Estimating the Completeness of Preserved Collections in Representing Global Biodiversity

Pieter Huybrechts‡, Maarten Trekels‡, Quentin Groom‡

‡ Meise Botanic Garden, Meise, Belgium

Corresponding author: Pieter Huybrechts (pieter.huybrechts@plantentuinmeise.be)

Abstract

There are an estimated 8.7 million eukaryotic species globally and knowledge of those organisms is organised about their scientific names and the specimens we have of those species (Sweetlove 2011, Mora et al. 2011). Likewise there are between 1.2 and 2.1 billion (10^9) specimens held in biodiversity collections globally (Ariño 2010). These collections constitute an infrastructure and scientific tool to understand, catalogue and study biodiversity. Yet we find it hard to answer the simple question, how many species are in a collection? This is not trivial to answer, collections are not completely inventoried, do not use the same taxonomy, and the volume of data is vast (Samy et al. 2013, Ariño 2010). We have developed a method that allows us to take a list of collections and to estimate the species richness contained within them. By doing this we will have a deeper insight into the scientific value of the world's biodiversity collections.

Dealing with non-homogeneous and non-random, but incomplete, sampling of sites is a common issue that occurs in many ecological studies (Magurran and McGill 2011, Colwell et al. 2012, Gotelli and Colwell 2001). By using techniques and toolboxes, such as iNEXT (Chao et al. 2014b) and vegan (Oksanen et al. 2020) we can estimate species richness under these conditions. In the case of collections we consider not only the digitized and published proportion of preserved collections, but make extrapolations to the specimens that have not made their way to the Global Biodiversity Information Facility (GBIF) yet.

Nevertheless, to calculate on such large datasets we need to employ innovative Big Data analytic tools. GBIF contains 1.8 billion observations that amount to 120 GB of data compressed. This can then be interrogated in the cloud or locally using tools such as Galaxy, which has made it possible to process large numbers of records in a single batch. We can now evaluate the biodiversity within collections, and divide the result by taxon and geographical region, and compare them to one another.

Ultimately, this work will allow individual collections and consortia to evaluate their coverage of biodiversity and help them better target their collecting strategies.

Keywords

specimen, natural history, big data, GBIF, extrapolation

Presenting author

Pieter Huybrechts

Presented at

TDWG 2021

Funding program

This work was facilitated by the Research Foundation – Flanders research infrastructure under grant number FWO I001721N

Conflicts of interest

References

- Ariño A (2010) Approaches to estimating the universe of natural history collections data. Biodiversity Informatics 7 (2). https://doi.org/10.17161/bi.v7i2.3991
- Chao A, Gotelli N, Hsieh TC, Sander E, Ma KH, Colwell R, Ellison A (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
- Colwell RK, Chao A, Gotelli NJ, Lin S-, Mao CX, Chazdon RL, Longino JT (2012) Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. Journal of Plant Ecology 5 (1): 3-21. https://doi.org/10.1093/jpe/rtr044
- Gotelli N, Colwell R (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecology Letters 4 (4): 379-391. https://doi.org/10.1046/j.1461-0248.2001.00230.x
- Magurran AE, McGill BJ (Eds) (2011) Biological Diversity: Frontiers in Measurement and Assessment. Oxford University Press [ISBN 9780199580668]
- Mora C, Tittensor D, Adl S, Simpson AB, Worm B (2011) How Many Species Are There on Earth and in the Ocean? PLoS Biology 9 (8). https://doi.org/10.1371/journal.pbio.1001127

- Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin P, O'Hara RB, Simpson G, Solymos P, Stevens MHH, Szoecs E, Wagner H (2020) vegan: Community Ecology Package. 2.5-7. URL: https://CRAN.R-project.org/package=vegan
- Samy G, Chavan V, Ariño A, Otegui J, Hobern D, Sood R, Robles E (2013) Content assessment of the primary biodiversity data published through GBIF network: Status, challenges and potentials. Biodiversity Informatics 8 (2). https://doi.org/10.17161/bi.v8i2.4124
- Sweetlove L (2011) Number of species on Earth tagged at 8.7 million. Nature https://doi.org/10.1038/news.2011.498