Biodiversity Informatics, **Training Modules**, 13, 2018, pp. 49-50

SAMPLE DATA AND TRAINING MODULES FOR CLEANING BIODIVERSITY INFORMATION

Marlon E. Cobos, Laura Jiménez, Claudia Nuñez-Penichet,
Daniel Romero-Alvarez, Marianna Simões
Department of Ecology and Evolutionary Biology and Biodiversity Institute, University of Kansas, Lawrence, KS, USA

Abstract.—Large-scale biodiversity databases have become crucial information sources in many analyses in biogeography, macroecology, and conservation biology, often involving development of empirical models of species' ecological niches and predictions of their geographic distributions. These analyses, however, can be impaired by the presence of errors, particularly as regards taxonomic identifications and accurate geographic coordinates. Here, we present an introductory data-cleaning exercise based on two contrasting datasets; we link these example data with a step-by-step guide to overcoming these problems and improving data quality for analyses based on these data.

Key words: accuracy, data cleaning, error, redundancy, precision, primary biodiversity data

The availability and management of large online biodiversity databases has become an exciting, although challenging, step in the development of an authoritative and comprehensive basis for biodiversity knowledge. Thanks to improvements in technology and data collection, steps that previously took years can now be accomplished in seconds (Robin 2012; Musa et al. 2013). The temptation to use this information and interpret analyses quickly, however, often undermines the fundamental necessity of checking data and addressing data quality carefully. This tension is a perpetual caveat to use of modern, open datasets, and has been particularly problematic as regards use of georeferenced, primary biodiversity data. Still, these data are the building blocks for many interesting analyses, perhaps most prominently, species distribution and ecological niche models (Peterson et al. 2011; Anderson 2015).

Primary biodiversity data have become much more accessible in recent decades, with a major transition from recalcitrance (Graves 2000) to enthusiasm. Many institutions (e.g., museums, herbaria, observational data initiatives) have digitized data associated with their work, and increasing numbers now make these data available via the Internet (Soberón and Peterson 2004). The Global Biodiversity Information Facility (GBIF), the Botanical Information and Ecology Network (BIEN), and the Distributed Information System for Biological Collections (*species*Link), are a few examples of repositories that provide online open access biodiversity information, now providing access to over a billion individual records. The benefit of these initiatives is clear; for instance, in 2016, 438 articles were published using data from GBIF alone in multiple fields including data management, evolution, biogeography, biodiversity, and public health (GBIF 2018). However, data quantity is compromised by frequent low data quality; occurrence data from these sources suffer from diverse errors that should be identified, assessed, and minimized before performing any analysis. Low accuracy, low precision, occurrences from outside species' ranges, abundant duplicate records, and taxonomic misidentifications, are just a few of the common errors found in these data (Rahm and Do 2000).

To present a means of building and assessing capacity to handle such problems, we provide a hands-on exercise for data cleaning, with two worked examples, coupled with recommendations and suggestions on how to identify and overcome the most common errors. The goal, of course, is to obtain a cleaned database that will be robust and reliable in different downstream analyses (Peterson et al. 2011). We have assembled two example datasets, one small (960 records) and one large (36,574 records), using records from GBIF—in each case, we have cleaned the data extensively, and then re-populated the dataset with typical classes of errors—we then provide a step-by-step guide to the process of cleaning and improving them. The databases and the data-cleaning manual are available at: http://hdl.handle.net/1808/26512.

These data files and the associated manual are not intended as a detailed treatment of biodiversity data quality, or to offer automated methods with which to solve these problems (Chapman 2005; Hijmans and

Elith 2013; Maldonado et al. 2015). Rather, we focus on individualized, hands-on, user assessment of occurrence datasets, and providing detailed training materials with which to build capacity to make such assessments possible. This publication can be used as a tool for educational purposes, as an exercise for a broader audience that is starting to explore the field, and/or as a reminder of this often-overlooked step (Anderson 2015). We are aware of the time-consuming nature of manual data cleaning, but, to avoid a 'garbage in, garbage out' situation, high-quality data should always be the goal, to allow researchers to develop informative experiments and operational models.

References

Anderson, R. P. 2015. El modelado de nichos y distribuciones: no es simplemente clic, clic, clic. Biogeografía 8:4–27.

Chapman, A. D. 2005. Principles and methods of data cleaning. GBIF, Copenhagen.

Graves, G. R. 2000. Costs and benefits of Web access to museum data. Trends Ecol. Evol. 15, 374.

Hijmans, R. J., and J. Elith. 2017. Species distribution modeling with R. Available at cran.r-project.org/web/packages/dismo/vignettes/sdm.pdf. Accessed 8 March 2018.

Maldonado, C., C. I. Molina, A. Zizka, C. Persson, C. M. Taylor, J. Albán, E. Chilquillo, N. Rønsted, and A. Antonelli. 2015. Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases? Glob. Ecol. Biogeogr. 24:973–984.

Musa, G. J., P.-H. Chiang, T. Sylk, R. Bavley, W. Keating, B. Lakew, H.-C. Tsou, and C. W. Hoven. 2013. Use of GIS mapping as a public health tool—from cholera to cancer. Heal. Serv. Insights. 6:111–116.

Peterson, A. T., J. Soberón, R. G. Pearson, R. P. Anderson, E. Martínez-Meyer, M. Nakamura, and M. B. Araújo. 2011. Ecological niches and geographic distributions. Princeton University Press, Princeton.

Rahm, E., and H. H. Do. 2000. Data cleaning: problems and current approaches. IEEE Data Eng. Bull., 23(4): 3–13.

Robin W. 2012. Robin's Blog. John Snow's famous cholera analysis data in modern GIS formats. Available at http://www.blog.rtwilson.com/john-snows-famous-cholera-analysis-data-in-modern-gis-formats. Accessed March 8 2018.

Soberón, J., and A. T. Peterson. 2004. Biodiversity informatics: managing and applying primary biodiversity data. Philos. Trans. R. Soc. Lond. 359:689–698.