Near real-time soybean phenology detection using proximally-sensed hyperspectral canopy reflectance and machine learning methods

Nicolas Rigalli[1], Enrique Montero Bulacio[1], Martin Romagnoli[1], Juan Martin Enrico[2], Silvina Bacigaluppo[2], Ricardo Pozzi[3], Raquel L. Chan[4], Maria E. Otegui[5], and Margarita Portapila[1]

[1]CIFASIS-CONICET-UNR, Rosario, Argentina
[2]INTA Estacion Experimental Oliveros
[3]CREA San Jorge Las Rosas
[4]IAL-CONICET-UNL, Santa Fe, Argentina
[5]CONICET-INTA-FAUBA, Pergamino, Buenos Aires, Argentina

1. Table of genotypes

Table 1. List of genotypes used for the experiments.

Short name	Full name	Company	Maturity Group	Growth habit	Flower colour	Year of Release
DM2200	DON MARIO 2200	ASOCIADOS DON MARIO S.A.	II	Indeterminate	Violet	2004
SRM3410	SRM 3410	SURSEM S.A	III	Indeterminate	Violet	2013
LDC3.7	LDC 3.7	ASOCIADOS DON MARIO S.A.	III	Indeterminate	White	2013
DM40R16	DON MARIO 40R16	ASOCIADOS DON MARIO S.A.	IV	Indeterminate	Violet	2016
SP4x4	SP 4X4	ASOCIADOS DON MARIO S.A.	IV	Indeterminate	Violet	2009
NS4955	NS 4955	SYNGENTA CROP PROTECTION A.G.	IV	Indeterminate	Violet	2014
NA5009	NIDERA A 5009 RG	SYNGENTA CROP PROTECTION A.G.	V	Indeterminate	White	2006
NS5258	NS 5258	SYNGENTA CROP PROTECTION A.G.	V	Indeterminate	White	2014
AW5815	5815IPRO	MONSANTO ARGENTINA S.A.I.C.	V	Determinate	Violet	2014
M6210	M6210 IPRO	D&PL BRASIL LTDA.	VI	Indeterminate	Violet	2013
RA644	RA 644	COOP.PROV.SERV.AGR.SANTA ROSA	VI	Determinate	White	2012
CZ6505	6505 B	BASF AGRIC. SOLUT SEEDS US LLC	VI	Indeterminate	Violet	2015
RA750	RA 750	COOP.PROV.SERV.AGR.CR.STA.ROSA	VII	Determinate	Violet	2018
W82	WILLIAMS	USDA-ARS/Illinois AES	III	Indeterminate		1982
NS3220	NS 3220	SYNGENTA CROP PROTECTION A.G.	III	Indeterminate	Violet	2016
ADM4800	ADM 4800	ASOCIADOS DON MARIO S.A.	IV	Indeterminate	Violet	1999
AS5308	AS 5308 I	ASOCIADOS DON MARIO S.A.	V	Indeterminate		2008
DM5.8	DON MARIO 5.8I	ASOCIADOS DON MARIO S.A.	V	Indeterminate	White	2006
NA4990	NIDERA A 4990 RG	SYNGENTA CROP PROTECTION A.G.	IV	Indeterminate	White	2006
b10H	b10H	BIOCERES SEMILLAS	III	Indeterminate		-
NAHaHB4	NA 4990 - HaHB4	BIOCERES SEMILLAS	IV	Indeterminate	White	-

2. Outlier identification and cleaning method

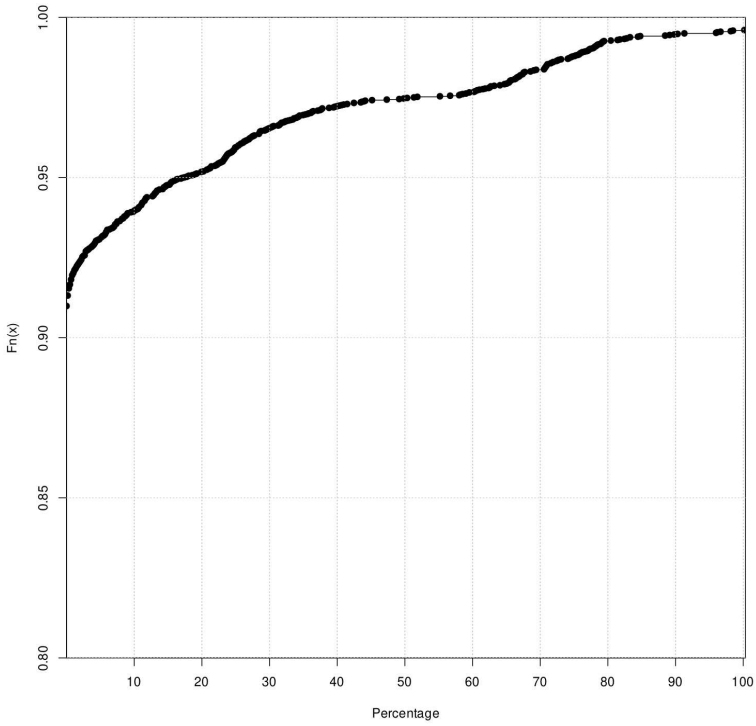

Figure 1. Empirical cumulative distribution function Fn(x), x is the percentage of outliers in a spectrum. For Fn(x=10%)=P[n<x=10%]=0.94 which means that 6% of the spectrum are excluded.

3. Table of samples after data cleaning

Table 2. Canopy spectral reflectance samples per date and genotype after data cleaning

Site	Date	AS5308	ADM4800	DM40R16	DM5.8i	NS3220	W82	B10H	NA4990	NAHaHB4	AW5815	CZ6505B	DM2200	LDC3.7	M6210	NA5009	NS4955	NS5258	RA644	RA750	SP4X4	SRM3410
Oliveros	11.01.18	--	--	25	--	--	--	--	--	--	34	28	27	28	27	30	26	28	28	27	26	27
	30.01.18	--	--	30	--	--	--	--	--	--	26	30	28	28	27	30	29	29	27	27	29	26
	23.02.18	--	--	33	--	--	--	--	--	--	30	30	33	33	31	34	31	32	37	29	31	31
	16.03.18	--	--	32	--	--	--	--	--	--	31	31	--	30	31	29	33	29	33	32	28	--
	08.01.19	--	--	46	--	--	--	--	--	--	--	--	--	45	--	45	47	--	--	--	45	46
	07.02.19	--	--	45	--	--	--	--	--	--	--	--	--	--	--	43	42	--	--	--	29	45
	13.03.19	--	--	43	--	--	--	--	--	--	--	--	--	--	--	43	44	--	--	--	43	43
Pergamino	01.02.18	30	26	29	31	29	26	27	--	--	--	--	--	--	--	--	--	--	--	--	--	--
	22.02.18	31	30	26	29	29	28	28	--	--	--	--	--	--	--	--	--	--	--	--	--	--
	15.03.18	34	32	31	31	31	32	27	--	--	--	--	--	--	--	--	--	--	--	--	--	--
	28.03.18	31	32	30	30	30	31	30	--	--	--	--	--	--	--	--	--	--	--	--	--	--
	18.01.19	23	35	35	34	34	34	36	35	32	--	--	--	--	--	--	--	--	--	--	--	--
	05.02.19	21	29	29	29	28	18	39	29	20	--	--	--	--	--	--	--	--	--	--	--	--
	26.02.19	19	30	30	30	29	30	30	30	29	--	--	--	--	--	--	--	--	--	--	--	--
	25.03.19	24	38	40	33	31	32	31	31	33	--	--	--	--	--	--	--	--	--	--	--	--
Santa Fe	19.01.18	--	--	--	--	29	31	30	--	--	--	--	--	--	--	--	--	--	--	--	--	--
	08.02.18	--	--	--	--	28	29	31	--	--	--	--	--	--	--	--	--	--	--	--	--	--
	01.03.18	--	--	--	--	30	29	31	--	--	--	--	--	--	--	--	--	--	--	--	--	--
	26.03.18	--	--	--	--	32	30	30	--	--	--	--	--	--	--	--	--	--	--	--	--	--
	16.01.19	--	--	--	--	--	61	58	55	59	--	--	--	--	--	--	--	--	--	--	--	--
	29.01.19	--	--	--	--	--	48	58	59	59	--	--	--	--	--	--	--	--	--	--	--	--
	19.02.19	--	--	--	--	--	59	59	59	59	--	--	--	--	--	--	--	--	--	--	--	--
	21.03.19	--	--	--	--	--	59	57	62	59	--	--	--	--	--	--	--	--	--	--	--	--

4. Vegetation indices

Rn, Rr, Rn2, Rre, R900 and R970 are the mean reflectance values in the intervals [850,880] nm, [640,670] nm, [775,805] nm, [705,735] nm, [875,915] nm and [955,985].

NDVI=(Rn-Rr)/(Rn+Rr)

NDRE=(Rn2-Rre)/(Rn2+Rre)

WI=R900/R970

EVI= 2.5(Rn-Rr)/(Rn+2.4Rr+1)

SIF=(a x d-c x b)/[(a-b) x a]

The coefficients of SIF index are based on the absolute reflectance of the plant material inside and outside a chosen O2 band. Consequently, a and c are the white reflectance material and sample measurements outside the absorption band while b and d are the white reflectance material and sample measurements inside the absorption band. 757.56 nm and 760.41 nm were chosen as outside and inside band wavelengths.

5. Modelling methods

ANN is based on layers of artificial neurons consisting of one input layer, zero to multiple hidden layers and one output layer. It iterates over the training set updating weights and estimating biases until a stopping criterion is met. In this work, a single-hidden-layer neural network was constructed. The number of neurons in the input layer was set according to the input feature dimension (i.e., auxiliary and spectral predictors plus reflectance spectrum). The number of neurons in the hidden layer was determined by a grid search.

SVM is built based on the concepts of Vapnik–Chervonenkis dimension and structural risk minimization. It basically works by finding the hyperplane that maximizes the margin between classes. Important parameters include the kernel function, which reflects similarity between data points (i.e., between reflectance values) and the cost loss function (regularization parameter). In this work, the linear function was used as the kernel function and the value of optimal cost was established after a grid search.

RF is based on the decision tree algorithm and bootstrap aggregation ensemble method. It constructs multiple decision trees and uses voting technique to obtain the final prediction. The key parameter of RF is the number of trees; the more trees, the higher the accuracy of the RF prediction. In this work, the number of trees was set at 500, as suggested by Breiman (2001). The number of variables available for splitting at each tree node was determined by a grid search.

PLSR is a multivariate statistical technique that combines the merits of multiple linear regression analysis, canonical correlation analysis, and principal component

analysis. It reduces the dimensionality and noise of high-dimensional data and transforms the collinear input features to a few non-correlated latent variables by performing component projection. The number of latent variables controlling the model complexity was set in this study by a grid search.

6. Available datasets per phenological stage

Table 3: Available datasets per phenological stage used for training and testing models. CY: Cross-year, CSY: Cross-Same-Year. T0, T1, T2, T3 and T4: test datasets

Site	Date	V	R1	R2	R3	R4	R5	R6	R7	R8	CY	CSY	T1	T2	T3
Oliveros	11.01.18	228	--	26	107	--	--	--	--	--	X	X	--	--	--
	30.01.18	--	54	83	--	88	141	--	--	--	X	X	--	--	--
	23.02.18	--	--	--	--	60	129	65	161	--	X	X	--	--	--
	16.03.18	--	--	--	--	--	32	126	91	90	X	X	--	--	--
Pergamino	01.02.18	95	103	--	--	--	--	--	--	--	X	X	--	--	--
	22.02.18	--	--	60	10	131	--	--	--	--	X	X	--	--	--
	15.03.18	--	--	--	--	65	153	--	--	--	X	X	--	--	--
	28.03.18	--	--	--	--	--	--	156	58	--	X	X	--	--	--
Santa Fe	19.01.18	32	58	--	--	--	--	--	--	--	X	X	--	--	--
	08.02.18	--	--	60	28	--	--	--	--	--	X	X	--	--	--
	01.03.18	--	--	--	--	--	90	--	--	--	X	X	--	--	--
	26.03.18	--	--	--	--	--	--	92	--	--	X	X	--	--	--
Oliveros	08.01.19	--	--	181	--	93	--	--	--	--	--	X	--	--	--
	07.02.19	--	--	--	--	43	--	161	--	--	--	X	--	--	--
	13.03.19	--	--	--	--	--	--	87	86	43	--	--	--	X	X
Pergamino	18.01.19	92	161	45	--	--	--	--	--	--	--	X	--	--	--
	05.02.19	41	9	10	124	58	--	--	--	--	--	X	--	--	--
	26.02.19	--	--	--	--	39	189	29	--	--	--	--	X		X
	25.03.19	--	--	--	--	--	24	175	94	--	--	--	--	X	X
Santa Fe	16.01.19	233	--	--	--	--	--	--	--	--	--	X	--	--	--
	29.01.19	79	145	--	--	--	--	--	--	--	--	X	--	--	--
	19.02.19	--	--	--	236	--	--	--	--	--	--	--	X		X
	21.03.19	--	--	--	--	--	--	237	--	--	--	--	--	X	X

7. Confusion matrices and statistics by class for the Cross Year (CY) model evaluation of phenological stages on T2 test dataset. ANN: Artificial Neural Network, SVM: Support Vector Machine, RF: Random Forest and PLSR: Partial Least Square Regression. SD: Sowing Day, LAT: Latitude, DOY: Day Of Year, AGDD: cumulative Growth Degree-Day, FS: Full Spectrum.

Sensitivity = [a / (a + c)]
Specificity = [d / (b + d)]

a: True Positive
b: False Positive
c: False Negative
d: True Negative

		ANN	SVM	RF	PLSR
SD.LAT.DOY.AGDD	FS.NDVI.NDRE.WI.EVI.SIF	0.39	0.44	0.46	**0.75**
	FS	**0.71**	0.42	0.46	0.49

a. SD.LAT.DOY.AGDD FS.NDVI.NDRE.WI.EVI.SIF **PLSR 0.75**

Pred.\Obs.	V	R1	R2	R3	R4	R5	R6	R7	R8
V	0	0	0	0	0	0	4	0	0
R1	0	0	0	0	0	0	0	0	0
R2	0	0	0	0	0	0	4	0	0
R3	0	0	0	0	0	0	0	0	0
R4	0	0	0	0	0	0	0	0	0
R5	0	0	0	0	0	0	2	1	0
R6	0	0	0	0	0	24	460	93	9
R7	0	0	0	0	0	0	29	81	9
R8	0	0	0	0	0	0	0	5	25

Overall Statistics Accuracy : 0.7587

Statistics by Class:

	V	R1	R2	R3	R4	R5	R6	R7	R8
Sensitivity	NA	NA	NA	NA	NA	0.0000	0.9218	0.4500	0.58140
Specificity	0.9946	1	0.994638	1	1	0.995845	0.4899	0.9329	0.99289

b. SD.LAT.DOY.AGDD FS.NDVI.NDRE.WI **ANN 0.71**

Pred.\Obs.	V	R1	R2	R3	R4	R5	R6	R7	R8
V	0	0	0	0	0	0	7	0	0
R1	0	0	0	0	0	0	0	0	0
R2	0	0	0	0	0	0	55	0	0
R3	0	0	0	0	0	0	0	0	0
R4	0	0	0	0	0	0	0	0	0
R5	0	0	0	0	0	4	10	0	0
R6	0	0	0	0	0	20	377	24	9
R7	0	0	0	0	0	0	50	141	30
R8	0	0	0	0	0	0	0	15	13

Overall Statistics Accuracy : 0.7172

Statistics by Class:

	V	R1	R2	R3	R4	R5	R6	R7	R8
Sensitivity	NA	NA	NA	NA	NA	0.1667	0.7555	0.7833	0.3023
Specificity	0.9906	1	0.9263	1	1	0.9862	0.8219	0.8587	0.9787

8. Confusion matrices and statistics by class for the Cross Same Year (CSY) model evaluation of phenological stages on T2 and T3 test datasets. ANN: Artificial Neural Network, SVM: Support Vector Machine, RF: Random Forest and PLSR: Partial Least Square Regression. SD: Sowing Day, LAT: Latitude, DOY: Day Of Year, AGDD: cumulative Growth Degree-Day, FS: Full Spectrum.

T2 test dataset

		ANN	SVM	RF	PLSR
SD.LAT.DOY.AGDD	FS.NDVI.NDRE.WI.EVI.SIF	0.46	0.66	0.45	0.83
	FS	0.47	0.83	0.45	0.53

a. SD.LAT.DOY.AGDD FS.NDVI.NDRE.WI.EVI.SIF **PLSR 0.83**

Pred.\Obs.	V	R1	R2	R3	R4	R5	R6	R7	R8
V	0	0	0	0	0	0	0	0	0
R1	0	0	0	0	0	0	0	0	0
R2	0	0	0	0	0	0	0	0	0
R3	0	0	0	0	0	0	0	0	0
R4	0	0	0	0	0	0	0	0	0
R5	0	0	0	0	0	0	4	0	0
R6	0	0	0	0	0	24	448	19	2
R7	0	0	0	0	0	0	47	159	22
R8	0	0	0	0	0	0	0	2	19

Overall Statistics Accuracy : 0.8391

Statistics by Class:

	V	R1	R2	R3	R4	R5	R6	R7	R8
Sensitivity	NA	NA	NA	NA	NA	0.000000	0.8978	0.8833	0.44186
Specificity	1	1	1	1	1	0.994460	0.8178	0.8781	0.99716

b. SD.LAT.DOY.AGDD FS **SVM 0.83**

Pred.\Obs.	V	R1	R2	R3	R4	R5	R6	R7	R8
V	0	0	0	0	0	0	0	0	0
R1	0	0	0	0	0	0	31	0	0
R2	0	0	0	0	0	0	4	0	0
R3	0	0	0	0	0	0	0	0	0
R4	0	0	0	0	0	0	0	0	0

```
        R5   0   0   0   0   0   0   1   0   0
        R6   0   0   0   0   0  24 450  12   0
        R7   0   0   0   0   0   0  13 157  26
        R8   0   0   0   0   0   0   0  11  17

Overall Statistics Accuracy : 0.8365

Statistics by Class:
             V    R1      R2      R3    R4    R5     R6     R7     R8
Sensitivity  NA   NA      NA      NA    NA    0.0000 0.9018 0.8722 0.39535
Specificity  1    0.9584 0.99463 1     1     0.9986 0.8543 0.9311 0.98435
```

T3 test dataset

		ANN	SVM	RF	PLSR
SD.LAT.DOY.AGDD	FS.NDVI.NDRE.WI.EVI.SIF	0.48	0.62	0.38	0.55
	FS	0.49	**0.71**	0.39	0.35

c. SD.LAT.DOY.AGDD FS **SVM 0.71**

```
 Pred.\Obs.  V  R1  R2  R3  R4  R5  R6  R7  R8
        V    0   0   0   0   0   0   0   0   0
        R1   0   0   0   4   0   0  31   0   0
        R2   0   0   0  34   0   0   4   0   0
        R3   0   0   0 198   0   0   0   0   0
        R4   0   0   0   0  25  18   0   0   0
        R5   0   0   0   0   2   7   1   0   0
        R6   0   0   0   0  12 188 479  12   0
        R7   0   0   0   0   0   0  13 157  26
        R8   0   0   0   0   0   0   0  11  17

Overall Statistics Accuracy : 0.7127

Statistics by Class:
             V   R1      R2      R3      R4       R5         R6      R7      R8
Sensitivity  NA  NA      NA      0.8390  0.64103  0.032864   0.9072  0.8722  0.39535
Specificity  1   0.97175 0.96933 1.0000  0.98500  0.997076   0.7018  0.9632  0.99080
```

9. Phenological phases

a. SD.LAT.DOY.AGDD FS SVM T1

"Accuracy 0.85"

```
 V R1+R2 R3+R4 R5+R6 R7+R8
V     0     0     0     0     0
R1+R2 0     0    38     0     0
R3+R4 0     0   223    18     0
R5+R6 0     0    14   200     0
R7+R8 0     0     0     0     0
               V     R1+R2     R3+R4     R5+R6 R7+R8
Sensibilidad  NA        NA 0.8109091 0.9174312    NA
Especificidad  1 0.9229209 0.9174312 0.9490909     1
```

b. SD.LAT.DOY.AGDD FS SVM T2

"Accuracy 0.92"

```
 V R1+R2 R3+R4 R5+R6 R7+R8
V     0     0     0     0     0
R1+R2 0     0     0    35     0
R3+R4 0     0     0     0     0
R5+R6 0     0     0   475    12
R7+R8 0     0     0    13   211
               V     R1+R2 R3+R4     R5+R6     R7+R8
Sensibilidad  NA        NA    NA 0.9082218 0.9461883
Especificidad  1 0.9530831     1 0.9461883 0.9751434
```

```
c. SD.LAT.DOY.AGDD      FS    SVM   T3
"Accuracy 0.89"
       V R1+R2 R3+R4 R5+R6 R7+R8
V      0     0     0     0     0
R1+R2  0     0    38    35     0
R3+R4  0     0   223    18     0
R5+R6  0     0    14   675    12
R7+R8  0     0     0    13   211
                V      R1+R2     R3+R4     R5+R6     R7+R8
Sensibilidad   NA         NA 0.8109091 0.9109312 0.9461883
Especificidad   1 0.9410815 0.9813278 0.9477912 0.9872047
[1] "----------------------"
```

10. Effects of contrasting climate regimes and extreme weather events on the quality of phenology detection

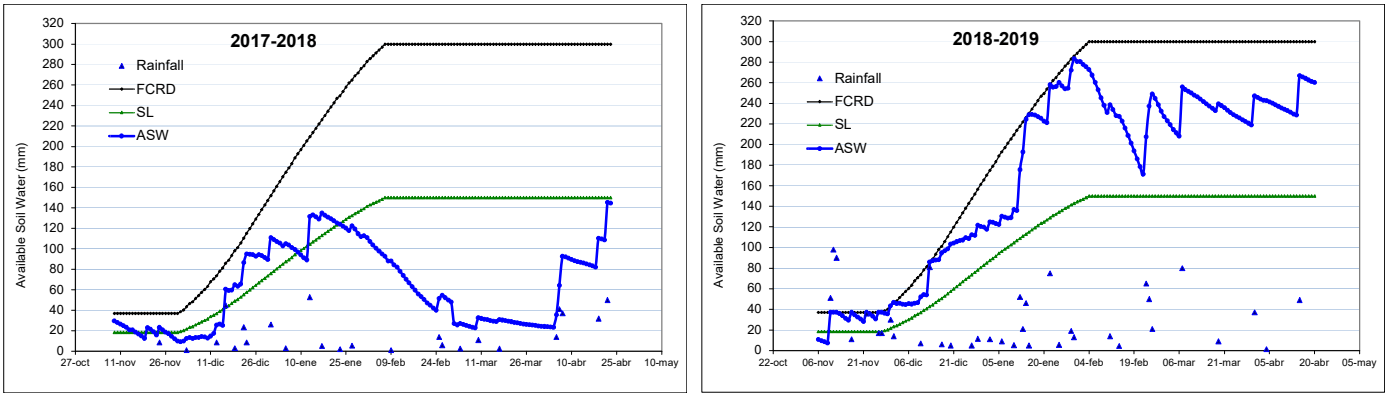

Figure 3. Daily available soil water for soybean crops sown at Oliveros Experimental Station during 2017-2018 and 2018-2019 planting seasons. FCRD: Field Capacity at the crop´s rooting depth, SL: Stress limit (50% of FCRD), ASW: Available soil water (difference between observed soil water content and permanent wilting point). When ASW is below SL the crop begins to show symptoms of water deficit

Figure 4. Soybean plot in R6 at Santa Fe site affected by hailstorm.