The effects of a cross-validation approaches on the model transferability of a soybean yield prediction model using UAV-based remote sensing

○Luthfan Nur Habibi (UGSAS, Gifu Univ.), Tsutomu Matsui and Takashi S. T. Tanaka (Fac. Appl. Biol. Sci., Gifu Univ.)

1. Introduction

Yield prediction model is an integral part of precision agriculture, especially in supporting decision making for site-specific crop management practices. Machine learning combined with UAV-based remote sensing data have been widely used in the research community in developing yield prediction model. Robust yield prediction models should have good transferability potential, meaning the model are capable to predict yield in a domain beyond the training regions (extrapolation) rather than only making predictions in the gaps between sampled observations (interpolation). Yield data often have underlying spatial and temporal structure between observations that could undermine model validation. The commonly used cross-validation (CV) technique was reported to lead to underestimation of model prediction error in yield prediction models as the procedure tends to dismiss any spatial structure within the data. Therefore, this study investigated an appropriate CV strategy for establishing transferable UAV-based yield prediction models in different spatial and temporal domains.

2. Materials and Method

Soybean yield data were taken from seven farmers' fields in Gifu Prefecture, Japan from 2018 to 2021. Spectral data from UAV-based imagery were also captured during the full blooming (R2) and initial seed-filling (R5) stages. Vegetation indices selected based on the Index Database (https://www.indexdatabase.de/) were calculated from the spectral data. Yield prediction models were established based on a set of predictor variables from the vegetation indices data. Two subset models were initiated, namely, a model using all vegetation indices data (*all features* model) and model processed with recursive feature elimination (*RFE* model). Three base learner algorithms, including LASSO regression, random forest, and XGBoost, were utilized for developing the model, and a stacked ensemble model formed with these base learners was also implemented. Three data splitting procedures for the CV were compared, including random CV (RCV), cluster-based spatial CV (SCV), and field-specific hold-out CV (LOFOCV). The established models were later tested on an independent field as a test dataset to evaluate the model transferability performance.

3. Results and Discussion

Yield prediction models established using RCV approach performed in a poor accuracy in predicting yield of the independent field. Meanwhile, spatially aware CV, including SCV and LOFOCV, could predicted the yield of the independent field within the ranges of validation accuracy. Spatial CV models were suitable for interpolation and extrapolation implementation, while LOFOCV was specifically appropriate for predicting yield outside the model's spatial domain. These results highlighted the general inability of RCV approach in developing transferrable yield prediction model. These findings was in accordance with Ferraciolli et al. (2019) that emphasized the importance of addressing the spatial dependence of data in developing yield prediction model. Furthermore, the selection of the algorithm and predictor variables used for constructing the models also affected the model transferability. Our results suggest that spatially aware CV should be used as the standard method in constructing transferrable yield prediction model rather than conventional RCV approach.

Reference

Ferraciolli, M.A., Bocca, F.F. and Rodrigues, L.H.A. (2019) Neglecting spatial autocorrelation causes underestimation of the error of sugarcane yield models. *Computers and Electronics in Agriculture*, Vol. 161, pp. 233–240. https://doi.org/10.1016/j.compag.2018.09.003