BioHackrXiv

An analysis of sex ratios using a biodiversity data cube

Quentin Groom[1] **and Maarten Trekels**[1]

1 Meise Botanic Garden, Nieuwelaan 38, 1860 Meise, Belgium

BioHackathon series:
Hacking Biodiversity Data Cubes for Policy
Brussels, Belgium
Project 12

Submitted: 22 Jul 2024

Published by BioHackrXiv.org

Abstract

This investigation uses biodiversity data cubes derived from the datasets mobilised by the Global Biodiversity Information Facility (GBIF), to conduct an analysis of sex ratios of ducks across Europe. Encompassing over 4 million occurrences extracted from nearly 5000 datasets, this study elucidates sex distribution patterns across various species, focussing on temporal and spatial dynamics. The aim of this study is to highlight the availability of open sex data and its potential usefulness in research and monitoring of sex ratios of wild organisms, particularly in sexual dimorphic species.

Introduction

The balance of sex ratios in animal populations is a critical indicator of ecological health and evolutionary dynamics (Payevsky, 2021). Alterations in these ratios can reflect environmental stresses, reproductive strategies, or human influence (Fryxell et al., 2015; Milner et al., 2007; Székely, 2023). Biodiversity data cubes, particularly those derived from the diverse data mobilised by GBIF, provide a pathway to investigate various biological phenomena (Oldoni et al., 2020). These data may originate from diverse sources, including citizen science projects, ringing(banding) activities, breeding bird counts, and hunting bag records. This variety introduces a degree of heterogeneity not typically encountered in more specialised surveys, yet it potentially enables comprehensive analyses across broader temporal and spatial scales.

This study leverages a data cube to examine sex ratio dynamics within the European ducks, aiming to uncover patterns that could inform conservation efforts and ecological understanding. However, the results presented here are only intended to draw attention to the availability of such data and the use of data cubes to analyse them. Therefore, the analysis gives readers an overview of the characteristics of the available data.

The family of ducks, geese, and swans (Anatidae) was chosen as an exemplar group because it contains species that are both sexually monomorphic (e.g. *Branta canadensis*) and dimorphic (e.g. *Somateria mollissima*). In sexually dimorphic ducks, sex is easy to determine, even at a distance, for a human observer, whereas sexually monomorphic Anatidae would usually have to be captured to determine their sex reliably. There are many other examples of sexually dimorphic vertebrates that could be analysed from data mobilised from GBIF, including many other species of bird, deer, fish, primates and pinniped. There are also many sexually dimorphic insects, including some Odonata, Coleoptera and Lepidoptera. And, although plants are not strongly sexually dimorphic there are many dioecious species, which can be sexed when flowering, examples include members of the genera *Asparagus*, *Cycas*, *Diospyros*, *Ginkgo*, *Juniperus* and *Salix*.

Here we describe a Jupyter notebook written to analyse sex from a data cube. This script is openly licensed and could be repurposed to explore the availability of sex data on GBIF for any taxa. Though for a more serious examination of sex ratios of organisms the users may wish to write their own code and also examine the datasets that contributed to this data cube more critically.

Materials and Methods

The data cube was generated from the SQL query run on the 2nd April 2024 detailed below. It extracted 4,038,527 aggregated rows from GBIF (GBIF.Org User, 2024). The data are restricted to GBIF family key 2986 that is the code for the Anatidae, the family that includes ducks, geese, and swans. The records were aggregated to a 10 km^2 grid using the European Environmental Agency reference grid system. Aggregated occurrences were also restricted to those after 1900 and to the continent of Europe. Records already identified within the GBIF infrastructure as invalid were excluded. The resulting dataset aggregates data from 4,985 published datasets on GBIF, involving 230 publishers. https://www.gbif.org/occurrence/download/0083528-240321170329656

> SELECT "year", gbif_eeargCode(10000, decimalLatitude, decimalLongitude, COALESCE(coordinateUncertaintyInMeters, 10000)) AS eeaCellCode, speciesKey, COUNT(*) AS 'count', SUM(CASE WHEN sex = 'FEMALE' THEN 1 ELSE 0 END) AS female_count, SUM(CASE WHEN sex = 'MALE' THEN 1 ELSE 0 END) AS male_count, SUM(CASE WHEN sex = 'HERMERMAPHRODITE' THEN 1 ELSE 0 END) AS hermaphrodite_count, MIN(COALESCE(coordinateUncertaintyInMeters, 10000)) AS minCoordinateUncertaintyInMeters FROM occurrence WHERE occurrenceStatus = 'PRESENT' AND familyKey = 2986 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 (identificationVerificationStatus IS NULL OR NOT (LOWER(identificationVerificationStatus) LIKE '%unverified%' OR LOWER(identificationVerificationStatus) LIKE '%unvalidated%' OR LOWER(identificationVerificationStatus) LIKE '%not able to validate%' OR LOWER(identificationVerificationStatus) LIKE '%control could not be conclusive due to insufficient knowledge%' OR LOWER(identificationVerificationStatus) LIKE '%unconfirmed%' OR LOWER(identificationVerificationStatus) LIKE '%unconfirmed - not reviewed%' OR LOWER(identificationVerificationStatus) LIKE '%validation requested%')) AND "year" >= 1900 AND continent = 'EUROPE' AND hasCoordinate GROUP BY "year", eeaCellCode, speciesKey ORDER BY "year" DESC, eeaCellCode ASC, speciesKey ASC;

We strategically excluded any data in the Darwin Core field `dwc:individualCount` from our analysis to simplify the initial approach, we aim to provide a foundational understanding of sex ratio variations and how they could be used to monitor population status and trends of biodiversity. However, a more focused study may wish to examine whether it is useful to consider `dwc:individualCount` in the aggregation step.

Shape files of the European borders were sourced from Natural Earth naturalearthdata.com.

The Jupyter notebook code used for the analysis has been archived on Zenodo (Groom & Trekels, 2024).

Our analytical framework is predicated on a selective extraction from the GBIF dataset, focusing on records designated as "PRESENT" while excluding data compromised by spatial inaccuracies. The analysis was facilitated by Python's scientific stack, including pandas for data manipulation, Matplotlib and seaborn for visualisation, and GeoPandas with Shapely for spatial analysis. This study processes the derived biodiversity data cube, augmenting it with necessary spatial and temporal attributes, and preparing it for analysis. Kriging was conducted using PyKrige with a hole-effect variogram_model (Müller et al., 2023). Versions of Python packages are detailed in table 1. Writing of the code was significantly assisted by the use of ChatGTP 4.

Table 1. Versions of Python packages used

Package	version	Website
geopandas	0.14.3	https://geopandas.org/en/stable/
matplotlib	3.7.1	https://matplotlib.org/
numpy	1.24.3	https://numpy.org/
pandas	2.0.3	https://pandas.pydata.org/docs/index.html
requests	2.31.0	https://pypi.org/project/requests/
seaborn	0.12.2	https://seaborn.pydata.org/

Results

The amount of sex data is likely to be higher for clearly sexually dimorphic species because it is easier to collect. To test this, the script extracts species for which the total number of records is over 10,000 to provide a substantial sample upon which to calculate proportions. The proportion of male plus female records was compared with the total number of all records for both sexually monomorphic (23) and dimorphic species (31) in this subsample. For monomorphic species there is an average of 5.5 records with a recorded sex per 1000 records total, whereas for dimorphic species that average is 138.7, that is 25 times more.

Most dimorphic species have a higher proportion of males, in some cases well over two males for each female (Table 2). However, two species *Mergellus albellus* and *Somateria mollissima* have more females than males.

Table 2. The proportion of males of the most widespread sexually dimorphic ducks in Europe.

Species	Vernacular name	proportion of males
Anas platyrhynchos L., 1758	Mallard	1.54
Aythya fuligula (L., 1758)	Tufted Duck	1.39
Anas crecca L., 1758	Eurasian Teal	1.54
Mareca strepera (L., 1758)	Gadwall	2.29
Bucephala clangula (L., 1758)	Common Goldeneye	1.12
Mareca penelope (L., 1758)	Eurasian Wigeon	2.10
Spatula clypeata (L., 1758)	Northern Shoveler	2.52
Somateria mollissima (L., 1758)	Common Eider	0.53
Aythya ferina (L., 1758)	Common Pochard	2.14
Mergus merganser L., 1758	Common Merganser	1.41

Since the turn of the millennium the volume of records with sex information has increased considerably (Fig. 1). The proportion of male birds is larger in recent decades, and generally there are more males than females, at least for the decades after the 1970s (Fig. 2). Exceptionally, in the 1950s there were considerably more females than males. This can be attributed to two datasets, both of which related to ringing and recovery data (Inventaire National du Patrimoine Naturel, 2020; van der Jeugd, H.P., 2022).

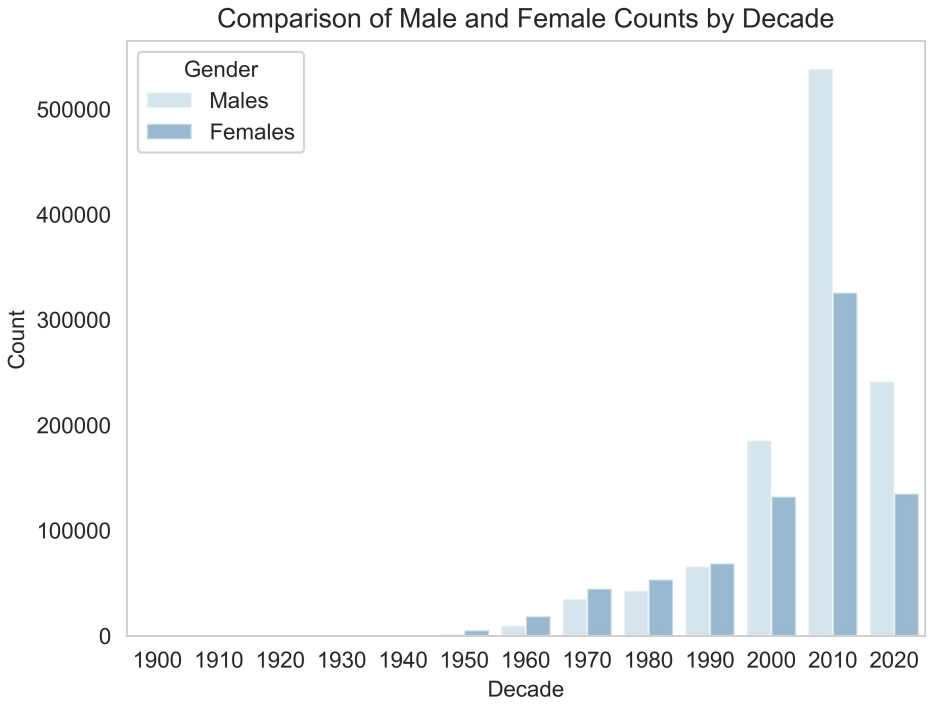

Figure 1: The total number of male and female records of *Anas platyrhynchos* in Europe aggregate to decade

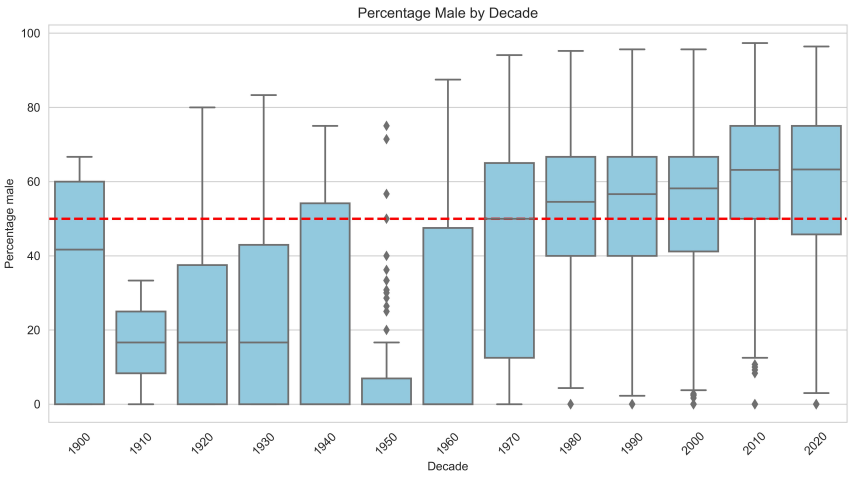

Figure 2: A time series of of the percentage of male *Anas platyrhynchos* in Europe aggregate to decade

Anas platyrhynchos is widely distributed across Europe, however, records that hold sex information are not homogeneously distributed. For example, Denmark, Estonia and Norway have large quantities of sex data, while Ireland, Portugal, Spain and many Eastern European countries have little (Fig. 3). Most countries have a higher proportion of males than females, with the exception of Finland and France. In Finland the data largely come from a ringing and recovery dataset (Finnish Biodiversity Information Facility, 2024) containing both records of

males and females. In France the data also largely come from a single ringing and recovery dataset (Inventaire National du Patrimoine Naturel, 2020). However, in this case, although the dataset contains over two million records, none of them are annotated as male, and about half a million are annotated as female. The large patches of female dominated areas in France, such as in the Camargue, are the result of this dataset (Fig. 3), as is the anomaly in the 1950s mentioned earlier. In addition, a notable feature of the distribution of sex data is that clusters of data in large cities, including London, Madrid, Prague, Paris and Vienna.

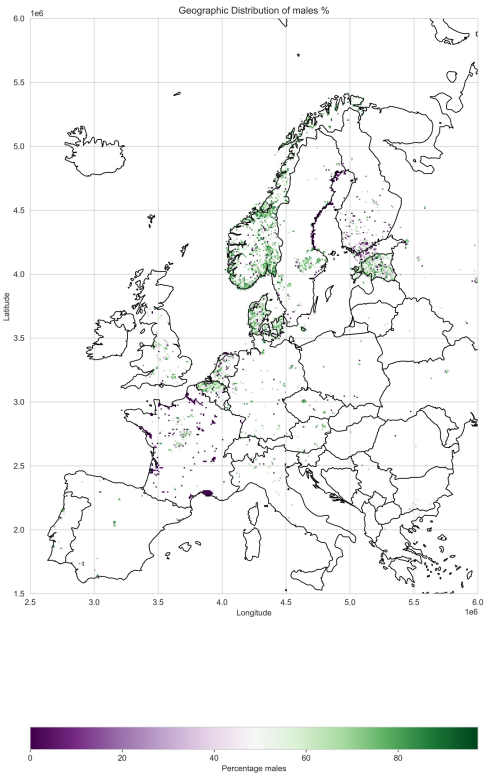

Figure 3: A map of the percentage of male ducks for *Anas platyrhynchos* in Europe aggregate to a 10 km using the EEA grid

Owing to the patchy distribution of records, interpolation can be a useful way to illustrate patterns in sex ratio, because it smooths some of the roughness of the data and gives more local weight to isolated points (Fig. 4).

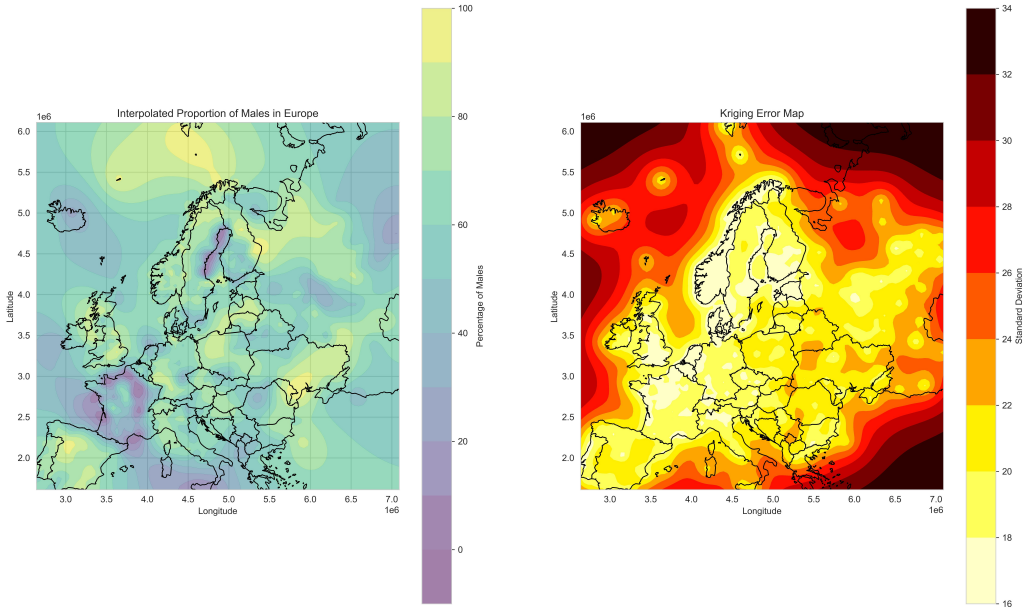

Figure 4: An interpolated map of the proportion of males of *Anas platyrhynchos* and the associated kriging error

Discussion

Geospatial and temporal visualisation has unveiled insights into the sex ratio distribution of Anatidae across Europe. Visualisation of male and female counts over decades has revealed discernible patterns and anomalies, indicating both temporal changes and spatial variations (Fig. 1 & Fig. 2).

Most species in our results are male biassed and that is consistent with other estimates in the literature (Donald, 2007; Owen & Dix, 1986; Pöysä et al., 2019). In the case of *Somateria mollissima* where ratios tend to be female biassed, this is also consistent with the literature (Lehikoinen et al., 2008).

The distribution map shows that sex data are not evenly distributed across Europe (Fig. 3). This is likely the result of national differences in the collection and mobilisation of data (Wetzel et al., 2018). The concentration of records from large urban areas may be the result of records from citizen scientists who tend to focus more on developed areas than professional scientists (Mandeville et al., 2022). Furthermore, the apparent female bias of datasets from ringing is evident. This may be because females are more available for ringing and recapture when they are brooding, nursing their chicks, and while they are moulting. It would be possible to remake the data cube excluding ringing and recapture the dataset to avoid this bias.

However, a significant challenge to accurately assessing these ratios stems from the differential detectability of males and females in many species. Factors such as behavioural differences, sexual dimorphism, and varying habitat preferences can substantially influence the likelihood of observing and recording individuals of each sex. This variability in detectability complicates the interpretation of raw sex ratio data, potentially skewing our understanding of population dynamics. By aggregating vast amounts of occurrence data, these cubes potentially allow for the application of analytical techniques to address the issue of differential detectability. Despite these limitations, other absolute measures of population health are difficult to derive and interpret because of large variation in survey effort both temporally and spatially. As sex ratio is a relative indicator it is potentially less influenced by survey effort and although it is still influenced by spatio-temporal variation in recording there may be ways to smooth this to extract a more accurate biological signal from the data.

Conclusion

The use of biodiversity data cubes derived from GBIF data represents a novel approach to biodiversity studies, allowing for large-scale analysis that was previously more time consuming. This research not only contributes insights into sex ratio analysis but also demonstrates the potential of biodiversity data cubes in advancing ecological and conservation science.

Biodiversity data cubes are an effective tool to analyse, and potentially monitor, sex ratios at scale. Large amounts of data are available, but are rather patchy across Europe. Care needs to be taken with the underlying datasets that some are not significantly biassed for one sex or another, which may happen due to the nature of the methods used in data capture, or the focus of the research.

Acknowledgements

B3 (Biodiversity Building Blocks for policy) receives funding from the European Union's Horizon Europe Research and Innovation Programme (ID No 101059592). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Commission. We extend our gratitude to the Global Biodiversity Information Facility for facilitating access to the data that formed the basis of this research, and to the myriad contributors whose efforts in data collection and curation have enriched the GBIF repository.

References

Donald, P. F. (2007). Adult sex ratios in wild bird populations. *Ibis*, *149*(4), 671–692. https://doi.org/10.1111/j.1474-919X.2007.00724.x

Finnish Biodiversity Information Facility. (2024). *Ringing and recovery database of birds (TIPU)*. Finnish Biodiversity Information Facility. https://doi.org/10.15468/kht1r3

Fryxell, D. C., Arnett, H. A., Apgar, T. M., Kinnison, M. T., & Palkovacs, E. P. (2015). Sex ratio variation shapes the ecological effects of a globally introduced freshwater fish. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1817), 20151970. https://doi.org/10.1098/rspb.2015.1970

GBIF.Org User. (2024). *Occurrence download*. The Global Biodiversity Information Facility. https://doi.org/10.15468/DL.AQUMSN

Groom, Q., & Trekels, M. (2024). *FowlPlay* (Version 1.2) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.12793722

Inventaire National du Patrimoine Naturel. (2020). *Suivi d'oiseaux sauvages par marquage en france - CRBPO - base de données de baguage et déplacements d'oiseaux de france / bird ringing and movement database for france. Centre de recherches sur la biologie des populations d'oiseaux, muséum national d'histoire naturelle, paris, france*. UMS PatriNat (OFB-CNRS-MNHN), Paris. https://doi.org/10.15468/yf8wjf

Lehikoinen, A., Christensen, T. K., Öst, M., Kilpi, M., Saurola, P., & Vattulainen, A. (2008). Large-scale change in the sex ratio of a declining eider Somateria mollissima population. *Wildlife Biology*, *14*(3), 288–301. https://doi.org/10.2981/0909-6396(2008)14%5B288:LCITSR%5D2.0.CO;2

Mandeville, C. P., Nilsen, E. B., & Finstad, A. G. (2022). Spatial distribution of biodiversity citizen science in a natural area depends on area accessibility and differs from other recreational area use. *Ecological Solutions and Evidence*, *3*(4), e12185. https://doi.org/10.1002/2688-8319.12185

Milner, J. M., Nilsen, E. B., & Andreassen, H. P. (2007). Demographic side effects of selective hunting in ungulates and carnivores. *Conservation Biology*, *21*(1), 36–47. https:

//doi.org/10.1111/j.1523-1739.2006.00591.x

Müller, S., Yurchak, R., Murphy, B., nannau, Ziebarth, M., Basak, S., Albuquerque, M., Vrijlandt, M., Peveler, M., Raigosa, D. M., Matchette-Downes, H., Porter, J., Rhilip, Staniewicz, S., Chang, W., & kvanlombeek. (2023). *GeoStat-framework/PyKrige: v1.7.1* (Version v1.7.1) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.10016909

Oldoni, D., Groom, Q., Adriaens, T., Davis, A. J., Reyserhove, L., Strubbe, D., Vanderhoeven, S., & Desmet, P. (2020). Occurrence cubes: A new paradigm for aggregating species occurrence data. *bioRxiv*, 2020–2003. https://doi.org/10.1101/2020.03.23.983601

Owen, M., & Dix, M. (1986). Sex ratios in some common british wintering ducks. *Wildfowl*, *37*, 104–112. https://wildfowl.wwt.org.uk/index.php/wildfowl/article/view/731

Payevsky, V. (2021). Sex ratio and sex-specific survival in avian populations: A review. *Biology Bulletin Reviews*, *11*(3), 317–327. https://doi.org/10.1134/S2079086421030099

Pöysä, H., Linkola, P., & Paasivaara, A. (2019). Breeding sex ratios in two declining diving duck species: Between-year variation and changes over six decades. *Journal of Ornithology*, *160*, 1015–1023. https://doi.org/10.1007/s10336-019-01682-7

Székely, T. (2023). Evolution of reproductive strategies: Sex roles, sex ratios and phylogenies. *Biologia Futura*, *74*, 351–357. https://doi.org/10.1007/s42977-023-00177-0

van der Jeugd, H.P. (2022). *Vogeltrekstation (NL) - historical data on timing of ringing of nestling birds. Version 1.3*. The Netherlands Institute of Ecology (NIOO-KNAW). https://doi.org/10.15468/6l4ban

Wetzel, F. T., Bingham, H. C., Groom, Q., Haase, P., Kõljalg, U., Kuhlmann, M., Martin, C. S., Penev, L., Robertson, T., Saarenmaa, H., Schmeller, D. S., Stoll, S., Tonkin, J. D., & Häuser, C. L. (2018). Unlocking biodiversity data: Prioritization and filling the gaps in biodiversity observation data in europe. *Biological Conservation*, *221*, 78–85. https://doi.org/10.1016/j.biocon.2017.12.024