An Analysis Tool for Reviewing Farm Economic Data Using SAS/AF®

Van Johnson, National Agricultural Statistics Service, Washington, DC

ABSTRACT

The well-being of farming is of great interest to many groups. There is a great need, therefore, for accurate and timely data on farm economics. The primary source of these data is the Agricultural Resource Management Survey (ARMS), conducted by the U.S. Department of Agriculture (USDA) National Agricultural Statistics Service (NASS). To ensure that data on the farm enterprise are accurate, raw data from the survey must be reviewed and corrected if necessary before it is summarized. The complex interrelationships between factors, inherent in farm economic data, provide a challenge to this review process.

This paper provides an overview of the structure and use of an Interactive Data Analysis System (IDAS) tool designed for reviewing ARMS data during and after data collection. The IDAS tool makes extensive use of graphics and numerical descriptions to depict the relationships between factors in the farm economic data. The application is designed for use by other disciplines as well as statistics; therefore, the user does not need an extensive background in statistics to be able to use the tool effectively.

INTRODUCTION

ARMS is an annual survey cosponsored by NASS and the Economic Research Service (ERS), both agencies of USDA. ERS provides specifications on data requirements and guidance for editing and presummary data analysis. NASS is responsible for the overall sample design, data collection procedures, and data editing and summarizing. After summarization, the data set is turned over to ERS for additional economic analyses.

Data collected through ARMS provide the only national perspective on the economic well-being of the farm sector. ARMS data cover the entire spectrum of farm economic issues, including resources, costs, and financial conditions. Specifically, the ARMS is used to do the following:

1. Gather information about the relationships among agricultural production, resources, and the environment
2. Help determine farm/ranch operation and operator characteristics, including off-farm income
3. Determine costs of production for various crop and livestock commodities
4. Help determine farm income and provide information on other farm financial measurements such as measurements for assets and debt
5. Assess agricultural chemical usage and collect information on related management practices

The farm economic data obtained through ARMS are used at several aggregate levels. The numbers in parentheses are the current number of levels of aggregation used in data analyses. Data may be aggregated by geopolitical boundary (3), operation type (2), farm type (16), or sales class (7). These levels of aggregation may be nested. For example, a data user may want to see fuel expenses by farm type for the United States, or cattle inventory by sales class for all states in a region within the United States. The data set also contains substate identifiers such as district and county to allow for more levels of aggregation where there are enough samples to make the summary statistics from that aggregate level meaningful.

ARMS has three data collection phases: (1) screening to identify in-business farm operations, (2) collecting crop production practices and chemical usage information, and (3) collecting farm finance data. This paper will focus on the analysis of data for the third phase, collecting farm finance data. The target population for the ARMS Phase III is the official USDA farm population and is defined as ìall establishments except institutional farms that sold or would normally have sold at least $1,000 of agricultural products during the year.î The survey is a multiple frame survey consisting of a list and a complementary area frame. In 2002, nearly 19,000 farm operations were sampled in the United States. Sample sizes in the various states ranged from 30 to 1,150. Statisticians in each state were responsible for providing a clean, edited data set to NASS headquarters.

SCOPE

The structure of any data analysis system should be determined by how the survey is designed and how the data will be used. In this paper, I will discuss the issues that were considered in the development and implementation of the data analysis system for ARMS as they relate to the areas of survey design and data usage. Under Survey Design Issues, I will look at the population of interest, the sample design, data collection, data editing, and summarization. I will conclude the section on survey design issues by looking at the impacts of the survey design on data analysis. Under Data Usage Issues, I will look at data requirements and the analyses needed to ensure that these requirements are met. I will conclude with a demonstration of some of the analysis capabilities of IDAS.

SURVEY DESIGN ISSUES

The survey must be designed to provide data to answer questions related to agricultural policy and farm economics. Agricultural policy is usually targeted to widely diverse areas of agricultural production. ARMS must be designed to obtain data on characteristics of farm operations that cover a broad range of values throughout the United States. The survey must be designed to ensure that there are sufficient data to provide meaningful statistics across all levels of aggregation. There are 16 farm types defined by NASS. All 16 types must

be represented in the sample. The sampling design must take into account that farms are not necessarily evenly spread out across the United States, nor are the farm types evenly distributed throughout the population of interest.

Population of Interest
The target population for ARMS III is the official USDA farm population. This population is defined as ìall establishments except institutional farms that sold or normally would have sold at least $1,000 of agricultural products during the year.î Analysis issues for the target population relate to whether a sampled operation fits the farm definition.

Sample Design
ARMS III is a multiple frame survey using a list and an area frame. Because we are interested in obtaining data on all farms in the continental United States, all farms must have some probability of selection. An area frame alone would meet the criterion of complete coverage but would be inefficient, because the distribution of farms can be clustered and farm types may not be evenly distributed throughout the target population. An extremely large sample would be required to ensure adequate coverage for all levels of aggregation. Because lists are never 100% complete, a list frame by itself is not adequate because all farms would not have a chance of being selected for the sample. However, sampling from a list would be more efficient because farms can be identified and targeted for sampling. Using the list with a complementary area frame allows targeting the sample to include all farm types and ensures that all farm operations have a chance of being selected for the sample.

The list frame sample is a stratified sample within states. The strata are defined by sales class and commodities produced. The area frame sample is a sample of nonoverlap tracts from the NASS June Agricultural Survey (JAS). The strata are the poststratified value of sales from JAS.

Data Collection
Data for each sampled operation are collected through personal interviews. Respondent burden is a major consideration for ARMS. The interview length is targeted for 1 hour but could take longer. Questionnaire versions differ depending on data requirements. Additional detail on costs of production for specific commodities is collected on a rotational basis.

Data requirements can vary between and within survey years. These data requirements are driven by the need for information to answer specific questions related to agricultural policy. The need for information on farm income, expenses, assets and debt, inventory values, and agricultural production is fairly constant over the years. A set of core questions are used to obtain this information. Additional questions are added to obtain economic data for areas of special interest. Much of the requested data are totals for broad categories; for example, seed expense for all crops, total marketing expenses, and so forth. Often data within the broad categories can be widely divergent. We might ask for total acres planted to all vegetable crops or asset value for trucks and cars. To get a clear picture of farm production expenses, we not only need operator expenses but also corresponding landlord and contractor expenses.

Reporting for ARMS is voluntary. Respondents may refuse to provide data for the farm operation. Others may agree to the interview but may refuse to answer some questions, resulting in partially completed questionnaires. In addition, reports could be lost because the sample was inaccessible or the sample units were out of scope for the survey.

Data Editing
When survey forms first arrive at each state office, the nonoverlap status of area records is verified and list records are checked for duplication. The forms are checked for completeness. Listings of descriptive statistics for several items are provided to help statisticians edit in values for incomplete items. Many states have there own sources of data for this purpose.

All states use the same computer edit for data validation, reasonable value checks, and relationship checks. Data validation checks verify that items were keyed correctly, item numbers were correct, and item codes were not duplicated. Data validation checks are generally consistent over all states. However, tests for reasonable value can vary widely depending on the state and item being checked. The reasonable value checks can do only so much to find possible outliers. With economic data, it is often difficult to determine whether an outlier is invalid data or an extreme observation. Relationship checks in a generalized edit cannot be comprehensive enough to cover all situations that may be encountered. Further, the edit checks look at relationships within a single record. For the checks to be effective, they also must be considered at the aggregate levels.

Summarization
Estimates are based on a reweighted estimator. The delete-a-group-jackknife procedure is used to estimate variances. Details of the estimators are beyond the scope of this paper but can be obtained by contacting the NASS Statistical Methods Branch .

Data Analysis
One of the goals of survey design is to reduce total error for the survey. The sampling error component of total error can be controlled through the use of an efficient sampling design. The editing component of the survey design and the IDAS tool for data analysis work together to reduce the nonsampling error component of total error.

Nonsampling errors are often a major portion of the total error and must be controlled wherever possible. Possible sources of nonsampling errors include misreported or misrecorded data, problems with data capture, omitted data, and item refusal and imputation.

There are three possible responses to each data item request on the survey instrument: (1) a positive value, (2) zero, or (3) item refusal. If the response is positive, the value is recorded

in the corresponding cell on the questionnaire. If the reported value is zero, the cell is left blank. A minus 1 is entered to indicate item refusal. A data value is considered to be valid if the item passed the validity, range, and/or relationship checks in the generalized edit. Data are considered invalid if the response fails one or more of the edit checks.

For a sampled operation, the following situations are possible for each requested data item:

1. An item was reported and the value is valid.
2. An item was reported and the value is invalid.
3. An item was not reported and zero is valid.
4. An item was not reported and zero is invalid.
5. An item was refused.

Note that the value for a data item may be considered valid even if the reported value is in error (nonsampling error). Conversely, the value for a data item could be invalid even though the reported value was actually correct (outlier).

NASS deems certain data items essential to answering the questions on the well-being of the farm sector. Data are required for these surveyed items. If the respondent refuses to report values for these items or the data have been omitted, values must be imputed for the missing data. Two types of imputation are used for ARMS: manual imputation and machine imputation. A large portion of the manual imputation occurs during the edit phase and is generally based on information external to the survey. However, as survey data become available, the IDAS tool can provide information for the manual imputation based on the reported data for the item of interest.

In general, machine imputation occurs on an item-by-item basis in a stepwise manner starting with region, sales class, and farm type. If the pool of values for region, sales class, and farm type is too small, the restriction on region is dropped and only sales class and farm type are used. If the second level pool is also too small, sales class is dropped and only farm type is used. This procedure ensures that data values are imputed at the most specific level available.

Along with data collection and editing, data analysis is an integral part of obtaining a clean, edited data set. Survey statisticians should not wait for data sets to be clean before starting data analysis. Portions of IDAS can be used to help clean up edited data files. The IDAS tool enables a statistician or data analyst to view whole farm data as well as view data relationships at various aggregate levels. IDAS is another source of values that can be used to edit in values for missing and incomplete data.

DATA USAGE ISSUES

NASS, with some guidance from ERS, sets estimates for farm production expenses in an official board process. These estimates include total expenditures and expenses for 16 major component items. These estimates are set at the national level. In addition, estimates for total expenditures are also set for crop farms, livestock farms, and sales classes at regional and national levels. The *Farm Production Expenditures* release is published in July and contains current year and revised previous year estimates.

ERS publishes several reports based on ARMS III data:

1. *Structure and Characteristics of U.S. Farms*
2. *Financial Performance of U.S. Farm Businesses*
3. *Economic Well-Being of Farm Operator Households*
4. *Farm Business Economics Report*

ERS also produces special reports in response to policy questions from Congress, the Executive Branch, and other interested parties.

Data Analysis

Data users need information for a complex set of interrelationships. Data items may be summarized for aggregate and disaggregate levels. Estimates vary for the same and different variables. Standard descriptive statistics on single items would not be sufficient to check for problems with relationships among multiple data items. For example, if feed expenses are reported, are expenses reported for livestock purchases or veterinary services? Are there livestock sales and marketing expenses? Relationships between the livestock items in the example must be reasonable. Data relationships must be reviewed.

Data analysts cannot look at all possible combinations of relationships; they need to concentrate on the most important relationships. As both data users and data analysts, statisticians and economists must work together to develop specifications that provide the types of checks that are required to validate data relationships.

IDAS DESIGN ISSUES

An IDAS must cover all the analysis issues in a coherent manner. Intended users will be interdisciplinary and may or may not have a strong background in statistics. The system must be robust and recognize that statistical outliers may be valid. Analysts must be able to look at data at multiple summary levels (i.e., state, regional, and national). The system must have drilldown capabilities to enable the analyst to view data relationships at the sampled operation (record) level. There should be the capability to view and record comments at the record level.

IDAS IMPLEMENTATION

The following screenshots illustrate how we incorporated survey design issues into IDAS. The main menu is the portal that leads to various areas of analysis (Figure 1). The general topics for these areas are risky records, production expenditures, capital expenditures, acreage ratios, other ratios, assets and debt, income, other listings, and expense ratios. Each area contains a set of items for which data relationships for various levels of aggregation can be reviewed. IDAS allows analysts to focus on different areas during the editing and analysis phases of the survey.

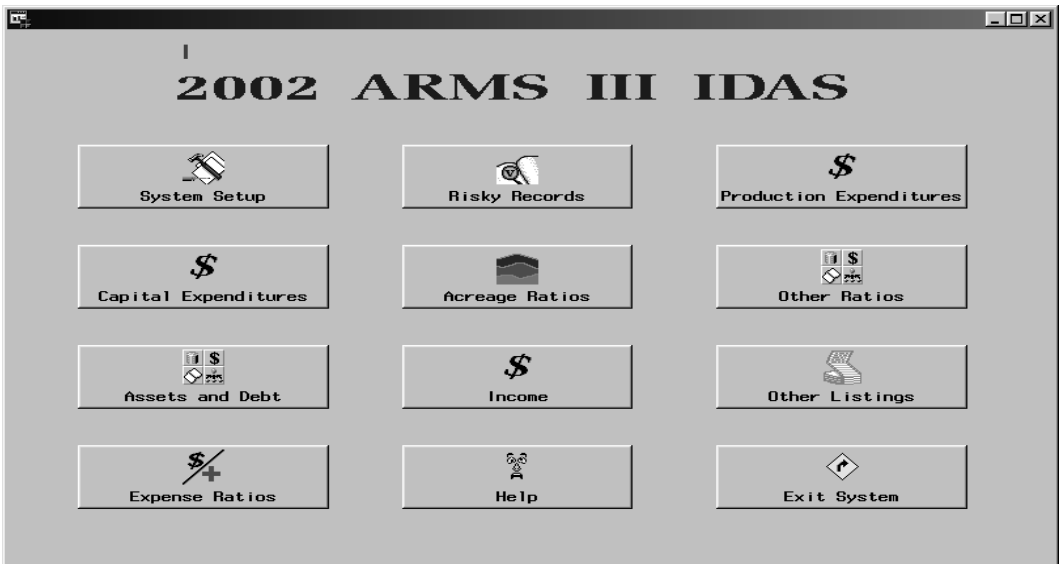


Figure 1 - ARMS III Main Menu

Each button represents a different area for relationship checks. The **System Setup** button is used to prepare the data for analysis and set regions when appropriate. Users can get help by clicking on the **Help** button.

Let us assume that you want to see whether reported seed expenses are reasonable. Expenses are reported as total dollars spent for the expense item or category. You cannot determine whether an expense is reasonable from the reported value alone. However, you could obtain a measure of reasonableness by looking at expense as a ratio to some measure of size; for example, seed expense per acre or seed expense as a percent of total expense.

Select one of the general topics from the main menu. Clicking on the button for the selected topic brings up a dropdown menu that lists all variables or subject areas available for analysis. Because you are interested in analyzing seed expenses, you could look at comparisons involving acreage ratios. Click on the **Acreage Ratios** button. A dropdown menu of available variables, such as seed expense, fuel expense, and insurance appears (Figure 2).

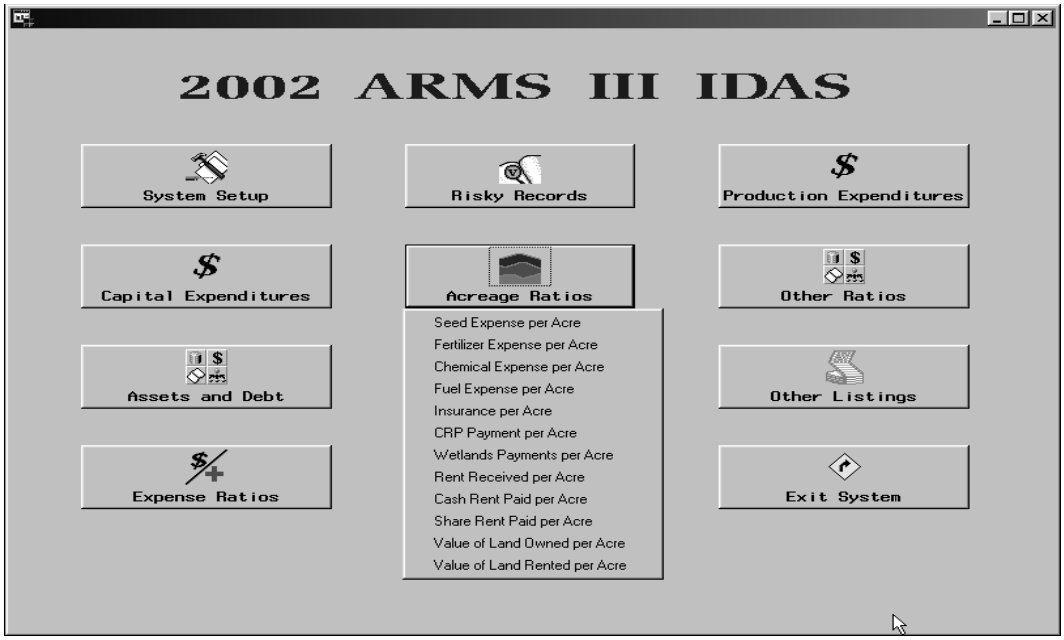


Figure 2 - Main Menu with Dropdown for Acreage Ratios

Selecting **Seed Expense per Acre** from the dropdown menu opens a screen with a graph of plotted values (Figure 3). Ratios can be plotted by economic region, state, or sales class. Buttons let you alter plot variables. Change the analysis variable by clicking on the **Item** button and choosing a new expense variable from the dropdown menu. Click on the **ID By** button to change the marker color to indicate farm type, sampling frame, presence or absence of comments, or whether the record was marked for additional analysis (pulled record). The **ID By** value is shown in the legend. Click on the **Category** button to change the level of aggregation. Changing the category changes the plotted-by variable on the X axis. The Figure 3 plot shows seed expense per acre by economic region. The ratio for one observation appears to be extreme. You need additional information to determine whether there is a problem with the data as reported.

Clicking on the **Category** button and selecting **Farm Type** from the dropdown menu changes the plotted variable on the X axis (Figure 4). From this plot, you see that most of the higher expense-to-acre ratios are concentrated in one type of farm. You can find the type of farm by clicking on the **Farm Types** button, after which a dropdown list of farm types appears. In this instance, you would discover that the high ratios are associated with nursery operations, where high expense-to-acre ratios would be expected. However, our suspect data point still seems to be extreme.

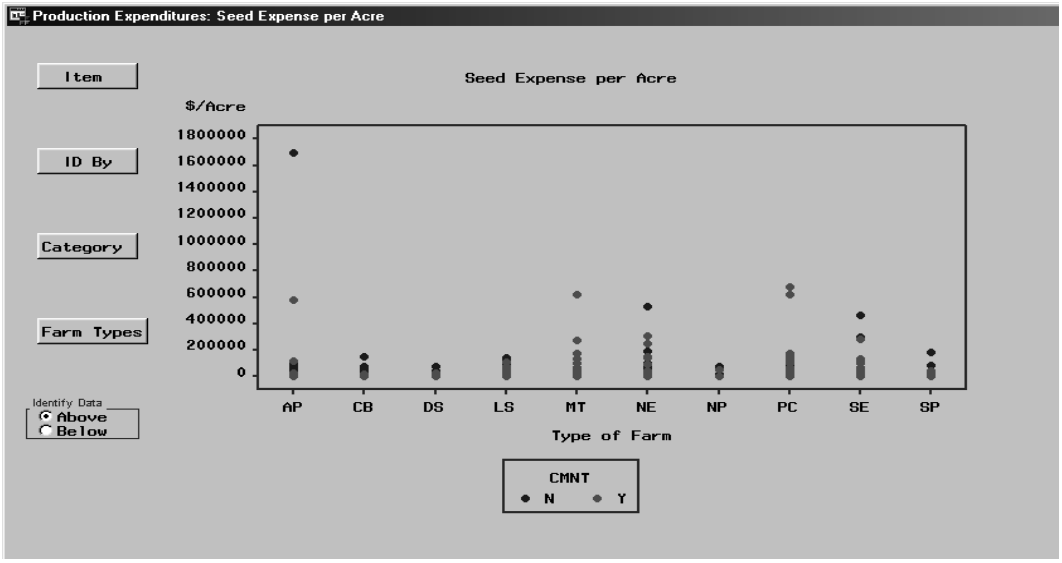


Figure 3 - Seed Expense per Acre by Economic Region

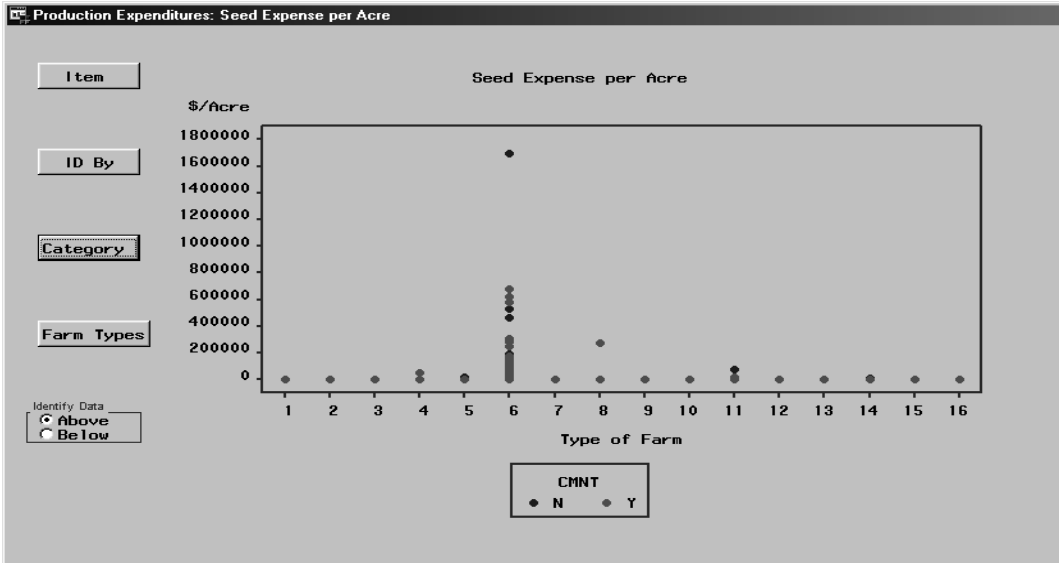
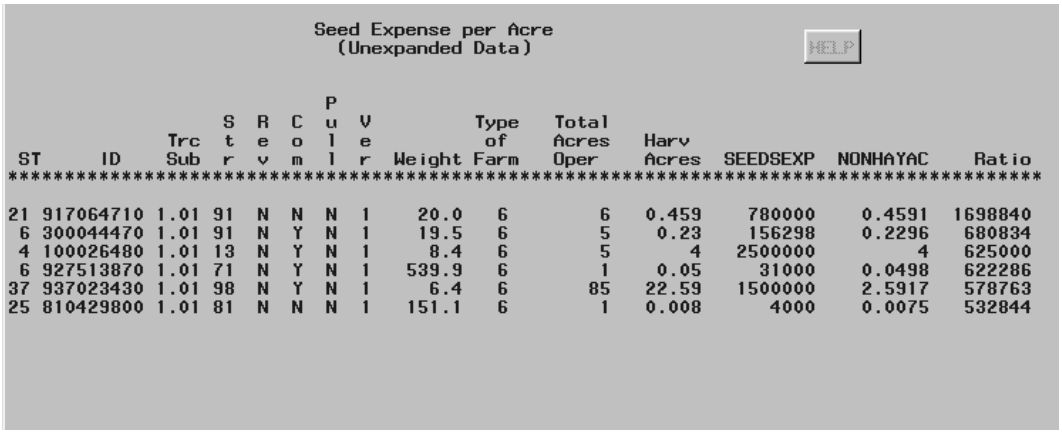


Figure 4 - Seed Expense per Acre by Farm Type

Clicking on a data point in the plot produces a list of all reports with ratios greater than or equal to the point selected (Figure 5). This list shows the expense variable and the acres used in the calculation. For the suspect data point (Row 1), the reported acreage is very small compared with the seed expense. However, you still do not have enough information to conclude whether there is a problem with the seed expense data for the operation. From here, you could click on the row for the record to open the first in a series of three screens that show expanded and unexpanded summary values for that record, or you could continue analysis on the seed expense item by returning to the main menu, clicking on the **Expense Ratios** button and observing whether the record appears to be out of line when compared with total expense.

The **Unexpanded Data** screen is the second of the three screens which show summary values for a record (Figure 6). These screens provide information on the type of operation and show summary values for specified expense and income categories. Based on the commodity cash income ($9.751 million) and other expenses for our suspect record, you might conclude that the seed expense was not out of line for this operation.

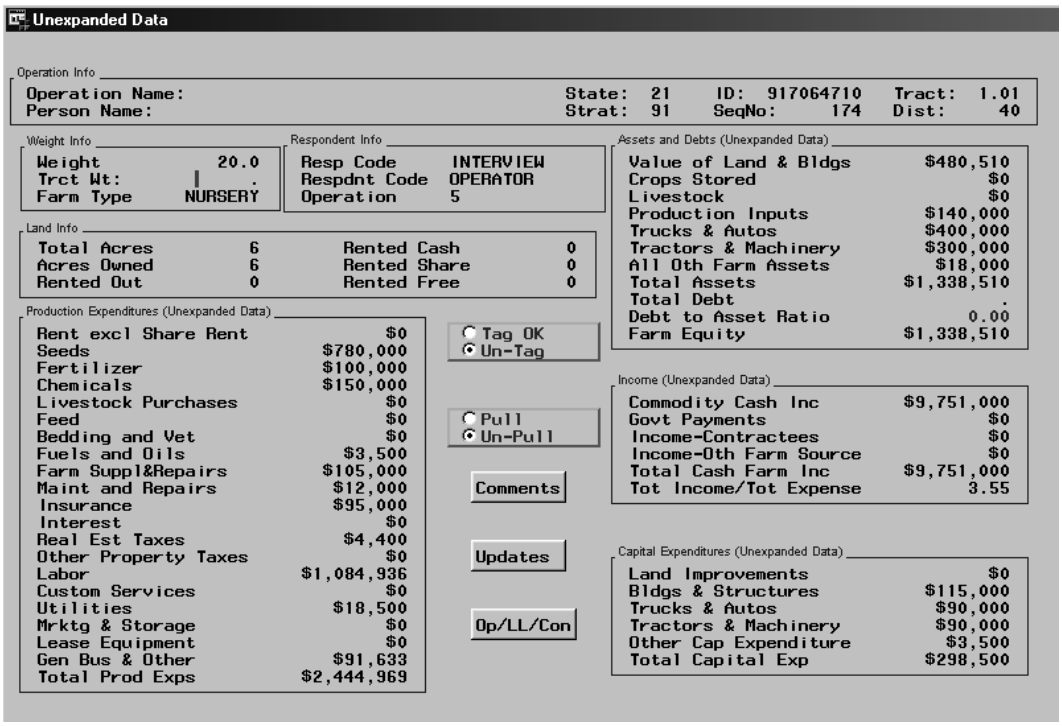


Seed Expense per Acre
(Unexpanded Data)

HELP

ST	ID	Trc Sub	Str	Rev	Com	Pul	Ver	Weight	Type of Farm	Total Acres Oper	Harv Acres	SEEDSEXP	NONHAYAC	Ratio
21	917064710	1.01	91	N	N	N	1	20.0	6	6	0.459	780000	0.4591	1698840
6	300044470	1.01	91	N	Y	N	1	19.5	6	5	0.23	156298	0.2296	680834
4	100026480	1.01	13	N	Y	N	1	8.4	6	5	4	2500000	4	625000
6	927513870	1.01	71	N	Y	N	1	539.9	6	1	0.05	31000	0.0498	622286
37	937023430	1.01	98	N	Y	N	1	6.4	6	85	22.59	1500000	2.5917	578763
25	810429800	1.01	81	N	N	N	1	151.1	6	1	0.008	4000	0.0075	532844

Figure 5 - Listing of Seed Expenses per Acre (Unexpanded Data)



Unexpanded Data

Operation Info

Operation Name:
Person Name:
State: 21 ID: 917064710 Tract: 1.01
Strat: 91 SeqNo: 174 Dist: 40

Weight Info

Weight	20.0
Trct Wt:	.
Farm Type	NURSERY

Respondent Info

Resp Code	INTERVIEW
Respdnt Code	OPERATOR
Operation	5

Assets and Debts (Unexpanded Data)

Value of Land & Bldgs	$480,510
Crops Stored	$0
Livestock	$0
Production Inputs	$140,000
Trucks & Autos	$400,000
Tractors & Machinery	$300,000
All Oth Farm Assets	$18,000
Total Assets	$1,338,510
Total Debt	.
Debt to Asset Ratio	0.00
Farm Equity	$1,338,510

Land Info

Total Acres	6	Rented Cash	0
Acres Owned	6	Rented Share	0
Rented Out	0	Rented Free	0

Production Expenditures (Unexpanded Data)

Rent excl Share Rent	$0
Seeds	$780,000
Fertilizer	$100,000
Chemicals	$150,000
Livestock Purchases	$0
Feed	$0
Bedding and Vet	$0
Fuels and Oils	$3,500
Farm Suppl&Repairs	$105,000
Maint and Repairs	$12,000
Insurance	$95,000
Interest	$0
Real Est Taxes	$4,400
Other Property Taxes	$0
Labor	$1,084,936
Custom Services	$0
Utilities	$18,500
Mrktg & Storage	$0
Lease Equipment	$0
Gen Bus & Other	$91,633
Total Prod Exps	$2,444,969

Tag OK / Un-Tag

Pull / Un-Pull

Comments

Updates

Op/LL/Con

Income (Unexpanded Data)

Commodity Cash Inc	$9,751,000
Govt Payments	$0
Income-Contractees	$0
Income-Oth Farm Source	$0
Total Cash Farm Inc	$9,751,000
Tot Income/Tot Expense	3.55

Capital Expenditures (Unexpanded Data)

Land Improvements	$0
Bldgs & Structures	$115,000
Trucks & Autos	$90,000
Tractors & Machinery	$90,000
Other Cap Expenditure	$3,500
Total Capital Exp	$298,500

Figure 6 - Summary Value Screen for Unexpanded Data

Clicking on the **Risky Records** button on the main menu will bring up a list of records that are possible outliers (Figure 7). The expanded values from these records is above a percentage threshold for the state, region, or national estimates. Although the data may be valid, these records will have undue influence on the estimate.

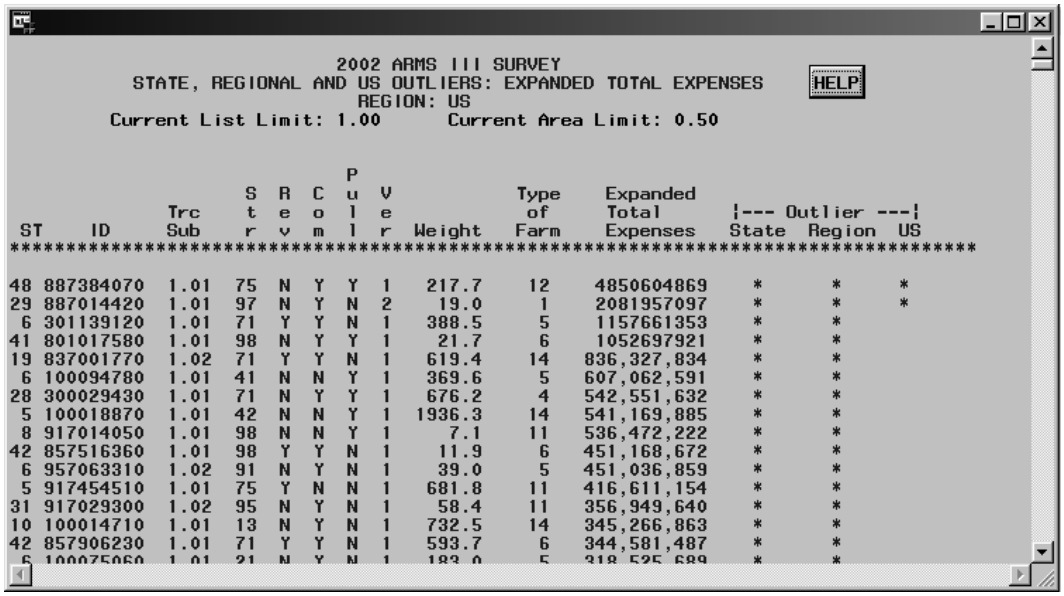
CONCLUSION

IDAS provides data analysts with a tool to review data with complex interrelationships. Analyst need not have a strong statistical background to use IDAS effectively. NASS not only uses IDAS with ARMS III, but also with other major surveys with similar types of data relationships, including quarterly acreage surveys, livestock inventories, and labor surveys.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Van B. Johnson
National Agricultural Statistics Service
Room 6436A
1400 Independence Avenue SW
Washington, DC 200250
Work Phone: (202) 720-6482
Fax: (202) 264-3725
Email: Van_Johnson@nass.usda.gov



2002 ARMS III SURVEY
STATE, REGIONAL AND US OUTLIERS: EXPANDED TOTAL EXPENSES
REGION: US
Current List Limit: 1.00 Current Area Limit: 0.50

HELP

ST	ID	Trc Sub	Strv	Rev	Com	Pull	Ver	Weight	Type of Farm	Expanded Total Expenses	Outlier State	Outlier Region	Outlier US
48	887384070	1.01	75	N	Y	Y	1	217.7	12	4850604869	*	*	*
29	887014420	1.01	97	N	Y	N	2	19.0	1	2081957097	*	*	*
6	301139120	1.01	71	Y	Y	N	1	388.5	5	1157661353	*	*	
41	801017580	1.01	98	N	Y	Y	1	21.7	6	1052697921	*	*	
19	837001770	1.02	71	Y	Y	N	1	619.4	14	836,327,834	*	*	
6	100094780	1.01	41	N	N	Y	1	369.6	5	607,062,591	*	*	
28	300029430	1.01	71	N	Y	Y	1	676.2	4	542,551,632	*	*	
5	100018870	1.01	42	N	N	Y	1	1936.3	14	541,169,885	*	*	
8	917014050	1.01	98	N	N	Y	1	7.1	11	536,472,222	*	*	
42	857516360	1.01	98	Y	Y	N	1	11.9	6	451,168,672	*	*	
6	957063310	1.02	91	N	Y	N	1	39.0	5	451,036,859	*	*	
5	917454510	1.01	75	Y	N	N	1	681.8	11	416,611,154	*	*	
31	917029300	1.02	95	N	Y	N	1	58.4	11	356,949,640	*	*	
10	100014710	1.01	13	N	Y	N	1	732.5	14	345,266,863	*	*	
42	857906230	1.01	71	Y	Y	N	1	593.7	6	344,581,487	*	*	
6	100075060	1.01	21	N	Y	N	1	183.0	5	318,525,689	*	*	

Figure 7 - Listing of State, Regional, and National Outliers