On the Role of Spatial Clustering Algorithms in Building Species Distribution Models from Community Science Data

This research discusses opportunities for developments in spatial clustering methods to help leverage broad scale community science data for building species distribution models (SDMs). SDMs are critical tools that inform the science and policy needed to mitigate the impacts of climate change on biodiversity. Community science data span spatial and temporal scales unachievable by expert surveys alone, but they lack the structure imposed in smaller scale studies to allow adjustments for observational biases such as imperfect detection. Spatial clustering algorithms can construct the necessary structure after surveys have occurred, but more work is needed to ensure that they are effective for this purpose. This research focuses on occupancy models and its rigid structure that assumes observations are organized into a set of sites, where each site has one or more visits that satisfy the closure assumption. Existing methods that attempt to group unstructured species observation into sites are either too constrained, leading to an exclusion of useful data, or too lenient, resulting in sites that violate closure. In this research, we formalize this challenge and introduce it as the Site Clustering Problem. We describe the role of spatial clustering for improving upon these existing methods and realizing the potential of large biodiversity datasets to build SDMs.