Crop identification in smallholder farms using machine learning & multi-sensor satellite data

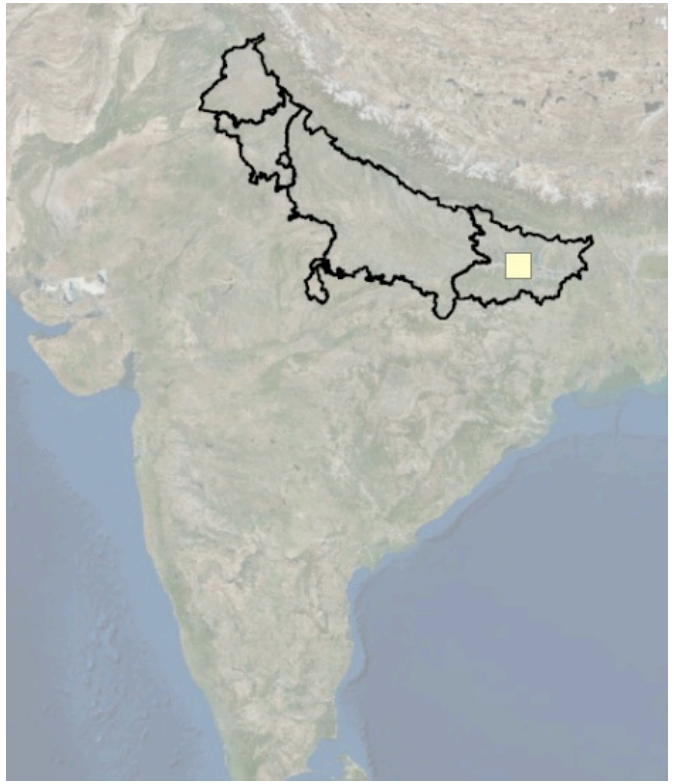
Preeti Rao[1], Meha Jain[1], Balwinder Singh[2], Amit Srivastava[2]

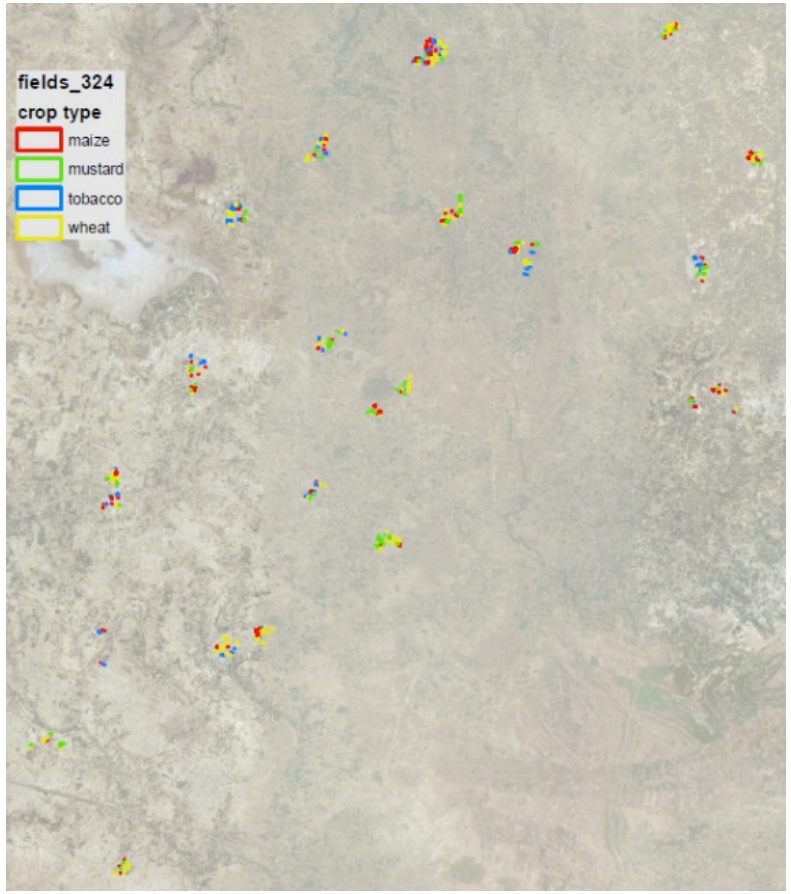
1: University of Michigan Ann Arbor

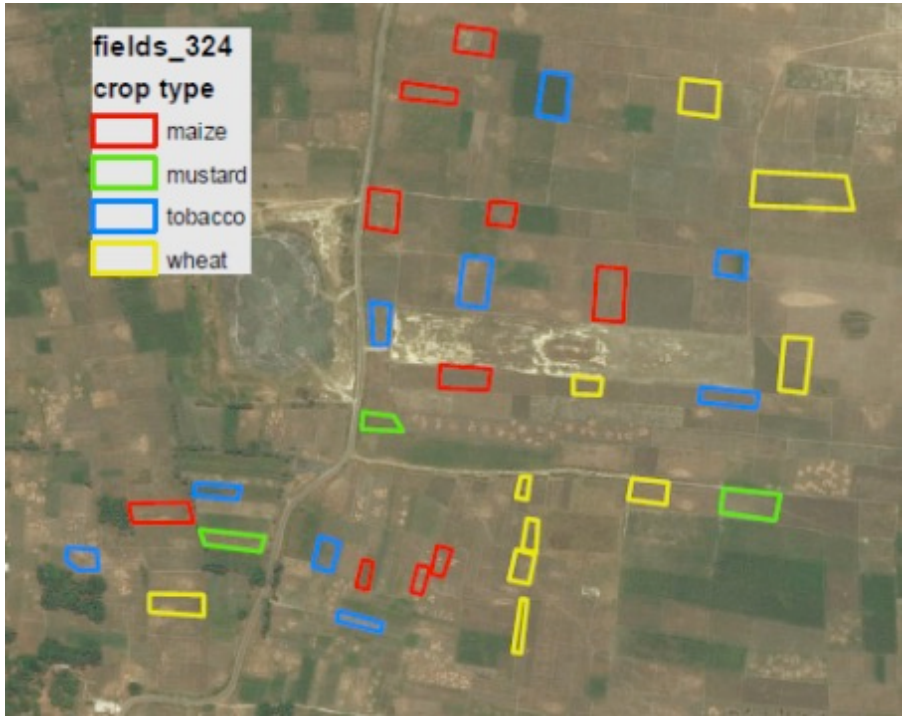
2: CIMMYT-India (International Maize & Wheat Improvement Center)

Abstract: Smallholder farming systems in the Indo-Gangetic Plains (IGP) are a major part of the rice-wheat production belt of India. Identifying the crop types across the entire IGP provides a critical dataset to help understand cropping patterns, crop yield intensities, and farmer adaptations to climate change. Our study area is a 20 x 20 km area in Eastern IGP where we collected crop type information for four major crops (maize, mustard, tobacco and wheat) during the winter growing season of 2016-17. The mean farm size in our sampled dataset of 324 fields is 745 m^2 with 64% of the fields smaller than the mean size. We compare the performance of three machine learning algorithms, Random Forests (RF), Support Vector Machines (SVM) and Artificial Neural Networks (ANN) to develop an ensemble classifier. We apply this ensemble to multi-sensor high-resolution optical (Sentinel-2 and Planet) and radar (Sentinel-1) satellite data to identify the four major crop types in our study area. We identify the critical number and timing of images essential for high classification accuracies. These learnings will be applied towards multi-temporal crop type classification in the entire IGP region.

Study area & smallholder farms







Indo-Gangetic Plains (rice-wheat belt) of India – Vaishali District in Bihar
Smallholder farms with diverse cropping patterns
Winter growing season – Nov 2016 to April 2017

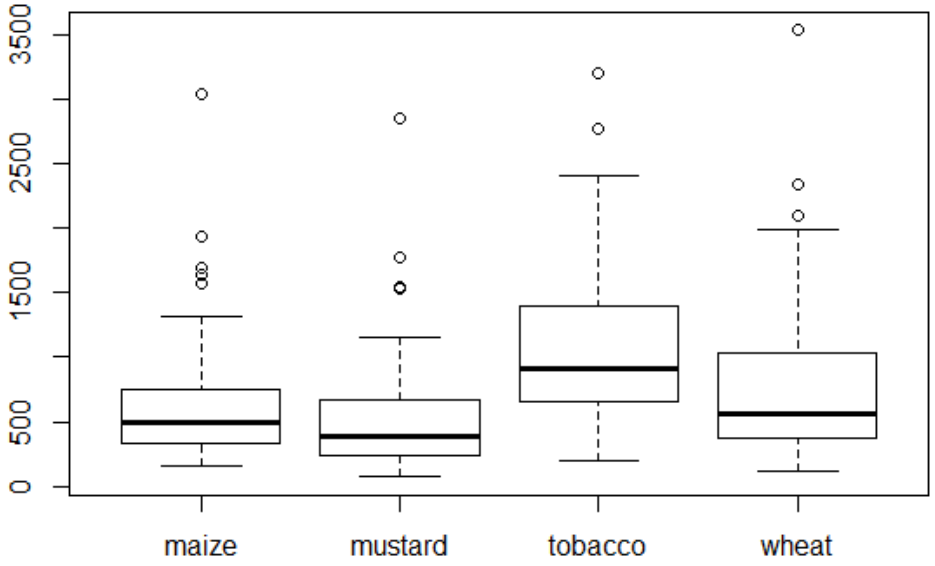
Field and satellite data

- 324 field polygons for 4 major crop types,
- Temporal satellite data from 3 sensors:
  - Planet = 40 bands [4 images * (4 bands + 6 indices)]
  - Sentinel-2 = 102 bands [6 images * (10 bands + 7 indices)]
  - Sentinel-1 = 60 bands [15 images * (2 bands + 2 indices)]
- R-package: Caret library

Crop type	field poly (n=324)
Maize	81
Mustard	65
Tobacco	58
Wheat	120

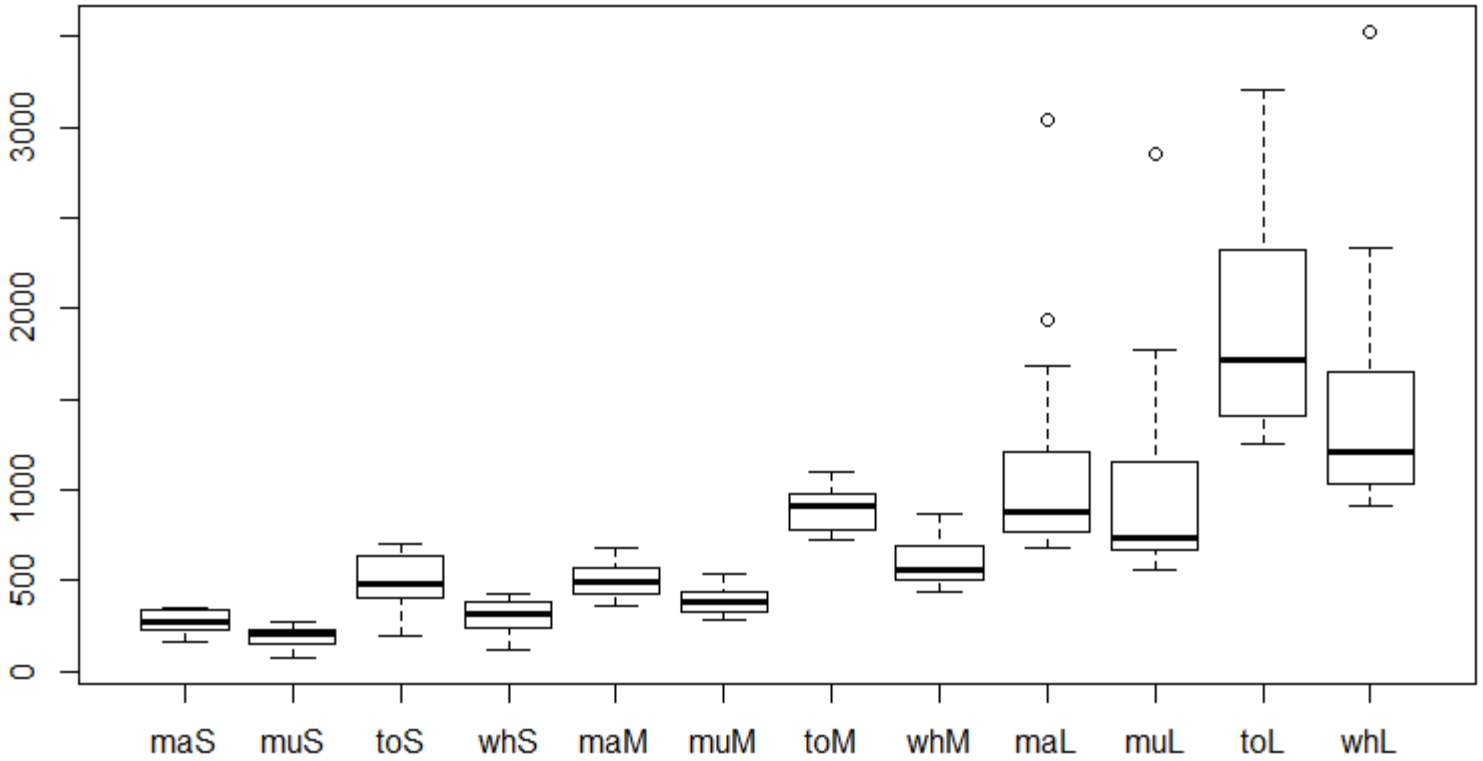
Satellite data	Image dates (mmdd)
Planet SR (bands BGRN, indices G-B NDVI, G-R NDVI, NDVI, PSRI, NPCI)	1115, 0218, 0320, 0409
Sentinel-2 SR (B2-B12, indices GCVI, NDVI, G-B NDVI, NDTI, PSRI, NPCI)	1119, 1129, 0118, 0207, 0217, 0309
Sentinel-1 (VV, VH, CR, BSR)	1120, 1202, 1214, 1226, 0107, 0119, 0209, 0212, 0221, 0224, 0305, 0308, 0317, 0320, 0401

Smallholder farm characteristics

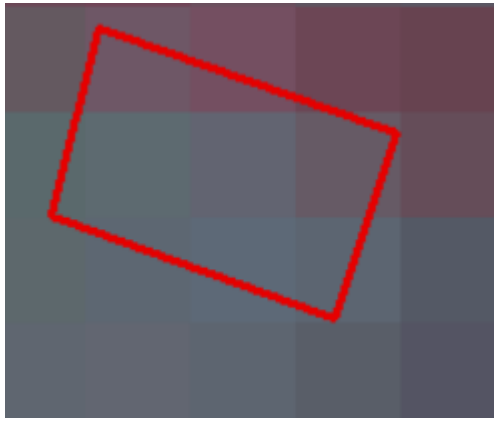


Maize: small is <= 355 m^2, medium is 355-675 m^2, large is > 675 m^2
Mustard: small is <= 280 m^2, medium is 280-535 m^2, large is > 535 m^2
Tobacco: small is <=725 m^2, medium is 725-1100 m^2, large is > 1100 m^2
Wheat: small is <= 425 m^2, medium is 425-910 m^2, large is > 910 m^2

ma = maize
mu = mustard
to = tobacco
wh = wheat
S = small
M = medium
L = large



Field poly vs. Planet (3 m) & Sentinel (10 m)



FieldArea_ m2	Min	Max	Mean	Total Area	# poly	3m pix in min poly	10m pix in min poly
Maize	162	3041	621	50,335	81	18	1.62
Mustard	74	2856	527	34,282	65	8	0.74
Tobacco	195	3203	1090	63,217	58	22	1.95
Wheat	117	3528	774	92,859	120	13	1.17
Total	74	3528	743	240,693	324		

Classification steps

- Equal data samples from each of the four major crop types were collected from 70% training and 30% test polygons
- Basic feature selection: removed variables with correlation > 0.9
- Data was sampled using the 10-fold cross-validation repeated 10 times and the best model evaluated on the basis of the Kappa statistic.
- For ANN, classification accuracy increased when the 3-sensor data was scaled (scaling didn’t make any difference for the other two models)
- For SVM, radial kernel performed better on the 3-sensor data
- RF performed best for mtry = 12 (sqrt of # variables) & ntree = 500

Training & test data

- # Training polygons (70%)

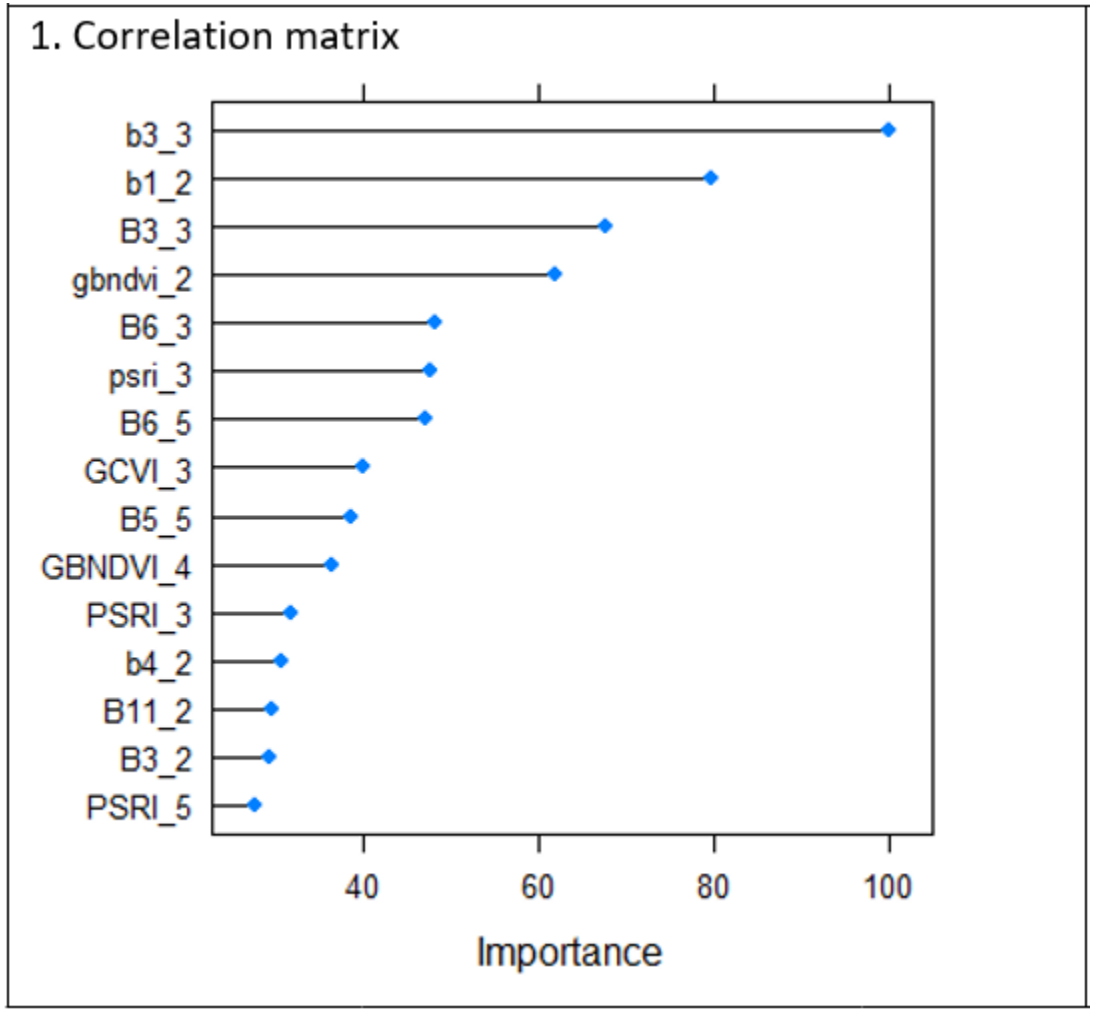
maize	mustard	tobacco	wheat
57	46	41	84

- # Test polygons (30%)

maize	mustard	tobacco	wheat
24	19	17	36

- At 3 m pixel locations: Planet + Sentinel-2 + Sentinel-1 band values
- Total pixels from each crop type (2040 pixels in mustard class)

Feature selection



- Top 15 variables after selecting features with correlation < 0.9
- 142 of 202 remained after removing the most highly correlated variables
- Final model runs with these 142 variables

Comparison of classification accuracies

Final model runs with optimized parameters: overall accuracy (kappa coefficient) for different combinations of satellite sensors and machine learning algorithms

	Accuracy (Kappa)		
	Planet	**Planet + Sentinel-2**	**Planet + Sentinel-2 + Sentinel-1**
Random Forest (RF)	0.813 (0.732)	0.806 (0.721)	0.850 (0.786)
Support Vector Machine (SVM)	0.781 (0.692)	0.822 (0.750)	0.859 (0.799)
Artificial Neural Network (ANN)	0.795 (0.710)	0.759 (0.660)	0.840 (0.775)

F1 score	Maize	Mustard	Tobacco	Wheat
RF	0.872	0.675	0.867	0.891
SVM	0.859	0.686	0.894	0.898
ANN	0.807	0.707	0.876	0.868

F1 score = (2*prodAcc*userAcc)/(prodAcc + userAcc)

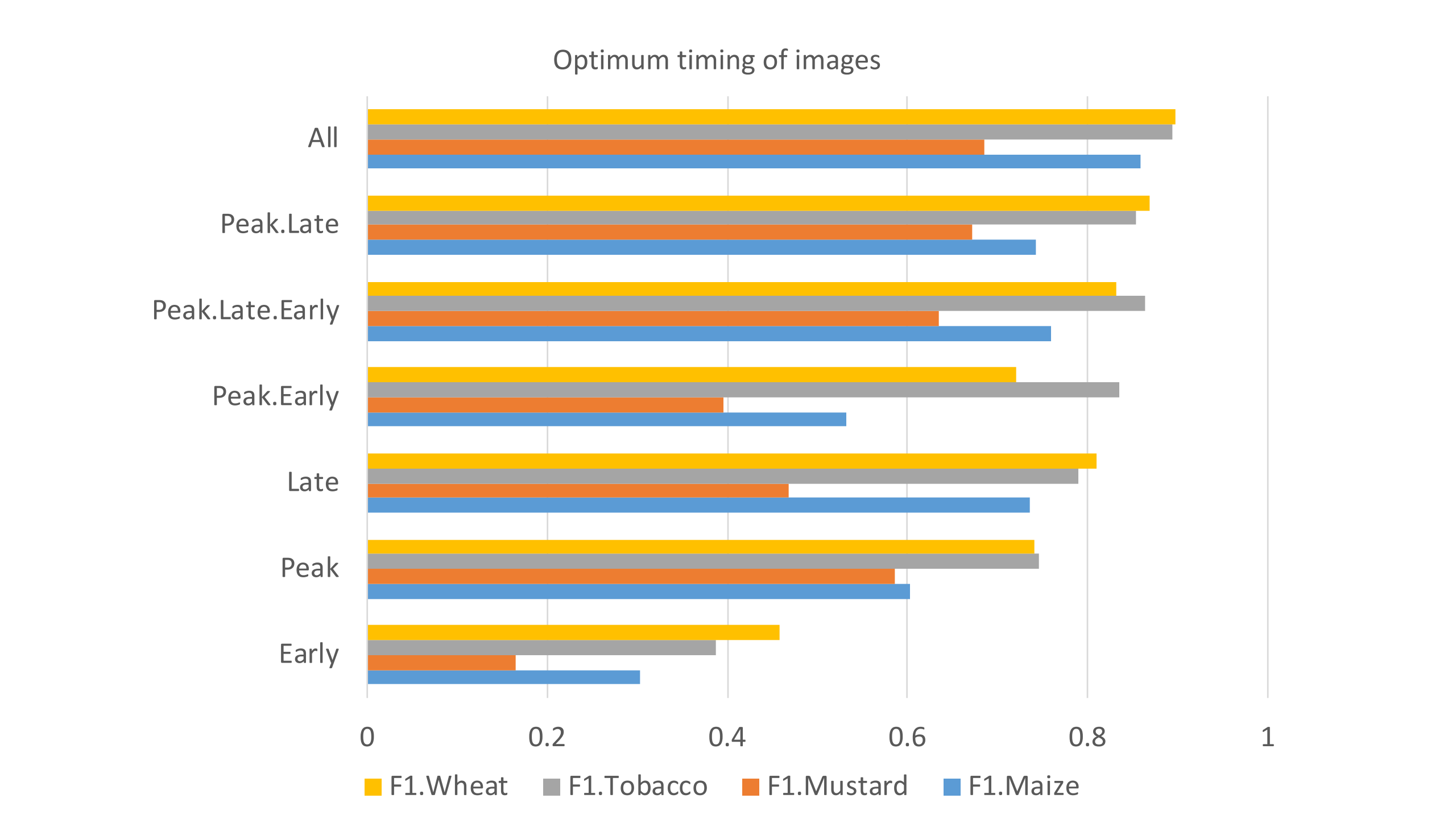
Optimum image analysis

- What is the best crop stage & min # images required for most accurate classification results
- SVM (best algorithm of the three) used to test the classification accuracies

Image date	Peak (mid Feb)	Late (late Mar/ early Apr)	Early (late Nov/ early Dec)	P + L	P + E	P + L + E
Planet	0218	0320, 0409	1115	0218 +0320 +0217 +0309 + 0212 +0320	0218 +1115 +0217 +1129 + 0212 +1202	0218 +0320 +1115 +0217 +0309 +1129 +0212 +0320 +1202
Sen-2	0217	0309	1119, 1129			
Sen-1	0212	0317, 0320, 0401	1120, 1202, 1214			

Optimum images: classification accuracies

SVM		Early	Peak	Late	Peak. Early	Peak. Late. Early	Peak. Late	All
Overall accuracy (kappa)		0.370 (0.105)	0.690 (0.571)	0.740 (0.635)	0.664 (0.521)	0.798 (0.713)	0.811 (0.733)	0.859 (0.799)
User accuracy	Maize	0.265	0.590	0.670	0.498	0.725	0.672	0.818
	Mustard	0.214	0.563	0.541	0.481	0.769	0.777	0.772
	Tobacco	0.389	0.672	0.746	0.819	0.821	0.843	0.874
	Wheat	0.456	0.822	0.839	0.708	0.827	0.877	0.895
Producer accuracy	Maize	0.354	0.617	0.816	0.571	0.799	0.829	0.905
	Mustard	0.134	0.610	0.413	0.335	0.540	0.593	0.617
	Tobacco	0.385	0.837	0.839	0.851	0.910	0.864	0.915
	Wheat	0.461	0.673	0.782	0.732	0.838	0.860	0.902
F1 score	Maize	0.303	0.603	0.736	0.532	0.760	0.742	0.859
	Mustard	0.165	0.586	0.468	0.395	0.634	0.672	0.686
	Tobacco	0.387	0.745	0.790	0.835	0.863	0.853	0.894
	Wheat	0.458	0.740	0.809	0.720	0.832	0.868	0.898

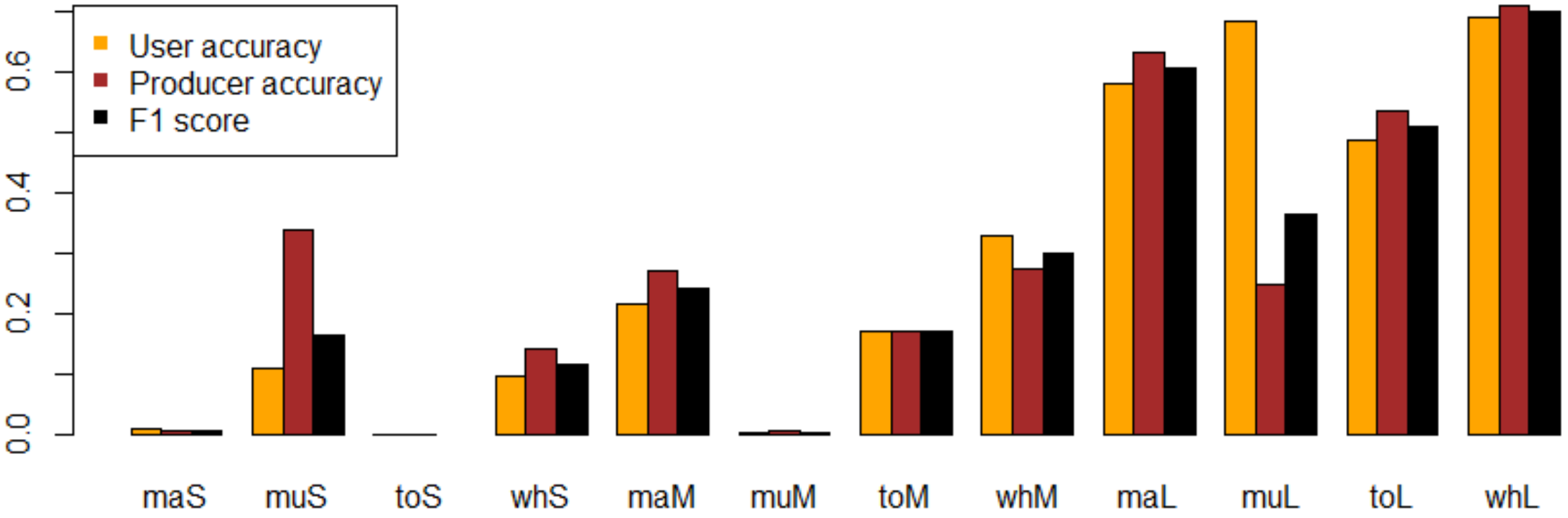

Optimum timing of images
All
Peak.Late
Peak.Late.Early
Peak.Early
Late
Peak
Early
0
0.2
0.4
0.6
0.8
1
F1.Wheat
F1.Tobacco
F1.Mustard
F1.Maize

Farm size & classification accuracies

Maize: small is <= 355 m^2, medium is 355-675 m^2, large is > 675 m^2
Mustard: small is <= 280 m^2, medium is 280-535 m^2, large is > 535 m^2
Tobacco: small is <=725 m^2, medium is 725-1100 m^2, large is > 1100 m^2
Wheat: small is <= 425 m^2, medium is 425-910 m^2, large is > 910 m^2

# Fields	Maize	Mustard	Tobacco	Wheat
Small	27	21	19	40
Medium	27	22	20	39
Large	27	22	19	41

Farm size & classification accuracies



- Large size fields of wheat, maize, tobacco perform the best
- Medium fields of wheat, maize and tobacco – next best
- Small size fields – only mustard does better than others

Conclusions

- SVM performs better than RF & ANN
- Combination of satellite data from the three sensors is the best
- Single image from each of the 3 sensors from late growing season & a combination of single peak & late image from each sensor classify the crop types almost as well as the complete dataset.
- It is difficult to classify small fields (300 – 700 m^2)
- Some crops are easier to identify (larger sample size or pheonology)
- Next step is to apply this SVM model or an ensemble of all three models to the larger IGP region.

Questions? Suggestions?