XXV Congresso de Iniciação Científica da Unicamp

October 18 to 20 Campinas | Brazil

2017

Neglecting spatial autocorrelation leads to underestimation of the error in the development of sugarcane yield models.

Matheus A. Ferracioli*, Felipe F. Bocca, Luiz H. A. Rodrigues.

Abstract

Sugarcane yield models, as most crop predicting models, are subject to the existence of spatial autocorrelation between observations. In this work, we used machine learning techniques to generate sugarcane yield models by splitting blocks of data, grouped by distance, in training and test sets in the cross validation phase, in contrast with separating single observations, as if they were independent. Although models generated using blocks of data led to a better estimation of the error in new data, both approaches generated similar error values.

Key words:

Boosted Regression Trees, Support Vector Machine, Random Forest.

Introduction

In the development of sugarcane yield models, autocorrelation between samples can be detrimental to model assessment. Spatial autocorrelation measures how much the correlation of values of a variable is attributed to its geographic location [1]. Its presence in data used for modeling crop productivity violates the premise of data independence, leading to similar observations being used for training and testing data, which may result in an optimistic prediction error in the final model [2]. A general strategy to increase independence is to split the data into blocks, based on spatial location [3].

In this work, we compared the usual protocol for splitting data sets with the spatial blocks protocol in the development of sugarcane yield models. We used data from three sugarcane mills, with varied areas and field configurations, studying the impact of spatial autocorrelation in the separation of folds for training and testing data in hyperparameter adjusting of the model, adapting a previously presented technique [2].

Results and Discussion

The data set consisted of production, meteorological, input and soil data for three mill areas, located in central Brazil. We used three regression techniques to develop yield prediction models: Boosted Regression Trees (BRT), Support Vector Regression (SVR) and Random Forest (RF).

We compared models developed using blocked data in spatial cross validation (SCV) with models developed using standard cross validation (STN) by measuring each workflow's model performance in a validation set. This set consisted of around 20% of the total of data for each mill, located at least 3km apart from other observations. In SCV, we structured data blocks using k-means, based on each observation geographic location. On average, there were 100 observations per block.

For each technique, we tuned hyperparameters using a random search algorithm [4]. In general, SCV generates less complex models. For USL mill using SVR, using SCV and STN had no impact on the final model. A summary of the results is displayed in Chart 1.

Chart 1. Mean absolute error [$t \cdot ha^{-1}$] in test and validation (Val) sets, for each regression technique.

Mill	WF	BRT		SVR		RF	
		Test	Val	Test	Val	Test	Val
UCR	SCV	13.46	17.88	12.92	18.32	14.05	17.50
UCR	STN	9.34	19.94	9.08	19.68	8.99	18.05
URC	SCV	12.84	16.29	13.35	15.65	12.65	17.44
URC	STN	7.63	16.12	8.71	16.16	7.67	17.37
USL	SCV	14.00	16.10	13.16	15.08	13.44	16.69
USL	STN	6.35	16.51	7.24	15.08	7.03	16.77

Even though the difference between test and validation error is smaller for the SCV workflow, in general, the difference between errors in the validation set for SCV and STN is small.

Conclusions

Our analysis indicated that addressing autocorrelation between observations when developing yield models resulted in a better estimation of the error, while this could not be directly used for creating better models.

Acknowledgement

This work was partially supported by Odebrecht Agro-industrial.

[1] GRIFFITH, D. A. **Spatial Autocorrelation and Spatial Filtering**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003.

[2] RUBIN, G.; BRENNING, A. Data mining in precision agriculture: management of spatial information. In: **Computational intelligence for knowledge-based systems design**. [s.l.] Springer, 2010. p. 350–359.

[3] ROBERTS, D. R.; BAHN, V.; CIUTI, S.; BOYCE, M. S.; ELITH, J.; GUILLERA-ARROITA, G.; HAUENSTEIN, S.; LAHOZ-MONFORT, J. J.; SCHRÖDER, B.; THUILLER, W.; WARTON, D. I.; WINTLE, B. A.; HARTIG, F.; DORMANN, C. F. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. **Ecography**, n. December 2016, p. 1–17, mar. 2017.

[4] BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. **The Journal of Machine Learning Research**, v. 13, n. 1, p. 281–305, 2012.