Evaluation of (Machine Learning) Prediction Models for spatially correlated data

Authors: Andreas Bender[1]* et al.

1 Department of Statistics, LMU, Munich, Germany

*corresponding author: andreas.bender@stat.uni-muenchen.de;
https://arxiv.org/abs/2310.15108

Abstract: Unbiased estimation of the generalization error (GE) is an important aspect of the evaluation of predictive models, including machine learning methods. The most common approach to do so is to employ resampling techniques like cross validation or bootstrapping. However, their standard application usually assumes independence and identical distribution of individual data points.

In settings where this is not the case, estimation of the GE through standard resampling techniques can lead to biased or over optimistic estimates. In this talk, we will summarize settings in which adaptations might be necessary, particularly clustered and hierarchical data, data sampled with unequal probabilities and spatially correlated data. Focusing on spatial data, we will illustrate how common resampling techniques can be adapted to different use cases, including a short illustration using R Code.