Wildlife Wrangler: A high-level data processing framework that supports the utilization of species occurrence data for biogeographical analyses

Nathan M Tarr[‡], Abigail Benson[§], Matthew J Rubino[‡]

‡ North Carolina Cooperative Fish and Wildlife Research Unit; North Carolina State University, Raleigh, North Carolina, United States of America
§ U.S. Geological Survey, Colorado, United States of America

Corresponding author: Nathan M Tarr (nmtarr@ncsu.edu)

Abstract

Biodiversity data are more findable and accessible to research communities due to efforts over the last 20 years by data infrastructures. Species occurrence data are especially valuable to conservation biogeography analyses, such as the U.S. Geological Survey's Gap Analysis Project (GAP), which assesses how much of individual species' habitat is protected. Prior to application, analysts must process data into desired storage formats, assess the attributes and quality of records, and exclude undesirable data. That processing can become cumbersome and disorganized when the quantity of records or taxa of interest are large. Thus, we developed the Wildlife Wrangler, a tool to facilitate the curation of species occurrence datasets.

The Wildlife Wrangler capitalizes upon the availability of other software packages, APIs, and data standards. It fills gaps in the collective functionality of existing resources with customized functions and tools. The core functionality consists of processes that acquire, filter, and determine the spatial extents of records ("footprints"). Data are primarily acquired from the Global Biodiversity Information Facility (GBIF) API, but the user can also retrieve bird data from the eBird Basic Dataset, which contains spatial information not available from GBIF. In cases where the user is only interested in subsets of available records, filter parameters ("filter sets") can be applied that remove undesirable records. Some filters are applied at the data acquisition step with data request parameters while others, such as unacceptable issue flags, are applied after data are retrieved. Certain spatial filters are supported: queries can be limited to within a single country or a user-defined area of interest, as well as user-defined spatial extents of occurrence for the study taxon. Additionally, the Wildlife Wrangler identifies date-coordinate duplicates while accounting for unequal coordinate precisions among records, allowing the user to exclude such duplicates.

The spatial locations of species observations are of the upmost interest for biogeographical analyses, which are often performed at coarse spatial resolutions at which the spatial precision of records is inconsequential. However, medium- and fine-scale analyses could be sensitive to uncertainty about the locations of the *individuals* that were observed relative to the observer and/or spatial coordinates of the record ("locational uncertainty"). Data providers deliver information on locational uncertainty through the process of georeferencing records (Chapman and Wieczorek 2020), and the Wildlife Wrangler compiles that information to identify each record's footprint, approximating or estimating values when necessary. The user can specify a maximum allowable locational uncertainty whereby imprecise records are omitted.

In addition to the core functionality, the Wildlife Wrangler includes auxiliary functions that aid in generation and use of the output datasets. One such function generates a shapefile of either record coordinates as points, record footprints as polygons, or a randomly selected coordinate from within each record's footprint. Other functions round spatial coordinates, calculate the nominal precision of coordinates, and return a well-known text representation of a polygon.

Additional features provide convenience during the dataset curation process. Query results include data summaries that characterize the retained data and help the user understand what they have acquired. Those summaries facilitate iterative refinement of the output dataset as they often reveal data quality issues that need to be accounted for. Record attributes that are useful when assessing data quality are retained in the output database and can provide a basis for weighting records. In addition to data summaries, the terms-of-use for providers and datasets are summarized.

We designed our framework to be transparent, efficient, repeatable, and accessible. The user can iteratively curate datasets, and the output contains sufficient detail for someone to trace back to decisions made during curation. We automated as many processes as possible for efficiency, consistency, and repeatability. We utilized open-source resources including programming languages (Python, R, and SQL), a database management system (SQLite), environment and package management (Conda), and version control (Git) so that the software would be freely accessible. Filter sets and taxon concept information are stored locally as JSON files and in the output SQLite databases so that queries can be rerun accurately.

Use of the Wildlife Wrangler requires some familiarity with scientific programming, spatial data, and relational databases. However, once installed, datasets can be curated with minimal effort. To perform a query, the user enters taxon information, filter parameters, and associated justification text into a Jupyter Notebook that acts as a form for queries. Running the query performs the data acquisition, filtering, summary, and storage processes. The output dataset is stored in a SQLite database that also contains several data summary tables, the filter set, and the taxon concept information. SQLite databases can easily be archived along with the HTML or PDF copies of the query form notebook.

The USGS recently approved a version of the software and we continue to work on refinements. Although we developed the Wildlife Wrangler for our applications to wildlife habitat conservation, we expect that other researchers may find it valuable.

Keywords

species observation records, conservation biogeography, gap analysis, species distribution modeling, habitat conservation

Presenting author

Nathan M. Tarr

Presented at

TDWG 2022

Conflicts of interest

References

- Chapman AD, Wieczorek JR (2020) Georeferencing Best Practices. GBIF Secretariat https://doi.org/10.15468/doc-gg7h-s853