Journal of Data Science 24 (3), 584–599
July 2026

DOI: 10.6339/26-JDS1235
Data Science in Action

Use of Farm Equipment Machine-Logged Data to Inform Crop Production Statistics☆

Sean Rhodes[1,*], David M. Johnson[1], Luca Sartore[2,1], Samuel C. Garber[1], Darcy Miller[1], and Denise A. Abreu[1]

[1]*United States Department of Agriculture, National Agriculture Statistics Service, Washington DC, 20250, USA*
[2]*National Institute of Statistical Sciences, Washington DC, 20033, USA*

Abstract

The United States Department of Agriculture's (USDA's) National Agricultural Statistics Service (NASS) conducted a pilot study in 2024 to obtain data collected onboard farm machinery and explore their uses for statistical purposes. NASS has recognized high value in these machine-logged data (MLD) systems as they can potentially augment, or even replace, traditional survey efforts while providing additional benefits of reducing respondent burden and improving crop-related estimates. This pilot study ultimately addressed four topics: 1) understanding the obstacles in obtaining MLD from farmers; 2) creating geographic workflows to manage inherent geospatial MLD; 3) developing the linkages to NASS's tabular list frame information; and 4) assessing the use of MLD to replace survey data for time-sensitive estimates. To study each topic, field-level information was gathered from the MLD systems of dozens of producers over hundreds of fields across the central United States (US) for the 2023 growing season. Results showed that 90% of the fields could be linked to a producer on the NASS list frame. Of those producers, the consistency of MLD versus traditional survey reporting was highly variable for those who were selected for a survey in 2023. Comparisons showed median MLD values were larger than historical NASS survey values. Approximately 48% of survey comparisons showed a difference of 25% or less between MLD and historical NASS survey values. MLD shows promise for use in official statistics; however, further analyses with additional producers' data and enhancements to MLD collection processes are needed before supplementing traditional survey methods.

Keywords *data linkage; geospatial; imputation; MyAgData®; precision agriculture; unstructured data*

1 Introduction and Background

For most of the 20th century, sample surveys and censuses have been the main data collection methods to produce official statistics. Beginning in the late 1900s, response rates have been trending downwards (Groves et al., 2004), while the cost of conducting surveys has been steadily increasing. Declining response rates and increasing costs of data collection pose challenges to

☆The findings and conclusions in this presentation are those of the authors and should not be construed to represent any official USDA or US Government determination or policy.

*Corresponding author. Email: sean.rhodes@usda.gov.

© 2026 The Author(s). Published by the School of Statistics and the Center for Applied Statistics, Renmin University of China. Open access article under the CC BY license.
Received September 29, 2025; Accepted May 29, 2026

many agencies producing official statistics. Various data collection strategies can be used to mitigate issues related to decreasing response rates. Previous studies have shown that in-person or multi-mode follow-up enumeration improve response rates (Heberlein and Baumgratner, 1978; Leeuw et al., 2007; Dillman et al., 2014), and this phenomenon is also true in agricultural populations (Tran et al., 2012; Thompson A, 2025). However, face-to-face data collection is the most expensive mode to implement, and effective operating budgets are declining for many survey organizations. Hence, survey statisticians have turned to alternative resources for data to help supplement survey data. For example, incorporating existing administrative data, such as data from official registries, into survey processes has been shown to be successful (Longva et al., 1998). Passive and system-to-system data collection strategies have also been explored to mitigate declining response rates while also reducing respondent burden (Hutchinson et al., 2023). Examples include point-of-sale scanner data, web-scraped data, and remotely sensed data (Bradley, 1997; Hutchinson and Scheleur, 2017; Dumbacher and Capps, 2016; Brimble et al., 2020; Murphy et al., 2022; Rhodes et al., 2023).

Fortunately, the growth of digital technology is providing alternate sources of data, which can help mitigate the decline in traditional survey response rates. In the agricultural sector for example, producers' planters, harvesters, and applicators are now commonly equipped with monitors and sensors that records data logs from many farming practices. Much of the equipment is also linked to Global Navigation Satellite Systems (GNSS), which provide locational information accurate within a few meters. Temporally and spatially detailed data on many farming variables, such as crop seeding, chemical applications, and yield rates are collected using this machinery. McFadden et al. (2023) studied producers from 1996 to 2019 on the adoption of agricultural technologies within their operation. Findings showed that over 50% of acres operated by the producers used GNSS auto-navigation technology. GNSS-tied soil maps, yield monitors, and other tools have seen an increase in adoption over time as well. Due to the financial cost of digital agriculture, larger operations have a higher adoption rate than smaller operations. Finally, most users of digital agriculture have preserved a "time-and-effort" reduction resulting in labor-savings.

Machine-logged data (MLD) from farming equipment have several potential uses in official agricultural statistics. Producers could upload their MLD in near-real time during data collection to complete their survey or Census questionnaires. Reducing the time needed to complete the questionnaire would decrease the response burden and cut enumeration costs of statistical offices. Accuracy of the producers' responses could rise due to the planted, harvested, and yield data coming directly from the piece of machinery. Moreover, the georeferenced data may be used as auxiliary data to represent covariates in imputation, modeling, or calibration processes. Snijkers et al. (2021) discussed several approaches that use MLD as input for official statistics.

Despite the many benefits, various challenges have been associated with the implementation and access of MLD from producers. Studies at National Statistical Offices and other organizations have been conducted on the use of MLD to explore crop measurements and analyses (Deines et al., 2021). Statistics Netherlands conducted a pilot study with approximately five producers to explore the use of a system-to-system approach to automate data collection for an agricultural business survey (Snijkers et al., 2024). An Application Programming Interface (API) from John Deere was utilized to have producers share their data through the system to fill out the questionnaire. Some respondents indicated that the machine data may not be correct. Furthermore, part of the data expected to be collected was missing from the tool used to collect the crop data. The small-scale test of five producers was designed to assess the usability of MLD during data collection, boost respondent trust, and pinpoint technical issues.

In this article, the United States Department of Agriculture's (USDA's) National Agricultural Statistics Service (NASS) conducted its own pilot project to obtain precision agriculture data and explore their use for statistical purposes. This study was built upon the use of farm MLD to help inform, supplement, and possibly reduce the number of surveys completed by US producers. The pilot study explored the use of MLD to impute producers' survey responses. Furthermore, it focused on overcoming the obstacles of obtaining, managing, and developing workflows with the MLD. Finally, the study assessed the quality of linking MLD to NASS's list frame and compared MLD to the producer-reported survey data for the same reference period on a much larger scale.

The remaining part of this article is organized as follows. Section 2 describes the obstacles associated with obtaining producers' MLD for the pilot study. Section 3 contains the development of workflows to manage geospatial data and their linkages to the NASS list frame information. Section 4 summarizes the results from matching and assessing how MLD compares to historical NASS survey responses. Section 5 outlines successes and challenges from the pilot study. Finally, an overview of the study and concluding remarks are found in Section 6.

2 Data

In 2024, NASS worked with MyAgData®, a leading company in precision agriculture, to obtain MLD from producers with whom MyAgData® had a working relationship. MyAgData's® cloud-based application has automated the data collection and reporting process required by the USDA to help satisfy producers' obligations for crop insurance and Farm Service Agency (FSA) program participation. MyAgData's® application has collected, translated, organized, and delivered field-level data for crop insurance, farm programs, private insurance products, carbon capture, crop traceability, and food safety.

2.1 Planter and Harvester MLD

MyAgData® provided MLD summarizing crop planter and harvester data at the field-level. The provided data were tied to GNSS based instruments aboard the farm equipment to give detailed locational information. Variables collected within the planter's MLD system included information such as seed type, row spacing, and ground speed. Most relevant to the pilot study was crop type information and knowing the boundaries of the area planted. Likewise, the harvester MLD included detailed information such as grain volume, moisture, and flow rate. Crop type, average yield, and geographic area harvested were the most important outputs to this study.

Data from 60 large-scale producers were provided by MyAgData® as geoJSON files (an open standard for geographical data with associated attribute information). There were 2,651 unique planted fields and 1,430 unique harvested fields across the following nine states: Arkansas, Georgia, Illinois, Indiana, Michigan, Nebraska, North Carolina, North Dakota, and Wisconsin (see Figure 1). The data provided were not representative samples of the population, but rather the best effort to obtain GPS-based information from as many producers with whom MyAgData® had a working relationship. Producers' name and/or operation information were not provided in the MyAgData® files.

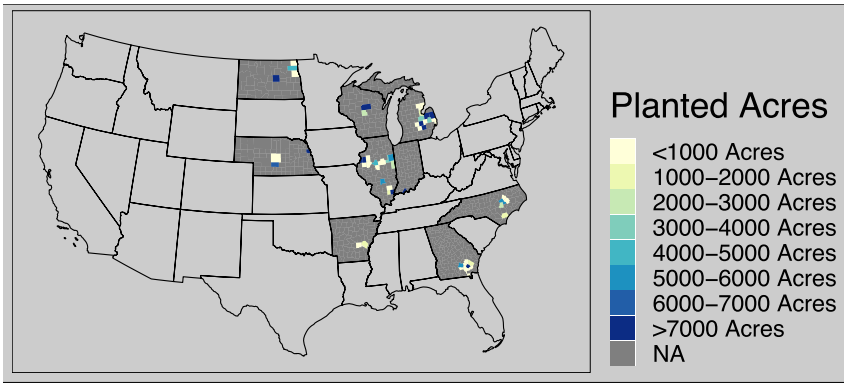


Figure 1: Map of US planted acreage from MLD provided by MyAgData® for the states of Arkansas, Georgia, Illinois, Indiana, Michigan, Nebraska, North Carolina, North Dakota, and Wisconsin. Shades of blue represent more planted acreage while yellow represents less.

2.2 USDA Farm Service Agency Common Land Unit Data

Through a cooperative agreement with the USDA Farm Service Agency (FSA), NASS obtains complete and up-to-date field-level boundary polygon information known as the Common Land Unit (CLU; Heald, 2002). CLUs are associated to FSA tabular "farm program" information, which relates each field to a particular producer or operation. The dataset is managed by FSA at the state-level in a geographic information system (GIS) in a Geodatabase format. A Geodatabase is a common proprietary format developed by the GIS software company Esri. Each state's CLU database can contain millions of field-level records. CLUs are hand-digitized over high-resolution aerial imagery to give a good representation of crop field boundaries. Boundaries are usually within a few meters of geographic accuracy. The FSA dataset is administratively confidential to the USDA.

2.3 USDA NASS List Frame Data

NASS maintains a list of all known farmers and ranchers in the US for sampling purposes. NASS's list frame has approximately five million records. Records within the list frame are coarsely georeferenced by site to a producer's mailing address. In other words, the list frame does not store detailed information on the location of the land that a producer farms. The list frame catalogs producers using a Person Operation Identification (POID). Additionally, for each POID, survey responses are also contained within the list frame.

The list frame and CLUs can be cross-referenced and linked by leveraging the CLUs' physical location, operator name, and supplementary factors. The approach of linking FSA data to NASS records is based on traditional NASS record-linkage processes. These processes are used to match names and addresses from the Geodatabase to the names and addresses on the NASS list frame. Overall, the CLUs and the list frame are administratively confidential to NASS.

2.4 Surveys for Comparison

To assess the quality of MLD-imputed survey responses for producing time-sensitive estimates, producers' MLD-imputed values were compared to historically reported survey values. Resulting

differences measured the imputation error associated with non-survey data (i.e., when MLD values were used instead of the historical survey values). Comparisons between MLD and historical survey values were conducted on two different NASS surveys: 1) the Crops Acreage, Production, and Stocks (APS) Survey; and 2) the Agricultural Yield Survey (AYS). Crops APS collects producers' information on planting, harvesting, and production of crops on a quarterly basis. The AYS collects monthly information on yields, but very limited information on the amount of planted and harvested acres. Producers are only sampled for Crops APS or AYS, and never for both. Comparisons focused on corn and soybean item-level survey variables from the 2023 Crops APS and AYS.

3 Methods

3.1 Geospatial Matching

MLD were geographically joined to the CLU data and subsequently linked to a producer on the NASS list frame, when possible. Multiple steps were needed to spatially join the data together for analysis. Those steps were as follows:

1. The MLD field-level geoJSON files were converted to the Esri Shapefile format using QGIS software. This format provided a geographic area boundary of the land where a planter or harvester operated. Related field-level statistics were also included for successive analyses on crop type and average yield.
2. The MLD and FSA CLU areas were geographically intersected using the Esri ArcGIS Pro software because the CLUs are stored natively as an Esri Geodatabase. Resulting in 17,722 field intersections.
3. The ratios between the intersected areas and their respective CLU areas were calculated as percentages. These ratios represented the percentages of CLU overlapping areas that ranged from zero to 100%. When several MLD polygons intersected a single CLU, their total intersected area was considered in the calculations. Conversely, when a single MLD polygon intersected several CLUs, only CLU-specific intersected areas were considered in the calculations.
4. The MLD-and-CLU data with CLU-overlap areas having percentages greater than a 90% threshold were considered geographically matched. Additionally, a minimum area of one acre (0.405 hectares) and a yield of less than 300 bushels/acre (roughly 20 metric tons/hectare) were required resulting in 4,735 field matches. The 12,987 fields not meeting these inclusion criteria were removed from further analyses. The development of the inclusion criteria was based on subject matter expertise and NASS' historical standards.
5. Where MLD-and-CLU records closely matched, their associated list frame records were linked through a unique producer identifier (POID) contained within both CLU and NASS data sets.

3.2 Calculating Machine-Logged Survey Values

Once the geospatial matching was completed, the MLD acreage was denoted by z_{ij} to highlight the linkages to producer i and CLU j. Let a_{ij} be the total acreage of producer i in CLU j, and let z_{ij} be the MDL acreage of producer i covered by CLU j. Then, the overlap ratio x_{ij} is defined as follows:

$$x_{ij} = \min\left(1, \frac{z_{ij}}{a_{ij}}\right).$$

This ratio ensured that only the portion covered by the machinery was assigned to a single producer. Furthermore, the overlap ratio was bounded at one because the excessive acreage was accounted for by another CLU. For record i, item-level MLD values $\hat{y}_i$ represent a producer's imputed responses to a single survey question. These values are calculated by the following equation:

$$\hat{y}_i = \sum_{j=1}^{n_i} x_{ij} a_{ij},$$

where n_i is the total number of CLUs associated with producer i. The formula above applies to planted and harvested acreage variables. Yield values are computed as the arithmetic average of a producer's harvested MLD fields for a single crop. Finally, item-level imputation values $\hat{y}_i$ are successively compared to true historical survey values y_i.

3.3 Assessment of Imputation Error

Several methods are available to quantify the imputation error between MLD and historical survey values. For example, when comparing the distributions via measures of central tendency, one can show which distribution tends to have greater or smaller values. Measures of central tendency are also used to evaluate the consistency and accuracy of the imputation approach. In addition, the variability of the two distributions can be examined by calculating measures of spread. Thus, smaller dispersion values are associated with a more precise approach. Despite these measures providing basic understanding between the two distributions, more sophisticated testing is needed for more reliable and meaningful conclusions. For all statistical analysis a significance level of ($\alpha = 0.01$) was used, ensuing a 99% confidence interval for the results.

The Empirical Moment Generating Function (EMGF) is foundational to the nonparametric test proposed by Collender and Chalfant (1986) to compare two distributions. This test simultaneously accounts for all empirical moments of two distributions allowing one to explore the distributional differences between the MLD and historical survey data. EMGF-based test was chosen for its simplicity, preciseness, and flexibility when working with small sample sizes, unlike t-tests and Kolmogorov-Smirnov tests. T-tests considers only mean differences, where the Kolmogorov-Smirnov test evaluates the maximum discrepancies between two distributions. The EMGF is specifically defined for a sample $\{s_1, s_2, \cdots, s_n\}$ as follows:

$$\hat{M}_S(t) = \frac{1}{n} \sum_{i=1}^{n} e^{t s_i},$$

which has estimated variance

$$\hat{\sigma}^2_{M_S}(t) = \frac{1}{n-1} \left[\sum_{i=1}^{n} e^{2t s_i} - \frac{1}{n} \left(\sum_{i=1}^{n} e^{t s_i} \right)^2 \right],$$

for a generic value of $t \in \mathbb{R}$. Therefore, when comparing two distribution functions the following test is considered:

$$T(t) = \frac{\sqrt{n} \left\{ \hat{M}_Y(t) - \hat{M}_{\hat{Y}}(t) \right\}}{\sqrt{\hat{\sigma}^2_{M_Y}(t) + \hat{\sigma}^2_{M_{\hat{Y}}}(t)}}, \tag{1}$$

where the standard error in the denominator is computed under the assumption that the mechanism of producing the MLD-imputed values y_i is independent from the mechanism that acquires

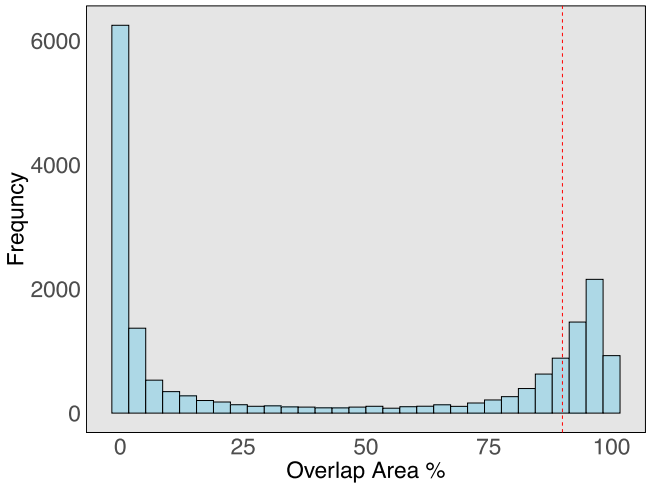


Figure 2: Distribution of overlap percentage of CLUs with MLD fields. Red dashed line represents the 90% cut off for inclusion in analysis.

historical survey data $\hat{y}_i$. The test in (1) is approximately distributed as standard Gaussian distribution (see Appendix C). However, because MLD-imputed and survey values are positive, the EMGFs of the log-transformed data are considered instead of computing them on the original data values. These data transformations are considered in this article to improve the computational stability of the test statistic. Therefore, the equivalence between the two data distributions is tested without losing validity on their log-transformed realizations.

The conclusions made from these tests are valid for the convenience sample considered in this article and cannot be used to make inference to the population. In general, more sophisticated weighted approaches are required to generalize the conclusions at the population level. Furthermore, other nonparametric tests of centrality can be performed to assess if the item-level differences of MLD-imputed and survey data are significant (Hollander and Wolfe, 1999). For example, the Wilcoxon signed-rank test statistic V for paired or matched data is computed by excluding the records i, such that $y_i = \hat{y}_i$, and it is defined as

$$V = \min\left(\sum_{i=1}^{n} R_i^+, \sum_{i=1}^{n} R_i^-\right),$$

where R_i^+ is the rank of $|y_i - \hat{y}_i|$ when $y_i > \hat{y}_i$, R_i^- is the rank of $|y_i - \hat{y}_i|$ when $y_i < \hat{y}_i$. In contrast to EMGF-equivalence tests, which are useful to determine if the inference based on MLD-imputed has the potential to be equivalent to traditional survey methods, nonparametric centrality tests at the item-level are used to determine if substantial bias is potentially introduced when replacing historical survey data with MLD-imputed values.

4 Results

Approximately 17,722 Common Land Units (CLUs) overlapped with the geospatial MLD polygons provided by MyAgData®. After applying the inclusion criteria described in Section 3.1, 4,735 CLUs (26%) of the original 17,722 CLUs remained for successive analysis (see Figure 2).

Table 1: Number of survey value comparisons.

	Corn	Soybeans	Total
Planted (Acres)	12	13	25
Harvested (Acres)	4	5	9
Yield (Bushels/Acre)	3	6	9
Total	19	24	43

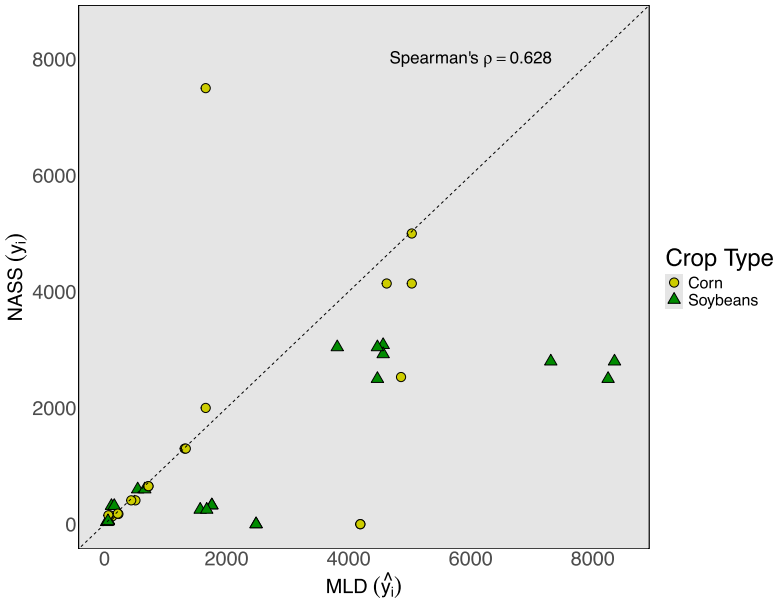


Figure 3: Scatter plot where the machine-logged (MLD) imputation values $\hat{y}_i$ are shown on the x axis and NASS historical survey values y_i are on the y axis. The dashed line represents a one-to-one association between the MLD and NASS historical survey values. Most points align with this line, with exception of approximately 17 points.

Of these 4,735 CLUs, 4,294 (90.6%) could be linked to a Person Operation Identification (POID) on the NASS list frame. At the list frame unit level, 58 (96.6%) out of 60 MyAgData® producers were matched to a producer on the NASS list frame.

Throughout 2023, 424 item-level MLD-imputed values $\hat{y}_i$ have been calculated for all producers with MLD. For the two surveys (Crops APS and AYS) and two crops (corn and soybeans), 43 item-level MLD-imputed values $\hat{y}_i$ were available to compare to historical NASS item-level survey values y_i (see Table 1).

Relationships between MLD-imputed values $\hat{y}_i$ and historical survey values y_i were investigated (see Figure 3). The Spearman's rho-correlation coefficient was used to evaluate the relationship between MLD-imputed values $\hat{y}_i$ and historical survey values y_i. There was a positive correlation between the two variables, and it was statistically significant ($\rho = 0.628$, with p-value $p < 0.01$ calculated with 43 data points). Overall, the results suggested that producers with higher historical survey values y_i would have larger MLD-imputed values $\hat{y}_i$.

Several survey variables (i.e., planted and harvested acreages, and yields) for two crops (i.e., corn and soybeans) were inspected to explore the differences between MLD-imputed and survey

Table 2: Results of Wilcoxon signed rank test comparing crops and variable type.

Crop Type: Corn		
Variable	V	p-value
Planted (Acres)	22	0.209
Harvested (Acres)	3	0.625
Yield (Bushels/Acre)	1	0.500
Crop Type: Soybeans		
Variable	V	p-value
Planted (Acres)	2	0.002*
Harvested (Acres)	3	0.312
Yield (Bushels/Acre)	5	0.293

Table 3: Median and standard deviation comparisons between MLD and NASS data for three variables and two crops.

Crop Type: Corn				
	$\hat{y}_i$ (MLD)		y_i (NASS)	
Variable	Median	St. Dev.	Median	St. Dev.
Planted (Acres)	1,657	1,979	977	2,377
Harvested (Acres)	884	2,076	855	1,827
Yield (Bushels/Acre)	216	55	175	25
Crop Type: Soybeans				
	$\hat{y}_i$ (MLD)		y_i (NASS)	
Variable	Median	St. Dev.	Median	St. Dev.
Planted (Acres)	2,483	2,598	600	1,342
Harvested (Acres)	1,567	2,964	600	1,397
Yield (Bushels/Acre)	59	11	50	9

values. Out of the 43 survey comparisons, 21 (48%) showed a difference of 25% or less between MLD and historical NASS survey values.

Wilcoxon signed rank tests were conducted to assess the presence of potential bias if historical values were replaced with MLD-imputed data. As shown in Table 2, a statistically significant difference between MLD-imputed and historical survey values ($p < 0.005$, calculated with 13 data points) was revealed for soybeans planted acreage. Large differences between MLD-imputed and historical survey values for corn planted, harvested, and yield and for soybean harvested and yield were not observed ($p > 0.01$, calculated with 30 data points).

The analyses of robust centrality measures, such as medians, and standard deviations provide further insights on the location and scale of the sample distributions. All medians of MLD-imputed values $\hat{y}_i$ for planted, harvested, and yield variables for both corn and soybeans are higher than historical survey values y_i (see Table 3). For example, the largest median discrepancy

Table 4: Summary statistics from the averaged equivalence-distribution tests based on the empirical moment generating function over 1000 values of $t \in [-1.5,\ 1.5]$.

Crop Type: Corn				
Variable	Average Difference	Average St. Error	Average z-score	p-value
Planted (Acres)	−945	9, 621	0.369	0.711
Harvested (Acres)	−451	7, 737	−0.243	0.807
Yield (Bushels/Acre)	−30	68	−0.113	0.909
Crop Type: Soybeans				
Variable	Average Difference	Average St. Error	Average z-score	p-value
Planted (Acres)	−6, 670	9, 890	−0.460	0.644
Harvested (Acres)	−4, 450	10, 471	−0.410	0.681
Yield (Bushels/Acre)	−1	10	−0.073	0.941

is observed between MLD and NASS median soybean planted acreage (approximately over 1, 800 acres); however, other discrepancies are not found to be statistically significant ($p < 0.01$) in later analyses (see the following paragraph). In addition, most standard deviations for MLD are higher than those for NASS values except for corn planted acreage. MLD-imputed values, $\hat{y}_i$, systematically exceed historical survey values, y_i, indicating the need for further evaluation.

EMGF-based tests are more comprehensive to compare two sample distributions. The evaluation of equivalence between the sample distributions of MLD-imputed values and historical survey values is conducted using EMGF tests. Several EMGF-equivalence tests were calculated using $K = 1000$ values of $t \in [-1.5,\ 1.5]$. The analysis was conducted on MLD and historical survey responses for all survey variables and crop types (see Table 4). Four metrics were used to measure the two distributions. First, the average difference across the 1000 values of t was calculated for the two distributions. If the average difference between the two EMGF is closer to zero, the two distributions have similar moments, and hence, they are statistically equivalent. Next, the average standard deviation quantifies the variability of the EMGF differences. For the 1000 values of t, z scores were calculated and averaged to produce a summary statistic that is normally distributed (see Appendix C). Therefore, the more they are away from zero, the more likely the two distributions differ. Finally, p-values were used to measure the evidence against the hypothesis of equivalent distributions. For $p > 0.01$, one can infer that there is enough evidence in the data to conclude that the two distributions are statistically identical. Overall, all EMGF tests were not statistically significant, providing a solid foundation for future studies to build on. (Additional data quality comparisons and further distribution testing analyses are provided in Appendix A and B.)

5 Discussion

The major accomplishment of the pilot study was the creation of the first pipeline at the national level in the US to link MLD to NASS historical survey values. Even though data from only 60 producers were analyzed, the pipeline and related procedures have been established to ingest more data for future large-scale integration. Comparing MLD-imputed data to historical NASS

values has provided insight into potential imputation errors that could be introduced if MLD were directly used in surveys. Although no statistically significant differences were found, the test for one item-level variable (i.e., soybean planted acreage) were inconclusive, and further investigation is needed to determine if the observed difference is due to imputation error or other factors such as sample size. In general, further investigation would need to encompass a wider range of crops, additional surveys, and aggregated county-level measurements.

Several technical lessons were learned from challenges addressed in the pilot study. The first one was related to the small number of producers that provided their data. Accessing producers' data can be complex due to securing permissions to share their MLD. The second challenge was related to conflicting observations; for instance, when a producer had high historical NASS data and low MLD-imputed values or vice versa (see Figure 3). Conflicting observations could be valid due to complex relationships between the owner and operator of the land, and the individual collecting the MLD. For example, when one producer uses a piece of equipment owned by another producer to plant or harvest, typically in exchange for a fee. Therefore, parsing MLD to correct individuals has been complicated. Addressing conflicting observations would require direct work with the individual producers to account for acres not included in the MLD.

Two significant aspects can contribute to a possible incorrect matching of MLD to a survey respondent. One aspect consists of geospatial data without name or operation information and makes it impossible to complete a questionnaire. Using the geospatial matching technique via the CLUs has yielded the ability to compare historical survey values to MLD. However, producers' field-level boundaries are not well defined; hence, the overlap between CLUs and the MLD can be quite different. Consequently, parts of fields are unaccounted for in the MLD-imputed values when compared to historical NASS survey values. The second aspect regards mapping the MLD to multiple frames and data structures can introduce error. All three separate frames (i.e., MyAgData®, FSA CLU, and NASS list frame) had substantial overlaps, yet none of them have a perfect one-to-one match. Consequently, as the MLD are compared from one frame to another, various miss-matches can occur. Over 300 MLD item-level imputations were not available in comparison to historical NASS survey values. For example, the producer with NASS survey values could have been excluded from the sample of either Crops APS or AYS in 2023. In another scenario, the producer could have been sampled but did not complete the 2023 questionnaire. Key factors contributing to mismatches also underscore the need for improved matching methods and broader data coverage.

6 Conclusion

This pilot study investigated if farm MLD could be used to directly impute for nonresponse by comparing it to historical survey data. Geospatial data from pieces of machinery to plant and harvest were matched to the respective producers on the NASS list frame. Nearly 90 % of the fields meeting the analysis inclusion criteria were matched resulting in MLD-imputed values for 58 of the 60 producers provided for the study. Values from two NASS surveys conducted in 2023 were compared to MLD-imputed values. Only corn and soybean variables were examined resulting in 43 item survey value comparisons between MLD and NASS data. Other crops and survey variables not included in this analysis are peanut acreage, cotton acreage, and total land operated resulting in approximately 95 item-level imputations. These other crops could be considered within the same statistical framework applied to corn and soybeans. Overall, there is a positive correlation between MLD-imputed values and historical survey values. Comparisons

of medians for planted, harvested, and yield variables were used to examine differences between the two sources of data. Soybean planted acreage via a Wilcoxon signed rank test showed a statistically significant difference between the MLD imputed and NASS survey medians. Evidence of a difference appears for that variable, while insignificant differences were found for the remaining variables. However, a series of EMGF equivalence tests showed statistically identical distribution between the MLD and NASS data for corn and soybean variables planted, harvest, and yield. In general, the replacement of NASS survey values with MLD would require additional processing and providing a new statistical framework under the assumption of sample exchangeability between the two sources of data (De Finetti, 2017).

Successful linkage of MLD to producers has established a foundation for exploring potential imputation error. Due to the small amount of producers' data, latency of receiving data, and potential error associated with data linkage, any operational implementation would require significantly more experimentation. Future research will focus on operationalizing methods tested in this pilot study and evaluating how to scale within a full NASS production environment. Planned work includes developing metrics to quantify acreage bias from linkage mismatches and assessing how such mismatches influence aggregated statistics. Additional effort is needed to document staff time, usage of computational resources, and data management steps required to process MLD for assessing production costs. For example, this pilot study required manual review of linkages and customized interfaces with external databases. Therefore, automation of these tasks is needed before implementing it on a national scale. Finally, future research will evaluate the performance of MLD beyond acreage statistics for other survey programs, including use of fertilizers and pesticides and other agricultural practices.

Supplementary Material

The supplementary material contains an R program that is made available in a zip file attached to this article.

A Data Quality Comparisons

Earlier results showed no substantial differences between MLD and NASS values, prompting further distributional comparisons. All values were transformed using $\log(1 + \text{value})$. For each variable, the empirical cumulative distribution functions (ECDFs) for NASS and MLD were computed. The plots in Figure 4 display the ECDFs log-transformed for each variable (where NASS values are depicted in blue and MLD in red). The comparisons of the ECDFs between the two data sources were used to study distributional differences. Therefore, additional statistics have been reported including the L1-norm and the Kolmogorov-Smirnov (KS) test and related p-values.

Results from comparing the L1-norm and KS test are provided in Table 5. The biggest difference between NASS survey data and MLD has been found on planted acreage for both corn and soybeans. By contrast, yield estimates are very similar between the two sources. Resulting patterns suggest potential improvements on how planting data are filtered or transformed to find what causes the discrepancies.

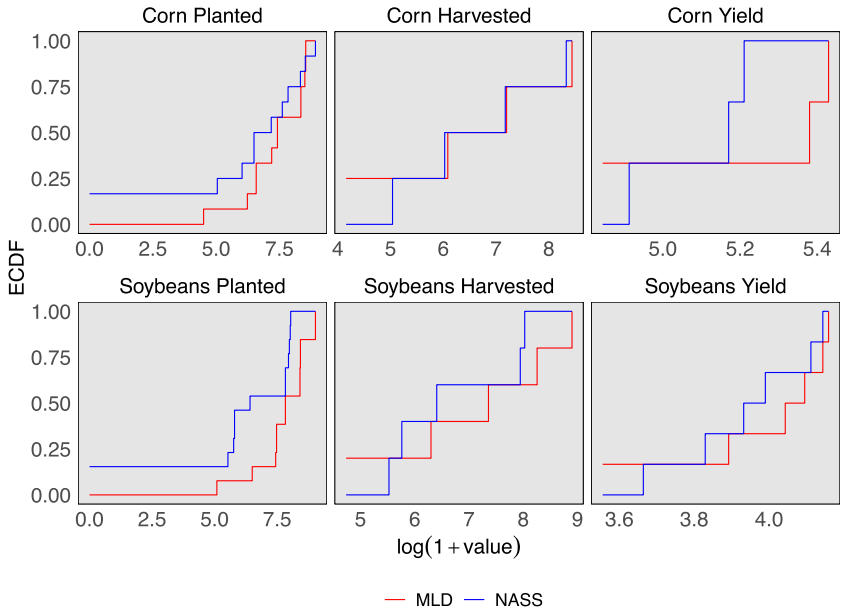


Figure 4: Empirical cumulative distribution functions (ECDFs) comparing NASS (blue) and MLD (red) across variables. The x axis shows $\log(1 + \text{value})$ and the y axis shows cumulative probability.

Table 5: Summary statistics from the equivalent-distribution tests based on the empirical moment-generating function.

Crop Type: Corn				
Variable	L1	L1 p-value	KS	KS p-value
Planted (Acres)	1.374	0.242	0.333	0.517
Harvested (Acres)	0.269	1.000	0.250	1.000
Yield (Bushels/Acre)	0.166	0.593	0.667	0.600
Crop Type: Soybeans				
Variable	L1	L1 p-value	KS	KS p-value
Planted (Acres)	1.878	0.023	0.462	0.122
Harvested (Acres)	0.695	0.811	0.400	0.873
Yield (Bushels/Acre)	0.073	0.771	0.333	0.896

B Power Analyses

A series of power analyses were conducted to quantify the minimum sample size required to achieve a statistical power of 0.8 and significance level of 0.01 for a chosen test in this pilot study. Entries in Table 6 represent the smallest sample size ($\hat{n}$), at which estimated power is the closest to the target of 0.8 under a nonparametric bootstrap setting. Let the following notation indicate test-specific sample size minima: $\hat{n}_W$ for Wilcoxon signed-rank, $\hat{n}_E$ for the EMGF-based equivalence test, $\hat{n}_{L1}$ for the L1-area permutation test, and $\hat{n}_{KS}$ for the Kolmogorov-Smirnov test.

Table 6: Minimum sample sizes reaching target power by test.

Crop Type: Corn				
Variable	$\hat{n}_W$	$\hat{n}_E$	$\hat{n}_{L1}$	$\hat{n}_{KS}$
Planted (Acres)	97	2	25	96
Harvested (Acres)	98	5	29	98
Yield (Bushels/Acre)	49	2	64	18
Crop Type: Soybeans				
Variable	$\hat{n}_W$	$\hat{n}_E$	$\hat{n}_{L1}$	$\hat{n}_{KS}$
Planted (Acres)	26	2	59	25
Harvested (Acres)	81	97	7	46
Yield (Bushels/Acre)	100	2	43	87

Analyses used $\log(1+\text{value})$ transformed paired MLD and survey observations when applicable. Results of this exploratory analysis rely on a convenience sample and its linkage quality to survey data, and they require larger design-based evaluation before any population-level generalization.

The power analysis results in Table 6 are compared with those in Table 1. All four tests for harvested acreage of corn and soybeans require sample sizes far exceeding the number of farms in the convenience sample. The EMGF tests for planted acreage and yield of both crops were conducted with more observations than the estimated minimum requirement of two. However, the remaining tests for planted acreage and yield across all crops analyzed would require substantially larger sample sizes. Overall, the EMGF test generally requires fewer observations to achieve the desired power of 0.8, whereas the other tests consistently demand much larger samples.

C Asymptotic Distribution of Integrated Test Statistics

Under the null hypothesis of distribution equivalence, the test statistics in (1) is asymptotically distributed as standard normal (Collender and Chalfant, 1986). However, when testing if the equivalence in distribution between MLD and historic survey data, the integrated test statistics

$$\mathcal{Z} = \int_{-\infty}^{\infty} T(t)\mathrm{d}t$$

is approximated using a definite integral and Riemann's definition:

$$\mathcal{Z} \approx \int_{-3/2}^{3/2} T(t)\mathrm{d}t \approx \frac{1}{K}\sum_{k=0}^{K} T\left(-\frac{3}{2}+\frac{3k}{K}\right). \tag{2}$$

Therefore, based on the central limit theorem, the results on right side of (2) is also distributed as a standard normal with convergence rate $O(K^{-1/2})$.

Acknowledgement

The authors would like to thank MyAgData® for providing expertise on the data. A. Dau contributed to numerous parts of the georeferencing process. Y. Chang, T. Murphy, and R.

Mueller provided valuable remarks on early drafts of this paper. In addition, the authors would like to thank the associate editor and the anonymous reviewers for their constructive feedback that improved the quality of this article.

The findings and conclusions in this presentation are those of the authors and should not be construed to represent any official USDA or US Government determination or policy.

References

Bradley R (1997). Potential Benefits from the Use of Scanner Data in the Construction of the CPI. In: 1996 ASA Proceedings.

Brimble P, McSharry P, Bachofer F, Bower J, Braun A (2020). Using machine learning and remote sensing to value property in Rwanda.

Collender RN, Chalfant JA (1986). An alternative approach to decisions under uncertainty using the empirical moment-generating function. *American Journal of Agricultural Economics*, 68(3): 727–731. https://doi.org/10.2307/1241557

De Finetti B (2017). *Theory of Probability: A Critical Introductory Treatment*. John Wiley & Sons.

Deines JM, Patel R, Liang SZ, Dado W, Lobell DB (2021). A million kernels of truth: Insights into scalable satellite maize yield mapping and yield gap analysis from an extensive ground dataset in the US corn belt. *Remote Sensing of Environment*, 253: 112174. https://doi.org/10.1016/j.rse.2020.112174

Dillman DA, Smyth JD, Christian LM (2014). *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. John Wiley & Sons, Indianapolis, Indiana, 17.

Dumbacher B, Capps C (2016). Big data methods for scraping government tax revenue from the web. In: *Proceedings of the Joint Statistical Meetings, Section on Statistical Learning and Data Science*.

Groves RM, Fowler FJ, Couper MP, Lepkowski JM, Singer E, Tourangeau R (2004). *Survey Methodology: Wiley Series in Survey Methodology*. 43–198. John Wiley & Sons, New York, USA.

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

Heberlein T, Baumgratner R (1978). Factors affecting response rates to mailed questionnaires: A quantitative analysis of the published literature. *Am Sociological Review*, 43: 447–462. https://doi.org/10.2307/2094771

Hollander M, Wolfe DA (1999). *Nonparametric Statistical Methods*. John Wiley & Sons, New York, NY, 2nd edition.

Hutchinson R, Scheleur S (2017). *Using Big Data to Enhance US Census Bureau Economic Data Products*. American Statistical Association, Alexandria, VA.

Hutchinson R, Scheleur S, Weidenhamer D (2023). Alternative data sources in the Census Bureau's monthly state retail sales data product. *Advances in Business Statistics, Methods and Data Collection*, 593–611. https://doi.org/10.1002/9781119672333.ch26

Leeuw ED, Callegaro M, Hox J, Korendijk E, Lensvelt-Mulders G (2007). The influence of advance letters on response in telephone surveys: A meta-analysis. *Public Opinion Quarterly*, 71(3): 413–443. https://doi.org/10.1093/poq/nfm014

Longva S, Thomsen I, Severeide PI (1998). Reducing costs of censuses in Norway through use of administrative registers. *International Statistical Review*, 66(2): 223–234. https://doi.org/10.1111/j.1751-5823.1998.tb00415.x

McFadden J, Njuki E, Griffin T (2023). Precision agriculture in the digital era: Recent adoption on US farms.

Murphy T, Rosales A, Abreu D, Young L (2022). Automatic imputation for an area survey. In: *Proceedings of the Joint Statistical Meetings, Survey Research Methods Section*. American Statistical Association, Alexandria, VA.

Rhodes S, Rosales A, Sartore L, Murphy T (2023). Uncertainty assessment for the imputation of an area survey. In: *Proceedings of the 2023 Joint Statistical Meetings*. American Statistical Association.

Snijkers G, de Jong T, Lam C, van Meurs C (2024). System-to-System Data Collection in business surveys applied to an agricultural survey: Small-scale pilot results.

Snijkers G, Punt T, De Broe S, Pérez JG (2021). Exploring sensor data for agricultural statistics: The fruit is not hanging as low as we thought. *Statistical Journal of the IAOS*, 37(4): 1301–1314. https://doi.org/10.3233/SJI-200728

Thompson A VHS (2025). Assessing Engagement and Outreach Initiatives: A Pilot Study for the Grain Stocks Report.

Tran HN, Gerling MW, McCarthy JS, O'Connor TP (2012). The Effectiveness of Automated Reminder Messages Used on the 2007 Census of Agriculture and on the National Animal Health Monitoring System 2011 Small Producer Study.