36

A machine learning method for estimating the probability of presence using presence-background data

Estimating the prevalence or the absolute probability of presence of a species from presence-background data has become a controversial topic in species distribution modelling. In this paper we propose a new method by combining both statistics and machine learning algorithms that overcomes many of the known existing problems. We have also revisited the popular but highly controversial Lele and Keim (LK) method by evaluating its performance and assessing the RSPF condition it relies on. Simulations show that the LK method is fragile and often fails to give reliable estimates even when its underlying RSPF condition is met. In addition it is shown through simulation studies that when ``local knowledge'' is available, the new method proposed here is able to accurately estimate the actual probability of presence, outperforming the LK method regardless of the type of true parametric regression functions used. The new local knowledge condition proposed in this paper introduces and extends the local certainty condition known in the context of machine learning and serves as the more generalised condition for accurately estimating the absolute probability of presence from presence-background data. Our conclusion emphasises that in order to make accurate estimations, any extra information has to come from the data itself rather than from introducing unfounded model assumptions. The latter only renders fragile estimation/prediction of the desired probabilities.