Geophysical Research Abstracts
Vol. 19, EGU2017-3256, 2017
EGU General Assembly 2017

The R package "sperrorest" : Parallelized spatial error estimation and variable importance assessment for geospatial machine learning

Patrick Schratz, Tobias Herrmann, and Alexander Brenning
Department of Geography, Friedrich Schiller University Jena, Löbdergraben 32, 07743 Jena, Germany

Computational and statistical prediction methods such as the support vector machine have gained popularity in remote-sensing applications in recent years and are often compared to more traditional approaches like maximum-likelihood classification.

However, the accuracy assessment of such predictive models in a spatial context needs to account for the presence of spatial autocorrelation in geospatial data by using spatial cross-validation and bootstrap strategies instead of their now more widely used non-spatial equivalent. The R package ***sperrorest*** by A. Brenning [IEEE International Geoscience and Remote Sensing Symposium, 1, 374 (2012)] provides a generic interface for performing (spatial) cross-validation of any statistical or machine-learning technique available in R.

Since spatial statistical models as well as flexible machine-learning algorithms can be computationally expensive, parallel computing strategies are required to perform cross-validation efficiently. The most recent major release of sperrorest therefore comes with two new features (aside from improved documentation): The first one is the parallelized version of ***sperrorest()***, ***parsperrorest()***. This function features two parallel modes to greatly speed up cross-validation runs. Both parallel modes are platform independent and provide progress information. ***par.mode = 1*** relies on the ***pbapply*** package and calls interactively (depending on the platform) ***parallel::mclapply()*** or ***parallel::parApply()*** in the background. While forking is used on Unix-Systems, Windows systems use a cluster approach for parallel execution. ***par.mode = 2*** uses the ***foreach*** package to perform parallelization. This method uses a different way of cluster parallelization than the ***parallel*** package does. In summary, the robustness of ***parsperrorest()*** is increased with the implementation of two independent parallel modes.

A new way of partitioning the data in ***sperrorest*** is provided by ***partition.factor.cv()***. This function gives the user the possibility to perform cross-validation at the level of some grouping structure. As an example, in remote sensing of agricultural land uses, pixels from the same field contain nearly identical information and will thus be jointly placed in either the test set or the training set. Other spatial sampling resampling strategies are already available and can be extended by the user.