

# EIA DISCLOSURE LIMITATION TEAM'S REPORT

## Overview

The Energy Information Administration's (EIA) Disclosure Limitation Team was formed at the direction of EIA's Strategic Goal 4 Committee. The Team was directed to recommend standard software that could be used across EIA to protect the confidentiality of survey data used to generate statistical tables.<sup>1</sup> EIA collects a variety of energy data under a promise of confidentiality to its respondents. Program offices apply sensitivity rules to protect tabular data from disclosing identity or attribute information about the respondents. EIA has established a statistical standard regarding nondisclosure. (<http://www.eia.doe.gov/smg/Standard.pdf>) Standard 2002-22, "Nondisclosure of Company Identifiable Data in Aggregate Cells," contains the procedures and policies to identify sensitive (i.e., *primary*) cell values that need protection. Subsequently, disclosure procedures, such as Tau Argus and DiAna, are used to identify and suppress values in cells for *secondary* confidentiality, values that otherwise could have been used to derive suppressed sensitive values in the tables.

## Preliminary Report

The Team considered the following factors to evaluate four software packages:

1. Tau-Argus software package;
2. Automated Cell Suppression System (ACS);
3. Disclosure Analysis (DiAna), a PC version of the mainframe software developed by the Census Bureau; and
4. Statistical Disclosure Limitation (SDL) Software which is a prototype applying the controlled tabular adjustment method, written in SPSS and developed by the Bureau of Transportation Statistics).

Non-monetary factors considered in the preliminary evaluation included:

- 1) the availability of documentation, technical support, and training
- 2) computer compatibility/platforms, and requirements for operating the software
- 3) ease of use and whether the software is currently used by other agencies.

This testing was focused only on suppression and the Tau-Argus and DiAna software systems. These two software packages were chosen because they had the best documentation out of the available software, both are compatible with EIA computer systems, and both rank high on ease of use. Among the considerations were that Tau-Argus required the additional purchase of one of two specific linear programming software packages and is limited to three dimensions. There are some four dimensional tables currently released by EIA program offices. DiAna has no licensing costs and can process and apply disclosure limitation methods on tables with four dimensions or higher. CNEAF currently uses a DOS PC version of the DiAna software for applying disclosure protection.

---

<sup>1</sup> The Team would like to express its gratitude to the Office of Oil and Gas, Petroleum Division for providing material and substantive assistance with testing procedures.

The Team further considered the following elements to evaluate Tau-Argus and DiAna software systems. By order of importance they were:

1. Use by other statistical agency (domestic or international)
2. Documentation and supplemental materials
3. Training support / training availability
4. Ease of Use
5. Computer compatibility / platforms, requirements, components
6. Dimensions of table handled, customizable, open source,
7. Critical reviews

## Discussion of Evaluation Procedure

Both Tau-Argus and DiAna software apply network flow theory to develop cell suppression patterns. The network flow model for cell suppression is self-auditing for two-dimensional tables in which there is a hierarchy in one dimension. The network flow method is not self-auditing for two dimensional tables with a hierarchical variables structure in both the row and column, and it is not self-auditing for three dimensional or higher dimensional tables that contain a hierarchical structure. DiAna uses a heuristic approach to apply network theory to develop cell suppression pattern on three and four dimensional tables. This necessitates auditing of the cell suppression pattern developed by DiAna for three and four dimensional prior to release in final publication. Disclosure Audit Software<sup>2</sup> (DAS) developed by Federal Committee on Statistical Methodology at the Office of Management and Budget, in conjunction with Confidentiality and Data Access Committee (CDAC), may, therefore, be used to verify the compliancy of a cell suppression pattern developed by DiAna for three and four dimensional tables, with respect to suppression rules applied by individual offices.

The Team used a three dimensional data set to evaluate both programs. The P% rule was used to evaluate the programs. Both programs were evaluated using the same protection criteria, i.e. percentage in the P% rule.

## Evaluation Results

The DAS program was used to audit the suppression patterns from both DiAna and Tau-Argus. Table 1 summarizes the results of the audits. Tau-Argus provided adequate protection to all of the primary or sensitive cells. DiAna had four (4) primary cells without adequate protection. Those four cells all represented the same entity in various cells (totals, subtotals, etc) within the table. The Team is not concerned by complimentary cells that fail the audit as long as they provided adequate protection for the primary cell. Tau-Argus suppressed fewer cell than DiAna. However, Tau-Argus tended to suppress more marginal totals than DiAna.

---

<sup>2</sup> EIA led an inter-agency project to develop DAS; key contributors included: Bureau of Labor Statistics, Bureau of Economic Analysis, Bureau of the Census, National Center for Education Statistics, Internal Revenue Service, and National Science Foundation.

**Table 1-Summary Audit Results**

Program	Cells Suppressed	Primary Suppressed	Cells Failing Audit	Primary Cells Failing Audit
DiAna	161	49	62	4
Tau Argus	228	46	12	0

DiAna suppressed fewer cell Tau-Argus. However, Tau-Argus tended to suppress more marginal totals as complimentary cells than DiAna, see Table 2. Both programs were evaluated using the default modes. Future work will experiment using different mode setting to minimize the number of totals used as complimentary cell in Tau Argus and increase the number used in DiAna, and evaluate the effectiveness of the modifications.

**Table 2- Number of Table Total Selected as Complimentary Cell**

Program	Comp. Totals Suppressed		
	Dim 1	Dim 2	Dim 3
DiAna	27	26	12
Tau Argus	45	71	62

## Program Details: DiAna

DiAna requires specific file formats that are not flexible. To run DiAna the user must specify the table with a series of three (3) descriptive files that explain the layout of the table and the hierarchy within the table. These files explain the layout of the table that is to be protected in terms of valid rows and columns, and row and column relationships. Once these files are created they can be reused and do not need to be created again as long as the table structure does not change.

The input file is the fourth file used in DiAna and it contains a record for every non-zero entry in the published table. The record consists of the descriptive data on the row, column, and level (for 3-D tables) of the entry, the aggregated total of the respondents that contribute to that entry, and the ID and values of the two largest respondents. This work was done using Microsoft Excel pivot tables to manipulate and aggregate the data. This file can be difficult and time-consuming to prepare, but the process may be automated if a cross-walk table, or similar mechanism, is created that identifies the row and column where data appear in the published table. The structure of this file must be precise and is inflexible. The program does not require the records to be in any specific order as long as the row and column relationships are defined correctly in the descriptive files.

Specifying the layout of the table involves more than just recreating the structure of the published table. Two-dimensional tables in which there is hierarchy in only one dimension are more straight forward than three-dimensional tables; however, the organization and structure of the inputs can still differ significantly from that of the published table. Processing a table as a 3-D table will require coding the data to look entirely different than in a two dimensionally represented published table. The data is organized along dimensions and levels and not rows and columns (in terms of rows and columns in the table). In a 3-D table layout, the rows and columns essentially represent different dimensions. With 3-D and 2-D tables that are coded differently than the published table, a mapping scheme will need to be developed to incorporate the suppression pattern into the publication table. Changing the layout of the table also means changing all of the files, including the data file which must be completely recoded.

There is a fifth file that reads the input files (the programs assumes the file is in the same location as the executable) and parameter values to use for applying the P% rule; *the p-percent rule declares disclosure whenever the cell value minus the contribution of the second largest contributor is less than (100+p)-percent of the contribution of the largest contributor.* This file also defines for DiAna how many dimensions and levels to expect and any other diagnostic files to print. To run the same table at multiple P% points, the user only has to change one line in this file.

DiAna is set up to run as a stand alone executable. If all of the input files are present and formatted correctly, all the user has to do is open the executable and enter the name of the file mentioned in the previous paragraph.

The output from DiAna is the same as the input file, except the entries that have been identified for suppression (either primary or complimentary) will have additional information added to their record. If the input file is coded to match the structure (rows and columns) of the published

table then it is relatively easy to convert the output into a format that resembles the publication; however, the output for the publication layout is not particularly useful as input for the audit software (DAS). DAS requires that the data be arranged by hierarchical structure, with the total of each level first and the sum of the components afterwards. The input data file used by DiAna can be coded this way also, but the results will need to be mapped back to the publication layout. DAS also needs a value (zero or non-zero) for every possible entry. DiAna does not require zeros for empty or blank entries in the table but the DAS does require a zero for any empty or blank table cells. Therefore the user must recode the output data from DiAna with zeros in order to run the audit program on the output file.

DiAna will also generate several other diagnostic files that function like logs that go into great detail on how the program interpreted the input files and what it did during the run. These files are useful to explain how, when and why each cell was chosen for suppression.

### Program Details: Tau-Argus

Tau-Argus is menu driven program that does not require specific file formats. The user makes a series of selections through a number of dropdown menus and check boxes to specify the format of the file, where the variables are located within the character strand, and if the variable is an explanatory variable, response variable, weighted variable, and other descriptive information on the structure of the table. . As with DiAna, Tau Argus requires files that specify hierarchical order and valid characters (row and columns in DiAna). Unlike DiAna, Tau Argus needs a separate file for each variable to define the valid characters and hierarchy (if present) for that variable. These two files have a specific format that they must use and like DiAna, once they are created, they can be reused.

Tau Argus aggregates the data according to the user's specifications. There is no need to aggregate and code the data before creating the input data file. However, it can be advantageous to recode some data elements if the data set has a hierarchical structure that does not appear in the published table. For example if the data set has product codes for low sulfur diesel (11), high sulfur diesel (12), and distillate (20), but the table only has columns for total diesel and distillate, then assigning product codes 11 and 12 to one product code such as (10) in the input file before it is read by Tau Argus will simplify the users' workload for preparing an output table later on.

After the data is specified, the user indicates what variables, suppression rules (Dominance or P%) and parameters to use for that particular table. The user provides this information through one window with point and click navigation. After the table is selected the program will show the user how it interpreted the specified table. After that table is displayed, the user then specifies what suppression module it would like Tau Argus to use to evaluate the table.

Tau Argus has four modules/programs (Hypercube, Modular, Optimal, Network) that can be used to determine suppression patterns. In previous versions of Tau Argus two modules Hypercube and Network flow used optimizers that were internal to Tau Argus to generate suppression patterns. The Modular and Optimal options require the purchase of one of two external commercial programs/solvers, Cplex or Xpress to generate a suppression pattern. The newest version requires an external solver for the Network module as well. The user's manual for the old version does recommend using one of the external solvers for 3-D tables.

Table 44 was processed with the Network module using Argus' internal solver in the older version. A temporary license of Xpress was negotiated and attempts were made to utilize the Xpress enabled modules. However, the tests were not successful that used the Xpress optimizer to interface with Tau. Attempts were made to solicit support from both Statistics Netherlands and Xpress, but neither was able to offer any useful customer support.

Tau Argus gives the user the ability to format the output in multiple user-defined two-dimensional tables that are stacked/layered/arranged by the third dimension in a CSV file. The tables have the user defined hierarchical structure described in the input file. These tables also have dashes for cells without data (zero data). The stacked/layered format of this file structure is extremely difficult to use for when working with 3-D tables. For simple 2-D tables that are arranged according to a hierarchical structure, this table can be used for publication.

The user can also choose to have the table outputted as a CSV file that will have a record (row of data) for every entry in the table. The entry will have columns that identify each dimension (user specifies the order of the dimensions), the aggregated total (dashes for zero data), and suppression code. Tau Argus has codes for unsuppressed, primary suppression, complimentary suppression, and zero data. The CSV file would make using the audit software, DAS, easier. Argus will collapse the hierarchal structure into one line regardless of the levels. The audit software requires a separate line/column for each level within the hierarchy. Therefore the output needs additional columns externally added to the file before the audit can be performed. For both DiAna and Tau Argus, the output needs to be rearranged or mapped back to the publication layout.

## **Future Work and Summary**

The next steps in this project will evaluate the program with different user settings. Tau-Argus has the capability to utilize other suppression methodologies than Network Flow to determine a suppression pattern. DiAna does not have this capability. Both programs will be evaluated using different setting to either increase or decrease the probability of totals being selected as complimentary cells for suppression. Future work will also evaluate both programs using different tolerances.

The results summarized in this paper were evaluated running both programs using in default mode with comparable tolerances for the same suppression methodology (network flow). Tau-Argus is has a user friendly interface and adequately protected all of the primary cells that needed protection. However, Tau-Argus in it default mode tended to suppress more complimentary cell and target marginal totals. Also because Tau-Argus as tested is menu driven, it would also be more difficult to run in a production cycle. The user interface is friendly, but not easily automated.

DiAna does not have a user friendly interface, however the program can be automated easily and therefore integrated into a production cycle. The program does require very specific file formats that require experience to conceptualize and generate. DiAna tended to suppress fewer complimentary cells and fewer marginal totals, however, the program did not provide adequate

protection for one entity in evaluation using the default mode and comparable tolerances to the Tau-Argus.