Automatic Imputation for an Area Survey

Tara Murphy[1], Arthur Rosales[1], Luca Sartore[1,2], Denise A. Abreu[1]
[1]National Agricultural Statistics Service, USDA, 1400 Independence Avenue, Washington DC 20250
[2]National Institute of Statistical Sciences, 1750 K Street NW Suite 1100, Washington, DC 20006

Abstract
The U.S. Department of Agriculture's (USDA) National Agricultural Statistics Service's (NASS's) June Area Survey (JAS) is an annual survey based on an area frame, which has complete coverage of the contiguous US. Data for this survey are collected via in-person interviews. NASS employs manual imputation for JAS nonresponse, which is becoming increasingly costly as response rates are declining. Moreover, it can be difficult to measure the data quality resulting from these efforts. We are proposing a new automatic imputation approach that uses a unique combination of data sources, including historic satellite imagery, digital geospatial archive of the sampled areas of interest, and administrative data. This paper evaluates the quality of the proposed automatic imputation approach.

Key words: Imputation, Area frame, Nonresponse, Geospatial, Administrative data, Multiple data sources

1. Introduction and Background

1.1. The June Area Survey

The United States Department of Agriculture's (USDA) National Agricultural Statistics Service (NASS) has developed and used area sampling frames since 1954 for the purpose of conducting surveys to collect information about US agriculture. The primary area frame survey conducted by NASS is the June Area Survey (JAS), which is based on a frame that covers all land in all US states except for Alaska. JAS data are primarily used to provide direct estimates for crop acreages and livestock inventories, as well as to measure sampling coverage of the NASS list frame. The JAS is one of the largest annual NASS survey projects and incurs the largest data collection costs to NASS, outside of the Census of Agriculture and reimbursable surveys (Cotter et al., 2010). This is due to the immense data collection effort required to conduct in-person interviews. The area frame is composed of primary sampling units (PSUs), which are stretches of land delineated by permanent boundaries, such as roads, rivers, etc. The PSUs are stratified by the land-use occurring within their bounds, particularly by the amount of crop cultivation. The JAS survey has a multi-stage design, where secondary sampling units called "segments" are selected for enumeration. These segments are approximately 1 square mile in size and may contain multiple "tracts" of land, where each tract represents a unique farm operation. Approximately 9,000 segments are sampled each year. Segments are not designed to contain complete farm operations and frequently contain only part of the land operated by a particular farm. Figure 1 provides an example of a segment, as well as the tracts within.

Figure 1. The segment boundary is drawn in red, and the tracts within are drawn and labeled in blue.

The JAS sample has a panel design with a five-year rotation scheme, where selected segments are measured for five consecutive years before being rotated out and replaced. Rotation is staggered so that segment rotation occurs every year and approximately 20% of the sample rotates out and in each year.

Of note, data collected for a segment applies to two different scopes: 1) all land specifically within the tracts of the segment, and 2) all acreage and items of interest associated with the farm operations identified within the segment. NASS estimation procedures rely on information from both components (Vogel, 1995). For the remainder of this paper, the focus is on the land-in-tracts component, unless otherwise stated.

1.2. Nonresponse

When compared with other NASS surveys, the challenge of nonresponse in the JAS is unique. For one, data collection has traditionally been based entirely on in-person interviews. This restricts the time and money that can feasibly be spent on contact attempts and refusal conversions. Additionally, the sampled segment only provides a broad area for which agricultural activity must be accounted for. Extensive screening activities are needed to identify in-scope land tracts for agricultural questions and potential farm operators, but these efforts may not always produce adequate contact information for conducting an interview. Moreover, land-use arrangements within a segment may change during the five-year period for which it remains in the sample. These changes may result in the need to add or remove farm-operations, or to account for land entering or leaving agricultural production within the segment. Finally, digital records of tract boundaries have historically never been created, making it difficult to link external, ancillary data to those tracts. All these factors have caused addressing tract-level nonresponse to be a difficult undertaking, where the tract itself can be ephemeral.

NASS has employed various techniques to address nonresponse in the JAS, including re-interview efforts, nonresponse weighting, visual observations by field interviewers, and manual imputation relying on previously reported data when available (Cox, 1993; Cotter et al., 2010). NASS has never applied automatic imputation techniques for JAS tract-level data, although imputation methodologies are applied for farm-level data (Wesley, 1991). However, imputing data for non-responding tracts was identified as a potential area for improvement in the process of estimating the number of farms in the US (Lopiano, 2010).

Declining response rates for the JAS in recent decades (Gerling et al., 2010; Price, 2017) have provided impetus for seeking innovative approaches for data collection and addressing nonresponse. The problem of JAS nonresponse was exacerbated by conditions surrounding the COVID-19 (SARS-CoV-2) pandemic. Data collection for JAS segments was halted completely in 2020, and 2021 data collection was performed via telephone interviews as opposed to field enumeration, which resulted in increased questionnaire editing and manual imputation due to higher nonresponse. Fortunately, greater technological capabilities as well as a focus on employing external data sources, such as administrative data, satellite imagery, and geospatial information, have supported the broader application of manual imputation for nonresponse in JAS tracts. However, this process relies on considerable staff labor hours and is subject to inconsistent decision-making among those staff when making imputation decisions. The unique challenges of JAS nonresponse and the absence of digital tract boundary records have prevented NASS from using automatic imputation techniques for JAS tracts in the past.

1.3. New Data

1.3.1. Digitized Tracts

In response to the large cost of the JAS, nonresponse issues, and recent data collection challenges, NASS allocated resources to a new JAS tract digitization effort. For this effort, all tracts in the 2021 JAS sample were digitized, and all new segments rotating into the sample in subsequent years are to be digitized immediately following the data collection period. This offers a new capability to perform computerized analysis of JAS tracts within segments, including the application of external data sources to assist with automatic imputation.

The newly digitized tracts are created by NASS field office staff, who are in possession of JAS photo enlargements of selected segments at the end of each data collection period. These photos are freshly updated with new tract level information such as updated boundary lines and tract IDs corresponding to tabulated tract-level data. The new digitized records are uploaded to a centralized database where they can be utilized for analysis. Figure 2 illustrates digitized tracts within a segment.

Figure 2. In this digital image, the segment boundary is drawn in red, and the tracts within are drawn in various colors.

1.3.2. Cropland Data Layer

One dataset that is particularly useful for staff conducting manual imputation is the Cropland Data Layer (CDL). The CDL is a geo-referenced, land cover classification dataset providing complete coverage of the contiguous US (Boryan et al., 2011). The CDL has been produced nationally by NASS every year since 2008. The data are composed of pixels that represent 30m x 30m squares of land and are encoded with over 100 crop type categories, as well as non-agricultural landcover categories such as urban and woodland. An example of the CDL is provided in Figure 3. When multiple years of the CDL are viewed in sequence, it is possible in some cases to make assumptions about crop rotation patterns within crop fields, which are reasonably reliable.

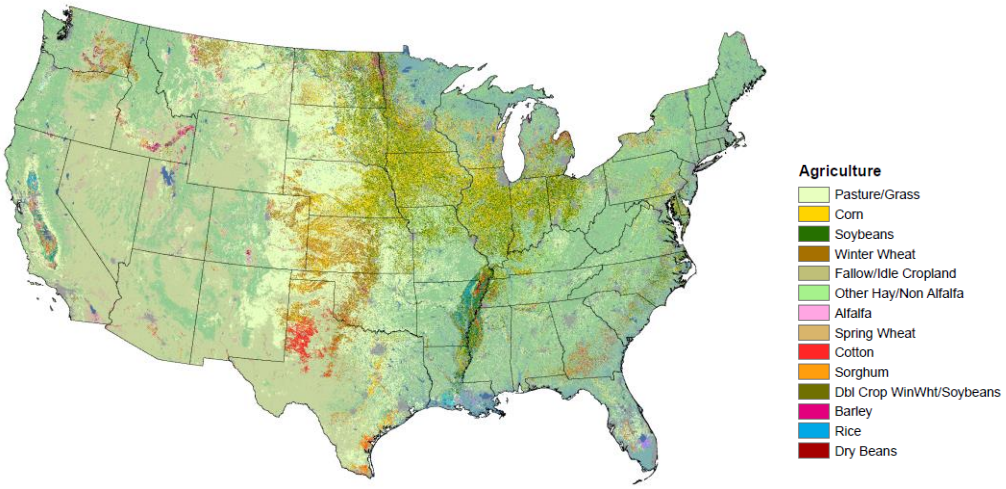


Figure 3. An example of the CDL, composed of pixels representing 30m x 30m squares of land, zoomed out to the contiguous US. The legend on the right is a partial list of major crop categories.

In the context of JAS imputation, a concern with the CDL is that it requires an entire year of satellite imagery and end-of-year ground truth data to be created. Therefore, available CDL data always lags behind the current growing season by a year.

1.3.3. Farm Service Agency Data

In recent years, NASS has had an increased interest in utilizing non-survey data to augment and support survey efforts. One such administrative data source is Farm Service Agency (FSA) Form-578 data. Farmers participating in USDA programs or purchasing crop insurance must report their crop plantings each year. These reports include crop field locations, where each crop field boundary is defined by a Common Land Unit (CLU) polygon (Heald 2002; USDA, 2022). The FSA creates new CLUs each year and updates them with new Form-578 data continuously throughout the growing season as farmers report to the over 2300 FSA offices across the US. Examples of FSA CLUs are provided in Figure 4.

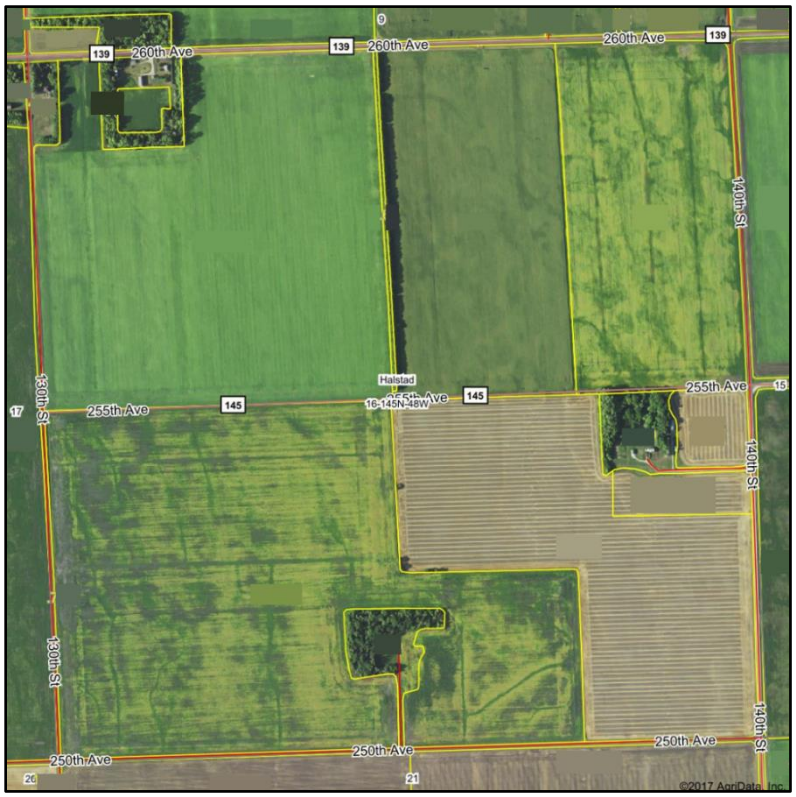


Figure 4. An example of FSA CLUs, outlined in yellow. Each polygon represents an agricultural field, excluding non-agricultural areas such as farmsteads.

1.3.4. Predictive Cropland Data Layer

In recent years, there have been efforts at NASS to create in-season land-cover classification products as alternatives to the CDL. One such product is the Predictive Cropland Data Layer (PCDL), which was newly developed in 2021. The PCDL is a raster dataset providing cropland predictions that are computed by combining information from both the CDL and the FSA data. The categorical predictions correspond to the most likely classes, that is, those with the highest transition probabilities. These probabilities are estimated by assuming a Higher-Order Markov Chain (HOMC) model that takes as input the best crop-rotation data available from the previous three years (Raftery, 1985). Because

FSA data are reported and the CDL data are estimated, the FSA data are preferred to the information provided by CDL, while the CDL data are used only for the areas without FSA coverage. The estimation is performed using a rolling window starting five years earlier than the year to predict, as shown in Figure 5. This approach is equivalent to data augmentation techniques that allow one to leverage the temporal dynamics for gaining precision by increasing "artificially" the information used to fit the model and tuning parameters (Frühwirth-Schnatter, 1994; Hyndman and Athanasopoulos, 2018).

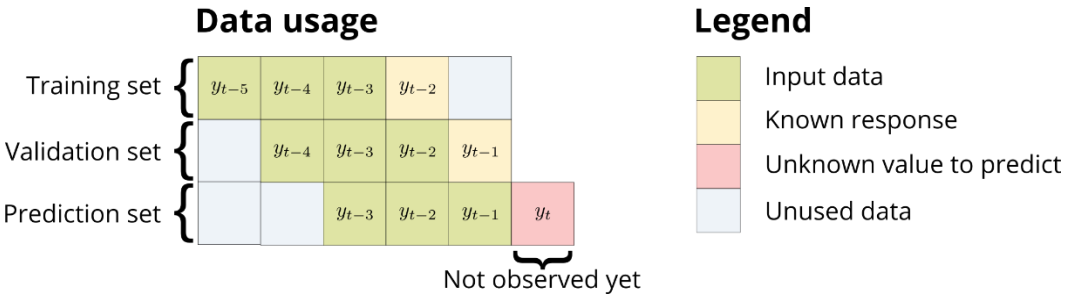


Figure 5. Visual of the training, validation, and testing (or prediction) datasets for the PCDL.

1.3.5. Entropy Layer

The Entropy Layer is a companion dataset to the PCDL. In fact, while the PCDL consists of current year, multi-class predictions over the continental US, the Entropy Layer provides a measure of uncertainty of the categorial predictions. This conversely translates to more confidence in the quality of the resulting predictions for the areas that are characterized by low entropy values (Shannon, 1951; Kolmogorov, 1956). The use of the Entropy Layer as a measure of the quality of the PCDL predictions can be verified by comparing the predictive accuracy of the model using past PCDL and FSA data. Although the relation between entropy and prediction accuracy may not be always stable through time, the combined information from the Entropy Layer and the PCDL provides several opportunities to develop more sophisticated approaches to improve surveys and related statistical approaches.

1.4. Automatic Imputation Approach

The targets for imputation are the crop-specific acreages within each tract. The unique benefit of using the PCDL for tract-level imputation is that it provides an in-season crop type prediction. Additionally, due to the "wall-to-wall" nature of the PCDL, we have complete availability of this external dataset regardless of JAS data collection. The new digitized tracts allow the PCDL data to be limited to the JAS tracts, and FSA data allow for an assessment of the quality of PCDL-based imputation for these tracts.

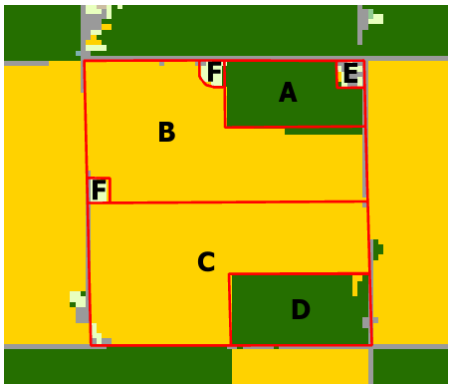
This paper explores a model-based approach to creating tract-level crop acreage predictions, which in turn could be used to impute for item-nonresponse automatically, as an alternative to the more labor-intensive and time-consuming manual imputation approach. Additionally, the viability of using tract-level summaries of PCDL values to impute JAS items directly is explored. The Entropy Layer is utilized to identify optimal candidates, allowing for an automatic imputation process.

2. Methodology

2.1. Data Preparation

For this study, 2019 and 2021 JAS data were utilized because of the availability of interview data, the creation of the 2019 and 2021 versions of the PCDL, and the creation of digitized tract boundaries performed for the 2021 JAS. Due to the five-year panel design of the JAS, the 2021 sample contains elements that were also in the 2019 sample. Thus, the 2019 JAS data used consisted only of those segments that were also in the JAS sample in 2021 and were digitized. Subsets were created using ArcGIS Pro geospatial software.

Once the subsets of digitized tracts were created, the national 2019 and 2021 PCDLs were tabulated within the digitized tract boundaries for the respective survey year tracts. PCDL pixels that straddled the boundary were considered in the tabulation only if the majority of the pixel was contained by the tract. This resulted in a data frame summarizing PCDL data for each tract, where each row represented a digitized tract, and each column represented a crop type or other land cover type. Each cell in this tabulation represented a pixel count within the tract boundary.



Tract ID	Cropland-indicator	Corn Acres	Soybean Acres	Other Crop Acres	Non-Crop Acres
A	1		33		
B	1	119			
C	1	122			
D	1		34		
E	0				1
F	0				2

Figure 6. An illustration of the tabulation of PCDL data within JAS digitized tracts. The different colored swaths of area represent a pixel in the PCDL, shaded by a land cover type. The red boundary lines represent digitized tracts. The table beneath demonstrates the result of the tabulation.

This process was repeated for the Entropy Layer data, providing summary statistics of entropy within each digitized tract. Next, the tabulated PCDL and Entropy Layer data were joined to JAS reported data. JAS data included tract ID, reported crop types and acreages, and other design information such as JAS strata (level of cultivation and urbanicity). The

two datasets were joined based on unique state-segment-tract ID. This process was completed using the R software.

FSA CLUs are geospatially referenced polygon vector data in their native format. Although farmers may report multiple crops within a CLU, only single-crop CLUs were used in this study so that the exact location of crop types could be used as ground truth. These single-crop CLUs were then rasterized to form a 30m x 30m grid that overlapped the CDL and PCDL grids exactly. The rasterization process resulted in removal of some very small CLUs that were smaller than a 30m x 30m pixel. This resulting dataset was tabulated at the digitized tract level with the same process used for the PCDL, Entropy Layer, and JAS data.

2.2. Exploratory Analysis

Automatic imputation of tract level information in JAS would be useful if it could accurately predict crop types and acreages within tract boundaries. NASS considers administrative FSA data to be the best available alternative to directly collecting field-level ground truth, in terms of accuracy and reliability. Therefore, a predictive model was developed where the acreage of a particular FSA crop type within digitized tracts was the dependent variable and the PCDL, Entropy Layer, and other JAS administrative variables that would normally be available at the time of JAS data collection were used as predictors. For this paper, the focus was on two major US commodities: corn and soybeans.

An exploratory analysis was performed to investigate the similarity between responding and nonresponding tracts in terms of key characteristics. In general, tracts were found to be similar across response and nonresponse cases for these characteristics.

2019

Stratum/Response Type	Responded	Nonresponse/ Imputed	Total
High Agricultural Area	8858 (61.9%)	4399 (64.5%)	13257 (62.8%)
Medium Agricultural Area	3927 (27.5%)	1795 (26.3%)	5722 (27.1%)
Low Agricultural Area	1437 (10.0%)	587 (8.6%)	2024 (9.6%)
Agri-Urban Area	61 (0.4%)	29 (0.4%)	90 (0.4%)
Non-Agricultural Area	16 (0.1%)	8 (0.1%)	24 (0.1%)
Total	14299 (100.0%)	6818 (100.0%)	21117 (100.0%)

2021

Stratum/Response Type	Responded	Nonresponse/ Imputed	Total
High Agricultural Area	9307 (62.6%)	8636 (63.6%)	17943 (63.1%)
Medium Agricultural Area	4058 (27.3%)	3644 (26.8%)	7702 (27.1%)
Low Agricultural Area	1434 (9.6%)	1244 (9.2%)	2678 (9.4%)
Agri-Urban Area	69 (0.5%)	41 (0.3%)	110 (0.4%)
Non-Agricultural Area	10 (0.1%)	13 (0.1%)	23 (0.1%)
Total	14878 (100.0%)	13578 (100.0%)	28456 (100.0%)

Table 1. 2019 and 2021 JAS tracts' response type by sampling stratum.

2.3. Cubist Model

For predicting the planted corn acreage and planted soybean acreage, Cubist (Kuhn et al., 2013) models were implemented using the R caret package (Kuhn et al., 2007). Cubist models are similar to decision trees, where each node is a conditional rule and each leaf is

a linear model. Cubist allows for multiple model trees to be considered (committees) and for individual predictions to be based on a set of similar instances (neighbors), if desired. A cubist model was fit for each crop and similar predictor variables were used in each case, as summarized in Table 2. The cubist models were run using 10-fold cross validation, and default settings for committee and nearest neighbor parameters, resulting in as many as ten committees and no nearest neighbors for prediction.

Predictor Variable	**Variable Description**
PCDL corn pixels	PCDL pixel count of single-cropped corn within digitized tract
PCDL combined corn pixels	PCDL pixel count of single and double-cropped corn within digitized tract
PCDL soybean pixels	PCDL pixel count of single-cropped soybeans within digitized tract
PCDL combined soybean pixels	PCDL pixel count of single and double-cropped soybeans within digitized tract
PCDL winter wheat pixels	PCDL pixel count of single-cropped winter wheat within digitized tract
PCDL spring wheat pixels	PCDL pixel count of single-cropped spring wheat within digitized tract
PCDL durum wheat pixels	PCDL pixel count of single-cropped durum wheat within digitized tract
PCDL non-crop pixels	PCDL pixel count of non-crop land use
PCDL crop pixels	PCDL pixel count of any crop land use
PCDL max crop	Dominant PCDL crop type
PCDL max crop 2	Second dominant PCDL crop type
PCDL crop count	Count of PCDL crop types in tract
Entropy Mean	Mean entropy of pixels in tract
Entropy Median	Median entropy of pixels in tract
Latitude	Latitude of segment
Longitude	Longitude of segment
Stratum	Sample design stratum code of segment
State	State code of segment
Digitized tract acres	Size of tract in digitized acres

Table 2. Cubist models' predictor variables for corn and soybeans models.

To investigate the plausibility of generating predictions before the in-season survey cycle, models were fit using 2019 PCDL and JAS administrative data to make predictions on the corresponding 2021 data. This eliminated the need for 2021 survey data to make predictions. Moreover, the 2019 and 2021 datasets were filtered down to fewer, simpler tracts to improve model performance at the sacrifice of number of tracts provided predictions. Mainly, those records that met recommended PCDL reliability based on Entropy Layer values and those records that only contained one PCDL land-cover type within its boundaries. The records in the new datasets were considered "low hanging fruit" records.

Number of PCDL crops in tract < 2
Mean entropy of tract < 0.1 (Sartore, et al., 2022)
Digitized tract acres between 10 and 1000 acres

Table 3. Definition of "low hanging fruit" records.

2.4. Beyond Cubist Model

Additionally, the performance of the tract-level summaries of PCDL values as a predictor of FSA corn or soybeans acreage was investigated. For this analysis, the data were again limited to "low hanging fruit" as described above. The PCDL values themselves are the result of advanced modeling and are designed specifically to predict FSA values for a given area in a way that is compatible with the CDL (i.e., 30m x 30m raster).

3. Results

Once the datasets were filtered to "low hanging fruit" records, the number of records used from 2019 to fit the model dropped from 21,117 records to 2,712 records, and the number of records used from 2021 for analysis dropped from 28,456 records to 1,510 records.

3.1. Cubist Models

For the FSA corn acreage prediction model, the R^2 was found to be 0.781 and the Mean Absolute Error (MAE) was found to be 4.783 acres against the FSA data from 2021. The sole important model variable was PCDL corn pixels, as shown in Figure 7. Figure 8 shows the modeled pixel-based corn acreage against the FSA corn acreage for 2021.

```
Attribute usage:
  Conds  Model

          100%    PCDL_CORN
```

Figure 7. Variable importance in the corn model in terms of percent of time a variable was used as a condition and/or in a linear model.

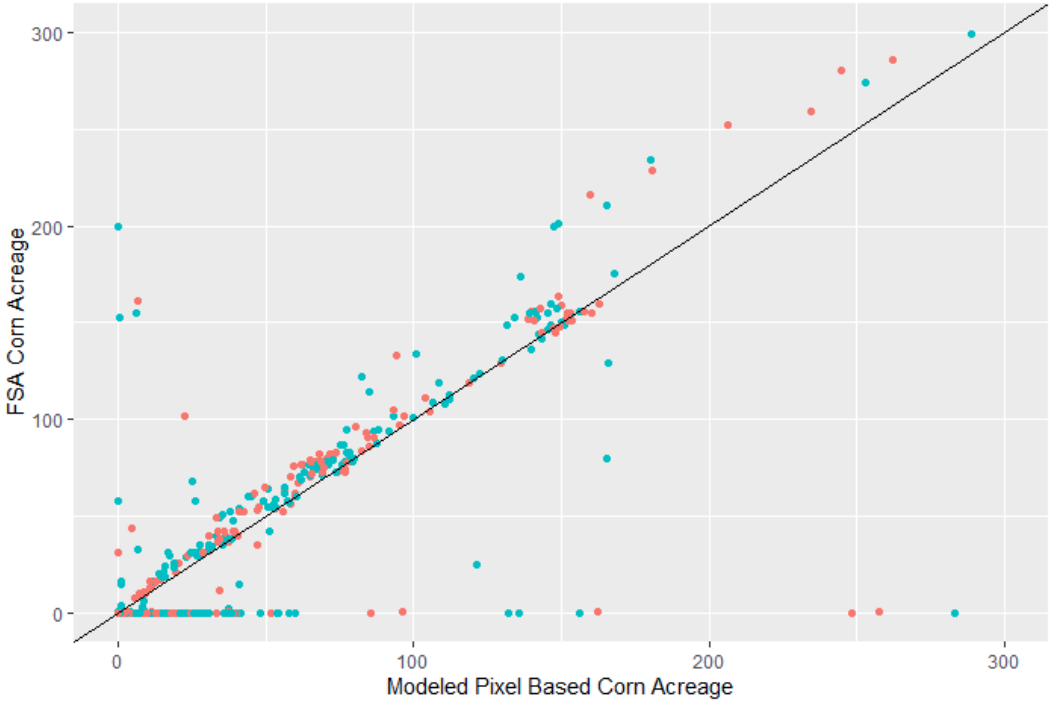


Figure 8. Modeled Pixel Based Corn Acreage against FSA Corn Acreage for 2021

For the FSA soybean acreage prediction model, the R^2 was found to be 0.86 and the MAE was found to be 2.82 acres against the FSA data from 2021. The important model variables included PCDL combined soybean pixels, PCDL soybean pixels, digitized tract acres, state, entropy mean, entropy median, latitude, and longitude, as shown in Figure 9. Figure 10 shows the modeled pixel-based soybean acreage against the FSA soybean acreage for 2021.

```
Attribute usage:
  Conds  Model

   98%    92%    PCDL_new_soy
    5%    99%    PCDL_SOYBEANS
    1%           MLONG
    1%           MLAT
           4%    digitized_tract_acres
           3%    state
           2%    ENTROPY_MEAN
           2%    ENTROPY_MEDIAN
```

Figure 9. Variable importance in the soybean model in terms of percent of time a variable was used as a condition and/or in a linear model.

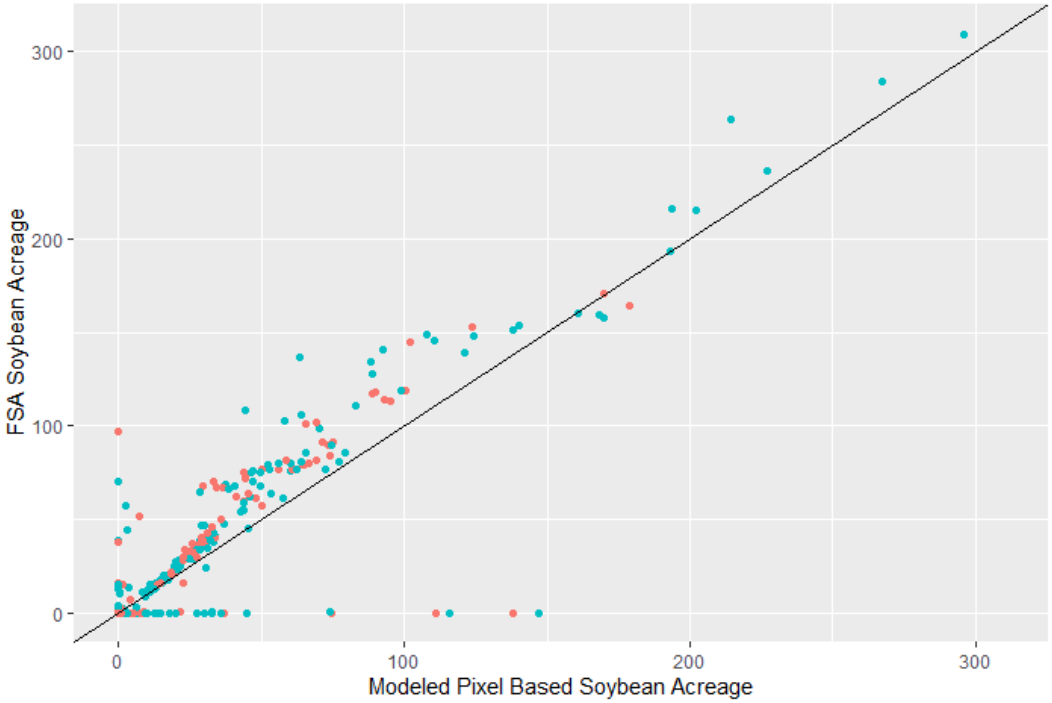


Figure 10. Modeled Pixel Based Soybean Acreage against FSA Soybean Acreage for 2021

3.2. Beyond Cubist Models

The accuracies for FSA corn acreage using the tract-level summaries of PCDL values included an R^2 of 0.807 and an MAE of 2.797 acres. Figure 11 shows the 2021 PCDL corn acreage against the 2021 FSA corn acreage. Recall the PCDL does not incorporate any in-season data and is merely a prediction based on crop rotation patterns.

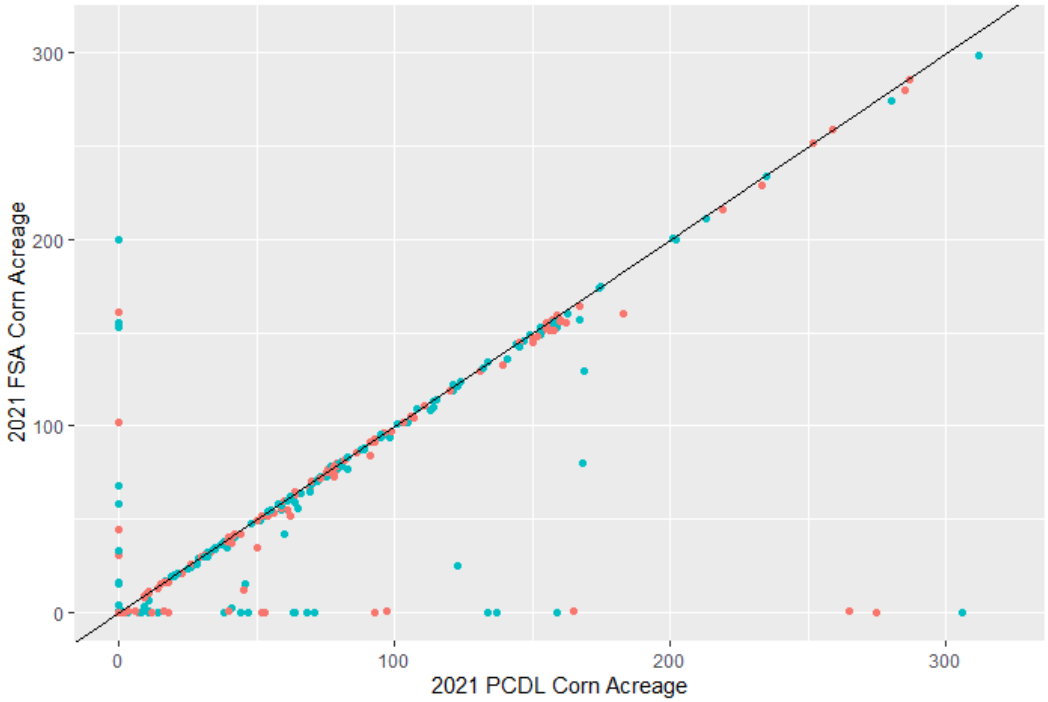


Figure 11. 2021 PCDL Corn Acreage against 2021 FSA Corn Acreage

The accuracies for FSA soybean acreage using just the PCDL included an R^2 of 0.884 and an MAE of 1.336 acres. Figure 12 shows the 2021 PCDL soybean acreage against the 2021 FSA soybean acreage.

Figure 12. 2021 PCDL Soybean Acreage against 2021 FSA Soybean Acreage

4. Discussion

For tract records identified as "low hanging fruit," the tract-level summaries of PCDL values outperformed the proposed cubist model predictions for corn and soybeans in 2021. Among both the cubist model predictions and the PCDL value summaries, many of the predictions were for the zero or low acreage cases for corn and soybeans. This indicates that either imputation approach may be useful for classifying tracts as not containing these commodities. While this study showed promising potential for US corn and soybeans, the potential usefulness of the PCDL for accurately predicting other crop types targeted by the JAS at the tract level needs to be investigated.

Limiting available data to only "low hanging fruit," resulted in highly reliable predictions for corn and soybeans, but it removed a vast amount of the original data. Future work is needed to investigate the tradeoff between accuracy achievable at the tract level and number of cases for which predictions can meet desired accuracy benchmarks. Perhaps error levels normally accepted by NASS for the JAS survey could be used to define new Entropy Layer thresholds or other factors to increase the number of records for which this imputation approach is acceptable. In general, future work is needed to determine optimal entropy for the JAS tract. This may depend on commodity, geography, or other features of tracts.

Although the tract-level summaries of PCDL values outperformed the cubist model predictions in this study, this approach is still blind to in-season events such as storms, economic scenarios, or other abnormalities, which may occur in given year that the PCDL may not be able to represent. Further work includes investigating the incorporation of additional auxiliary data to use in modeling, such as economic data (e.g., fertilizer prices, and future prices paid to farms) and environmental data (e.g., temperature, soil moisture, vegetation indices, storm locations and severity).

All input data and FSA data were rasterized to the same 30m x 30m grid, which provided many benefits to the ease of processing and preparation. However, it is possible that this approach results in an overstatement of agreement between predicted acreage and actual planted acreage since the grid does not perfectly align with actual field locations and size. This potential systematic bias resulting from differences between actual field locations and sizes and rasterized data is not accounted for in this study's findings.

The advent of new data sources has opened the possibility of in-season automatic imputation. The PCDL provides same-year predictions of crop types spanning the entire US; the digitized tracts accurately limit the data to the scope of JAS survey items; and the Entropy Layer provides a meaningful measure of uncertainty associated with PCDL prediction for a given tract. Using these data in tandem allows researchers to automatically target JAS tracts for imputation and to provide reliable estimates of crop acreages for major US commodities.

Acknowledgements

The findings and conclusions in this report are those of the author(s) and should not be construed to represent any official USDA or U.S. Government determination or policy. This research was supported by the intramural research program of the USDA, NASS.

References

Boryan, Claire, et al. "Monitoring US agriculture: the US department of agriculture, national agricultural statistics service, cropland data layer program." *Geocarto International* 26.5 (2011): 341-358.

Cotter, Jim, et al. "Area frame design for agricultural surveys." *Agricultural survey methods* (2010): 169-192.

Frühwirth-Schnatter, Sylvia. "Data augmentation and dynamic linear models." *Journal of time series analysis* 15, no. 2 (1994): 183-202.

Gerling, Michael W., HoaiNam N. Tran, and Terry P. O'Connor. *The Road to Understanding Nonresponse in the June Area Survey*. No. 1496-2016-130667. 2010.

Heald, J. "USDA establishes a common land unit." *ArcUser Online* (2002).

Hyndman, R. J., and Athanasopoulos, G.. *Forecasting: principles and practice*. OTexts, 2018.

Kolmogorov, A. (1956). On the Shannon theory of information transmission in the case of continuous signals. *IRE Transactions on Information Theory, 2*(4), 102-108.

Kuhn, Max, et al. "The caret package." *Gene Expr* (2007).

Kuhn, Max, and Kjell Johnson. *Applied predictive modeling*. Vol. 26. New York: Springer, 2013.

Lopiano, Kenneth K., et al. "Adjusting the June Area Survey for Non-response and Misclassification." *Proceedings of the Joint Statistical Meetings*. 2010.

Raftery, Adrian E. "A model for high-order Markov chains." *Journal of the Royal Statistical Society: Series B (Methodological)* 47, no. 3 (1985): 528-539.

L. Sartore, C. G. Boryan and P. Willis, "Developing Entropies of Predictive Cropland Data Layers for Crop Survey Imputation," *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 2022, pp. 1404-1407, doi: 10.1109/IGARSS46834.2022.9884059.

Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell system technical journal, 30*(1), 50-64.

USDA FSA, “Common land units (CLUs),” 2022.

Vogel, Frederic A. "The evolution and development of agricultural statistics at the United States Department of Agriculture." *JOURNAL OF OFFICIAL STATISTICS-STOCKHOLM-* 11 (1995): 161-161.

Zimmer, Stephanie, Jae Kwang Kim, and Sarah Nusser. "Automatic stratification for an agricultural area frame using remote sensing data." *Proceedings of the 59th ISI World Statistics Congress*. 2013.