IMPROVING DIGITAL SOIL MAPS FOR SITE-SPECIFIC SOIL FERTILITY MANAGEMENT USING FEATURE SELECTION

C. Ferhatoglu and B.A. Miller
Iowa State University, Ames, IA
canerf@iastate.edu (515) 708-9712

ABSTRACT

In this study, the effectiveness of six types of FS methods from four categories (filter, wrapper, embedded, and hybrid) were compared. These FS algorithms chose relevant covariates from a set of 1049 environmental covariates for predicting five soil fertility properties in ten fields, in combination with ten different ML algorithms. The resulting model performance was compared by three different metrics (R^2 of 10-fold cross validation (CV), robustness ratio (RR; developed in this study), and independent validation with Lin's concordance correlation coefficient (IV-CCC)). Wrapper (BorutaShap) and embedded (Lasso-FS, Random forest-FS) methods with decision-tree based ML algorithms usually led to the optimal models. FS improved CV, RR, and IV-CCC compared to the models built without FS for most fields and soil properties. Wrapper (BorutaShap) and embedded (Lasso-FS, Random forest-FS) methods usually led to the optimal models. The filter-based ANOVA-FS method mostly led to overfit models, especially for fields with smaller sample quantities. Decision-tree based models were usually part of the optimal combination of FS and ML. Considering RR helped identify optimal combinations of FS and ML that can improve the performance of DSM compared to models produced from full covariate stacks. FS can assist building better predictive soil models to create better digital soil maps, which in return can improve the farm management (e.g., fertilization, liming, and manuring).

Introduction

Digital soil mapping (DSM) has been widely used to map various soil properties and classes for the last few decades [1]. A strategy for DSM is the process of using predictive statistical models (e.g., machine learning (ML)), utilizing the relationships between georeferenced soil lab data and environmental predictors (aka covariates) [2]. Performance of ML relies heavily on the covariates used to represent true soil-landscape relationships [3]. Thus, covariate (aka feature) selection is an important aspect for DSM. Objectives of FS, as a data pre-processing strategy, include building simpler models, reducing the effect of the curse-of-dimensionality, and improving prediction performance [4]. Previous studies on FS in DSM have focused on less dynamic and less heavily managed soil properties (e.g., soil classes and soil organic matter) [5, 6, 7] compared to soil fertility properties (e.g., soil-test-P and -K). In our study, the effectiveness of six types of FS methods from four categories (filter, wrapper, embedded, and hybrid) were compared. These FS methods chose relevant covariates from a set of 1049

environmental covariates for predicting five dynamic soil fertility properties in ten fields, in combination with ten different ML algorithms.

Methodology

Study Area and Input Datasets

The study sites were ten agricultural fields located within a research farm near Ames, Iowa, USA. A total of 992 soil samples, collected from a depth of 0–15 cm between 2018 and 2020 were used in this study. All samples from each field (A-J) were collected on a single date (Figure 1). Samples were analyzed for nitrate-nitrogen (NO_3^-), soil-test phosphorus (P), soil-test potassium (K), buffer pH (BpH), soil organic matter (SOM). The covariate set included 1049 spatial variables from digital terrain analysis (DTA) and spectral bands of RS (aerial and satellite imagery). Same covariates were used as in Ferhatoglu and Miller [8]. All covariates were resampled to 3 m spatial resolution and spatially aligned. Environmental covariate values were then paired with soil lab data at the sampling locations and transferred to a csv format for FS process. Selected covariates were used in ML algorithms to create predictive soil models.

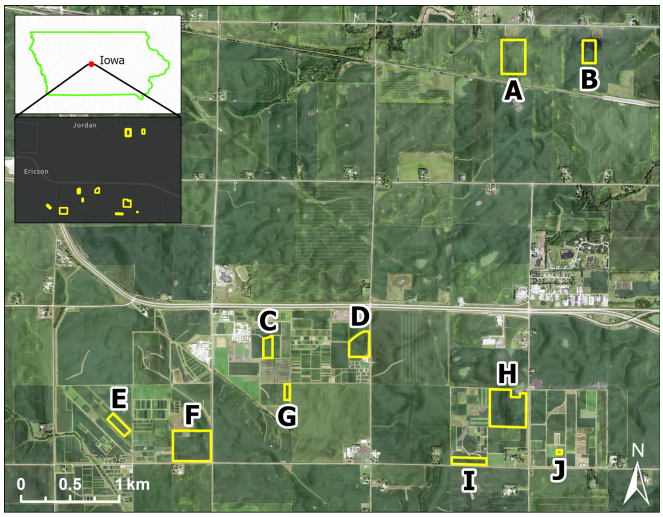


Figure 1. The study fields (A–J). The size of the fields ranged from 0.4 ha to 13.1 ha. Soil samples were collected from the fields with a grid-sampling design.

Feature Selection and Modelling

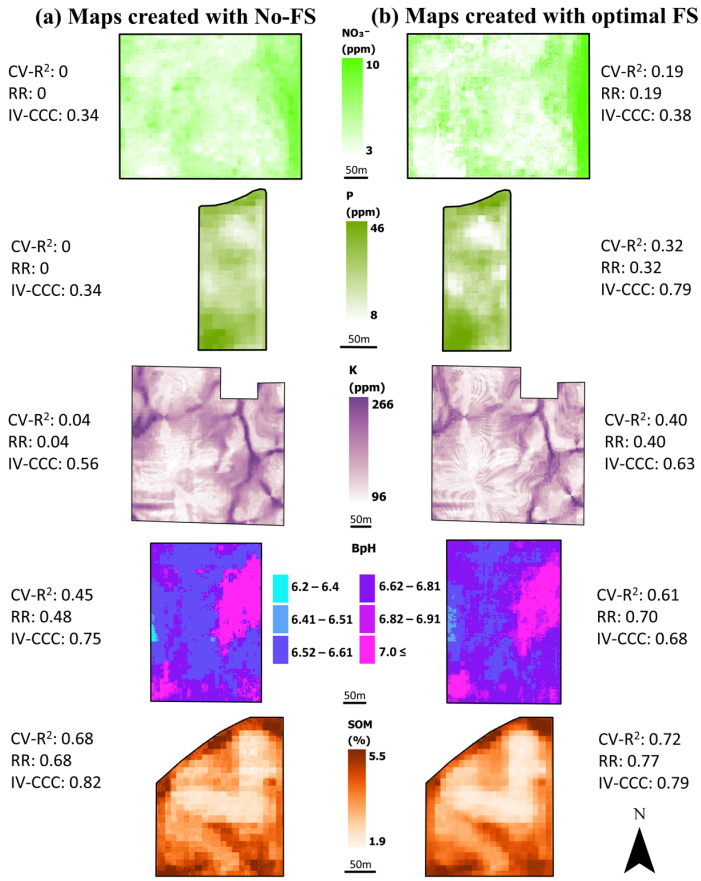
Six different FS methods were applied to identify relevant covariates: (1) Combined-filter-FS, (2) ANOVA-FS, (3) BorutaShap-FS, (4) Random Forest FS (RF-FS), (5) Lasso-FS, and (6) Hybrid-FS. Combined-filter-FS and ANOVA represented filter FS strategies while BorutaShap-FS, RF-FS, and Lasso-FS were embedded FS strategies. Hybrid-FS method represented hybrid FS strategy. More details about these FS methods can be found in in Ferhatoglu and Miller [8]. Using each covariates selected, ten ML algorithms (Lasso regressor, support vector regressor (SVR) with polynomial kernel, and multi-layer perceptron regressor (MLP), random forest (RF) regressor, extra-trees-regressor (ETR), CatBoost, AdaBoost, LightGBM, gradient boosting (GB), a voting regressor based on the nine ML algorithm above) from scikit-learn [9] with default parameters were used to model soil properties and compare FS methods by their interaction with ML algorithms. Voting regressor ranked predictions of the nine ML algorithms based on R^2 score on the validation set (IV: 20%) to weight respective model predictions in the final prediction.

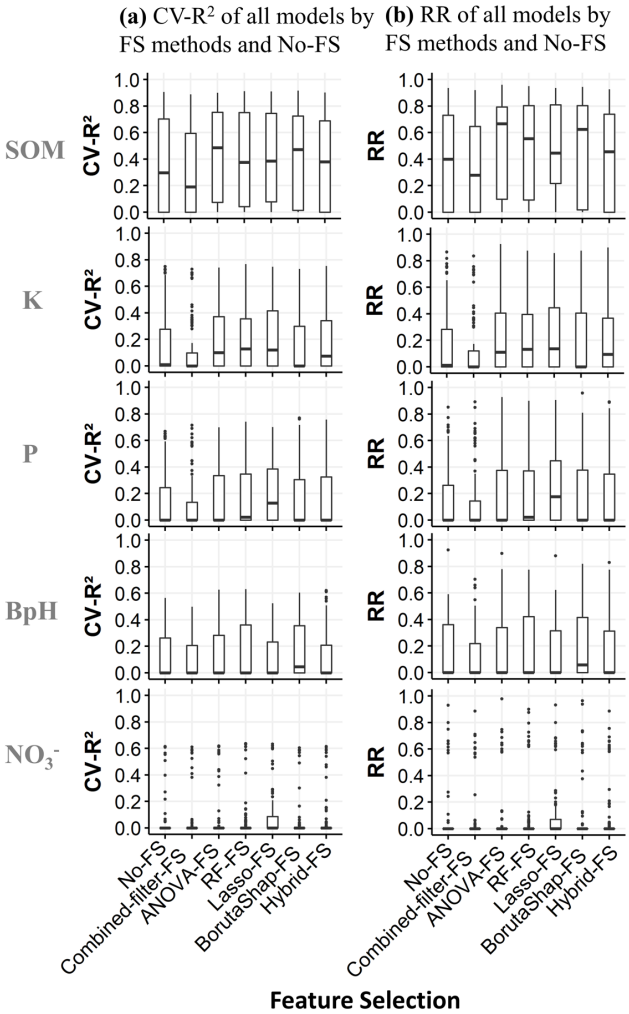
Experimental Design & Evaluation

Firstly, six FS methods were applied to the covariate stack for each target soil property and sample set (i.e., individual fields plus all fields combined. Full covariate stack was the control treatment. Models were then built from ten ML algorithms for each of those treatments, yielding 70 ML models per sample set and soil property. To simplify the evaluation process and interpretation of results, three stages of evaluation metrics were applied to identify the highest performing models. First, the models were evaluated by the R^2 of 10-fold cross-validation (CV). Ten models with the highest CV-R^2 score were selected for subsequent analysis based on a new metric introduced in this study to measure the robustness of the model (RR: R^2 of 10-fold CV/ R^2 of goodness-of-fit, which was 80% of samples). The five models exhibiting a likelihood to be robust were subsequently evaluated for prediction performance. The model with the highest CCC [10] based on IV was determined to be the optimal model for the respective field and soil property.

Results

Overall, FS methods consistently reduced the covariate stack to less than half of the original quantity. The largest to lowest reduction in covariate stack size was made by as follows: BorutaShap-FS > Lasso-FS ≈ Hybrid-FS > ANOVA-FS, Combined-filter-FS ≈ RF-FS. Models built from covariate stacks reduced by FS mostly performed better in CV than those built without FS for all properties (Figure 2A). NO_3^- was particularly challenging for producing predictive models, where the median CV-R^2 for No-FS and all FS methods was zero. Patterns between CV and RR were similar, which suggests stronger CV performance could be connected to RR performance (Figure 2A). In the second step of the evaluation, models from some FS methods remained competitive, while others were more often cut due to lower performance in terms of RR (data not shown). Although ANOVA-FS had the highest frequencies in the first step for K, BpH and NO_3^-, the difference between ANOVA-FS and other FS methods became smaller. Models produced from covariate stacks reduced by FS methods outperformed models built from covariate stacks without FS in most cases for independent validation (data not shown). IV-CCC scores for the final models were higher than full models for nine (SOM), eight (K), six (P), five (BpH), and all (NO_3^-) sample sets. BorutaShap-FS, ANOVA-FS, RF-FS, Lasso-FS were commonly optimal FS methods among the sample sets. Digital soil maps developed with the full covariate stack tended to be smoother than the maps created by using FS with exceptions in some fields (e.g., SOM map in field D) (Figure 2B). Despite differences observed in the evaluation of the models' prediction performance, all maps produced from covariate stacks reduced by FS had similar patterns to their No-FS counterparts. Figure 2B presents some examples comparing maps developed with and without FS.


(a) Maps created with No-FS
(b) Maps created with optimal FS
CV-R²: 0
RR: 0
IV-CCC: 0.34
NO₃⁻ (ppm)
10
3
50m
CV-R²: 0.19
RR: 0.19
IV-CCC: 0.38
CV-R²: 0
RR: 0
IV-CCC: 0.34
P (ppm)
46
8
50m
CV-R²: 0.32
RR: 0.32
IV-CCC: 0.79
CV-R²: 0.04
RR: 0.04
IV-CCC: 0.56
K (ppm)
266
96
50m
CV-R²: 0.40
RR: 0.40
IV-CCC: 0.63
CV-R²: 0.45
RR: 0.48
IV-CCC: 0.75
BpH
6.2 – 6.4
6.41 – 6.51
6.52 – 6.61
6.62 – 6.81
6.82 – 6.91
7.0 ≤
50m
CV-R²: 0.61
RR: 0.70
IV-CCC: 0.68
CV-R²: 0.68
RR: 0.68
IV-CCC: 0.82
SOM (%)
5.5
1.9
50m
CV-R²: 0.72
RR: 0.77
IV-CCC: 0.79
N


(a) CV-R² of all models by FS methods and No-FS
(b) RR of all models by FS methods and No-FS
SOM
K
P
BpH
NO3-
CV-R²
RR
0.0
0.2
0.4
0.6
0.8
1.0
No-FS
Combined-filter-FS
ANOVA-FS
RF-FS
Lasso-FS
BorutaShap-FS
Hybrid-FS
Feature Selection

(A) (B)

Figure 2. (A) Comparisons of performance for models produced from the different FS treatments, evaluated by (a) CV-R^2 and (b) RR. Except for Combined-filter-FS, FS methods consistently outperformed the No-FS treatment. For the most part, evaluation of models by RR followed similar patterns to those of CV-R^2, which suggests the higher CV-R^2 may also tend to have smaller differences between the goodness of fit R^2 and CV-R^2. (B) Examples of maps created by the optimal models built from covariate stacks with (a) No-FS and (b) FS. Applying FS generally led to less smooth maps compared to maps created with full covariate stacks. However, there were exceptions such as the SOM map shown in these examples. Maps shown reflect soil fertility levels present on the sampling dates: NO_3^- for field F (8 June 2019), P for field C (12 July 2019), K for field H (25 June 2018), BpH for field A (29 June 2020), and SOM for field D (16 July 2019).

Conclusions

Models produced from covariate stacks reduced by FS methods were less likely to be overfit and tended to have better performance in IV-CCC. Although there was no single optimal FS method among sample sets or soil properties, wrapper and embedded FS strategies produced the optimal model more frequently than the hybrid and filter FS strategies. Given the economic and environmental promise of precision agriculture, combined with the increasingly finer temporal resolution of remote sensing, there is an opportunity to apply these methods to provide farmers with better soil fertility maps.

ACKNOWLEDGMENTS

This work would not have been possible without the cooperation of several research groups within the Department of Agronomy at Iowa State University. Special thanks to the A.K. Singh Soybean Breeding and Genetics Group and the Geospatial Laboratory for Soil Informatics. Support for Caner Ferhatoglu was provided by the Department of Agronomy at Iowa State University.

REFERENCES

Minasny, B.; McBratney, Alex.B. Digital Soil Mapping: A Brief History and Some Lessons. *Geoderma* 2016, *264*, 301–311, doi:10.1016/j.geoderma.2015.07.017.

McBratney, A.B.; Mendonça Santos, M.L.; Minasny, B. On Digital Soil Mapping. *Geoderma* 2003, *117*, 3–52, doi:10.1016/S0016-7061(03)00223-4.

Xiong, X.; Grunwald, S.; Myers, D.B.; Kim, J.; Harris, W.G.; Comerford, N.B. Holistic Environmental Soil-Landscape Modeling of Soil Organic Carbon. *Environmental Modelling & Software* 2014, *57*, 202–215, doi:10.1016/j.envsoft.2014.03.004.

Flynn, T.; de Clercq, W.; Rozanov, A.; Clarke, C. High-Resolution Digital Soil Mapping of Multiple Soil Properties: An Alternative to the Traditional Field Survey? *South African Journal of Plant and Soil* 2019, *36*, 237–247, doi:10.1080/02571862.2019.1570566.

Behrens, T.; Scholten, T. Digital Soil Mapping in Germany—a Review. *Journal of Plant Nutrition and Soil Science* 2006, *169*, 434–443, doi:10.1002/jpln.200521962.

Campos, A.R.; Giasson, E.; Costa, J.J.F.; Machado, I.R.; Silva, E.B. da; Bonfatti, B.R. Selection of Environmental Covariates for Classifier Training Applied in Digital Soil Mapping. *Rev. Bras. Ciênc. Solo* 2019, *42*, doi:10.1590/18069657rbcs20170414.

Luo, C.; Zhang, X.; Wang, Y.; Men, Z.; Liu, H. Regional Soil Organic Matter Mapping Models Based on the Optimal Time Window, Feature Selection Algorithm and Google Earth Engine. *Soil and Tillage Research* 2022, *219*, 105325, doi:10.1016/j.still.2022.105325.

Ferhatoglu, C.; Miller, B.A. Choosing Feature Selection Methods for Spatial Modeling of Soil Fertility Properties at the Field Scale. *Agronomy* 2022, *12*, 1786, doi:10.3390/agronomy12081786.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-Learn: Machine Learning in Python. *the Journal of machine Learning research* **2011**, *12*, 2825–2830.

Lin, L.I.-K. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* **1989**, *45*, 255–268, doi:10.2307/2532051.