Improving the Quality of Survey Data Through an Interactive Data Analysis System

Robert Hood, National Agricultural Statistics Service, Washington, D.C.
Mark Apodaca, National Agricultural Statistics Service, Washington, D.C.

Abstract
The National Agricultural Statistics Service (NASS), an agency of the U.S. Department of Agriculture, conducts surveys to provide accurate and reliable agricultural forecasts and estimates for a variety of commodities. To ensure and improve the quality of its survey data, NASS has recently begun to develop and use SAS-based information systems.

Currently, much time is spent editing incoming data with little time available for analysis before data are summarized. Present analysis tools are limited to data listings and outlier printouts. While useful, they are somewhat limited in the problems they flag, and resolution of problems generally involves time-consuming review of paper questionnaires or data files. The time between data collection and summarization is very limited and must be used as efficiently as possible.

The authors have developed a SAS-based application to interactively analyze survey data to improve the quality of data going to summary. This system, developed with extensive use of SAS/AF and SAS/EIS software, takes advantage of object oriented concepts to build user-friendly graphical user interfaces (GUI's). These GUI applications allow our system to be used with very little training or documentation and without any knowledge of the SAS system. This paper gives an overview of the system, its development, and coding examples of some techniques used in the system.

I. Introduction

The National Agricultural Statistics Service (NASS), an agency of the U.S. Department of Agriculture, conducts surveys to provide accurate and reliable agricultural forecasts and estimates for a variety of commodities. As a result of new technology and the desire to ensure and improve data quality, NASS has recently begun to develop and use information systems based on SAS/AF and SAS/EIS software.

NASS has 45 state statistical offices (SSO's) which are responsible for collecting and editing data for all 50 states. Currently, SSO's spend a large amount of time editing incoming data and thus have little time for analyzing the data before it is submitted to headquarters for summary. Analysis tools available to the SSO's are limited primarily to data listings and outlier printouts (which are not available until a few days before all data must be submitted for summarization). These listings are useful, but limited in the data problems they flag, and resolution of problems generally involves reviewing paper questionnaires or data files. Data collection, editing, and analysis time is very limited during a survey and must be used as efficiently as possible. The availability of an information system that provides current and historic data comparisons and analysis at the click of a mouse should make the process much more productive.

Using a SAS-based information system in operational programs is a relatively new undertaking for NASS. Work on a prototype system began only a couple of years ago with the purpose being to learn more about SAS/EIS and SAS/AF software and to determine if practical applications could be developed. The results of this initial endeavor were presented at the Twentieth Annual SUGI Conference in the paper "Interactive Analysis of Survey Data Using SAS/AF and SAS/EIS Software" by Robert Hood. The agency response to using such systems has been very positive. In fact, the system discussed in this paper is currently being used in 16 state offices and another module should be in place in other states by March 1996. The long range goal is to develop such SAS-based information systems for every SSO for all the major commodities covered by our surveys.

All data presented in this paper are fictitious.

II. The Interactive Data Analysis System (IDAS) Module for Hog Inventory

A. Description

The primary purpose of IDAS is to help identify "risky records" as early in the data collection period as possible. These are records that may warrant

verification or further investigation based on predefined criteria. These include, but are not limited to, records with large expansion factors, records with data out of range, records with large changes in inventory, and estimated records. IDAS further provides macro analysis capability in which historically unusual expansions for individual strata or stratum types can be directly traced to individual records with the click of a mouse. This type of drill-down or hierarchical analysis is completely unavailable elsewhere in our survey processing system.

Currently, large amounts of time are spent on ***editing*** incoming data with little time available for the ***analysis*** of the survey data. The current data review capabilities consist of data listings and potential outlier printouts (POPS) which are made available to SSO's toward the end of the data collection period. These printouts are quite large, and the review process is an overwhelming paper-oriented task. In contrast, IDAS is totally interactive, it identifies different types of possible outliers, it is fast, visually appealing and easy to use. Most importantly, the system identifies risky records early in the data collection process, giving the statistician more time to review the suspect record, and if necessary, the opportunity to recontact the respondent before the end of the data collection period. In addition, the use of IDAS during the survey period will enable the statistician to get a "feel" for the incoming survey data which leads to a better understanding of the final survey indication.

IDAS provides data analysis options during the data collection period and after all data have been collected. During the data collection period several data listings, based on certain criteria, are generated and made available for review. These include the potential outlier prints for total inventory and breeding stock as well as listings of (1) operations which reported hogs in the previous quarter and currently report no hogs, (2) operations which reported no hogs in the previous quarter and currently have hogs, (3) estimated operations, (4) operations whose respondent differed from the previous quarter, (5) operations which report any type of contract hogs, and (6) estimated records.

Post data collection analysis contains not only the above data listings but also a listing of operations whose reported hog inventory represents at least five percent of the stratum indication; bar charts representing survey indications, quarter to quarter ratios, stratum types (Hogs, Crops, Capacity, and Area) and individual stratum indications. Significant changes in stratum indications from quarter to quarter may help explain the change in the current indication. In addition, by breaking down the survey indication into parts (i.e., stratum types and stratum totals) the statistician can get a better picture of the pieces that contribute to the whole.

A. System Setup

Before using IDAS for the first time for each survey, some preliminary setup steps are required. Each state must define their current list strata, set the criteria used to identify records in the data listings, and execute a SAS program. Two setup programs are available which generate the data listings and graphs used in the analysis system depending on the timing of the use of the system. The System Setup consists of six options.

1. *Define List Strata*

   Each state is responsible for defining their list strata prior to using IDAS.

2. *Define Data Selection Criteria*

   Each State has the ability to customize the selection criteria used to subset the data to identify records for the data listings. Default values based on historical levels are provided so that a reasonable number of records will be identified for review. If these criteria are changed, the appropriate setup program must be run again before the changes will take effect. Figure 1 shows the program entry that is used to set these values. This screen shows what data listings are available along with the selection criteria for each listing.

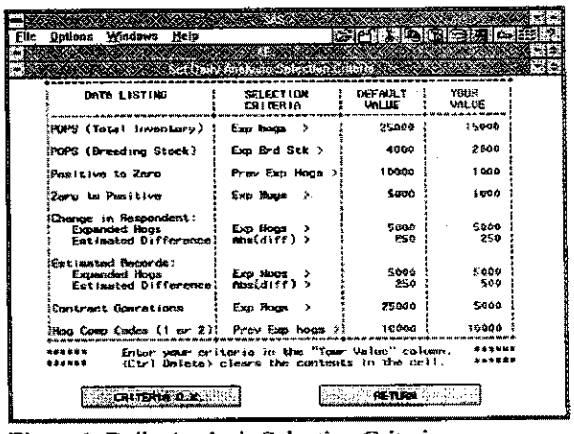

Figure 1. Daily Analysis Selection Criteria

3. *Convert Transport File to SAS Data Set*
   The survey data are collected then uploaded to a mainframe for batch edits. Currently, the IDAS system uses a subset of these edited data, which must be downloaded to a PC. (We do not have SAS/Connect.) The conversion of the transport file to a SAS data set must be executed each time before new data can be used in the system, but not each time someone uses the system. The appropriate setup program must also be run with each new batch of data.

4. *Setup Program (During Data Collection)*
   During the data collection period, this program is used to create all data listings. The analysis options available during the data collection period are contained in the DAILY DATA ANALYSIS and REVIEW RISKY RECORDS icons. The program expands the data using previous quarter's stratum expansion factors and takes about 10 minutes to run.

5. *Setup Program (Post Data Collection)*
   After all the data have been collected, this program is run to create final data listings and graphs with the current quarter's expansion factors. All analyses options are available after this setup is executed, which takes about 15-20 minutes.

6. *Post Survey Cleanup*
   This option is used after all analyses are completed for the current quarter. It deletes and archives files not needed for the next quarter.

B. Analysis Options

Review Risky Records
All records that meet predefined criteria will be identified here. This option is available at all times during the survey period. The records contained here are a subset of the records displayed with the various options in the Daily Data Analysis data listings. These records should be the first set of records reviewed as they meet more stringent criteria. Extended tables, which allow the user to select a record to display additional information, are used for all data listings. Figure 2 shows the extended table for the Risky Records listing. All other data listings have a similar format.

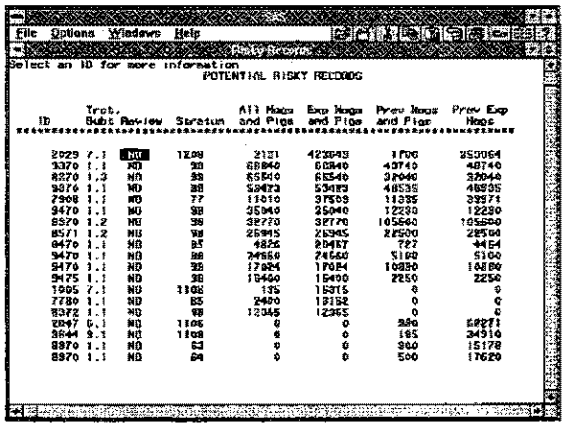

Figure 2. Risky Record Listing.

Additional information can be obtained for a specific ID by left clicking on that ID within the extended table. Figure 3 shows the frame that is displayed when an ID in the extended table is selected. Subsequent screens displaying record level information are traffic lighted (color coded) to point out inconsistencies or strange relationships between current and previous quarter data.

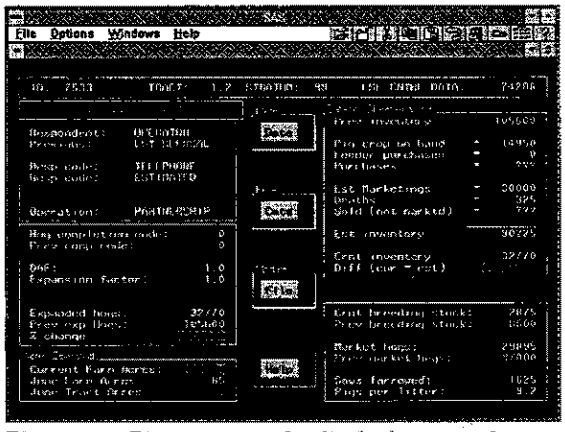

Figure 3. AF/Frame Entry for displaying more data for a selected record from an extended table listing.

You can mark records that have been reviewed by clicking on a check box within the frame. This changes the value of the REVIEW column (see Figure 2) from NO to YES, which helps prevent confusion over which records have been reviewed and which ones have not.

The "Enter Cmts" pushbutton allow the reviewer to enter comments or explanations for the selected record. Comments for all records, along with appropriate identification information are written to a

text file, which can be printed. Comment files are archived in the Post Survey Cleanup. The "Update File" pushbutton is similar. This option is used to create a text file containing any updates or edits that need to be made. The file can be printed to allow for easy key entry and re-editing.

The "More Data" pushbutton is available for records that were also surveyed the previous quarter. This links to another AF/Frame entry shown in Figure 4. This screen shows the breakdown of inventory components for the current and previous quarter.

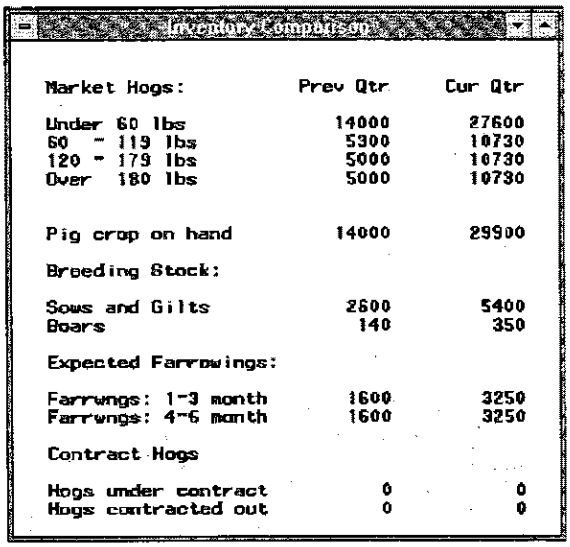

Figure 4. AF/Frame entry displaying more data.

Daily Data Analysis

This option, available at any time during the survey, allows you to review records that fall into specific categories. The criteria used to identify records for each data listing can be customized through the SYSTEM SETUP option. There are eight listings available and an option to change how the records are sorted. The available listings are:

1. *POPS Total Inventory*
2. *POPS Breeding Stock*
3. *Zero => Positive Records*
4. *Positive => Zero Records*
5. *Change in Respondent*
6. *Estimated Records*
7. *Hog Completion Codes (1 or 2)* (Completion Codes within the survey can affect how imputation is done. A completion code of 1 indicates that the sampled unit has hogs, but the section of the survey pertaining to hogs was incomplete for some reason. A completion code of 2 indicates that the hog section was incomplete and the prescence of hogs on the operation is unknown.)

All of these data listings are similar in function to the Review Risky Record data listing.

Post Data Collection

After all data have been collected, the Post Data Collection Setup program should be executed. In the post data collection setup, various graphs are created which can be replayed using the SAS/Graph Output class in a SAS/AF: Frame entry. These graphs show overall indications, as well as stratum type and individual stratum indications. You can drill down from these charts to the individual record level information. Printing options for all screens are also being developed.

There are three options available.

1. *Hog Strata Outliers* - Displays operations in certain strata that have expanded hog inventory greater than 5% of the entire stratum expansion. Drill down capability is available for any ID in this listing as described in the Review Risky Records option.

2. *Survey Indications (Charts)* - Displays a bar chart comparing the current indication to the previous quarter indication. Additional charts can be accessed through a list menu activated by a push-button. Figure 5 shows the AF/Frame entry that displays the indications by strata type. Also shown super-imposed on the frame is the resulting data box that is displayed when the user selects the text label for a strata type.

3. *Strata Indications (Charts/Top Five)* - Displays bar charts comparing current to previous quarter indications by stratum type. Additional information is displayed in a frame when the user selects the text label for any of the bars. The user can then "drill down" to see the actual observations that contribute to the bar. The listing of observations has the same functionality as all other data listings that have been discussed. Section III discusses how this application was developed.

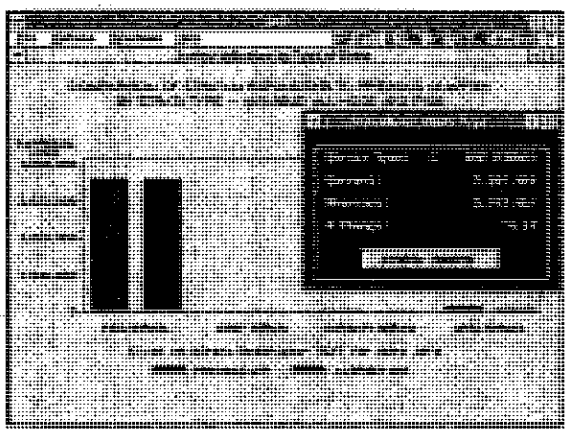

Figure 5. Survey Indications (Charts) Frame.

C. Other Features

There are several other features in this system that have not been discussed due to the size limitation of the paper. One feature is the online help, which can be found by selecting the HELP icon on the main menu. This describes the system and each of its components. Other features found under the EXIT icon on the main menu include:

1. *Tag Records* - Executes a SAS program to "tag" all records that have been checked as OK. This changes the value of the "review" variable from "NO" to "YES" on all data listings that contain the tagged ID, so users will know that the record has already been reviewed.

2. *Create Comment/Update/Review Files* - Create text files for records for which the user entered comments or update instructions. Also creates a text file of IDs for records that have been reviewed.

3. *Create Data Listings* - Creates text files of any or all of the available data listings.

4. *Exit System* - Exits IDAS and the SAS system.

III. SCL CODING EXAMPLES

Following is a brief discussion of a couple of techniques that are used often in our system. The two examples given below are presented as the PROBLEM and the SOLUTION. The solutions presented here are by no means the only ones possible, but they met the needs of our system.

A. SAS/Graph Output Class

Problem: To display a graph and to "drill down" to more specific information based on the user selecting a region of the graph.

Solution: This could easily be done with the SAS/AF: Frame Graphics Class, except that the graph options required for our graphs were not available. Notably (at least for version 6.08) we found it impossible to create bar charts requiring the "group" and "subgroup" options using the Graphics Class. Our solution was to create the graph with the options that were needed using SAS/Graph and then replay it using the SAS/Graph Output class. We also found it easier (and necessary in some cases) to use SAS/Graph to set axis, symbol, and legend options.

One use of the SAS/Graph Output class is to display graphics that have been previously created and stored in a GRSEG catalog entry. It is convenient to replay graphics in this manner without using PROC GREPLAY, but what else does it buy you? With SCL, it buys a lot of options. For example, we wanted to display a bar chart on the screen and have the user select the label of the bar in order to "drill down" to see more information. To do this, we used the "_GET_INFO_" method. This method returns the identifier for an SCL list which contains information about the point that was selected on the graph. The object that displays the graph is named GRAF. The object's name is used in the SCL program as a label. The code in the labeled section executes each time the object is selected, or in this case, each time a user clicks on the graph. The SCL code is shown below.

```
INIT:
 length seltype $16;
 _msg_='Select a stratum type for more information';
RETURN;

GRAF:
 call notify('graf',_'get_info_',grafid);
 call putlist(grafid,'GRAF INFO',2);
 seltype=getnitemc(grafid,'TEXT');
 if seltype not in ('HOG STRATA','CROP
STRATA','CAPACITY STRATA',
  'AREA STRATA') then
  do;
    alarm;
    _msg_='You must select a strata type';
   return;
  end;
```

```
  else do;
  if seltype='HOG STRATA' then do;
    type=1;
    call display('strtyinf.frame','seltype,type);
  end;
else if seltype='CROP STRATA' then do;
    type=2;
    call display('strtyinf.frame','seltype,type);
  end;
  else if seltype='CAPACITY STRATA' then do;
    type=3;
    call display('strtyinf.frame','seltype,type);
  end;
  else if seltype='AREA STRATA' then do;
    type=4;
    call display('strtyinf.frame','seltype,type);
  end;
RETURN;

TERM:
RETURN;
```

The '_GET_INFO_' method results in an SCL list identifier whose contents can be display with the following statement:

```
call putlist(grafid,'GRAF INFO',2);
```

The resulting list is

```
GRAF INFO:( SPOT=<invalid list id> [0]
            X=547
            Y=290
            TEXT='HOG STRATA'
            SPOTTYPE=30
            SPOTID=57
          ) [269]
```

To extract the information needed, we use the 'GETNITEMC' statement which returns a character value identified by its name in an SCL list. The statements says to get the character value for 'TEXT' from the list identifier 'grafid'. Based on the text selected, we assign a value to the variable "type" in the IF-THEN statement and pass the value to another FRAME entry that allows you to see more information. To receive parameters from a CALL DISPLAY routine, use the ENTRY statement. The first line of the INIT section of the SCL program for STRTYINF.FRAME is: ***entry seltype $ 16 type 1;***.

Hotspots could not be used for obvious reasons. First, the graphs are not created by the developer and therefore the hotspots could not be created. The graphs are created in the state offices only when all the data has been collected. Second, with the SAS/Graph Output Class, we can create any type of graph and use the '_GET_INFO_' method to process user input. We found this method to be particularly useful for getting around some of the limitations imposed by the Graphics Frame options for bar charts and to simulate drill down capability.

B. Extended Tables as Selection Lists

Problem: To display the values of certain variables from a SAS data set for a specific number of observations, and to allow the user to select one record to see additional information and to record comments.

Solution: Extended tables

Extended tables are PROGRAM entries that can be used in a variety of ways including displaying data in tables, updating a data set, or as custom selection lists. While there are numerous ways to display data, the IDAS system uses extended tables because of their ability to serve as selection lists at the same time.

Note: You must specify the EXTENDED TABLE attribute in the BUILD procedure's general attribute (GATTR) window in order to use the SETROW routine.

Building an extended table that displays multiple rows of data is quite simple. First, you delimit the area of the screen that is always visible by following it with a line containing three logical 'NOT' signs in the first three columns. This area can be used for labels for the columns, titles, or other informational text and is nonscrollable. The area below the line of the logical not signs is scrollable. Then, below the delimiter line, you have to define only one row of fields in what is called a *logical row*. It is in this row that you specify the field names for the variables to be displayed. These names do not have to be the same as the SAS variable names You can give the window variable a different name by using the ALIAS field in the ATTR window. This 'alias' is the name by which the field is referenced in the SCL program. Figure 6 shows the BUILD Display window with the nonscrollable area, the delimiter line, and the logical row for one of the IDAS data listings. Figure 7 shows the field attribute window (ATTR) for the field LSFID.

The data that fill the extended table can come from a variety of sources including: SAS data sets, SCL lists, arrays in the SCL program, or external files. The tables can also have either a fixed number of rows (static) or a variable number of rows (dynamic). If you don't know the size of the data set or SCL list or if the number of observations can change, then use a dynamic extended table. The extended tables in IDAS are all static tables that use SAS data sets. The number of observations that are displayed are fixed. Shown with the screen captures of the BUILD display window and the Field Attributes window is an example of the SCL code used to create and fill a static dynamic table.

Display Window:

```
+-------------------------------------------------------------------------+
|                                                                         |
|                            POTENTIAL OUTLIER PRINT                      |
|                             Expanded Total Hogs                         |
|                                                                         |
|                                                                         |
|                                                                         |
|                                    All Hogs    Exp Hogs  Breeding   Market|
|     ID    Reviewed  Stratum        and Pigs    and Pigs    Hogs       Hogs |
|  **********************************************************************  |
|^^^                                                                      |
|  &ID_____   &REVIEW   &STRATA &LHOGTOTL__   &EXPTHOGS_ &BREEDSTK &MRKTHOGS|
|                                                                         |
+-------------------------------------------------------------------------+
```

Figure 6. Build Display Window for Potential Risky Records Data Listing.

Field Attributes:

```
-------------------------------------------------------------------------
  Field name:  ID        Frame: 2   Row: 1      Col: 3      Length: 9
       Alias:  LSFID       Choice group:                Pad: _
        Type:  NUM              Protect: YES
      Format:                     Just: RIGHT
    Informat:
 Error color:  RED          attr:  REVERSE      Help:
        List:
     Initial:
     Replace:
     Options: CAPS AUTOSKIP
-------------------------------------------------------------------------
```

Figure 7. Field Attribute Window (ATTR) for Variable LSFID.

SCL Program:

```
INIT:
  _msg_='Select an ID for more
       information';
  dsidpop = open('td.pops');
  call set(dsidpop);
  numobs = attrn(dsidpop,'nobs');
  call setrow(numobs,1);
  selid=_blank_;
  seltrct=_blank_;
RETURN;

TERM:
  if dsidpop then rc = close(dsidpop);
RETURN;


GETROW:
  rc = fetchobs(dsidpop,_currow_);
RETURN;

PUTROW:
  selid=id;
  seltrct = lcrid;
  if selid ne _blank_ and seltrct ne
 _blank_ then do;
    if screen = 1 then
        call display ('univall.frame',
          selid,seltrct);
    else if screen = 3 then
         call display('univz2p.frame',
          selid,seltrct);
    end;
    else do;
       alarm;
        _msg_='Click on an ID to view
              more information';
      end;
  selid=_blank_;
  seltrct=_blank_;
RETURN;

MAIN:
     _msg_='Select an ID for more
  information';
RETURN;
```

CONCLUSION

The IDAS module for hog inventory was field tested by one state in March 1995, and by five states in June 1995. Full scale operational use of this system began in September 1995 in the 16 major hog producing states and continues on a quarterly basis. All state offices found the system useful and requested additional modules for other commodities. Testing for a cattle on feed module took place in February 1996, and several other modules are being considered. NASS' long range goal in its pursuit to improve the quality of survey data is to have such SAS-based information systems for every SSO for every major commodity.

The response from SSO users regarding the IDAS module for hogs has been overwhemingly positive. In fact, states that use the system for hogs, as well as some states to whom the system is not available, have requested modules for additional commodities.

We feel that SAS-based interactive systems such as IDAS can play an important role in our strategy to ensure and improve the quality of our survey data.

ACKNOWLEDGEMENTS

SAS, SAS/AF, SAS/EIS, and SAS/GRAPH are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

REFERENCES

Hood, Robert (1995), "Interactive Analysis of Survey Data Using SAS/AF and SAS/EIS Software". Proceedings of the Twentieth Annual SAS Users Group International Conference.

SAS Institute Inc. (1993), *SAS/EIS Software: Reference, Version 6, First Edition*, Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1993), *SAS/AF Software: FRAME Entry Usage and Reference, Version 6, First Edition*, Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1991), *SAS Screen Control Language: Usage, Version 6, First Edition*, Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1991), *SAS Technical Report P-222, Changes and Enhancements to Base SAS Software, Release 6.07*, Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1990), *SAS Screen Control Language: Reference, Version 6, First Edition*, Cary, NC: SAS Institute Inc.

Robert Hood (e-mail: rhood@ag.gov)
Mark Apodaca (e-mail: mapod@ag.gov)
USDA/NASS
Research Division
3251 Old Lee Highway
Room 305
Fairfax, VA 22030