GeoEnrich: A New Tool For Scientists To Painlessly Enrich Their Species Occurrence Data With Environmental Variables

Authors : Gaetan Morand[1], Julien Barde[2]
[1]UMR MARBEC (IRD, Ifremer, CNRS, Université De Montpellier), Saint-Pierre, Réunion,[2]UMR MARBEC (IRD, Ifremer, CNRS, Université De Montpellier), Sète, France

E-mail Address: gaetan.morand@ird.fr

ID: 12182

Submission:

Background

As biodiversity undergoes a fast decline, it is essential to understand the life cycles of threatened species to be able to protect them and ensure a sustainable use of marine ecosystems. From water temperature and salinity, to eddy kinetic energy or ocean heat content, environmental variables are key to understanding species' distributions, movements and migrations.

Yet the complex formats and large sizes of environmental datasets still represent barriers to scientists willing to study the influence of these factors. An additional level of complexity is added as species occurrences datasets tend to get larger due to the modernization of data collection techniques. Often, those challenges can lead to the use of downgraded versions of the data, such as averages over long time periods or reduced geographical extents, which can impair the ability to understand important processes. The present research is at the crossing of marine biology, oceanography, and data science. It removes the current technical barriers and empowers scientists with full data access so they can set the scene for the future of ocean conservation.

Method

We designed and developed a new tool that enables scientists to easily associate environmental variables co-occurring with their species records (or any other geo-referenced data). It allows specifying a geographical buffer (e.g., 100km) as well as a time window (e.g., from one month before the occurrence to the day of occurrence) around species records, and download all the data meeting these requirements. The tool gives access to datasets documenting 30 different variables including satellite observations such as sea surface temperature and chlorophyll ,as well as modeled variables such as ocean currents at multiple depth levels. Moreover, users can easily add additional data sources.

Data for the requested locations and time windows are then downloaded from remote servers, such as NOAA's or Copernicus', using the OpenDAP protocol that queries netCDF files. This method is intended for large occurrences datasets, so it implements local storage to avoid redundant requests to remote servers. As more and more occurrences are being enriched with remote data, the need to download new data decreases, and with it the processing time and bandwidth requirements.

Results

This new tool is available in open-source and is free to use. For basic needs, a web interface allows anybody to request a summary of the variables of interest: for each occurrence, they can get the corresponding variable value, as well as the average, standard deviation, minimum and maximum over the user-specified buffer. Such a request provides enough data to perform exploratory analyzes.

For more advanced needs, Python and R packages allow users to download complete datasets over the entire buffers to have an accurate view of the oceanographic landscape at the time and place of species' occurrence. Additional functionalities are also available to let users prepare the output environmental data for subsequent analyzes, such as the ability to export data as netCDF files or as pictures that can be used for visualization or to train deep learning algorithms.

Conclusion

Being able to easily download complete environmental data is often a required step when trying to understand ecological processes. However, this task is complex even for experts, and as such it often impedes scientists' capacity to include all relevant data into analyzes.

Here, we designed and developed a new software tool that provides scientists with all the necessary context to investigate the influence of environmental parameters. It therefore opens the way to more robust studies on species' distributions and migrations. These may in turn help us understand the present and future consequences of anthropogenic changes on marine biodiversity, and guide us towards more sustainable resource management.