Crop classification and uncertainty assessment at 10-m resolution using Google Earth Engine

Luca Sartore[1,2], Claire Boryan[2]

[1]National Institute of Statistical Science (NISS)
[2]United States Department of Agriculture (USDA)
National Agricultural Statistics Service (NASS)

lsartore@niss.org

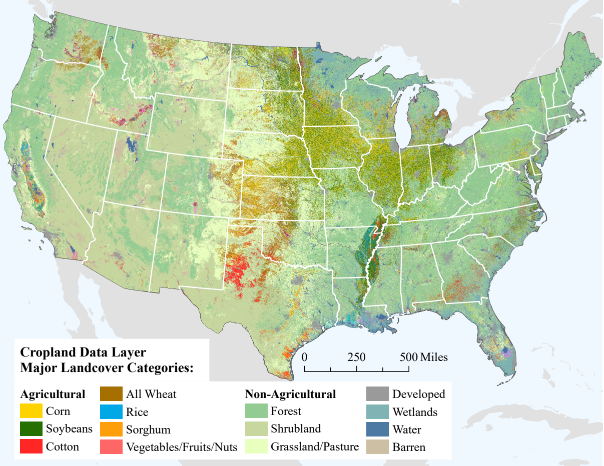
Government Advanced Statistical Programming 2020
November 6, 2020

Disclaimer and acknowledgments

- This research was supported in part by the intramural research program of the U.S. Department of Agriculture, National Agricultural Statistics Service.

Cropland Data Layer (CDL; Boryan et al., 2011)

Administrative data and remote sensing images (not available in June) are used as inputs to create the crop specific land product



Objectives

Early crop classification (June) would be valuable to NASS for a variety of operational purposes

An Early Season CDL (ESCDL) has been recently developed at 10-m resolution using Sentinel-2 images before administrative data become available

Uncertainty layers can provide information on the stability of the classification results showing areas that might be either misclassified or very challenging to model

Classification approach

Bootstrap methods (Efron and Tibshirani, 1994) provide the opportunity to study the predictive distribution for a given classifier

Classification results y^* are computed as

$$y^* = \underset{k \in \{1,\ldots,K\}}{\arg\max} \ \hat{p}_k$$

where

- K denotes the total number of classes
- $\hat{p}_k$ represents the **prediction frequency** of the class k defined as

$$\hat{p}_k = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}_{\{k\}}(\hat{y}_b)$$

where

- B denotes the total number of bootstrap iterations
- $\hat{y}_b$ represents the predicted class for the bootstrap iteration b
- $\mathbb{1}_{\{k\}}(\hat{y}_b)$ is the indicator function returning 1 if $\hat{y}_b = k$, and 0 otherwise

Uncertainty assessment

The **standardized empirical entropy** is a measure of uncertainty for y^*, and it is computed from the predictive distribution as

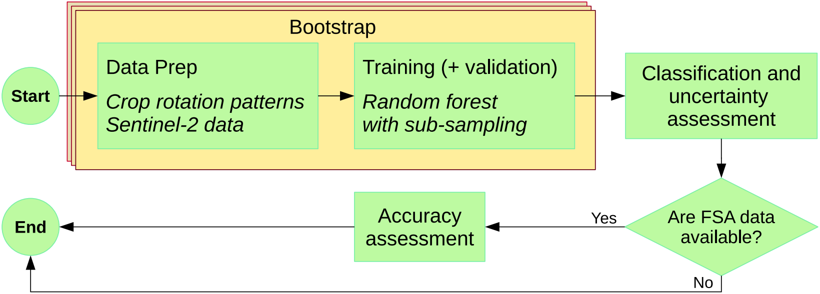
$$\hat{H} = \frac{-1}{\log K} \sum_{k=1}^{K} \hat{p}_k \log \hat{p}_k$$

where

- K denotes the total number of classes
- $\hat{p}_k$ represents the prediction frequency of the class k

The quantity $\hat{H}$ reflects the randomness of the empirical predictive distribution used to generate y^* rather than measuring its predictive accuracy

Implementation of the classification algorithm



Early season crop classification for Illinois

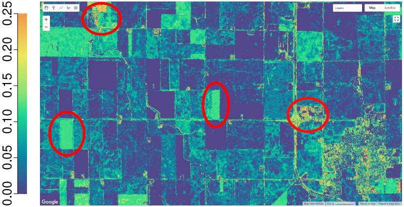
Classification settings used for training

- 16 bootstrap iterations
- 5,000 data points per training set
- 16 trees per forest trained at 90 m
- 2,500 random points per tree

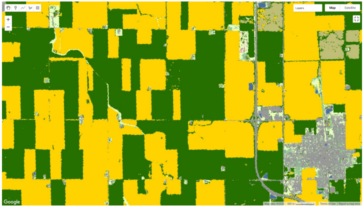
About 18 hours are required to classify the full state of Illinois

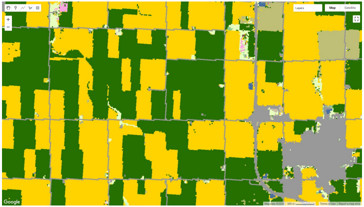
Crop classification in 2019

ESCDL (left) and uncertainty (right) layers at 10-m in June



End of season CDL at 10- (left) and 30- (right) m resolution





Corn Soybeans

Early classification accuracies (June)

The overall accuracy, and the producer and user accuracies for both corn and soybeans are computed with administrative data from Farm Service Agency (FSA)

$$\textbf{Producer accuracy} = \text{True neg.}/(\text{True neg.} + \text{False pos.})$$
$$\textbf{User accuracy} = \text{True pos.}/(\text{True pos.} + \text{False neg.})$$

Table: Accuracies (%) for ESCDL at 10-m resolution in Illinois

		Producer		User	
Year	**Overall**	**Corn**	**Soybeans**	**Corn**	**Soybeans**
2017	82.14	89.90	74.61	78.99	87.56
2018	83.80	90.80	77.73	80.28	88.66
2019	78.85	83.21	85.40	80.61	77.41

Conclusion

- Crop identification in June, which is early in the phenological cycle for corn and soybean in Illinois, is difficult; however, the uncertainty layer provides useful information to identify fields in which greater classification uncertainty occurs.
- The production and use of historical crop rotation patterns at 10-m resolution has been a necessary step in providing more accurate classification results
- Preliminary results show the potential of the ESCDLs in current and future NASS operations

Selected References

Boryan, C., Yang, Z., Mueller, R., and Craig, M. (2011). Monitoring US agriculture: the US Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer program. *Geocarto International*, 26(5):341–358.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. The Wadsworth statistics/probability series. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA.

Congalton, R. G. and Green, K. (2019). *Assessing the accuracy of remotely sensed data: principles and practices*. CRC press.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Freund, Y., Schapire, R. E., et al. (1996). Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Citeseer.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R. (2017). Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27.

Kuhn, M. and Johnson, K. (2013). *Applied predictive modeling*, volume 26. Springer.

Navarro, J. A. (2017). First experiences with google earth engine. In *GISTAM*, pages 250–255.

Thank you!

Questions?

Luca Sartore, PhD lsartore@niss.org
Claire Boryan, PhD claire.boryan@usda.gov