S1: Algorithm Comparison

This document is part of the supplementary material for the article, "A machine learning-based approach for surface soil moisture estimations with Google Earth Engine" by Greifeneder, Notarnicola, and Wagner.

The selection of Gradient Boosting Regression Trees (GBRT) was based on previous experience[1] and practical issue related to the specific requirements of the work.

In a first step, a comparison of GBRT with other algorithms (Random-Forest (RF), AdaBoost (ADB), Support-Vector-Regression (SVR)) was carried out where GBRT and RF showed the best performance of prediction accuracy, in terms of R^2 as well as the Root-Mean-Squared-Error (RMSE). Another significant and more practical advantage of GBRT was the significantly reduced computational effort (compared to the other methods) for both training and prediction. During the development, it was possible to spend more effort on algorithm optimization. Moreover, GRBT has proved to manage well different types of input data (continuos and categorial).

Each method was tested based on the training and validation framework described in the article in section 3.3. The table below shows a summary of the results of this comparison:

Method	LOGO R^2	LOGO RMSE [m^3m^{-3}]	Test-set R^2	Tet-set RMSE [m^3m^{-3}]	Training time* [sec.]	Prediction time (testset) [sec.]
GBRT	**0.73**	**0.05**	**0.81**	**0.04**	**5.20**	**0.05**
SVR	0.68	0.06	0.77	0.05	363.62	7.38
RF	0.70	0.05	0.81	0.05	340.99	3.07
ADB	0.38	0.08	0.41	0.08	57.46	0.55

**Excluding LOGO-CV and feature selection.*

It shows that GBRT and RF perform very well; both cross-validation and test scores are almost identical. There are significant differences between the two methods regarding the time required for training and prediction. The choice of the best hyper-parameters can explain this effect:

Algorithm	Hyper-parameters	
GBRT	Learning-rate	0.1
	Number of estimators	100
	Fraction of samples used for fitting	0.5
	Maximum depths of individual regression trees	10
	Early stopping after n iterations with no change	10
SVR	Kernel	Radial-Basis-Function

[1] Luca Pasolli et al., "Estimation of Soil Moisture in Mountain Areas Using SVR Technique Applied to Multiscale Active Radar Images at C-Band," *Selected Topics in Applied Earth Observations and Remote Sensing* 8, no. 1 (2015): 262–83, https://doi.org/JSTARS.2014.2378795; Felix Greifeneder, Claudia Notarnicola, and Wolfgang Wagner, "Using Machine Learning and SAR Data for the Upscaling of Large Scale Modelled Soil Moisture in the Alps," in *11th European Conference on Synthetic Aperture Radar, EUSAR 2016* (Hamburg: VDE, 2016), 1108–11; Iftikhar Ali et al., "Review of Machine Learning Approaches for Biomass and Soil Moisture Retrievals from Remote Sensing Data," *Remote Sensing* 7, no. 12 (2015): 16398–421, https://doi.org/10.3390/rs71215841.

	γ	0.316
	C	1
	ε	0.01
RF	Number of estimators	1000
	Maximum depths of individual regression trees	no limit
ADB	Learning rate	0.01, 0.1, 0.2
	Number of estimators	500

The RF algorithm seems to favour more complex structures as the best results were achieved with a high number of fully grown base estimators. The scatterplots in Figure 1 highlight the similarity between the prediction accuracy of GBR and RF. Moreover, it shows that SVR performs very similar with a slightly higher level of noise.

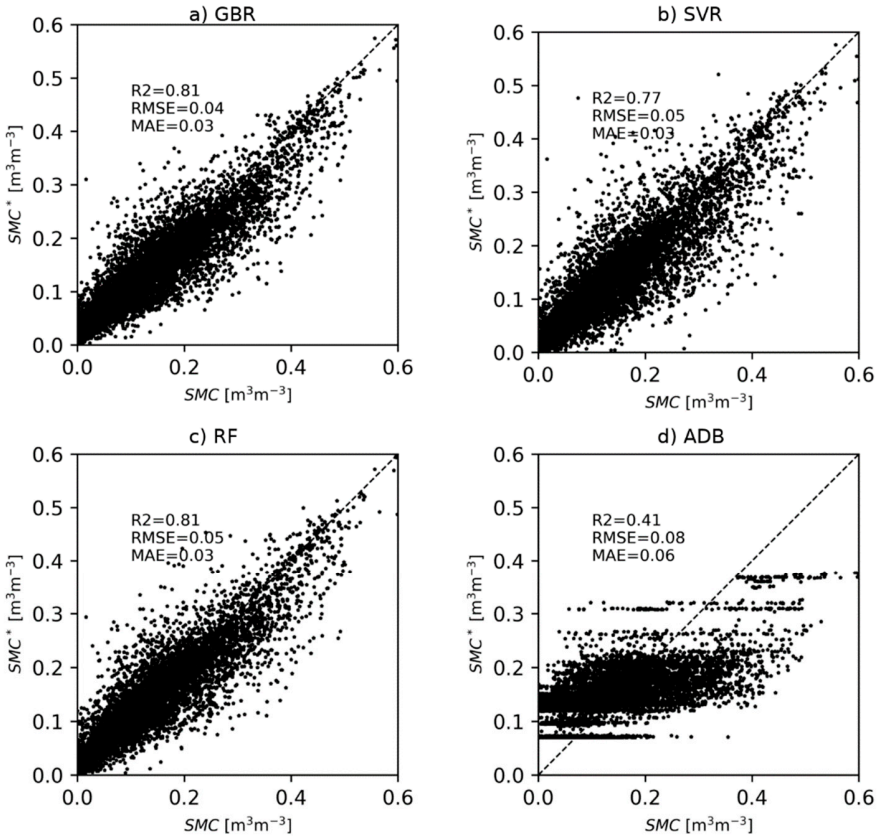

Figure 1: Comparison between true (x-axis) and estimated (y-axis) SMC based on the four different ML algorithms.