Pre-Season Crop Type Forecasting with an Application to Area Survey Imputation

Jonathon Abernethy[1], Arthur Rosales[1], Tara Murphy[1]
[1]United States Department of Agriculture, National Agricultural Statistics Service, 1400 Independence Ave SW, Washington, DC 20250

Abstract
The United States Department of Agriculture's (USDA's) National Agricultural Statistics Service (NASS) provides timely and accurate statistics in service to United States (U.S.) agriculture. The June Area Survey (JAS), whose sample is drawn from the NASS area frame, provides an early season estimate of crop-specific planted acreage for a variety of crops. These estimates, with the June Crops Acreage, Production, and Stocks Survey, informs the official June acreage estimate. For the JAS, tracts of land are drawn, and surveyed farmers report the crop types planted in each tract. As in any survey, tract-level nonresponse is possible. Currently, NASS uses manual imputation to mitigate tract-level nonresponse, where staff use historical data to fill in missing crop types. An alternative, automated approach is proposed that does not require human intervention. The approach uses machine learning models to predict the new crop type using historical crop rotations on parcels of land called crop sequence boundaries (CSBs). The CSB-level predictions are aggregated to the tract level to provide the imputation. The model-based approach is shown to outperform the manual one for a variety of crop types.

Key Words: Machine Learning, Agriculture, Automatic Imputation, Forecasting, Area Survey

1. Introduction

The United States Department of Agriculture's (USDA's) National Agricultural Statistics Service (NASS) conducts hundreds of surveys and publishes reports on a wide variety of topics relating to agriculture in the United States (U.S.). These reports are available to the public and are used by farmers and ranchers, universities, federal and state agencies, and internal USDA stakeholders. One such report is the June Acreage report, which as its name implies, is published at the end of June. This annual report provides acreage, harvesting, yield, and weather information for the planting season leading up to the publication date.

Statistics provided in the June Acreage report are the result of an expert review process that is conducted under the direction of the NASS Agricultural Statistics Board (ASB). The process includes two surveys, the June Crops Acreage, Production, and Stocks Survey and the June Area Survey (JAS). The JAS is an area-based survey, where plots of land called tracts are sampled according to a survey design. Once selected, data on agricultural activity within the tract is collected from the farmer by trained interviewers. These data, along with weights from the survey design, are used to provide state-level estimates for the agricultural quantities of interest. Part of the tract level data collected

are acreages planted to various crops. These acreages are part of the data series that inform the published planted acreages in the June Acreage report.

Nonresponse is a fact of life in many surveys and the JAS is no exception. Currently, NASS field office statisticians manually impute the tract-level data for non-responding farmers. This manual imputation process is time consuming and expensive, so an alternative, automated approach is desirable. Since the JAS is an area survey, the digitized tracts can be overlayed with geospatial data useful for imputation.
In this paper, the focus is on reliably imputing missing crop-acreage data in selected tracts using field level predicted crop acreages. Many models for field level crop type prediction exist. A recent approach uses machine learning to predict crop types planted in fields called crop sequence boundaries (Abernethy et al., 2023). Crop sequence boundaries (CSBs) are areas of land selected to contain homogenous crop rotation patterns (Hunt et al., 2024). The historical crop rotation in the boundary can be combined with a LightGBM (Ke et al., 2017) model to predict the crop planted within the boundary during the current planting season.

The proposed imputation process is described as follows. First, the LightGBM model using historical crop rotations is used to predict the current crop type that will be planted within the CSB. Next, the JAS tracts and CSBs can be overlayed. After this, the imputed acreage value for a given crop is the sum of the areas of CSBs within the tract predicted to be planted to the crop. Finally, tracts are selected for imputation based on reliability of the model forecasts using the average entropy of the model predictions within the tract.

1.1 Objective and Outline

The objective of this paper is to combine a modern machine learning approach with geospatially referenced field boundaries to forecast crop planted areas that can be used to impute missing information caused by nonresponse within the JAS tracts. This automatic approach can then be compared with the traditional manual approach using administrative ground truth data augmented with end-of-season gridded landcover-type data.

In Section 2, the JAS is described, with a focus on the tracts, crop acreage data collected, and current imputation processes. In Section 3, a machine learning based approach using field analogues called CSBs is proposed for automatic imputation of tracts with missing data. The experimental setup is described and the automatic imputations are compared to the manual ones with respect to the ground truth in Section 4. Finally, Section 5 concludes.

2. June Area Survey

2.1 Design

The JAS is conducted annually during the month of June. Since it is an area survey, the units sampled are parcels of land. To collect the sample, the continental U.S. is divided into strata based on the level of agricultural activity. Within each stratum, the land is again divided into substrata, based on similar types of agricultural activity. Within each substratum, the land is divided into primary sampling units. Stratified random sampling is used to select a set of primary sampling units.

Once a primary sampling unit is selected, it is divided into parcels of land called segments. Each segment is typically (but not always) a one square mile plot of land. Next, one segment is randomly sampled from each of the previously selected primary sampling units. The sampled segments are divided into tracts which represent unique land operating arrangements. The survey data are collected from all farmers who own or operate a tract within a selected segment. Note that it is possible for segments and tracts to not be on agricultural land, in which case no data is collected.

A JAS segment with tracts representing unique land operating arrangements is provided in Figure 1 below. The segment boundary is show in red, while the tract boundaries are shown in blue. The unique land operating arrangements are labeled A-H. Note that tracts can contain agricultural and non-agricultural land (in this case trees). This segment was randomly sampled according to the survey design and is located in Lancaster County, Pennsylvania.

Figure 1: A JAS segment in Lancaster County, PA with boundaries in red. The segment is divided into tracts (boundaries in blue) that represent unique land operating arrangements (A-H).

2.2 Data Collected

Once an agricultural JAS tract is selected, the owner is interviewed about the agricultural activity conducted within the tract. Apart from contact information, data collected include ownership and operating structure, information on workers, acreage of crops planted (or to be planted), acreage of crops harvested (or to be harvested), livestock information, crop storage, government program participation, economic data such as sales and land value, technology use (e.g. internet use, tablet use, etc.), and demographic information. The focus of this paper is acreage of crops planted (or to be planted), where the goal is to impute missing planted acreage values when a farmer non-response occurs.

The JAS collects acreage for a variety of crops, including corn, sorghum, barley, winter wheat, spring wheat, durum wheat, rye, rice, millet, hay, soybeans, peanuts, sunflowers, canola, flaxseed, safflower, cotton, sugar beets, sugarcane, tobacco, dry beans, chickpeas, lentils, peas, and potatoes. Along with the survey weights, this crop acreage data is used to provide planted acreage at the state and national level. Note that every crop does not appear in every state (for example Louisiana has a published value for rice acreage while Illinois does not).

2.3 Manual Imputation

Like many surveys, the JAS is subject to non-response. This can occur when the farmer refuses to participate in the survey or is otherwise inaccessible. Partial response is also possible. For example, a farmer completing an online survey may report corn but fail to report presence or absence of other crops. Regardless, when farmers do not respond to a survey item, the item must be imputed.

Currently, missing items are imputed manually by NASS staff. Interviews with respondents can occur either in person or remotely (by telephone or online). The method of interviewing determines the method of manual imputation. If the missed interview is in person, the interviewer will attempt to fill out the survey form based on visual assessment. For example, if the interviewer sees corn growing in a ten-acre field within the tract then ten acres of corn can be added to the imputation. If the missed response does not involve an in-person interview, then imputation is performed by office staff using historical data, administrative data, or conventional imputation techniques using survey respondent data.

3. Automatic Imputation Approach

3.1 Crop Sequence Boundaries

The input data for the proposed automatic imputation approach are the NASS Crop Sequence Boundaries (CSBs) (Abernethy et al., 2023; K. Hunt, 2024; K. A. Hunt et al., 2024). These algorithmically delineated field polygons represent fields with consistent cropping history over a fixed period of time (Abernethy et al., 2023; K. A. Hunt et al., 2024). Coverage of the CSBs includes the entire continental U.S. where crop land is present. They are available to the public in eight-year windows, with the oldest being 2008-2015 and the newest as of this writing being 2016-2023. More information on how the CSBs are created can be found in Abernethy et al., 2023; K. A. Hunt et al., 2024.

Figure 2 shows a one-year snapshot of a set of eight-year CSB polygons in Bond County, Illinois. Note the field boundaries (black) which in this case contain corn (yellow), soybeans (green), winter wheat (brown), and alfalfa (pink). Also note how the non-agricultural land (mostly roads and trees) does not have CSBs. Finally, notice how some homogenous cropping areas contain more than one CSB. This is a result of looking at a one-year snapshot of eight-year homogenous crop type fields. Multiple CSBs in these areas indicate that the crop type in the area was not homogenous in at least one year of the eight-year time series of cropping history.

Figure 2: Corn (yellow), soybean (green), wheat (brown), and alfalfa (pink) CSBs in Bond County, Illinois.

A three-year snapshot of a hypothetical set of CSB polygons is provided in Figure 3 below. The crop types in this simple example are corn (yellow) and soybeans (green). One can see that in the current year (year three), the L-shaped region has one common crop type but is composed of four CSBs. This is because the rotation history over years one, two, and three is only homogenous within the red boundaries that define the CSB polygons. CSBs are designed not to have multiple crop types planted within them during any given year. This one crop per year rule makes the CSBs useful for crop type forecasting applications as demonstrated in Section 3.2.

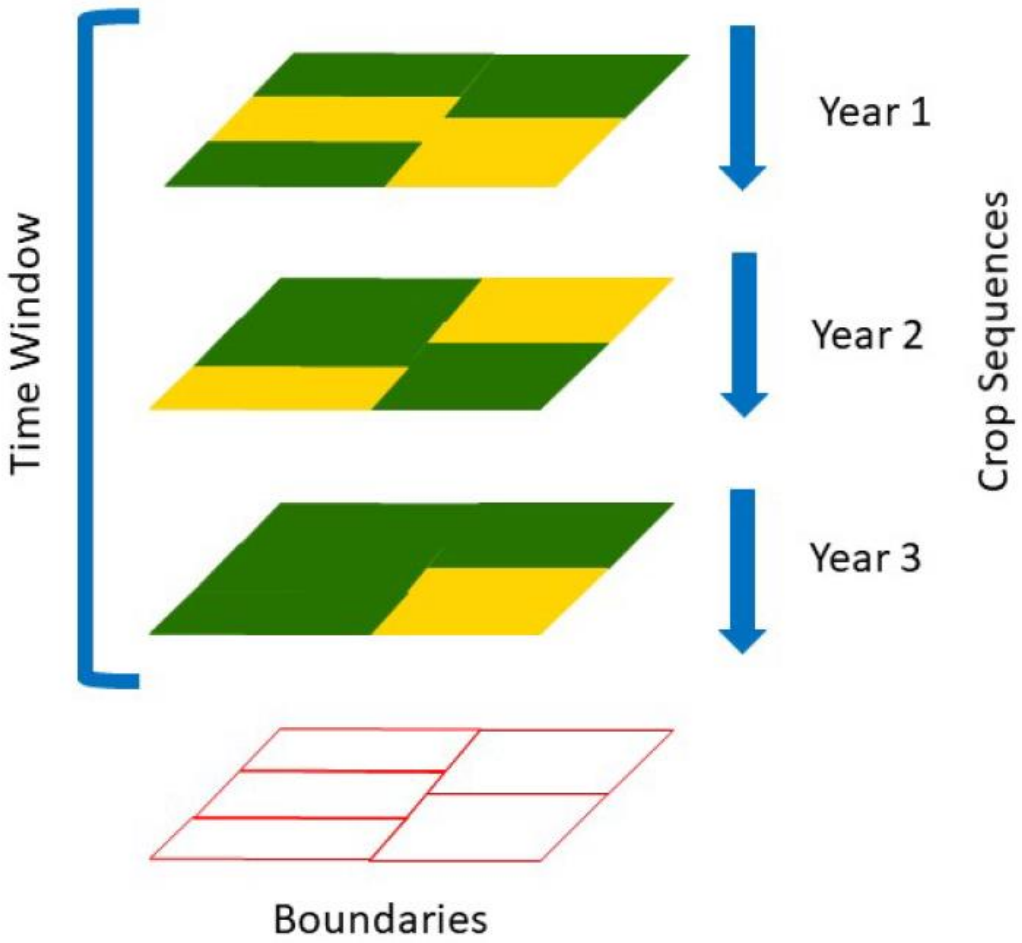


Figure 3: Hypothetical three-year CSBs (red) defined by areas with homogenous corn (yellow) and soybean (green) crop rotation history.

3.2 CSB Crop Type Forecasting Using Machine Learning

Given a set of CSBs with a fixed timespan window $\boldsymbol{w}$, the goal is to use the historical crop rotation to predict the crop type inside each CSB during the current year, $\boldsymbol{w+1}$. First assume a set of CSB polygons, enumerated $\boldsymbol{i = 1, 2, \ldots, n}$, a set of crop types to predict, enumerated $\boldsymbol{j = 1, 2, \ldots m}$, and a year within the crop rotation history $\boldsymbol{t = 1, 2, \ldots, w}$. Historical crop rotation data $\boldsymbol{X_{it}}$ is available, where $\boldsymbol{X_{it}}$ is an indicator vector of length $\boldsymbol{m}$, with entries $\boldsymbol{x_{ijt} = 1}$ indicating the presence of crop $\boldsymbol{j}$ in CSB $\boldsymbol{i}$ at time $\boldsymbol{t}$. The vector $\boldsymbol{X_{it}}$ can only have one non-zero entry. The indicator structure of $\boldsymbol{X_{it}}$ is possible because the CSBs are designed to contain only one crop type each year (see Section 3.1). For any alternative polygon, the crop types inside would have to be treated as an $\boldsymbol{m}$ vector of proportions of each crop type within the field.

Using the historical data above, the goal is to predict the crop types in each CSB during the current year, $\boldsymbol{X_{i(w+1)}}$. To do this in a simple way, an added assumption is needed that the CSB contains only one crop during the current year. While historical homogeneity over the window $\boldsymbol{w}$ does not guarantee a one crop CSB during the current year, it can be

encouraged by choosing a large enough $\boldsymbol{w}$. A large $\boldsymbol{w}$ suggests stability, as a farmer planting only one crop type in a CSB each year over a long history suggests this behavior is likely to continue. Furthermore, the competitive performance of CSB-based models with an eight-year window ($\boldsymbol{w = 8}$) in relation to pixel-based alternatives that are less sensitive to this assumption in Abernethy et al., 2023 suggests that CSBs suddenly becoming multi-crop are rare and have minimal impact on predictive accuracy. Under this assumption the response variable follows a multinomial distribution given the rotation history, which allows the use of a variety of off-the-shelf statistical or machine learning models.

The supervised machine learning setup for this application is nearly identical to that used in Abernethy et al., 2023. The only difference is that different crop types are predicted to accommodate imputing crop types important to the JAS (see Section 4.2 for the crop types examined). To review the setup, the input data are eight-year window CSBs. The predictor variables use six years of rotation history, the area of the CSB in acres, and the NASS agricultural statistics district in which the CSB is located. Note that an agricultural statistics district is a NASS-defined set of counties within a state with similar agricultural practices. For the machine learning algorithm, a gradient boosting approach called LightGBM (Ke et al., 2017) via the lightgbm R package is used to train multiclass models to predict the crop types. Tuning parameters for LightGBM are validated using time series cross validation over a moving window. In particular, candidate models are trained using years one through seven. The best model is selected using years two through eight. The final model with best tuning parameters is trained using years two through eight. Finally, years three through eight are fed into the model to predict the crop type in the unknown year nine.

3.3 Imputation

Given the predictions described in Section 3.2, imputation can be achieved by a simple overlaying process. Since the JAS is area-based, it is possible to geo-reference the tracts. Geo-referenced tracts have been available internally to NASS since 2021. These tracts can be rasterized to a given extent (e.g. state) for comparison with other geospatial data products. The rasterized dataset is a grid of pixels within the state where each pixel is either labeled with a tract id or a value indicating that no tract is present.

To complete the imputation, the CSBs were rasterized in a similar manner. The predicted crop type can be joined to the rasterized CSB by id resulting in a grid of rasterized crop type predictions. These predictions are overlayed with the rasterized tracts. The imputed values are the sum of the crop probabilities within the tract. All datasets were rasterized using 900 square meter pixels, as this presented the best compromise between accurately representing the JAS tracts and CSB polygons while also being computationally tractable.

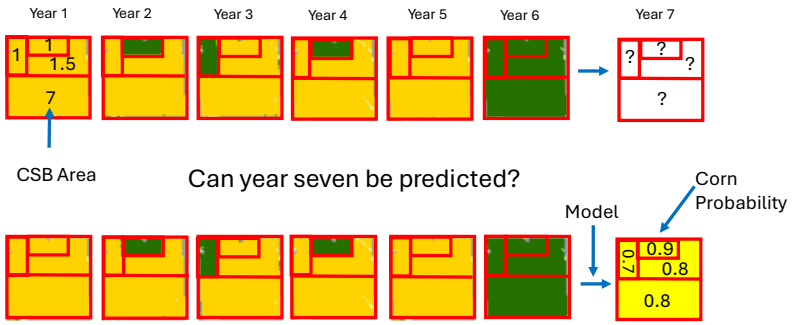
The modeled crop type probabilities can also be used to provide a measure of uncertainty for the imputation of each tract. In particular, the average entropy $\boldsymbol{H}$ in the tract is used here and defined below.

$$\boldsymbol{H} = -\frac{\sum_{\boldsymbol{CSB} \in \boldsymbol{Tract}} \sum_{\boldsymbol{i} \in \boldsymbol{crop\ types}} \boldsymbol{Area_{CSB}} * \boldsymbol{p_{CSB}}(\boldsymbol{type} = \boldsymbol{i}) \ln\left(\boldsymbol{p_{CSB}}(\boldsymbol{type} = \boldsymbol{i})\right)}{\sum_{\boldsymbol{CSB} \in \boldsymbol{Tract}} \boldsymbol{Area_{CSB}}}$$

In this case the sums run over all CSBs in the tract and all crop types that could be planted within the CSB. The CSB area, $Area_{CSB}$, is measured in acres. The crop type probabilities $p_{CSB}(type = i)$ are derived from the LightGBM model described in Section 3.2. These entropies can be used to only impute tracts where model certainty is high (entropy close to zero up to some practical threshold, see Section 4.4). Areas where the model is uncertain (entropy is large) can be passed to the manual imputation process. This can assure quality automatic imputations while still reducing staff workload.

Entropies in areas where CSBs are not present are assumed to be zero. The CSBs are designed to cover all cropland, so a lack of a CSB means the area likely has no crops, which would suggest the zero-entropy assumption is reasonable. However, the CSBs are not perfectly accurate, so there will be some error in this assumption. Obtaining more accurate entropies for areas with no CSB is a subject of future work.

An example summary of the imputation process is illustrated in Figure 4, which shows a hypothetical tract with four CSBs. Six years of history are provided for each CSB, each following a corn-soybean cropping pattern. The goal is to use this history to predict the crops planted in the tract during year seven. In this case, the history is used to train a machine learning model that predicts the corn probability (and by extension the soybean probability) for each CSB for year seven. These probabilities can be used for tract level imputations and uncertainties (see Figure 4).



$$\text{Imputed Corn Area} = 0.7 * 1 + 0.9 * 1 + 0.8 * 1.5 + 0.8 * 7$$

$$H = \frac{0.7*\ln(0.7)+0.3*\ln(0.3)+0.9*\ln(0.9)+0.1*\ln(0.1)+\{0.8*\ln(0.8)+0.2*\ln(0.2)\}*8.5}{-10.5}$$

Figure 4: Six-year crop rotation history for a hypothetical tract (outer red boundary) with four CSBs (inner red boundary) each following a corn (yellow) soybean (green) system. Prediction of year seven is accomplished with LightGBM model. Modeled corn probabilities are multiplied by CSB areas and summed to get imputed tract corn area. Entropy H is also calculated to get tract level uncertainty.

4. Experiment Set Up and Results

4.1 Ground Truthing

Ground truthing is used to compare the accuracy of the automatic model-based crop type predictions with historical manual imputations. The ground truth used is USDA's Farm Service Agency (FSA) Form 578 administrative data. FSA data are a georeferenced set of U.S. crop fields supporting commodity and conservation programs. These data are obtained from farmers who participate in FSA's crop insurance program. The FSA 578 data are updated every growing season and historical data are available as far back as 2008. Data for each farmer who participates in the program is usually available later in the season (July and August). The FSA 578 data are administratively confidential and not available for public dissemination (Heald, 2002; USDA-FSA, 2017).

The subset of FSA data used in Boryan et al., 2011 to train and validate the NASS Cropland Data Layer is used in this study. This subset contains all FSA fields with only one crop within the field. The FSA acres within a tract for each crop type can be obtained by rasterizing the FSA ground truth and adding the pixel acreage of each crop type within the tract. Since these data only include USDA program participants and exclude multi-crop fields, the FSA coverage of all crop acres is not complete. This means that June tracts will typically not have full ground truth coverage. The incompleteness requires a modified error function as described in Section 4.3.

In addition to the FSA data, the National Land Cover Dataset (NLCD) is also used to represent non-agricultural data (Homer et al., 2012). In particular, areas with no FSA coverage that contain NLCD nonagricultural land are included in the ground truthing as such. The 2019 NLCD raster is used for this study.

4.2 Study Area and Crop Types

The model predictions and manual imputations are compared using all high ground truth coverage agricultural tracts in the continental U.S., with the exception of North Dakota, which was excluded due to clerical issues. High ground truth coverage means that the tract has at least 90% ground truth coverage. The years for which comparisons are made are 2021, 2022, and 2023.

The CSBs, JAS tracts, and FSA ground truth data all provide slightly different crop-type information. Many, but not all, of the crop types overlap between the three sources. Furthermore, many crop types are rare at the tract level, so sample sizes are not large enough for meaningful comparison. For these reasons, crop types that appear on all three sources and appeared in at least 50 tracts on average over the three-year period were chosen for this study. The selected crop types were corn, cotton, peanuts, rice, sorghum, soybeans, spring wheat, and winter wheat.

4.3 Error Function

Ideally, error between each approximation (predictive model and manual imputation) could be described using the absolute difference, shown below.

$$E_{jk}^{*} = \left| A_{sjk} - A_{Fjk} \right|$$

In this case the error E_{jk}^{*} for crop j in tract k is the absolute difference between the acreage A_{sjk} predicted by approximation s for crop j in tract k and the ground truth FSA

acreage A_{Fjk}. This error function assumes complete coverage of the FSA ground truth within a tract.

As mentioned in Section 4.1, the FSA ground truth data do not have complete coverage in general. This means that the ground truth can only provide lower and upper bounds. For each crop, the lower bound is the FSA ground truth acreage of the crop and the upper bound is the FSA ground truth acreage of the crop plus any additional acreage in the tract containing no FSA data. This suggests the use of the threshold error function below.

$$E_{jk} = \max\left(0, A_{Fjk} - A_{sjk}, A_{sjk} - A_{Fjk} - M_k\right)$$

The extra term M_k is the acreage of the tract where there is no ground truth. The function E_{jk} will produce a non-zero error whenever an approximation produces a crop acreage less than the ground truth lower bound or greater than the ground truth lower bound plus the no data acreage. There is a region of uncertainty when an approximated crop acreage is between the lower bound and the sum of the lower bound and acreage with no ground truth. In this region of uncertainty, the error is set to zero. The reason for only selecting tracts with 90% or more ground truth coverage in Section 4.2 is to minimize the effect of the ground truth under coverage and resulting zero inflation in the errors.

Once the errors E_{jk} are obtained for crop j and tract k, the ratio of average errors R_j^τ is reported for each crop. The error ratio R_j^τ is defined below:

$$R_j^\tau = \frac{\sum_{k \in tracts(\tau)} E_{jk}^{model}}{\sum_{k \in tracts(\tau)} E_{jk}^{survey}}$$

In this case $tracts(\tau)$ refers to the set of all tracts with at least 90% ground truth coverage and entropy less than or equal to τ. The quantity E_{jk}^{model} is the model error for crop j in tract k, and E_{jk}^{survey} is the survey error for crop j in tract k. It is desirable to choose an entropy threshold τ to keep the error ratio less than or equal to one, as this means that the average error of the automatic imputation is no worse than that of the manual one for tracts with entropy below τ.

4.4 Results

Counts for each crop type by year are provided in Table 1 below. The count refers to the number of tracts that contain nonzero FSA ground truth for each crop type. Note that all crop types in Table 1 have at least 40 instances each year, averaging at least 50 instances over three years. The most common crops, corn and soybeans, are planted in over a thousand tracts each, while the least common, peanuts, are only planted in slightly more than 50 tracts on average.

Table 1: Count of Tracts Containing Specified Crop

Crop	Count 2021	Count 2022	Count 2023
Corn	1824	1629	1478
Cotton	256	266	258
Peanuts	45	53	57
Rice	100	49	122
Sorghum	130	124	124
Soybeans	1804	1687	1433
Spring Wheat	87	92	96
Winter Wheat	349	376	336
Total Tracts	4155	4076	3866

The results of the comparison for all three years combined are provided in Figure 5 below. The x axis in Figure 5 is a threshold on the entropy, as described in Section 4.3. In particular, all tracts with entropy less than or equal to the threshold are retained for imputation by the model. The model-based entropies range from roughly zero to two. The y-axis is the ratio of average errors described in Section 4.3. It ranges from about zero to about 1.3. Each colored dot represents the ratio of average errors between the model and manual imputations for each crop at the specified entropy threshold. The vertical black line specifies an entropy cutoff of about 0.6. The horizontal black line represents an error ratio of one.

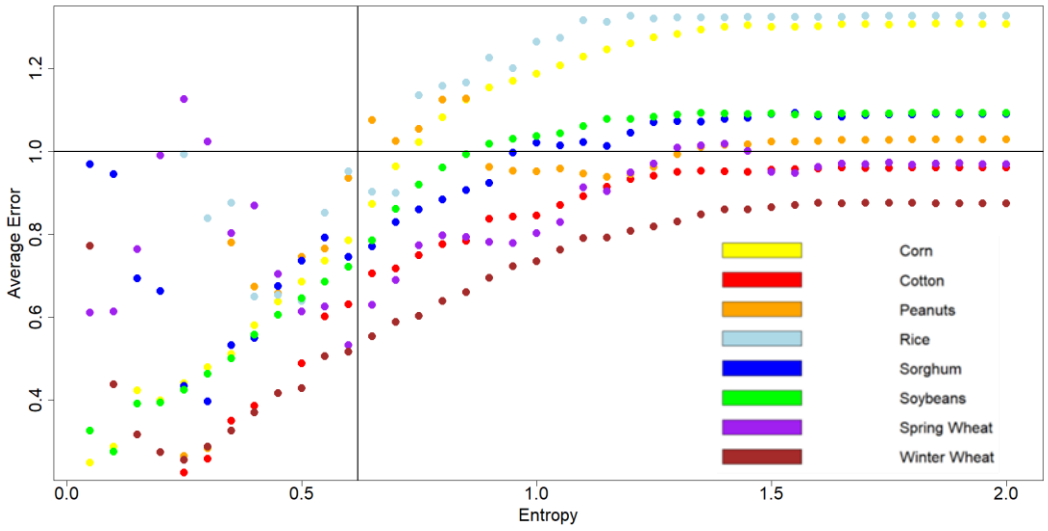


Figure 5: Ratio of average errors by entropy threshold using combined data from all years. Using an entropy threshold of about 0.6 keeps the average errors below one for all crop types.

Figure 5 demonstrates that for most crops it is not possible to guarantee an error ratio below one while retaining all tracts for imputation. However, it is possible to achieve an error ratio below one by selecting tracts with a model-provided entropy below about 0.6. Most crops follow the expected pattern where the error ratio increases as the entropy increases. Deviations from this pattern near zero occur for some crops with smaller tract count in Table 1. This may be because occurrence in fewer tracts would also suggest

fewer CSBs in the model to train for these rare classes. Thus, the errors on these crops even when the model is confident may result from class imbalance during model training.

Figure 6 depicts the same information as Figure 5 but broken down by year. The results are similar to the pooled data for all years, although there are a few cases where the threshold method is not as effective. These cases are cotton and sorghum in 2022 and peanuts in 2023. The average error ratio for rice is also slightly above (but very close to) one in 2021 and 2023. Regardless, using the 0.6 entropy threshold keeps the average error below one every year for the three most prevalent crop types (corn, soybeans, and winter wheat) and nearly so for the fourth most prevalent, cotton.

Temporal variability between years does exist in Figure 6 and may result from a variety of factors. One likely factor is the prevalence of in-person imputations versus office imputations each year. Furthermore, year-to-year variability on the model side may also exist. In particular, economic or weather incentives can cause farmers to alter their planting decision from what their past cropping history may suggest. These in-season factors will likely affect the quality of model predictions trained only on historical cropping patterns.

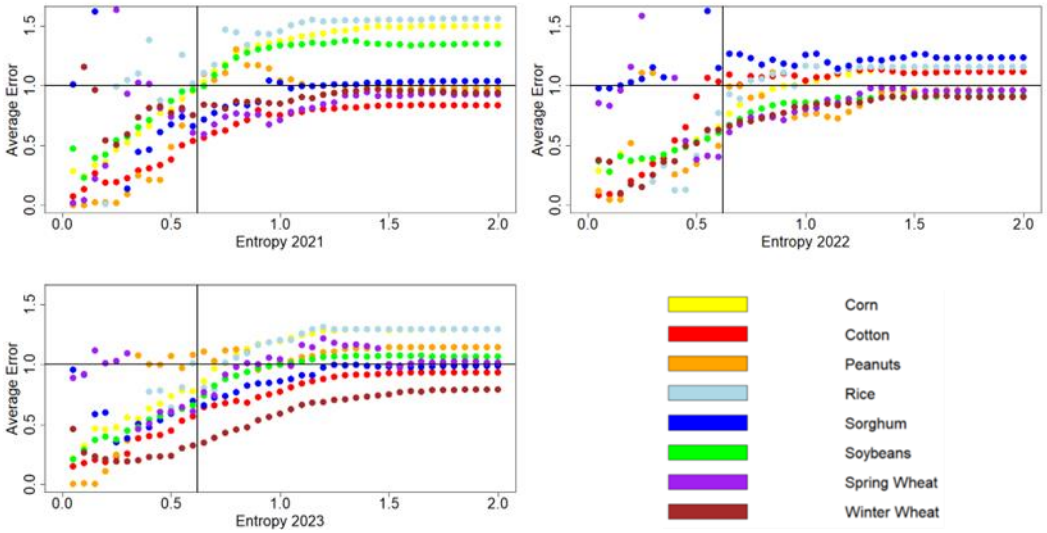


Figure 6: Ratio of average errors by entropy threshold broken out by year. Using an entropy threshold of about 0.6 keeps the average errors below or very close to one for most crop types.

While thresholding can improve the average model imputation error with respect to the manual imputation, it does require that tracts with high entropy be skipped over. This leads to the question of how many tracts are lost for potential imputation. If only a small percentage of available tracts can be automatically imputed, then the procedure may not be very useful in practice. Table 2 lists the percentage of tracts that meet the maximal entropy criteria of 0.6 by year.

Table 2: Percent of Tracts Containing Specified Crop Retained After Thresholding

Crop	Percent 2021	Percent 2022	Percent 2023
Corn	55	58	56
Cotton	43	32	47
Peanuts	38	19	39
Rice	28	20	25
Sorghum	19	23	27
Soybeans	53	60	59
Spring Wheat	17	17	10
Winter Wheat	35	36	38
Total Tracts	56	58	58

From Table 2 it can be seen that over half of the total tracts are kept, even when thresholding. This could suggest a substantial reduction in manual work overall. As for crop type, the more common crop types benefit most. Corn and soybeans consistently see over 50% of tracts containing these crops acceptable for automatic imputation. Winter wheat and cotton also see substantial potential for automatic imputation, with at least 30% in-scope tracts each year. The worst results occur with the rarest crops, which is also indicative of class imbalance, as the model may be more accurate for the majority classes and therefore low entropy tracts may occur most often in tracts with minority class counts of zero.

5. Conclusion and Future Work

A machine learning model using fields with homogeneous cropping patterns called crop sequence boundaries was proposed to assist in imputation of the NASS JAS. This automatic approach could potentially help relieve the staff time and cost related to the current use of manual imputation. While the model-based imputation error is not lower than the manual error when using all tracts, it can be reduced by only imputing tracts with sufficiently low model-based entropy. Choosing tracts to impute based on an entropy threshold keeps the average model errors lower or near equal to the manual ones for most crops and years. This automatic approach could potentially lead to an over fifty percent reduction in tracts that need to be manually imputed.

Several avenues of future work exist. The first is incorporating available ground truth information into the automatic imputation process. FSA-based crop-type information is typically not complete until August; however, some data are available by June, particularly winter wheat data. Other near ground truth data may exist as well. For example, historical JAS data suggesting non-crop land in a tract is unlikely to change year over year. For example, farmsteads, barns, silos, equipment storage areas, etc. are unlikely to be converted into cropland. These ground truth sources can be combined with the model predictions to potentially yield more accurate imputations.

The second is to break out model performance by manual imputation type and item-level response. For example, failed in-person interviews where the interviewer records observed crops may be more accurate than office-based manual imputation. Furthermore, partially imputed data that contain some true farmer provided responses will likely be more accurate than data for which all crop data were imputed. Currently it is not possible to separate these cases using available data. The model may perform better when used only in cases that would have been completed by remote office-based imputations, should it become possible to identify these cases.

Another potential for future work is the addition of more useful variables to the predictive model. Most relevant would include economic data like crop prices and futures, nearby ethanol production, changes in government programs, etc. Weather data, such as precipitation, soil moisture levels, and temperature, also impact planting decisions. The current challenge in including these data is that historical, field-level crop-type data are typically only available starting in 2008 for the continental U.S. Using six-year crop rotations means the first year available for training a model is 2014. This means that there is not a large enough temporal sample size to adequately capture the influence of these variables, which vary greatly in time. One solution option would be research in producing CSBs prior to 2008.

Acknowledgements

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. We would also like to acknowledge Valbona Bejleri, Yang Cheng, and Linda Young for their effort in reviewing this paper for quality prior to submission.

References

Abernethy, J., Beeson, P., Boryan, C., Hunt, K., & Sartore, L. (2023). Preseason crop type prediction using crop sequence boundaries. *Computers and Electronics in Agriculture*, *208*, 107768. https://doi.org/10.1016/J.COMPAG.2023.107768

Boryan, C., Yang, Z., Mueller, R., & Craig, M. (2011). Monitoring US agriculture: the US Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer Program. *Http://Dx.Doi.Org/10.1080/10106049.2011.562309*, *26*(5), 341–358. https://doi.org/10.1080/10106049.2011.562309

Heald, J. (2002). *USDA Establishes a Common Land Unit*. https://www.esri.com/news/arcuser/0402/usda.html

Homer, C. G., Fry, J. A., & Barnes, C. A. (2012). The National Land Cover Database. *U.S. Geological Survey*. https://doi.org/10.3133/fs20123020

Hunt, K. (2024). *USDA - National Agricultural Statistics Service - Crop Sequence Boundaries (CSB)*. https://www.nass.usda.gov/Research_and_Science/Crop-Sequence-Boundaries/index.php

Hunt, K. A., Abernethy, J., Beeson, P. C., Bowman, M., Wallander, S., & Williams, R. (2024). Crop sequence boundaries using USDA National Agricultural Statistics Service historic cropland data layers 1. *Statistical Journal of the IAOS*, *40*(2), 237–246. https://doi.org/10.3233/SJI-230078

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, *30*. https://github.com/Microsoft/LightGBM.

USDA-FSA. (2017). *Common Land Unit Information Sheet*. https://www.fsa.usda.gov/Assets/USDA-FSA-Public/usdafiles/APFO/support-documents/pdfs/clu_infosheet_2017_Final.pdf