csv2xyd 2.0 User Manual

by: Jonathan Liria

Introduction

Biodiversity repositories are essential for scientific research, providing valuable information for ecology, biogeography, and conservation studies. A key component of biogeographical research is identifying areas of endemism, which involves analysing many datasets that frequently contain typographical errors, insufficient data, and taxonomic discrepancies; **csv2xyd** is a Python-based software designed to process large CSV files and convert them into XYD format, ideal for endemism analysis with NDM/vNDM. It provides advanced functionalities for preprocessing, filtering, and combining biodiversity data while offering spatial analysis capabilities.

System Requirements

- Python 3.12.4 or higher
- Required libraries: Tkinter 0.1.0, Pandas 2.2.2, Dask 2024.7.0, FuzzyWuzzy 0.18.0, Folium 0.17.0, NumPy 2.0.0, GeoPandas 1.0.1, Shapely 2.0.5.
- Recommended hardware: 16 GB of RAM or more for handling large datasets.

Windows 10 / Linux (Ubuntu 22.04-2 LTS) Installation

1. Install Python from python.org.
2. Install the necessary libraries by running command prompt:

   ```
   pip install tkinter pandas dask fuzzywuzzy geopandas shapely
   folium numpy
   ```

3. Download the software from the GitHub repository and run the `csv2xyd` application.

Main Menu Overview

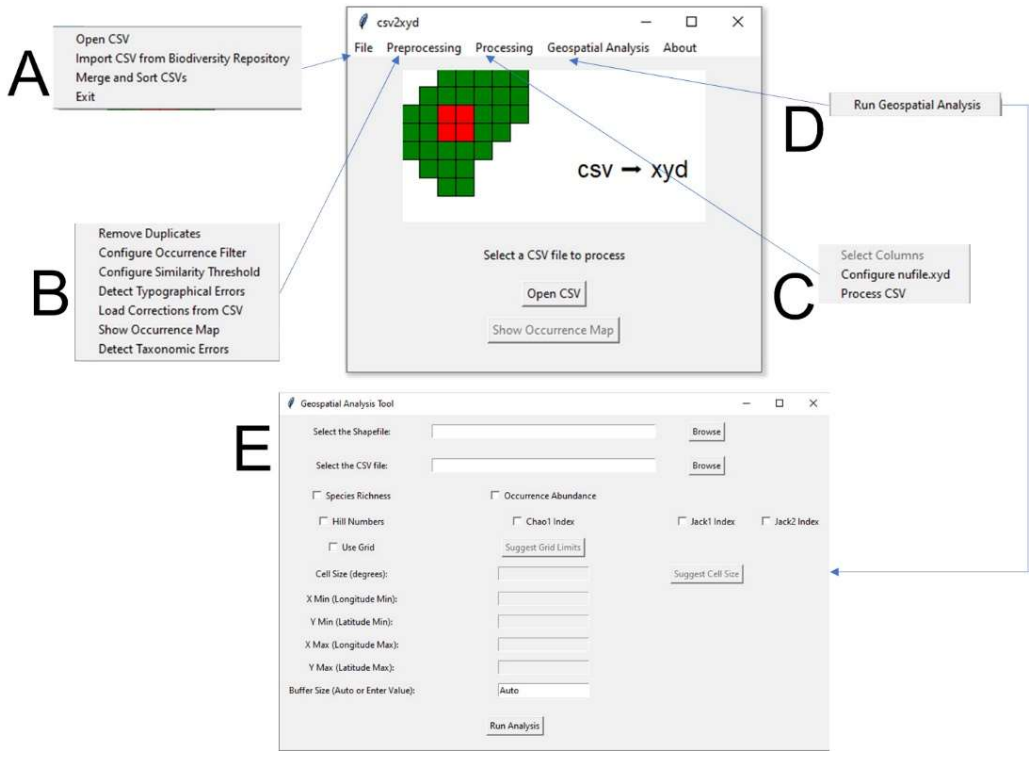

Figure 1. Main menu overview of csv2xyd: A) File, B) Preprocessing, C) Processing, D) Geospatial Analysis, and E) Details of the options in Geospatial Analysis.

File Menu (Fig. 1A)

1. Open Semi-Formatted CSV

- Use this option to open a CSV file with essential columns like species (genus and species), latitude, and longitude. Additional columns for higher taxonomic levels (e.g., phylum, class, order) are also supported.
- Example: Open the file "World_Arthropoda.csv", choose the delimiter (comma or tab), and the program will display the number of occurrences and estimated memory usage.

2. Open CSV from a Biodiversity Repository

- Load CSV files from repositories like GBIF. The program allows you to choose the relevant columns (e.g., phylum, class, order, species) and apply filters such as valid coordinates or specific taxonomic groups.
- Example: Open "0024888-240626123714530.csv", select columns, filter for "Amphibia", and save the filtered file as "Chordata_amphibia.csv".
- This is a list of some commands for filters:

Command	Description	Use
texto	Includes rows/occurrences that contain exactly the specified text.	Insecta in the **class** column will include all rows where **class** field contains "Insecta" Insecta,Arachnida in the **class** column will include all rows where **class** field contains "Insecta" or "Arachnida".
!texto	Excludes rows/occurrences that contain exactly the specified text.	!cf. in the **species** column will exclude all rows where **species** field contains "cf." !.*cf.* in the **species** column will exclude all rows where **species** contains the “cf.” text in any position.
textonly	Filters rows/occurrences where the column contains only text.	textonly in the **phylum** column will select rows where this taxonomic category contains text.
isnumeric	Filters rows/occurrences where the column contains only numeric values.	isnumeric in the longitude column will select rows where longitude contains numeric values.
.	Filters rows that contain any character.	.+ in a column will select all rows where the column contains any text.

3. Merge and Sort CSV Files

- This feature allows merging multiple CSV files and sorting the combined dataset by specific fields.
- Example: Merge "File_1.csv", "File_2.csv", etc., and sort them by genus and species, saving the result as, for example "File_merge.csv".

4. Reorder Columns

- After merging CSVs, you can rearrange the column order to suit your needs.

Preprocessing Menu (Fig. 1B)

1. Remove Duplicates

- Removes duplicate species entries with the same coordinates.

2. Occurrence Filter

- Set a minimum number of occurrences per species (default is 10), ensuring that only species with sufficient data are processed.

3. Detect Typographical Errors

- This option detects species names with small differences, which may indicate typos. A similarity threshold (default: 90%) can be adjusted. The program generates a report (possible_typo_errors.csv) for review and correction.

4. Show Occurrence Map

- Displays a random selection of occurrences on an interactive map using Folium. The default sample size is 10%, but this can be adjusted.

5. Detect Taxonomic Errors

- Identifies species with inconsistent taxonomic assignments, generating a report for review.

Processing Menu (Fig. 1C)

1. Process CSV for XYD

- Convert a CSV file into the XYD format required for NDM/vNDM analysis. Select columns (e.g., genus and species), define grid limits (gridx, gridy), and fill values to generate the XYD file.

2. Export to TNT

- The program also generates a TNT file that can be processed with `gettaxo.run` macro to include higher groups in the endemism analysis.

Geospatial Analysis Menu (Figs. 1D,E)

1. Run Geospatial Analysis

- Perform exploratory spatial analysis on species occurrences and polygons. Load a shapefile or generate a custom grid to match the XYD file's boundaries.
- Adjust the grid size and analyze richness, diversity, and species evenness across the spatial distribution of data.

Example Workflow

1. **Opening and Filtering a CSV**: Load a CSV file, select relevant columns (e.g., species, latitude, longitude), and apply filters (e.g., valid coordinates or a specific taxonomic group). Open CSV file downloaded from biodiversity repositories (e.g., GBIF, CRIA, etc).

2. **Merging Datasets**: Combine multiple CSV files (e.g., merge form different sources datasets).
3. **Preprocessing**: Remove duplicates, select occurrence filter, detect typographical and taxonomic errors, and visualize a random subset of occurrences on a map.
4. **Exporting XYD**: Once preprocessing is complete, convert the data into XYD format for use in NDM/vNDM.
5. **Geospatial Analysis**: Run spatial analyses using custom grids or shapefiles, adjusting the grid size to optimize the spatial representation of your data.

Contact and Support

For support or to contribute to the development of **csv2xyd**, visit the GitHub repository or contact the lead developer at jonathan.liria@gmail.com.

Licensing

The software is open-source and licensed under GNU General Public License (GPL).

References

Cohen, S. 2011. FuzzyWuzzy: Fuzzy string matching in Python. https://github.com/seatgeek/fuzzywuzzy

Folium developers 2017. Folium: Python Data. Leaflet.js Maps. https://python-visualization.github.io/folium/

Gillies, S. 2007. Shapely: Geometric objects, predicates, and operations. https://shapely.readthedocs.io

Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020. Array programming with NumPy. Nature. 585, 357-362. https://doi.org/10.1038/s41586-020-2649-2

Jordahl, K., et al. 2020. GeoPandas: Python Tools for Geographic Data. https://geopandas.org

Lundh, F. 1999. An introduction to tkinter. www.pythonware.com/library/tkinter/introduction/index.htm

McKinney, W. 2010. Data Structures for Statistical Computing in Python. in Proceedings of the 9th Python in Science Conference (eds. van der Walt, S. & Millman, J.) 56–61 https://doi.org/10.25080/Majora-92bf1922-00a

Rocklin, M. 2015. Dask: Parallel Computation with Blocked algorithms and Task Scheduling. Proceedings of the 14th Python in Science Conference (pp. 126-132). https://doi.org/10.25080/Majora-7b98e3ed-013

Szumik, C., Casagranda, M.D., Roig-Juñent, S. 2006. Manual de NDM/VNDM: Programas para la identificación de áreas de endemismo. Instituto Argentino de Estudios Filogenéticos, Año V, Vol. 3. Argentina. https://www.lillo.org.ar/phylogeny/endemism/Manual_VNDM.pdf

Szumik, C.A. and Goloboff, P.A. 2015. Higher taxa and the identification of areas of endemism. Cladistics, 31: 568-572. https://doi.org/10.1111/cla.12112

Szumik, C.A., Cuezzo, F., Goloboff, P.A., Chalup, A.E., 2002. An optimality criterion to determine areas of endemism. Syst. Biol. 51, 806–816. https://doi.org/10.1080/10635150290102483

Szumik, C.A., Goloboff, P.A., 2004. Areas of endemism: an improved optimally criterion. Syst. Biol. 53, 968–977. https://doi.org/10.1080/10635150490888859

Van Rossum, G., Drake, F. L. 2009. Python 3 Reference Manual. CreateSpace. Scotts Valley, CA. 242p.