Exploring Spatial Cross-Validation (CV) Techniques for Enhanced Crop Yield Prediction Models

Daniel Chvat[1], Landon Dierkes[2], Grace Abraham[3], Kristen North[4], Andy Wang[5], Gabriel Sipos*, Grace McDonnell*, Dr. Papia Rozario*, Dr. Rahul Gomes[+]

[1]University of California- Los Angeles, Computer Science, [2]Madison Area Technical College, Computer Science, [3]North Carolina State University, Computer Science, [4]El Camino College, Computer Science, [5]University of Wisconsin-Madison, Mathematics and Computer Science, *University of Wisconsin-Eau Claire, Geography and Anthropology, [+]University of Wisconsin-Eau Claire, Computer Science

Introduction:

- Variable Rate Agriculture is the intelligent application of agricultural techniques to reduce waste and improve efficiency.
- Effective utilization of VRA, accurate measurements of soil conditions are required.
- This process is traditionally done manually, but Remote sensing techniques provide an alternative to in-situ testing.
- Unmanned Aerial Vehicles (UAV's) can scan large areas of land using Lidar based remote sensing. interpretation of this data could provide an avenue to accurately predict soil conditions without the need for in-situ soil sampling.

Fig. 1: Geospatial Data was collected using DJI Phantom 4 UAV

- Machine Learning models interpret complex data quickly & efficiently.
- One of the largest problem in applying ML algorithms to this type of data is the high potential for model over fitting.
- Our study is focused on exploring the ability of different interpolation techniques and spatial cross validation to improve accuracy of machine learning models that train on spatial data.

Study Area:

- Data was collected at the University of Minnesota Southern Research and Outreach Center (44°04'41.0"N 93°31'29.0"W).
- From 2020 – 2022, hybrid maize crop was grown

Southern Minnesota's Geography:

- Predominately flat land
- Warm/humid continental climate (Köppen class: Dfa)

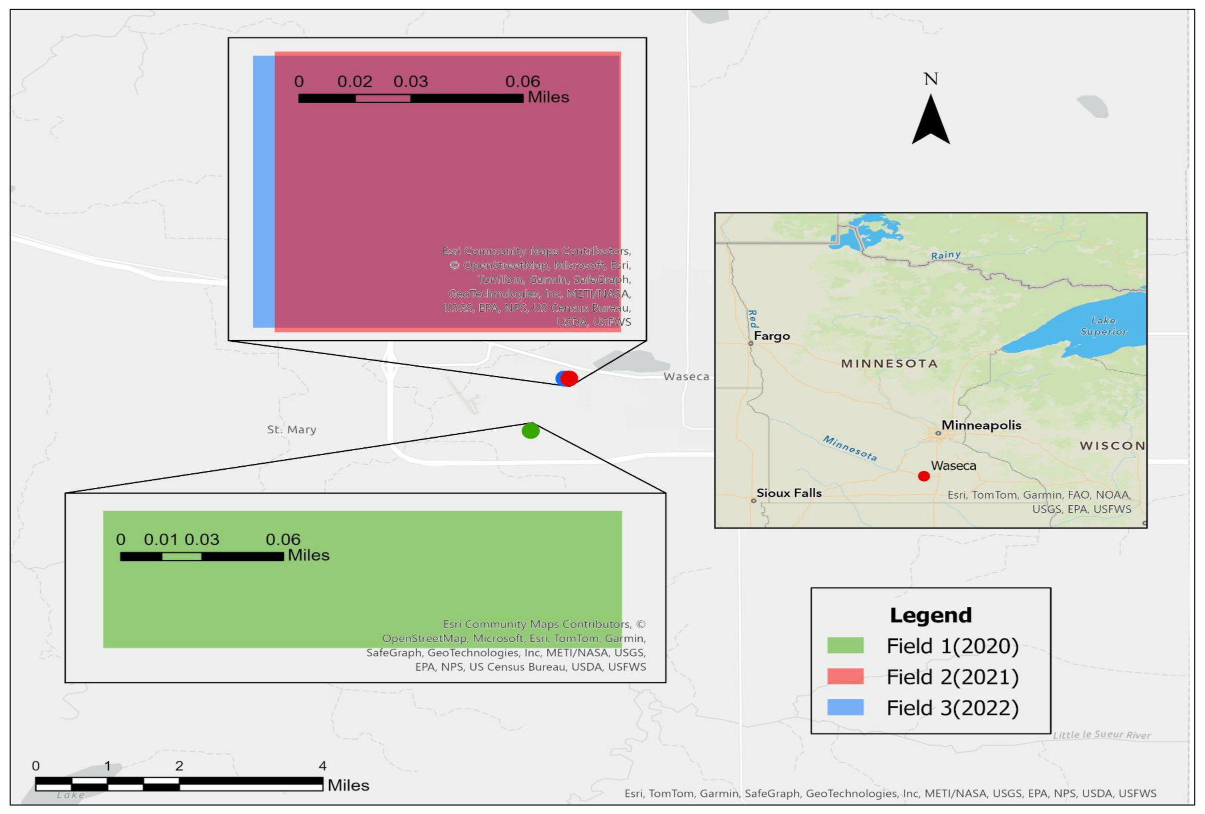

Fig. 2: Study area map

Data Collected:

- Ortho-mosaics
- Digital elevation models
- Plot boundary shape files
- RGB Vegetative Indices
- Weather and Soil data
- Extracted plant heights
- Harvested crop yield dry mass
- Manual height measurements

Fig. 3: Point data collected in 2020

Methods:

Four-Fold CV:

- Dataset is split into training and validation sets
- Four folds created by alternating data in training and validation sets

Random CV:

- Training - Validation sets randomly sampled using 70/30 train-validate split

GroupKFold CV:

- Averaged bands are classified through K-Means in ArcGIS
- One cluster is assigned as validation, rest is used for training

Spatial+ CV:

- Each feature in averaged bands is classified through K-Means
- Individual feature clustering labels are combined to a clustering label
- One cluster is assigned as validation, rest is used for training

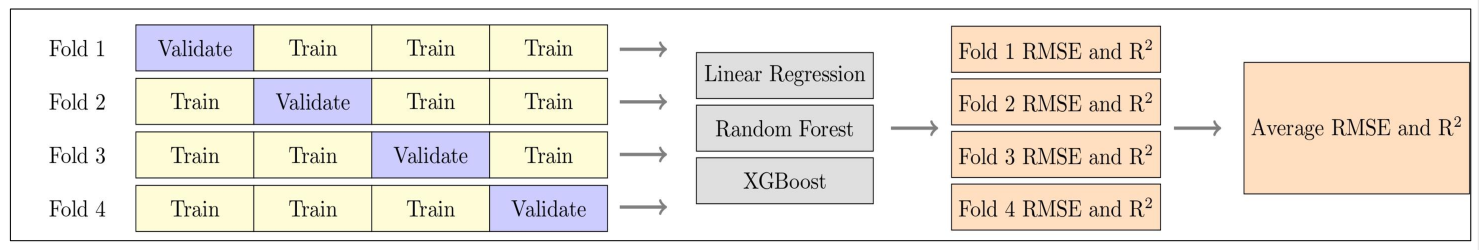

Fig. 4: Illustration of four-fold cross validation

Feature Extraction using ArcGIS Pro:

- Field rasters imported from online repository
- Spatial autocorrelation used to predict attribute values in regions of interest
- Polygons were generated and features extracted using extract by mask tool
- Extracted the feature value spreadsheet across all attributes and years
- Added a coordinate system for spatial data interpretation outside the ArcGIS Pro environment using Python 3.X

Index:	Meaning:
Red	Red RGB Index
Green	Green RGB Index
Blue	Blue RGB Index
BI	Brightness Index
GLI	Green Leaf Index
RGRDI	Normalized Green-Red Difference Index
VARI	Visible Atmospherically Resistant Index
BGI	Blue-Green Index
ExG	Excess Green Index
ExR	Excess Red Index
ExB	Excess Blue Index
ExGR	Excess Green-Red Index
MGRVI	Modified Green Red Vegetation Index
RGBVI	Red Green Blue Vegetation Index
GRRI	Green Red Ratio Index
VEG	Vegetation Index

Table 1: Indices used

Feature Importance with Linear Regression:

- Simple Linear Regression model was used to determine how well each feature predicted mean dry yield. Examples shown below

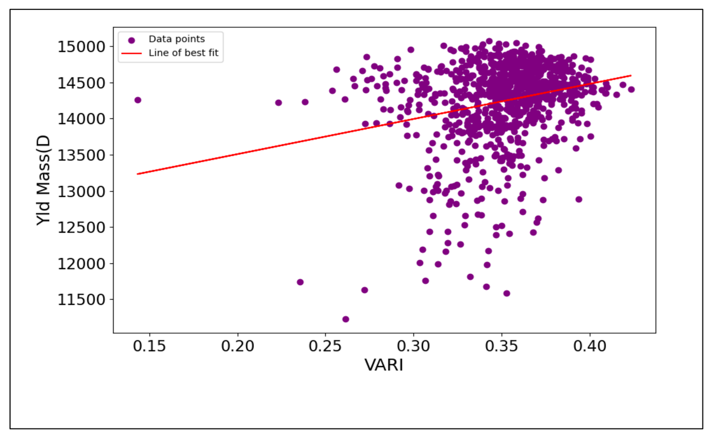

Fig. 5a: Scatterplot correlation of VARI to Mean Yield on 6/15/2022

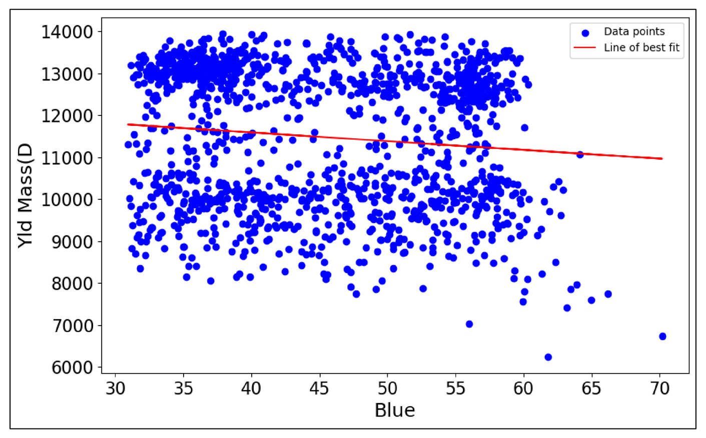

Fig. 5b: Scatterplot correlation of Blue to Mean Yield on 6/23/2020

Feature Importance with Random Forest:

- Random Forest model fitted on a 70/30 train-test split generated feature important values for vegetative indices

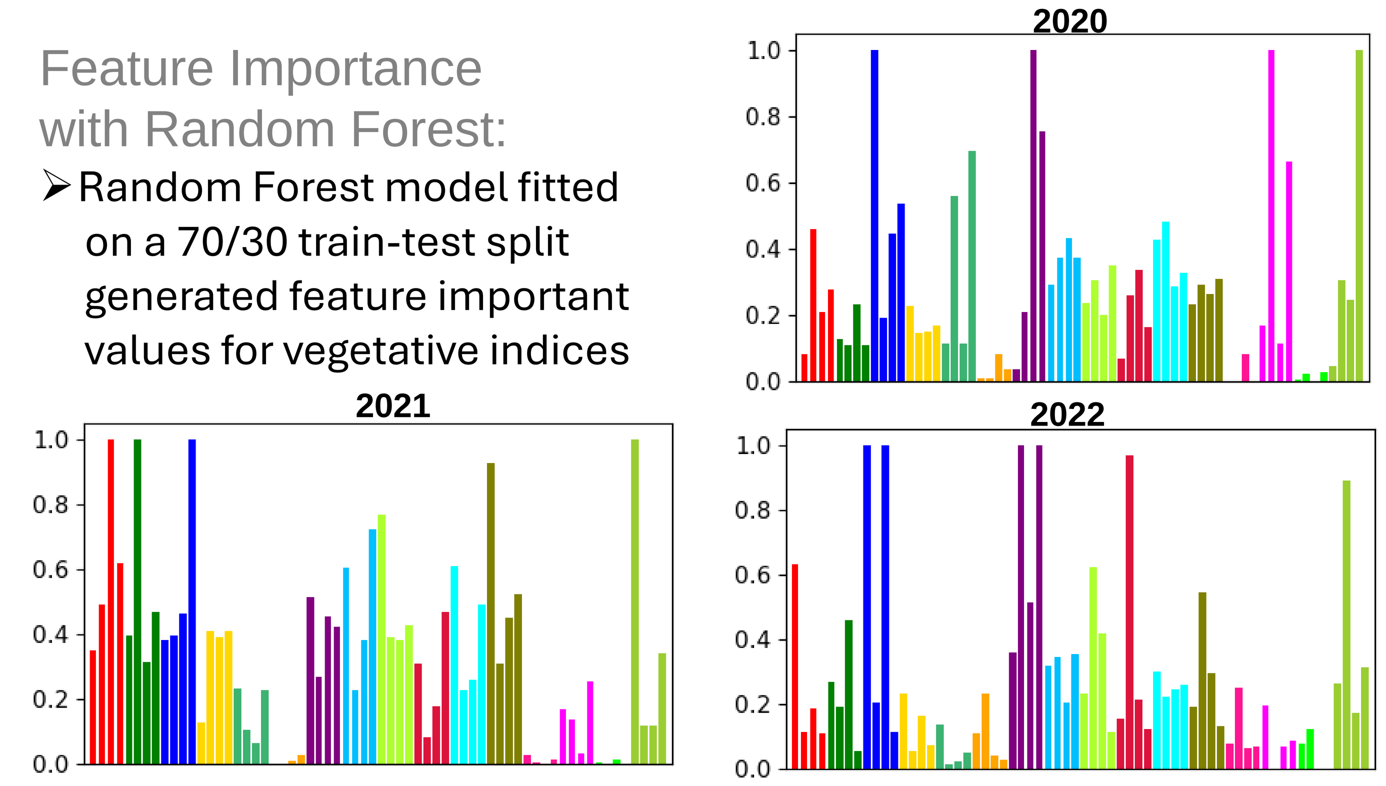

Fig. 6: Normalized mean feature importance histograms for each year.

Results:

Preliminary Feature CV Results:

Model	Random		GroupKFold		Spatial+	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
2020						
Linear Regression CV	0.540	0.145	−0.978	0.206	0.008	0.190
Linear Regression TEST	−0.415	0.255	−1.912	0.355	−1.286	0.316
Random Forest CV	0.825	0.089	−0.513	0.169	0.527	0.130
Random Forest TEST	−0.713	0.280	−0.220	0.233	−0.386	0.249
XGBoost CV	0.801	0.095	−0.506	0.174	0.455	0.138
XGBoost TEST	−0.613	0.272	−0.316	0.241	−0.394	0.251
2021						
Linear Regression CV	0.515	0.120	−0.387	0.177	−0.211	0.171
Linear Regression TEST	−3.318	0.348	−3.873	0.372	−4.614	0.392
Random Forest CV	0.710	0.093	0.244	0.134	0.522	0.114
Random Forest TEST	0.394	0.134	0.234	0.149	0.327	0.141
XGBoost CV	0.700	0.094	0.141	0.144	0.481	0.119
XGBoost TEST	0.303	0.144	0.351	0.139	0.347	0.139
2022						
Linear Regression CV	0.273	0.133	−1.129	0.163	−0.496	0.174
Linear Regression TEST	−5.860	0.389	−4.930	0.356	−5.429	0.373
Random Forest CV	0.569	0.103	−0.656	0.156	0.141	0.126
Random Forest TEST	0.171	0.137	−0.034	0.153	0.089	0.144
XGBoost CV	0.533	0.107	−1.193	0.178	−0.022	0.133
XGBoost TEST	0.159	0.138	−0.040	0.153	−0.009	0.151

Table 2: Performance metrics for cross validation across 2020, 2021, and 2022

Feature Importance Analysis:

Feature Name	LR R^2 Means	Feature Name	RF FI Means
VARI	0.08483	Blue	0.56146
MGRVI	0.07704	VARI	0.53990
NGRDI	0.07623	VEG	0.39495
ExR	0.07515	BGI	0.38248
GRRI	0.07434	Red	0.37534
VEG	0.05934	ExGR	0.36786
RGBVI	0.05632	ExG	0.36619
GLI	0.05604	ExB	0.33718
Red	0.05191	Green	0.30306
Blue	0.05104	ExR	0.27251
ExB	0.04953	RGBVI	0.23457
BGI	0.04935	BI	0.20693
BI	0.04752	GLI	0.19037
Green	0.04587	MGRVI	0.04259
ExG	0.03340	NGRDI	0.04121
ExGR	0.03329	GRRI	0.01670

Table 3: Comparison of mean R^2 values generated with Linear Regression to the mean feature importance values generated with Random Forest. Highlighted features considered important in previous works.

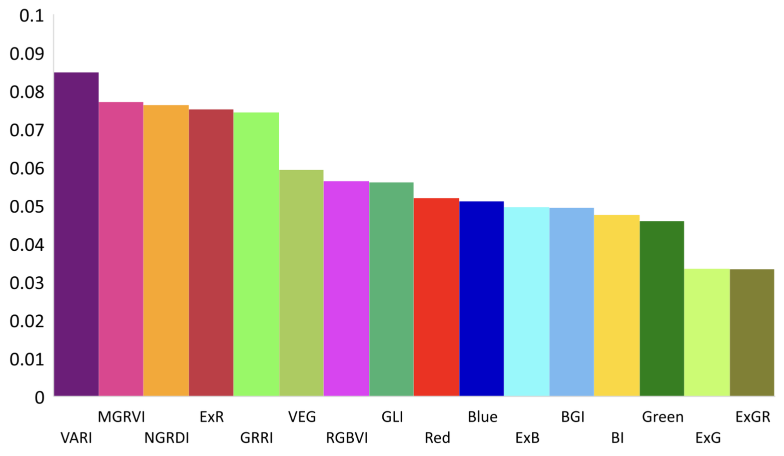

Fig. 7: Mean R^2 values from Linear Regression

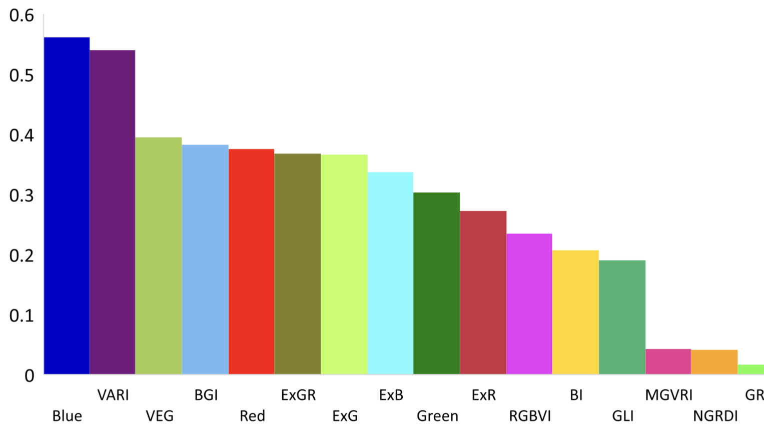

Fig. 8: Mean feature importance values from Random Forest

Discussion:

- Results indicate CV is an important tool for detecting spatial variation.
- CV and Test Results are more similar when spatial components are incorporated into building folds.
- Spatial models in general can reduce overfitting which is essential for crop estimation as fields can have variable characteristics.
- Random Forest and XGBoost reduce overfitting as shown by similar R^2 and RMSE values between CV and Test results.
- Feature Extraction: Linear Regression consistently ranked prominent features in other literature higher compared to Random Forest.

Future Research:

- Inconsistencies observed in 2020's data require further investigation.
- Training CV model with selected features to improve model performance.

Acknowledgments:

This research was funded by the National Science Foundation (NSF) Research Experience for Undergraduates (REU) grant OAC-2150191. We would also like to thank the Office of Research and Sponsored Programs (ORSP) at UW-Eau Claire for student support. The computational resources of this study was provided by the Blugold Center for High-Performance Computing under NSF grant CNS-1920220

We thank the Office of Research and Sponsored Programs for supporting this research, and Learning & Technology Services for printing this poster.