Major Initiatives to Improve the U.S. Energy Information Administration's Statistical Disclosure Limitation Procedures

David Kinyon[1]
[1]U.S. Energy Information Administration,
1000 Independence Ave. SW, Washington, DC 20585

Abstract
We discuss two major initiatives that the U.S. Energy Information Administration (EIA) is leading to improve our statistical disclosure limitation procedures: (1) modernizing our cell suppression software and (2) establishing a formal procedure for conducting disclosure reviews of our data products. To modernize our cell suppression software, we first acquired the research prototype of the Census Bureau's linear programming (LP) cell suppression software in July 2017. Since February 2019, we have been successfully using our modified version of the Census Bureau's LP prototype in production to perform disclosure analysis for year-end estimates of U.S. photovoltaic module shipments by state and territory. In February 2021, we successfully tested our modified version of the Census Bureau's LP prototype in our modernized processing environment, and we plan to use this version of the prototype in our new production environment. In a related effort, we plan to form a Disclosure Review Board and incorporate formal disclosure reviews as part of our production processes to ensure that we consistently apply appropriate statistical disclosure limitation techniques to our data products.

Key Words: statistical disclosure limitation, cell suppression, linear programming, energy surveys, Disclosure Review Board

Introduction

In March 2017, the U.S. Energy Administration (EIA) began the process of modernizing the cell suppression software used for disclosure limitation by developing a project plan. This project was part of a five-year plan for improving the statistical methodologies that are used for energy programs conducted by EIA's Office of Energy Statistics. The modernization of our cell suppression software was an important part of this five-year plan because of the need for statistical disclosure limitation procedures at EIA, which is discussed in the first section of this paper. The project plan outlined the software licenses that we would need to purchase to replace our current Disclosure Analysis (DiAna) software with the Census Bureau's linear programming (LP) cell suppression software. In the second section of this paper, we discuss cell suppression software based on the network flow methodology, such as DiAna, as well as modern software options that perform cell suppression using linear programming.

After our project plan was approved by EIA's Information Technology Governing Board, we drafted a Memorandum of Understanding (MOU) between EIA and the Census Bureau to share disclosure limitation software, and this MOU was signed by both agencies in June 2017 and was renewed in February 2023. In the third section of this paper, we will discuss our plan to modernize our cell suppression software including software acquisition, testing, and implementation in our production

[1] The analysis and conclusions contained in this paper are those of the author and do not represent the official position of the U.S. Energy Information Administration or the U.S. Department of Energy.

environments. We will also discuss programming enhancements that we made in our modified version of the prototype, as well as a status report on the production use of the modified version of the prototype.

In a related effort, which is discussed in the fourth section of this paper, we began two initiatives in 2023 to ensure that we consistently apply appropriate statistical disclosure limitation techniques to our data products. First, we proposed to EIA's Chief Process Officer that statistical reviews of our data products, including formal disclosure reviews, be incorporated as part of EIA's new Product Review and Approval Process. Second, we included the development of a plan to form a Disclosure Review Board (DRB) in the Office of Energy Statistics' Strategic Plan for Fiscal Year 2023. The proposed DRB would play an important role in proposing standards, best practices, and policies on the use of statistical disclosure limitation methods and resolving issues that arise in conducting formal disclosure reviews for EIA's data products.

1. The Need for Statistical Disclosure Limitation Procedures at EIA

Title 15 of the U.S. Code authorizes mandatory data collection by EIA. Specifically, EIA's mandatory data authority is provided in two sections of Title 15:

- Section 772, which established the mandatory requirement of owners and operators of businesses in the U.S. to report energy supply and consumption data to the EIA Administrator.
- Section 764, which established the EIA Administrator's powers to plan, direct, and conduct mandatory and voluntary energy programs that are designed and implemented in a fair and efficient manner. These powers include duties to collect, evaluate, assemble, and analyze energy information on U.S. reserves, production, demand, and related economic data, while obtaining the cooperation of business, labor, consumer, and other interests.

In the instructions for EIA's survey forms, we inform the respondents to our surveys when and for which data elements we apply statistical disclosure limitation procedures to statistical aggregates so that a respondent's data cannot be identified. Typically, these procedures involve cell suppression. The promises that we make to protect the confidentiality of data reported to EIA are important for us to keep in order to maintain the trust of EIA's survey respondents and the high quality of our data products.

The legal authority that establishes the situations in which we apply statistical disclosure limitation procedures to statistical aggregates produced from EIA's surveys is based in the Confidential Information Protection and Statistical Efficiency Act of 2018 (CIPSEA) and Exemption 4 of the Freedom of Information Act (FOIA). CIPSEA was originally enacted in 2002 as Title V of the E-Government Act of 2002, and it was reauthorized under the Foundations for Evidence-Based Policymaking Act of 2018. The FOIA exemption protects trade secrets or commercial or financial information that is confidential or privileged. For EIA's surveys, the FOIA exemption protects data provided by our respondents that, if released, could cause competitive harm for our respondents in the marketplace. All data items collected from 10 EIA surveys are protected under CIPSEA, and we perform statistical disclosure limitation procedures for all statistical aggregates produced from these data. Under CIPSEA, the data we collect must be used for statistical purposes and may not be disclosed to the public in identifiable form. We apply statistical disclosure limitation procedures for statistical aggregates produced from select data items from almost 40 other EIA surveys that are protected under the FOIA exemption. Although EIA uses data protected under the FOIA exemption for statistical purposes, these data may be shared by law with other federal agencies for nonstatistical purposes such as administrative, regulatory, law enforcement, or adjudicatory purposes. Also, data collected under the FOIA exemption may be subject to time limits on disclosure protection if the release of these data in the future would no longer cause competitive harm to our respondents.

2. Cell Suppression Software Based on Network Flow or Linear Programming

At EIA, we typically use cell suppression to protect sensitive table cells prior to publication in order to prevent disclosures of identifiable and sensitive data, where these table cells typically contain totals of volumetric data. For these sensitive cells, the corresponding table cell values are dominated by only a small number of contributors, which frequently happens in establishment surveys with skewed business populations.

We typically identify sensitive cells using the P% Rule, where the value of P depends on the survey and is confidential, but other linear sensitivity rules such as the (n, k) Rule and threshold rules may be used instead of, or in combination with, the P% Rule. Federal Committee on Statistical Methodology (2005) describes linear sensitivity rules that are used by federal statistical agencies like ours to identify sensitive table cells for suppression, which are called *primary cell suppressions*.

The basic P% Rule may be described as follows. If we denote the total value for a given cell by T and the values for the top two contributors to the cell by C1 (largest) and C2 (second largest), then a cell is considered sensitive and is a primary suppression if the remainder, T – C1 – C2, is less than P% of C1. In this situation, we do not want the second largest contributor to be able to derive the largest contributor's value within less than P% of C1, which is the amount of required protection. By testing this situation, we ensure that any contributor's value is sufficiently protected.

In tables containing additive relationships between individual cells, sensitive cell values may not be protected by suppressing only the table cell values for sensitive cells. In this case, additional suppressed cells, called *complementary cell suppressions*, are required. We typically need software to identify these cells for tables that are more complicated than simple one-dimensional tables, which show data having only a single additive relationship.

2.1 Disclosure Analysis (DiAna) Software

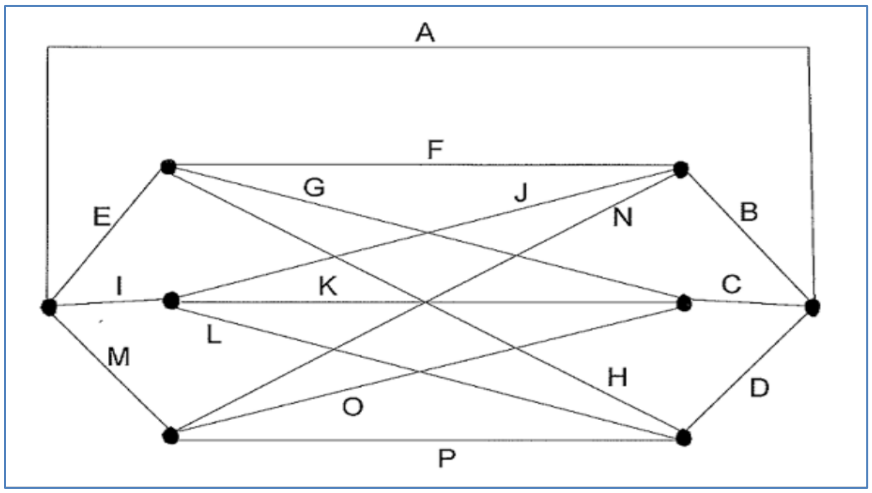
In the 1980s, Larry Cox, Jim Fagan, and Brian Greenberg at the Census Bureau developed the network flow methodology for tabular data based on the idea of a cell's capacity to protect a primary suppression, which is discussed in Cox, et al. (1986). Dating back to Robert Hemmig's network flow program for the 1987 Economic Census conducted by the Census Bureau, network flow methods were favored for determining complementary cell suppressions because they consumed less computing time. Robert Jewett continued to make enhancements to his cell suppression program that was written in FORTRAN and used for the 1992 Economic Census. Jewett's program incorporated the Minimal Cost Flow algorithm that was written in FORTRAN by Professor Darwin Klingman at the University of Texas. The details of Jewett's program for the 1992 Economic Census are presented in Jewett (1993).

Since Jewett's FORTRAN program was originally developed, software based on Jewett's program has been used at federal statistical agencies in the U.S. including EIA, the National Agricultural Statistical Service (NASS), and the Census Bureau. As a result of the proposal made by EIA's Mark Schipper and Ruey-Pyng Lu to the Federal Committee on Statistical Methodology's Confidentiality & Data Access Committee, Richard Graham developed the Disclosure Analysis (DiAna) software in 1998. DiAna was a variation of Robert Jewett's network flow cell suppression software that was written for the PC in FORTRAN and SAS. The SAS code was used to create the ASCII input files for the FORTRAN code because FORTRAN cannot read SAS data sets. EIA used this version of DiAna until 2014, when EIA acquired another variation of Jewett's software from NASS, which is still used at EIA for some of our surveys. This new version of DiAna was the result of a collaboration between NASS and the National Institute of Statistical Sciences and was written completely in PC SAS.

The basic principles of the network flow methodology for complementary cell suppression on which Jewett's original program and the DiAna software are based is described in Sullivan (1992). The objective is to ensure the protection of the sensitive data values at a minimum cost, which requires assigning a cost to each nonsensitive cell for being suppressed in the table. In this way, we can

publish our tables with maximum utility while protecting the sensitive cells. Typically, the nonsensitive cells' data values are each assigned as the cost in the network flow methodology, in which a two-dimensional table is represented as a network flow diagram. Jewett (1993) shows such a diagram, which is displayed in Diagram 1, where the arcs that are assigned a letter represent table cells, and nodes represent additive relationships in the table. For a given sensitive cell, identifying a set of complementary suppressions to form the suppression pattern involves finding the closed path of arcs associated with minimum cost that includes the arc for the sensitive cell and satisfies the required protection for the sensitive cell in flowing units through the suppression pattern in the network.

Diagram 1: Example of a Network Flow Diagram



Although we continue to use the PC SAS version of DiAna in production for some of our surveys, this software is outdated because of advances in computing technology that now allow for a linear programming (LP) solution. Given that the methodology on which variations of Jewett's software are based applies directly to two-dimensional tables, this software is known to under-suppress or over-suppress for more complicated table structures such as three-dimensional tables containing hierarchical structures and linked tables (i.e., same data in multiple tables), when attempting to back track from two-dimensional cross sections.

DiAna software requires four different input files to develop the complementary cell suppression pattern for a two-dimensional table. The software internally combines information coming from these four files to mathematically recreate the table structure for processing. The four files are:

- *Row identifier file*: The file contains a list of valid row numbers.
- *Row relations file*: The file identifies hierarchical relations between different rows.
- *Column relation file*: The file is used to show columns that are equal to a sum of other columns. Multiple records in this file may be used to sequentially process multiple linked two-dimensional tables.
- *Table cell values file*: The file needs to be created every time a new table is to be generated and contains a record for each nonzero cell present in the table. Each record on this file contains a row identifier, column identifier, total cell value (DiAna does not tabulate the micro data), largest cell contributor value, second largest cell contributor value, and optional third dimension code. It also contains a preference indicator that may be used to indicate a preference for cell suppression that includes options for prioritizing a cell in the suppression pattern or excluding (or "freezing") previously published data from the suppression pattern.

In addition to these four files, the software also requires multiple parameters that are used during the table processing including: the value of P in the P% Rule (the software can first identify primary

suppressions, if desired, instead of identifying them on the table cell values file), the level of details in the print output, and various limits on variable arrays used internally by the software.

2.2 Cell Suppression Software Based on a Linear Programming Algorithm

Software that incorporates a linear programming (LP) algorithm to identify complementary suppressions attempts to minimize an objective function subject to linear constraints that express the additive relationships in the publication tables. As a result of setting up the problem to be solved in this way, multidimensional data are better supported, and we can ensure that identifiable and sensitive data are not disclosed in linked tables, where the same data appear in multiple tables that have different additive relationships.

From my experience working in the Economic Statistical Methods Division at the Census Bureau, I was familiar with the excellent work that was done by a team of researchers at the Census Bureau to develop LP cell suppression software that could handle the volume and complexity of tables published from the quinquennial Economic Census. This research team was led by Philip Steel and included James Fagan, Vitoon Harusadangkul, Paul Massell, Richard Moore Jr., John Slanta, and Bei Wang. In September 2015, the team was awarded the prestigious Department of Commerce Silver Medal for their work in developing this software.

Before developing their own LP cell suppression software, the Census Bureau's research team evaluated other software options that were developed by other government agencies. Steel (2013) discusses Tau Argus, which has been used by Statistics Netherlands and other countries in the European Union, but Tau Argus was neither open source at the time nor set up for batch, network server processing with security protocols. Steel (2015) mentions that the Census Bureau's LP cell suppression software is similar to the LP software developed by Statistics Canada. At the time, there were concerns as to whether Statistics Canada's LP software, which was called CONFID2 and was written in SAS, could efficiently process the volume and complexity of tables published from the Economic Census.

2.3 Census Bureau's Linear Programming Cell Suppression Software

Steel (2015) describes the Census Bureau's linear programming (LP) cell suppression software, which consists of a research prototype and a production version that use input and output files based on the format used in Jewett's original program. The common format makes it easier for users of software based on Jewett's original program to switch to the Census Bureau's LP cell suppression software. However, the comma-delimited input files are easier to use, the row identifier file is not needed, and there is a new input file of cells to be excluded from the suppression pattern (or "frozen") because they were already published.

The prototype is programmed in SAS and AMPL with the CPLEX linear solver, and the Census Bureau used it in production for the 2010 Manufacturing Energy Consumption Survey (MECS). The five SAS programs prepare ASCII inputs for the two AMPL programs because AMPL cannot read SAS data sets. The prototype is set up for three-dimensional tables, allowing for only a single relationship for the third dimension, and 50,000 cells, which we believed would be sufficient for EIA's surveys. However, the prototype could be modified to handle full three-dimensional tables or even four-dimensional tables.

The production version is programmed in C++ and interacts with CPLEX through a graph object, and it has been used in production for the Economic Census, the Business R&D and Innovation Survey, the Annual Survey of Manufactures, and MECS since the 2014 cycle. The production version involves less interaction between C++ and CPLEX than the prototype's AMPL and CPLEX, which is more efficient especially for large input files. However, the production version uses a more restrictive format for input files compared to the prototype.

Steel (2015) also describes the three stages of processing used in the LP cell suppression software. These stages borrow concepts from the network flow methodology.

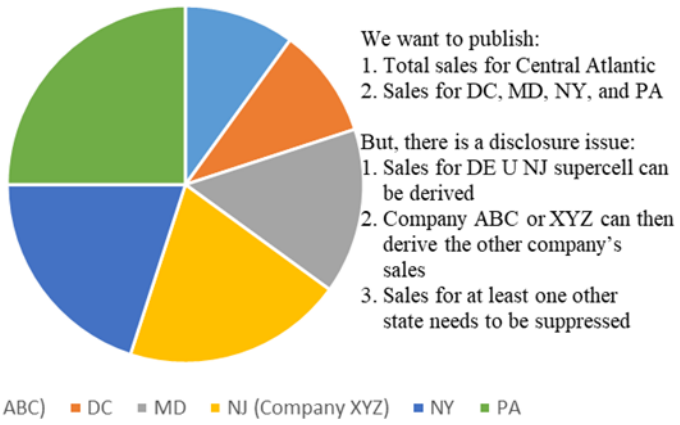
In the first stage, the LP solution is applied sequentially to a queue of all of the primary suppressions, similar to variations of Jewett's software. A linear objective function, f, is minimized based on minimum cost flow and represents a balance between the number of suppressions and the amount of cell values that are suppressed.

Wang and Klement (2008) describe the mathematical set up for the LP problem. For 3-dimensional tables, $f = \sum_{i,j,k} c_{i,j,k}\left(x^{+}_{i,j,k} + x^{-}_{i,j,k}\right)$, where i denotes a given row, j denotes a given column, k denotes a given level, $c_{i,j,k}$ denotes the cost coefficient for the cell (0 for suppressed cells or the cell value for unsuppressed cells having a value), and $x^{+}_{i,j,k}$ and $x^{-}_{i,j,k}$ denote the flow into and out of the cell, respectively. Linear constraints are formed on sums of cells to marginal totals, cell capacity (i.e., bounds on the flow into and out of a cell), and the flow into and out of the target primary suppression, where the flow into the target primary suppression is the amount of required protection and the flow out of the target primary suppression is 0.

To reduce over suppression and processing time, a given solution is checked to see if additional primary suppressions are protected besides the target primary suppression. This process is called "skip-P", where "P" refers to a target primary suppression, and is an improvement compared to variations of Jewett's software because these additional protected primary suppressions are skipped in the queue to improve processing efficiency. The solution after processing the queue of primary suppressions may still contain unnecessary complementary suppressions, so an additional step is included to remove these complementary suppressions by constructing a second problem in a way that limits the solution based on the original problem and uses an inverted cost function.

In the second stage, the suppression pattern from the first stage is examined to identify aggregates, called *supercells*, which fail the P% Rule. Supercells are described in Massell (2011) and are composed of unions of suppressed cells in an additive constraint, which is often referred to as a *shaft*. Checking for sensitive supercells is another important improvement compared to variations of Jewett's software. In applying the P% Rule for a given supercell, the top two contributors are determined using the information in the table cell values input file for the top two contributors to its component suppressed cells. This procedure was done to make the processing as efficient as possible by not having to input a separate file of the top two contributors for each supercell after the supercells are determined. We have found that the top two contributors for a supercell tend to be correctly identified using this procedure, so it seems to work well in practice.

Figure 1. Annual Sales for Central Atlantic (Sensitive Supercell DE U NJ)



The simplest example of a sensitive supercell is when each individual component cell has only one contributor, which is often referred to as *company protection*. For an example of company protection, consider the hypothetical situation depicted in Figure 1, which represents annual sales by state for the Central Atlantic Region. Here, we are referring to the District of Columbia as a state

for simplicity. As is typically the case, we would like to publish total sales for the region and sales for as many states as we can. We have to suppress sales for Delaware because the sales consist of just the sales for Company ABC, and we have to suppress sales for New Jersey because the sales consist of just the sales for Company XYZ. We would like to publish sales for the other four states consisting of the District of Columbia, Maryland, New York, and Pennsylvania. However, there is a disclosure issue here because the sales for the supercell defined by the union of Delaware and New Jersey can be derived. Then, Company ABC or Company XYZ could derive the other's reported sales. So, we have to suppress the sales for at least one other state to protect the data reported by Company ABC and Company XYZ.

In the third stage, the LP solution is applied sequentially to a queue of the sensitive supercells, which is another improvement compared to variations of Jewett's software. For a given sensitive supercell, one or more nonsensitive cells in the additive relationship, called *siblings*, are identified as complementary suppressions and added to the suppression pattern.

3. EIA's Plan to Modernize Our Cell Suppression Software

In fiscal year 2017, we began an initiative to modernize our cell suppression software at EIA. This initiative was an important part of both our IT Modernization Project and the Office of Energy Statistics' five-year plan for improving the statistical methodologies for its programs. As mentioned in the previous section, DiAna is known to under-suppress or over-suppress for more complicated table structures than two-dimensional tables. As an improvement, the Census Bureau's LP cell suppression prototype better handles multidimensional and linked tables.

Our IT Modernization Project has been an important initiative at EIA to modernize our production processing environment for surveys conducted by the Office of Energy Statistics. Prior to 2021, our priorities for the IT Modernization Project primarily focused on modernizing our data collection system and database structure. As a result, we decided to first test the Census Bureau's LP prototype in our legacy production environment for a couple of our surveys that would not be modernized for some time and had simple publication table structures. In January 2018, we first purchased licenses for AMPL with CPLEX for our legacy production environment. In February 2021, we purchased licenses for AMPL with CPLEX for our modernized production environment, which is independent from our legacy production environment.

The results that we expected to see in replacing DiAna with the LP prototype in production achieved two major goals:

- EIA would be in a much better position to prevent disclosures of identifiable and sensitive data, while publishing as much data as possible.
- EIA would be better able to maintain the cell suppression software and support it through software updates provided by the Census Bureau, version control, training, and documentation of the software and disclosure avoidance methodology.

Besides the methodological improvements and familiar network flow concepts and file formats discussed in the previous section, key factors in EIA choosing the Census Bureau's LP prototype were as follows:

- We estimated savings of $500,000 to EIA as a result of benefitting from the research and development done by the Census Bureau's research team, including reviews of software developed by other agencies and processing improvements discussed in the previous section.
- The Census Bureau agreed to provide EIA staff with demos of the prototype and documentation on how to use the software. The demos and documentation were really helpful in learning the methodology and understanding how to use the LP prototype.
- EIA and the Census Bureau signed a Memorandum of Understanding (MOU) that specified that there would be no charge for the LP prototype.

- We already had a contract with SAS, and affordable options were available for acquiring licenses for AMPL with CPLEX. In January 2018, we purchased an AMPL dual-socket server license and 5 CPLEX single-user licenses for our legacy production processing environment that together cost $61,500. A sever license for CPLEX would have been much more expensive.

3.1 Software Acquisition, Testing, and Implementation in Production Environments

The timeline for software acquisition and initial testing was as follows:

- *June 2017*: EIA and Census Bureau signed the MOU to share software for data collection, processing, and dissemination including disclosure protection. We requested and subsequently received approval for purchasing licenses for AMPL with CPLEX as part of the June 2017 Federal Information Technology Acquisition Reform Act (FITARA) list for the Department of Energy.
- *July 2017*: We acquired Census Bureau's LP prototype on a Census Bureau protected flash drive after obtaining approval from the Census Bureau's Disclosure Review Board and Office of Information Security. EIA's Information Systems Security Officer conducted a code review and vulnerability assessment based on authorization-to-operate (ATO) and security assessment standards provided by the Federal Risk and Authorization Management Program (FedRAMP).
- *December 2017*: We worked with our IT staff to successfully test the prototype in our legacy production environment, using temporary licenses for AMPL with CPLEX on a trial basis to ensure that the software met our needs. We processed the three-dimensional German data example, which was included with prototype, in about 30 minutes. This result was really encouraging, given that this example was included to put the software through its paces. The German data example consists of almost 18,000 cells and over 40 additive row relationships.
- *January 2018*: Our IT staff installed permanent license files for AMPL with CPLEX in our legacy production environment after the licenses were purchased.
- *February 2018*: We successfully ran the German data example in our legacy production environment using permanent license files for AMPL with CPLEX.

For fiscal year 2018, we tested the prototype in our legacy environment for two surveys for which EIA performed cell suppression on simple publication tables:

- Feedstocks consumed from Monthly Biodiesel Production Survey (EIA-22M).
- Shipments by state/territory from Annual Photovoltaic Module Shipments Report (EIA-63B).

For each survey, we prepared the input files for one survey cycle, ran the prototype in our legacy environment, and got results in about two minutes, which was really fast compared to the test run using the more involved German data example. We worked with the Census Bureau on problems that we found in how the SAS programs prepare the input to AMPL programs for tables that are only one-dimensional (EIA-22M) or two-dimensional (EIA-63B), as well as limitations in the supercells processing. Also, we attended demos at the Census Bureau in the first quarter of 2018 and started our internal documentation on the prototype, which we later shared with the Census Bureau to assist with their internal documentation.

3.2 EIA's Modifications to the Census Bureau's LP Prototype

In fiscal year 2019, we modified the Census Bureau's LP prototype to improve the processing of supercells for our surveys. The modifications that we made to the SAS programs for supercells processing that are included as part of the prototype were as follows:

- We corrected the array bounds for reading in data that was causing an error and preventing supercells processing inputs to be produced.
- We added code to handle cells with only one contributor that was causing missing values.
- We checked that siblings identified for suppression were not previously published, using the previously unused preference indicator on the table cell values input file.

- The prototype assumed that a sibling could always be found to provide required protection, but this assumption may not be satisfied, especially when the data are not all published at the same time. So, we modified the code to suppress the total when no such sibling could be found.
    - For the EIA-63B, we publish data monthly and then release an annual report that contains annual totals in July of the following year. When we perform the disclosure analysis for the annual report in February of the following year, we sometimes cannot find a sibling, among the two or three months of the year that have not been published yet, that provides sufficient protection for the supercell consisting of the suppressions across months of the year.
    - A hypothetical example is given in Table 1. In this example, we have already published data from January through September, and we held back the release of data for October through December so we can publish the annual total, when possible. However, for this state, the data need to be suppressed for October through December, as indicated by "W", and the supercell, which is determined by aggregating all 7 months of suppressed data, is sensitive. There is no additional unpublished monthly data that we could suppress to protect the sensitive supercell, so we must suppress the total.
- We were getting duplicates when suppressing multiple siblings, so we changed the code to store the amount of remaining protection needed after subtracting a sibling's capacity.
- Identified siblings for suppression were not integrated into the suppression pattern from the first stage, so as a quick fix, we added these siblings to the end of the queue of primary suppressions from the first stage for the third stage of processing. This is not the most efficient way to handle these additional complementary suppressions, which could matter when processing data from the Economic Census, but it works well for EIA's tables. As expected, we found that the siblings were skipped in the queue of primary suppressions.

We gave the Census Bureau a brief overview of these changes. With the February 2023 signing of the revised MOU between EIA and the Census Bureau to share disclosure limitation software, we plan to meet later this year to share our changes to these programs in detail when time allows. Then, the Census Bureau may decide whether to enhance their software based on the modifications that we made to the prototype.

Table 1. Hypothetical State-Level Monthly Photovoltaic Module Shipments

Month	*Previously Published Monthly Shipments*	*Unpublished Monthly Data*
January	100,000	
February	50,000	
March	75,000	
April	W	
May	3,000	
June	W	
July	2,000	
August	W	
September	W	
October		W
November		W
December		W

Notes: "W" indicates that the value of the cell is withheld to prevent the disclosure of sensitive data. The supercell defined by aggregating suppressed data for the 7 months is sensitive. No sibling is available so we must also suppress the annual sales.

3.3 Status Report on the Production Use of EIA's Modified Version of the Prototype

From February 2019 to February 2023, we used EIA's modified version of the prototype in production for the EIA-63B year-end estimates of U.S. photovoltaic module shipments by state and territory in our legacy processing environment. When we used the prototype for the first time in February 2019, we got the same results as our former SAS programs that performed the statistical disclosure limitation, except that the prototype also identified an infeasibility in that a cell was characterized as both previously published and a primary suppression. We corrected this error in our input files.

In July 2020, we received approval by EIA's Information Technology Governing Board to purchase AMPL with CPLEX for our modernized processing environment. In February 2021, after receiving FITARA approval for obtaining AMPL with CPLEX, we purchased an AMPL dual-socket server license and 4 single-user CPLEX licenses for the modernized processing environment, which together cost $40,600. As part of AMPL's maintenance and support service, we received a volume discount based on the number of licenses under support, plus we were able to transfer our unused fifth single-user CPLEX license from our legacy processing environment to our modernized processing environment. We set up EIA's modified version of the prototype in our modernized processing environment and successfully tested the prototype using the German example provided by the Census Bureau. Prior to the test, I gave demos of the prototype to our IT staff in June 2020 and January 2021.

For both our legacy and modernization environments, there are annual charges for AMPL's maintenance and support service that include, besides volume discounts that were mentioned above, access to AMPL and CPLEX software updates, regeneration of license files necessitated by hardware changes, and technical assistance with installation and execution. The annual rate is currently 20% of the license price at the time of renewal.

Since we began using EIA's modified version of the prototype, we have had changing priorities for production use in our new modernized environment. There were other priorities with the IT modernization of the Monthly Report of Biofuels, Fuels from Non-Biogenic Wastes, Fuel Oxygenates, Isooctane, and Isooctene (EIA-819), which includes the now discontinued EIA-22M. Also, for the foreseeable future, cell suppression software is not needed for the new EIA-819 survey because the new feedstock tables produced from the survey contain no additive relationships in order to maximize publication of data by type of feedstock. Currently, we are planning to test the LP prototype later this year in our new SAS production environment.

4. EIA's Initiatives to Ensure Consistency in Conducting Disclosure Reviews of Our Data Products

EIA is implementing a new Product Review and Approval Process that is led by EIA's Chief Process Officer and includes statistical reviews of data products, as required by Office of Management and Budget (2006). These statistical reviews include formal disclosure reviews based on a checklist approach. The proposed *Disclosure Review Board (DRB)* would play an important role in proposing standards, best practices, and policies on the use of statistical disclosure limitation (SDL) methodologies to support the proposed *Data Stewardship Executive Committee (DSEC)*. The proposed DSEC would develop and establish policies at EIA that involve several areas of data stewardship including protection of confidential and sensitive data as required by law. Also, DRB would collaborate with EIA's program offices in the Office of Energy Statistics (OES) to resolve issues in conducting disclosure reviews. As noted in Interagency Council on Statistical Policy (2023), consulting with an agency's DRB, when practicable, is an important task for when an agency is balancing between the overall sensitivity, accessibility, and quality of a proposed data asset. Currently, some of the disclosure reviews for EIA's data products are conducted on an ad-hoc basis and are not included in production schedules, which results in inconsistency in conducting these reviews over time and puts a strain on the staff that conducts them. Also, program offices sometimes run cell suppression programs like a black box without fully understanding how these programs

work. So, training EIA staff on SDL methodologies and software would also be an important role for the DRB.

To develop an outline of our proposed plan to form a Disclosure Review Board (DRB) at EIA, we relied on our experiences in working with DRBs at other statistical agencies, as well as charters from selected statistical agencies that included the Census Bureau (2018), Census Bureau (2019), Bureau of Labor Statistics (2018), National Center for Education Statistics (2023), National Agricultural Statistics Service (2023), National Center for Science & Engineering Statistics (2022), and National Center for Science & Engineering Statistics (2023). Also, Scheuren et al. (2000) provided some additional background information on many of these statistical agencies' DRBs. For this outline of our proposed plan, we focused on the following functions and membership components of the DRB.

We proposed that the DRB's functions would include:

- Proposing SDL policies, standards, and best practices to DSEC.
- Communicating SDL policies, standards, best practices, and methodologies internally and externally, which includes training EIA staff.
- Researching SDL methodologies by collaborating internally and externally.
- Reviewing selected data products for potential disclosure, based on information provided by subject-matter and technical experts in OES on statistical methodologies or requests for DRB review made by the program offices in OES.
- Collaborating with the Office of Information Technology to maintain a list of approved SDL software.
- Reviewing proposed data products prepared by other federal agencies or external researchers having approved access to EIA's protected data for statistical purposes, which will help us fulfill our legal obligations under the Foundations for Evidence-Based Policymaking Act of 2018.

We proposed the following four components of DRB membership:

- *DRB Chair*, who would be the Director for the Office of Statistical Methods & Research, provide senior-level technical guidance, adjudicate disagreements at DRB meetings preventing consensus, and represent DRB on the proposed DSEC.
- *DRB Vice-Chair*, who would be included in all communications regarding DRB submissions, serve as the back-up point of contact for submissions for DRB review and the back-up for the DRB Chair, work with the DRB Chair to review DRB submissions prior to the DRB meeting, communicate with the program offices regarding their DRB submissions as appropriate, and bring issues to the attention of DSEC for their consideration, when appropriate.
- *DRB Coordinator*, who would be the point of contact for all submissions for DRB review made by the program offices, work with the DRB Vice-Chair and DRB Chair to establish the DRB agenda, and schedule regular meetings of the DRB.
- Other members would include two senior mathematical statisticians, two junior mathematical statisticians, and four subject-matter analysts of varied experience from the program offices. IT staff could be included as members or consultants to DRB.

Pending initial approval of DRB and DSEC by EIA's senior leadership, we next plan to propose these functions for DSEC before discussing the details of DSEC membership and potential subcommittees of DSEC other than DRB:

- Reviewing and providing guidance to DRB on proposed standards, best practices, and policies.
- Resolving issues that are appealed by the program office when a consensus cannot be reached between DRB and the program office that is responsible for publishing the data product.
- Working with senior leadership to ensure appropriate funding for functioning of the DRB and SDL research.

- Providing guidance on non-SDL data stewardship policies related to use of administrative and third-party data, CIPSEA, and an expanded data sharing program that allows access to EIA's protected data by approved external researchers for statistical purposes under the Foundations for Evidence-Based Policymaking Act of 2018.

EIA's current policy is that we do not share protected data with external researchers. However, we are revisiting our policy based on the Foundations for Evidence-Based Policymaking Act of 2018. To that end, we are planning to work with the Census Bureau in 2024 to pilot use of the Federal Statistical Research Data Centers for approved access to protected CIPSEA data for statistical purposes.

Conclusion

Modernizing our cell suppression software has been a high priority at EIA because of the large number of surveys for which we perform statistical disclosure limitation on published statistical aggregates. Our modified version of the Census Bureau's LP prototype is an affordable solution that is more efficient than the Disclosure Analysis (DiAna) cell suppression software that we have been using for years, handles complicated table structures better, and checks for sensitive supercells, while using familiar formats for our input and output files. The Census Bureau provided demos and documentation on the Census Bureau's LP prototype, which allowed us to quickly learn the methodology and understand how to use the LP prototype. We have been successfully using EIA's modified version of the prototype in production for the EIA-63B year-end estimates of U.S. photovoltaic module shipments by state and territory in our legacy processing environment. We set up and successfully tested EIA's modified version of the prototype in our new modernized processing environment, and we are planning to test the LP prototype later this year in our new SAS production environment.

In a related effort, we began two initiatives in 2023 to ensure that we consistently apply appropriate statistical disclosure limitation techniques to our data products. These initiatives involved including formal disclosure reviews as part of EIA's new Product Review and Approval Process and developing a plan to form a Disclosure Review Board.

Acknowledgements

The author thanks the U.S. Census Bureau for sharing their LP cell suppression prototype. In particular, the author thanks Philip Steel for quickly preparing the software for our use and for giving us informative demos and documentation on the LP prototype. The author thanks Marvin Atchley, Timothy Butka, Debra Coaxum, Pamela Edmond, Thomas Leckey, Alexander McLean, Igor Pedan, Nanda Srinivasan, Joseph Wilson, and Crystal Wunder at EIA for their help and support in acquiring the LP prototype and necessary software licenses, as well as Nathan Agbemenyale, Thomas Broene, Adebowale Sijuwade, and Jason Worrall at EIA for their valuable contributions in testing the LP prototype and using the software in production for the EIA-63B year-end estimates of U.S. photovoltaic module shipments by state and territory. Also, the author thanks Samson Adeshiyan, Pushpal Mukhopadhyay, Adebowale Sijuwade, Jason Worrall, and Orhan Yildiz for their helpful comments on this paper.

References

Bureau of Labor Statistics (2018). Disclosure Review Board Charter. Unpublished internal memorandum that is available in the *Federal Committee on Statistical Methodology's Data Protection Toolkit*.

Cox, L., Fagan, J., Greenberg, B., and Hemmig, R. (1986). Research at the Census Bureau into Disclosure Avoidance Techniques for Tabular Data. *Proceedings of the 1986 Joint Statistical Meetings*, Survey Research Methods Section, Chicago, IL.

Census Bureau (2018). Census Bureau Data Stewardship Program: Organization of the Disclosure Review Board. *Data Stewardship Policy Document DS025*. Published on Census Bureau website under *Data Stewardship Policies*.

Census Bureau (2019). Data Stewardship Executive Policy Committee Charter. Published on Census Bureau website under *Data Stewardship Executive Policy Committee*.

Federal Committee on Statistical Methodology (2005). Report on Statistical Disclosure Limitation Methodology. *Statistical Policy Working Paper 22*, Washington, DC.

Interagency Council on Statistical Policy (2023). *CIPSEA 3582 Draft Regulations*.

Jewett, R. (1993). Disclosure Analysis for the 1992 Economic Census. Unpublished Economic Programming Division internal documentation, U.S. Census Bureau, Washington, DC.

Massell, P. (2011). Modernizing Cell Suppression Software at the U.S. Census Bureau. *Proceedings of the 2011 Joint Statistical Meetings*, Survey Research Methods Section, Miami Beach, FL.

National Agricultural Statistics Service (2023). Data Access and Disclosure Review Board and Standard Application Process Appeals Committee. Unpublished internal memorandum. *Operations Memorandum No. TM-01-23*.

National Center for Education Statistics (2023). Disclosure Review Board (DRB). Unpublished internal memorandum that is available in the *Federal Committee on Statistical Methodology's Data Protection Toolkit*.

National Center for Science & Engineering Statistics (2022). Data Governance Board Charter. Published on NCSES website under *NCSES Data Governance*.

National Center for Science & Engineering Statistics (2023). Confidentiality Governance Group Charter. Unpublished internal memorandum.

Office of Management and Budget (2006). Office of Management and Budget Standards and Guidelines for Statistical Surveys.

Scheuren, F., Hoy, E., Zarate, A., Stamas, G., McMIllen, M., and Therriault, G. (2000). Panel on Disclosure Review Boards of Federal Agencies: Characteristics, Defining Qualities and Generalizability. *Proceedings of the 2000 Joint Statistical Meetings*, Indianapolis, IN.

Steel, P. (2013). Re-development of the Cell Suppression Methodology at the US Census Bureau. Paper presented at the *2013 Joint United Nations Economic Commission/Eurostat work session on statistical data confidentiality*, Ottawa, Canada.

Steel, P. (2015). Techniques to apply cell suppression to large sparse linked tables and some results using those techniques on the 2012 (US) Economic Census. Paper presented at the *2015 Joint United Nations Economic Commission/Eurostat work session on statistical data confidentiality*, Ottawa, Canada.

Sullivan, C.M. (1992). The Fundamental Principles of a Network Flow Disclosure Avoidance System. *Statistical Research Division Research Report Series, No. RR-92/10*, Census Bureau, Washington, DC.

Wang, B. and Klement, S. (2008). Disclosure Protection – A New Approach to Cell Suppression. *Proceedings of the 2008 Joint Statistical Meetings*, Business and Economics Statistics Section, Denver, CO.