STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

COMMISSION OF THE EUROPEAN COMMUNITIES

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

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

Working Paper No. 21
English only

Topic II: Impact of new technological developments in software, communications and computing on SDC

AMERICAN FACTFINDER: U.S. BUREAU OF THE CENSUS WORKS TOWARDS MEETING THE NEEDS OF UERS WHILE PROTECTING CONFIDENTIALITY

Contributed paper

Submitted by the U.S. Bureau of the Census[1]

Summary : American Fact Finder (AFF) is the United States Census Bureau's new online data dissemination system that accesses data from the 1990 Decennial Census, Census 2000 Dress Rehearsal, the American Community Survey, the 1997 Economic Census, and eventually Census 2000. Most of the data files accessed are summary files with matrices of aggregated data. A special capability of the system is the production of tabulations from a query of microdata files that are behind a firewall (Tier 3). The dissemination of tabulations on-line from a query of the full microdata files requires special techniques for disclosure limitation. The rules and techniques described in this paper are applied to Census 2000 full microdata files Tier 3 access.

I. INTRODUCTION

1. The Census Bureau is the pre-eminent collector and provider of timely, relevant and quality data about the people and economy of the United States. In more than 100 surveys annually and 20 censuses a decade, evolving from the first census in 1790, the Census Bureau provides official information about America's people, businesses, industries and institutions. The Census Bureau guarantees the confidentiality of individual responses for persons for 72 years, as required by federal law (Title 13, Section 9 of the U.S. Code.) The cooperation of citizens, enterprises and other respondents in providing appropriate data needed for necessary statistical compilations largely depends on the Census Bureau achieving a balance between protecting individual privacy and allowing distribution of information in a useful and timely manner. To insure that any data accessible to external users - those outside the Census Bureau - will not disclose information on individuals or entities the Census Bureau has developed a set of disclosure limitation rules and techniques. The purpose of this document is to describe these disclosure limitation rules and techniques for American FactFinder.

II. THE AMERICAN FACTFINDER SYSTEM

2. Prior to 1960 the only data released by the Bureau of the Census were in the form of tables. Tabulations were at the block level for data collected using the short form - the full hundred per cent decennial data. They were at the tract level for data collected using the long form - the sample decennial data. With the proliferation of inexpensive and powerful computer systems and storage it was discovered that some research goals can best be met only if the data were in microdata form. In 1963 the first Public Use Microdata Samples file was generated from a sample of the 1960 decennial census and released to the public. Public Use Microdata Samples contain individual records of responses to questionnaires with unique identifiers (names, addresses, etc.) removed so that the confidentiality of respondents is protected.

[1] Prepared by Sam Hawala (e-mail: sam.hawala@census.gov).

3. As part of the Clinton Administration's initiative to make government more efficient and accessible to the public the Census Bureau announced, in October 1998, the new Internet data-delivery system that significantly expands user access to the agency's vast data resources. This new system complements the Census Bureau's existing Internet site by giving the public online access for the first time to the Census Bureau's largest data collection programs. The new system is referred to as "American FactFinder" (AFF). It is being built under contract with the Census Bureau by IBM Global Services Corp., principal contractor, responsible for systems integration and user-interface design.

4. The first data released via AFF were preliminary reports from the 1997 Economic Census, 1990 Census of Population and Housing files, American Community Survey test and demonstration data and results of the Census 2000 Dress Rehearsal conducted in 1998.

5. The full range of Census 2000 data products will become available via AFF beginning in January 2001, with the release of the state population totals for reapportionment and the detailed population totals (to the census block level) for redistricting. Census blocks are the smallest geographic area for which the Bureau collects and tabulates decennial census data. The blocks are formed by visible physical boundaries (streets, roads, railroads, and bodies of water) and/or by cultural features.

6. Within certain limits described in this paper, AFF users are allowed to define their own geographic areas for non-standard tabulations. Nonstandard tabulations and tabulations for small geographic areas introduce a greater risk of disclosure of individual information. Threats to individual information may affect people’s willingness to cooperate with the censuses and surveys conducted by the Census Bureau. The Census Bureau has therefore made a considerable effort over the years to protect confidential data. The research and development of the disclosure limitation rules and techniques presented in this paper aim at keeping sound the Census Bureau's record for maintaining confidentiality.

III. DISCLOSURE LIMITATION RULES AND TECHNIQUES

7. Through AFF, and within the limits of the data provided, the public will be able to obtain summarized data from Census 2000 over the internet. Data will be provided in table format, displayed back to the user, and in the softcopy file equivalent of the table. When the Census Bureau defines the tables and enables users to choose from the corresponding list of tables, then the resulting table output is referred to as "Tier 2 data". When external users define their own tables, the resulting table output is referred to as "Tier 3 data". In both Tier 2 and Tier 3 the data feeding the tables is "microdata" about individual persons and households. To insure that any data table accessible to external users will not disclose microdata, the Census Bureau uses data recoding, data swapping and query filtering, covering both Tier 2 and Tier 3 data. All the tables in Tier 2 have been approved by the Census Bureau's Disclosure Review Board. Summarized data from Tier 3 will be provided only if it passes disclosure limitation rules.

III.1 Data Recoding

8. Variables such as detailed race, occupation, industry, Hispanic origin, group quarters are re-coded into new variables that show less detail. All continuous variables, such as household/family income, individual income types, cost of electricity, gas, water, fuel, property tax, mortgage payments, gross rent, are top-coded. Cut-off values for top-coding (or sometimes bottom coding) depend on the structure of the distribution of the variable to be top-coded.

9. Re-coded variables are added to the files used by AFF and accessed according to the query and who is making the query. Internal users may use all of the variables and records, as required by their work. External users are diverted to re-coded variables depending on confidentiality issues such as:

- the population groups, geography and variables requested and,
- the rules and population thresholds the Census Bureau requires for the tabulation as a whole and the cells in the tabulation to meet confidentiality requirements.

III.2 Data Swapping

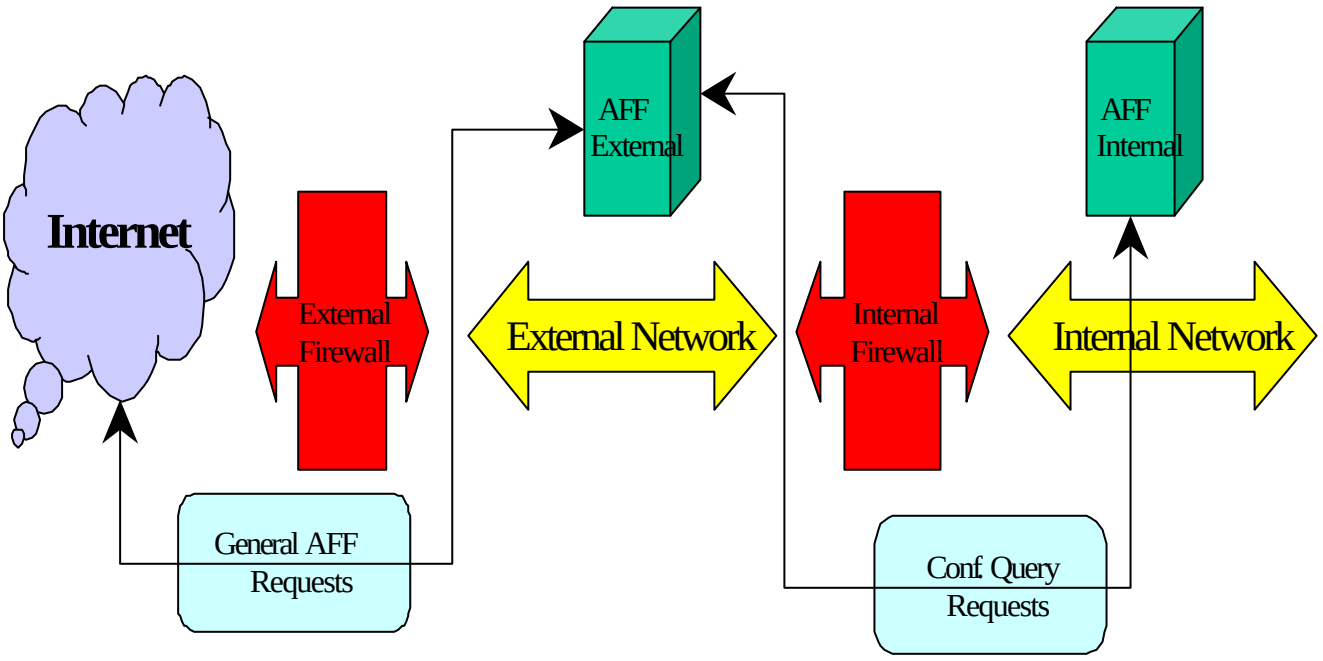
10. The swapping technique was used for the 1990 Census of Population and Housing. This technique will be used for Census 2000 one hundred per cent - short form - and independently for the Census 2000 sample - long form - data. The technique consists of swapping pairs of records selected as having the greatest disclosure risk. In particular, records that are unique with respect to a set of variables are marked for swapping. Entire household records are swapped from two different geographic locations.

11. The variables that make a record unique are referred to as key variables. A record will be selected for swapping with a probability inversely proportional to block size. Records of households with unique race categories in the block will have an increased probability of being swapped. The swapped records match on a set of demographic characteristics but are in different census blocks.

12. AFF will use the swapped data files as input. All tables to be publicly released (on paper, on tape, through AFF, etc.) to anyone outside the Census Bureau will be generated from the swapped data files.

III.3 AFF Technology

13. The Census Bureau Internet firewall is configured to permit web-service requests which originate on only AFF's external server, and which terminate on AFF's internal server. Communication from any other external machine to the internal server is blocked by the firewall.



14. Any request from an external user is routed along the path “General AFF requests” as shown in the illustration. By definition, there is no confidential information sent along this path. The external server receives requests, which require Title13 data, and re-transmits them over the path marked “Conf. Query requests”. Again, no confidential information is sent along this path. Network packets on this second path cannot be observed by outsiders because the Internet Router does not propagate those packets to the Internet. No traffic from the Internet can talk to the Internal Server because the firewall permits communication only between the two American FactFinder servers. There is no need to encrypt information on either path because there is nothing confidential to be observed.

15. Most of the data files accessed through AFF are summary files with matrices of aggregated data. These are found in Tiers 1 and 2. The special capability of the AFF system is the custom-made

production of tabulations from microdata files behind the firewall. This is called Tier 3. Through Tier 3, the user can obtain very specific data for very specific geographic areas. We discuss next the disclosure limitation rules designed for Tier 3 queries. The rules are implemented as a set of filters. There are two types of filters: Query filters and Results filters.

III.3.1 Query Filters

16. The purpose of the query filter is to detect those queries that will not pass disclosure limitation before they are submitted for execution. This saves Tier 3 system resources and saves time for external users by telling them relatively quickly whether or not their query has a chance to pass disclosure limitation rules.

- Cross-tabulations must be created from geographic areas and/or a Census Bureau's predefined list of non-geographic variables.
- The query's geographic variable must meet a minimum threshold. The system determines if the query requests small areas - blocks, block-groups or user-defined geography with a population size that is less than average tract size (4060 in 1990), medium areas (population size 4060-99,999) or large areas (population size 100,000 or more.)
- According to the population size of the area or areas requested, the system permits the use of appropriate combinations of short, medium or long lists of predefined categories of race, Hispanic origin, group quarters and other sample variables in the cross tabulation. Only top-coded variables may be accessed.
- When reporting cell values for a split area, the splits are ignored - the data given is for the entire area.
- The maximum number of variables used to create an overall table is three, excluding the geographic variable.
- Depending on the cross-tabulation variables, the user must select from a list of derived measures possible (means, medians, ...) to be reported within a cell.
- Derived measures are provided only if the corresponding counts are provided.
- If the estimated execution time or size of output exceeds a limit set by the Bureau, the query is disallowed.

17. If the query filter rejects the user's original query, the user is given a choice to change the population threshold involved, or to quit. For example, if a user submits a query using race-medium re-code, which would pass a medium area population threshold, and the area population threshold for this query turns out to be a tract, then the user is given the choice of using a race-short re-code appropriate for the small area population threshold.

18. If a query passes all the disclosure limitation rules for the query filter, the query is passed from the external server outside the firewall to the internal server inside the firewall to the full microdata files for computation. The full microdata files contain all of the predefined categories for race, Hispanic origin, group quarters and modified sample data variables.

III.3.2 The Results Filter

19. The results filter provides a final check on the values in the cells of the resulting table. The table is provided back to the external user if and only if the table passes all the disclosure limitation rules in the results filter.

- When geographic sub-tables are requested in the query, the disclosure limitation rules are checked separately for each geographic area - corresponding to the geographic sub-table -, as if the user had submitted separate queries, one for each geographic area.
- The median cell size in a requested table cannot be less than a parameter set by the Bureau.
- The mean cell size in a requested table cannot be less than a parameter set by the Bureau.
- The ratio of the number of cells with a count equal to one to the total number of cells in a requested table cannot be more than a parameter set by the Bureau.

- When a user defines re-codes for a query, the Tier 3 system first computes the results as if no re-codes had been requested, and checks these results against the disclosure limitation rules. Secondly, if the results satisfy the rules, they are aggregated to reflect the re-codes requested.

20. The Census Bureau will continue to test the foregoing AFF Tier 3 filters on the Decennial 2000 data as that data become available for any needed adjustments.

Reference

Zayatz, Laura, and Rowland Sandra (2000), "Disclosure Limitation for Census 2000", Proceedings of the Section on Survey Research Methods, American Statistical Association, to appear.