WP. 43
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Geneva, Switzerland, 9-11 November 2005)

Topic (vi): Software for statistical disclosure control

DISCLOSURE ANALYSIS FOR THE CENSUS OF AGRICULTURE

Supporting Paper

Submitted by the National Agricultural Statistics Service, U.S. Department of Agriculture, United States [1]

[1] Prepared by Robert T. Smith (Robert_smith@nass.usda.gov).

Disclosure Analysis for the Census of Agriculture

Robert T. Smith

Census and Survey Division, National Agricultural Statistics Service, U.S. Department of Agriculture

Abstract: The National Agricultural Statistics Service used the network-flow disclosure methodology for the 2002 Census of Agriculture. This paper discusses how the statisticians and computer programmers worked closely together to achieve a very successful application of this disclosure methodology. The paper describes how the magnitude and complexity of the agricultural data structure were the impetus for the creation of a system to assist the analyst in the development and analysis of the input parameter files. Enhanced diagnostic tools allowed the analyst to continually review disclosure patterns and provide feedback that helped the computer programmers tailor the system specifically to the agriculture data set. This paper presents the details on these tools and discusses some of the modifications to the program logic that was a result of their use.

1 Background

The National Agricultural Statistics Service (NASS) used the network flow disclosure methodology for the Census of Agriculture. The network flow programs were originally developed at the U.S. Census Bureau for the economic censuses. The programs were modified by NASS for application to the 2002 Census of Agriculture and its follow-on programs including the Census of Puerto Rico, the Farm and Ranch Irrigation Survey and the Census of Aquaculture.

In the agency=s first application of this methodology it was important that the agriculture analyst provide feedback to the disclosure system so that it could be adapted more effectively to agricultural data. The publication tables were very complex, and it was very difficult to create the data file of >linear relations= which specified how an entry in one publication table could be derived from entries in that table or in other tables. The analysts were involved in the project from the beginning specification of the linear relations to the review of the final suppression patterns. It was this involvement and our ability to modify the system that made this a very successful project.

The application of this methodology to the census consisted of three broad areas; the creation of the input parameters and files, the modification and running of the disclosure programs, and the review of the diagnostics that identified any necessary modifications of the parameters or the program itself. Other than a brief overview of the system, this paper will not discuss the internal workings of the network flow system since that has been documented elsewhere. It will discuss how we achieved a very successful application of the methodology by a joint effort between agricultural and disclosure analysts. We were able to design an adaptable system that could respond to the inquiries and needs of the agricultural analysts. This resulted in a continual cycle of review and modification that ultimately yielded a high quality product. This paper will give the highlights of this cooperative effort in the context of the development of the input parameters, the diagnostics and the resulting modifications to the system.

2 Project Scope

The census of agriculture is a very large and complex project which is taken to obtain agricultural statistics for each county or county equivalent, state, and the Nation. The census of agriculture is the leading source of facts and statistics about the Nation=s agricultural production and provides a detailed picture of U.S. farms and ranches every five years. It is the only source of uniform, comprehensive agricultural data for every state and county or county equivalent in the U.S. The census data have very wide usage and are routinely used by farm organizations, businesses, State departments of agriculture, elected representatives and legislative bodies at all levels of government, public and private sector analysts, and colleges and universities.

The final census product comprises publications for each of the 50 states and the U.S. The published data products are both hard-copy and web-based and include 61 U.S. tables, 51 all states tables, 3050 state level tables and 2550 all counties by state tables. These tables contain more than 18 million cells. These are the data cells that required disclosure analysis.

3 System Overview

You could think of the agriculture census data structure as a large two-dimensional table where the rows refer to the geographic areas (U. S., states, and counties) and the columns refer to different agricultural statistics. In this structure, there are 3129 rows and 6002 columns. There are 1885 linear relations which define how a column is the sum of other columns.

3.1 Program Logic

Fortunately, we can divide this data set into subsets and process them separately. For example, we can process the U. S. and states in one computer run and then process the counties in following runs. In addition, each linear relation can be used to create an individual sub-table that can be checked for disclosures. When we do this, the largest single sub-table we process has 255 rows and 50 columns.

These sub-tables are dependent because a statistic found in one sub-table often appears in other sub-tables. If it is suppressed in one table, it must be suppressed in the other tables in which it appears, and additional complementary suppressions may be chosen in the other tables to make sure the data cell is protected.

The disclosure methodology is applied to these sub-tables. Four input files define their structure. Two files provide the geographic row information which comprises the U.S. total, a grouping of states, an individual state, or the counties in a state. The third file defines in terms of matrix numbers the columns as the linear relations within the publication tables. A fourth file contains a record for each cell in the table and includes its value, initial suppression flag, and other information necessary for the disclosure analysis.

From these files a disclosure table (sub-table) is created and converted into a network, initial suppressions are identified, capacities and cost are calculated and the minimum cost flow subroutine called to determine the complementary suppressions. The data file records are updated

to correspond to the new complementary suppressions. This entire process is repeated until all linear relations have been processed once. If we either suppress a cell or increase the protection on a cell which appears in an earlier processed table, backtracking must be done to recheck the earlier table for disclosures. The disclosure run is completed when all backtracking has been done.

3.2 Order of Runs

The census data are published at the U.S., New England Region, States and County levels. All of these levels could not be processed in a single disclosure run. The first run of the program assigned suppressions to the U.S., New England Region and the States. The rows of the disclosure tables referred to the U.S. total, the New England regional subtotal, and the fifty state totals; the columns referred to the different agricultural statistics. The six New England states summed to the New England subtotal.

After the suppression patterns for the states were finalized, we did a disclosure run for each of the fifty states to assign suppressions to the counties. The first row contained the state data and the other rows had the data for the counties; the columns again referred to the agriculture statistics. Since the state data in the first row had already been processed, all of the cells in the first row were frozen; that is, no new suppressions were added to the first row in these disclosure runs. Had these cells not been frozen, new state suppressions may have been added which would have required redoing the U.S. and state disclosure run.

4 Development and Analysis of the Linear Relations

The development of the linear relations was the most time consuming and labor intensive activity of the entire disclosure process. The linear relations describe the summation relationships that exist among the cells of the published census tables and are defined in terms of numeric cell identifiers.

The development of the linear relations could not begin until there was a stable draft of the census publication table shells. An inter-divisional team of agriculture analysts developed the publication table shells that described all aspects of the census tables including row and column descriptors, detailed tabulation instructions for each cell, and other information needed to program the summary and tabulation system. This effort took more than 18 months and produced 120 table shells.

Concurrent with this process was the creation of the numeric cell identifiers, referred to as matrix numbers, which were used to specify the linear relations. Each unique data cell was assigned a six-digit matrix number; if the cell appeared in multiple tables, it was assigned the same number in those tables. Similar agricultural commodities and characteristics were assigned numbers within a predetermined range to assist the analyst in their review of the disclosure data.

The linear relations were composed of a single matrix number that represented the summation cell followed by matrix numbers representing the component or interior cells that summed to the

summation cell. The agricultural data relations were more complex than in earlier applications of the disclosure methodology where the structure was defined by North American Industry Classification System (NAICS) codes which had an inherent logical additive structure. Many of the agriculture relations are unstructured. Groups of cells which summed were not necessarily contiguous within a table and sometimes occurred over multiple tables which made the identification process difficult and required detailed subject matter knowledge. Because of these complexities and the large size of the project, a system of linear relation programs was developed to assist the agricultural analysts in their development of the relations. This aspect of the system was crucial to the successful application of network flow methodology to the census.

The linear relation system took two forms: output to assist the analyst during the development of the relations and programs to analyze the relations for efficient structure and organization. Since the relations were developed by various analysts over an extended period of time, it was important to consolidate the updating of the relation file through a single source; an easier way to add new relations and to update existing relations accomplished this. Data products were created to assist the analyst in developing the relations. They included a complete set of table shells with each cell populated with its matrix number, a matrix number dictionary that gave the description and detailed tabulation instructions in terms of the census questionnaire item numbers, a complete listing by matrix number of all publication tables containing the cell, and a listing of all linear relations giving their matrix numbers, verbal descriptions and tabulation instructions. Another listing provided by matrix number the published table number in which it appears. This listing sometimes revealed a difference in table numbers for related relations that upon investigation revealed a missing relation. Another listing gave for each matrix number the relations that use each number which was an additional help in identifying inconsistent patterns in matrix numbers and relations that could indicate missing or incomplete relations. All of these files were used by the analysts to develop and validate the linear relations logic.

Two files were extremely useful while developing the relations. Analysts reviewed a file of matrix numbers and their verbal descriptions which were not used in any relation; these were called independents. In most situations, it was clear from the verbal description whether the matrix number had been omitted from a relation. The other file was a listing of matrix numbers that are in relations but are not in the file of published matrix numbers. This file should be empty; if it were not, it indicated a problem with the relations.

The relations were grouped into independent blocks of matrix numbers; matrix numbers composing relations within a block were not used in relations outside the block. Initially, this was done to speed computer processing by isolating the run to self-contained groups of matrix numbers; however, this was not needed for speed. The blocking did allow for more efficient testing of the programs by isolating runs to single blocks. The disclosure table output was also blocked to make it easier for analysts to review since similar types of data appeared together.

While these activities assisted the analyst in their development of a complete relations list, the following were some of the activities used to analyze the listings and create a file of relations that would work best with the disclosure program to reduce the potential for extraneous complements.

4.1 Restructuring the Relations

Even though technically correct when considered individually, the relations may not be in the best structure for disclosure processing when considered as a group. Identifying these relations and modifying their structure helped reduce the number of complementary suppressions.

Assume we have these two linear relations, where the numbers in the relations refer to matrix number mentioned earlier.

Relation 1: 10=20+30+40+50+60
Relation 2: 70=40+50+60

To reduce the number of complementary suppressions, it would be better to change the first relation in this way.

Relation 1a: 10=20+30+70

Suppose that matrix number 40 is a primary suppression; we may suppress matrix number 30 as a complement in the first relation. However, if the first relation is restructured as shown, there is no suppression to protect.

4.2 Reordering Relations

In certain situations, the relation processing order may contribute to the number of complements selected by the program. When a matrix number is a summation in one relation and an interior number in another relation, the relation in which it is interior should be processed first. As much as possible, these types of relations were identified and resorted in a logical Atop-down@ order to make sure the relations were ordered in this manner.

If the relations are not ordered in this way, it is easy to construct examples showing how the improper order of the relations may cause over-suppressions. Suppose that in relation 1a and 2 above the matrix numbers 30 and 50 are primary suppressions and the value of matrix number 20 is much larger than the value of matrix number 70. If relation 1a is processed first the matrix number 70 is chosen as a complement and no new complements are required in relation 2. However, if relation 2 had been processed first then either matrix number 40 or 60 would have been chosen and matrix number 70 would still have been chosen in relation 1a.

4.3 Combining Linear Relations

Some relations in the census have subcategories that are published together with an embedded subtotal. This occurs frequently with size breakouts such as for acreage when the higher acreage categories are subtotaled and all are published together in the same table. Under certain conditions this situation could create a disclosure. Suppose the subcategories are expressed as two relations and there is a one-respondent primary suppression in each relation. There is a risk

that if the subtotal is chosen as a complement then the two one-respondent suppressed cells are left unprotected since either respondent can calculate the others value. If this had been checked as a combined relation, another cell would have been selected as a complement since the program will not let one-respondent primary suppressions protect each other. To avoid this potential problem, relations of this type are combined by the program prior to disclosure processing.

4.4 Complete File of Linear Relations

At the conclusion of this process, 1885 relations had been identified; the largest containing 50 matrix numbers. When combined with the geographic dimension this relation created the largest single disclosure table of more than 12,000 cells. These relations generated more than 96,000 disclosure tables that contained more than 18 million cells awaiting disclosure analysis.

The relations file showed all of the relations grouped into homogeneous blocks for easier access and review by analysts. The matrix numbers in each relation were printed along with a complete verbal description and detailed tabulation instructions.

5 Enhanced Features of the System

5.1 Capacity and Cost Parameters

A major reason for the success of the disclosure application was the agriculture analysts' review of the suppression patterns, their feedback to the disclosure analysts and our ability to modify the system. When suppression patterns were unacceptable, rather than modifying the cost function in the minimum cost flow algorithm, we chose to modify its input parameters of capacity and cost. The capacity of a cell is the amount of protection a cell can give the initial suppression and the cost is the value that is assigned to each cell. The parameters were set so that the algorithm gave suppressions patterns that both adequately protected the initial suppressions while still preserving the most important statistics we wanted to publish.

Some of these modifications are discussed below.

5.2 Freeze/Maximum Capacity Program

The agriculture census data set was too large to process in a single computer run, so we had to break it into subsets and process them separately. The first computer run checked the U. S. and state data for disclosures and assigned suppressions to many of the state-level statistics. Then we did disclosure runs to assign suppressions to the data for the counties within each state. In these computer runs, we still had to check each linear relation separately and make sure a value suppressed in one relation had complementary suppressions chosen to protect it in every other relation in which it appeared.

Given the way the program works, this would sometime lead to the program wanting to suppress additional values at the state level. Of course, we could not allow new state suppressions to be chosen because we would then have to redo the U.S. and state disclosure run. To address this

problem, we decided to >freeze= all of the state-level suppressions after the U.S. and states disclosure run. Then we used a computer program to decide if any data for the counties must be frozen before the next set of disclosure runs were done.

For example, if a data cell for a state had not been suppressed and if it is equal in value to one of the counties in the state, then the county must also be frozen. If the newly frozen county cell is in a linear relation and if it is equal in value to the total of the relation, then the total must also be frozen. If the data for a county is close in value to an unsuppressed state cell, then the county data may be suppressed during the computer run but its >capacity= to protect another suppression must be limited. As a result, the computer program for doing all of this is quite complicated, but it must be done when processing large data sets that are divided into subsets.

5.3 Cost Function Options

There are two cost function options that could be selected during the parameter settings for the disclosure run. The standard option does not modify the cost to take into consideration other linear relations when processing a specific relation. This is the setting that has been used in previous applications of the methodology. The modified option adjusts the cost based on whether suppressing that cell would help or hurt other tables that contain that cell; in other words, it attempts to 'look ahead' to see if a suppression would cause problems in other column relations. For example, if the other table had no suppressions in the row that contains the cell, then suppressing that cell would hurt the other table because we would then have to suppress another cell in that row; a much higher cost is given in this situation. If the other table had only one cell suppressed in that row, then suppressing the new cell would help the other table and a lower cost is assigned.

5.4 Preference Codes & Cost Adjustment Factors

We used the preference codes and cost adjustment factors to make the suppression patterns more to our liking. The preference codes were used to identify unpublished cells that can be suppressed with no harm to the publications or to identify cells that must not be suppressed, usually because they are frozen. The cost adjustment factors were used to assign relative levels of importance to cells based on the analysts' recommendations. The costs for cells of lower importance were decreased by incremental factors so that the cells would be more likely chosen as complementary suppressions. The costs for important cells were increased so that they would be less likely to be chosen. Sometimes analysts requested that the summation be suppressed before the interior cells. This was done by decreasing their cost relative to the interior cells= costs.

During the disclosure review analysts identified commodities that were of such importance at a specific geographic level that they should not be suppressed as complements if other patterns of complements could adequately protect the initial suppression. These commodities included such items as Valencia oranges in Florida or grapes and olives in California. The cost for these commodities was increased by a factor to discourage their selection as complements; however, their selection was not strictly forbidden in a situation where there were not other potential complements available. Numerous commodities were processed with this technique.

For some disclosure tables, analysts identified cells which were selected as complements that, because of their size relative to their row total, were important for that row. To decrease the chance that these types of cells were selected as complements their costs were increased by a factor if their disclosure table had more than three columns and the value of the cell was more than 80 percent of the row total.

6 Diagnostic Tools

The diagnostic files gave the analysts the ability to review the suppression patterns and to identify the situations that resulted in the modifications that have been described above. These files were developed to provide detailed information on the suppressions and why they were selected. The two major files were the disclosure table file and the suppression pattern file. The table file was used by both agricultural and disclosure analysts while the suppression pattern file was used mainly by the disclosure analysts. Both were used extensively throughout the disclosure process and were invaluable when explaining the reason why a particular suppression occurred.

The table file gave the disclosure tables for each of the 1885 relations and identified all the suppressed cells within each table. The table files were created for the US/State disclosure run and for each of the 50 State/County runs. The rows of the disclosure tables were the U.S./State or State/County geographies depending on the disclosure run and the columns were the matrix numbers for the specific relation. Each table contained the relations number, block number and verbal descriptions for the matrix numbers composing the relation. The disclosure tables were sorted by block so that similar agricultural commodities were grouped together for the analysts to review. The core of each table contained the published census values followed by the suppression flags which identified the primary and complementary suppressions.

Many relations shared the same matrix numbers due to the cell overlap between publication tables and the complexity within those tables. The table file showed all suppressions for the matrix numbers composing the specific relation regardless of when the suppression was selected. To facilitate the analytical review, the suppressions were coded to indicate whether they were selected when the current relation was processed, during processing of an earlier relation, a later relation or during backtracking. This was of particular help to the analyst as they reviewed the patterns. Prior to implementing this coding scheme we would receive questions about why such a large cell was chosen to protect a very small cell; when in fact, the large cell was required for protection of an initial suppression in another relation that shared a matrix number with the currently processed relation. The coding scheme increased the confidence among analysts that the system was operating properly.

Additional codes were included adjacent to the suppression flags to indicate the preference code or cost adjustment factor. Since these affect the cost of suppressing a cell, knowing these codes helped analysts understand why a particular cell was chosen as a complementary suppression. Each table file provided counts for the number and value of primary and complementary suppressions. Separate counts were given for primary suppressions with one or two respondents to help analysts determine whether the data were being sliced too thinly.

The suppression pattern file was used jointly with the disclosure table file and mainly by a disclosure analyst to determine the suppression patterns for complex situations. For each initial suppression the file lists all cells in the suppression pattern. The data included on the pattern file allows the analyst, even in the most complex situations, to reconstruct the actual suppression pattern. The selection of each complementary suppression can be justified. One of the most important pieces of information gave the suppression pattern in which the cell received its maximum protection or carried its greatest flow. This information was used to address the analysts' questions on why the protection was so high on an initial suppression. By examining the suppression pattern, one can determine which initial suppression pattern gave that protection to the cell. The file also gave the number of units flowing though the complementary suppression. This information was used to determine the importance of the cell relative to other cells in the pattern in protecting the initial suppression.

The joint use of these two files was essential to the success of the disclosure project. The ability to quickly respond to the inquiries from the agricultural analysts gave them confidence in the disclosure system. The analyst may not have always liked our answers but they could see why it happened. On numerous occasions their use of these files identified issues that resulted in a parameter modification or a change in the program. The benefit of this interaction to the success of this project cannot be overstated.

7 Performance Issues

The disclosure system is written in FORTRAN. During census production the disclosure system ran on an IBM R50 UNIX server with 1997 architecture. To run the disclosure system for the entire census which consisted of more than 18 million cells took 3.5 hours. We were able to run the entire census twice each day and review diagnostics which were extremely beneficial during the development and analytical review process. If we were testing specific aspects of the program we could isolate the run to specific relations or blocks and reduce the run time to minutes.

After the census we moved the disclosure system to an IBM P690 UNIX server with 32 processors and 132 gigabytes of memory. Because of the greater power of this machine, the entire census disclosure system ran in 20 minutes.

8 Future Work

All of the FORTRAN programs are being converted to SAS to make it easier for the agency to support the disclosure system in the future. The two-dimensional programs have been completed and run on the entire census with identical results as the FORTRAN programs. As expected, the SAS version runs slower taking approximately 10 times longer on the IBM P690 machine than the original FORTRAN version. However, this is approximately the run time experienced on the slower IBM R50 machine for the FORTRAN version during the census and is deemed to be acceptable. Currently efforts are continuing to convert the remaining programs to SAS.

References

Jewett, R.S. (2003) "Developing the Linear Relations", *unpublished internal manuscript*, Washington, D.C.: USDA, National Agricultural Statistics Service, Census and Survey Division.

Jewett, R.S. (2004) "Disclosure Analysis for the 2002 Agriculture Census", *unpublished internal manuscript*, Washington, D. C.: USDA, National Agricultural Statistics Service, Census and Survey Division.

Jewett, R.S. (2004) "Description of the 2002 Agriculture Census Disclosure System", *unpublished internal manuscript*, Washington, D. C.: USDA, National Agricultural Statistics Service, Census and Survey Division.

National Agricultural Statistics Service, USDA (2004) "2002 Census of Agriculture, Summary and State Data, Volume 1, Geographic Area Series".