Use of Satellite Imagery to Predict Corn Phenology at a Regional Scale

L. Nieto, R. Schwalbert, and I.A. Ciampitti

Summary

Existing methods to report phenology are expensive, labor-intensive, time-consuming, and often not very accurate, especially at some specific crop growth stages. The objective of this study was to develop large-scale phenology models via utilization of satellite imagery data and machine learning techniques for the southwest (SW) agricultural crop reporting district of Kansas. Different satellite images collected from Landsat were utilized as the main input to obtain different vegetation indices (normalized difference vegetation index, NDVI; enhanced vegetation index, EVI; green chlorophyll vegetation index, GCVI; normalized difference water index, NDWI; and global vegetation moisture index, CVMI). Vapor Pressure Deficit (VPD), temperature, precipitation, and growing degree units (GDU) were evaluated for improving phenology prediction models. A large set of ground truth data with information about day of the year, crop phenology, and field location was provided by Crop Quest Inc. (Dodge City, KS) from 2014–2018 and utilized to train two different statistical models (Random Forest and Support Vector Machine) to catalog corn fields, and build a phenology evolution model for this crop.

Introduction

During the crop-growing season, the U.S. Department of Agriculture (USDA) via its agency, National Agricultural Statistics Service (NASS), releases a weekly report concerning the Crop Progress and Report Conditions (CPRC), providing an estimate of the crop phenology and overall condition of selected crops in major producing states. Phenology crop progress estimates are based on survey data collected each week from an extensive network of regional agricultural agents based on their field observations. Although this is a useful source of information, this task is labor-intensive, time-consuming, and biased on the data collection process. In addition, in some regions of the United States the CPRC are released after the crop is planted, decreasing the prediction power of estimating planting and emergence progress of the crop. Therefore, in an effort to improve the overall prediction of crop phenology and to resolve potential issues related to data bias and missing information, utilization of satellite imagery can play a key role in this work (Figure 1).

The objective of this research study was to explore and test the utilization of different classifiers to find the most accurate approach to predict crop phenology by integrating data, such as field survey (ground-truthing), remote sensing, and weather, via utilization of machine learning techniques.

The project is focused on the Southwest Agricultural Statistics District (SW), KS. The ground-truth consist in a large dataset owned by CropQuest, focusing on crop phenology for corn fields during the 2014–2018 growing seasons. This dataset was comprised of the following features: 1) geolocation of each field; 2) date of visit; 3) crop phenology; and 4) crop (e.g., corn in this study). Approximately 60,000 observations were made (Figure 2a) in Kansas, and approximately 25,000 observations just in the SW region of Kansas (Figure 2b).

Procedures

Briefly, the data preparation presented the following steps: 1) corn fields from the SW region in Kansas were selected from the dataset; 2) the different phenology stages from the original dataset (more than 20 categories) were combined into nine classes, these classes follow the most critical moments for field management practices (Table 2); 3) the geolocation of each field was utilized to locate the farms and the CONUS layer by Yan and Roy (2015) was used to recreate the boundaries of the fields presented in the dataset and transform these points into a shapefile; 4) satellite imagery (Landsat mission) from each farmer field was selected due to its spatial resolution, using one image per month, masking clouds, and selecting the best pixels to calculate the different vegetation indices; and 5) weather information (Table 1).

The different vegetation indices were selected according to the purpose of this research and to enhance some variables in the canopy. As an example, Cai (2018) stated that NDVI is based on the fact that healthy plants usually have a greater reflectance in the near infrared (NIR) than visible bands. The problem with the NDVI is that it tends to saturate at high biomass levels. The EVI was designed to reduce the influence of some atmospheric effects, including the blue band, into the calculation. The GCVI has been found to have most linear relationship with leaf area index (LAI) for corn and soybeans than other indices. The NDWI was developed to approximate canopy water thickness, based on the rationale that the shortwave infrared (SWIR) band is sensitive to leaf water and soil moisture. Finally the GVMI index is more suitable when looking at the global water content.

In terms of weather information, the Gridded Surface Meteorological dataset merges the high-resolution spatial data from PRISM with high temporal resolution data from NLDAS. From this data layer, we extracted metrics related to precipitation, minimum and maximum temperature, and VPD. Using these data, a GDU model was applied as:

$$GDU = [(Max.\ temp.\ (°F) - Min.\ temp.\ (°F))/2] - 50°F\ (base\ temp.).$$

All data layers were merged with the ground truth data, providing a final output of a table with the spectral bands, vegetation indices, weather data, and phenological growth stages in each georeferenced point, per month during the growing season. All computations mentioned were performed into a Google Earth Engine Environment (GEE) platform. The GEE is a cloud-based platform optimized for parallel processing of geospatial data for environmental data analysis, supporting work with large datasets. The GEE code editor allows us to rapid visualize the spatial analyses using JavaScript. The final table with all the information obtained in the GEE platform was then moved to the R environment to train the classifiers.

The two models selected to test in this study were:

1. Random Forest (RF).
2. Support Vector Machine (SVM).

The first classifier (RF) was selected due its performance with a large amount of data. In this classifier, each tree is a representation where the leaves are the class labels and the branches are the mergers of features that lead to those class labels. Then is trained by a random subset of the original dataset and the final classification is computed by aggregating results of all tree predictors. The second classifier (SVM) can solve problems in classification by looking for the global optimum and taking advantages from all the dimensions existing in the data to solve problems that a simpler model cannot achieve.

Results

The accuracy (number of all correct predictions divided by the total number of predictions) was selected as a parameter to compare the behavior of the models. The values for this specific parameter range between 0 and 1, 1 being the best scenario, where the model is able to predict one class 100% of the times.

The results considering accuracy for yearly analysis not using weather and using weather information are shown in Tables 3 and 4 respectively.

A second analysis was executed for each month during the growing season, from May to September, again not using weather and using this variable in the analysis. (Tables 5 and 6).

Conclusions

Several conclusions can be drawn from this preliminary analysis. First, the weather dataset is critical when training models. The use of this parameter helps to increase the accuracy of both models (Random Forest and Support Vector Machine), especially during the critical period of the crop (June-August), but with a positive impact throughout the entire growing season.

Second, a special treatment was necessary for the 2018 data. The phenology prediction model was built with crop data that presented a dissimilar weather condition relative to 2018, an anomalous year (e.g., high temperatures early in June). Thus, the phenology prediction model could improve as the data evaluated and added to the model could include broader weather variation.

References

Cai, Y., Guan, K., Peng, J., Wang, S., Seifert, C., Wardlow, B., & Li, Z. (2018). A high-performance and in-season classification system of field-level crop types using time-series Landsat data and a machine learning approach. Remote Sensing of Environment, 210, 35-47.

Ceccato, P., Gobron, N., Flasse, S., Pinty, B., & Tarantola, S. (2002). Designing a spectral index to estimate vegetation water content from remote sensing data: Part 1: Theoretical approach. *Remote sensing of environment*, *82*(2-3), 188-197.

Ciampitti, I. A., Elmore, R. W., Lauer, J. (2016). Corn Growth and Development. Kansas State University, MF3305.

Gao, B. C. (1996). NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sensing of Environment, 58(3), 257-266.

Gitelson, A. A., Viña, A., Arkebauer, T. J., Rundquist, D. C., Keydan, G., & Leavitt, B. (2003). Remote estimation of leaf area index and green leaf biomass in maize canopies. Geophysical Research Letters, 30(5).

Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X., & Ferreira, L. G. (2002). Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sensing of Environment, 83(1-2), 195-213.

Hunt Jr, E. R., Rock, B. N., & Nobel, P. S. (1987). Measurement of leaf relative water content by infrared reflectance. Remote Sensing of Environment, 22(3), 429-435.

Ozdogan, M., Yang, Y., Allez, G., & Cervantes, C. (2010). Remote sensing of irrigated agriculture: Opportunities and challenges. Remote sensing, 2(9), 2274-2304.

Peñuelas, J., Pinol, J., Ogaya, R., & Filella, I. (1997). Estimation of plant water concentration by the reflectance water index WI (R900/R970). International Journal of Remote Sensing, 18(13), 2869-2875.

Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. Remote Sensing of Environment, 8(2), 127-150.

Xiao, X., Boles, S., Liu, J., Zhuang, D., & Liu, M. (2002). Characterization of forest types in Northeastern China, using multi-temporal SPOT-4 VEGETATION sensor data. Remote Sensing of Environment, 82(2-3), 335-348.

Yan, L., & Roy, D. P. (2016). Conterminous United States crop field size quantification from multitemporal Landsat data. Remote Sensing of Environment, 172, 67-86.

Table 1. Datasets used to study corn phenology indicators

Vegetation indices	Weather information
Normalized difference vegetation index (NDVI) (Tucker, 1979)	Precipitation (Pr)
Enhanced vegetation index (EVI) (Huete, 2002)	Maximum temperature (Tmx)
Green chlorophyll vegetation index (GCVI) (Gitelson, 2003)	Minimum temperature (Tmin)
Normalized difference water index (NDWI) (Gao, 1996)	Vapor pressure deficit (VPD)
Global vegetation moisture index (CVMI) (Ceccato, 2002)	Growing degree units (GDU)

Table 2. Class division for corn growth and phenology stages

Class number	Phenology stages*	Observations
1	V0-V1	Planted
2	V2-V4	Vegetative
3	V5-V8	
4	V9-V16	
5	R1-R3	Reproductive
6	R4	
7	R5	
8	R6	
9	H	Harvested

*Ciampitti *et al.* 2016.

Table 3. Yearly accuracy for Random Forest and Support Vector Machine considering all the variables except weather

Year	Random Forest	Support Vector Machine
2018	**0.79**	0.77
2017	**0.88**	0.68
2016	**0.90**	0.79
2015	**0.89**	0.71
2014	**0.87**	0.67

Table 4. Yearly accuracy for Random Forest and Support Vector Machine including weather parameters

Year	Random Forest	Support Vector Machine
2018	0.85	0.63
2017	0.89	0.92
2016	0.92	0.73
2015	0.91	0.93
2014	0.86	0.87

Table 5. Monthly accuracy for Random Forest (RF) and Support Vector Machine (SVM) considering vegetation indices and no weather

Year	Model	May	June	July	Aug	Sep	
2014	RF	0.75	0.6726	0.807	0.9198	0.6512	
	SVM	0.8571	0.7083	0.814	0.9321	0.6366	
		May	June	July	Aug	Sep	
2015	RF	0.93	0.8571	0.4545	0.6094	0.8864	
	SVM	0.97	0.8571	0.4909	0.6011	0.9038	
		May	June	July	Aug	Sep	
2016	RF	0.9727	0.57	0.7233	0.4697	0.8281	
	SVM	0.9727	0.5222	0.7547	0.5251	0.8281	
		J 1	J 2	July	A 1	A 2	Sep
2017	RF	0.7514	0.6042	0.733	0.5597	0.4583	0.7995
	SVM	0.7715	0.599	0.7961	0.5767	0.473	0.7226
		May	June	July	Aug	Sep	
2018	RF	0.8927	0.5183	0.3729	0.4731	0.936	
	SVM	0.8927	0.4878	0.4746	0.5484	0.8722	

Table 6. Monthly accuracy for Random Forest and Support Vector Machine considering all the variables (vegetation indices and weather information)

Year	Model	May	June	July	Aug	Sep	
2014	RF	0.86	0.84	0.84	0.9	0.77	
	SVM	0.61	0.95	0.98	0.95	0.94	
		May	June	July	Aug	Sep	
2015	RF	0.97	0.77	0.67	0.84	0.89	
	SVM	0.99	0.97	0.8	0.95	0.98	
		May	June	July	Aug	Sep	
2016	RF	0.89	0.87	0.81	0.82	0.77	
	SVM	0.79	0.81	0.71	0.55	0.76	
		J 1	J 2	July	A 1	A 2	Sep
2017	RF	0.89	0.75	0.79	0.76	0.75	0.77
	SVM	0.98	1	0.9	0.97	0.94	0.9
		May	June	July	Aug	Sep	
2018	RF	0.91	0.68	0.45	0.58	0.93	
	SVM	0.9	0.62	0.4	0.55	0.92	

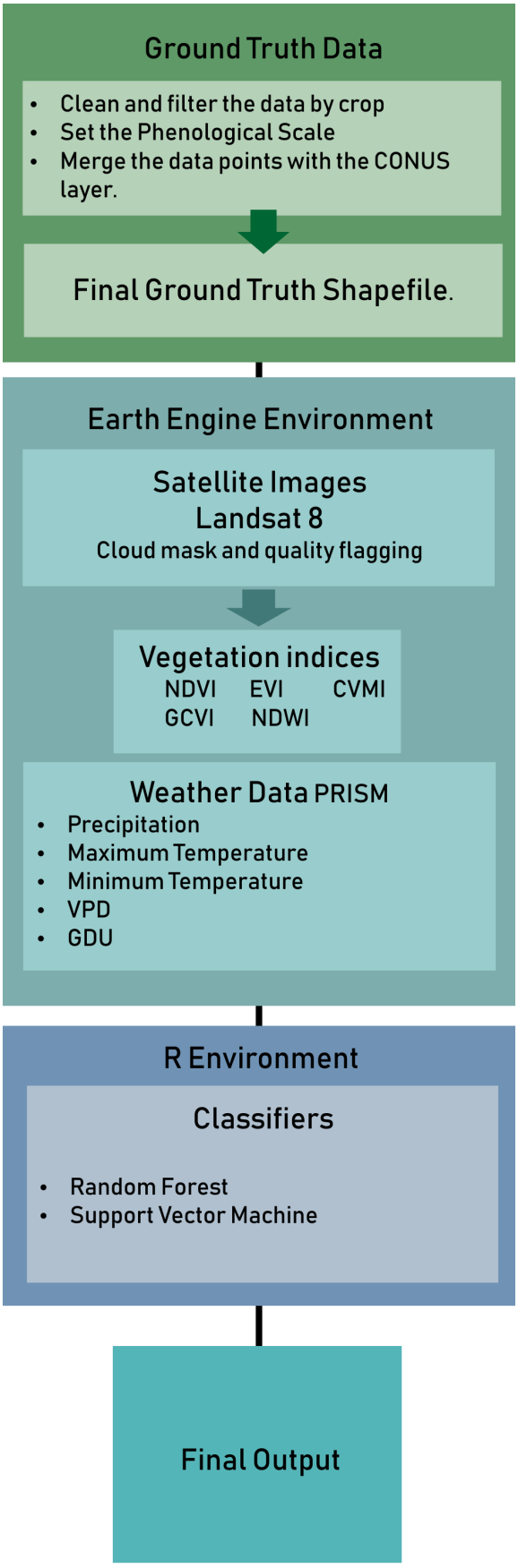

Figure 1. Workflow.

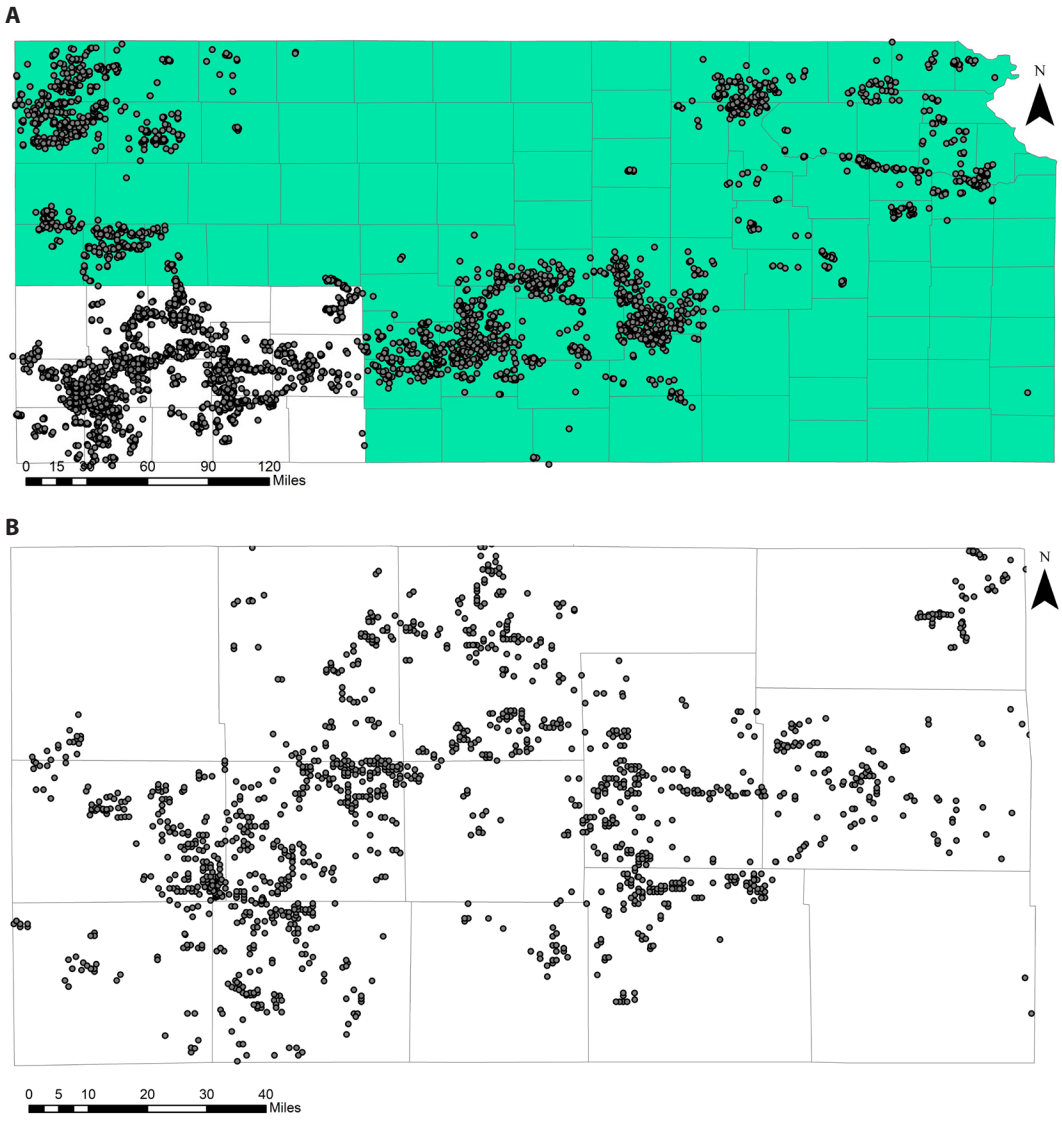

Figure 2. A) Data point distributions in Kansas in 2014. B) Data point distributions in the Southwest Agricultural District in 2014.