Research And Methodology Directorate

A History of the Economic Census and Disclosure Avoidance

By Laura McKenna

Issued April 2019

United States™ Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

INTRODUCTION[1]

The U.S. Census Bureau conducts the Economic Census under Title 13, U.S. Code, Section 9 mandate to not "use the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied; or make any publication whereby the data furnished by any particular establishment or individual under this title can be identified; or permit anyone other than the sworn officers and employees of the Department or bureau or agency thereof to examine the individual reports (13 U.S.C. § 9 (2007))." To compile a more comprehensive and complete data resource, the Census Bureau uses data obtained from other sources to supplement data collected in Economic Censuses and surveys. Through an agreement with the Internal Revenue Service, the Census Bureau obtains Federal Tax Information data whose disclosure is prohibited by Title 26 of U.S. Code. Administrative records provided by other federal and state agencies are also used in conjunction with commercial data purchased from various data brokers to further support the Census Bureau's mission.

The Census Bureau applies Disclosure Avoidance (DA) techniques to its publicly released statistical products in order to protect the confidentiality of its respondents and their data. None of the information in this paper is confidential.

HISTORY OF THE ECONOMIC CENSUS

The Census Bureau has measured U.S. economic activities since the first census of manufactures in 1810. Since then, the nation's economy has grown more diverse and complex, and the Census Bureau now conducts the Economic Census, which over time was expanded to include retail and wholesale trade, construction industries, mining, and a broad array of services, <www.census.gov/history/www/programs/economic/economic_census.html>. Early in the nineteenth century, Congress ordered census takers to ask questions on manufactured products and goods. The 1905 Economic Census was the first to be conducted by mail. There are a few 1905 state-level tables on the Internet, and it appears that the states where data were sparse were combined into a category of "all other states." In 1930, the Census Bureau conducted the first census of business covering retail and wholesale trade. A note from this publication says, "The Census Bureau is prohibited by law from publishing any statistics which might make possible the disclosure of operations of individual establishments. As a rule, statistics are given for each industry represented by three establishments or more. In some cases, however, a single establishment produced so large a proportion of the combined output of three or more establishments in a particular industry in a state that the figures for the industry could not be given without disclosing approximately the operations of the dominant establishment" (it is not clear what criterion was used to determine dominance).

In 1947, the first official Census of Manufactures was conducted, followed by the 1948 Census of Business. In 1954, the Economic Census integrated various kinds of businesses. Since 1963, administrative records have provided information for very small firms. Also in 1963, the Census of Transportation began collecting data on travel and transportation of commodities.

Currently, the Economic Census is conducted in years ending in "2" and "7." In 1992, there was a major expansion of the Economic Census to include finances, insurance, real estate, communications, and utilities, which account for more than 20 percent of the U.S. gross domestic product, <www.census.gov/history/www/programs/economic/economic_census.html>. The 1997 Economic Census was the first to use the North American Industry Classification System (NAICS) developed by the United States, Canada, and Mexico. Previously Standard Industrial Classification (SIC) codes were used. In 1995, the Census of Agriculture was transferred to the U.S. Department of Agriculture, National Agricultural Statistics Service.

MAGNITUDE DATA

Data from the Economic Census is published in the form of additive magnitude tabular data. Magnitude data are aggregates of quantities of interest from establishments within a table cell. Each cell includes the number of establishments operating in a given geographic area broken down by NAICS, and the total sum of a value of interest such as value of shipments for those establishments, <www.census.gov/library/working-papers/2002/adrm/massell-01.html>. Most tables are two- or three-dimensional in the Economic Census, many have hierarchical relationships (subtotals), and many tables overlap.

DA protection is given at the company (firm) level and in such a way that ensures that values for individual establishments are protected as well. Values are commonly highly skewed at both the individual establishment and company level. Totals

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

of establishment values within a given company in a cell figure into the calculation of cells that are at risk of disclosure, their required protection, and the capacities of other cells in the table to protect them.

Certain data cells, called primary suppressions (or sensitive cells), in tables from the census are withheld because a data user could use them to very closely estimate a value reported by an individual establishment (Cox, 1981; Sullivan, 1992). The table cell values withheld are replaced with the letter "D" for disclosure. Most census tables are additive and show totals and subtotals of certain cells. To ensure that primary suppression values cannot be obtained through addition and subtraction due to the additivity of the tables, other cell values, called complementary suppressions, must also be withheld, <https://nces.ed.gov/FCSM/pdf/SPWP22_rev.pdf>. These values are also replaced with the letter "D." Values for cells that are not suppressed remain unchanged. Unlike the magnitude cell values, establishment counts are not considered to be disclosures, so counts may be published in all table cells, <www.census.gov/programs-surveys/economic-census/technical-documentation/methodology/disclosure.html>.

DA PRIOR TO THE 1992 ECONOMIC CENSUS

Cell suppression software has more than a 40-year history at the Census Bureau, and programs have undergone an interesting evolution. According to Cox (2000), the earliest large-scale use of automated suppression programs based on a mathematical theory was for the 1977 Economic Census. This program was based on combinatorial algorithms for protecting confidential data within a single two-dimensional (2D) table. At that time, the number of suppressed cells was the measurement for information loss (each suppressed cell had a cost of 1). Disclosure protection in three way tables was done heuristically by "stacking" constituent two-way tables (think of laying one on top of another). Suppression was performed on the layered 2D tables separately, and then the heuristic attempted to ensure the third dimension was also protected. Cell sensitivity (and primary cell suppression) was based on an (n, k)-rule. A cell was deemed sensitive if the largest n values of the contributing establishments accounted for more than k percent of the total cell value, <https://nces.ed.gov/FCSM/pdf/SPWP22_rev.pdf> and parameters were kept confidential. The suppression module was called INTRA, denoting that suppression was "intra-table." INTRA was used by the Census Bureau for the 1977 and 1982 Economic Censuses and several surveys conducted around that time.

In 1987, Larry Cox and Brian Greenberg proposed using mathematical networks for complementary cell suppression. Networks are mathematical models based on flow of a quantity along the arcs of a graph, and there is a natural interpretation of a 2D table as a network, <https://www2.amstat.org/sections/srms/proceedings/papers/1991_060.pdf>. They worked on this idea with Dr. Bruce Golden from the University of Maryland and his students as well as a small group of researchers at the Census Bureau. A cell suppression program based on network flow theory was written by Bob Hemmig and used for the 1987 Economic Census. A new and improved suppression program, also based on networks, was written by Bob Jewett and used for the 1992 Economic Census (see the next section). The computational module of this program, the network flow algorithm, is a Fortran subroutine called MCF, written by Professor Darwin Klingman at the University of Texas circa 1980 (MCF stands for Minimal Cost Flow). When using this type of optimization program, a cost must be assigned to the suppression of each complementary suppression cell. At this time, the cost of suppressing a cell was changed from 1 to the cell's value (Massell, 2001). Cell suppression programs at the Census Bureau call optimization routines (such as MCF) to find the optimal set of complementary suppressions in terms of achieving protection while minimizing information loss.

DA FOR THE 1992 ECONOMIC CENSUS

A more detailed history of the 1992 Economic Census is available at <www.census.gov/history/pdf/1992econhistory.pdf>. The DA methods, techniques, and software were changed dramatically for the 1992 Economic Census. General programs were used for most parts of the Economic Census, but were altered for the census of agriculture as it represented a slightly different problem, <www.census.gov/history/www/programs/agriculture/census_of_agriculture.html>. As in previous years, cell suppression was the Census Bureau's DA method of choice to protect data products from the 1992 Economic Census, <https://onlinelibrary.wiley.com/doi/pdf/10.1002/net.3230220407>.

The MCF subroutine and the entire network flow suppression program based on it were fast computationally. For 2D tables, the network method is known, under certain conditions on the distributions of contributors to the cell values, to create a

suppression pattern that is not undersuppressed, i.e., a pattern with a sufficient number of complementary suppressions to ensure that the primaries have (at least) the desired amount of protection for each of the contributors (Cox, 1995). Undersuppression means the primary suppressions did not receive the desired level of protection. Oversuppression means that unnecessary cells were suppressed as complementary suppressions. For 2D tables, there was a widespread acceptance of the network-based suppression programs. However, for three-dimensional (3D+) tables, it is known that use of network flow methods is a rough heuristic, so rough that the possibility of undersuppression at the cell level exists and has actually occurred (although often there is only a small amount of it). It is known that Linear Programming- (LP-) based suppression always provides adequate protection at the cell level for standard tables of all dimensions. For this reason, Jim Fagan and Laura Zayatz wrote, in 1991, an LP-based suppression subroutine callable from Jewett's suppression program; it called an LP subroutine called XMP that was acquired from Jim Kelley of the University of Maryland (one of Bruce Golden's students). This suppression program was not fast enough for production work, and Jewett decided to try another approach for 3D tables. Jewett wrote a 3D suppression program using the idea that Bob Hemmig had used for the 1987 Economic Census; namely stacking tables and using the network flow approach for 3D tables even though it had the possibility of undersuppression, <www.census.gov/srd/sdc/Jewett.disc.econ.1992.pdf>.

Another great problem that the Census Bureau faced was the fact that so many very large tables were interrelated, and the MCF package could not process all of them at the same time. This lead to a process called "backtracking" where parts of the data were run through the software separately and repeatedly (with suppressions carried from one part to another) in order to make sure all of the same cells were either suppressed or published in all parts. This process was difficult and time consuming.

The first audit program for assessing the results of a cell suppression program was written by Laura Zayatz in 1992, <www.census.gov/srd/sdc/Jewett.disc.econ.1992.pdf>.

This audit program used the LP program XMP mentioned above to calculate the feasibility interval (reflecting the actual protection level) associated with each suppressed cell (either primary or complementary), <www.census.gov/srd/papers/pdf/rr92-02.pdf>. A feasibility interval represents the minimum and maximum value that a suppressed cell could have with a given suppression pattern in a table or set of tables. The program then compared the feasibility interval with the desired protection interval (the minimum and maximum values that are necessary to protect a primary suppression) to see if the latter was contained in the former. If the desired interval was contained in the actual interval for each suppressed cell, the cell suppression was deemed a success. The main purpose of the audit program was to determine and write out a list of those suppressed cells whose desired protection interval was NOT contained in the actual protection interval. In that case, the audit determines if the actual protection interval did have a width at least as large as that of the desired interval; this is called sliding protection, which many people feel is acceptable. Of course, the most serious situation is where there is little or no protection afforded a suppressed cell; such cases indicate either some trivial data processing error in some input or a programming error with the suppression program or possibly an inherent weakness with the methodology of the suppression program that leads to undersuppression. At that time, if a primary suppression did not receive enough protection, additional complementary suppressions were added by hand, and this process was also difficult.

Unfortunately, in 1992, the network flow suppression software used for every other part of the Economic Census did not work well for the Census of Agriculture, <https://www2.amstat.org/sections/srms/proceedings/papers/1991_060.pdf>. Most of the agriculture tables were 3D and were linked in very complicated ways, often missing a row or column (sometimes internal cells and sometimes totals) in a table that could be partially filled in using another table. Most other Economic Census tables were 2D and strictly hierarchical, and that structure was much easier to process using network flow software. The LP software (XMP) could have taken into account the unusual structure of the tables, but it ran too slowly. For the Census of Agriculture, the Census Bureau used a heuristic approach for complementary cell suppression much like the one that had been used in previous years, but it was improved to reduce over suppression. As stated previously, in 1995, the Census of Agriculture was moved to the Department of Agriculture.

DA FOR THE 1997 ECONOMIC CENSUS

As mentioned previously, the 1997 Economic Census was the first to use the NAICS developed by the United States, Canada, and Mexico (previously SIC codes were used). Because data users often compare Economic Census data from census to census (for example, 1992 to 1997), the Census Bureau wanted to produce "bridge tables" to help the users make the transition to NAICS and give them the ability to make such a comparison that would aid them in the change from SIC codes to NAICS. This proved extraordinarily difficult because of the different industry codes and cell suppression patterns used to protect both years of data. An SIC code may have been divided into two or more NAICS codes and vice versa. All of the relationships between cells in tables from both years were very complicated. The Census Bureau could only produce comparison tables for very large geographic areas, and the technique used for coordinating suppressions was very conservative and resulted in oversuppression.

DA FOR THE 2002 ECONOMIC CENSUS

In 1999, the Census Bureau acquired a license for a commercial LP and mixed integer programming package called CPLEX. The package was written in C++ and was computationally much faster than MCF and XMP. Its routines could be called from either a FORTRAN program or a C program. Also, because it was a LP package, it could be used on tables of all dimensions. This was a great breakthrough but needed quite a bit of testing, and software was not ready for the 2002 Economic Census. It was used for the 2007 Economic Census.

In 2002, the Census Bureau stopped using the (n, k) rule to determine primary suppressions, and began using the P% rule instead, <https://nces.ed.gov/FCSM/pdf/SPWP22_rev.pdf>. The P% rule is used to ensure that no one can estimate an establishment or firm's true reported value within P%.

DA FOR THE 2007 ECONOMIC CENSUS

The program to protect the 2007 Economic Census called the optimization LP software CPLEX. It was a great improvement over MCF, reducing oversuppression and undersuppression.

In 2010, staff from the Center for Disclosure Avoidance Research and the economic program areas formed a cell suppression improvement team. Their purpose was to document problems with the then current cell suppression software so they could begin addressing them. All of the problems involved undersuppression or oversuppression. Most of the problems were due to the complexity of the tabular data. The rates of both undersuppression and oversuppression were already very small, but the Census Bureau's goal is to make them both zero.

The rest of this section is taken directly from <https://www2.amstat.org/sections/srms/proceedings/y2011/files/301855_67474.pdf>. It outlines the three most important problems that were identified.

"The following are requirements for Census Bureau suppression that are beyond those typically found with cell suppression for magnitude data tables.

A "A company consists of one or more establishments, typically at different locations throughout some region. For statistical purposes, the key fact is that the Census Bureau has data for each establishment for a given company. These establishment values form the contributions to cell values. Thus, a given establishment contributes to at most one cell of a table, but if there are k establishments for that company, those k establishment values may contribute to as many as k different interior cells in the table. The basic type of protection done in cell suppression involves protecting only the individual contributions to cell values, i.e., it does not involve protecting sums of associated values. For economic data, this is called 'protection at the establishment level.' The type of protection the Census Bureau is required to do is more complicated. It involves the basic type of protection described above, plus protection of the sums of establishment values associated with each company. This type of sum-level protection is often called, for economic data, 'protection at the company level.' 'Protection' of these data involves suppressing carefully chosen cell values, so that estimation of the cell contributions (or sums of such associated with a given company) cannot be made better than some accuracy threshold.

B "Protection of linked Tables. 'Linked tables' here refers mainly to tables generated from the same microdata file that have some cells in common, i.e., tables that overlap (e.g., tables that have a column in common). Tables can also be linked via additive relations (e.g., [Table1] = 3*[Table2] + 5*[Table3]).

C "Use of an audit program that can model linked tables to determine if full protection of sensitive cells has been achieved, when theory related to

the mathematical program does not guarantee full protection of establishment or company values, e.g., when backtracking is used. Ideally, the audit program should test protection of contributions rather than just cell values."

This is part of the complexity of protecting data at the company (firm) level."

DA FOR THE 2012 ECONOMIC CENSUS

Prior to the 2012 Economic Census, the cell suppression and auditing programs were rewritten to use C++ instead of FORTRAN.

Problems A and B described in the section on DA for the 2007 Economic Census have always been very difficult to address. Perfect solutions may never be found given the amount of data that the Census Bureau publishes for each census. Improvements have been made to the Census Bureau software to reduce these problems. The improved software was not ready for the 2012 Economic Census, but was in place for the 2017 Economic Census as described below. Problem C was much less of a problem when the Census Bureau began using CPLEX. CPLEX is an LP package that can handle any number of dimensions or relationships between tables. Still, CPLEX cannot process the very large amount of data in an Economic Census at one time (one run of the software on all data), and overloading it can make the software run too slowly. An auditing program needed to be in place when checking cell suppression patterns between parts of the data that were processed separately and when testing new cell suppression software. CPLEX would work perfectly fine for almost any application of cell suppression (other than the census due to its size).

DA FOR THE 2017 ECONOMIC CENSUS

To address the problems in the section on DA for the 2007 Economic Census, the cell suppression team made two algorithmic improvements, <https://www2.amstat.org/sections/srms/proceedings/y2011/files/301855_67474.pdf>.

The first was defining a new data structure called the "supercell" which would improve the protection of primary suppressions. A supercell is the union of all interior primary suppressions and complementary suppressions that exist in an additive constraint. Interior cells are cells in a table that are not totals. This cell union is associated with the contributions to each cell and sums establishment contributions into company sums. The P% rule may then be applied at the company level, rather than the establishment level. This made protection at the company level easier and avoided undersuppression sometimes caused by the problem of protecting data at the company level.

Secondly, the cell suppression algorithm was also improved by finding a way to include the capability to capture two or more linked (interrelated) tables and process them as a single group. This can reduce undersuppression and oversuppression, as well as the amount of backtracking needed.

Work continues to document any other weaknesses in the suppression methodology and software so that they may be addressed in the future. The team is simplifying the code and developing functional specifications for each feature of the program to help with future modifications. They also simplified the input for those that need to run the software in production.

For more detailed information on company-level protection, supercells, linked tables, measuring information loss, and processing special types of data (rounded data and negative values), see <https://www2.amstat.org/sections/srms/proceedings/y2011/files/301855_67474.pdf>.

CONCLUSION

Recently, the Census Bureau has embarked on an aggressive effort to replace its legacy DA methods with modern DA techniques based on formal privacy methods, <https://privacytools.seas.harvard.edu/formal-privacy-models-and-title-13>. Current methods will gradually change with the introduction of formal privacy (Nissim et al., 2018). Most of the current Census Bureau's DA research is focused on formal privacy for all types of data (Nissim et al., 2007). An algorithm operating on a private database of records satisfies formal privacy if its outputs are insensitive to the presence or absence of any single record in the input (Dwork, 2006). The Disclosure Review Board[2] is quickly learning about formal privacy and how it protects Census Bureau data products.

[2] All Census Bureau data products must be approved before dissemination by the Disclosure Review Board.

REFERENCES

L. Cox, "Linear Sensitivity Measures in Statistical Disclosure Control," *Journal of Statistical Planning and Inference*, Volume 5, 1981, pp. 153–164.

L. Cox, "Network Models for Complementary Cell Suppression," *Journal of the American Statistical Association*, Volume 90, 1995, pp. 1453–1462.

L. Cox, "A Comprehensive Methodology for Complementary Cell Suppression in Tabular Data," (unpublished), 2000.

C. Dwork, "Differential Privacy," International Colloquium on Automata, Languages, and Programming (ICALP), 2006, pp. 1–12.

P. Massell, "Cell Suppression and Audit Programs used for Economic Magnitude Data," RR2001/01, *Statistical Research Report Series*, U.S. Census Bureau, 2001.

K. Nissim, S. Raskhodnikova, and A. Smith, "Smooth Sensitivity and Sampling in Private Data Analysis," Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing, 2007, pp. 75–84.

K. Nissim, T. Steinke, A. Wood, M. Altman, A. Bembenek, M. Bun, M. Gaboardi, D. O'Brien, and S. Vadhan, "Differential Privacy: A Primer for a Non-technical Audience (Preliminary Version), Harvard University Privacy Tools for Sharing Research Data, 2018, <http://privacytools.seas.harvard.edu>.

C. Sullivan, "An Overview of Disclosure Principles," RR-92/09, *Statistical Research Division Research Report Series*, U.S. Census Bureau, 1992.