Spatial prediction with spatially clustered data by combining residual kriging and instance-based transfer learning

Jeremy Rohmer[1]

[1]{BRGM, 3 av. C. Guillemin - 45060 Orléans Cedex 2 - France}
Correspondence to: J. Rohmer (j.rohmer@brgm.fr)

Data points are rarely uniformly distributed in space due to bias in the sampling procedures or due to difficulties to do measurements in the field (e.g. difficulties in accessing to the sites, time constraints, lack a common sampling design between the successive measurement campaigns). This leads to some spatial structures that often take the form of clusters, i.e. data points may over-represent some regions while under-representing or even missing others. This might negatively impact the capability of the spatial machine learning model to predict test points that lie out of the domain of the (training) data points. To address this problem, we borrow instance-based methods from the transfer machine learning community. Under the so-called "covariate shift adaptation" assumption, we re-weight the training data in order to match the test probability distribution. The weights are related to the ratio between the test (target) and the training (source) densities and can be calculated either by minimizing the distance between the distributions (e.g. with Maximum Mean Discrepancy) or using probabilistic classification methods. We propose to use the weights to correct the sample bias at two levels: (1) during the fitting of the spatial machine learning model; here we focus on random forest regression models that can incorporate weights during the splitting of the trees so that observations with larger weights are selected with higher probability in the bootstrap samples for the trees; (2) the spatial distribution of the residuals by relying on a weighted Gaussian process regression model with adjustments of the simple kriging equations similar to those of the normalized kernels. We investigate the benefits of this twofold correction by testing them to simulation studies defined by randomly varying the number, size and spatial extent of the clusters as well as the number of training points outside the clusters. To do so, two large datasets are used, namely the soil organic carbon stock (0–30 cm soil depth) at the European scale from SoilGrids and the virtual species suitability surface for the Iberian peninsula from WorldClim bioclimatic variables. In both cases, we show a twofold improvement in terms of prediction accuracy (with a decrease of the relative absolute error up -10%), as well as in terms of prediction reliability with a sharpening of the distributional prediction as indicated by a decrease of the continuous rank probability score up -20%.