DATA IMPUTATION STUDY ON

OKLAHOMA DES

by

RAYMOND R. BOSECKER

Sampling Studies Section
Sample Surveys Research Branch
Statistical Reporting Service
United States Department of Agriculture

October 1977

CONTENTS

	Page
INTRODUCTION	1
IMPUTED DATA	1
SOURCE OF IMPUTED DATA	4
SUMMMARY	8

DATA IMPUTATION STUDY ON OKLAHOMA DES

INTRODUCTION

In conjunction with the 1976 December Enumerative Survey (DES), the Oklahoma SSO recorded all data edited into the DES and the source of this edited data. Manual data imputation is required in the DES to complete any questionnaire with missing data. Following is a brief analysis of their data resulting from edit action. We would encourage each State to carefully examine the impact that missing data imputation is having on their estimates. Too often the effect of edit action is forgotten once the questionnaires are made "complete". Your faith in the estimate and your determination to take corrective action may be influenced by the amount of data imputed and the reliability of the source of the data.

IMPUTED DATA

Illustration 1 provides a relative measure of the impact of imputed data on certain Oklahoma 1976 DES indications. Data were considered imputed when the response code was refusal or no response.

Percent nonresponse is not always a good measure of the proportion of data imputed to the indications. In most cases in Illustration 1 the percent of data imputed is less than the percent of tract operators who refused or were inaccessible. Notable exceptions were tract acres and the weighted expansions for imputed bulls and beef cow replacement heifers as a percent of their respective totals.

Dairy cattle comprise a small percentage of total Oklahoma cattle so perhaps it was not surprising that no milk cows were imputed. However, no other heifers 500 pounds plus were imputed and only 4.6 percent of total steers 500 pounds plus resulted from edit action. Perhaps some of the imputed replacement heifers should be other heifers or maybe those who refused are not cattle feeders. There is no way to tell for sure.

Illustration 1: Proportion of Survey Indications from Imputed Data for Selected Items, Oklahoma 1976 DES

Item	Percent of Total	3 6 9 12 15
DES Questionnaires	(9.5)	xxxxxxxxxxxxxxxxxxxxxxxxxxxx
Tract Indications		
Acres	(11.6)	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
Cattle and Calves	(7.5)	xxxxxxxxxxxxxxxxxxxxx
Beef Cows	(8.2)	xxxxxxxxxxxxxxxxxxxxxxx
Milk Cows	(0)	
Bulls	(8.7)	xxxxxxxxxxxxxxxxxxxxxxxx
Replacement Beef Heifers	(7.6)	xxxxxxxxxxxxxxxxxxxx
Other Heifers	(0)	
Steers	(7.3)	xxxxxxxxxxxxxxxxxxx
Calves	(7.4)	xxxxxxxxxxxxxxxxxxx
Weighted Indications		
Cattle and Calves	(7.8)	xxxxxxxxxxxxxxxxxxxx
Beef Cows	(8.5)	xxxxxxxxxxxxxxxxxxxxxxx
Milk Cows	(0)	
Bulls	(10.5)	xxxxxxxxxxxxxxxxxxxxxxxxxxx
Replacement Beef Heifers	(13.2)	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
Other Heifers	(0)	
Steers	(4.6)	xxxxxxxxxxxxx
Calves	(7.7)	xxxxxxxxxxxxxxxxxxx
Cows to Calve	(8.3)	xxxxxxxxxxxxxxxxxxxxx
Calves Born Since 1/1/76	(7.0)	xxxxxxxxxxxxxxxxxx
Cattle and Calf Deaths	(5.8)	xxxxxxxxxxxxxxx

Illustration 1 is based on expanded totals. To examine the data supplied for refusals and inaccessibles we can compare the averages of the raw data for these reports with the means of reported data from operators, their spouses, or other reporters.

The means of reported and edited data for each respondent group are shown in Table 1 for selected variables from the DES.

TABLE 1: Mean Values for Selected Items by Respondent Category, Oklahoma 1976 DES

RESPONSE	NUMBER OF REPORTS	MEAN TOTAL ACRES	MEAN TRACT ACRES	MEAN TOTAL CATTLE	MEAN WTD. TRCT CATTLE	MEAN TRACT CATTLE
Operator	543	1007	184	112	23	25
Spouse	98	460	127	46	16	17
Other	75	910	185	103	19	15
Refusal	44	1925	297	104	24	28
Inacc.	31	695	157	42	10	6
Combined	791	969	182	100	22	23

RESPONSE	MEAN WTD. TRCT CALF INV.	MEAN TRACT CALF INV.	MEAN WTD. TRCT CALVES BORN	MEAN WTD. TRCT BEEF REPLACEMT	MEAN TRACT BEEF REPLACEMT	MEAN WTD. TRCT STEERS	MEAN TRACT STEERS
Operator	7	7	9	1.5	1.5	2.5	3.5
Spouse	5	7	6	1.0	1.0	.9	1.3
Other	9	5	6	.3	.4	.9	.9
Refusal	8	9	9	2.0	1.7	.8	2.5
Inacc.	3	1	4	.3	.2	.9	.7
Combined	7	7	8	1.3	1.3	2.0	2.8

From the above table it appears that operations were smaller where the enumerator accepted data from the operator's wife or was unable to contact anyone. On the other hand, operators who refused to cooperate appear to have had substantially larger operations than average. The acreage data includes extreme operators while the cattle data excludes them. This corresponds to the way the DES is summarized. Extreme operator data were not analyzed.

Even after excluding extreme operators, the tract and weighted cattle means are larger than those of the operator reported data. Means for refusals and inaccessibles are entirely different since a large proportion of the inaccessibles are zero reports. Note also that the tract and weighted averages were generally closer together for the first two respondent groups than for the last three. The tract and weighted averages per report were farthest apart for steers 500 lbs. and over. Also, unlike the other indications, means of refusals for steers were below operator reported means.

SOURCE OF IMPUTED DATA

In addition to the quantity of imputed data, one should also consider the quality of the data substituted for a missing report. The Oklahoma SSO recorded the edit source, whether the data was observed and who was the observer. We are now ready to examine the bases upon which the data in the last two categories of Table 1 are imputed into the system by the statistician. The frequencies of each source used for editing expressed as a percent of total number of reports imputed are shown in Table 2.

Table 2: Frequency of Imputed Reports by Observer and Edit Source for Selected Items, Oklahoma 1976 DES

Observer-Edit Source	Frequency of Occurance: Farm Acres Owned (%)	Farm Acres Rented (%)	Tract Acres (%)	Tract Cattle (%)	Total Cattle (%)
1. Enum - Enum Notes	47	47	11	67	52
2. Enum - JES	15	12	69	0	0
3. Other - Enum Notes	17	13	14	14	13
4. No Obs. - Enum Notes	3	8	2	3	5
5. No Obs - JES	11	10	4	3	5
6. No Obs - Stat Edit	7	10	0	13	25
	100	100	100	100	100

Approximately half the imputed questionnaires were based on enumerator observations; two-thirds for tract cattle. Enumerator notes from others, including ASCS, county agents, neighbors, relatives or friends, accounted for about 15 percent of the reports. Around 20 percent of the reports requiring imputation of total farm acres owned had no observations and nearly 30 percent of reports with imputed acres rented were without observation. Approximately 35 percent of the reports with imputed entire farm cattle could not be observed. Only six percent of the imputed tract acreage reports had no direct observation. Thirteen percent of the reports with imputed tract cattle were due to stat edit and another six percent contained data from enumerator notes or JES without observation.

The proportion of the total data, weighted by the expansion factor, which was imputed from each source is presented in the following table.

Table 3: Proportion of Imputed Data by Observer and Edit Source for Selected Items, Oklahoma 1976 DES

Observer-Edit Source	Proportion of Imputed Data: Farm Acres Owned (%)	Farm Acres Rented (%)	Tract Acres (%)	Tract Cattle (%)	Weighted Cattle (%)
1. Enum - Enum Notes	56	18	7	72	47
2. Enum - JES	5	1	76	0	0
3. Other - Enum Notes	9	13	6	3	8
4. No Obs. - Enum Notes	15	48	3	2	5
5. No Obs. - JES	11	14	8	1	7
6. No Obs - Stat Edit	4	6	0	22	33
	100	100	100	100	100

Except for rented acreage most of the imputed data was based on observation. However, 22 percent of the imputed tract cattle data and 33 percent of the imputed weighted cattle resulted from edit action with no assistance from observation, notes or JES data. From Illustration 1, about 7.5 percent of the tract and weighted survey cattle indications were from imputed data so approximately 1.5 to 2.5 percent of the indications are solely from stat edit. The heaviest impact of imputed data without observation falls on the weighted indication because not only are more of the cattle on the entire farm unobservable but also a large portion of the imputed entire farm acres which influence the weighted indication cannot be observed.

The proportion of imputed data by source for the remainder of the cattle items are shown in Table 4. Individual items vary quite a bit in the proportions imputed with and without observation. Births and deaths cannot be observed at the time of the interview but some of these data were based on notes from enumerators.

Table 4: Proportion of Imputed Data by Observed and Edit Source for Cattle Subgroups, Oklahoma 1976 DES

Code	All Beef Cows (%)	All Bulls 500 + lbs (%)	All Heifers for Beef Cow Replacement (%)	All Steers 500 + lbs (%)	All Calves (%)	All Expected to Calve (%)	All Calves Born (%)
1. Enum - Enum Notes	44	41	26	53	30	0	0
2. Enum - JES	0	0	0	0	0	0	0
3. Other - Enum Notes	9	9	15	16	11	0	0
4. No Obs. - Enum Notes	6	7	12	0	7	37	48
5. No Obs. - JES	6	8	0	31	8	0	7
6. No Obs. - Stat Edit	35	35	47	0	44	63	45
	100	100	100	100	100	100	100

Code	All Cattle and Calf Deaths (%)	Tract Beef Cows (%)	Tract Bulls 500 + lbs (%)	Tract Replacement Heifers (%)	Tract Steers 500 + lbs (%)	Tract Calves (%)
1. Enum - Enum Notes	0	53	53	40	72	44
2. Enum - JES	0	0	0	0	0	0
3. Other - Enum Notes	0	5	0	0	28	1
4. No Obs. - Enum Notes	36	4	4	7	0	0
5. No Obs. - JES	7	4	4	0	0	6
6. No Obs. - Stat Edit	57	34	39	53	0	49
	100	100	100	100	100	100

SUMMARY

There is no way of knowing what the survey value would have been if everyone in the sample for the Oklahoma DES had responded. However, we can get a better feel for how the survey indication is being influenced by imputation for missing data. The following questions can be answered by data analysis:

1. What percentage of the total indication comes from imputed data? (See Illustration 1)

2. On the average, how does the imputed data compare with reported data? (See Table 1)

3. Are there certain items where the proportion of the data imputed or the relationship between imputed data and reported data are different from the other items? (Perhaps steers and/or heifers for beef cow replacement in Illusration 1 and Table 1)

4. What are the sources of the imputed data and what proportions of the total data imputation do they comprise? (See Tables 2, 3 and 4)

The answers to these questions raise other questions which can only be answered by the survey and commodity statisticians.

1. Is the proportion of missing data that must be imputed causing doubts about survey results?

2. Do the relationships between imputed data and reported data appear reasonable given the control data and background information available for each respondent group?

3. Are there good reasons for some items to have a much larger share of imputed data than other items?

4. Can more be done to prevent refusals?

5. Can more be done to increase current information about refusals so that a lower proportion of missing data is supplied blindly?

6. How much faith do I have in the data imputed from each of the edit sources?

Analysis after the survey period may be helpful for future surveys but if coding and summarization could be accomplished at the time of the survey the stat interpretation, estimate and comments could include the impact of data edited into the survey. As long as the impact is slight or can be explained then we can feel comfortable with the survey indication. If the outcome of some parts of the survey has been altered by imputed data for no apparent reason then the statistician can evaluate a range of possible outcomes. This range may be wider than the sampling error associated with the point estimate. In any event, it is not knowing the consequences of our edit action which causes concern.