TOWARDS DEVELOPMENT OF A TOOL FOR THE AUTOMATED ASSESSMENT OF THE SPATIAL ACCURACY OF NATURE OBSERVATION DATASETS

Filip Varga[*1,2], Ana Kuveždić Divjak[3] and Dragica Šalamon[1]

[1] Faculty of Agriculture, University of Zagreb, Croatia
[2] Centre of Excellence for Biodiversity and Molecular Plant Breeding (CroP-BioDiv), Croatia
[3] Faculty of Geodesy, University of Zagreb, Croatia

*correspondence E-mail: fvarga@agr.hr

Keywords: open data; spatial accuracy; big data; toponyms; pipeline concept

1. Introduction

Croatia is one of the most species-rich countries in Europe, thanks largely to its geographical position, where several biogeographical regions overlap, each with its own specific climate, geomorphology, and ecology (Radović et al., 2006). With more than 37,000 taxa within the major taxonomic units, the research on distribution, ecology and conservation, population genetics etc. generates a large amount of spatial data distributed across a large number of taxon-specific databases which are constantly increasing.

Majority of nature obervation databases are a by-product of research projects funded by the government or are set up by non-governmental organisations with little funding. Their data are mostly publicly available, with some exceptions, such as data on endemic and endangered taxa. They often use the help of citizen science initiatives to expand their knowledge of target taxa (Virić Gašparić et al., 2022). Against this background, most nature observation databases, especially the smaller ones, do not have quality control for new records, or only partially apply quality control methods, depending on budget constraints. Input errors related to logical consistency, completeness or duplicate entries are usually mitigated in the initial steps of database architecture development in the form of mandatory fields, input formats and cross-checks with existing records (Dalcin et al., 2012).

Inexperience of researchers in using GPS equipment, such as lack of knowledge of the coordinate reference system (CRS) in which the equipment collects data, writing down coordinates manually instead of using export options on GPS devices often lead to errors in entering records into the database. Errors in the spatial accuracy of records (largely geographical coordinates) are much more difficult to determine. It requires advanced knowledge of GIS (geographical information system) and still takes a lot of time to manually determine whether a particular questionable record has been correctly entered into the database. High-quality spatial data for plant and animal taxa may be the most valuable information for researchers building spatial distribution models to predict the potential

distribution of endangered and endemic taxa, as well as potential risks that invasive species pose to biodiversity, and for government stakeholders involved in decision-making processes for the conservation and sustainable use of species and habitats (Glasnovic et al., 2018; Valencia-Rodríguez et al., 2021).

The aim of this paper was to explore and propose the possibility of developing a pipeline concept for an automated tool to assess the spatial accuracy of nature observation records.

2. Research approach

The first step in our research was a small-scale analysis of the quality of spatial records for Dalmatian pyrethrum (*Tanacetum cinerariifolium* (Trevis.) Sch. Bip.) from the largest Croatian plant database, the Flora Croatica database (FCD) (Nikolić, 2022). The species is endemic and protected by law but has also been cultivated in Dalmatia for a long time as it produces a natural insecticide, pyrethrin, which makes it interesting for both conservation biologists and the agricultural sector. Data on Dalmatian pyrethrum included observational data, herbarium vouchers, published research data and photographs from the natural habitat and was scrapped using Beautiful Soup version 4.9.3. (Python library). The positional accuracy for 906 records in total was checked manually. Points located in water bodies (Dalmatian pyrethrum is a terrestrial species) or outside national boundaries were searched in QGIS version 3.16.9. (QGIS Development Team, 2021).

The second part of the research was to design a pipeline concept suitable for nature observation databases that could be implemented in various open-source software to make it accessible to a wider scientific community. This involved developing theoretical steps for the pipeline detailing the assessment process, exploring different solutions to the specificities of nature observation data, such as the need for geoparsing, i.e., assigning geographic coordinates to free-text descriptions of locations (a commonly used format for describing sampling locations), and finding the best open-source databases with reference toponyms to use in the assessment.

3. Results and discussion

3.1. Preliminary positional accuracy study

The preliminary results showed a relatively high quality of data, with only 4.1% of records classified as spatially inaccurate (1 record was outside national boundaries and 36 records were in water bodies). We had expected the highest positional inaccuracy to occur in nature observation data, which make up the largest share of records (67.8%). For published research data and herbarium specimens, we expected the quality of the specimens to be high, as herbarium curators and researchers check their data extensively before including them in herbarium collections or publishing articles. However, the lowest positional accuracy was found in published research data (10.1% of records are spatially inaccurate), which is alarming.

We are well aware of the limitations of this analysis. Its small scale does not necessarily reflect the quality of the entire database or the state of spatial data in other nature observation databases. The fact that the vast majority of data on Dalmatian pyrethrum in the last two decades (74.5%) was collected using GPS (according to the metadata) may further obscure the

actual quality of the spatial records. If we extrapolate the percentage of low-quality data from this study to the entire FCD, we arrive at a large number of questionable records in terms of spatial quality, namely more than 4,000. Such a large number of records that need to be manually checked would further burden the already understaffed and underfunded institutions that maintain the databases in question.

3.2. Pipeline concept

The first and crucial step in designing this pipeline was to determine the best possible open-source spatial database of Croatian toponyms. The majority of nature observation research data is collected in Croatia, and the location descriptions, including toponyms, are often so small that they cannot be found using global spatial search engines such as Google Maps or OpenStreetMap. For this reason, we strongly recommend using national toponym registers whenever possible. Not only do they contain a large number of records, but they are also standardised and regularly updated. The Register of Geographical Names in the Republic of Croatia consists of 124,018 spatial records and can be downloaded through the WFS service in shapefile format provided by the State Geodetic Administration (State Geodetic Administration, 2021).

The second step we explored was geoparsing. Since nature observation records contain a description of the sampling location in addition to geographical coordinates, and often in free form (e.g., "Dalmatia, Šibenik archipelago, Žirje island, 100 metres from Straža"), we needed to find an efficient way to convert the text identifying a location into a unique geographical reference (from the toponym register), which can then be cross-checked with the sampling coordinates and determine their accuracy. Some of the geoparsing tools we find promising are Mordecai, a full text geoparsing system (Halterman, 2017) and the opencage package for R software (Possenriede et al., 2021). Both are open source and implemented in Python and R respectively, which are known and used worldwide. They can be connected to QGIS and automate the spatial analyses required in this pipeline.

The tool theoretically requires three main components to function properly. The first is the toponym reference dataset in spatial format (shapefile, GeoJSON). Ideally, users should be able to replace the toponym reference dataset with another one so that the tool can be used beyond the Croatian borders. The second component is the sample dataset, which the user provides in tabular form, and which is compared with the toponym reference dataset and checked for spatial accuracy. The third component consists of the software environment in which the whole process takes place. Installation of the tool and use by the user should be easy, and all necessary dependencies should be installed with the tool (Brack et al., 2022). The assessment process described in **Figure 1** is as follows:

1. The user enters the tabular sample dataset into the software environment and defines the CRS of the sample dataset.
2. The tabular sample dataset is converted to spatial format. If necessary, CRS is converted to match the CRS of the toponymic reference dataset.

3.Each record of the sample dataset (description of sampled locations) is geoparsed using the toponym reference dataset. The geographical distance between the sampled location (from the sample dataset) and the reference toponym is calculated.
3.1.If multiple reference toponyms are found in a single record, their names, geographical coordinates, and distance between each reference toponym and sampled location are stored in the sample dataset.
3.2.If no toponyms are found in a single record, this information is stored in the sample dataset and the user receives a notification. The process continues with evaluation of the next record.
4.The process ends when the last record in the sample dataset has been processed.
5.When the assessment of the entire dataset is completed, the user has the option to export:
5.1.the initial sample dataset in tabular form with the addition of the name, geographical coordinates, and distance of the nearest reference toponym to the sampled location for each record.
5.2.The extended sample dataset in tabular form with addition of the name, geographical coordinates and distance of all reference toponyms found for each record. In this dataset, each record is represented with multiple rows depending on the number of reference toponyms found in the location description.
5.3.Initial sample dataset in spatial format (shapefile, GeoJSON)
5.4.Extended sample dataset in spatial format (shapefile, GeoJSON)

Automated assessment of spatial data accuracy could benefit producers (as well as curators) of nature observation data by giving them insight into, and identifying, the number of records that should be assessed more closely for spatial accuracy. The tool could also be of great importance to researchers working on spatial distribution and niche modelling, as it will allow them to identify spatial records for target taxa that they should omit from their datasets to improve the accuracy of the models they develop. The main challenge for the future is the implementation of geoparsing methods for the Croatian language and the evaluation of the success of geoparsing (Gritta et al., 2020).

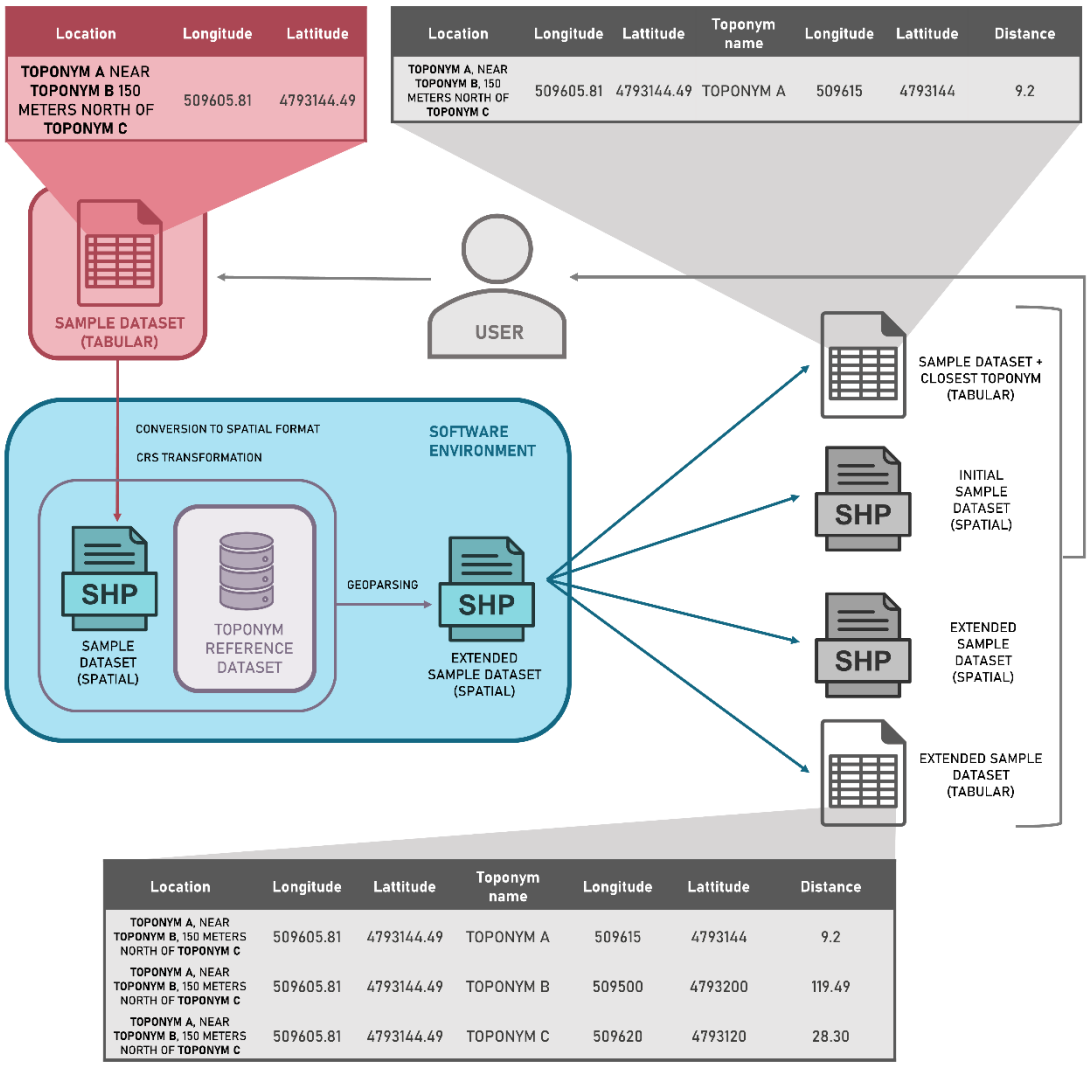

Figure 1. Theoretical pipeline for automated assessment of spatial accuracy of records in nature observation databases.

4. Conclusions and future work

The proposed concept of a tool for the automated assessment of the spatial accuracy of nature observation data sets serves as a design draft that will surely be expanded and modified throughout the implementation process and testing stages. Inclusion of geoinformatics experts and software developers in the next phase of development will provide us with feedback on design flaws and steps necessary to successfully implement this pipeline in a robust assessment tool. After development, variety of nature observation datasets and potentially datasets from other sectors will be tested in order to determine how widely can the tool be used for improvement od spatial records quality on a massive scale.

Acknowledgments: This research is part of TODO project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857592.

References

Brack, P., Crowther, P., Soiland-Reyes, S., Owen, S., Lowe, D., R. Williams, A., Groom, Q., Dillen, M., Coppens, F., Gruning, B., Eguinoa, I., Ewels, P., Goble, C. (2022). Ten simple rules for making a software tool workflow-ready. PLOS Computational Biology 18(3): e1009823. https://doi.org/10.1371/journal.pcbi.1009823

Dalcin, E., Silva, L., Cabanillas, C., Loures, M., Monteiro, V., Send Email To Zimbrao, I., Souza, J. (2012). Data Quality Assessment at the Rio de Janeiro Botanical Garden Herbarium Database and Considerations for Data Quality Improvement. In Proceedings of 8th International Conference on Ecological Informatics (ISEI) 1–11. Brasilia, Brasil.

Glasnovic, P., Temunović, M., Lakušić, D., Rakić, T., Grubar, V. B., Surina, B. (2018). Understanding biogeographical patterns in the western Balkan Peninsula using environmental niche modelling and geostatistics in polymorphic *Edraianthus tenuifolius*. AoB PLANTS 10(6). https://doi.org/10.1093/aobpla/ply064

Gritta, M., Pilehvar, M. T., Collier, N. (2020). A pragmatic guide to geoparsing evaluation: Toponyms, Named Entity Recognition and pragmatics. Language Resources and Evaluation 54(3): 683. https://doi.org/10.1007/s10579-019-09475-3

Halterman, A. (2017). Mordecai: Full Text Geoparsing and Event Geocoding. Journal of Open Source Software 2(9): 91. https://doi.org/10.21105/JOSS.00091

Nikolić, T. (Ed.). (2022). Flora Croatica Database. Available at http://hirc.botanic.hr/fcd

Possenriede, D., Sadler, J., Salmon, M. (2021). opencage: Geocode with the OpenCage API. R package version 0.2.2. Available at https://cran.r-project.org/package=opencage

QGIS Development Team. (2021). QGIS Geographic Information System 3.16.9. Open Source Geospatial Foundation Project. Available at http://qgis.osgeo.org

Radović, J., Čivić, K., Topić, R. (Eds.). (2006). Biodiversity of Croatia. Zagreb, Croatia: State Institute for Nature Protection, Ministry of Culture - Republic of Croatia.

State Geodetic Administration. (2021). Register of Geographical Names in the republic of Croatia.

Valencia-Rodríguez, D., Jiménez-Segura, L., Rogéliz, C. A., Parra, J. L. (2021). Ecological niche modeling as an effective tool to predict the distribution of freshwater organisms: The case of the *Sabaleta Brycon henni* (Eigenmann, 1913). PLOS ONE 16(3): e0247876. https://doi.org/10.1371/journal.pone.0247876

Virić Gašparić, H., M. Mikac, K. M., Pajač Živković, I., Krehula, B., Orešković, M., Galešić, M. A., Ninčević, P., Varga, F., Lemić, D. (2022). Firefly Occurrences in Croatia – One Step Closer from Citizen Science to Open Data. Interdisciplinary Description of Complex Systems 20(2): 112–124. https://doi.org/10.7906/INDECS.20.2.4