Potential Research Data Center Methodological Topics

Last Revised: **September 6, 2006.**

The Census Bureau has a long history of collaboration with external researchers. These efforts have focused on many areas of the Census Bureau's work. For example, these collaborations have led to improvements in industry and occupation coding, improvements in the measurement of total factor productivity, construction and analysis of job creation and destruction statistics, and creation of demographically detailed establishment-based statistics. These collaborations were effective because of a community of interest between the researchers and their counterparts at the Census Bureau, such as improving the basic data products of the Census Bureau as a result of the research. Using the Census Bureau's extensive microdata inventory in creative ways does lead to direct, demonstrable, improvements in its public-use economic and demographic data products.

It is our hope to foster more such collaborations between external and Census Bureau researchers through the Research Data Center network. In order to do so, we prepared an illustrative, though not exhaustive, list of research areas that the Census Bureau considers high priority. Since every external research proposal to the RDC network must show how access to confidential Census Bureau microdata has the potential to benefit the Census Bureau, we invite researchers to treat this list as a request for proposals to address these important problems. The methodology topics cited below influence the quality and validity of virtually all analyses that are performed with Census Bureau public data. We hope researchers can use these topics to help identify ways that their proposed research can be directly beneficial to those research programs inside the Census Bureau that are concerned with a particular topic.

Click on each topic below for specific examples and links to relevant Census Bureau web pages. For more information, contact:

- The RDC Administrator of the site where the proposed research would take place http://webserver01.ces.census.gov/index.php/ces/1.00/researchlocations;
- Census Bureau experts http://www.census.gov/contacts/www/contacts.html.

This list is updated as we receive new topics from the Census Bureau.

1. Unit and Item Response and Nonresponse

Survey nonresponse rates have been increasing, leading to concerns about the accuracy of survey estimates. For example, from 1990 to 2004 initial contact nonresponse rates approximately doubled for selected household surveys, including the Quarterly Consumer Expenditure Survey (from 12.0% to 23.3%), Current Population Survey (from 5.7% to 10.1%), and Survey of Income and Program Participation (from 7.3% to 14.9%). Response rates also are concerns for economic data. For example, response rates for the Economic Census declined from 86% in 1997 to 84% in 2002. In the Medical Expenditure Panel Survey – Insurance Component, about 7 percent of establishments do not report type of ownership in 2003, 8 percent do not report age of the firm, and 15 percent do not report the proportion of low-wage employees.

Errors introduced by unit nonresponse may bias survey estimates when nonresponse is high if those who participate in surveys differ from those who do not. Standard nonresponse adjustment procedures typically assume that nonrespondents are similar to respondents, but the literature does not always support this assumption. The Census Bureau is interested in both unit and item response rates in its surveys and censuses, including ways to increase response rates by improving data collection procedures. Microdata available to RDC researchers include indicators of response and of imputations for nonresponse. See also *Editing, imputation, and weighting*.

For more information on nonresponse issues in household surveys, see the Interagency Household Survey Nonresponse Group, http://www.fcsm.gov/committees/ihsng/ihsng.htm

2. Editing, Imputation, and Weighting

Studies that provide insights on the best ways to impute missing items for surveys and censuses would help the Census Bureau improve its products. The Census Bureau often relies on historical relationships among variables as the basis of its edit checks and imputation, yet some of these relationships appear to change over time, some more quickly than others. We are interested in comparisons of imputed and reported data, and assessments of alternative imputation methods. See also *Unit and item response and nonresponse*. Studies could:

- Improve our understanding of the nature of nonresponse and its effects on data quality;
- Identify "best practice" editing and imputation techniques among survey professionals and show how they can be applied to Census Bureau data;
- Provide information to develop new or improved practices;
- Assess how good edit checking and imputation can reasonably be expected to be.

For business data, our editing and imputation processes would benefit from studies that:

- Provide objective, observation-based knowledge about relationships among variables (e.g., steel mills that have positive values of shipments must have employees) and about how historical relationships change (e.g., at business cycle "turning points");
- Reveal the relative degree of homogeneity among plants in specific industries or industry-groups, especially if there is an alternative way to identify groups of homogenous plants. For example, such studies could describe the manufacturing sector, illustrating both similarities and differences between various sub-sector groupings of plants, or describe specific sub-sector areas, ranging from a 3-digit North American Industry Classification System (NAICS) subsector to a single 6-digit industry to other possible groupings of plants.
- Measure and evaluate the impact of procedures that down-weight or modify the values of influential observations, as described in http://www.bls.gov/bls/fesacp3060906.pdf.

For household surveys, we want to know how well our data editing processes work.

We also want to assess how effective weighting and imputation methods are at reducing nonresponse bias. Studies could:

- Compare imputed and reported data;
- Assess alternative ways to impute for missing data.
  - Compare model-based and hot-deck methods when the frame or administrative records provide many correlated variables. Hot deck imputation uses classification and sorting to select a nearest neighbor within the same class and as close as possible in the sort to be a "donor" to the sample subject with missing data. Using this approach, at best four or five correlated variables will affect the choice of the donor. In surveys like the National Survey of College Graduates, where many variables are available from the frame, a missing variable like income can have ten highly correlated covariates with limited collinearity. In that case, a linear model would seem to be a good estimator for the missing value. Is a linear or non-linear model or regression a better estimator than the hot-deck?
  - Assess multiple imputation. See http://www.federalreserve.gov/pubs/oss/oss2/scfindex.html for an example.
- Assess alternative ways to weight survey data.

3. Error Profiles

Describing and cataloging survey errors and comparing estimates to other known quantities, such as administrative records, would provide the Census Bureau with important information about the quality of its data. The profiles could:

- Describe the kinds of errors associated with any specific survey; an example for the Current Industrial Reports (CIR) data would be "Sampling vs. Reporting vs. Processing Errors in the XXX CIR: Which Should We Worry About First?"
- Describe how a specific type of error varies across surveys, such as "Reporting Errors in CIRs: Where are They Large and Why?", and "The Cost of Errors: Which Errors Are Affecting the Annual Survey of Manufactures (ASM) Estimates More?".

Some existing error profiles are available on-line: American Community Survey:

http://www.census.gov/acs/www/UseData/index.htm.
American Housing Survey: http://www.census.gov/hhes/www/housing/ahs/dataquality.html.
Current Population Survey: http://www.bls.census.gov/cps/basic/perfmeas/folder.htm.
Survey of Income and Program Participation: http://www.sipp.census.gov/sipp/source.html.
Census 2000: http://www.census.gov/pred/www/.

4. General Methodological Issues

Evaluate and suggest how the Census Bureau can improve:

- Substantive estimates – how processing steps or aspects of the data collection process, such as attrition, affect the estimates;
- Routine data or information products, such as reports and tabulations, or specific data items within them.
- Survey estimation techniques, including evaluation of alternative estimation strategies, such as small-domain estimators. The Census Bureau's design-based estimation paradigm was developed and works well for estimation in certain situations with large samples (e.g., for many national level estimates), but was not developed to do small area estimation, or to deal with large amounts of missing data, outliers, etc. In such settings, other approaches (e.g., model-assisted or model-based estimation) may offer opportunities for improvement. Examples include:
  - Study the optimum use of population and housing unit controls for the American Community Survey (ACS) and other surveys.
  - Investigate bias and uncertainty in population controls to help develop error estimates for population estimates; also, study the impact of these errors on survey estimates.
  - Research on estimators that incorporate administrative data to improve ACS estimates for very small areas.
  - Research on improving small area estimation for Census Bureau survey applications such as the Small Area Income and Poverty Estimates (SAIPE), the Small Area Health Insurance Estimates (SAHIE), etc.
  - Investigate the feasibility of using model-based or model-assisted estimation techniques in the monthly residential construction program (to use additional information from the large sample of building permits to improve estimation of housing starts, completions, and sales).

- Understand measurement errors. Recent Census Bureau experience points to measurement errors (that is, errors of observation arising from the interviewer, the respondent, the questionnaire, or the mode of data collection) as major sources of inaccurate and inconsistent data. Basic research is needed to better understand sources of measurement errors. Research on sources and magnitude of measurement errors can include:
  - Evaluate effects of mode of data collection on quality and comparability of survey data, to support guidelines for standardizing survey instruments across modes.
  - Conduct research on fundamental sources of survey measurement problems (e.g. recall error), drawing on theory and methods in relevant scientific disciplines, such as psychology and linguistics.
- Improve population estimates:
  - Integrate new data sources and statistical modeling to model migration. This includes measuring the annual inflow of migrants to the United States, estimating the annual outflow of migrants from the United States, and internal migration.
  - Measure population on a current residence basis or develop models to reconcile current residence population with usual residence population.

5. Dataset-Specific Data Quality Issues

The Census Bureau is interested in assessing data quality issues such as potential biases in its surveys. Examples include:

- Decennial Census:
  - Improve mode consistency – that is, how can multi-mode surveys and census (surveys conducted by mail, telephone, in person, and Internet) be designed to get the best data possible in all modes *and* get the same answer from the same respondent regardless of mode (see *Understand measurement errors*, above);
  - Improve survey response from linguistically isolated populations in all survey modes;
  - Improve coverage: The decennial census suffers from errors due to omission of persons who should be counted and to

erroneous enumerations of persons, the latter including enumeration of persons who should not be counted at all, enumeration of persons in the wrong place, and enumeration of persons multiple times (duplication). Research on all aspects of coverage are needed:
  - Prevent and correct for duplication at all stages of the census and coverage measurement process, from address list development to final coverage estimation.
  - Improve determination of Census Day residence—more basic research on errors in, e.g., recall and reporting of moves and other problematic residence situations.
  - Develop coverage measurement methods for group quarters.
  - Statistical research on gross census coverage errors, i.e., separate estimation of census omissions and erroneous enumerations.

- American Community Survey (ACS):
  - The effect of seasonality on ACS estimates;
  - Alternate methods for adjusting financial variables for inflation;
  - Alternate methods for accumulating multiple years of information for small areas, or to construct estimates such as poverty;
  - Use of models and external data along with ACS data to develop better ACS estimates.

- Current Population Survey (CPS) Rotation Group Bias. The CPS is a panel survey that interviews each housing unit eight times: once each month for four months, eight months of no interviews, and once each month again for four months. It is known that unemployment data differ, depending upon which rotation month, of the eight, the respondents are in. Why does the difference exist and which month produces the most accurate unemployment data? http://www.bls.census.gov/cps/basic/perfmeas/nonresp.htm.

- National Crime Victimization Survey (NCVS) Time-in-Sample and Mode Bias. There has been some indication that crime reporting in the NCVS goes down the longer a respondent is in the survey. This phenomenon could also be related to mode of contact, as it seems to increase with Computer Assisted Telephone Interviewing. What is the extent of time-in-sample bias and how is it related to mode?

- Survey of Income and Program Participation (SIPP) Attrition Bias and Seam Bias. SIPP is a longitudinal survey that interviews each respondent every four months for three to five years.
  - Attrition Bias. During the course of the survey, some sample persons "drop out" for a variety of reasons, introducing nonresponse bias due to attrition, especially to longitudinal analyses. How much nonresponse bias is there for individual waves and for longitudinal analysis, particularly on income and program participation estimates? Can models be developed to correct certain estimates for attrition bias?
  - Seam Bias. Each interview asks about the previous four months. Transitions between life circumstances such as employment and unemployment or marriage and divorce are most likely to be reported as occurring at the "seams" between 4-month waves. How can models best account for this misreporting?

- Medical Expenditures Panel Survey – Insurance Component (MEPS-IC).
  - Evaluate possible methods to impute MEPS-IC variables that are not currently included in imputation procedures. Developing imputation procedures requires resources, so imputation procedures for missing or invalid values exist only for variables included in the MEPS-IC published estimates.
    - Can reliable imputations be developed for variables (such as the fraction of workers over age 50) that are collected in the MEPS-IC but not currently included in published estimates?
    - If so, it would be vauable to evaluate whether the newly imputed variables would produce useful new estimates. For example, whether there would be interesting differences in whether or not insurance is offered, premiums charged, types of insurance plans typically offered to workers in that age group compared to those offered to younger workers.
  - Use MEPS-IC microdata to identify likely sources of differences (e.g. coverage of sample, definitions, timing) between MEPS-IC estimates and estimates from other sources collected by the Census Bureau (e.g. employer

health benefits estimates from the Current Population Survey) or external sources (e.g., Kaiser/Health Research and Educational Trust Employee Health Benefits Survey).

- Improve methods for estimating retiree health insurance coverage, all health expenditures, other variables.
- How should MEPS-IC handle workers acquired through Professional Employment Organizations and temporary employment?
- Assess whether MEPS-IC overestimates the percentage of workers with insurance because coverage data are collected at the company rather than the establishment or work unit level.
- Reclassify historical MEPS-IC records from SIC to NAICS to allow more consistent analysis of time trends.

- Industrial Research and Development Survey (R&D)
  - Frame Enhancement. A major concern about the R&D survey is whether the frame includes all private entities that perform research and development. The Census Bureau's Business Register lacks an establishment or firm level variable that accurately predicts R&D activity. This leads to inefficient sampling since many surveyed firms have no R&D activity. It also leads to concerns that the frame is incomplete as the survey may miss many firms that do have R&D activity. The Census Bureau is interested in the use of other data sources that may help in constructing variables that accurately predict whether a business conducts R&D. Other potential sources of data include: Compustat, the Bureau of Economic Analysis's Foreign Direct Investment data, Federal Procurement Data, and patent data.
  - Increased Accuracy in Industry Coding. As a firm level survey, the industry codes that are applied to the R&D activity are the industry codes for the firm. To the extent that the R&D activity occurs at an establishment in an industry different from that of the firm as a whole, industry R&D statistics may be misleading. The Census Bureau is interested in projects that can help verify and improve the firm industry codes in the R&D survey.
  - Quality Controls of Data. The R&D survey microdata are available back to the 1970s. The Census Bureau is interested in consistency checks within a survey year (across data items) and across survey years (time series consistency). The Census Bureau is particularly interested in accurately measuring the R&D activities of large, complex multinational companies.

6. Household Data Quality Issues

- The Hispanic Social Security Number Effect. Analysis of mortality throughout the United States including the National Longitudinal Mortality Study has shown that Hispanics tend to have higher survival rates than other ethnicities, despite higher correlations with other variables know to correlate with lower survival rates, such as lower income and incidence of smoking. Some evidence has emerged that it may be an artificial effect caused by Social Security Number errors. Is this an artificial effect or a real effect?
- Under-reporting issues—Income and health insurance coverage (particularly Medicaid), are under-reported in SIPP and CPS, compared to estimates from other sources. How serious is this under-reporting, by characteristics such as components of income and sources of health insurance coverage, as well as characteristics of individuals who do, and do not, report them? Can models be developed to correct the survey data for underreporting?
- Defining economic well-being—The Census Bureau is interested in research into expanding the definition of economic well-being beyond money income—both exploring new methods and evaluating current ones. http://www.census.gov/hhes/www/income/income.html.

7. Business Data Quality Issues

- Using the Supply Chain Data to Answer Other Questions. The Census Bureau collected data on the 2002 Economic Census about supply chains. The Census identified key supply chain activities (e.g., bundling or kitting, pick and pack, warehousing, local and long-distance delivery, and processing returned merchandise), and asked respondents to indicate whether and by whom these activities had been performed—i.e. whether they had been performed by the responding establishment, by an affiliated or an unaffiliated establishment, or not at all. The survey also included the four questions on inventory ownership and management.

  The supply chain data are a rich and largely untapped resource for statisticians, economists, and others interested in the

structure and operations of U.S. industries and businesses. The data could be used to assess the implications of supply chain innovations for national statistical programs. The data may also contain answers to other interesting questions—e.g., effects of functional differentiation and integration on the performance of manufacturing establishments, the forms and incidence of business-to-consumer electronic commerce, the changing role of particular supply chain industries and functions (e.g., logistics consulting services; product design and engineering services). The Census Bureau encourages researchers to use the data to explore these and other supply chain questions.

- Economic measurement. These topics include appropriate *methods for measuring economic variables* such as outputs, inputs and productivity using the Census Bureau's data and *recommendations for better data collection methods*. These topics can be applied to all sectors of the economy. In fact, the concepts and measurement of outputs, inputs, and productivity in the business and service sectors are particularly important.

  For example, theoretically, gross output is defined as the total value of shipments (TVS) plus changes in inventories; however, most empirical studies using Census Bureau plant-level data have used TVS as a proxy for gross output without adjustment for inventory changes. This is because inventories collected by the Census Bureau can be based on different accounting methods chosen by the firm (e.g., last-in-first-out and first-in-first-out. Using these data (without appropriate adjustment) would introduce errors in the measurement of output. Thus, research that involves methods for converting these inventories data into "consistent" data based on a single accounting method would provide significant benefits to the Census Bureau. This would help measure correctly output and inventories.

  Similarly, exploring the adequacy of current measures of capital stocks, and recommending improvements, would provide benefits to the Census Bureau. For example, data on book values of capital, and most kinds of investment, are no longer collected in non-census years.

- Structures of establishments and firms. These topics include studies that involve *methods for identifying* accurately births, deaths and ownership changes of establishments and firms and *recommendations* for ways to collect accurate data. For example, studies that extend and update the existing "ownership change database" for the manufacturing sector to the services sector would provide direct benefits to the Census Bureau.

- Assess data collected in new or pilot surveys. These studies would evaluate the new data, inform the Census Bureau about the quality of the data, and make recommendations to the Census Bureau's data collection programs. For example:
  - The 1999 Annual Survey of Manufactures (ASM) Computer Network Use Supplement collected for the first time detailed data on the use of information technology (IT) and electronic commerce;
  - The Annual Capital Expenditure Survey collected for the first time in 2003 data on detailed forms of high-tech capital;
  - A pilot survey is being planned for the Pollution Abatement Costs and Expenditures Survey. Studies to develop editing and imputation algorithms and sampling specifications, and other survey development work, would supplement the expertise of Census Bureau staff.
  - Reporting Units. A problem for economic surveys is the definition of statistical (reporting) units and their potential mismatch with the structural units of a company. The organization of company records may make it more difficult or impossible for the respondent to provide data according to the Census Bureau's desired statistical units. This could lead to poor quality estimates. For example, companies in some services industries cannot report data by geographic area for products or services distributed via a network. The Census Bureau is interested in understanding how businesses keep their records, including which units keep which kinds of information. In other words, when is it optimal to collect data at the establishment level and when is it optimal to collect data at the enterprise (firm) level.
  - Compare to external data. Other sources of business microdata may include information that is related to that collected by the Census Bureau. Linking such data to Census Bureau internal microdata can

provide important information about Census Bureau data. For example, linkage studies could evaluate:

- How differences in definition, timing, aggregation, or sources of errors contribute to differences between Census Bureau and external microdata;
- How useful public microdata are for imputation or non-response adjustment in Census Bureau surveys;
- The viability of using information from public filings to reduce respondent burden;
- How to interpret differences in estimates across data sources.

8. Additional Resources

U.S. Census Bureau Technical Documentation and Technical Working Papers
http://www.census.gov/prod/www/workpaps.html

American Statistical Association
http://www.amstat.org/index.cfm?fuseaction=main
Proceedings of the Survey Research Methods Section
http://www.amstat.org/sections/SRMS/Proceedings/

Federal Committee on Statistical Methodology – Methodology Reports
http://www.fcsm.gov/reports/

Guidance on Agency Survey and Statistical Information Collections
http://www.whitehouse.gov/omb/inforeg/statpolicy.html under "Standards"

The ASA / NSF / Census Bureau Research Fellow Program
http://www.census.gov/srd/research.pdf.

Last Revised: September 6, 2006