Use of spatially, temporally, and taxonomically biased species occurrence records in biodiversity science

Shogo Ikari[1], Buntarou Kusumoto[2], Osamu Komori[3,4], Hsin-Hsiung Bill Huang[5],

[1] Graduate School of Engineering and Science, University of the Ryukyus.

[2] Faculty of Agriculture, Kyushu University, Japan

[3] Department of Science and Technology, Seikei University, Japan

[4] School of Statistical Thinking, The Institute of Statistical Mathematics, Japan

[5] Department of Statistics and Data Science, University of Central Florida, USA

Species occurrence records (i.e., when, where, and which species were observed) constitute foundational data for estimating species diversity and distribution. At the macroscopic level, estimating distribution necessitates the standardization and integration of heterogeneous data sources.

In our session, we first assessed the characteristics of biases in species occurrence records from different data sources. Among them, the most reliable data encompass specimen information meticulously documented by taxonomists and precise geospatial coordinates. Moreover, community survey data collected by other field researchers (such as vegetation surveys) also exhibit significant potential. In recent years, the contributions of citizen scientists in amassing extensive data have been acknowledged for their ability to efficiently capture large-scale information. Each data source manifests variances in taxonomic and spatial configurations. For instance, while professional scientists may concentrate on scientifically innovative locales and taxonomic groups, citizen science may demonstrate preference towards easily accessible regions and taxa recognizable even by amateur enthusiasts. In this study, our aim is to evaluate the completeness and equitability of taxonomic representation of Japanese natural history data (specimens, community survey, and citizen science) using a diversity estimation theory based on Hill numbers. We elucidate the inherent biases in each data source and discuss on strategies for effectively harnessing heterogeneous datasets.

Second, we evaluate species occurrence records in terms of the description of biodiversity patterns at macroscales. The lack of information on the spatial distribution of species (Wallacean shortfall) is an inherent part of biodiversity research. Attempts have been made to reduce uncertainty arising from incomplete distribution data, such as species distribution modeling and statistical methods to estimate species richness (e.g. Chao's estimator).

However, assessment of completeness of the visualized biodiversity patterns themselves have rarely been studied. Here, the current study proposes a novel method to quantify the completeness of taxonomic richness patterns by assessing the asymptotics of changes in taxonomic richness patterns obtained from datasets of occurrence records which are accumulated over time. Specifically, we tested whether a taxonomic richness map does not change by appending the data utilized for its compilation. If the variance of taxonomic richness in the present-day known patterns is well-accounted for by that in the past known patterns (as measured by a high R-squared value yielded by the linear regression), then the known taxonomic richness patterns are assumed to be invariant to further data accumulations, hence highly complete. This method was applied to taxonomic richness patterns generated by global occurrence datasets of GBIF and OBIS. Also, we used simulated datasets to test its detectability of completeness of species richness.

Thirdly, we consider a new estimating algorithm for species distribution modeling. In the field of species distribution modellings, Maxent is recognized as one of the most powerful tools, in which the model's parameters are estimated by sequentially maximizing the likelihood. However, computing the likelihood becomes extremely expensive when the number of locations in the study area is large. Therefore, we propose a new species distribution modeling approach based on gamma-divergence, in which the normalization term in the gamma-loss function is approximated using cumulant coefficients for the study area. This results in a computationally efficient estimation of the model's parameters, as well as high estimation accuracy. Moreover, by using the cumulant-based approximation, we show that the estimation of Maxent is equivalent to Fisher linear analysis when the normality assumption holds. We demonstrate the proposed method using simulation studies and a dataset of Japanese vascular plants.

Finally, Dr. Huang will discuss his research on spatiotemporal algorithms for threat detection and biodiversity data analysis, highlighting significant advancements made through projects like the ATD Challenge from 2021 to 2023. His team's innovative approaches, particularly their development of a quantile-based hybrid model combining nearest-neighbor Gaussian processes with temporal fusion transformers, have dramatically improved the predictability of complex events, demonstrating the profound impact of integrating advanced statistical techniques in biodiversity science.