STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

COMMISSION OF THE EUROPEAN COMMUNITIES

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

Joint ECE/Eurostat Work Session on Statistical Data Confidentiality
(Skopje, The former Yugoslav Republic of Macedonia, 14-16 March 2001)

Working Paper No. 8
English only

Topic I: Application of statistical disclosure control methodology and software in business statistics and social and demographic statistics

APPLICATION OF STATISTICAL DISCLOSURE CONTROL METHODOLOGY TO AGRICULTURAL DATA

Contributed paper

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

Abstract: With the transfer of the Census of Agriculture to the National Agricultural Statistics Service within the Department of Agriculture it was recognized that the addition of the census and its follow-on surveys to the agency's existing programs brought different issues and concerns regarding the confidentiality of data. The agency conducted a review of the multiple policies for disclosure control methodology to ensure the confidentiality of individual respondent data. This paper will discuss the complexities of disclosure in the realm of the census of agriculture and its follow-on surveys. The paper will deal mainly with disclosure methodology for primary and complementary suppressions for magnitude data, these are generally the data in the tables that are published. A discussion of the methodology used in previous censuses and the reasons contributing to the decision to move to a network/minimum cost flow system will be presented.

I. INTRODUCTION

1. This paper gives an overview of the application of disclosure methodology to census of agriculture data over the last several censuses[2] - what we have done in the past and where we are heading in the future. Special attention is given to the methods required to apply disclosure limitation methodology to census of agriculture data.

II. CENSUS OF AGRICULTURE

2. The census of agriculture is taken to obtain agricultural statistics for each county, State, and the Nation. It is conducted on a 5-year cycle collecting data for years ending in 2 and 7. The data products from the census of agriculture are available on three different media: printed report, CD-ROM, and the Internet. The printed reports comprise 51 volumes - one for each state and a national summary. Within each

[1] Prepared by Robert T. Smith, Jr. Census and Surveys Division.

[2] Agricultural censuses for 1992 and earlier were conducted by the U.S. Department of Commerce, Bureau of the Census. However, the 1997 Appropriations Act transferred the responsibility from the Bureau of the Census to the U.S. Department of Agriculture (USDA), National Agricultural Statistics Service (NASS). The 1997 Census of Agriculture is the first census conducted by NASS.

state volume, there are 52 state level tables and 39 county level tables; these 51 volumes comprise approximately 15 million tabular cells of data.

III. HISTORY OF DISCLOSURE METHODOLOGY FOR THE CENSUS OF AGRICULTURE

3. The confidentiality of the data reported to the census of agriculture is protected by law and can be used only for statistical purposes. No data are published that would disclose the operations of an individual farm. Specifically, U.S. Code, Title 7, Section 2276 states "... [that no person] may use such information for a purpose other than the development or reporting of aggregate data in a manner such that the identity of the person who supplied such information is not discernible and is not material to the intended uses of such information; or disclose such information to the public, unless such information has been transformed into a statistical or aggregate form that does not allow the identification of the person who supplied particular information."

4. The censuses of agriculture have used cell suppression to protect their published tabular data. The manner in which this has been done has evolved over the course of the censuses. Prior to 1978, the cell suppressions were performed manually by analysts using a technique called “nearest-smallest method.” For the 1978 Census, a portion of the manual procedure was automated; however, most of the complementary suppression still remained a manual application. It was not until the 1982 Census that the disclosure system was fully automated. Minor revisions were made to the system for the 1987 Census.

5. Prior to the 1992 Census, research was conducted to determine the optimal approach for disclosure. At that time, due to the uniqueness of some of the agricultural requirements, research indicated that a customization approach was required to address these issues. For both the 1992 and 1997 Censuses a rule-based system was used to identify complementary suppressions. The rule-based system was a non-mathematical system that chose a complement in each row and column containing a primary suppression. The system was essentially a very large and complex bookkeeping system. It had the ability to look ahead within groups of relationships within the aggregation structure so that a minimum number of complementary cells were chosen; however, it did not control on the value of those cells that were selected as complements. This system could minimize the number of complementary suppressions; however, this was at the expense of the total data value suppressed. In selecting a minimum number of complementary cells, it could over-suppress on data value. To ensure that the data were fully and adequately protected, required additional methodology which involved the application of additional programs to verify the levels and patterns of protection. This added additional processing difficulties at a critical point in the census cycle.

IV. DISCLOSURE METHODOLOGY FOR THE 2002 CENSUS OF AGRICULTURE

6. The disclosure system being developed for the 2002 Census of Agriculture will maintain the methodology used for frequency data and primary cell suppression methodology for magnitude data; however, the complementary cell suppression methodology will change. The network method will be used to select the complementary cell suppressions.

7. Historically, the census of agriculture does not suppress farm count; only the aggregated data values associated with those farm counts. In other words, the number of farms reporting an item is not considered a release of confidential information and is provided even though other information may be withheld. If a tabular cell represents an aggregation of less than three farms, the cell value is suppressed. If the tabular cell represents three or more farms, the primary disclosure methodology is applied to determine whether the data

cell is sensitive; that is, whether, due to the distribution of its reported data, it is at risk of disclosing individual data about one of its contributors. The standard disclosure rule used to identify whether a cell is a disclosure risk is the p-percent rule. This rule identifies a cell as a disclosure risk if the second largest respondent in the cell can use the cell value to estimate the largest respondent's value to within p-percent. Such a cell is considered a primary disclosure and is suppressed from the publication. This methodology has been used in the last several censuses to identify tabular cells that are at risk for disclosure.

8. If there were no additional information on the suppressed cell, external to the cell itself, then the data associated with that cell would be protected; however, this is never the case. Because of the many relationships that exist within the census of agriculture's data structure, complementary suppressions are necessary to adequately protect the data cells identified as sensitive either because there were fewer than three farms contributing to the cell or the primary disclosure methodology indicated that the cell was at risk of disclosure. For the 2002 Census of Agriculture, NASS has adopted a network/minimum cost flow methodology to identify the complementary suppressions that are needed to adequately protect these primary suppression cells in the census publication tables.

9. This methodology converts a two-dimensional table to a mathematical network. The arcs in the network represent the cells in the table. The nodes in the network, where the arcs come together, represent the additive relationships among the cells in the table. A closed path of cells in the table corresponds to a set of connected arcs in the network. The minimum cost flow methodology is used to select the complementary suppressions to protect the primary suppression in the table. The methodology selects from among all closed paths of cells in the table (connected arcs in the network) that contain the primary suppression, the one path that minimizes the total cost. The cost of a path is defined to be the sum of the data values of the tabular cells that correspond to the arcs in the network. Every arc in the selected path corresponds to a cell that should be suppressed as a complementary suppression in the table.

V. AGGREGATION AND DATA STRUCTURE

10. The purpose of this section is to give a general description of the framework within which the disclosure system functions for the census of agriculture. Implementing the disclosure methodology for an operation as large as the census requires that it be closely integrated with the tabulation and publication systems. The tabulation system for the census will produce approximately 5000 publication tables. The 1997 Census publications contained approximately 15 million cells; the 2002 Census is expected to be at least as large.

11. The disclosure system is part of a larger system that includes the summary (tabulation) module and the publication module. Figure 1 shows the flow of the data among these modules. A table summary module exists for each published table and uses a standardized method for summarizing the edited record level data for each tabular cell directly into the table database which is input to the table publication module. The table module also writes a record to the master matrix database for each published cell in the table. The master matrices for all the tables are then joined to create the concatenated master matrix database.

12. The master matrix database is one of the critical components of the disclosure system. It is the depository of information on all tabulated cells. It has three primary purposes: it is needed for the disclosure module, it is used to verify that cells appearing in multiple tables have the same value and same disclosure status, and to the extent that subsequent publications use previously tabulated data, data can be extracted from it for those publications. The concatenated master matrix database consists of a record for each cell in each table. Each record in the master matrix database includes: master matrix number, geographic identifier,

table/row/column identifier, published cell value, largest record level value, second largest record level value, (these last two items are required for primary disclosure), and the disclosure flag (added by disclosure module). If a cell appears in multiple tables then it also appears within the master matrix database the same number of times. The master matrix number (M), geographic identifier (G) and table/row/column (T,R,C) identifiers are attached to each record in the database. The master matrix number and the geographic identifier are unique to the summarized cell. If the summarized cell appears in multiple tables it will be assigned the same (M,G) index in the database; however, the (T,R,C) index for that cell will change with the tables. The unduplicated master matrix database is created by removing the (T,R,C) index and collapsing on the (M,G) index. The resulting database is input to the disclosure module.

13. The results of the complementary disclosure module (i.e., the cells composing the complementary suppression pattern) are posted to the concatenated master matrix database. The disclosure status of the tabular cells are then input to the publication module. In this manner, when cells appear in multiple tables their disclosure status is consistently applied across all tables.

14. The primary disclosure methodology operates independently within each data cell of the unduplicated master matrix database and is sensitive to the relationships that exist among the reported data within each cell. On the other hand, the complementary disclosure methodology operates among the data relationships that exist among all the cells. The framework is particularly important to the understanding of the relationships within the data structure that are an integral part of complementary disclosure methodology.

15. Data relationships exist not only within these published tables but across many of them. The tables on which the network/minimum cost flow methodology is applied are not the actual publication tables, but rather, the tables that are defined by the linear relationships that exist within the aggregation structure. The tabulation cells within the master matrix are linked by the relationships defined within the aggregation equations or linear relationships. In the previous census of agriculture, there were approximately 1700 linear relationships contained within the aggregation structure; the next census will contain even more. After the primary disclosures are identified, the complementary disclosure methodology is run on each of the 1700+ relationships.

16. It is within this system of relationships that the complementary disclosure system operates. These aggregation equations represent all the relations that exist for all data items within the census. It is for this reason that the complementary disclosure methodology is applied on the tables generated by those relationships across the appropriate geographical areas. This process can also be thought of slightly differently, and this alternate description is more applicable when thinking of the disclosure module and how the suppressions that are identified within the linear relationships are applied to the actual publication tables. A summarized cell exists within two spaces. The first, a two dimensional space, is defined by the master matrix number (M) and the geographic identifier (G). Each cell appears only once in this space. The complementary disclosure module will operate in this (M,G) space. A summarized cell also exists within a three dimensional space which is defined by its table/row/column (T,R,C) location. Cells in this space are not necessarily unique; that is, if a cell appears in multiple tables then it will be in this space multiple times. The tabulation module operates within this space.

17. The critical point in this concept is the one-to-many mapping that exists between the two-dimensional (M,G) space and the three-dimensional (T,R,C) space. It is this mapping that allows the complementary suppressions (as well as the primary suppressions) to be posted in the appropriate tables. In other words, when a complementary suppression pattern is identified, it is this mapping relationship that will ensure that the cells composing the suppression pattern are suppressed in all tables in which they appear. The input to this process starts at the earliest point in the tabulation system; that is, at the time when the tables are

defined, the table specifications are developed and the links are established between the tabular cells to master matrix database. This information is used to establish this linkage between the spaces.

18. An example helps to illustrate the procedure. Every data aggregation that is required by a publication exists within this aggregation structure. This aggregation structure also carries information on the disclosure status of the cells. Every cell of this structure can be described by a unique identifier. The linear relationships that exist within this structure can then be described by a system of linear equations of the form $a = b + c + ... + x$. When disclosure methodology is applied to this aggregation structure it interacts at two levels. First, primary disclosure techniques are applied to the individual cell level, that is, at the $a, b, c, ..., x$ levels. However, when we apply the complementary disclosure techniques not only are the data from the individual cells required but also the linear equations that define the relationships among the cells. Complementary disclosure techniques must be applied to each one of these relationships and at the same time take into account the geographical structure that exist within the aggregation structure.

19. Figure 2, Part 1 shows a representation of the linear relationships that exist within the census data structure. For example, the item *Harvested Cropland - Acres* is the sum of the following acreage categories: 1 - 9, 10 - 19, 20 - 29, 30 - 49, 50 - 99, 100 - 199, 200 - 499, 500 - 999, and 1000+ cropland harvested acres. This is expressed as the linear relationship R3, *Harvested Cropland - Acres*, by the equation 407400 = 407600 + 407800 + 407900 + 408000 + 408100 + 408400 + 408800 + 409100 + 409200 where the numbers refer to location within the master matrix data base of the specific data value.

20. Groups of these relationships (Part 2) are defined which are mutually exclusive relative to their master matrix numbers within the relationships composing the group. A table is created for each relationship within each group. The sum and addends of the relationships define the columns and the geographic variable define the rows of each table. The cell values and their primary disclosure statuses are extracted from the unduplicated master matrix database. Note that the primary suppressions were determined at the time that the master matrix database was populated and those actions are now carried to these tables where the complementary disclosure methodology suppresses a pattern of cells to protect the primaries.

21. The state of Texas contains 254 counties; relationship R3, *Harvested Cropland - Acres*, generates a 254 x 10 cell table (Step 3). This table contains a primary suppression identified from the earlier operation (Geo 2 x 407900). The network methodology converts this table to a mathematical network and the minimum cost flow methodology then determines from among all possible closed paths containing the primary suppression the one that minimizes cost; the cost being defined as the sum of the cells= data values. It should be noted that data analysts may have determined that certain cells not be included in a complementary suppression pattern due to their importance within the geographic area. These are referred to as preferences and are handled by setting very high cost on those cells so that the patterns including them have a high cost and therefore are never selected. These results are then posted to the master matrix database and then the next relationship (R21) within that Group 1 is processed.

22. This procedure continues until all the relationships in all the groups have been processed. The cells composing the suppression patterns are posted to the master matrix data base for each relationship. When publication tables are generated from this aggregation structure, the disclosure status of a cell is then consistently applied throughout the multiple tables that contain that cell value. This framework helps to ensure, once a cell is identified as either a primary or complementary suppression, the suppression is consistently applied across all tables in the current publication or any subsequent publications.

References

1997 Census of Agriculture, Volume 1, Geographic Area Series. U.S. Department of Agriculture, National Agricultural Statistics Service, March 1999.

Arends, Bill and Robert T. Smith et al. *PRISM Disclosure Subteam, Report to the Strategic Planning Council.* Internal Report, National Agricultural Statistics Service, Department of Agriculture, December 1999.

Jewett, Robert. *Disclosure Analysis for the 1992 Economic Census.* Internal Report. U.S. Bureau of the Census, U.S. Department of Commerce, 1993.

Smith, Jr., Robert T. and Tom Birkett. *Tabulation System 2002*. Internal Memorandum. National Agricultural Statistics Service, Department of Agriculture, August 2000.

Sullivan, Colleen M. and Laura Zayatz. *A Network Flow Disclosure Avoidance System Applied to the Census of Agriculture.* Internal Report. U.S. Bureau of the Census, U.S. Department of Commerce, September 1991.

FIGURE 1

Tabulation, Disclosure & Publication Systems

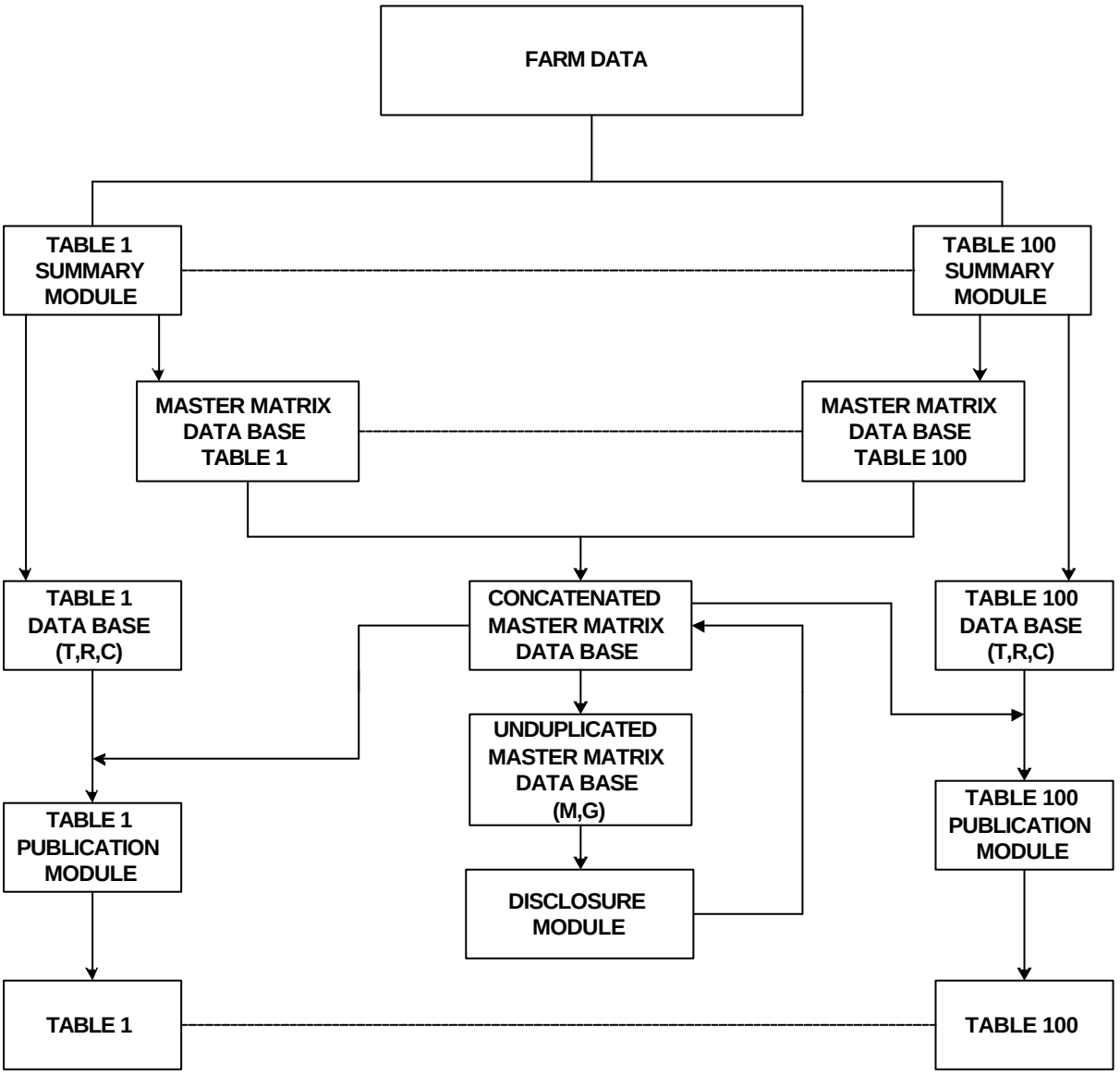

FIGURE 2

1. Linear Relationships

No. **Linear Relationships** (Aggregation Cell = Summation of Aggregation Cells)

R1 397800 = 398500 + 399300 + 399900 + 400500

R2 345000 = 345200 + 345300 + 345400 + 345500 + 345700 + 345900 + 346000 + 346300 + 346400 + 346600 + 346800 + 346900

***R3 407400 = 407600 +407800 +407900 +408000 + 408100 + 408400 + 408800 + 409100 + 409200**

R4 347200 = 347300 + 348100 + 348200 + 348300+ 348400 + 348500 + 348600 + 366300

R5 345000 = 347300 + 348100 + 348200 + 348300 + 348400 + 348500 + 348600 + 366300 + 366500 +370700 + 366700 + 370800 + 373900 + 374000

-
-
-

R1700 (1700 Relationships)

2. Groups of Relationships Representing Mutually Exclusive Master Matrix Numbers

Group 1	**Group I**	**Group N**
***R3**	R50	R1100
R21	R52	R1200
R1234	R65	
	R75	
	R78	

3. Table on Which Network/Minimum Cost Flow Is Applied

*** Relation R3: Harvested Cropland - Acres** (254 x 10 = 2540 cells)

		M									
		407400	**407600**	**407800**	**407900**	**408000**	**408100**	**408400**	**408800**	**409100**	**409200**
	Geo 1	X	X	X	X	X	X	X	X	X	X
	Geo 2	X	X	X	**P**	X	X	X	X	X	X
G	**.**										
	.										
l.											
	Geo 254	X	X	X	X	X	X	X	X	X	X

P = primary suppression from earlier phase

4. Converted to mathematical network.

5. Minimum cost flow methodology applied to determined complementary suppression pattern.

6. Cells composing the selected complementary suppression pattern posted to master matrix database.

7. Publication module.