Dealing with Messy Data

October 22 2019

Office of Risk Assessment and Cost-Benefit Analysis (ORACBA)
Office of Chief Economist (OCE)
Science Policy and Risk Forum

Jeff Bailey, Chief
Summary, Estimation, and Disclosure Methodology Branch
Methodology Division
USDA/NASS

National Agriculture Statistics Service (NASS)

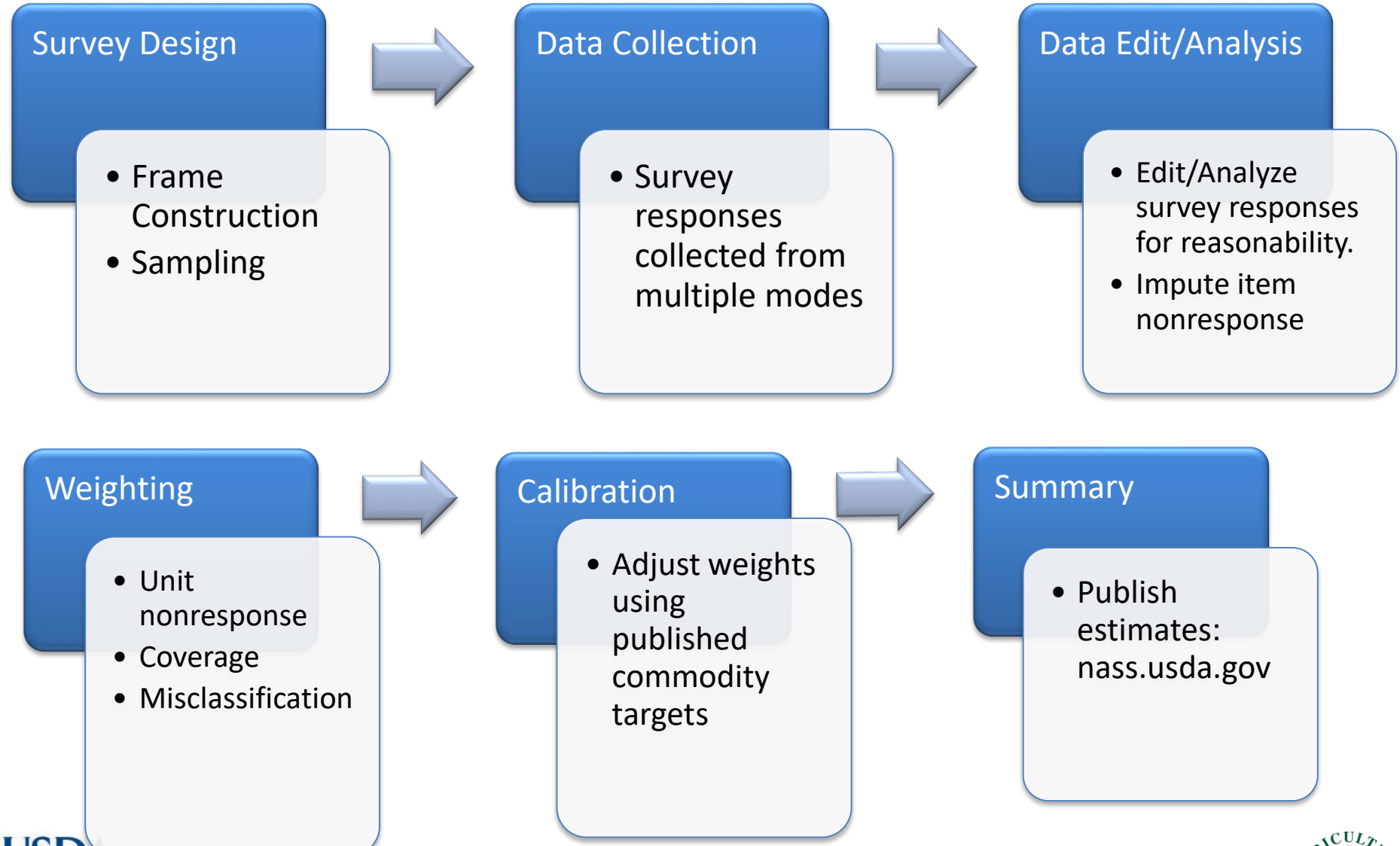
The NASS Mission:

The NASS mission is to provide timely, accurate, and useful statistics in service to U. S. agriculture.

Messy Data Outline

- Survey Quality
- Finding Data Errors
  - Edit
  - Analysis
- Handling Nonresponse
  - Item Nonresponse
  - Unit Nonresponse

NASS Survey Process Flow



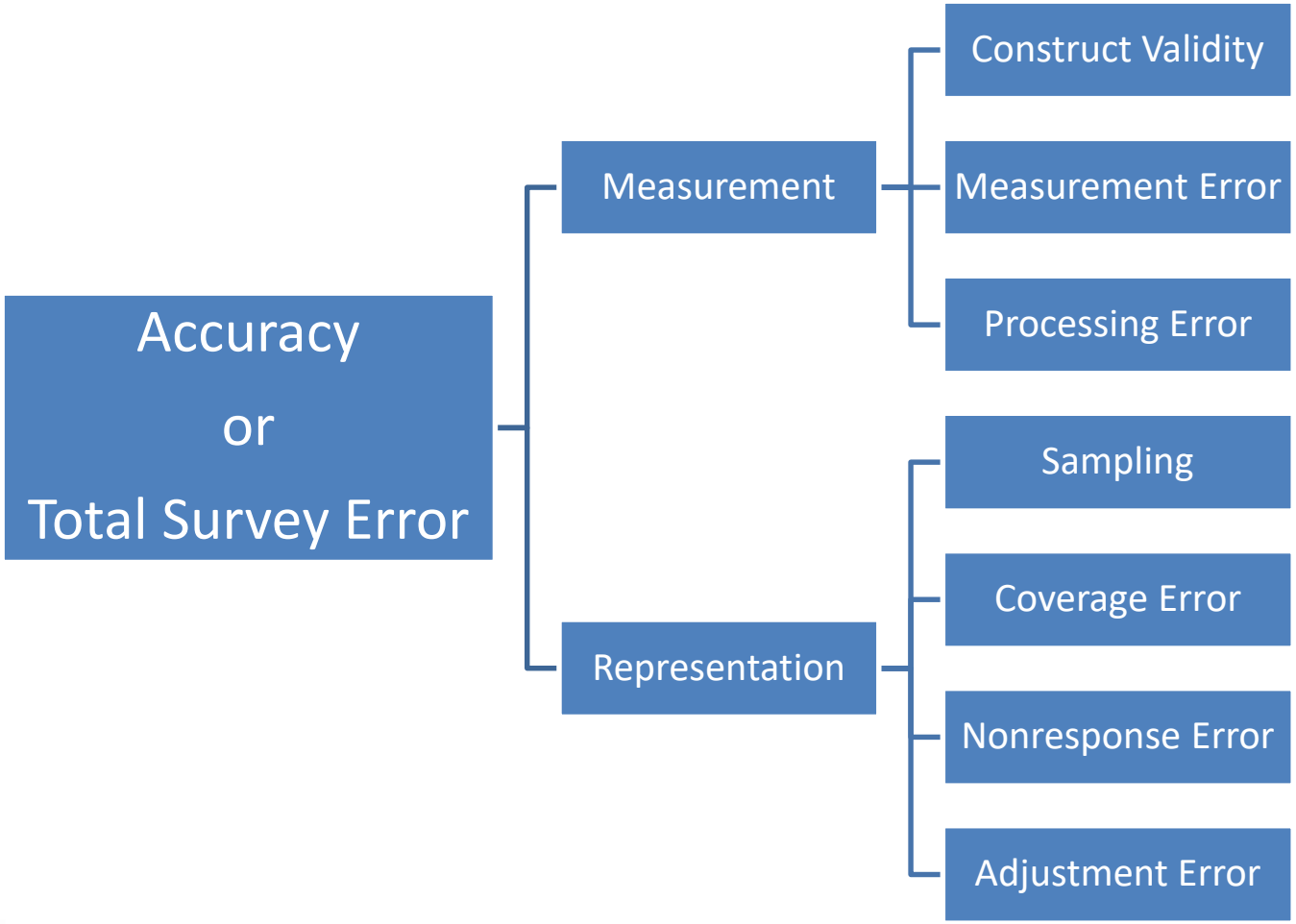
Survey Quality

Quality Dimension*	Description
Comparability	Are data source comparable to each other?
Coherence	Do the data form a coherent body of information that can be combined with other data?
Relevance	Do the data answer the questions you are asking?
Accuracy	Are the data describing what they were designed to measure?
Timeliness	How much time has elapsed since the data were collected?
Accessibility	Can user easily obtain and analyze the data?
Interpretability	Do the data make sense in terms of users' hypotheses?

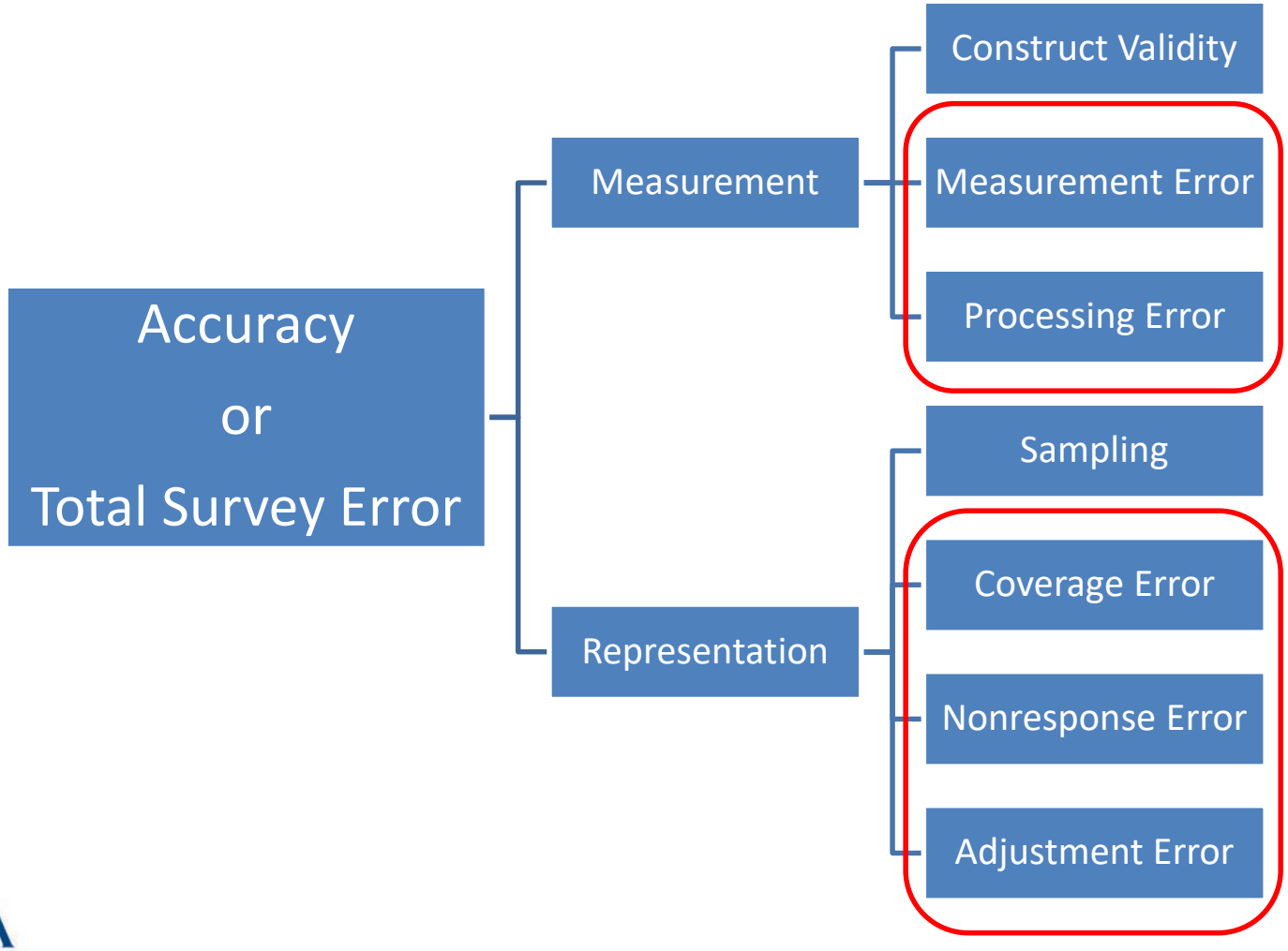
*Survey Quality by Sue Ellen Hansen, Et.al.

Fit for Use

Accuracy


Accuracy
or
Total Survey Error
Measurement
Construct Validity
Measurement Error
Processing Error
Representation
Sampling
Coverage Error
Nonresponse Error
Adjustment Error

Accuracy


Accuracy
or
Total Survey Error
Measurement
Representation
Construct Validity
Measurement Error
Processing Error
Sampling
Coverage Error
Nonresponse Error
Adjustment Error
Finding Errors
Nonresponse

Messy Data Outline

- Survey Quality
- Finding Data Errors
  - Edit
  - Analysis
- Handling Nonresponse
  - Item Nonresponse - Imputation
  - Unit Nonresponse - Reweighting

Importance of Editing and Analysis

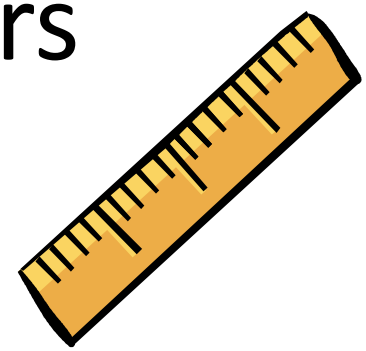
"Garbage In, Garbage Out"

Editing and analysis of survey data are important components of generating high quality indications.

Editing is Critical for quality estimates

- Must review the data
- Provide information about data quality

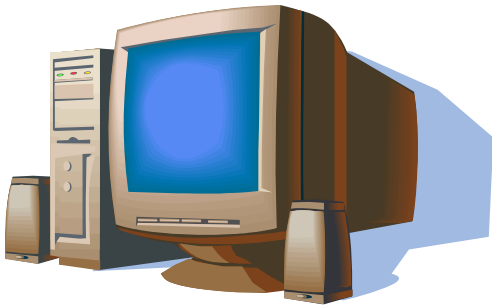
Non-Sampling Errors



- Measurement Error
  - Respondent reports incorrectly
  - Hard to understand questions
  - Memory recall
  - Unit of measures errors
  - Reference period
  - Overlooked questions

Non-Sampling Errors

- Processing
  - Data capture (key entry or OCR)
  - Coding of responses
  - Editing
  - Programming errors



What is Editing?

- Rules or Logic: Edits for items on the questionnaire
  - Univariate or Range Restrictions

    $C_1 < Y < C_2$ (number of cows between 1 and 1,000)
  - Bivariate

    $C_1 < Y_1 / Y_2 < C_2$ (calculated yield 10 and 100)
  - Balance Edits

    $Y_1 + Y_2 + Y_3 <= Y_4$ (Cows + Bulls + Calves = Total)
  - Statistical Edits

    Y > 2(SE) from the mean

How to Edit?

- Iterative:
  - Computer flagged and Manual correction, data entry correction, re-edit
- Interactive:
  - Computer assisted (Blaise, CSPro etc.)
- Influential:
  - Selective Edit, editing of only Influential or Significant records
- Automatic:
  - Programmatic fixing of errors
- Macro Editing/Analysis:
  - Across records, aggregate or distributional

NASS Editing/Analysis

- Some simple edits incorporated into the computer interviews
- Work is distributed among Regional Field Offices (RFOs) and HQ
- Done by subject matter specialists
  - Know the commodity
  - Know the sample
  - Know the questionnaire and edit
  - Know the estimators and the indications produced

NASS Editing Systems

- Designed to generate a “clean" data file
- Primarily flag records for review
  - Warnings 
  - Critical Errors 

- Large surveys logic written to fix data

Philosophy

FIX WHAT MATTERS

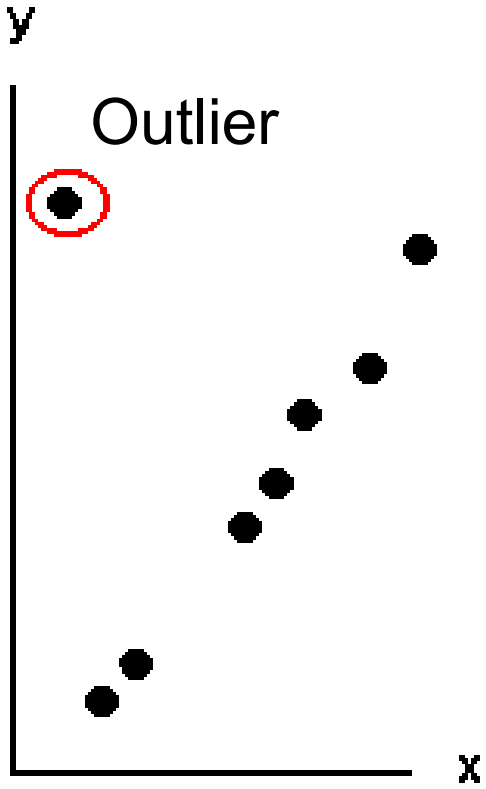
Fixing all known errors may not improve the final results. Focus on reducing large impactful errors.

What is Data Analysis?

- Data analysis is the process of reviewing survey data with analytical tools
  - To understand the current data
  - Find outliers.

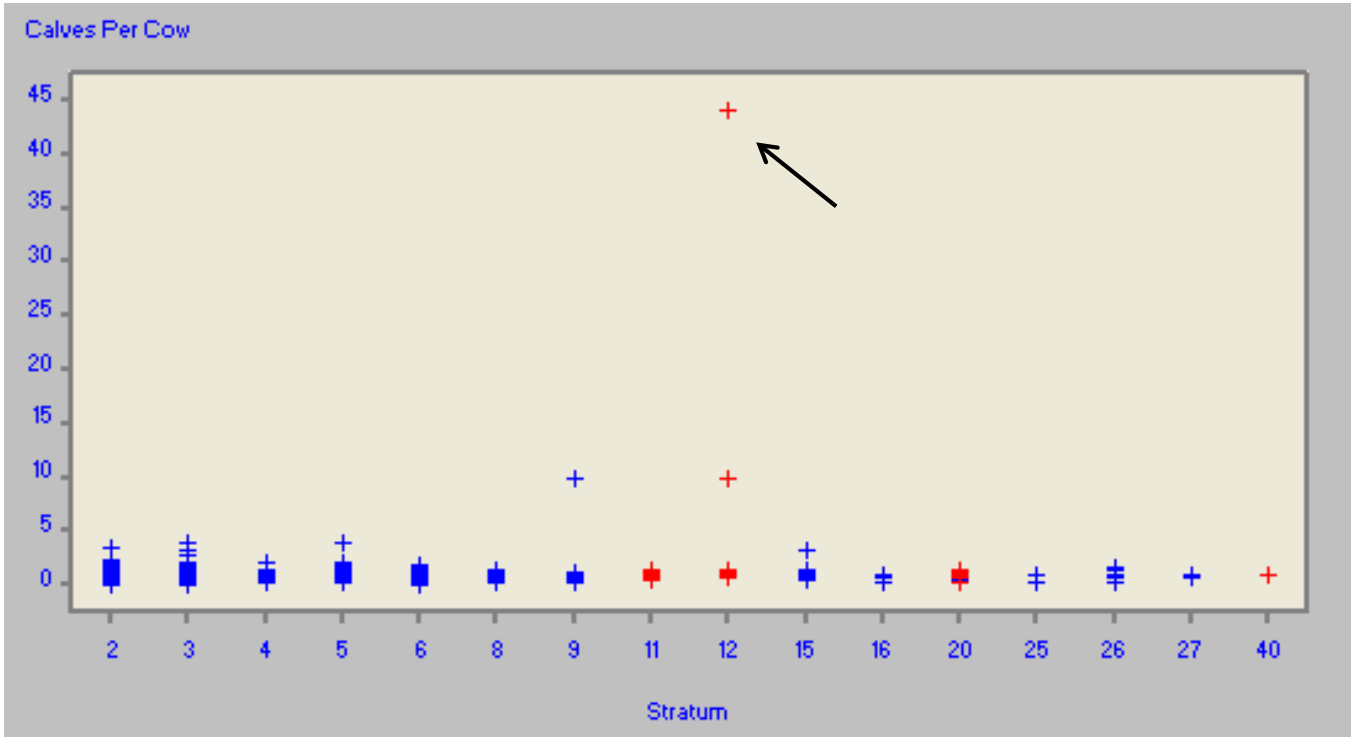
What is an outlier?

- A data value that is markedly different than the rest of the data
- An outlier may be correct
- Reasonable to expect outliers in the population

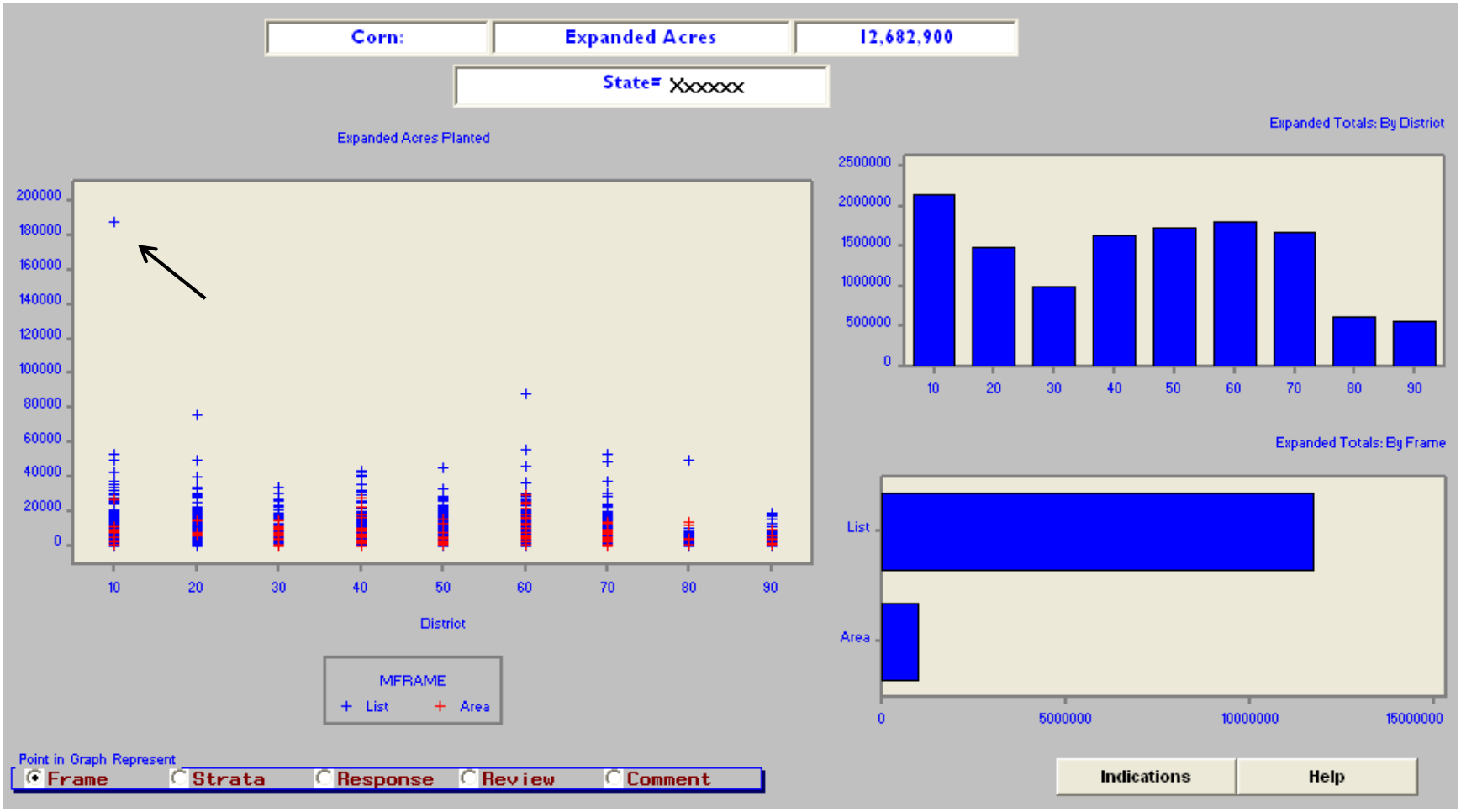


Identifying Outliers

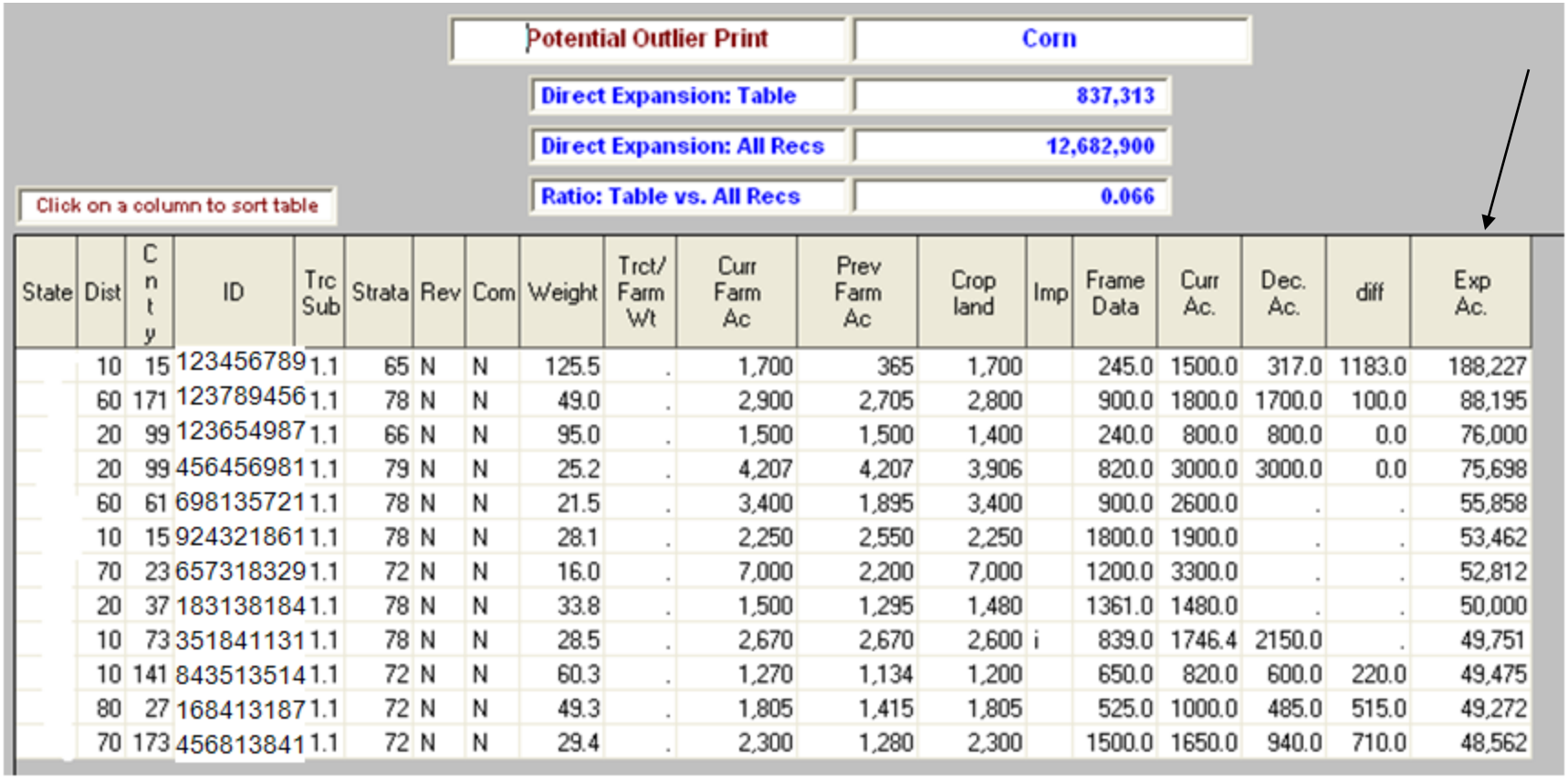
- Graphical Identification
  - Subjective
  - NASS Interactive Data Analysis System (IDAS)



Review survey data's expanded values



Review survey data's expanded values

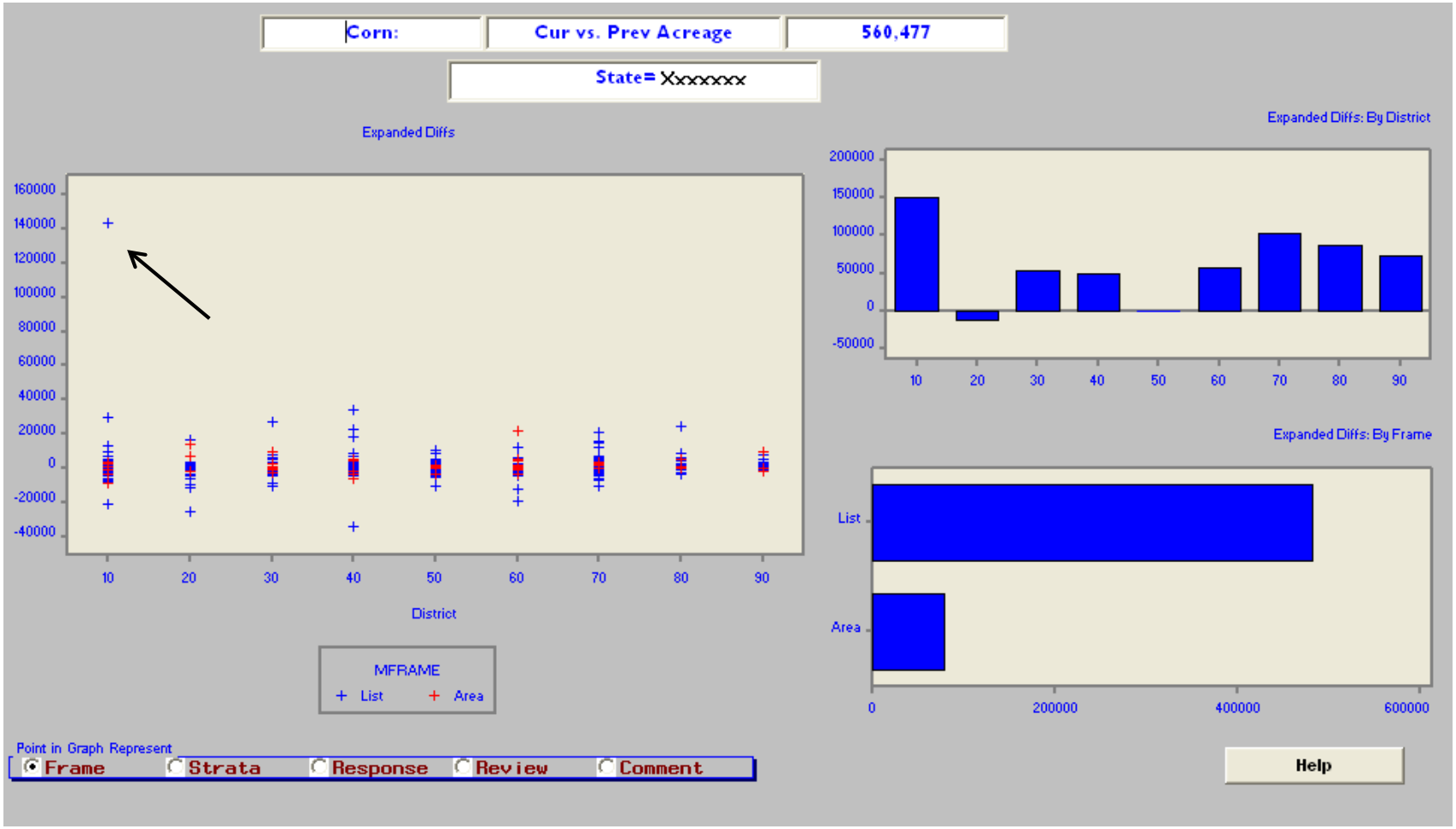


Potential Outlier Print	Corn
Direct Expansion: Table	837,313
Direct Expansion: All Recs	12,682,900
Ratio: Table vs. All Recs	0.066

Click on a column to sort table

State	Dist	Cnty	ID	Trc Sub	Strata	Rev	Com	Weight	Trct/ Farm Wt	Curr Farm Ac	Prev Farm Ac	Crop land	Imp	Frame Data	Curr Ac.	Dec. Ac.	diff	Exp Ac.
	10	15	123456789	1.1	65	N	N	125.5	.	1,700	365	1,700		245.0	1500.0	317.0	1183.0	188,227
	60	171	123789456	1.1	78	N	N	49.0	.	2,900	2,705	2,800		900.0	1800.0	1700.0	100.0	88,195
	20	99	123654987	1.1	66	N	N	95.0	.	1,500	1,500	1,400		240.0	800.0	800.0	0.0	76,000
	20	99	456456981	1.1	79	N	N	25.2	.	4,207	4,207	3,906		820.0	3000.0	3000.0	0.0	75,698
	60	61	698135721	1.1	78	N	N	21.5	.	3,400	1,895	3,400		900.0	2600.0	.	.	55,858
	10	15	924321861	1.1	78	N	N	28.1	.	2,250	2,550	2,250		1800.0	1900.0	.	.	53,462
	70	23	657318329	1.1	72	N	N	16.0	.	7,000	2,200	7,000		1200.0	3300.0	.	.	52,812
	20	37	183138184	1.1	78	N	N	33.8	.	1,500	1,295	1,480		1361.0	1480.0	.	.	50,000
	10	73	351841131	1.1	78	N	N	28.5	.	2,670	2,670	2,600	i	839.0	1746.4	2150.0	.	49,751
	10	141	843513514	1.1	72	N	N	60.3	.	1,270	1,134	1,200		650.0	820.0	600.0	220.0	49,475
	80	27	168413187	1.1	72	N	N	49.3	.	1,805	1,415	1,805		525.0	1000.0	485.0	515.0	49,272
	70	173	456813841	1.1	72	N	N	29.4	.	2,300	1,280	2,300		1500.0	1650.0	940.0	710.0	48,562

Review differences from prior survey



Review differences from prior survey

Potential Outlier Print

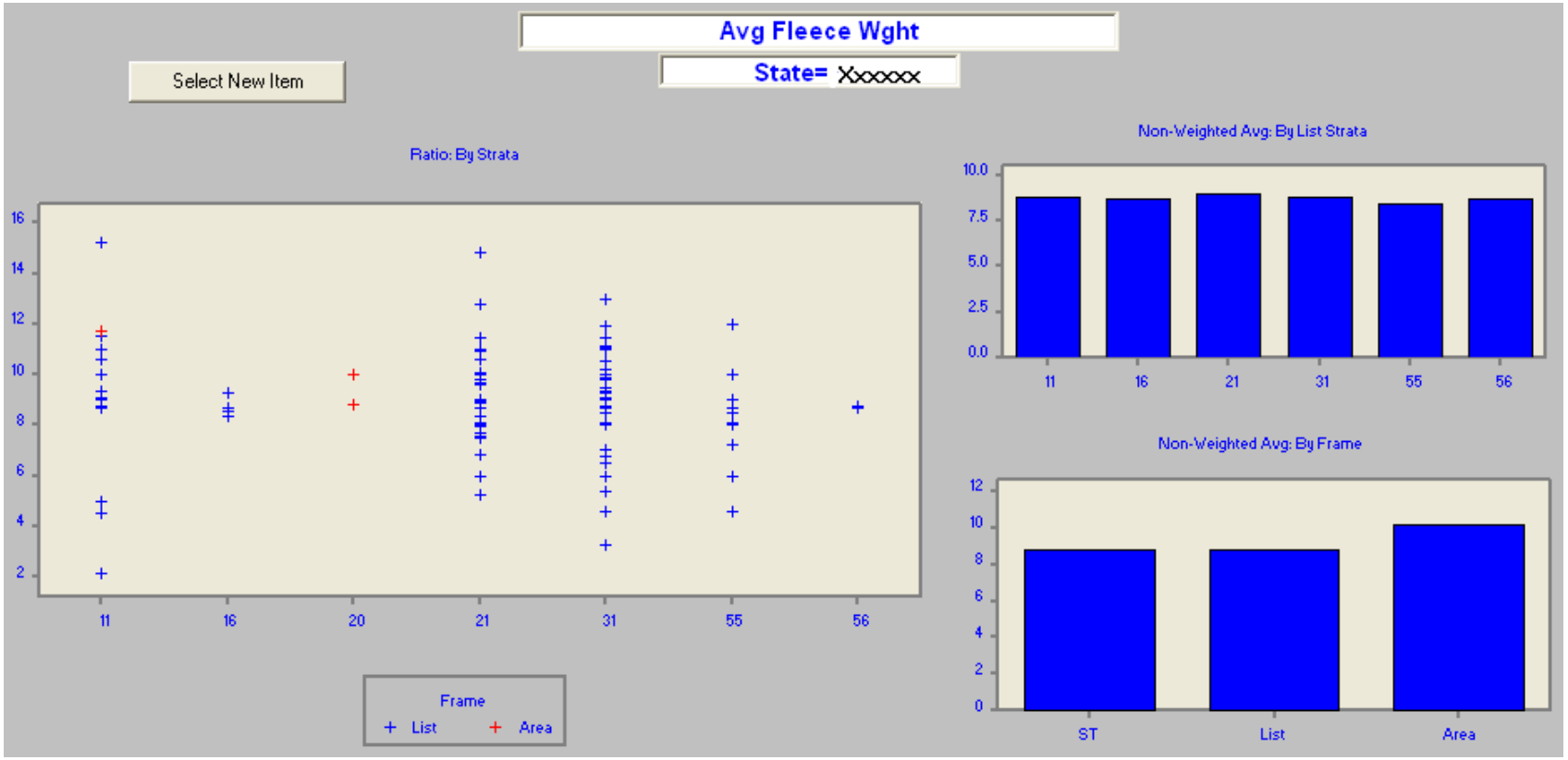
Corn

Current vs. Previous Data

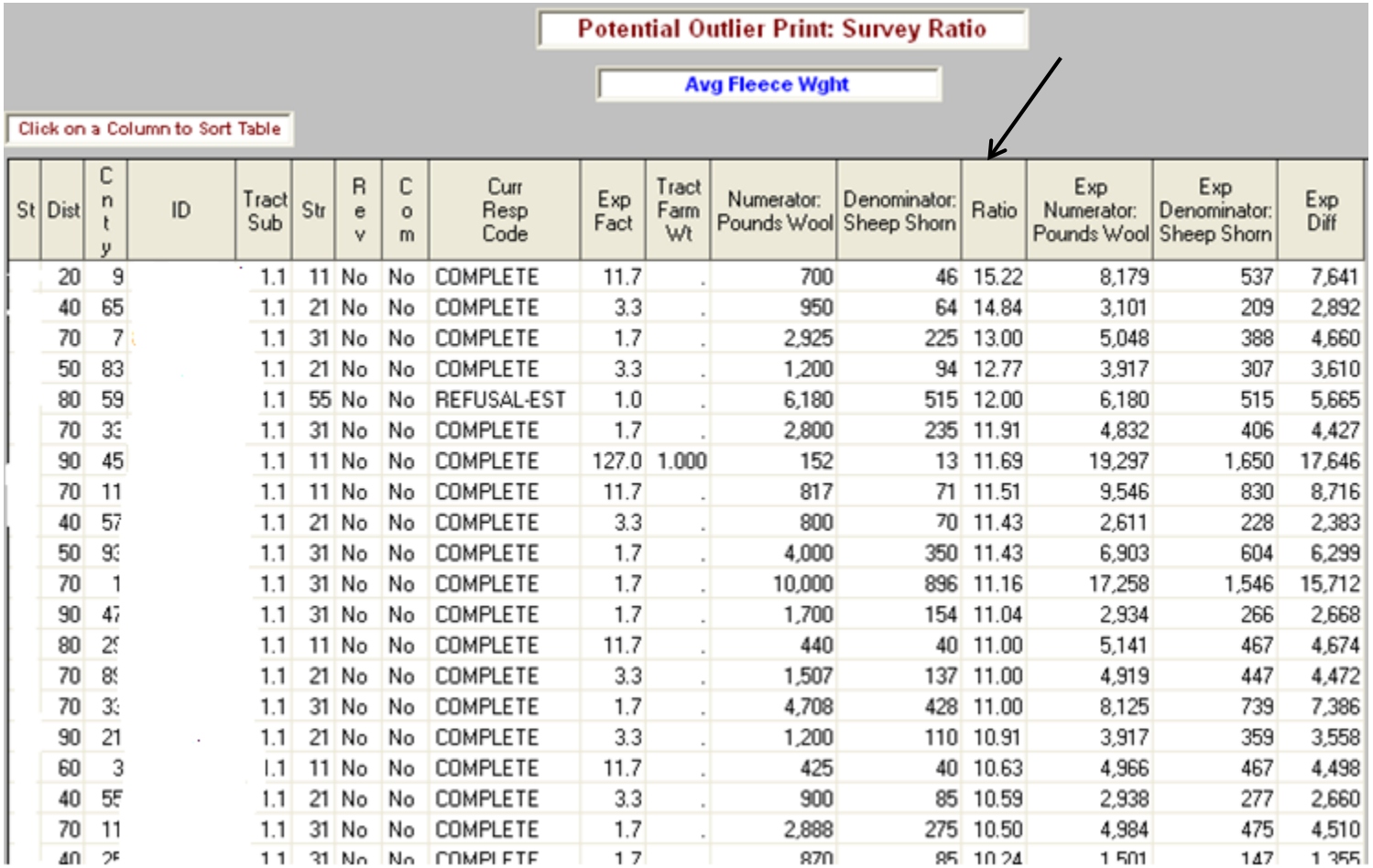
Click on a column to sort table

State	Dist	County	ID	Trc Sub	Strata	Rev	Com	Weight	Curr Farm Ac.	Dec Farm Ac.	Frame Farm Ac.	Curr	Prev	Ratio	Diff	Exp Diff
	10	15	123456789	1.1	65	N	N	120.9	1,700	365	317	1,500	317	4.732	1,183	143,059
	40	179	651581048	1.1	78	N	N	65.8	960	450	667	880	360	2.444	520	34,210
	10	201	186461843	1.1	65	N	N	77.2	800	250	225	380	0	.	380	29,351
	30	67	513843513	1.1	79	N	N	12.6	3,830	170	3,400	2,300	120	19.17	2,180	27,537
	80	27	168435135	1.1	72	N	N	47.7	1,805	1,415	850	1,000	485	2.062	515	24,580
	40	203	384614310	1.1	65	N	N	53.2	832	336	338	420	0	.	420	22,361
	60	61	841387718	1.1	12	N	N	364.8	61	61	.	61	0	.	61	22,072
	70	173	389877874	1.1	72	N	N	29.7	2,300	1,280	1,830	1,650	940	1.755	710	21,061
	60	167	534578654	1.1	78	N	N	37.7	1,000	2,020	1,033	500	1,000	0.500	-500	-18,872
	10	11	548725416	1.1	78	N	N	42.0	237	742	1,080	180	685	0.263	-505	-21,210
	20	89	198341254	1.1	72	N	N	30.5	840	840	843	0	837	0.000	-837	-25,516
	40	179	687008438	1.1	72	N	N	22.7	1,500	1,506	1,399	0	1,506	0.000	-1,506	-34,159

Review data ratios within current survey



Review data ratios within current survey



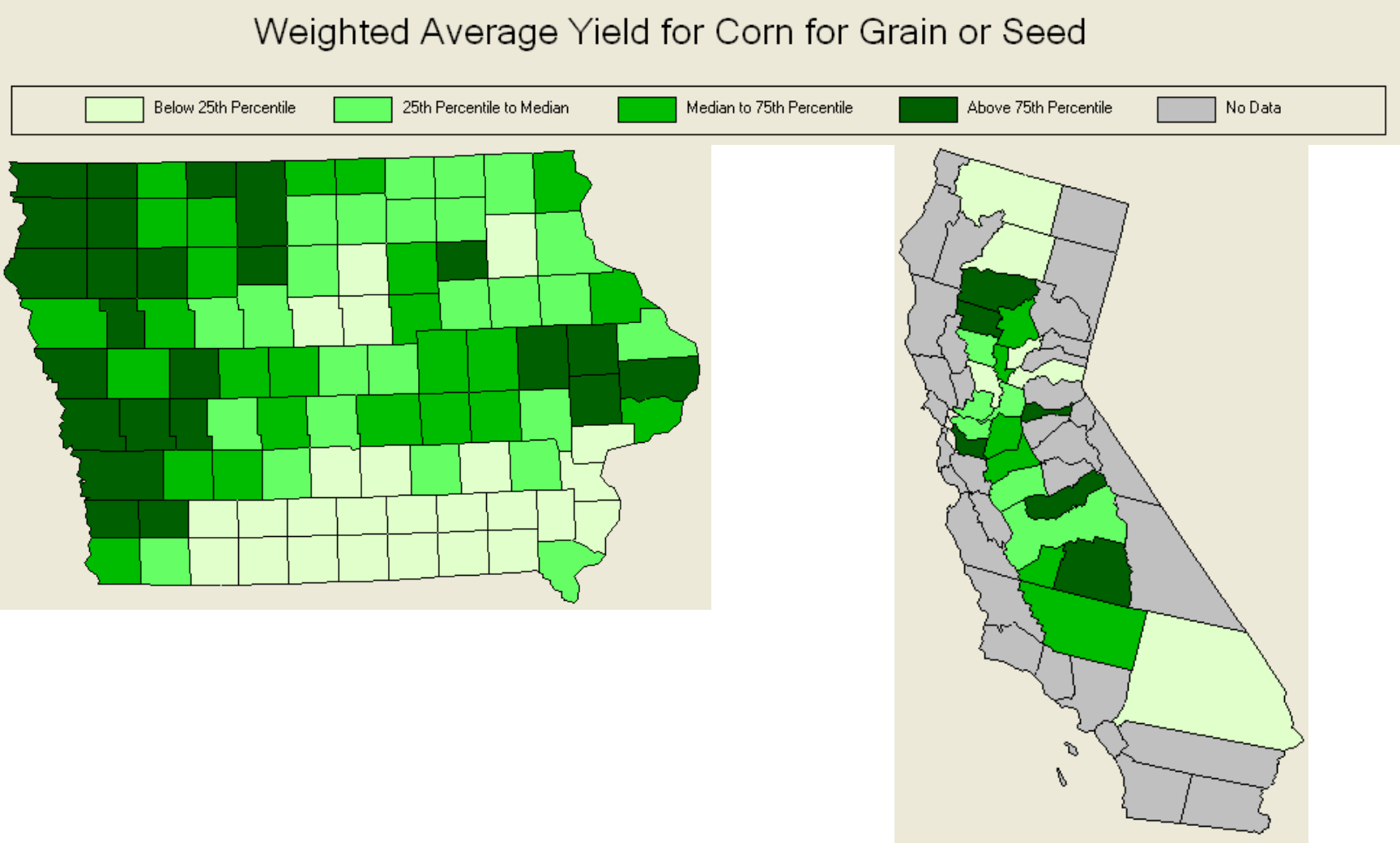
Potential Outlier Print: Survey Ratio

Avg Fleece Wght

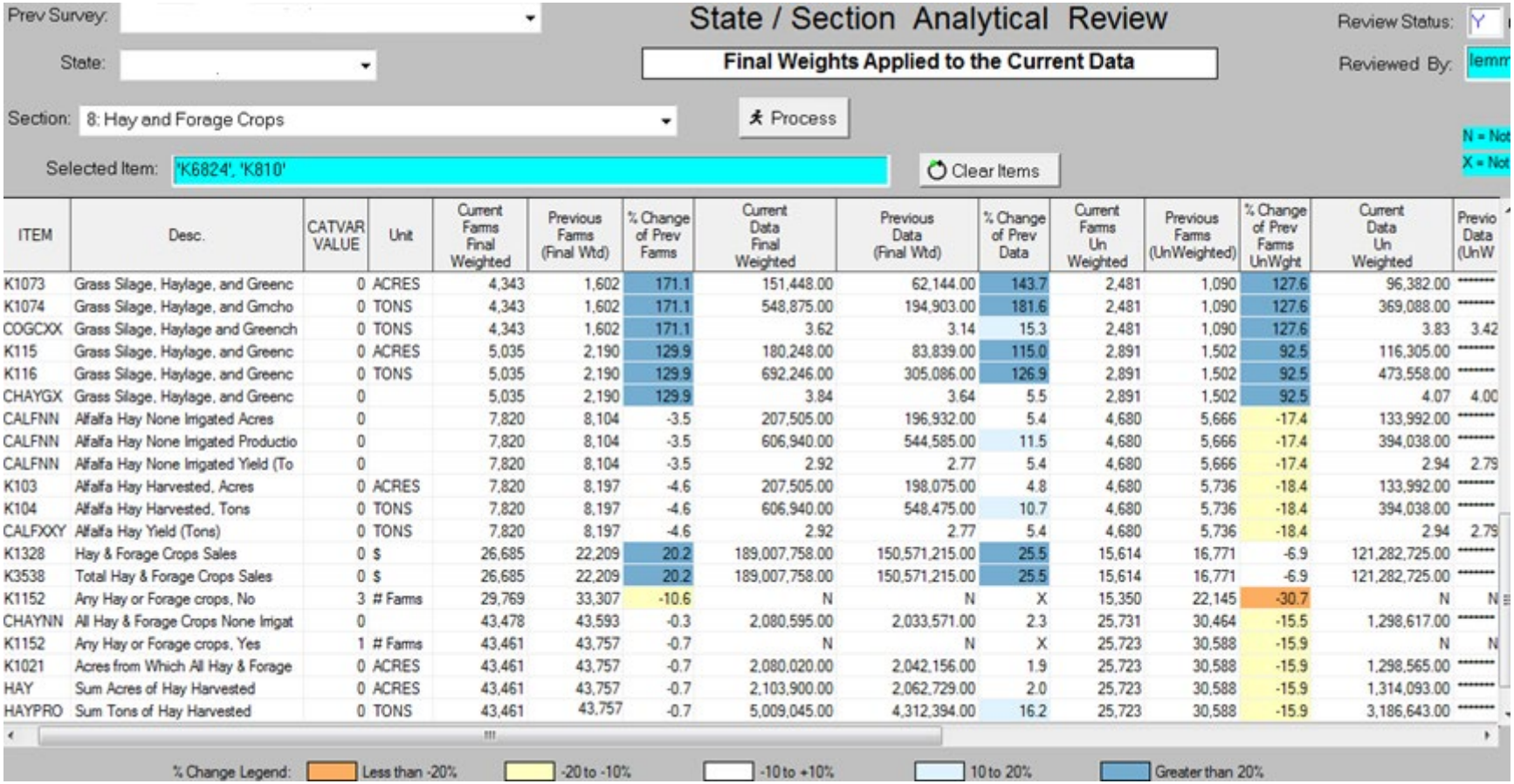
Click on a Column to Sort Table

St	Dist	Cnty	ID	Tract Sub	Str	Rev	Com	Curr Resp Code	Exp Fact	Tract Farm Wt	Numerator: Pounds Wool	Denominator: Sheep Shorn	Ratio	Exp Numerator: Pounds Wool	Exp Denominator: Sheep Shorn	Exp Diff
	20	9		1.1	11	No	No	COMPLETE	11.7	.	700	46	15.22	8,179	537	7,641
	40	65		1.1	21	No	No	COMPLETE	3.3	.	950	64	14.84	3,101	209	2,892
	70	7		1.1	31	No	No	COMPLETE	1.7	.	2,925	225	13.00	5,048	388	4,660
	50	83		1.1	21	No	No	COMPLETE	3.3	.	1,200	94	12.77	3,917	307	3,610
	80	59		1.1	55	No	No	REFUSAL-EST	1.0	.	6,180	515	12.00	6,180	515	5,665
	70	33		1.1	31	No	No	COMPLETE	1.7	.	2,800	235	11.91	4,832	406	4,427
	90	45		1.1	11	No	No	COMPLETE	127.0	1.000	152	13	11.69	19,297	1,650	17,646
	70	11		1.1	11	No	No	COMPLETE	11.7	.	817	71	11.51	9,546	830	8,716
	40	57		1.1	21	No	No	COMPLETE	3.3	.	800	70	11.43	2,611	228	2,383
	50	93		1.1	31	No	No	COMPLETE	1.7	.	4,000	350	11.43	6,903	604	6,299
	70	1		1.1	31	No	No	COMPLETE	1.7	.	10,000	896	11.16	17,258	1,546	15,712
	90	47		1.1	31	No	No	COMPLETE	1.7	.	1,700	154	11.04	2,934	266	2,668
	80	29		1.1	11	No	No	COMPLETE	11.7	.	440	40	11.00	5,141	467	4,674
	70	89		1.1	21	No	No	COMPLETE	3.3	.	1,507	137	11.00	4,919	447	4,472
	70	33		1.1	31	No	No	COMPLETE	1.7	.	4,708	428	11.00	8,125	739	7,386
	90	21		1.1	21	No	No	COMPLETE	3.3	.	1,200	110	10.91	3,917	359	3,558
	60	3		1.1	11	No	No	COMPLETE	11.7	.	425	40	10.63	4,966	467	4,498
	40	55		1.1	21	No	No	COMPLETE	3.3	.	900	85	10.59	2,938	277	2,660
	70	11		1.1	31	No	No	COMPLETE	1.7	.	2,888	275	10.50	4,984	475	4,510
	40	25		1.1	31	No	No	COMPLETE	1.7	.	870	85	10.24	1,501	147	1,355

Review data geographically



Aggregates Review



Prev Survey:

State:

State / Section Analytical Review

Final Weights Applied to the Current Data

Review Status: Y

Reviewed By: lemm

Section: 8: Hay and Forage Crops

Process

N = Not

X = Not

Selected Item: 'K6824', 'K810'

Clear Items

ITEM	Desc.	CATVAR VALUE	Unit	Current Farms Final Weighted	Previous Farms (Final Wtd)	% Change of Prev Farms	Current Data Final Weighted	Previous Data (Final Wtd)	% Change of Prev Data	Current Farms Un Weighted	Previous Farms (UnWeighted)	% Change of Prev Farms UnWght	Current Data Un Weighted	Previo Data (UnW
K1073	Grass Silage, Haylage, and Greenc	0	ACRES	4,343	1,602	171.1	151,448.00	62,144.00	143.7	2,481	1,090	127.6	96,382.00	********
K1074	Grass Silage, Haylage, and Gmcho	0	TONS	4,343	1,602	171.1	548,875.00	194,903.00	181.6	2,481	1,090	127.6	369,088.00	********
COGCXX	Grass Silage, Haylage and Greench	0	TONS	4,343	1,602	171.1	3.62	3.14	15.3	2,481	1,090	127.6	3.83	3.42
K115	Grass Silage, Haylage, and Greenc	0	ACRES	5,035	2,190	129.9	180,248.00	83,839.00	115.0	2,891	1,502	92.5	116,305.00	********
K116	Grass Silage, Haylage, and Greenc	0	TONS	5,035	2,190	129.9	692,246.00	305,086.00	126.9	2,891	1,502	92.5	473,558.00	********
CHAYGX	Grass Silage, Haylage, and Greenc	0		5,035	2,190	129.9	3.84	3.64	5.5	2,891	1,502	92.5	4.07	4.00
CALFNN	Alfalfa Hay None Irrigated Acres	0		7,820	8,104	-3.5	207,505.00	196,932.00	5.4	4,680	5,666	-17.4	133,992.00	********
CALFNN	Alfalfa Hay None Irrigated Productio	0		7,820	8,104	-3.5	606,940.00	544,585.00	11.5	4,680	5,666	-17.4	394,038.00	********
CALFNN	Alfalfa Hay None Irrigated Yield (To	0		7,820	8,104	-3.5	2.92	2.77	5.4	4,680	5,666	-17.4	2.94	2.79
K103	Alfalfa Hay Harvested, Acres	0	ACRES	7,820	8,197	-4.6	207,505.00	198,075.00	4.8	4,680	5,736	-18.4	133,992.00	********
K104	Alfalfa Hay Harvested, Tons	0	TONS	7,820	8,197	-4.6	606,940.00	548,475.00	10.7	4,680	5,736	-18.4	394,038.00	********
CALFXXY	Alfalfa Hay Yield (Tons)	0	TONS	7,820	8,197	-4.6	2.92	2.77	5.4	4,680	5,736	-18.4	2.94	2.79
K1328	Hay & Forage Crops Sales	0	$	26,685	22,209	20.2	189,007,758.00	150,571,215.00	25.5	15,614	16,771	-6.9	121,282,725.00	********
K3538	Total Hay & Forage Crops Sales	0	$	26,685	22,209	20.2	189,007,758.00	150,571,215.00	25.5	15,614	16,771	-6.9	121,282,725.00	********
K1152	Any Hay or Forage crops, No	3	# Farms	29,769	33,307	-10.6	N	N	X	15,350	22,145	-30.7	N	N
CHAYNN	All Hay & Forage Crops None Irrigat	0		43,478	43,593	-0.3	2,080,595.00	2,033,571.00	2.3	25,731	30,464	-15.5	1,298,617.00	********
K1152	Any Hay or Forage crops, Yes	1	# Farms	43,461	43,757	-0.7	N	N	X	25,723	30,588	-15.9	N	N
K1021	Acres from Which All Hay & Forage	0	ACRES	43,461	43,757	-0.7	2,080,020.00	2,042,156.00	1.9	25,723	30,588	-15.9	1,298,565.00	********
HAY	Sum Acres of Hay Harvested	0	ACRES	43,461	43,757	-0.7	2,103,900.00	2,062,729.00	2.0	25,723	30,588	-15.9	1,314,093.00	********
HAYPRO	Sum Tons of Hay Harvested	0	TONS	43,461	43,757	-0.7	5,009,045.00	4,312,394.00	16.2	25,723	30,588	-15.9	3,186,643.00	********

% Change Legend: Less than -20% | -20 to -10% | -10 to +10% | 10 to 20% | Greater than 20%

What to do with an outlier? Verify the Reported Data

- Verify the Reported Data
  - If an error is found: correct it!
- Otherwise
  - Adjust weights
  - Remove from models
  - Adjust estimates

Impacts of Outliers

- Survey Indications
  - In what direction
  - To what degree
- Measures of Precision
  - Standard Error (SE)
  - Coefficient of Variation (CV)
- Nonresponse Adjustment

Messy Data Outline

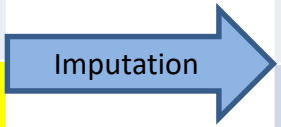
- Survey Quality
- Finding Data Errors
  - Edit
  - Analysis
- Handling Nonresponse
  - Item Nonresponse - Imputation
  - Unit Nonresponse – Imputation or Reweighting

What is Item Imputation?

The process of replacing missing data with substituted values.

BEFORE
Clean dataset with missing data

ID	Variable 1	Variable 2
1	10	33
2	?	74
3	25	?
4	15	?



AFTER
Clean dataset with imputed values

ID	Variable 1	Variable 2
1	10	33
2	27	74
3	25	70
4	15	52

Why is there missing data?

Refusal to answer the item in question

- Too personal
- Too sensitive

Too difficult to answer

- Poor memory or inadequate records
- Too difficult to calculate

Accidentally skipped

Other unknown reasons?

Common Item Imputation Techniques

- Manual
- Means
- Ratio
- Hot Deck/Cold Deck
- Multivariate

Manual Imputation

Replacing missing data with external information or historical data

- May be used when data are known at least approximately.
- Generally a simple process but not always statistically defensible.
- May be the easiest way to estimate extreme operators

Common Item Imputation Techniques

- Manual
- Means
- Ratio
- Hot Deck/Cold Deck
- Multivariate

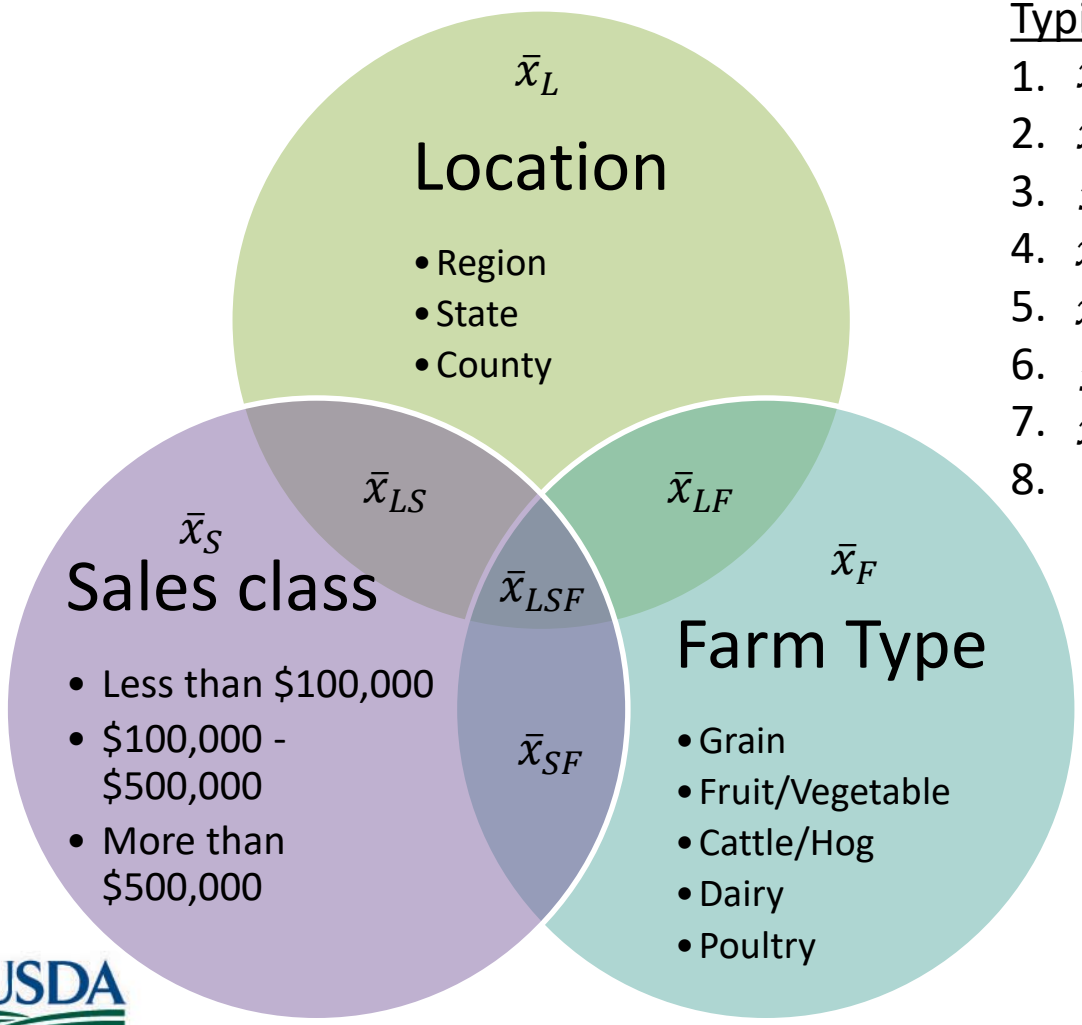
Mean Imputation

Replacing missing data with the mean of clean reported data

- Un-weighted means is the most common
- Best practice is to group the records with similar attributes

Mean Imputation Grouping

Which $\bar{x}$ to use?!



Typical Grouping Hierarchy

1. $\bar{x}_{LSF}$ Location, Sales Class, Farm Type
2. $\bar{x}_{SF}$ Sales Class, Farm Type
3. $\bar{x}_{LS}$ Location, Sales Class
4. $\bar{x}_{LF}$ Location, Farm Type
5. $\bar{x}_F$ Farm Type
6. $\bar{x}_S$ Sales Class
7. $\bar{x}_L$ Location
8. $\bar{x}$ National

- Picking the best $\bar{x}$ often depends on if enough records exist.
- Combining subgroups into broader categories is an option to get enough records

Mean Imputation

An Example:

Grain Farms with **less than $10,000 sales** in **Western Region**						
	Farm 1	Farm 2	Farm 3	Farm 4	Farm 5	Farm 6
Taxes	$10	$15	?	$27	$33	$20
Expenses	$89	$74	$13	?	$36	$100
Wages	?	$50	$44	$150	$102	$170

Mean Taxes = $21

$$\frac{10+15+27+33+20}{5}$$

Mean Expenses = $62

$$\frac{89+74+13+36+100}{5}$$

Mean Wages = $103

$$\frac{50+44+150+102+170}{5}$$

Grain Farms with **less than $10,000 sales** in **Western Region**						
	Farm 1	Farm 2	Farm 3	Farm 4	Farm 5	Farm 6
Taxes	$10	$15	$21	$27	$33	$20
Expenses	$89	$74	$13	$62	$36	$100
Wages	$103	$50	$44	$150	$102	$170

Common Item Imputation Techniques

- Manual
- Means
- Ratio
- Hot Deck/Cold Deck
- Multivariate

Ratio Imputation

Replacing missing data with values calculated from ratio of data from different reports

- Used in monthly surveys using a ratio of current to previous month
- Assumes similar relationship among different operations.

Ratio Imputation

Monthly Survey Results						
	Current Month			**Previous Month**		
	Farm 1	Farm 2	Farm 3	Farm 1	Farm 2	Farm 3
Production	?	120	190	130	110	170
Yield	40	?	50	60	45	55

Production Ratio

$\frac{120+190}{110+170} = 1.10$

Yield Ratio

$\frac{40+50}{60+55} = 0.78$

Apply Ratios to Previous Reports

Farm 1 Production

$130 * 1.10 = 143$

Farm 2 Yield

$45 * 0.78 = 35.1$

Monthly Survey Results						
	Current Month			**Previous Month**		
	Farm 1	Farm 2	Farm 3	Farm 1	Farm 2	Farm 3
Production	143	120	190	130	110	170
Yield	40	35.1	50	60	45	55

Mean & Ratio Imputation

Advantages/Disadvantages

Advantages	Disadvantages
Easy to implement	Artificially lowers variance
Easy to debug	More statistically sound methods available
Flexible	One record can really drive imputation
Creates imputations within edit limits	

Common Item Imputation Techniques

- Manual
- Means
- Ratio
- Hot Deck/Cold Deck
- Multivariate

Hot Deck / Cold Deck

Nearest Neighbor Selection

Hot Deck Imputation

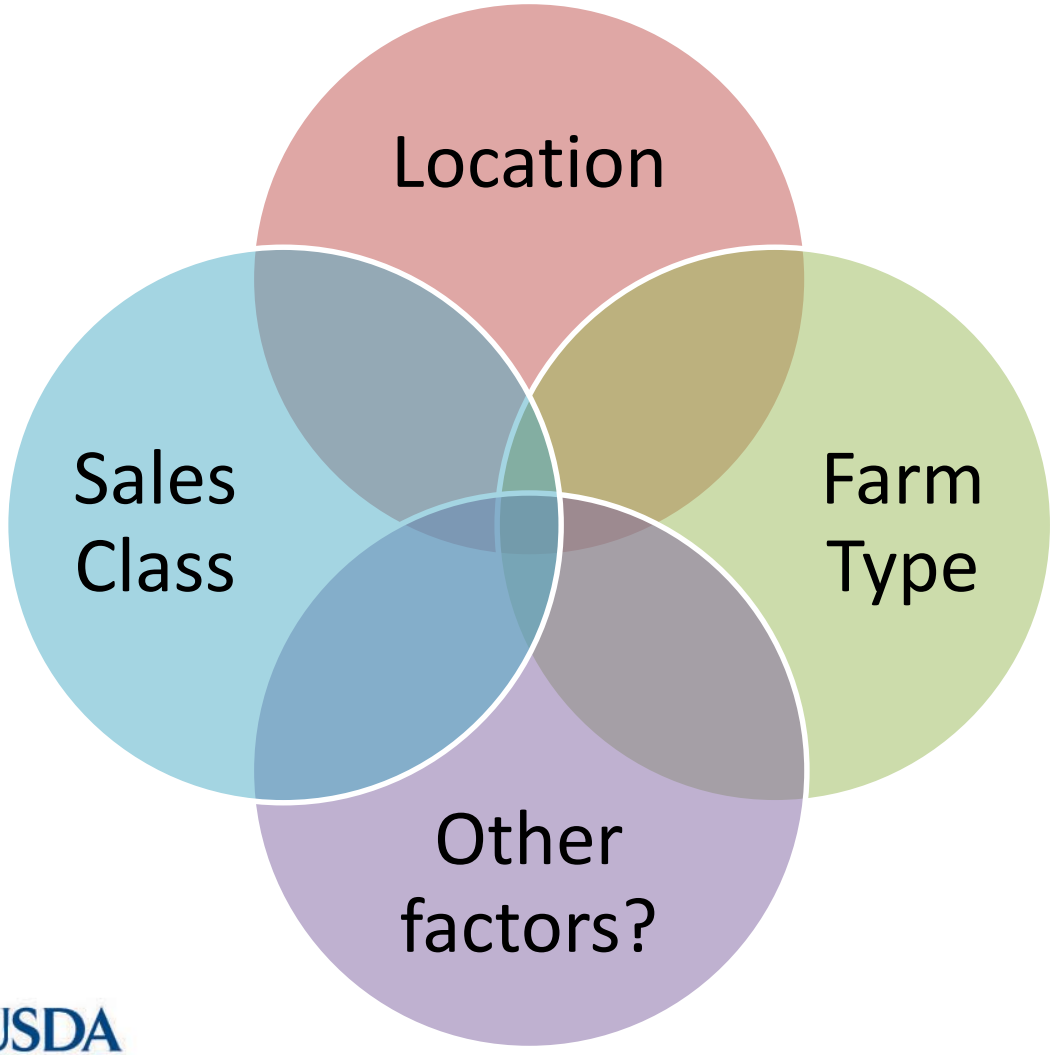
Replacing missing data with values from a similar record in the *same* dataset

Cold Deck Imputation

Replacing missing data with values from a similar record in a *different* dataset

Hot Deck/Cold Deck

Selecting a “similar” record from a donor pool



DONOR POOL is a group of complete records that have similar characteristics as the record requiring imputation.

Different algorithms (like Nearest Neighbor) can be used to find a similar record.

Different variables can potentially use different scoring algorithms.

Hot Deck Imputation

An Example:

Grain Farms with **less than $10,000 sales** in **Western Region**						
	Farm 1	Farm 2	Farm 3	Farm 4	Farm 5	Farm 6
Taxes	$10	$15	?	$27	$33	$20
Expenses	$89	$74	$13	?	$36	$100
Wages	?	$50	$44	$150	$102	$170

Farm 1 similar to Farm 2
- Use $50 wages from Farm 2 in Farm 1

Farm 3 similar to Farm 5
- Use $33 taxes from Farm 5 in Farm 3

Farm 4 similar to Farm 6
- Use $100 expense from Farm 5 in Farm 4

Grain Farms with **less than $10,000 sales** in **Western Region**						
	Farm 1	Farm 2	Farm 3	Farm 4	Farm 5	Farm 6
Taxes	$10	$15	$33	$27	$33	$20
Expenses	$89	$74	$13	$100	$36	$100
Wages	$50	$50	$44	$150	$102	$170

Common Item Imputation Techniques

- Manual
- Means
- Ratio
- Hot Deck/Cold Deck
- Multivariate

Multivariate Imputation

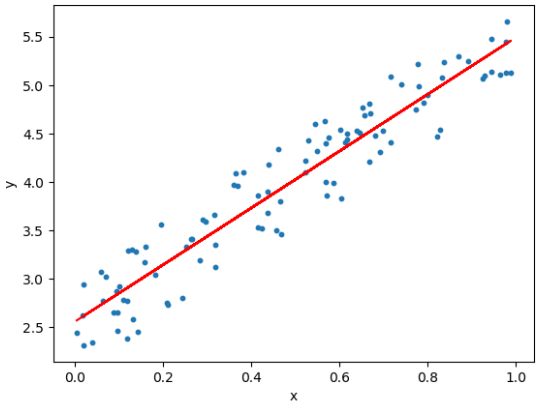
Replacing missing data with values calculated from regression models

- Typically uses linear regression to fit data to missing values
- Uses both complete and incomplete cases to help predict the missing values.

The Basic Form of Multivariate Imputation

TYPEFARM
CROPLAND
OWNED AC
RENTED AC
.
.
.
.
.
.
.
LAND VALUE
ASSETS

Imputed variables and covariates



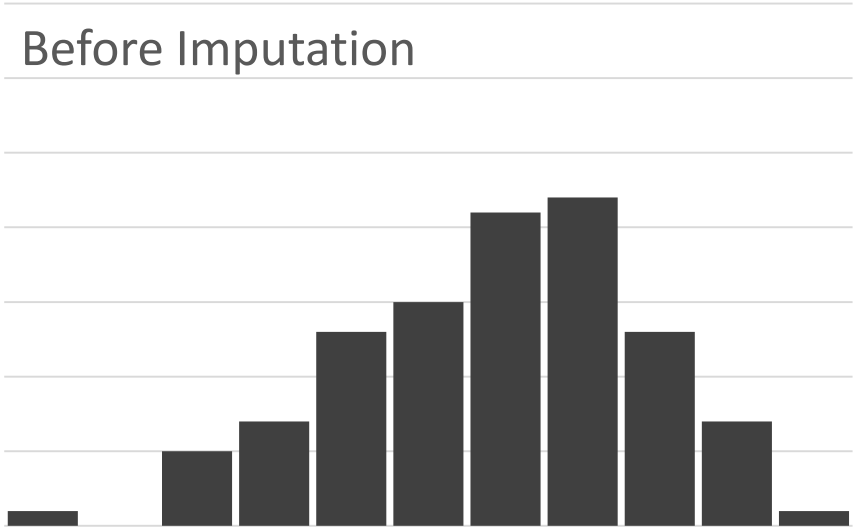
Iterative Sequential Regression

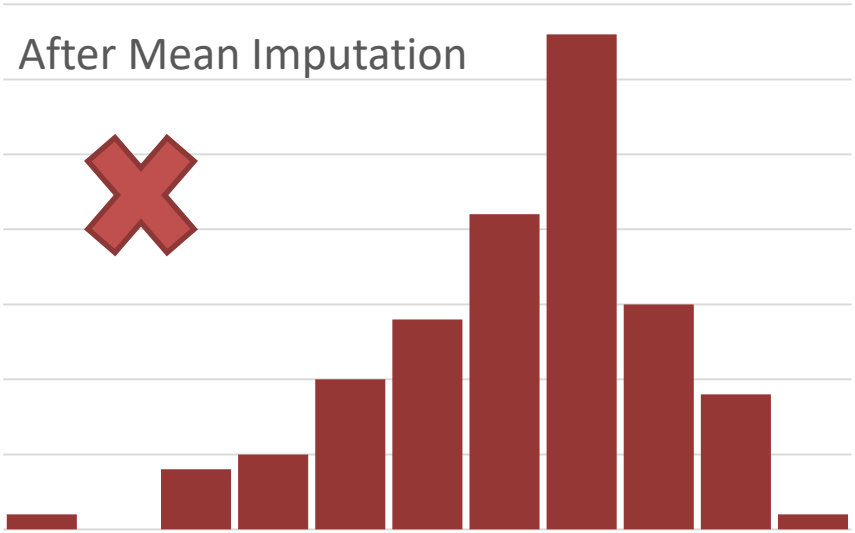
- Each variable run thru unique regression model

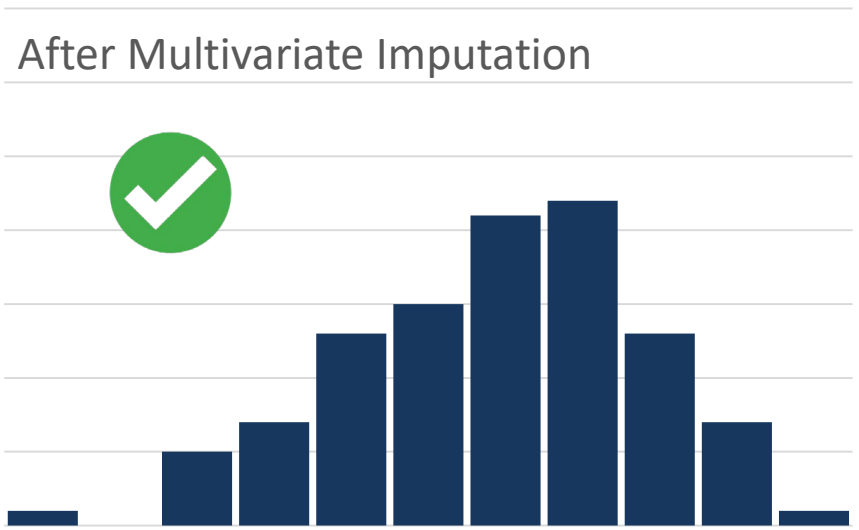
PRODUCTION
EXPENSES
ASSETS
.
.
.
FUEL COST
TAXES
YIELD
.
.
.

Imputed variables

Why Multivariate Imputation?







Unit Nonresponse

Primarily Refusal and Inaccessibles

- Can be done by Imputation (whole record)
- Commonly done by reweighting

Reweighting

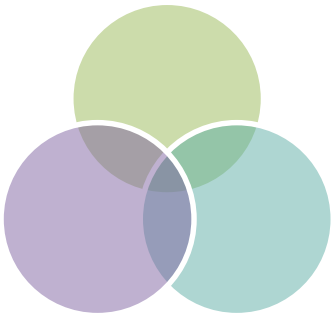
A very simple example:

- If we sent 100 questionnaires where 90 were completed and 10 were nonrespondents then the nonresponse weight = 100/90 = 1.11.
- To summarize the data multiply all the record's data by 1.11.

$$Survey\ Estimate = \sum_{i=1}^{\substack{completed \\ reports}} nonresponse\ weight_i * item\_data_i$$

In Production:

- Group records into homogeneous groups since nonresponse can be dependent on different attributes.
- Adjusting weights of good reports.



Complex Reweighting

$$DE = \sum \frac{N}{n} A_c \frac{n}{n_a + n_{na}} \bullet \frac{n_a}{n_a + n_{ah}} \bullet \frac{n_h}{n_{gh}} y_{gh}$$

N = number of units in the population
n = number of units in the sample
A_c = post stratification weight for post strata
n_a = number of known ag operations in the sample
n_{na} = number of known non-ag operations in the sample
n_h = number of known commodity operations in the sample
n_{ah} = number of known non-commodity ag operations in the sample
n_{gh} = number of positive responding commodity operations in the sample
y_{gh} = value of the positive responding commodity operation

*Business Status
*Presence of item

Calibration

A re-weighting algorithm that minimizes the change in the sampling weights so that several important weighted survey items match official published totals. (Bench marking)

- Input weights to the calibration routine are the sampling weights
- Unit non-response adjustment can be done prior to calibration or incorporated.
- Calibration helps correct for any disproportionate response from a particular farm type or sales class

Census of Agriculture Weights

- Composed of three adjustments
  - Nonresponse (nr)
  - Misclassification (m)
  - Coverage (c)
- Integerized
- For COA, max weight is 6

$$w_i = nr_i m_i c_i$$

Fully adjusted weight

Questions?