Applying Cluster Analysis in the Study of Large-Scale Species Occurrence Data

Kun-Chi Lai [1], Yueh-Chih Chen [2], You-Sheng Li [3] and Kwang-Tsao Shao [2]

[1] Department of Computer Science, National Chengchi University
[2] Biodiversity Research Center, Academia Sinica
[3] Research Center for Information Technology Innovation

TaiBIF stands for Taiwan Biodiversity Information Facility, the Taiwan Node of the Global Biodiversity Information Facility. One of TaiBIF's major tasks is to integrate species occurrence data which include records of animal specimens in museums, plant specimens in herbaria, ecological surveys, and species observations. The majority of the occurrence data in the TaiBIF integration platform come from TELDAP (specifically the specimen collections of the Biosphere & Nature thematic group, Taiwan e-Learning and Digital Archives Program). Another source is the non-TELDAP institutions with their digitized specimen or observational data.

The TaiBIF data portal has integrated 26 datasets so far, resulting in more than 1.5 million species occurrence data with 85% of them geospatial referenced. This study utilizes 8,526 Cyprinidae occurrence data from 11 datasets and uses different types of clustering algorithms to produce different spatial visualization results.

The study explores the comparative differences between the results obtained from the clustering algorithms and the expert opinion range maps of Cyprinidae. It tries to resolve the problems of efficacy and poor visualization when large scales of species occurrence data are presented in Google Maps and hopes to identify a quick and efficient way to present species distribution data, which in turn can help researchers to extract knowledge from large amounts of data so that the knowledge can be tapped for ecological conservation efforts in the future.

Keywords: Biodiversity informatics, cluster analysis, Cyprinidae, species occurrence data, visualization, zoogeographical