World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:14, No:12, 2020

Recommendations for Data Quality Filtering of Opportunistic Species Occurrence Data

Authors : Camille Van Eupen, Dirk Maes, Marc Herremans, Kristijn R. R. Swinnen, Ben Somers, Stijn Luca

Abstract : In ecology, species distribution models are commonly implemented to study species-environment relationships. These models increasingly rely on opportunistic citizen science data when high-quality species records collected through standardized recording protocols are unavailable. While these opportunistic data are abundant, uncertainty is usually high, e.g., due to observer effects or a lack of metadata. Data quality filtering is often used to reduce these types of uncertainty in an attempt to increase the value of studies relying on opportunistic data. However, filtering should not be performed blindly. In this study, recommendations are built for data quality filtering of opportunistic species occurrence data that are used as input for species distribution models. Using an extensive database of 5.7 million citizen science records from 255 species in Flanders, the impact on model performance was quantified by applying three data quality filters, and these results were linked to species traits. More specifically, presence records were filtered based on record attributes that provide information on the observation process or post-entry data validation, and changes in the area under the receiver operating characteristic (AUC), sensitivity, and specificity were analyzed using the Maxent algorithm with and without filtering. Controlling for sample size enabled us to study the combined impact of data quality filtering, i.e., the simultaneous impact of an increase in data quality and a decrease in sample size. Further, the variation among species in their response to data quality filtering was explored by clustering species based on four traits often related to data quality: commonness, popularity, difficulty, and body size. Findings show that model performance is affected by i) the quality of the filtered data, ii) the proportional reduction in sample size caused by filtering and the remaining absolute sample size, and iii) a species 'quality profile', resulting from a species classification based on the four traits related to data quality. The findings resulted in recommendations on when and how to filter volunteer generated and opportunistically collected data. This study confirms that correctly processed citizen science data can make a valuable contribution to ecological research and species conservation.

Keywords : citizen science, data quality filtering, species distribution models, trait profiles
Conference Title : ICSRD 2020 : International Conference on Scientific Research and Development
Conference Location : Chicago, United States
Conference Dates : December 12-13, 2020