Journal of Official Statistics
Vol. 2, No. 1, 1986, pp. 57–70

Current Research Problems Encountered at the U.S. Bureau of the Census

Kirk M. Wolter[1]

Abstract: Fifteen current research projects at the U.S. Bureau of the Census are discussed. The projects are in five broad categories, including survey methodology; census undercount and adjustment; time series analysis; disclosure methodology; and mathematics, automation, and artificial intelligence.

Key Words: Sample survey; census undercount and adjustment; time series; disclosure; artificial intelligence.

1. Introduction

This is a cursory review of a number of research areas undergoing current development at the U.S. Bureau of the Census. For organizational purposes, I have grouped the areas into five broad sections

2. Survey Methodology
3. Census Undercount and Adjustment
4. Time Series Analysis
5. Disclosure Methodology
6. Mathematics, Automation, and Artificial Intelligence.

In what follows, I give a brief description of each of 15 specific research areas within these broad categories. I will give some discussion of the importance of the research areas and where they fit in the Census Bureau's program of surveys, censuses, and other statistical products. I also provide some indication of the status of the work and of the methodological or technological approaches being employed.

At the end of each project description, I provide the name of our principal investigator on the project. Details about the project may be obtained directly from this person at the Statistical Research Division, U.S. Bureau of the Census, Washington, D.C. 20233.

2. Survey Methodology

Historically, much of the Census Bureau's research has been devoted to the development of (or improvements in) survey methodology. The accomplishments of Hansen, Hurwitz, and their colleagues in the 1940s and 1950s are well-known. This tradition of innovative work on survey methodology continues in the decade of the 1980s, and in this section, I describe six areas of current research. Precipitating factors that led to the development of these projects include

[1] Chief, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C. 20233.

(i) dwindling federal budgets for statistical activities cause emphasis on telephone surveys and on other less expensive modes of data collection;
(ii) the Census Bureau is currently redesigning a number of its major household surveys based on information from the 1980 decennial census;
(iii) the Census Bureau has recently started a new longitudinal survey and this raises a number of new and unsolved methodological problems.

2.1. *Telephone Survey Activities*

The Census Bureau is involved in research and development for telephone surveys on two fronts. For surveys of business establishments, the effort is focused on the development of an Integrated Survey Processing Network (ISPN) that will carry out a number of survey data collection functions using automation and high technology. For surveys of households and individuals, the focus is on computer assisted telephone interviewing (CATI) and random-digit-dialing (RDD) sampling. Both of these efforts will make major use of centralized telephone interviewing.

The Census Bureau is currently streamlining the list sample mailout and data collection operations for its surveys of wholesale and retail establishments by transferring these responsibilities from its 12 regional offices to a consolidated facility in Jeffersonville, Indiana. The ISPN entails the development and maintenance of a data base containing all information required for mailing and data collection. This master file will be accessed by a network of microcomputers in Jeffersonville with software to

a) prepare mail forms,
b) check-in mail responses,
c) capture and edit mail responses,
d) perform CATI (call scheduling, interview probes and skips, editing, etc.) for both delinquents and edit failures, and
e) generate management information reports for process control.

ISPN's production testing for the mailout and data collection modules started in 1984. Complete implementation of all modules will be phased in over the next five years.

CATI for demographic surveys is being tested for two general applications: "warm-contact" CATI and RDD CATI.

Several of the Census Bureau's large current surveys, such as the Current Population Survey (CPS) and the National Crime Survey (NCS), involve repeated interviews of the same units over time. The first interview is conducted in person and subsequent interviews (so-called warm-contacts) are usually done by telephone if possible. Interviewers receive their assignments from one of 12 regional offices and work out of their homes to complete the enumeration. The Bureau has initiated a program of research and development to explore the cost, quality and timeliness benefits of centralization of the telephone interviews with or without CATI. The primary goal of the research is to determine the feasibility and desirability of centralized warm-contact CATI and to resolve important operational uncertainties about the system. These include (a) the ability to transfer cases between the centralized facilities and the regional offices within existing time constraints, (b) the impact on the existing regional offices and field interviewers of the transfer of all telephone cases to a centralized facility leaving them personal interviewing cases only, and (c) the change in costs and quality due to centralization.

The first test of warm-contact CATI began in June, 1985 at the Bureau's new telephone research laboratory in Hagerstown, Maryland. A test sample of approximately 3 000 households per month (phased-in over four

months) were interviewed four times, with the first interview in person. The major objectives for the study are to develop the basic system and to test both the case transfer and the CATI components of the system.

Another project for the Hagerstown office is the development and testing of a system for conducting interviews by CATI in conjunction with RDD sampling. In this case, the initial interview is a "cold-contact" by telephone. Most of the research is confined to feasibility testing, and response rate investigations. The NCS questionnaire will serve as the test vehicle. A sample of 450 cases per month will be selected using RDD methods. Interviewing started on April 1, 1985 and will continue until September.

The Statistical Research Division (SRD) contact person regarding this research area is Paul Biemer.

2.2. *Using All Unbounded Interviews in the National Crime Survey*

The National Crime Survey (NCS) began in July of 1972 and is conducted by the Bureau of the Census for the Bureau of Justice Statistics. In this survey, the members (12 years or older) of a selected housing unit are interviewed seven times at six-month intervals. At any given time, the NCS sample consists of one incoming rotation group that is interviewed for the first time (first time-in-sample).

The first of the seven interviews, the bounding interview, is used only to set a time frame in order to avoid duplicating reported crimes in subsequent interviews. It is not used to produce the estimates of victimizations. This is a rather costly feature of the NCS design, since about one-seventh of the collected data is not used in tabulating results for publication. This feature was adopted because studies conducted before the survey began showed that, without bounding, crimes occurring before the reference period were frequently reported in the reference period.

On the other hand, all data from returning rotation groups are used to produce the estimates. In other words, the interviews are treated as bounded, that is, as if the interviewer can remind the respondent of the incident reported in the previous interview. However, since the NCS uses a probability sample of housing units (the physical structure) rather than of households (persons living in the housing unit), not all of the interviews conducted in returning rotation groups are bounded. One example is the situation where a new household moves into a selected housing unit.

Our current research involves studying the bounding effects so that: (1) the bounding interview data can be appropriately adjusted in order to be used in the estimation procedure; and (2) data from unbounded interviews from returning rotation groups can be appropriately adjusted in order to lead to more accurate estimates. One method that is currently being investigated is to estimate the difference in crime level between two consecutive years by using all interviews (bounded and unbounded), the assumption being that the biases from the bounding interviews would cancel. Other types of estimators that will be studied include ratio- and model-based estimators. The comparison of the mean square errors of the different estimators is the focus of the analysis.

This research has two main goals. The first one is to reduce the variance of the annual estimates (or equivalently reduce the sample size while keeping the variance the same) by using the bounding interviews. The second goal is to reduce the bias of the annual estimates by appropriately adjusting the unbounded interviews that are presently used in the estimation procedure (i.e., those from returning rotation groups). Applications of this study may include other panel and longitudinal surveys where interviewers inquire about occurrences during a specified recall

period and where an event can be erroneously telescoped into that period. One example is the Consumer Expenditure Survey (CES), which, like the NCS, does not make use of the first interview in the estimation procedure.

The SRD contact person regarding this research area is Lawrence Ernst.

2.3. *Measuring Redesign Effects*

Ongoing surveys are periodically redesigned to reflect changes related to the population of interest. Redesigns are often phased-in over several data collection periods. Data collected during and after implementation of the redesign may be affected simultaneously by changes in the population and by the redesign itself. If so, estimates produced during this period are not directly comparable to pre-redesign estimates. When the redesign effects can be estimated, adjustments can be applied to the survey estimates to make them directly comparable to pre-redesign results.

We have taken a linear model approach to this problem. Although the exact form of the model will depend on the survey, it will always include a parameter representing the characteristic to be estimated from the sample at each time point (e.g., an unemployment rate or a crime rate). The remaining independent variables can be classified as basic survey related effects or redesign related factors which are thought to affect the estimation of the parameter of interest. For example, if the survey design requires individuals to be interviewed for several consecutive time periods, then there may be a time-in-sample effect on the response (sometimes referred to as rotation group bias). This would be classified as a basic survey related effect since it existed prior to the redesign and will persist in some form after the redesign is completed. Examples of redesign related effects include the effect of a change in sampling frames, the behavioral effect of inexperienced interviewers in new sample areas and interviewers to be terminated in outgoing sample areas, and the effect of administrative burdens and disruptions associated with the redesign implementation. The dependent variable in the model represents the response corresponding to the characteristic of interest for individuals in a particular demographic group (e.g., unemployment among teenagers or criminal victimizations among blacks).

Parameter estimation is accomplished by the method of generalized least squares where the estimated weight matrix should reflect the sample design. If all model parameters are estimable, this approach provides estimates of the characteristic of interest and the remaining effects as well as estimated standard errors. Additional information and special studies may be necessary to decompose the aggregate redesign effect into its major components.

Models have been developed for the CPS (which measures labor force characteristics) and the NCS redesigns. Both will be utilized to varying degrees along with other methods to monitor the effects of these redesigns on estimates of selected characteristics. The CPS redesign is currently in progress and the NCS redesign began in July, 1985.

The SRD contact person regarding this research area is Edward Gbur.

2.4. *Minimum Variance/Composite Estimation*

Several of the household surveys conducted by the Census Bureau employ a rotation sampling scheme with a partial overlap of the sample from one time period to the next. For many key characteristics, estimates at two different time periods from the same panel are highly correlated. Under such conditions, estimators which are linear combinations of the estimators from each panel over all available time periods can result in estimates with greater precision than estimators (called elementary estimators) which use only the data from the time period to which the estimate refers.

Two classes of estimators that use data from other time periods will be considered here. The minimum variance linear unbiased (MVLU) estimators have the obvious advantage implied by the name. However, they have the drawback of computational complexity principally caused by the necessity to store and manipulate elementary estimates from each panel for each time period. As a result, MVLU estimators have not yet been used for any Census Bureau survey.

A computationally simpler alternative to MVLU estimation is composite estimation. A composite estimator for the current time period is a linear combination of the composite estimator for the preceding time period and elementary estimators from each panel for the current and previous time periods. At present, the only household survey at the Census Bureau that uses composite estimation is the basic portion of the CPS, the monthly labor force survey.

Recently, work has begun at the Census Bureau to determine whether composite or minimum variance estimation would be appropriate for other household surveys. As part of this work, it has been shown that MVLU estimators can be expressed in a form rivaling composite estimators in computational simplicity for a survey for which each panel is interviewed for two successive time periods and then permanently dropped, with a 50 percent overlap between successive time periods. This type of rotational pattern is present for those supplements to the CPS on specific topics that are asked on an annual basis, and also for annual estimates for the new Survey of Income and Program Participation (SIPP). Current MVLU and composite estimation research is focusing on these surveys with the emphasis on MVLU estimators, since computation complexity will not be a problem.

It has been observed for several of the household surveys that the expected value of an elementary estimate from a panel is dependent on the number of occasions that the corresponding panel has been a sample. In such circumstances, MVLU and composite estimators will in general have different expected values than the average of the elementary estimators from each panel. If there are large time-in-sample effects for the surveys under study it may be necessary to either modify the MVLU estimators to reduce the resulting problem or not use this type of estimation at all.

Currently, for the CPS supplements empirical work is in progress using data from the October School Enrollment Supplement to obtain correlations for key characteristics and to study time-in-sample effects for that survey. For SIPP, similar empirical work is not possible as yet because the survey has not been in existence long enough to have even two full years of data. At present, research areas and data needs are being determined for SIPP.

The SRD contact person regarding this research area is Lawrence Ernst.

2.5. *Current Industrial Report Research*

The Current Industrial Report (CIR) surveys are used to estimate the total production of a wide variety of commodities produced in the U.S. Information from these surveys is used by government economists in macroeconomic input-output analyses to gauge the performance of the U.S. economy. Reports produced by these surveys are used by trade associations to disseminate information about sales of specific commodities of interest to their members. The Census Bureau is currently developing a methodology for variance estimation for the special sampling designs often employed by CIR surveys.

Of the approximately 100 CIR surveys, about 30 percent use a "cutoff" survey design. That is, firms deemed to be above a specific cutoff (determined by amount of sales or

number of employees) are included in the survey. This type of design is efficient since small establishments, which are thought to contribute little to the totals (usually less than two percent), are not sampled. Since small establishments below cutoff give data of dubious quality, if they respond at all, the loss incurred by not including them in the sample is perceived to be trivial. Ratio estimation procedures are then employed to estimate the total for all firms, both above and below the cutoff.

Although this type of cutoff survey design is thought to be efficient, methodology to assess the quality of these estimates and the survey design strategy is undeveloped. Research is underway to obtain precision estimates when cutoff sampling is appropriate as well as to obtain decision criteria by which alternate designs may be used when cutoff survey designs are not optimal. The prediction-theory approach to survey sampling is being explored as a means of producing variance estimates for the CIR survey estimators. The notion of variability assumed here will have a somewhat different conceptual basis than that employed in the Census Bureau's many probability samples.

The SRD contact person regarding this research area is Nash Monsour.

2.6. *Imputation for the Survey of Income and Program Participation*

The Survey of Income and Program Participation (SIPP) is a longitudinal survey in which the members of each sampled household are interviewed every four months for two and a half years. At each interview a person is asked questions pertaining to the previous four month period, and a question may have one response for the entire period (wave variable) or one response for each month (monthly variable). The questions asked are divided into subsets of related questions, with each subset called a record type. Nonresponse for a person at an interview may consist of total nonresponse, one or more record types missing (record nonresponse) or individual questions missing (item nonresponse).

SIPP is designed to collect data in order to estimate longitudinal parameters such as duration of unemployment and length of time receiving food stamps, as well as cross-sectional information such as individual earnings and family wealth. It is crucial that the imputation methods maintain longitudinal as well as cross-sectional consistency among all the observed and imputed variables.

Some initial research on longitudinal imputation was completed at the Census Bureau. This research concentrated on maintaining longitudinal patterns for one variable at a time for item nonresponse. The method derived is a matching procedure which uses only the values of the given variable reported in other months. It works as follows:

1. Take 12 months of data from one rotation group and a single variable.
2. For a specified person who has one or more months missing, randomly select from the persons with 12 complete months a donor that matches the specified person in all its reported months.
3. Fill in the missing months for the specified person with the corresponding values from the donor.

Current research is examining the feasibility of model-based imputation for item and wave nonresponse using any appropriate multivariate relationships between variables. For categorical variables we are fitting logit models in which the explanatory variables for a specified month are

1. Several previous and future months of the variable being modeled.
2. The values of closely related response variables in the specified and the closest reported months on either side of the specified month.

3. Current wave values of selected demographic variables.

For continuous variables we are fitting regression models with the same types of explanatory variables. Imputations using these models will then be evaluated on their ability to maintain the observed distributions of the variables being imputed as well as on their cost versus other imputation methods.

The SRD contact person regarding this research area is Paul Biemer.

3. Census Undercount and Adjustment

In any census of human populations, the number counted seldom equals the true number of people. The Census Bureau's estimates show that the 1970 Census counted two percent fewer people than it should have. Current estimates are that the 1980 census figures were off by less than one percent at the national level.

In recent years, the acknowledged undercount has become a source of concern because it may affect a state's representation in Congress and the federal grants-in-aid that it receives. Some federal funding programs use census numbers to distribute money to local jurisdictions. If the population missed by the census were spread evenly across the country, there would be little effect upon resource allocation. Because the census tends to miss more of the minority populations, places with a high percentage of minority residents may not receive their fair share of funds.

Up to now, the Census Bureau has not adjusted the official counts for a measured undercount. This position came under legal challenge in 1980, and future court decisions could require the Census Bureau to adjust census counts. In the meantime, we are continuing research aimed at developing reliable adjustment methods.

3.1. Direct Coverage Estimation

Coverage estimation is often based on case-by-case matching studies. A sample is drawn from a source independent of the census. The proportion of this sample that is not found in the census through matching procedures is used to estimate the proportion of the total population not enumerated in the census. Post-enumeration surveys, administrative record checks, and reverse record checks are all forms of case-by-case matching.

It is also possible to estimate coverage at the national level by using demographic analysis. This method uses birth and death records, immigration data, administrative records, and the results of previous censuses to construct an estimate of the census day population. The difference between the estimate and the census count is interpreted as an undercount.

The Census Bureau has conducted large coverage evaluation operations after every census since 1950. Case-by-case match studies were conducted following the 1950, 1960, and 1980 censuses; demographic analysis has been used since 1950.

We are committed to a program of research that aims to develop a methodology for measuring undercount that is sufficiently accurate so that in the 1990 census, adjusted counts are more accurate than the counts achieved in the basic enumeration. Major tests of the undercount measurement methods will be conducted in 1985, 1986, 1987, and 1988. The present emphasis of this research program is on the post enumeration survey method, using blocks (or small areal units) as the unit at the last stage of sampling.

Two of the major difficulties encountered in the 1980 matching study were

- matching errors, particularly false non-matches, and
- a substantial number of cases for which there was insufficient information to determine the match status.

Much of the present research and testing program aims to eliminate, or at least greatly reduce, the size of these two difficulties. Better field work, better matching algorithms (see Section 6.1), and better statistical modeling are all ingredients of this program.

The SRD contact person regarding this research area is Howard Hogan.

3.2. *Census Adjustment*

Census adjustment involves modifying the counts obtained in the basic census enumeration of both large and small areas using measures of census undercount applicable to large geographic areas or broad categories. The initial research on adjustment involves an assessment of the level of information on the 1980 undercount currently available as well as an investigation of alternate levels of information that might be obtainable in 1990 via a post enumeration survey. The research involves modeling 1980 undercount data as well as pseudo undercount variables as a means of simulating the outcome of adjustment strategies. The strengths and weaknesses of various adjustment strategies are being explored.

Current research on adjustment is focused on total population as the variable of interest. Should adjustment be decided upon, associated variables such as housing stock and other demographic related characteristics may need to be adjusted as well. Part of our modeling efforts are focused on the lowest geographic level for which net undercount estimates are available, namely the area administered by the District Office (DO). The District Office is a census administrative unit averaging half a million persons.

Modeling the undercount requires a vehicle, a post enumeration survey for example, to estimate undercount at some level of aggregation. An important issue is then whether such estimators of undercount are unbiased. If so, then regression models may be fit to the undercount data and used to predict undercount rates at the appropriate level of aggregation. The resulting smoothed estimates may be used to adjust at the block level using the synthetic method.

The SRD contact person regarding this research area is Cary Isaki.

4. Time Series Analysis

The Census Bureau has a long tradition of research on the analysis and interpretation of economic time series. This tradition continues, and time series ideas and models are now being explored in new areas, e.g., in projecting population.

4.1. *Concurrent Seasonal Adjustment*

When the current month's datum is seasonally adjusted by means of a factor which is calculated from data up through the present month, as contrasted with the traditional method which only uses data up through the most recent December, the resulting adjustment is called a concurrent adjustment. The traditional adjustment is projected factor adjustment. Theoretical work at Statistics Canada and the University of Warwick, and empirical studies at Statistics Canada, the U.K. Central Statistical Office and the Census Bureau support the use of concurrent adjustment in place of the traditional projected factor method, supporting the common-sense dictum that all available data should be used.

The empirical studies confirm the theoretical prediction that values for levels and month-to-month changes obtained from concurrent adjustments more closely resemble (on the average) the best estimates obtained years later when sufficient future data are available that X-11's symmetric seasonal filter can be applied to obtain the adjustment. The Census Bureau's study also suggests that concurrent adjustments are more robust against large deviations from the X-11 final values than are projected factor adjustments.

Based on these investigations, Statistics Canada, several U.K. statistical offices and the Census Bureau have begun to implement concurrent seasonal adjustment.

The major incompletely resolved issue at the present time is: What previously published numbers should be revised when a current-factor adjustment is calculated for a newly obtained datum? Software to help the investigation of this issue has been developed and the analysis done so far suggests that it may be adequate to revise only the same-month-a-year-ago seasonal adjustment. The quantities investigated were the estimates of three different (percentage) changes: year-to-year, current month-to-previous month and its year-ago analogue, year-ago-to-13-month-ago change. It was observed that, if no revisions are made, substantial distortions sometimes occur in the estimates of year-to-year change, whereas if the year-ago value is revised, good improvements are seen, on the average, both in the estimates of year-to-year change and in the estimates of year-ago-to-13-month-ago change. Additional, but smaller, improvements in the latter estimates can be obtained by revising the 13-month-ago number.

The SRD contact person regarding this research area is David Findley.

4.2. *Population Projections*

For many years, demographers at the Bureau of the Census have made projections of demographic data using well-established and sound demographic techniques. The demographic data are, however, time series data. With the formulation of a cohesive time series modeling strategy by Box and Jenkins in the 1970's and the availability of state-of-the-art time series modeling software in the 1970's, the opportunity is available for using time series techniques to augment the traditional analysis done by demographers.

There are several important areas of research that we are pursuing at the Census Bureau. First, for long-term national projections, a methodology to project cohort fertility curves and generational life tables would be valuable. Second, tracking procedures are needed to translate monthly vital statistics data (such as fertility rates) to yearly values, for comparison with projections. Third, improved methods for projecting short time series, with associated ranges of uncertainty, are needed for head-of-household rates. More long-term research areas for incorporating time series methodology include convergence functions for fertility and mortality, by race, toward long-term assumptions; better techniques for subnational populations, to analyze the degree of convergence of state vital statistics to the national average; and improved techniques for obtaining high and low estimates around a middle projection, including whether "confidence" intervals are obtainable.

To date, progress has been made toward each of the first three goals mentioned. We have developed a methodology to project age specific fertility rates, which captures period and cohort effects. A gamma curve is used to parameterize the fertility curve, and time series techniques are used to project the parameters of the curve. Age-specific fertility rates are then read from these curves, allowing the estimated completion of birth cohorts whose future outcome (such as attained fertility of those women born in 1967) is as yet unknown.

Time series characterization of final monthly data is the first step in linking actual monthly values to yearly projections. We have provided time series models for the final data for U.S. general fertility rates, including a first-time statistical treatment of the effects of the calendar composition of the months in demographic data. Regarding projections for head-of-household rates, we have work in progress to evaluate alternative time series models and

their forecasts for a sample of 21 (out of 120) head-of-household rates. Also, we have shown that confidence intervals for forecasts from such short time series using current technology are of dubious reliability.

The SRD contact person regarding this research area is David Findley.

5. Analysis of Disclosure Risk

Research in statistical methods for assessing and limiting the disclosure risk for public use microdata is currently under way. In comparison to research in statistical methods for enhancing confidentiality for macrodata (e.g., tabulated data) research in confidentiality for microdata is relatively undeveloped. Although there has been much discussion about confidentiality for microdata, very few statistical methods have been developed for limiting the risk of disclosure. Preserving respondent confidentiality is a keystone in maintaining the Census Bureau's reputation as a collector of high quality data and as a source of useful quantitative information of the highest integrity.

No comprehensive solutions to the problem of limiting the risk of disclosure for microdata have been developed. Some unanswered questions are: How is the disclosure risk to be limited if we must publish microdata that benefits the research community at large which has very broad yet specific research interests and special data needs? How can we furnish the detailed data required by the research community and yet be reasonably assured that the identity of survey respondents will not be revealed?

Among the methodologies that have been used at the Census Bureau to limit the disclosure risk is limiting or suppressing data that may be crucial for the identification of respondents. For example, the amount of geographic specification on each respondent's record is usually limited to describe areas which have a population that exceeds 100 000. In the case of the new Survey of Income Program Participation (SIPP), the details specifying geographic areas have been limited to even larger locations. Because SIPP public use microdata tapes have an extraordinary amount of sensitive data listed for each respondent, no respondent can be identified as belonging to a geographic area having a population smaller than 250 000. This strategy of limiting geographical detail clearly limits the disclosure risk since in large highly populated areas it is likely that there are people who will share identical or indistinguishably similar characteristics with most, if not all, of our respondents. Thus, the disclosure risk is limited by ambiguity.

A second method currently employed to limit the disclosure risk is "top coding." For example, some of our respondents have salient attributes which could reveal their identity. For example, suppose a "special sample individual" resides in a geographic area, the particulars of which are reported in our file. In addition, suppose almost everyone in this identified area has a low income. If we were reporting total incomes for our respondents, and our "special sample individual" has an income which is extraordinarily high, then releasing his income could jeopardize the confidentiality of his data. To avoid this situation the Census Bureau typically "top codes" the income value. That is, the actual reported value is suppressed and a less extreme "top code" value is reported instead. The likelihood of the disclosure of our "special sample individual" is thereby reduced.

These and other methods have been used at the Census Bureau. However, there are no objective methods of determining what "top code" values should be. Nor do we know what the population of the smallest identified geographical areas should be to reduce the disclosure risk to an acceptable level.

Current research on these confidentiality problems is concerned with statistical modeling so that the risk of disclosure may be measured in terms of probability. By using these models, the effects of identifying geographic areas of specified population size may be assessed in terms of its impact on the probability of disclosure.

The SRD contact person regarding this research area is Nash Monsour.

6. Mathematics, Automation, and Artificial Intelligence

Throughout the past four decades, the Statistical Research Division (SRD) has made a number of significant contributions to statistical theory and methods. Indeed, through the mid 1970s almost all of the division's work was concerned with the application of statistical theory and reasoning to the solution of problems relevant to the Bureau's data programs and products. In the past 10 years, however, the scope of the division's work has expanded greatly. Now, several of the areas undergoing most vigorous development are closely allied with the disciplines of mathematics, computer science, and artificial intelligence. Of course, there is no clear distinction between these various scientific fields, and it is fair to say that the division's work is at the interface of statistics and these other related disciplines.

In this final section, I discuss four ongoing projects in these interface areas. Each of these projects is motivated by the basic desire to achieve more accurate and timely data in the U.S. population and economic censuses.

6.1. Record Linkage Research

Record linkage is the process of examining two independent computer files with the object of locating pairs of records that agree on various identifiers. For the Census Bureau, this process can be executed on two files containing individual names, addresses and demographic characteristics. Record linkage is important for census undercount determination, address list compilation and general census evaluation.

Record linkage research is focused on the development of a matching algorithm that will accomplish the above goals in a statistically justifiable manner. To this end the following major activities must be initiated:

1. Constructing a data base that can be used for calibration, validation and testing of the matching process.
2. Development of iterative methods to obtain information on the discriminating power of the various identifiers and their associative error rates.
3. Design and implementation of computer algorithms to perform the actual matching.
4. Development of a statistical foundation for the record linkage process.

The results of this research will be:

1. More accurate undercount determination and coverage analysis.
2. Replacement of costly clerical procedures with automated methods.
3. A statistically valid process, replacing previous ad hoc techniques, which can be used to make inferences.

A test of the 1990 decennial census was conducted in April 1985 in Tampa, Florida. A post-enumeration survey occurred in July 1985. These two files will be matched using the newly developed algorithms. An exhaustive series of quality-control operations are being planned to verify matches and nonmatches.

The basic foundations for the record linkage process is the Fellegi-Sunter decision procedure. A number of additional theoretical issues are being investigated to provide a completely operational system. These issues include blocking (the partitioning of the files into mutually exclusive blocks to limit the number of records that must be compared); distance metrics (a model for determining the

differences between misspelled names or the closeness of numeric data); statistical dependencies (certain fields may be statistically dependent on both value states and errors, e.g., given name and sex); local variations (names may have different frequencies of occurrence in different geographic locations); assignment (the actual assignment of the match pairs); error analysis (the ability to predict the percentage of false matches and false nonmatches); and parameter estimation (computing weights for all fields and analyzing effects of estimation error on the decision outcomes).

The SRD contact person regarding this research area is Matthew Jaro.

6.2. *Automated Coding*

All data collection organizations encounter applications which require the analysis of natural language text for the purpose of selecting a numeric code from a classification system. Codes are more amenable to computer manipulation. Discrete categories are needed for tabular presentation and aggregation to higher levels. In a clerical coding operation the coder refers to a list or dictionary of concepts which defines the translation. This clerical assignment of codes is subject to human error and subjective judgments which are not consistent when tracked over time. An incomplete list imposes a burden on the coders. Recruiting, training, and managing large numbers of coders imposes a burden on management.

In the Census Bureau's surveys and censuses of individuals, there are three categories of coding: Industry and Occupation (I&O), General (GEN), and Place of Work and Migration (POW/MIG). Research has concentrated on I&O because it is the coding area in which clerical coders have the most trouble.

The 1980 Population Census spent $27.4 million on clerical coding. Over 3 000 people were employed for more than 10 months and the error rates were higher than we would like (industry = 13 percent, occupation = 18.9 percent). Current surveys spend over $1.2 million per year on these activities.

The SRD is developing computerized procedures for selecting I&O categories on the basis of natural language responses to questions in censuses or surveys. The latest evaluations using current survey responses indicate that the automated coder is performing at the same error level as the 1980 clerical coders. I believe we can significantly reduce these error rates and have a viable automated coding product in place for use in the 1990 Census.

In the 1985 test census in Jersey City, New Jersey, all I&O entries will be coded both by the computer and by experienced clerks. Differences will be evaluated and new phrases added to the knowledge base. Further testing of the automated coder will occur in test censuses in 1986, 1987, and 1988.

Similar developments are planned for GEN and POW/MIG coding, although work in this area has not progressed as far as in the area of I&O coding.

The SRD contact person regarding this research area is Martin Appel.

6.3. *Automated Editing Research*

All survey and census programs are subject to nonresponse and erroneous reporting, whereas complete and accurate data are needed for most statistical purposes. The data collection agency often has the optimal vantage point and attendant obligation to provide valid imputations for missing values and to edit spurious responses. The development of statistically precise and mathematically rigorous edit and imputation systems is essential in meeting this objective and is vital in providing users with high quality data products.

Although the implementation of an edit and imputation system is highly survey specific, coherent methodologies can be developed that integrate diverse features and needs into

a structured framework. Within such a framework, various imputation strategies, such as the hot-deck, flexible distance matching, and statistical modeling, can be incorporated. State-of-the-art edit systems draw upon mathematical optimization techniques and outlier analysis to incorporate prior knowledge and concurrent information. Development and implementation of such systems requires that mathematical and statistical researchers work jointly with subject-matter specialists familiar with the survey environment.

For the past several years, Census Bureau researchers have been developing a general-purpose edit and imputation system for continuous data under ratio edits. The upper and lower limits for these ratio edits are, of course, survey specific and are based on statistical analysis and subject-matter expertise. These specifications are entered into the system as parameters. In addition, imputation rules for each field to be imputed are entered into well-defined and structured modules by the individual users of this system. The methodology for determining fields to alter on edit failing records and the procedures to ensure that all imputed values for missing fields are consistent do not change from one user to another and are embedded within the system.

The basic methodology within the system follows along the lines developed by Fellegi and Holt. One starts with user-supplied explicit ratio edits and generates all implied edits. Using all edits (explicit and implied) the system finds a weighted minimal set of fields to alter on edit failing records. For records with missing values, the mutally consistent fields are used in conjunction with all edits, and they determine a feasible region in which each imputed value must lie. The values to be imputed are derived from the survey-specific imputation rules contained within flexible and structured imputation modules.

This system was initally designed for use on the Annual Survey of Manufactures (ASM). The specific modules were removed from the program, and with minor modifications, this system was successfully used to process the Enterprise Summary Report and the Auxiliary Establishment Report of the 1982 Economic Censuses. It was also successfully used to process most of the 1982 Economic Census of Puerto Rico.

At the Census Bureau, all large scale automated edit and imputation systems for economic data run in batch mode, and based on the actions taken by the automated system, records are selected for analyst review. An analyst then examines the performance of the automated system and further adjusts individual records as needed. Typical causes for analyst review are large edit changes or changes on records for large establishments.

We have developed an interactive version of the core edit system for on-line use by analysts during the review process. The interactive system allows an analyst to target one or more fields for revision, observe the feasible region, select amongst the system generated imputation options, delete alternative fields, and observe (while on-line) the impact of any changes. If a field was deleted because of edit failures in a record, this interactive version of the system can be used to generate alternative sets of fields to delete.

The goal has been to develop an edit and imputation system that blends statistical and subject-matter expertise, integrates edit constraints with imputation strategy, and can accommodate a variety of users. Knowledge gained working with users in the subject-matter areas, learning their needs, and understanding the facets of their expertise has been crucial.

The SRD contact person regarding this research area is Brian Greenberg.

6.4. *Mathematical Cartography*

The Census Bureau is one of the major map makers in the United States. Conducting

censuses and surveys of the nation's population and industry requires accurate locational information that may be readily understood and utilized by enumerators. Maps with sufficient detail to find housing units or business establishments are indispensable.

To meet its mapping needs, the Census Bureau has pioneered in the field of automated cartography. The SRD has derived the mathematical theories for the automated routines for map-making, and currently continues to develop new tools and techniques.

SRD researchers used the principles of topology, geometry, and set theory to construct a complete mathematical model of maps. We built prototype geographic information systems based on that model, and are now working with the Census Bureau's Geography Division on implementing a full-scale system called the TIGER system. TIGER stands for "Topologically Integrated Geographic Encoding and Referencing" system.

Current research centers on the development of an automated map compilation or conflation system which permits the matching and merging of two separate computer map files through interactive graphics. The system will be used to conflate the United States Geological Survey's (USGS) digital data base of the entire United States with the map files produced by the Census Bureau for conducting the 1990 Census. Special geometric transformations are applied to census maps to bring them into exact alignment with USGS maps so that data on each file may be shared and differences may be highlighted for field verification. The conflation system will computerize the very difficult task of map comparison and information transfer for 5 500 maps which cover over 60 percent of the U.S. population. The automated system will not only speed up the map merging operation, but will also eliminate errors of omission and duplication inherent in a manual operation and provide uniform approaches to resolution of problem areas.

Other areas of applied mathematics supporting cartographic development at the Census Bureau are: lattice theory, which provides an algebraic tool for manipulating geographic hierarchies; curve representation, which is studied to permit quality drawing of maps; and digital data base strategies, which are needed for efficient handling of the large quantitites of data found in maps.

The SRD contact person regarding this research area is Alan Saalfeld.

7. Summary

There are a great number of research issues and problems at the Census Bureau, both in statistics and in various interface areas. In this paper, I have tried to describe briefly 15 research areas undergoing active development. Of course, there are many other areas that I did not describe where active developments are occurring, and many more areas that I would like to develop in the future.

Received June 1985
Revised November 1985