Spatial+: a new cross-validation method to evaluate geospatial machine learning

Spatial prediction machine learning model's evaluation and cross-validation

Geospatial prediction studies, such as soil mapping, ecological modeling, have extensively employed Machine Learning (ML) models. The evaluation of the model is a crucial step. To obtain reliable results, a test set that unbiasedly represents the prediction locations is needed. However, obtaining an additional test set is frequently impractical. Thus, the available samples must be partitioned into training and validation subsets to implement the evaluation.

The random k-fold cross-validation (RDM-CV) is a commonly used method for evaluating the ML model, but it may not be suitable for geospatial prediction, where available sample data and prediction locations usually have obvious differences. For example, samples may be distributed in spatial clusters, or the model may be required to predict a new area, i.e., extrapolation. RDM-CV cannot account for the differences by its random split. As a result, the model's evaluation result of RDM-CV will be over-optimistic.

Since the 2000s, a series of spatial CV methods have been proposed such as buffer CV, weighted CV, and block CV (BLK-CV). All these methods reflect the differences between training and validation samples by keeping them far away and not spatially autocorrelated. However, in such a manner, the differences can only be considered from the geographic space, while many of them which affect the model's prediction are derived from the feature space. Thus, in this work, we propose a new CV method to guarantee the split training and validation subsets provides a more accurate representation of the actual existed differences.

Spatial+ cross-validation

As figure 1 shows, the proposed method (spatial+ cross-validation, SP-CV) is composed of two stages to consider both the geographic and feature spaces to comprehensively reflect the differences between the samples data and prediction locations.

The 1st stage addresses spatial autocorrelation issues by using agglomerative hierarchical clustering (AHC) to divide the available samples into blocks. AHC is a "bottom-top" clustering method that always merges the closest samples pair or sub-clusters pair. The 1st stage improves the block's division by addressing the samples' spatial distribution.

The 2nd stage accounts for the differences in both the geographic space (locations) and feature space (covariates and target variable). It uses cluster ensembles (CE) to split folds. First, all blocks acquired from the 1st stage are separately clustered based on locations, target variable, and covariates respectively. Then, as shown in figure 1, the CE is used to combine them together to reflect the differences in both the geographic and feature spaces.

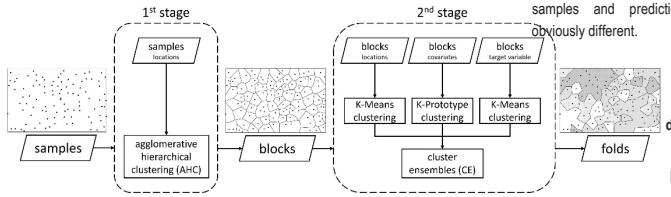

Figure 1: The flowchart of the spatial+ cross-validation.

Experimemts & Results

In this research, the proposed SP-CV was compared with the traditional CV – RDM-CV and the typical spatial CV – BLK-CV. The experiments workflow is shown as figure 2.

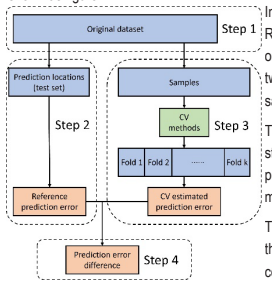

Figure 2: The workflow of the experiments.

In the experiments, for comparing RDM-CV, BLK-CV, and SP-CV, the original dataset should be split into two parts at first: the test set, and the samples.

The test set is used to calculate the standard prediction ability (i.e., prediction error) of the spatial ML model.

The samples are used to implement three CV methods to acquire corresponding estimated prediction ability.

The experiments were implemented on two datasets: 1) Brazil Amazon basin above ground biomass (AGB) dataset, with 28 covariates and 1km^2 resolution 928 * 1642 layer. 2) California houseprice dataset with 9 covariates and 20640 records.

The Amazon AGB dataset is used to simulate the actual situation of clustered samples. Thus, its experiments setup (test set and samples) is shown like figure 3(a). The California houseprice dataset is used to simulate the actual situation of extrapolation, its experiments setup is shown as figure 3(b).

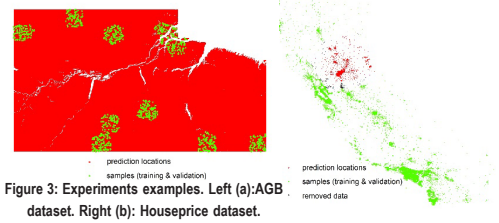

Figure 3: Experiments examples. Left (a):AGB dataset. Right (b): Houseprice dataset.

Random forest was adopted as the spatial ML prediction model. The experiments were repeated 10 times to reduce the random effect. Root-mean-square-error (RMSE) was chosen as the metric of the prediction error.

Compare all CV methods prediction errors.with the standard prediction error, the error differences – d_{CV} – can be calculated. In all experiments, the proposed SP-CV has the lowest d_{CV}, which suggests that SP-CV has the best performances for evaluating the spatial ML model when samples and prediction locations are obviously different.

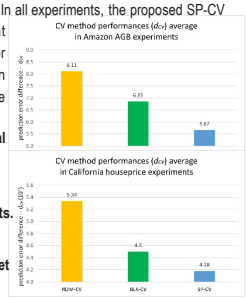

Figure 4: The final results.

Top (a): AGB dataset experiments.

Bottom (b): Houseprice dataset experiments.

Yanwen Wang, y.wang-4@utwente.nl
Mahdi Khodadadzadeh,
Raúl Zurita-Milla

Wang, Y., Khodadadzadeh, M., & Zurita-Milla, R. (2023). Spatial+: A new cross-validation method to evaluate geospatial machine learning models. International Journal of Applied Earth Observation and Geoinformation, 121, 103364. https://doi.org/10.1016/j.jag.2023.103364

Code & data: https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:278047

UNIVERSITY OF TWENTE.

ITC