Lee, Henk Nel, Ulf Ottosson, Ernst Retief, Chevonne Reynolds, Peter G. Ryan, and Sidney Shema. 2022. 'The African Bird Atlas Project: A Description of the Project and BirdMap Data-Collection Protocol'. *Ostrich* 93 (4): 223–32.

Chamberlain, Scott, Damiano Oldoni, and John Waller. 2022. 'Rgbif: Interface to the Global Biodiversity Information Facility API'.

Desmet, Peter, Damiano Oldoni, Matthew Blissett, and Tim Robertson. 2023. 'Specification for Species Occurrence Cubes and Their Production: B-Cubed Project Deliverable D2. 1'.

Dove, Shawn. 2025. 'B3gbi: General Biodiversity Indicators for Biodiversity Data Cubes'. Manual. https://github.com/b-cubed-eu/b3gbi.

García-Roselló, Emilio, Jacinto González-Dacosta, and Jorge M. Lobo. 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.

GBIF.Org User. 2025a. 'Occurrence Download'. Text/tab-separated-values,application/zip. The Global Biodiversity Information Facility. https://doi.org/10.15468/DL.DDZHRC.
———. 2025b. 'Occurrence Download'. Text/tab-separated-values,application/zip. The Global Biodiversity Information Facility. https://doi.org/10.15468/DL.MN4YBB.
———. 2025c. 'Occurrence Download'. Darwin Core Archive,text/tab-separated-values,application/zip. The Global Biodiversity Information Facility. https://doi.org/10.15468/DL.95KUJB.
———. 2025d. 'Occurrence Download'. Text/tab-separated-values,application/zip. The Global Biodiversity Information Facility. https://doi.org/10.15468/DL.PC226A.
———. 2025e. 'Occurrence Download'. Darwin Core Archive,text/tab-separated-values,application/zip. The Global Biodiversity Information Facility. https://doi.org/10.15468/DL.4Q36E5.
———. 2025f. 'Occurrence Download'. Darwin Core Archive,text/tab-separated-values,application/zip. The Global Biodiversity Information Facility. https://doi.org/10.15468/DL.BD4VRP.

Onkelinx, Thierry, Olivier Dochy, Glenn Vermeersch, and Koen Devos. 2024. Techreport. INBO Brussel, Herman Teirlinckgebouw, Havenlaan 88 bus 73, 1000 Brussel: Instituut voor Natuur- en Bosonderzoek. https://doi.org/10.21436/inbor.102669823.

Reddy, Sushma, and Liliana M. Dávalos. 2003. 'Geographical Sampling Bias and Its Implications for Conservation Priorities in Africa'. *Journal of Biogeography* 30 (11): 1719–27.

Rooyen, Johan van, and Les Underhill. 2020. 'Systematic Atlasing in Hessequa–Report on the First Cycle of Seasonal Monitoring: Systematic Atlasing in Hessequa'. *Biodiversity Observations* 11:11.3: 1-14.

Trimble, Morgan J., and Rudi J. van Aarde. 2012. 'Geographical and Taxonomic Biases in Research on Biodiversity in Human-modified Landscapes'. *Ecosphere* 3 (12): 1–16.

Troudet, Julien, Philippe Grandcolas, Amandine Blin, Régine Vignes-Lebbe, and Frédéric Legendre. 2017. 'Taxonomic Bias in Biodiversity Data and Societal Preferences'. *Scientific Reports* 7 (1): 9132. https://doi.org/10.1038/s41598-017-09084-6.

Van Rooyen, Johan Albertus. 2018. 'Systematic Atlasing in Hessequa-Moving from Mapping to Monitoring'. *Biodiversity Observations* 9:9.10: 1-13.

11 Annex

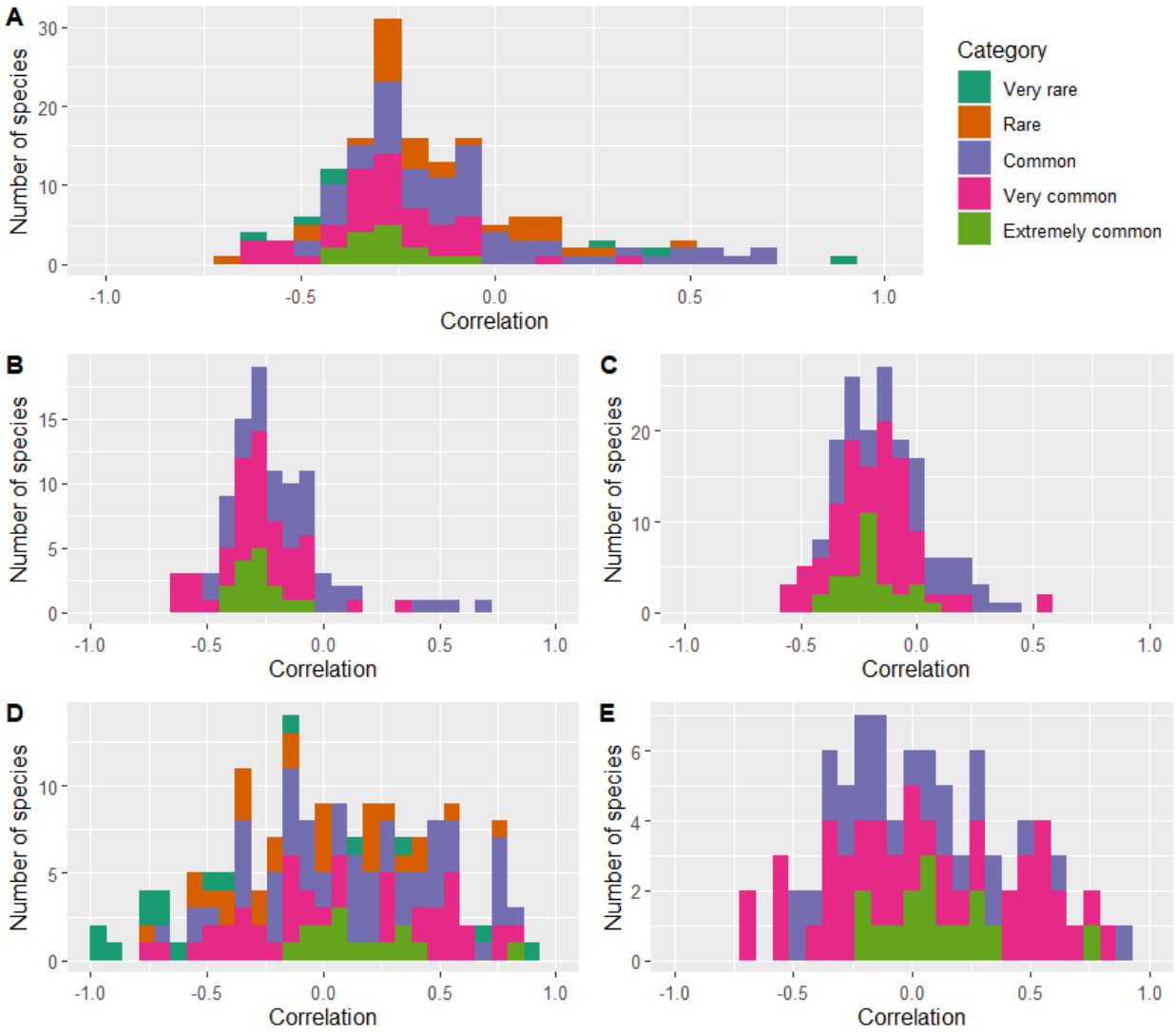

Figure A1: Correlation per species of the trend over the different years between the structured and unstructured data. A) Not filtered, filtered for B) only species that were analysed in the ABV framework, C) a specific set of rules to exclude non-informative squares and data deficient species, D) weighed by the total number of occurrences per time period and E) a combination of the specific set of rules and the weighing.