IOWA STATE UNIVERSITY
Department of Agronomy

Caner Ferhatoglu & Bradley A. Miller

Improving Digital Soil Maps for Site-specific Soil Fertility Management Using Feature Selection

Introduction

- The promise of precision agriculture has not been fully met because, in part, the challenges of obtaining the needed information at fine enough resolution. Once fully realized, precision agriculture offers benefits for crop yield, input use efficiency, and environmental sustainability. Soil fertility maps are essential for informing soil fertility management. However, current methods for mapping soil fertility are not efficient enough to produce the quality of maps needed to fully support future variable rate technologies.
- **Digital soil mapping (DSM)** is an attractive option to manage site-specific soil fertility thanks to its capabilities for creating highly accurate and reliable soil maps with fine spatial resolution (Iticha and Takele, 2019) (Figure 1).

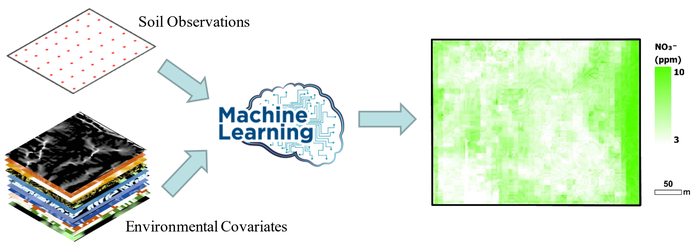

Figure 1. Digital soil mapping (DSM) refers to the creation of soil maps based on statistical learning algorithms (e.g., machine learning (ML)). These algorithms are used to identify the relationships between observed soil data and environmental covariates at soil sampling locations. Then, these relationships are used to make predictions at unsampled locations to create a soil map.

- Growing availability of environmental covariates due to advancements in remote sensing (RS) technologies is making it challenging to select and focus on the most important covariates. Using only relevant covariates in ML models is crucial where the curse of dimensionality can negatively impact the modeling accuracy and reliability with small datasets (Figure 2).

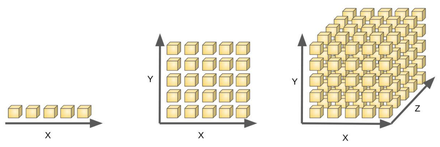

Figure 2. Curse of dimensionality indicates that "the quantity of samples needed to estimate an arbitrary function with a given level of accuracy grows exponentially with respect to the quantity of potential covariates (i.e., dimensionality) of the function" (Bellman, 1957).

Objectives

- Evaluate the response of model performance to feature selection (FS) under different ML algorithms for the spatial modelling of five dynamic soil fertility properties.
- Investigate the effect of sample quantity (i.e., the original sample sets, which were based on separate fields and all fields combined) for spatial modeling.

Materials & Methods

- Study was conducted in ten central Iowa fields (Figure 4). A total of 992 soil samples collected from a depth of 0–15 cm between 2018 and 2020 were used in this study.
- Samples were analyzed for nitrate-nitrogen (NO_3^-) by flow injection method, soil-test phosphorus (P) by Bray-1, soil-test potassium (K) by neutral ammonium acetate method, buffer pH (BpH), and soil organic matter (SOM) by loss on ignition (LOI). All soil samples were collected using a grid sampling design.
- A total of 1,049 environmental variables were used as potential covariates. The covariates were from digital terrain attributes based on 3 m and 10 m digital elevation models (DEMs) and time-series satellite (Sentinel-2) & aerial imagery (NAIP: National Agriculture Imagery Program) (Figure 3).

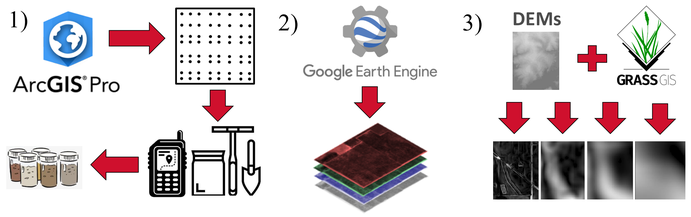

Figure 3. Operations to create input datasets (i.e., soil data and environmental covariates). 1) creation of soil data. 2) downloading spectral bands and their corresponding vegetation indices based on time-series satellite and aerial imagery products using Google Earth Engine platform. 3) creation of digital terrain attributes with varying analysis scales in GRASS GIS based on 3 m and 10 m DEMs from Iowa's LiDAR program (2010).

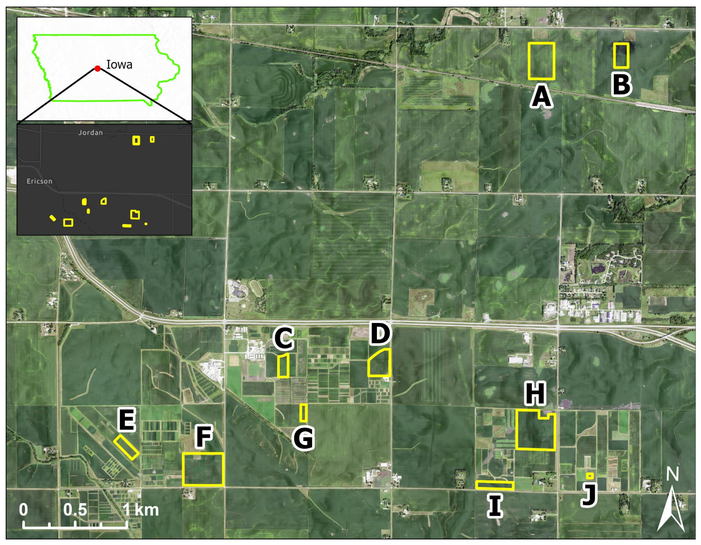

Figure 4. Map of the study fields (A–J). Size of the fields ranged from 0.4 ha to 13.1 ha. Soil samples were collected from these fields with varying grid-sample spacings (5 m to 37.5 m grids).

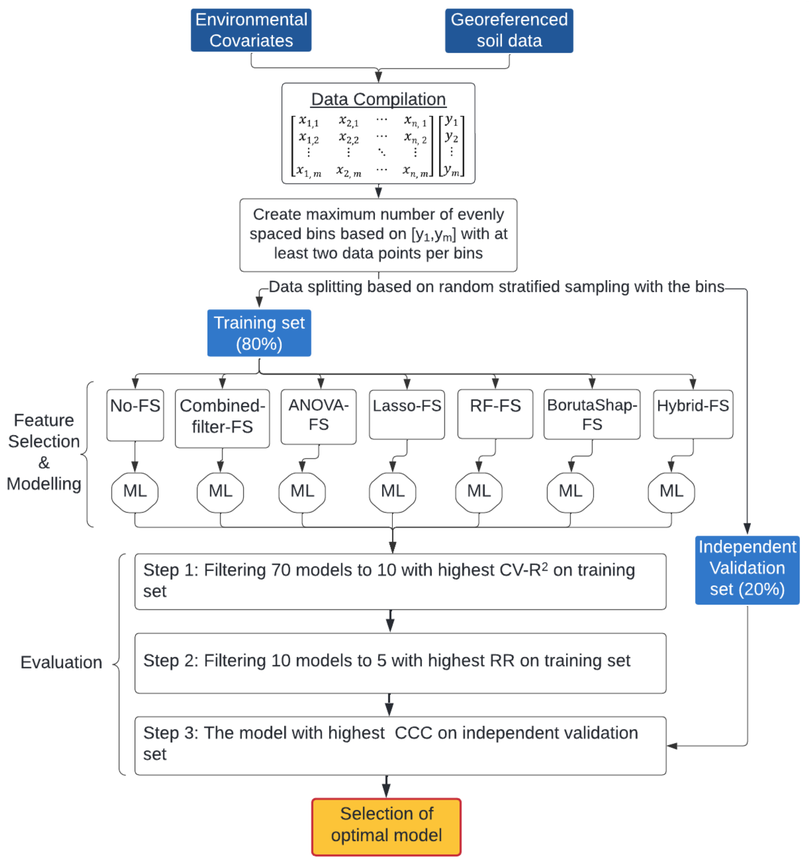

Figure 5. Overall workflow and evaluation process for selecting optimal models from the combinations of FS and ML algorithms tested. This process was applied to each sample set, which included five soil properties for ten individual fields and all fields combined. The reason for using three different evaluation steps was to sequentially evaluate characteristics of model quality. Step 1 tested the models' dependence on locations used for model training with cross-validation-R^2 (CV-R^2)). Step 2 tested the stability of results when different training sets were used with the robustness ratio (RR): CV-R^2/ goodness-of-fit). Finally, step 3 tested the accuracy of the predictions in the resultant soil map with an independently held-out sample set (20% of all samples) with Lin's concordance correlation coefficient (CCC)). For each of these metrics, higher values indicate better models and resultant maps.

Acknowledgements

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. Support for Caner Ferhatoglu was provided by the Department of Agronomy at Iowa State University.

Results

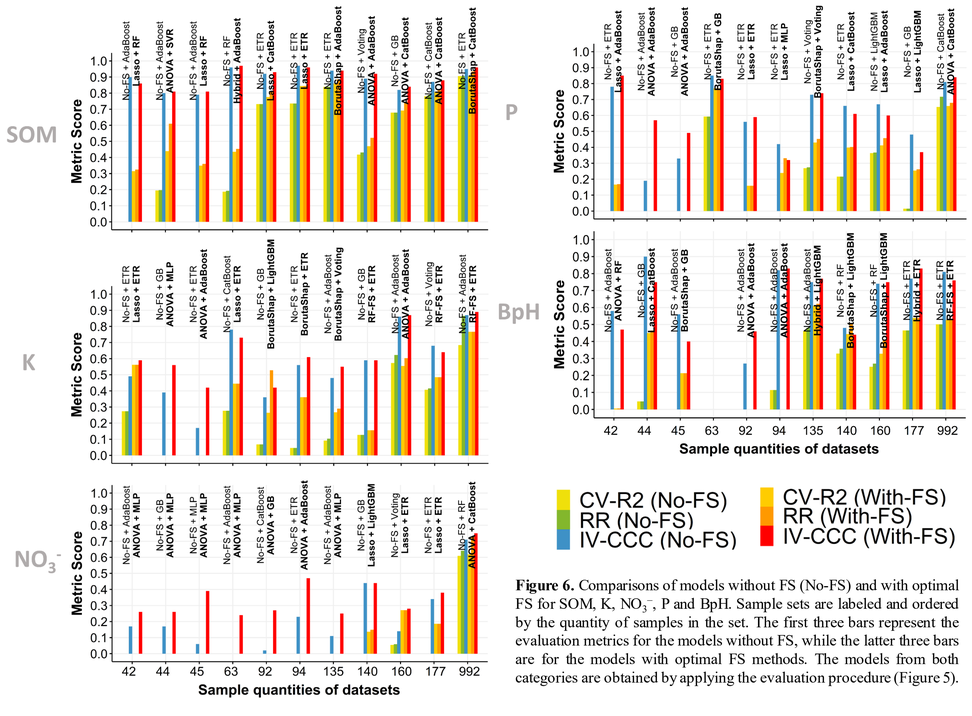

Figure 6. Comparisons of models without FS (No-FS) and with optimal FS for SOM, K, NO_3^-, P and BpH. Sample sets are labeled and ordered by the quantity of samples in the set. The first three bars represent the evaluation metrics for the models without FS, while the latter three bars are for the models with optimal FS methods. The models from both categories are obtained by applying the evaluation procedure (Figure 5).

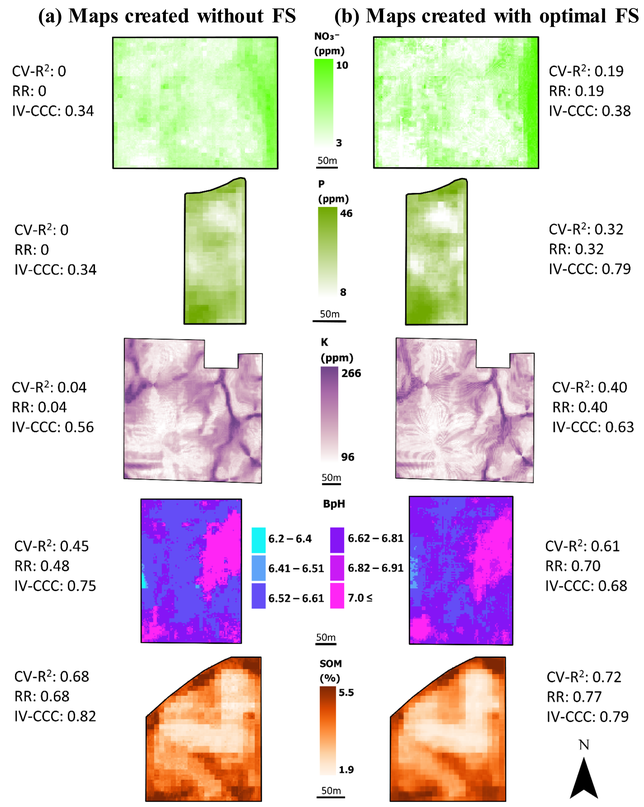

Figure 7. Examples of maps created by the optimal models built from covariate stacks with (a) No-FS and (b) FS. Applying FS generally led to less smooth maps compared to the maps created with full covariate stacks. However, performance across all metrics usually increased for the maps produced with FS. 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

- FS improved model robustness (RR) and prediction accuracy of soil fertility maps. Wrapper and embedded FS strategies with tree-based ML produced the optimal models/maps.
- IV-CCC results were higher for maps based on more samples in a sample set, but this relationship was weakly correlated for each of the FS methods.
- Our methods can significantly improve the reliability and accuracy of soil fertility mapping, which can help farmers to optimize their crop yield and input use efficiency.

Geospatial Laboratory for Soil Informatics