Fruit and Nut Production and Disposition Inquiry Bias Analysis

National Agricultural Statistics Service
Methodology Division
Summary, Estimation, & Disclosure Methodology Branch

Tom Pordugal
Lindsay Drunasky
Andrew Dau
Jeff Bailey

October 2020

BACKGROUND

USDA NASS conducts an annual Fruit and Nut Production and Disposition Inquiry (PDI) survey in February. Prior to the February 2020 survey, the survey was conducted in May. The survey collects information on total acres, bearing acres, harvested production (and yield), production dispositions, and income from specific marketing channels for a variety of non-citrus tree fruits and nut crops. Most of the NASS publications are driven by data collected via survey. The results from the PDI survey is published in May with the focus on bearing acres and production disposition, where the total acres are excluded from publication.

The Fruit and Nut PDI survey uses a multivariate probability proportional to size (MPPS) sampling design with the sampling weights calibrated to the frame state totals. Nonresponse groups are created using the probability of selection as a measure of size. The summary uses a traditional reweighted estimator to adjust the weights of the usable records to account for unit nonresponse. This is the most common unit nonresponse adjustment used for many of NASS surveys. As nonresponse increases, the risk of bias grows, and the reweighted estimator may not sufficiently adjust for nonresponse.

NASS completed a nonresponse bias study for the Fruit and Nut PDI survey in June 2016 which was submitted with the previous OMB docket renewal. The previous study used proxy data to “complete” the non-respondents and recalculated the total acreage estimates. The results suggested there was some downward bias in the total acreage estimates. Overall, 28 estimates were overestimated while 65 were underestimated, a 30% over and 70% under comparison. If the survey estimates were unbiased, we would expect roughly the same amount of states above and below the complete estimate. The proportion of estimates underestimated was statistically different from 50 percent (alpha=.01, using binomial distribution). Note that the “complete” dataset is not a perfect estimation of our population parameter, but it is the best obtainable comparison.

PLAN

This year we evaluated several options to assess the bias for the Fruit and Nut PDI survey and decided upon a plan to impute the tree fruit acreage based upon the frame data of known producers. The frame data for tree fruit acreage is relatively stable, and fruit trees likely are in the ground for many years. The model would then use the relationship of reported data to frame data for good respondents to impute for the nonrespondents. Other secondary covariates would also be introduced into the model as needed. Assuming this method works as expected, we could adjust for nonresponse using this imputation method operationally to estimate tree total area.

Total acres are collected and estimated for the survey but are not included in the publication. Bearing acres are published but not maintained as frame data. The ‘bearing acres to total acres’ ratio from the summary could be applied to the modeled total acres to calculate a “modeled” bearing acres. The surveys of interest for this study are the 2018 and 2019 crop year surveys conducted in April of 2019 and February of 2020, respectively. The crops of interest are apples and peaches as those are the two most common crops in the survey. This study includes the seven apple states and seven peach states that are in both publications. California peaches were also published in those two years, but the peach data is collected by type (clingstone and freestone), and the sampling method created complications in combining the two together in the model as all peaches.

SAS PROC MI

In the recent past, NASS has evaluated different types of multiple imputation (MI) methods and concluded that commercial off the shelf (COTS) imputation software has advantages to custom in-house imputation code (Dau and Miller, 2018). COTS software requires less maintenance, possesses greater flexibility, and has the potential to be utilized across multiple surveys. In this particular study, SAS PROC MI using the Predictive Mean Matching (PMM) method performed the best under three types of simulated missingness as well as when imputing for different variable types (continuous vs. categorical). As a result, SAS PROC MI PMM was recently implemented for imputing financial and production expenditure data in the Agricultural Resource Management Survey (ARMS) Phase 3 (USDA, 2020).

The data to be imputed in the Fruit and Nut PDI survey is semi-continuous and has a missing data pattern that is assumed to be missing at random. Based on the previously mentioned research, SAS PROC MI PMM was the clear choice for imputing tree fruit acreage for this survey. SAS deploys PROC MI within its SAS/STAT product. For this research, SAS 9.4 with SAS/STAT 14.1 was used (SAS, 2015).

METHODS AND ANALYSIS

To build the fully conditional specification (FCS) PMM model, covariate parameters were defined based on certain complete variables in the data set with total acreage as the response variable. Note that the apple and peach data sets did not have the same variables to serve as covariates. The modeled imputed variable (total acres) used as many as three covariates with the sampling frame data as the main covariate that was used in all models. Additional covariates used in some states were year, agricultural statistics district (ASD), and size. The year covariate was the year when the sampling frame data was last updated and was transformed into a categorical variable. The district was a categorical size variable created using the 2017 Census of Agriculture bearing acres at the district level. The size was based on the distribution of the main covariate.

For modeling apples, the seven states included Washington (WA), New York (NY), Michigan (MI), Pennsylvania (PA), California (CA), Virginia (VA), and Oregon (OR). The seven peach states included South Carolina (SC), Georgia (GA), New Jersey (NJ), Pennsylvania (PA), Colorado (CO), Michigan (MI), and Washington (WA). Table 1 marks with an ‘X’ the second and third covariate and the size grouping variable, if used in the model. Any covariate exclusion was based on limiting resources in the data.

Table 1. List of covariates used in model study*

Year=2019

Covariates	Apple States							Peach States						
	WA	NY	MI	PA	CA	VA	OR	SC	GA	NJ	PA	CO	MI	WA
Year	X		X		X	X	X							
District	X		X	X										
Size	X	X	X	X	X	X	X				X	X	X	X

Year=2018														
Covariates	Apple States							Peach States						
	WA	NY	MI	PA	CA	VA	OR	SC	GA	NJ	PA	CO	MI	WA
Year	X		X	X	X	X	X							
District														
Size	X	X	X	X	X	X	X			X	X	X	X	X

*Frame data was used as a covariate in all models.

The entire data sets were not input in the models. Data were subset into two data set groups. The first (excluded) group contained outliers that could bias the model results and consisted of records that were considered extreme operations to that crop. This group also included records that reported zero total acres for the crop. The second group for imputation were all other records.

The SAS PROC UNIVARIATE was used to examine the distribution of the covariate frame data going into the model. Ordinal size groups based on percentile ranges of the data distribution from box plots and the uniform distribution were made based on the output of the UNIVARIATE procedure. Next, analysis of the reported survey data and the frame data before model imputation was conducted and adjustments were made at the size group level when reported data were either above or below the frame data. For example, if the average ‘reported to frame data’ ratio was 1.2 (above the frame data) at the state size level for apples, then the frame data was multiplied by this factor prior to being used as a covariate for the imputation model. Then a second PROC UNIVARIATE checked data distributions again and to reset the size groupings, if needed, before going into the PROC MI models.

The models processed the incomplete missing data and the data patterns on modeled total acres were examined to see how well the model performed using the covariates. The imputed data were then combined back with the excluded data group described above. Note that all of the imputed data had positive total acres. To eliminate any bias, an algorithm was used to tease out the survey zeros, since it is quite possible that survey zeros exist in the non-respondent pool as well. First, the proportion of respondents that had frame data for apples and reported a valid zero on the survey at the strata level was identified. For example, given 400 reports for the crop and all of them had frame data for the crop, but only 300 reported positive crop data on the survey, then the ratio is equal to 0.75. Second, non-respondents with the crop on the frame were determined. A random uniform number was generated for each non-respondent on the scale of 0 to 1 by nonresponse group. If the random uniform number exceeded the ratio described above, then the crop value was set to zero, representing that the non-respondent was a survey zero and maintaining the relationship of survey zeros between the non-respondents and the respondents.

RESULTS

The data going into the imputation model were those records that had total frame data for the crop. PROC UNIVARIATE was used to create ordinal size groupings based on specific intervals in the data distribution using percentiles. After imputation processed, the model output were evaluated including the missing data patterns and average size evaluations were made to check the data patterns on the previously missing data. Appendix 1 shows an example of regression output from the Apple PROC MI model for apple acreage. In the Missing Data Patterns section of Appendix 1, the resulting average imputed value for Group 2 was 148.3 as compared to the average frame value of 136.4. For Group 1 the average reported value was 200.0, nearly identical to the average frame value of 199.8.

After running the PROC MI models, the resulting dataset was combined with the excluded group and reprocessed through the summary to expand the data. For total acres, there was no further nonresponse adjustment since all records had either reported or imputed data or were a valid zero for that crop.
The imputed summary consistently showed a higher total acres estimate when compared to the operational summary. The imputed summary reran for all the apple and peach states in 2018 and 2019 and comparisons were made to the operational summary. Most of the 28 imputed total acres estimates overestimated the operational summary, as expected. Twenty-three estimates were overestimated while five estimates were underestimated, an 82% over and 18% under comparison (see Table 2).

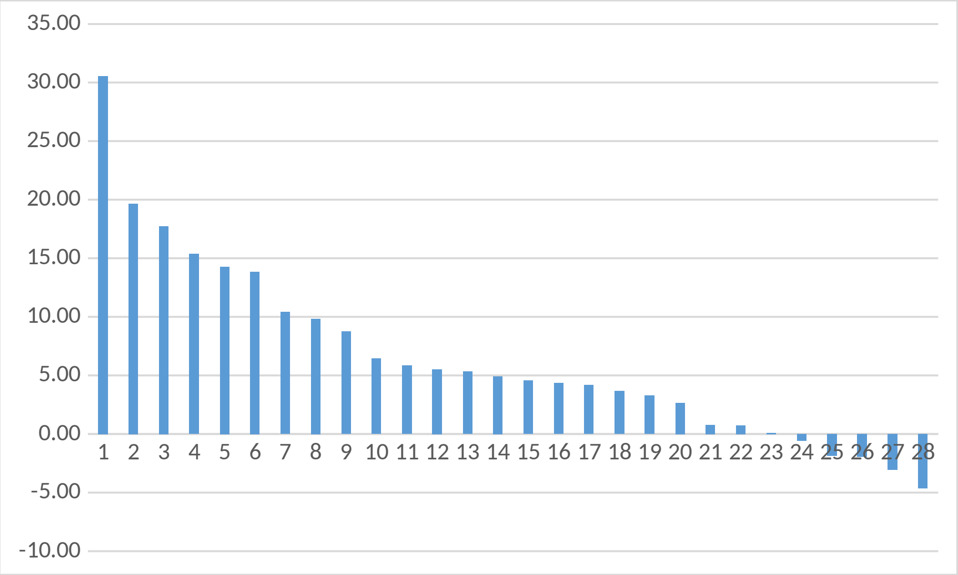
Table 2. Comparison between 2018 and 2019 imputed and operational total acres estimates

Crop	Total States	Overestimated	Percent Over	Underestimated	Percent Under
Apple	7	12	86%	2	14%
Peach	7	11	79%	3	21%
Total		23	82%	5	18 %

Using the confidence level of 95% (i.e. α = 0.05), an inverse binomial distribution calculation was made as BINOM.INV(n, p, $1–\alpha$) = BINOM.INV(28, 0.50, 0.95) = 18. This result meant that if 18 or more of the total acres estimates were overestimated then with 95% confidence the results bias towards overestimation. Since 23 estimates were overestimated, the actual alpha level is much smaller than 0.05.

The percent differences between the original reweighted estimates and imputed estimates are shown graphically in Figure 1. The average percent difference was 5.74% for apples and 7.18% for peaches.

Figure 1. Percent difference between reweighted and imputed estimates for all 28 comparisons.



In addition to survey indications, administrative data from other USDA agencies is available during the estimation process. In the surveyed states, strong administrative data exists from Farm Service Agency (FSA) and Risk Management Agency (RMA) for apple and peach total acreage. The 2017 Census of Agriculture estimates were also available just before the Noncitrus Fruits and Nuts 2018 Summary report was released. While the imputed estimates were consistently higher than the estimates from original operational summary, they were still generally lower than the final estimates. Fruit and Nuts PDI is a list only survey that assumes complete coverage of the target population. It is possible that we also have an undercoverage error that is preventing the survey estimates from reaching the level of the administrative data.

CONCLUSION

The operational summary uses a traditional reweighted estimator to adjust the weights of the usable records to account for unit nonresponse. The results of the PROC MI analysis show that the reweighted estimator does not sufficiently adjust for nonresponse with the resulting estimate being biased downward. The modeled total acres and the computed bearing acres can be incorporated into the operational summary to provide the commodity statistician with more accurate survey indications for the estimation process. This report only evaluates imputation of the total and bearing acreage for apple and peach trees. Additional research is needed to determine the methodology and develop models for the remaining crops and variables in the survey before implementation. Also, this output was generated using a singular imputation to match the fact that a singular dataset is used in our official estimates. More research should be conducted to determine the feasibility of multiple imputation techniques in our production setting.

REFERENCES

Dau, A. and Miller, D. (2018). “Dancing With the Software: Selecting Your Imputation Partner”. 2018 Joint Statistical Meetings Proceedings.

Groves RM. Nonresponse Rates and Nonresponse Bias in Household Surveys. Public Opinion Quarterly. 2006; 70(5):646–675.

SAS Institute Inc. (2015). SAS/STAT® 14.1 User's Guide. Cary, NC: SAS Institute Inc.

USDA NASS. (2020). Farm Production Expenditures Quality Measures. Retrieved from https://www.nass.usda.gov/Publications/Methodology_and_Data_Quality/Farm_Production_Expenditures/07_2020/fpxq0720.pdf

Appendix 1: PROC MI Example Model Output

```
***********************************************************************************************************************
The MI Procedure

Model Information

Method                               FCS
Number of Imputations                1
Number of Burn-in Iterations         20
Seed for random number generator     1868

FCS Model Specification

Method                     Imputed Variables

Regression                 Size Frame Year District
Regression-PMM (K= 5)      APPLE

Missing Data Patterns

Group    APPLE       Size    Frame   Year      District         Freq     Percent

    1    X           X       X       X         X                 240       34.43
    2    .           X       X       X         X                 457       65.57


      ---------------------------------Group Means-----------------------------------
Group           APPLE            Size           Frame            Year        District

    1      200.098333        2.341667      199.848750        1.237500        1.037500
    2               .        1.901532      136.439168        1.538293        1.019694



Regression Models for FCS Predicted Mean Matching Method

Imputed                   -Imputation-
Variable     Effect                  1

APPLE        Intercept       -0.072837
APPLE        Frame            0.939036
APPLE        Year             0.035893
APPLE        District        -0.004191
APPLE        Size            -0.074954

***********************************************************************************************************************
```