Characterizing Discrepancies in Reported Acreage between the Census of Agriculture and June Agricultural Survey

Michael E. Bellow[1] and Heather Ridolfo[2]

[1]National Agricultural Statistics Service, USDA, 1400 Independence Ave. SW, Washington, DC 20250
[2]National Agricultural Statistics Service, USDA, 3251 Old Lee Hwy, Fairfax, VA 22030

Abstract

The USDA's National Agricultural Statistics Service (NASS) launched a research effort to identify the causes of significant discrepancies in reporting of land related variables (in particular total farm acres operated) between the 2012 Census of Agriculture (COA) and June Agricultural Survey (JAS). NASS conducts the JAS (a probability-based sample survey of U.S. farm operators) annually and the COA (a complete enumeration of U.S. farms and ranches) every five years. JAS records were matched to corresponding ones from the COA (both unedited and edited), with those having absolute acreage differences exceeding a preselected threshold categorized as discrepancies and subjected to further investigation. The degree of influence of explanatory variables such as type of farm, number of operators and average drought level during the JAS data collection period on percentage and size of discrepancies associated with total acres operated was evaluated using descriptive statistics and logistic regression.

Key Words: Census of Agriculture, June Agricultural Survey, logistic regression, matching.

1. Introduction

In 2007, the National Agricultural Statistics Service (NASS) conducted a Classification Error Survey (CES) to try and identify why operations reported differently for the Census of Agriculture (COA) and June Agricultural Survey (JAS) (Abreu, Dickey and McCarthy, 2009). This study targeted operations classified as farms in the JAS and non-farms in the COA or vice versa, but also focused on operations reporting total farm acres operated that differed by more than 25 percent between the June survey and the census. Such operations were categorized as *discrepant*. Based on reinterviews conducted with 147 such operations, more discrepancies were found to be attributable to misreporting in the JAS than the COA. Large acreage differences were mainly due to respondent errors (e.g., providing estimates and excluding specific types of land such as pasture and woods), enumerator errors and differences in respondents between the June survey and census. A small number of discrepancies were the result of actual changes in acreage over the period between the JAS and COA. A limitation of this study is the fact that it was confined to five states and relied on a small sample of operations, so the findings may not be representative of the overall population.

In 2012, large differences were once again found between JAS and COA values of total acres operated reported at the individual farm level for a number of operations. These observations motivated NASS to conduct a study on a much larger scale than the 2007 CES (i.e., encompassing the entire country with the exception of Alaska). This investigation would be purely exploratory but with potential (based on findings) for future follow-up activities. Using matched JAS and COA samples, the study used quantitative techniques to answer the following research question: What factors are associated with the discrepancies in reported acreage operated between the 2012 JAS and COA?

A number of factors may influence acreage differences between the JAS and COA. This early exploratory study focuses on factors that can be readily investigated using survey data alone (such as farm type and number of operators). Consideration of factors more difficult to measure (e.g., respondent error) will be deferred to a future phase of the research effort.

The *farm type* (i.e., primarily crop or livestock) impacts how much land is needed and how the land is utilized in maintaining the operation. For example, farms and ranches specializing in livestock may be more likely to rent land or to use different land over the course of the year for grazing, and this may contribute to discrepancies in reported acres operated between the JAS and COA. Additionally, farms with multiple operators are more likely to have different individuals who provided data for the June survey and census, respectively. Some operators may have more knowledge of land use and are thus able to report the number of acres operated more accurately. *Operator tenure* (number of years that the principal operator has been operating the farm) may be related to discrepancies in the sense that the longer the tenure, the greater the likelihood of factors that could complicate reporting such as more land being acquired and different land acquisitions (owned, rented, share agreements). Drought conditions at the time of JAS data collection were used as a proxy for land usability. Operations that experienced drought during that period may not have usable land when data collection begins for the COA, which could lead to discrepancies in acres reported.

Survey characteristics may also impact data quality. Specifically, the *mode* of data collection and *elapsed time* between the JAS and COA are possible contributing factors. The JAS is an interviewer-administered survey, and such surveys typically attain higher data quality as interviewers can assist in comprehension of questions and reporting valid responses. The COA is primarily a mail survey although data are also collected via the web and CATI later in the data collection period. Interviewer-administered modes (face-to-face/CAPI, phone/CATI) are expected to have a lower percentage of discrepant records than other modes. The longer the period between JAS and COA data reporting, the higher the anticipated percentage of discrepant records due to changes in land acquisition and use over time.

2. Data and Methods

The main data sets used for this study were from the 2012 JAS and COA. The JAS is an area frame based sample survey that obtains data used to produce acreage estimates for various commodities. Data are collected on U.S. crops, livestock, grain storage capacity and type and size of farms. NASS' area frame is a nearly complete sampling frame where every acre of land has a known probability of selection. The sample units are designated land areas (called segments), typically one square mile (640 acres). Segments are further subdivided into tracts of land, each representing a different operating arrangement. The tracts are screened in advance to determine which ones are part of an agricultural operation (including both land inside and outside of tract boundaries). Field enumerators are assigned to visit tracts within sampled segments and collect data on all agricultural activity taking place within them. Farm operators are interviewed by the enumerators (either face-to-face or by phone) over a two-week period starting in early June, with information obtained not only about land within a segment but about the entire operation. The NASS area frame is also used by the COA to measure undercoverage of the census.

NASS conducts the COA, a complete enumeration of farms and ranches in the U.S., twice each decade in years ending in ‘2’ and ‘7’. Data are collected on a number of items, including land use and ownership, operator characteristics, income, expenditures and farming practices. Census forms are mailed to all known and potential agricultural operations in the nation starting in December and data obtained primarily via mail return over the next several months. As established by the U.S. Congress, a farm is defined as an operation for which at least $1,000 of agricultural products were produced and sold (or would normally have been sold) during the census year.

In preparing for data analysis, JAS records were first matched to corresponding ones from the COA based on known associations between census post office ID’s and JAS segment/tract identifiers. Due to multiple linkages occurring for certain COA and JAS records (especially those associated with large corporate operations), some modification of the original datasets was necessary in order to create the combined JAS/COA datasets which required a one-to-one correspondence. There were some COA records linked to multiple JAS records, each reporting data for the entire operation. For such records, if the COA operation type was ‘incorporated’, then the corresponding JAS values for total acres operated were averaged. If the COA operation type was not ‘incorporated’, the record was excluded from the combined data set unless the linked JAS records were identical (in which case the common value of total acres operated was used). A small number of JAS records were linked to multiple COA records due to the corresponding operations being geographically split in the JAS frame – for such records the COA values of total acres operated were summed. JAS records were separately matched to unedited and edited COA records to create two combined data sets.

The metric we used to identify discrepancies is called the *adjusted percent difference (APD)* and defined as follows:

$$\begin{aligned} APD &= 100(T_{(COA)} - T_{(JAS)}) \,/\, (T_{(COA)} + 100) \quad \text{if } T_{(COA)} > T_{(JAS)} \\ &= 100(T_{(JAS)} - T_{(COA)}) \,/\, (T_{(JAS)} + 100) \quad \text{otherwise} \end{aligned}$$

where $T_{(JAS)}$ and $T_{(COA)}$ are the reported values of total acres operated from the 2012 JAS and Census, respectively. This metric adjusts for the effect of scale so that (for example) if $T_{(COA)} = 7$ (acres) and $T_{(JAS)} = 5$ then the *APD* is 1.9 whereas if $T_{(COA)} = 700$ (acres) and $T_{(JAS)} = 500$ the *APD* is 25. Clearly, the latter is a more significant difference despite the fact that the standard percent difference is identical (29) for both. The same criterion of 25 percent or higher *APD* as in the earlier study was used to categorize operations as *discrepant*.

Operation characteristics were measured using *farm type*, *number of operators* and *operator tenure*. *Farm type* is a dichotomous variable that indicates whether the operation is primarily a livestock or crop farm (1 = livestock, 0 = crop). *Number of operators* is defined to be the number of individuals involved in making the day-to-day decisions, while *operator tenure* is the number of years that the principal operator has been involved in operation of the farm or ranch.

Drought level, obtained from the University of Nebraska's *Drought Monitor Classification Scheme* (University of Nebraska, 2015), is a continuous county-level measure of drought intensity over a time frame that includes the enumeration period for the 2012 JAS (May 29 to June 25). The data were provided in terms of percent of a county's area classified in each of six categories: 0 = no drought, 1 = abnormally dry, 2 = moderate drought, 3 = severe drought, 4 = extreme drought and 5 = exceptional drought. From this information, an average county level drought level was computed for use as an explanatory variable in subsequent analyses.

Survey characteristics were measured using the *mode* of the COA and *elapsed time* between the JAS and COA. *Mode* is a categorical variable that indicates the mode of data collection used for the census. The COA is collected primarily through mail but also via face-to-face (FTF) interviewing (including CAPI), telephone interviewing (including CATI) and the web. *Elapsed Time* is the number of days between data recording for the JAS and COA, which had to be estimated for some operations due to incomplete information.

3. Results

Table 1 summarizes the effect of COA data editing (conducted by NASS after data collection to correct errors and inconsistencies) on the number and percentage of discrepancies found between JAS and matched COA records. Note that of the 6,601 records identified as discrepant in the unedited combined data set , the editing process resulted in a change in total acres operated for 1,351 (20.5%). Of these edited discrepant records, 745 (55%) were resolved (i.e., converted to non-discrepant by editing). On the other hand, 102 non-discrepant records were broken (i.e., converted to discrepant by editing), accounting for 11% of the 929 non-discrepant records edited. The net result of data editing was to lower the overall number of discrepant records from 6,601 to 5,958 (only a 2.5% reduction).

Table 1. Effect of Data Editing on Number and Percent of Discrepancies in Total Acres Operated between 2012 JAS and COA

Item	Data Set	
	JAS/Unedited COA	**JAS/Edited COA**
No. Records	25,983	25,983
Discrepant Records	6,601 (25.4%)	5,958 (22.9%)
Discrepant Records Edited	1,351 (20.5%)	-
Discrepancies Resolved	-	745 (55.0%)
Non-Discrepant Records	19,383	-
Non-Discrepant Records Edited	929	-
Non-Discrepancies Broken	-	102 (11.0%)

In the remainder of this section, analyses performed on the combined data set of JAS and *edited* COA records are discussed. Table 2 shows the number and percent of discrepant records for six explanatory variables by specific values or ranges of values. *Percent Discrepant* (*PD*) (last column) is calculated based on the ratio between number of discrepant records and number of records in the category. Note that a slightly higher percentage of livestock farms than crop farms were identified as discrepant. *PD* was appreciably higher for farms with four or more operators (29.2) than for those with fewer than four, while only ranging from 21.6 to 24.4 for the *operator tenure* categories. With regard to *average drought level*, *PD* increased steadily with increasing severity of drought before leveling off at the highest category (‘*extreme or worse’*). The proportion of discrepancies was nearly five percent higher for *phone/CATI* (28.6) than for the next highest COA data collection mode category (*FTF/CAPI*). *PD* also increased with the categories of *elapsed time* (days) from JAS to COA including a jump of 3.2% from ‘*250-349’* to ‘*350 or more’*.

Table 2. Number and Percent Discrepant for Explanatory Variables

Variable	Category	Number in Category	Discrepant	
			Number	Percent
Farm Type	Crop	16,523	3,666	22.2
	Livestock	9,460	2,292	24.2
Number of Operators	1	14,157	3,216	22.7
	2	8,835	1,974	22.3
	3	2,134	518	24.3
	4 or more	857	250	29.2
Operator Tenure (Years)	<5	1,202	284	23.6
	5-14	4,686	1,123	24.0
	15-29	7,925	1,764	22.3
	30-39	6,625	1,432	21.6
	40 or more	5,545	1,355	24.4
Average Drought Level	None	12,784	2,633	20.6
	Abnormally Dry	8,151	1,950	23.9
	Moderate	3,399	869	25.6
	Severe	1,044	295	28.3
	Extreme or Worse	484	136	28.1
Mode of COA	Mail	20,036	4,471	22.3
	Web	2,810	614	21.9
	Phone/CATI	1,723	492	28.6
	FTF/CAPI	1,293	306	23.7
Elapsed Time (Days)	< 225	5,406	1,146	21.2
	225-249	10,778	2,368	22.0
	250-349	8,533	2,093	24.5
	350 or More	1,266	351	27.7

Figures 1 through 6 (end of paper following the references) are box plots of *adjusted percent difference* for the six variables in Table 2. These plots illustrate the relative size of acreage differences over different values or categories of the variables. The scale of the vertical axis is compressed in each of the plots in order to emphasize values of *APD* falling between 0 and 30. Note especially the generally increasing relationships between *APD* and both *average drought level* (Figure 4) and *elapsed time* (Figure 6).

Logistic regression was performed using a dichotomous variable (*Y*) of discrepancy for total acres operated (1 = discrepant, 0 = not discrepant) as the dependent variable and the following set of independent variables:

1. *Livestock farm* (1 if predominantly livestock, 0 otherwise)
2. *Number of operators*
3. *Operator tenure* (years)
4. *Average drought level*
5. *Mode = Phone/CATI* (1 if phone/CATI, 0 otherwise)
6. *Mode = Web* (1 if web, 0 otherwise)
7. *Mode = FTF/CAPI* (1 if FTF/CAPI, 0 otherwise)
8. *Elapsed time* (days) from JAS to COA data reporting

Note that items 5 through 7 are binary variables derived from the categorical *mode* variable. Table 3 shows: 1) test statistics and p-values from *Wald* χ^2 tests of whether the regression coefficient for a given independent variable is significantly different from zero, and 2) *odds ratios* and associated 95 percent confidence intervals. The *odds ratio* (*OR*) for an independent variable can be interpreted as follows: for a change of one unit in a given independent variable (with all of the others holding constant), the *OR* for a positive outcome (i.e., $Y = 1$) can be expected to change by the value of the regression parameter associated with that variable.

A confidence interval for the *OR* that does not include the value '1' suggests that the independent variable in question is influential with regard to predicting discrepancies/non-discrepancies. Table 3 indicates that livestock farms were more likely than crop farms to have an *APD* of at least 25 percent between the COA and JAS. The higher the level of drought experienced by operators during the JAS data collection period, the greater was the probability of a given record being discrepant. The *phone/CATI* mode of data collection was the most likely one to lead to discrepancies, while *elapsed time* was also significant. On the other hand, *number of operators*, *operator tenure* and the *web* and *FTF/CAPI* data collection modes were not influential with regard to acreage discrepancies.

Table 3. Logistic Regression Test Statistics by Independent Variable

Independent Variable	**Wald Test**		**Odds Ratio**	
	χ^2	p-Value	Value	95% Conf. Int.
Livestock Farm	20.1	<.0001	1.148	[1.081-1.219]
No. Operators	2.6	0.11	1.027	[0.994-1.06]
Operator Tenure	2.19	0.14	1.002	[1.0-1.004]
Average Drought Level	86.1	<.0001	1.137	[1.107-1.169]
Mode = Phone/CATI	6.9	.009	1.198	[1.047-1.371]
Mode = Web	0.31	.58	0.973	[0.883-1.072]
Mode = FTF/CAPI	0.26	.61	0.963	[0.832-1.114]
Elapsed Time (JAS to COA)	12.0	.0005	1.001	[1.001-1.002]

4. Discussion

The purpose of this study was to follow up on the 2007 CES which found large differences in reported acreage for total acres operated between the Census of Agriculture and June Agricultural Survey. Building upon that earlier work, a large, representative sample of matched JAS and COA records was used to explore factors that may be associated with the discrepancies found in 2012. The percentage of discrepant records was evaluated by specific values or ranges of values of a set of candidate explanatory variables. Logistic regression was then performed using a dichotomous variable indicating discrepancy or non-discrepancy as the dependent variable and eight explanatory variables as regressors. Four of the independent variables (*livestock farm*, *average drought level*, *phone/CATI* and *elapsed time* between the JAS and COA) showed significant association with discrepancies in reported acreage as measured by *Wald* χ^2 tests and *odds ratios*.

Several questions still remain: 1) why are these factors associated with acreage discrepancies, 2) are there other influential factors that we have not already considered, and 3) what action(s) can be taken to reduce or eliminate the discrepancies?

In the next phase of this research effort, we plan to: 1) investigate odd or unusual patterns (e.g., more than 60 records having total land = 1 in the COA but exceeding 100 in the JAS), and 2) employ data mining techniques such as classification trees or cluster analysis. Data mining enables sifting through large data sets (such as the COA) to try and detect operational characteristics related to reporting errors. For example, McCarthy and Earp (2009) used classification tree methodology to identify operations showing consistent reporting errors in the 2002 Census of Agriculture. One limitation of this research is the presence of additional factors that could be associated with acreage discrepancies, for example actual change in acres (buying/selling/renting out), respondent changes and reporting errors that cannot be measured using JAS and COA data alone. Alternative approaches may be necessary in order to evaluate the impact of such factors and make recommendations regarding potential remedial measures.

5. References

Abreu, Denise A., Dickey, Nancy J. and McCarthy, Jaki S. (2009). “2007 Classification Error Survey for the United States Census of Agriculture”. Washington DC: National Agricultural Statistics Service. RDD Research Report RDD-09-03.

McCarthy, Jaki S. and Earp, Morgan S. (2009). “Who Makes Mistakes? Using Data Mining Techniques to Analyze Reporting Errors in Total Acres Operated”. Washington, DC: National Agricultural Statistics Service. RDD Research Report RDD-09-02.

University of Nebraska-Lincoln. 2015. U.S. Drought Monitor Classification Scheme. http://droughtmonitor.unl.edu/AboutUs/ClassificationScheme.aspx. Accessed June 11, 2015.

Figure 1. Box Plot of APD by Farm Type

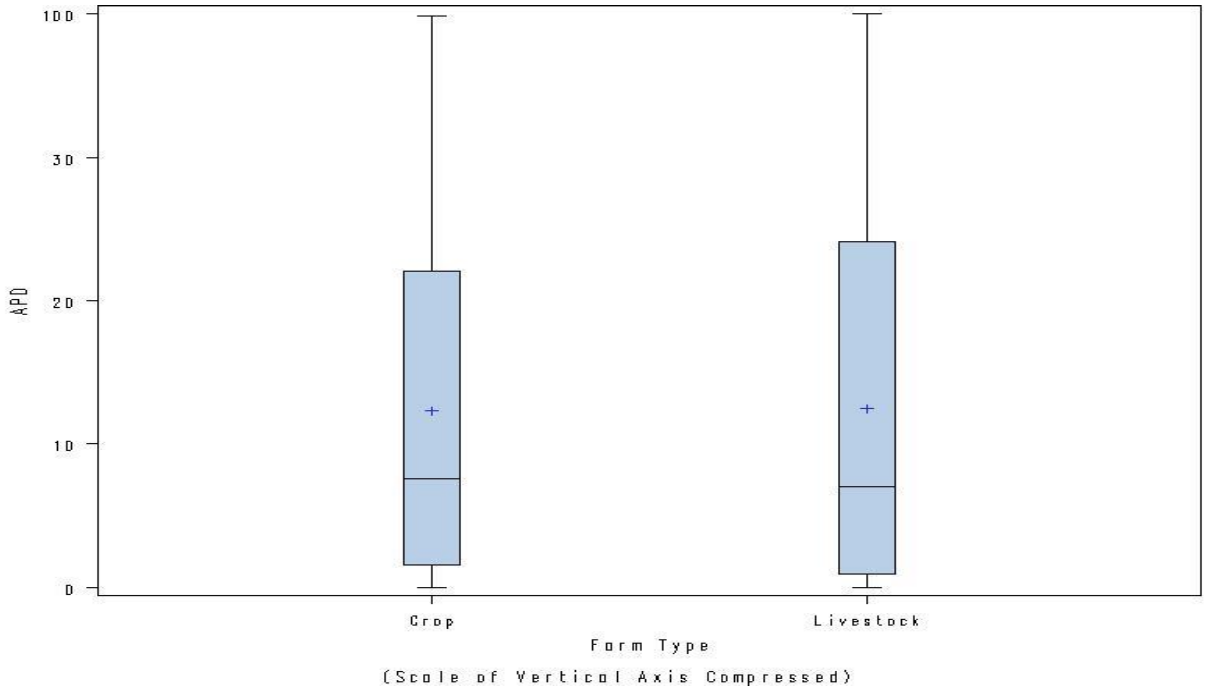


Figure 2. Box Plot of APD by Number of Operators

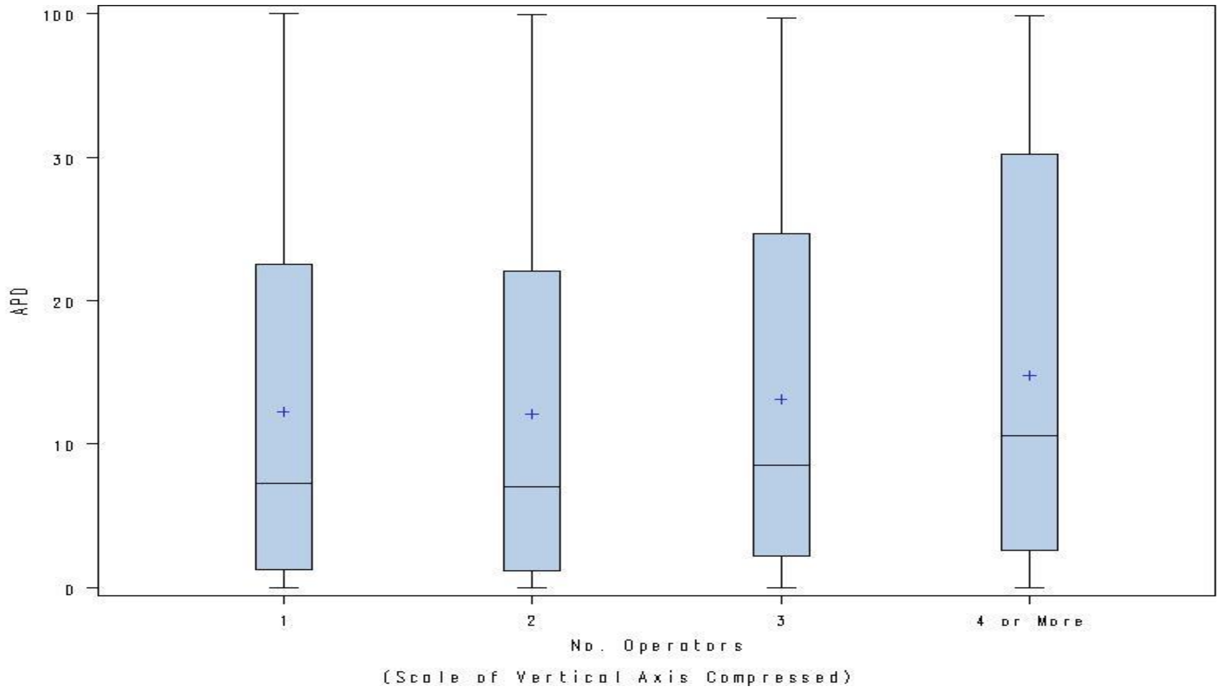


Figure 3. Box Plot of APD by Operator Tenure

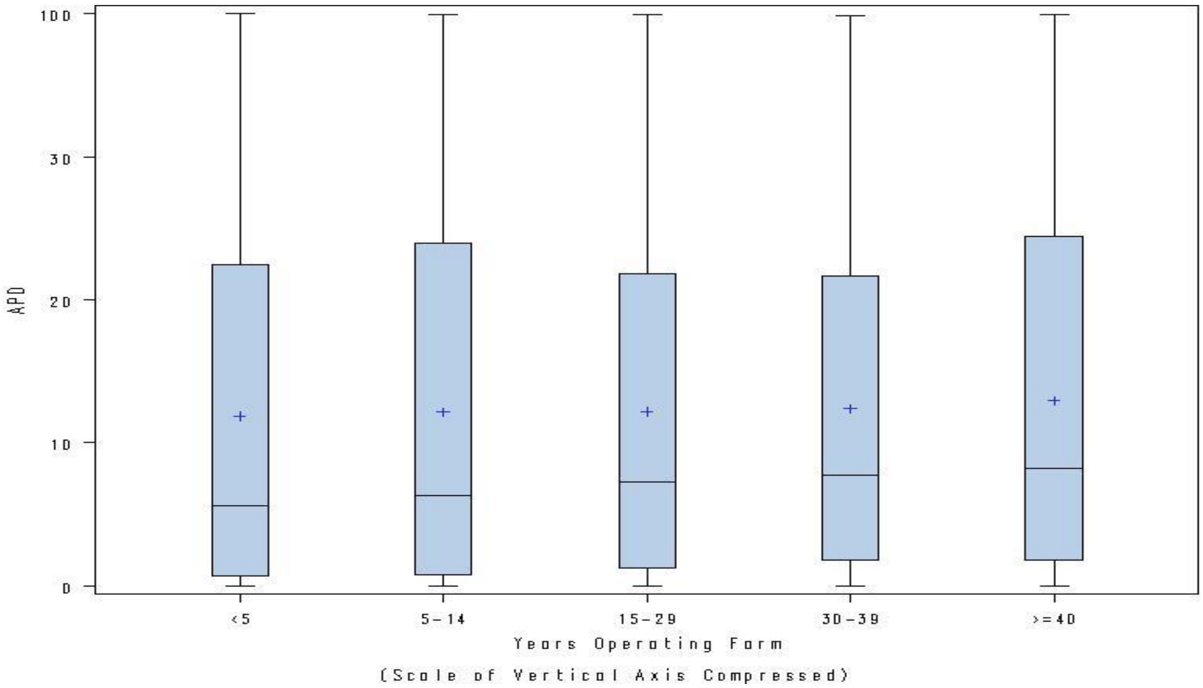


Figure 4. Box Plot of APD by Average Drought Level

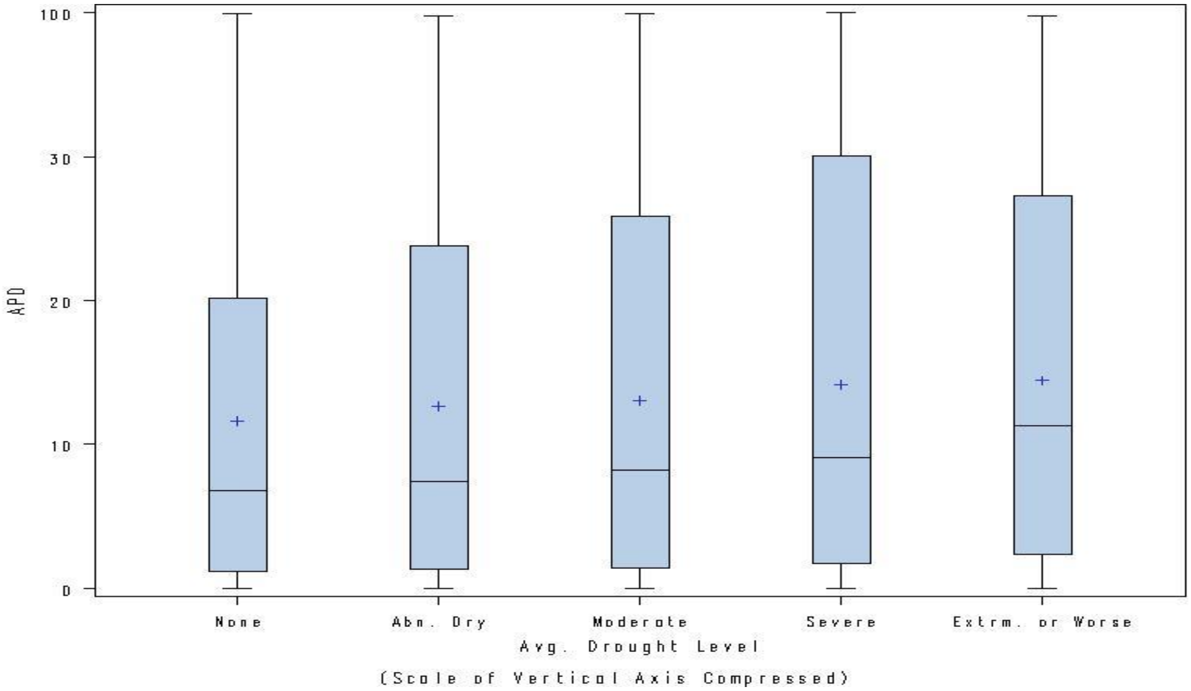


Figure 5. Box Plot of APD by COA Data Collection Mode

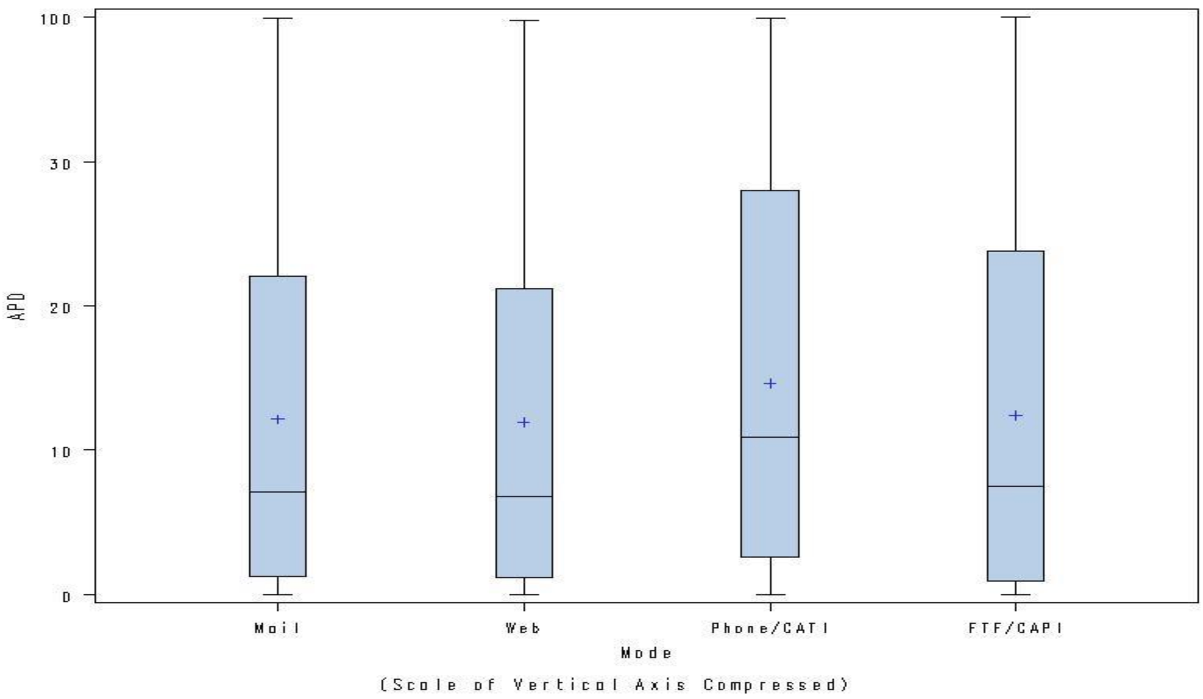
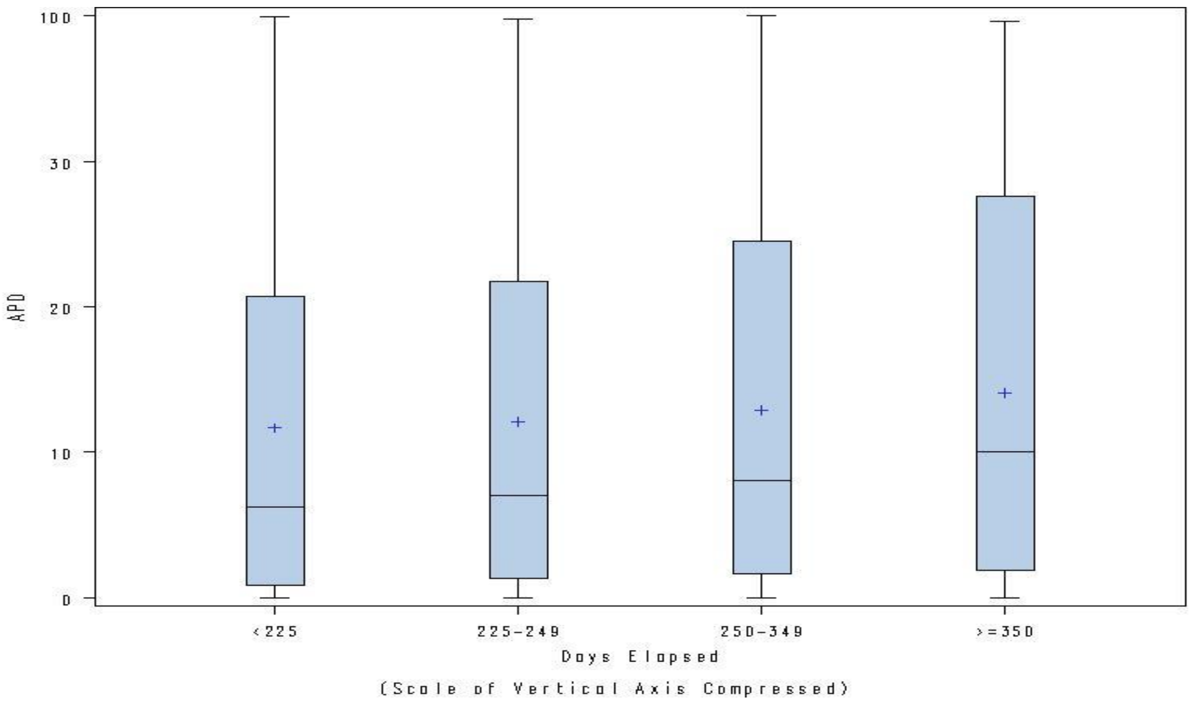


Figure 6. Box Plot of APD by Elapsed Time (JAS to COA)


APD
<225
>=350
Days Elapsed
(Scale of Vertical Axis Compressed)