# DEVELOPING ENTROPIES OF PREDICTIVE CROPLAND DATA LAYERS FOR CROP SURVEY IMPUTATION

*Luca Sartore**

National Institute of Statistical Sciences

*Claire G. Boryan, Patrick Willis*

National Agricultural Statistics Service
United States Department of Agriculture

## ABSTRACT

A novel approach for crop-specific prediction of future crop planting and the development of corresponding uncertainty measures for all predictions is proposed. Using transition probabilities, predictive crop categories are first developed to predict crop-specific planting in the pilot study state of Illinois. Corresponding entropy layers are developed concurrently and can be used to flag survey sample units based on the level of uncertainty associated with the crop predictions. This allows survey units to be prioritized for imputation or for data collection based on whether the predicted value is sufficiently reliable. Further, the predicted acreage is assigned only to those survey units for which the prediction has the potential to be the sufficiently accurate. This approach can provide a solid methodology for reducing survey costs and farmer response burdens without introducing estimation bias or incurring severe losses of statistical efficiency. This has far-reaching implications for the sample design, quality, and timeliness of results for future surveys.

***Index Terms*—** Crop prediction, Entropy layers, Cropland Data Layers, Survey imputation

## 1. INTRODUCTION

Early season crop predictions (ESCP) are key inputs to crop monitoring and yield assessment models and decision support systems and are used by agribusiness. Recent research has explored how ESCP may be used to augment traditional large-scale surveys conducted by the United States Department of Agriculture (USDA) National Agricultural Statistics Service (NASS). Although crop-specific prediction has been conducted [1][2], these methods do not incorporate measures of quality for individual crop predictions, which significantly increases prediction utility specifically when used for survey imputation. This paper describes two methodological advances: 1) novel crop-prediction results through the application of Transition Probabilities (TP) and 2) a new method that provides entropy metrics for all crop predictions. The entropy metrics, defined as $H = \int p(x) \log\{p(x)\} \mathrm{d}x$, identify the predictability of the crop prediction results.

A variety of supervised methods are commonly used for land cover classification [3]. These methods include Maximum Likelihood, K-Nearest Neighborhood, K-means, Parallelepiped, Minimum Distance to the Mean, Decision Trees, Artificial Neural Network and Fuzzy classifiers. These supervised methods rely on sampling ground reference data to improve classification and computational performance. Land-cover prediction differs from supervised land cover classification because predictions are usually generated without satellite imagery [1]. Therefore, prediction of land cover categories is computed from historic land cover data and other ancillary information acquired in the past.

In this study, TP was the model selected to produce the predictive crop classifications known as predictive Cropland Data Layers (PCDLs) in this paper. TP can use all available crop rotation data inputs rather than just a sample of the data, which is common in supervised classification [4]. Using all available crop rotation data results in higher accuracies than algorithms that are limited by sampling schemes.

Some crop rotation patterns are uncommon and, consequently, more difficult to predict. Therefore, the categorical outcomes associated with certain areas have different degrees of uncertainty. In this work, entropy is used to quantify uncertainty, and an entropy layer is produced for each prediction layer. The relationship between uncertainty and prediction accuracy can be leveraged and applied in an area-frame survey, particularly for sample selection and imputation.

The objective of this study is to produce predictive crop specific land cover classifications using TP at 30 meters as well as produce entropy layers for the corresponding predictions. Further, an example illustrates how the PCDLs and entropies can be used in tandem to target samples for imputation in a large-scale survey.

The paper is organized as follows: The study area and test data are described in Section 2. The study methodology is introduced in Section 3 followed by results and discussion in

*The findings and conclusions in this paper are those of the authors and should not be construed to represent any official USDA or US Government determination or policy. This research was supported in part by the intramural research program of the US Department of Agriculture, National Agriculture Statistics Service.

Section 4. Finally, the conclusions are presented in Section 5.

## 2. STUDY AREA AND DATA

### 2.1. Study Area – Illinois, United States

Illinois, United States (U.S.) is the study area for this research and is highlighted in red in Figure 1. Illinois is located in the center of the agriculturally intensive Corn Belt in the Midwest of the U.S. Illinois' land area is about 35 million acres (14 million hectares) of which 24 million acres (10 million hectares) are cropland. In terms of U.S. production, Illinois ranks near the top of states for total planted area, where corn and soybeans are the two major crops.

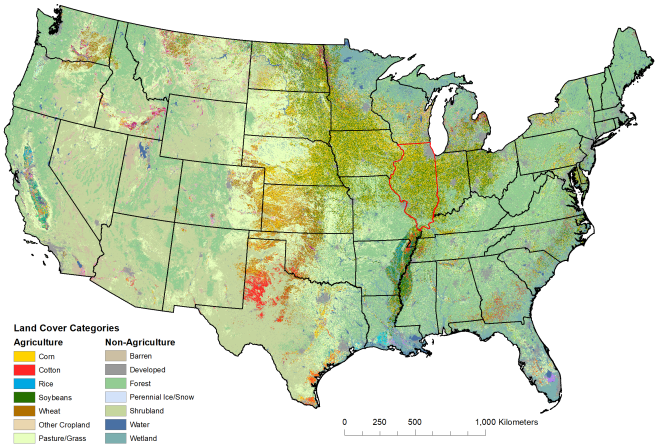


**Fig. 1**. National Agricultural Statistics Service – Cropland Data Layer. Illinois is highlighted in red.

### 2.2. Cropland Data Layers – 2008 forward

The NASS Cropland Data Layers (CDL) are raster, crop-specific land cover data sets produced at a 30-meter resolution. Moderate resolution satellite imagery, including Landsat 8 and Sentinel 2 A&B, acquired throughout the growing season are the satellite input data to a See5 decision tree classifier. Farm Service Agency Common Land Unit (CLU) and 578 administrative data are used as the crop specific training and validation data [5]. CDLs for all states in the conterminous U.S. are available since 2008. The CDLs are highly accurate achieving accuracies of 85-95% for major crops in large crop production states [6]. The CDLs are released on an annual basis following the growing season on the Cropland CROS web application `https://cropcros.azurewebsites.net/`.

### 2.3. Farm Service Agency Common Land Unit Data

Although the publicly available past CDLs are used as time series inputs for crop prediction, the corresponding past years' Farm Service Agency (FSA) Common Land Unit and 578 administrative data are used when available instead of the CDLs as these data are in situ ground reference data provided by farmers. Each year, farmers participating in a USDA program or purchasing crop insurance report the crops planted and the location of their fields, defined by the CLU polygons, to more than 2300 FSA offices across the U.S. [7][8]. In Illinois, approximately 99% of the cropland is identified in terms of CLU polygons. Farmers can report planting a single crop or multiple crops within a CLU. Only single crop CLUs are used as inputs for crop prediction for this study as the exact location of the crop can be determined.

### 2.4. NASS June Area Survey Segments

NASS's primary area frame-based survey is the June Area. Survey in which approximately 9,000 one square mile sample units (segments) are visited by survey interviewers at the beginning of each growing season to collect crop type and acreage information. Segments are randomly selected for in-person enumeration with a much larger proportion of segments selected in areas with greater than 50% cultivated land. For this study, the JAS segments in 2018, 2019 and 2021 are used to illustrate how the new predictive CDLs and entropy layers can be used to identify segments that may not require in-person enumeration and can be imputed.

### 2.5. Validation Data

The corresponding FSA CLU and 578 data, which are available at the end of the growing season, are used to validate the crop types in the same year's predictive CDL. Although the predictive CDLs can be produced before the growing season, the validation data are not available until the end of the growing season. The United States Geological Survey National Land Cover Database set points (for nonagricultural categories) are used as ground reference to validate the nonagricultural (water, urban, forest) categories [9].

## 3. METHODLOGY

### 3.1. Predictive CDLs

Conditioning on the relative frequencies of previously observed crop rotation patterns, a TP model is used to produce the state-level PCDLs that correspond to the highest predictive probability. The transition probability of a categorical stochastic process $X_t$ at the time $t$ is defined as $\pi_{j_1,\ldots,j_p,i} = \Pr(X_{t_i} = i | X_{t_1} = j_1, \ldots, X_{t_p} = j_p)$, where $p$ denotes the time length (in years) of the input pattern, and $i, j_1, \ldots, j_p \in$

$S$ represents the states of the process described by a sets $S$ of possible categorical outcomes.

The relative frequency of each land-cover change is proportional to the number of times that a specific pattern of length $p$ has been observed in the study area through time. However, if a new pattern has not been observed in the past data, shorter time windows are used. In the worst-case scenario, the counting process stops with the assessment of a non-time dependent distribution estimated as if dealing with a stationary distribution.

This simple procedure can be efficiently implemented on recent parallel computing technology using simultaneously the single-instruction multiple-data (SIMD) paradigm on several processing cores. Large areas can be processed within a few minutes without requiring the use of a sampling scheme to build the training set.

### 3.2. Entropy Layers

The evaluation of the output-class predictability can be related to the concept of intrinsic uncertainty of the model estimated for a given pattern. Thus, the entropy layer provides information that spatially varies accordingly to the land-cover changes previously observed.

Common patterns are more predictable than uncommon ones, therefore they have much lower entropy values and their predictions are often more accurate. Although the entropy layers can be used as a proxy for prediction inaccuracy, they should not be confused as such.

### 3.3. Targeting JAS Segments for Imputation

The PCDL and corresponding entropy layers are useful tools for guiding the acreage imputation of area-frame surveys, since they provide early indications of the crop types that are likely to be planted. Furthermore, the PCDL can also be used for more efficient and cost-effective sample-allocation procedures, or for reducing respondent burden by omitting survey questions regarding quantities available from other sources of information. However, the quality of the information provided at the segment level requires further assessment before being used for imputation in a survey.

To develop guidelines and thresholds for targeting high-quality segments, the ground reference data and the PCDL have been summarized within each JAS segment providing the acreage for major crops in Illinois (mostly corn, soybeans, and wheat). The entropy layer has also been summarized using the arithmetic average of the entropies within each JAS segment. These summary statistics allow one to conduct preliminary analyses on the acreage-error heteroscedasticity at the segment level as a function of average entropy. Therefore, one can derive a threshold based on a given empirical state-level quantile of the segment-average entropy such that certain conditions are satisfied. For instance, possible conditions can be set on the number of targeted segments or on the crop-specific expected variability of acreage error.

It has been observed (as shown in Section 4) that the smallest average entropies are associated with lower acreage errors and that the 20% empirical quantile can be a good threshold candidate selecting lower-entropy JAS segments for imputation procedures.

## 4. RESULTS AND DISCUSSION

Table 1 shows the corn and soybean producer and user accuracies for 2018, 2019 and 2021. The June Area Survey was not conducted in 2020 so is excluded from this study. Producer accuracy indicates omission or false negative error and the user accuracy indicates commission error or false positive error [10]. Corn and soybeans are mostly present in binary rotation patterns. Most of the accuracies in Table 1 are over 80%, indicating good prediction performances of the TP model. This is true even during a year affected by extreme weather, such as 2019, when heavy precipitation and spring snow melt resulted in flooding, saturated fields, and in some cases, farmers were unable to plant.

**Table 1**. Producer and user accuracy for corn and soybeans computed with respect to FSA ground reference data.

| | Corn | | Soybean | |
|---|---|---|---|---|
| Year | Producer | User | Producer | User |
| 2018 | 86.7% | 88.1% | 86.2% | 83.5% |
| 2019 | 84.2% | 83.3% | 88.3% | 76.7% |
| 2021 | 78.3% | 85.2% | 86.2% | 79.8% |

Table 2 shows the mean absolute error (MAE) and mean error (ME) for corn and soybeans measured in acres per segment. Both MAE and ME have been computed using only the segments flagged for imputation. This assumes that there are no errors from unflagged segments, which undergo in-person enumeration. The MAE is the mean prediction error, which is comprised of both bias and variability of the predictive acreage assessment, and the ME indicates the presence of bias. Both the MAE and ME vary across time; however, the MAE for corn is more stable than the MAE for soybeans. The ME in 2018 shows a negative bias for corn and a positive bias for soybeans, while the MEs for the following years are negative for corn and almost zero for soybeans. The segment-level errors (acreage per segment) reported in Table 2 are quite small in general; in fact, the largest MAE is about the 3% of the average size of a JAS segment (about 648 acres) in Illinois.

The relationship between acreage error and segment-level entropy for corn (Figure 2) and soybeans (Figure 3). In both cases, the errors are well-centered on zero, indicating the absence of bias in the predicted acreage. However, the most interesting aspect of these figures is the heteroscedastic be-

**Table 2**. Mean absolute error (MAE) and mean error (ME) for corn and soybeans measured in acres per segment.

| Year | Corn | | Soybean | |
|---|---|---|---|---|
| | MAE (ac / seg.) | ME (ac / seg.) | MAE (ac / seg.) | ME (ac / seg.) |
| 2018 | 25.49 | -15.95 | 25.06 | 16.81 |
| 2019 | 20.65 | -9.37 | 15.37 | -0.84 |
| 2020 | 17.39 | -2.15 | 16.54 | -0.09 |

havior of the error; in fact, as the average entropy computed at the segment level increases, the error variability increases. Furthermore, the chosen threshold at the 20% empirical quantile (shown as a red vertical line) seems to provide a reasonable level to flag segments that provide more accurate acreage predictions. Consequently, only segments below the red vertical line would be recommended for imputation.

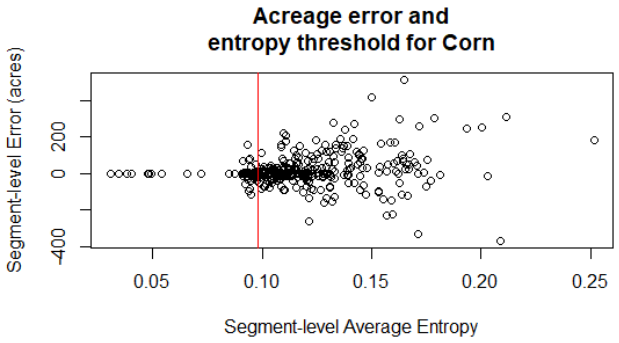


**Fig. 2**. Average entropy versus acreage error at the segment-level for corn in 2021. Red vertical line shows the chosen threshold.

## 5. CONCLUSIONS

This paper proposes the use of TP for crop-specific prediction and the development of corresponding entropy layers that can be used to assess the level of crop prediction uncertainty. The paper introduces a method to prioritize survey units for imputation or for data collection based on whether the predicted value is sufficiently reliable. This novel approach can provide a solid methodology for reducing survey costs and farmer response burdens without introducing estimation bias or incurring severe losses of statistical efficiency.

## 6. REFERENCES


[1] Chen Zhang, Liping Di, Li Lin, and Liying Guo, "Machine-learned prediction of annual crop planting in the us corn belt based on historical crop planting maps," *Computers and Electronics in Agriculture*, vol. 166, pp. 104989, 2019.


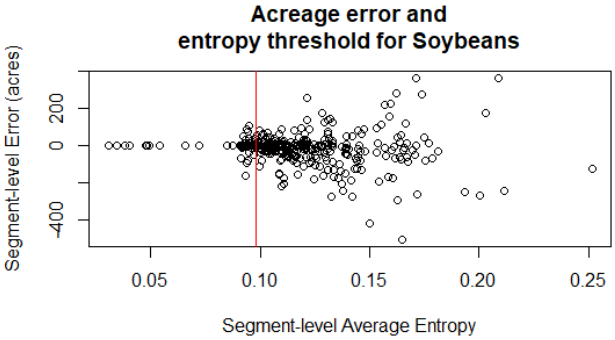


**Fig. 3**. Average entropy versus acreage error at the segment-level for soybeans in 2021. Red vertical line shows the chosen threshold.


[2] David M Johnson and Richard Mueller, "Pre-and within-season crop type classification trained with archival land cover information," *Remote Sensing of Environment*, vol. 264, pp. 112576, 2021.

[3] Minu Nair and JS Bindhu, "Supervised techniques and approaches for satellite image classification," *International Journal of Computer Applications*, vol. 134, no. 16, 2016.

[4] Lucia Morales-Barquero, Mitchell B Lyons, Stuart R Phinn, and Chris M Roelfsema, "Trends in remote sensing accuracy assessment approaches in the context of natural resources," *Remote sensing*, vol. 11, no. 19, pp. 2305, 2019.

[5] Claire Boryan, Zhengwei Yang, Rick Mueller, and Mike Craig, "Monitoring us agriculture: the us department of agriculture, national agricultural statistics service, cropland data layer program," *Geocarto International*, vol. 26, no. 5, pp. 341–358, 2011.

[6] USDA NASS, "Cropland data layer metadata," 2021.

[7] USDA FSA, "Common land units (clus)," 2021.

[8] J Heald, "USDA establishes a common land unit. ESRI ArcUser Online," 2002.

[9] James Wickham, Stephen V Stehman, Daniel G Sorenson, Leila Gass, and Jon A Dewitz, "Thematic accuracy assessment of the nlcd 2016 land cover for the conterminous united states," *Remote Sensing of Environment*, vol. 257, pp. 112357, 2021.

[10] Russell G Congalton and Kass Green, *Assessing the accuracy of remotely sensed data: principles and practices*, CRC press, 2019.