Best Practices and Challenges in Performing Disclosure Reviews at the U.S. Energy Information Administration

2024 FCSM Research and Policy Conference
David Kinyon, Supervisory Mathematical Statistician
October 23, 2024 | Hyattsville, MD

The analysis and conclusions contained in this presentation 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.

Outline

- Need for disclosure protection at the U.S. Energy Information Administration (EIA)
- Best practices for performing our disclosure reviews, which primarily involve cell suppression
- Challenges in performing our disclosure reviews
  - Additional reviews as part of potential future expansion of EIA's data sharing program
  - Agreed-upon statistical disclosure limitation (SDL) methodology for the 2026 Manufacturing Energy Consumption Survey

Need for Disclosure Protection Methodology at EIA

- EIA conducts almost 50 surveys in which data are protected by law
- EIA informs respondents to our surveys when disclosure limitation procedures (typically, cell suppression) are applied to statistical aggregates
  - These promises are important to keep in order to maintain the trust of EIA's survey respondents and the high quality of our data products
  - All data items collected from 10 surveys protected under the Confidential Information Protection and Statistical Efficiency Act of 2018 (CIPSEA)
    - Data must be used for statistical purposes
    - CIPSEA does not allow for time limits on disclosure protection
  - Select data items from almost 40 surveys protected under an exemption to the Freedom of Information Act (FOIA)
    - Data may be used for nonstatistical purposes
    - FOIA exemption allows for time limits on disclosure protection

Best Practices for Performing EIA's Disclosure Reviews

- Most of EIA's data products are tabular, involving volumes based on quantities collected in establishment surveys of typically skewed populations
- EIA also publishes public-use files for the Residential Energy Consumption Survey and the Commercial Buildings Energy Consumption Survey
- SDL procedures are based on methodologies presented in FCSM's Statistical Policy Working Paper 22 and include the following procedures
  - *Cell suppression:* withhold sensitive data and other data to prevent a data user from backing into the sensitive data using the published data and the relationships between table cells
  - *Combining table columns or rows:* aggregate table cells involving at least one sensitive cell
  - *Removing identifiers and limiting geographic detail*: limit detail in public-use files
  - *Top coding, bottom coding, recoding into intervals:* categorize data as greater than a certain value, less than a certain value, or in intervals
  - *Noise infusion* - perturb the data, typically at the microdata level using a multiplicative factor, by purposely introducing an element of noise/error

Primary Cell Suppressions of Volume Data

- The main method used by EIA to identify sensitive cells is the P% Rule, where the value of P depends on the survey and is confidential
  - 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
  - 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
  - This simple rule can be applied using a computer program or a spreadsheet
  - Other rules, such as the (n,k) Rule or threshold rules, may be used instead of, or in combination with, the P% Rule

- Additional cells called *complementary suppressions* may need to be suppressed to protect the primary suppressions if the primary suppressions alone do not provide sufficient protection for each other

Identifying Complementary Suppressions

- If there are few sensitive cells and the structure of the tables is not very complicated (e.g., 2-D tables that are not linked through common marginal totals) then a manual procedure may be acceptable
- In general, this is a difficult problem, and we rely on computer programs
  - Disclosure Analysis (DiAna) software based on network flow algorithm works well for 2-D tables, but it is known to under-suppress or over-suppress for more complicated table structures (e.g., 3-D tables containing hierarchical structures and linked tables)
  - EIA's modified version of the Census Bureau's linear programming (LP) cell suppression prototype better handles multidimensional and linked tables and checks for sensitive *supercells*, which are made up of unions of suppressed cells in an additive constraint and fail the P% Rule
    - Since February 2019, we have used our modified version of the LP prototype in production for the Annual Photovoltaic Module Shipments Report, which incorporates previously published data from the Monthly Photovoltaic Module Shipments Report

Challenges in Performing EIA's Disclosure Reviews

- Potential for future expansion of EIA's data sharing program under the Foundations for Evidence-Based Policymaking Act of 2018 to include external researchers using our protected data for statistical purposes
  - For each application, proposed output must be reviewed for disclosure risk
  - Disclosure reviews of special tabulations must take into account previously published data
- Collaboration between EIA and Census Bureau on agreed-upon SDL methodology for 2026 Manufacturing Energy Consumption Survey (MECS)
  - MECS is a sample survey of U.S. manufacturing establishments conducted every 4 years to collect more than 60 pages of information related to energy consumption and expenditures
  - Interagency agreement for Census Bureau to collect and process data under Title 13, U.S.C.
  - For 2026 MECS, Census Bureau determined to switch from cell suppression to Evans, Zayatz, Slanta (EZS) multiplicative noise based on their research for 2018 MECS
  - EIA would like to collaborate with Census Bureau to extend research to include all 2018 and 2022 MECS tables and analyze effects on all published estimates and their standard errors

Issues with Noise Infusion for Establishment Surveys

- Data users often want to see “real” data published by government agencies
  - How much noise is acceptable for a given statistic in terms of data quality?
  - How to inform data users as to the range of noise incorporated in a given statistic? At a minimum, should we perform primary suppression and notify data users that any derived statistics should be used with caution due to poor data quality?
- Balancing the effects of noise based on a key statistic may be problematic
  - Business populations tend to be highly skewed
  - Many data items are often collected (e.g., MECS), and they may not all be highly correlated
  - Business populations may change over time due to growth, acquisitions, and births/deaths
- Period-to-period changes are often of interest to data users (e.g., MECS)
  - Statistically significant period-to-period changes may become not significant due to the increase in the variance of statistics as a result of noise infusion
  - Direction of noise for large units must typically remain the same over time

Thank You!

- Thank you for you time and attention during this presentation!

- For questions after the presentation, please email me at David.Kinyon@eia.gov