Managing the census data base—Data description, acquisition, and manipulation

by JAY-LOUISE WELDON
New York University
New York City, New York

ABSTRACT

The U.S. Bureau of the Census collects and maintains a broad base of socio-economic and demographic data on population and housing in the United States. Summarizations and statistical information from this data base are made available to the public in printed reports. In addition, portions of the data are made available in machine-readable form, and thus are appropriate input for further research studies.

For most potential users, however, many obstacles exist in making use of the machine-readable Census data. These obstacles may be classified into three areas: the identification and selection of relevant data items from the available data to support a given research objective, the acquisition of accurate and complete data, and the physical storage and processing of such a large data base.

This paper discusses each of these problem areas and identifies the factors involved. Illustrations of possible solutions are presented, based on the author’s experience with the Census data base at the University of Pennsylvania. Guidelines for dealing with such problems are provided for other prospective users. Finally, some general recommendations are made for avoiding or reducing these types of problems in future Census data products.

INTRODUCTION

The basic Census data base, as maintained by the U.S. Bureau of the Census, contains the actual (or encoded) responses for each individual who completed a Census questionnaire in the 1970 Census of Population and Housing. Portions of this data base are made available to the public in three forms:

(i) printed census reports, which summarize various responses for certain geographic areas;
(ii) summary data tapes, which also contain response data summarized by geographic area but in computer-readable form; and
(iii) sample data tapes, called the Public Use Sample, which contain selected data records from the basic Census data base, also in computer-readable form.

The printed reports are the most frequently available and immediately usable of these three forms. These reports are available in most libraries and provide summarizations of basic housing or population characteristics for the most commonly accessed geographic regions. For example, from the reports a researcher may gather data on counts of housing units or on population by age or sex for counties or states. While convenient, the printed reports may not suffice for many researchers. For example, data may be required on more specialized data items, or by smaller geographic areas. For computer analysis, the researcher may wish to obtain a large volume of data records in computer-readable form. In these latter cases, the summary tapes and the Public Use Sample are required as data sources.

The summary data tapes, organized into six parts called “Counts,” contain the summarizations from which the printed reports were developed. The six Counts differ with respect to the data tabulations they contain and the size of the geographic areas for which these data are summarized. Each tabulation represents the result of cross-tabulating one or more questions from the Census questionnaire. As such, more than one data item (or cell) may be contained within one tabulation. For example, the tabulation “Population by Race” requires four data cells, corresponding to three racial groups plus the total. The geographic area for which the tabulations are summarized are identified in each record by a set of codes. The summary data tapes contain summarizations for states, counties, and Standard Metropolitan Statistical Areas (SMSA’s), as well as for smaller areas, such as census tracts, blockgroups, and Minor Civil Divisions (MCD’s).

The Public Use Sample tapes contain data records of actual questionnaire responses with most identifying information removed. These records are selected according to different sample sizes and for different
geographic areas, and grouped into six files according to these two characteristics. The data items in the Public Use Sample records are coded responses rather than counts or tabulations. However, coded geographic information is used to identify each record, as in the summary data.

The decision to use and/or acquire either type of computer-readable Census data forces the researcher/user to confront a number of problems. The problems encountered can be classified into three general areas: data description, acquisition, and manipulation. Problems in data description arise when the user attempts to identify and/or select the data items and geographic areas most appropriate for his/her research. Data acquisition problems center around the user's ability to obtain complete and accurate data in a form compatible with his/her computer installation. Manipulation problems are largely due to the overwhelming size of the Census data base (over two thousand reels of computer tape, as originally produced from the Bureau of the Census). The user/implmentor is faced with questions of how to store and access this large data base in the least expensive fashion.

This paper addresses each of the three problem areas and presents guidelines for potential census users in each area. These guidelines are based on experience with the implementation and use of the Census data base in a university environment. In addition, recommendations are made for improving future Census data products and services.

IDENTIFICATION OF DATA TABULATIONS AND AREAS

Before evaluating alternatives for Census data acquisition and manipulation, the prospective user must be able to identify and select those data items which are required for his research. The two-dimensional logical structure of the computer-readable Census data base requires that a potential user select both the tabulations he wishes to use and geographic areas for which these tabulations are summarized.

Selection of tabulations

Problems in identifying and selecting appropriate tabulations from those available arise due to (i) duplication of variables, and (ii) questions of data reliability. The same variable (e.g., housing unit value) may appear in several summary data counts, though in a slightly different context in each (see Table I). User aids, such as the 1970 Census Data Finder, should be consulted to locate the tabulations that may be of interest under more general subject headings. In every case, the user should refer to the formally defined Census Concepts to insure proper interpretation of the terms used, e.g., "mixed parentage" under race.

<table>
<thead>
<tr>
<th>Tabulation Number</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregate $ value (of housing unit)</td>
<td>17</td>
</tr>
<tr>
<td>Units for which value is tabulated by value class, by race</td>
<td>35</td>
</tr>
<tr>
<td>Aggregate $ value for units with all plumbing facilities</td>
<td>27</td>
</tr>
<tr>
<td>Value, type of household, age and race of head</td>
<td>57</td>
</tr>
<tr>
<td>Aggregate $ value</td>
<td>1</td>
</tr>
<tr>
<td>Value</td>
<td>22</td>
</tr>
<tr>
<td>Aggregate $ value</td>
<td>1</td>
</tr>
<tr>
<td>Value, occupancy status, and race of head</td>
<td>52</td>
</tr>
</tbody>
</table>

Reliability problems include questions of sampling, allocations, and actual errors in the data. Since a minimal number of questions (e.g., five on Population) were asked of all Census respondents, all other tabulations are based on questions asked of a sample of respondents. These sample counts are weighted to approximate the total, but actual totals and weighted totals may not agree in every case. Further, users of tabulations or small geographic areas, such as block-groups or enumeration districts, may find that the totals on a sample question reflect the responses of an unacceptably small number of actual responses. For example, a 5 percent sample item on a 300-household census tract would reflect only 15 actual households. For items that are further divided, e.g., into rental or owner-occupied, the count will be even lower.

"Allocations" are substitutions made by the Bureau of the Census for data items missing from the original Census responses. This procedure is documented and data indicating the allocations that were made exist within the Census data base. However, a user who makes a hasty selection of tabulations without investigating the impact of these pseudo-data may be jeopardizing his results. A high proportion of allocations for any data item affects the reliability of all tabulations utilizing that item.

Certain tabulations may be missing entirely from the summary data tapes. In cases of small area tabulations, the Bureau of the Census often suppresses data items that may impair confidentiality. For example, income tabulations might be suppressed for block groups containing only two individuals older than 65. These situations are indicated by a series of suppression codes placed into the data record at the time of suppression. The user must simply be aware of their meanings and make sure any access software handles these zero-filled fields accordingly.

Incorrect tabulation data discovered by the Bureau of the Census are generally reported in Bureau publications, such as the Data User News (formerly called

From the collection of the Computer History Museum (www.computerhistory.org)
Small Area Data Notes). However, it behooves the user to be alert to these announcements, since error correction is the responsibility of the owner of the data, not the Bureau. Further, such errors are reported as they are discovered and are not reflected in the primary documentation accompanying the receipt of Census data. Thus, a diligent user is forced to scan back issues of Census publications to spot errors that may affect his analyses.

Selection of geographic areas

In selecting geographic areas, the user is faced with problems involving geographic codes, record-sequencing, area matching, and missing data. All geographic areas represented in the Census summary data are identified by numerical codes. The Census Bureau provides a user aid called the Master Enumeration District List (MEDList) in which each code is associated with the corresponding state, county and area or place names. In order to access the summary data for any area, the user must know the codes of the areas desired, since no names exist on the tapes. While codes are also used in the Public Use Sample, no corresponding list exists for these codes. To further complicate the issue, the codes used in the P.U.S. are not always the same as those in the summary data.

A user who is dealing with regions or areas that do not coincide with census geographic areas must match the census designations with the region desired. Doing this may require overlaying the boundaries of these “special” areas on Census maps or using the Address Coding Guide (ACG). The latter is a list of block faces and their associated census tract designations originally used to properly sort Census responses. Using the ACG may enable a user to convert his “special” regions into tabulated areas. Or, at worst, he can describe these regions in terms of addresses and request a special tabulation from the Bureau.

Even for fairly standard geographic areas, the sequencing of the summary tapes may prove an obstacle to easy access. A “sequencing key” field, made up of various geographic codes plus category codes added for the summary, exists in each data record and the order of the records in the file is based on this field. As a result, the order of the data records may differ substantially from the order a user might expect. The physical sequence, of course, influences the length of search and access time for data records in a sequentially organized data file, thus affecting the user’s processing costs.

Missing data records that are lost due to processing errors (e.g., in tape copying) or are truly non-existent (because no tabulations exist for that Census tract) present another costly dilemma for the user. Such omissions are usually discovered as the data are processed, and thus may force the user to suspend or redesign his analyses in mid-stream. Unfortunately, preliminary completeness checks are also costly unless they can be performed as a by-product of tape copying (e.g., a list of all geographic areas represented on file).

ALTERNATIVES FOR DATA ACQUISITION

To the potential Census data user, the next step is obtaining access to the data he or she needs. In addition to the Census Bureau itself, there are several other types of agencies through which Census data are available. These agencies fall roughly into three categories: Census-designated processing centers, data suppliers, and suppliers of computer services.

Organizations which are designated by the Census Bureau as processing centers provide data retrieval and tape copying services for census data users. These centers may be private or public, and operate for profit or on a nonprofit basis. They are not controlled or certified by the Bureau of the Census, but once recognized they are added to a list of such centers which is available from the Bureau on request. Since each center is operated independently, a user may find from one tape reel to the entire Census available at the centers in his location.

Various private companies serve as intermediaries between the Census Bureau and Census data users. These organizations operate in a “value-added” mode, supplying users with data that is in a more efficient format than that provided by the Bureau. A most extensive variety of services, such as special tabulations and software products, also available from such agencies. Two organizations which supply Census data and services in this vein are the National Data Use and Access Laboratories of Rosslyn, Virginia (DUALabs), and the National Planning Data Corporation of Ithaca, New York.

Some organizations that market computer batch processing or time-sharing offer Census data services to their users. Such services include access to selected portions of the Census data and/or software access and display packages. Those services available over time-sharing networks, such as CENSAC on National CSS, Inc. and CENSTAT on CDC’s CYBERNET, usually include data access plus software. Smaller services bureaus, such as UNI-COLL, Inc., Philadelphia, Pa., offer customized software systems to process data obtained by the users.

The advantages to using any of these secondary sources for Census data are ease of access and the availability of special services, e.g., consulting or software. However, ease of access is limited to those data made available by the organization selected. Many of these agencies concentrate on a few oft-requested tabulations or on one geographic region. Even those from which all data is available may charge more for access to data items that are rarely used. Further, the software services provided are usually standardized, with
customized processing or special tabulations available at extra cost.

A user who wishes to obtain the original tabulations and do his own processing has the choice of dealing directly with the Census Bureau or with one of the secondary sources that provide tape copying services. The advantage of the latter is the receipt of data in a much more efficient form. For example, as offered by the Bureau of the Census, most Counts are supplied on a one state per tape reel basis—quite uneconomical in terms of storage and processing costs.

MANIPULATION OF THE CENSUS DATA BASE

The Census Bureau had maximum transferability as its objective in structuring the files of Summary Data and the Public Use Sample. In response to this objective, the data are made available in the most standardized fashion: in sequential organization, on magnetic tape, recorded at low density, encoded in one of two standard character codes (EBCDIC or BCD), often one state per reel, and with equal, fixed-length records. These characteristics, however, inflate the size of the Census data base and make its manipulation most inefficient. To improve storage and processing efficiency, the user should investigate changes in form and size, storage medium, and organization.\(^\text{18}\)

The size of the Census data base is unnecessarily inflated by the use of equal-length data records and standard character codes. The former necessitates a large proportion of blank characters in many files, to fill fields unused in those records. The latter requires larger data fields for numeric values than would be necessary if binary representation were used. After making the above changes the user may be well-advised to pursue further reductions in data base size through the application of data compression techniques.\(^\text{9}\) Zero suppression and simple character packing (two digits per byte) achieved compression ratios from 43 percent to 73 percent in the University of Pennsylvania Census data base.\(^\text{10}\)

An appropriate change in storage medium, e.g., from magnetic tape to disk, must be determined with respect to the user's processing installation. The factors involved include the amount of data to be stored, the relative costs of tape versus disk processing, and the extent and frequency of processing anticipated.\(^\text{11}\) In the University of Pennsylvania environment, transfer of user-requested tabulations from tape to disk as needed resulted in a sizable reduction in processing costs for a modest (and temporary) increase in storage charges.

Efficiency in accessing Census data can be further improved by changes in the organization of the data. Access to the Census data, as originally produced, is constrained by all the limitations of sequential files: variable access times for data dependent on location within the file, linear search for desired data, and a minimum number of data relationships actually represented by the file structure (e.g., geographic proximity or association by household). However, radical variation from the original organization puts the user and his data at variance with the published documentation. While this may be acceptable for a small portion of data, users maintaining a full Census data base, e.g., for a university community, would find the extra documentation burden unwieldy.

The approach to change in Census data organization at the University of Pennsylvania was to retain the original sequence of data records while regrouping these records into separate sub-files, and building an index to the new sub-files. Sub-files were defined either by geographic region (e.g., counties) or by record count (e.g., 500 records). In both cases, access costs for data in these files was substantially (50 to 60 percent) less than for access to the original state file. The level of indexing employed, however, still required sequential search within the sub-files and use of the index to locate specific records was limited by the sequencing imposed on the original file.

CONCLUSIONS AND RECOMMENDATIONS

The problems that face potential Census data users with respect to data description, acquisition, and manipulation indicate that the Census data base is not readily available to most researchers. With the next decennial Census rapidly approaching, it behooves those who have experienced some of these problems to suggest ways in which they might be ameliorated.

One recommendation would be that the Census Bureau assume the role of data base administrator (DBA) for the Census data base. That implies that while making the data available to a wide range of users, the Bureau would retain control over the data base. A radical means of doing this, that is now technically feasible, would be to make the Census data base available over a nation-wide network which users pay to access. In this way, the Bureau would physically retain control over the data and be able to assure their physical integrity and structural validity. Access methods and utility programs could then be standardized and shared amongst all Census users.

Another approach would have the Census Bureau retain only logical control over the data base. This would be an extension of the Bureau's current mode of operation. Portions of the Census data base would be physically transferred to users (for a fee) but the Bureau would be responsible for promoting accurate and efficient manipulation of Census data by applying the tools of documentation and standards.

To fulfill such a role, Census documentation would have to be improved in the following ways:

(a) It must be structured and cross-referenced as an integrated hierarchy, where the user could start at a general subject area and easily locate...
all relevant information down to the nature of specific tabulations and areas;
(b) it must be dynamic, so that continuing notices, error advices, and new data products could be easily reflected when they are pertinent; and
(c) it should be priced in such a way that participating research organizations and libraries would be motivated to purchase the entire set, not isolated fragments.

The concept of standards with respect to Census data would also have to be broadened. In addition to formatting standards required for transferability, the Bureau could provide standard access routines, oft-requested utility routines (e.g., for elementary statistics) or interface routines for the most common statistical packages, e.g., SPSS or BIOMED. For users with incompatible processing environments, these standards could be set forth in the form of guidelines or logic flowcharts, from which users could create their own software.

As with any other data base, once administrative tasks are removed from each group of diverse users and returned to a centralized agency of control, access and integrity will be improved. To do less with a data base of the size and importance of the Census, is to implicitly restrict public access to a valuable national resource.

ACKNOWLEDGMENT

The author wishes to express her appreciation to Mr. David F. Sanford, Wharton Econometric Forecasting Associates, Philadelphia, Pa., for his consultation and advice on this paper.

REFERENCES
