Developing Community Statistical Systems With American Community Survey Summary Profiles and Administrative Records

Cynthia M. Taeuber
U.S. Census Bureau and Jacob France Institute, University of Baltimore, Baltimore, MD 21202-5779

KEY WORDS: **Sampling error, nonsampling error, universe, comparability, data documentation, data quality**

Introduction

In the last decade, local governments have greatly expanded their use of administrative records for management of programs as statistical files to evaluate the results of program choices, to determine priorities among needs, to challenge anecdotal evidence used to make policy, and to make strategic plans. They are developing systematic information to understand trends and interactions – that is, community statistical systems.

As the Census Bureau releases results from the Census 2000 long form and the American Community Survey, more analysts are making comparisons with administrative data.

We expect estimates from the two surveys to differ from administrative records. It isn't that the results from one data set are "right" and the results from the other data set are "wrong." Both have weaknesses and strengths, and the data are collected in different ways, for different purposes, and have different types of errors. Administrative records have information about a subset of the total population, such as the people enrolled in a particular program. Statisticians design federal surveys to respond to policy questions; state and local governments and businesses collect administrative records primarily to manage programs, not to answer policy questions

The paper examines reasons for differences, including data collection methods, sources of error, confidentiality, and differences in universes, coverage, time periods, and questions. Even when concepts seem that they should be similar, such as the number of poor children and the number of children receiving public assistance, it is comparing the proverbial apples and oranges and ending up with kumquats. This paper then considers methodological research needed to develop community statistical systems with a comparable core set of statistics and to understand when and how it is possible to use slightly dissimilar data bases.

Some jurisdictions have developed community statistical systems to track population, health, housing, crime, business, and environmental trends, and to establish interaction effects. The statistics are geographically-based summaries from decennial census data, small-area population estimates, and administrative records, infrastructure, and physical attributes of the areas. Once annually updated statistics of population and housing characteristics become available from the American Community Survey, data users can incorporate the profiles into the community statistical systems to produce a picture of the direction and level of trends. Sometimes the information is for "internal use only," but often, the public can access the summarized statistics and maps.

An idealized concept of an enhanced system of community data sets is a core set of comparable variables from surveys and administrative records to use with automated analytical and display software and one that maintains the confidentiality of individual information. Analysts can use a set of comparable statistics in dynamic models of change to inform policy decisions and help determine strategies by providing improved estimates and projections and better understanding of interaction effects. The models could be econometric or needs assessment models as well as mapped interaction models. We don't have such a system of comparable statistics now and analysts will have to refine the methodology for such models from what has been done thus far.

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

A system of community statistics would track the direction of population and housing along with other characteristics of an area, and would be able to compare situations among areas across the nation. It would be able to "generate a profile of short- and long-term outcomes" of programs, produce statistics about population subgroups at risk of requiring assistance, the duration of episodes of need, and improve our understanding of how, for example, the economic environment affects the success of some programs.[1]

The systems communities have developed thus far are specific to a city and are not comparable across areas. Efforts are underway now to develop the next generation of community statistical systems, a network with a core data set (beyond what is available from federal sources now) that is comparable across areas.

The current systems have the beginnings of a comparable core population and housing data set from the decennial census long form, small-area population estimates, and eventually, the American Community Survey. The sample surveys produce estimates, that is, generalizations, or inferences about the total population that are key in any discussion of comparable community statistical systems. They also use the registry system of the U.S. vital statistics system and the few nationally comparable administrative record sets, such as the free/reduced-price School Lunch Program. The next step is to develop comparable, or essentially similar, statistical files from administrative records.

The difficulty is how to create comparable statistical files from dissimilarities such as definitions, coverage, reference periods, and so on, or at least how to create statistical files that are similar enough to use for comparisons of key trends (such as employment and wages). We expect estimates of population and housing characteristics from the decennial census and the American Community Survey to differ from the results of administrative records compiled for the management of programs. The data are collected in different ways and for different purposes and have different types of errors. A critical next step is to determine what the differences are among data sets and find ways to improve comparability where it is possible.

Factors that affect comparisons include data collection methods, sources of error, avoidance of the disclosure of personal information, and differences in universes, time periods, and questions. Examples of administrative records that one might compare with summarized profiles from the American Community Survey and the decennial census, especially the long form sample, include those related to public assistance, employment and unemployment, school enrollment, income, use of services for the homeless, prison rolls, public transportation ridership, births, information from licenses for occupations from medical professions to cosmetologists, deeds and local property tax records indicate house values and the year a structure was built, the number of owners and renters, vacant housing units, and the housing costs of mortgages, rents, and utilities.

The appropriate statistics to use depends on the questions you are trying to answer. Conclusions need to account for differences among data sets. Data users need to understand from where the data come, how they are produced, what they measure, and their relative advantages and disadvantages for different purposes.

The discussion below is of general factors that cause differences in the results between administrative records and estimates from the decennial census or the American Community Survey.[2] Why there are differences vary among

[1] Martin H. David, "Monitoring Income for Social and Economic Development," in Burt S. Barnow, Thomas A. Kaplan, and Robert A. Moffitt (eds.), **Evaluating Comprehensive State Welfare Reforms: The Wisconsin Works Program**, Albany, NY: Rockefeller Institute Press. Culhane, Dennis P. and Stephen Metraux. 1997. Where to from Here? A Policy Research Agenda Based on the Analysis of Administrative Data. In *Understanding Homelessness: New Policy and Research Perspectives*, ed. Dennis P. Culhane and Steven P. Hornburg, 345 – 346.

[2] Documentation of concepts, methods of data collection and processing, and the accuracy of the data are available for the data set on the Census Bureau's web site at www.census.gov. Because administrative records have not been treated as statistical files generally, statistical documentation for administrative records can be very difficult to obtain.

administrative record data sets. We can't completely disentangle the exact contribution of every factor to the differences, but we can measure part of the differences.[3]

Sources of Error in Data Sets

Every data set has errors that affect the accuracy of the statistics an agency publishes. There are two major categories of errors that affect the accuracy of a sample survey such as the American Community Survey and the decennial census long form: sampling error and nonsampling errors. Administrative records have nonsampling errors. The question for each statistic is: how accurate, how close are the results to the true value?

Sampling error: American Community Survey data products show the confidence interval next to the survey estimate. This makes it easy for data users to determine whether apparent differences between the survey estimate and the administrative records are actually explained when sampling error is considered.

Example: According to Maryland's welfare payments records, over calendar year 1989, an average of 1,824 children in Charles County, MD received welfare payments. The 1990 census long-form estimate of poor children for calendar year 1989 was lower, with only 1,664 poor children. At first it seems there is a mistake because we expect more poor children than welfare recipients because not all poor people are eligible or apply for public assistance. The long-form sample estimate is not an exact count – it is an estimate based on a sample of households. When the margin of error due to sampling in the census is computed, the results are as expected. The 90-percent confidence interval was 1,471 to 1,857 poor children in calendar year 1989. The 1,824 children who received welfare fell within that range as we expected.

Nonsampling errors are a major source of difference between survey results and administrative records. Nonsampling errors may be introduced during any of the complex operations used to collect, process, and publish statistics and are often not well measured.

Nonsampling errors are of four types: (1) measurement errors; (2) coverage; (3) nonresponse errors; and (4) processing errors. They include, for example, missing some people and double counting others, respondents giving incorrect answers or not answering some questions, imprecise questions, interviewers leading the respondent's answer or giving incorrect information, interviewing the wrong unit, and not capturing or coding the responses correctly.

State agencies check administrative records that generate cash or noncash benefits for program participants for fraud, clerical errors, and management errors, one of the few measurements of error for administrative records. They have increased electronic checking of information in recent years and that has reduced inconsistencies among many types of administrative records. By contrast, surveys suffer from "recall" errors (e.g., income responses may be less accurate than tax records).

Agencies tend not to provide little or no documentation of data collection and processing methods for most administrative records. Information for administrative records may come from a variety of sources (a caseworker, the client, or events). Forms, rules, and concepts change often, but it is unusual for an agency to provide formal documentation and it is difficult for data users to obtain. State documentation systems are often in the heads and desk drawers of state employees and critical information often departs with the employee, making historical analyses very difficult.

Data collection cycles are generally different. The American Community Survey contacts a portion of the sample throughout an entire year and asks questions that may refer to the day, the week, or 12 months before the respondent fills out the form. For example, the American Community Survey asks about total earnings from the 12 months before the form is filled.

[3] For example, sampling error, undercount, and differences in the definition of income between the 1990 census and Maryland's welfare records (AFDC) contributed to differences in the number of poor children and the number receiving AFDC benefits. See: Cynthia Taeuber, Jane Staveley, and Richard Larson, "Issues in Comparisons of Decennial Census Poverty Estimates With Public Assistance Caseloads in Maryland," prepared for the National Association for Welfare Research and Statistics conference in Baltimore, MD, August 2001.

Unemployment insurance (UI) records reflect individual quarterly earnings. While differences in the collection cycles means the distributions from the two data sets are not strictly comparable, one can still study their relationships.[4]

Geographic disparities in the assignment of residence between surveys and administrative records are a significant barrier in comparisons between data sets. While the American Community Survey is based on a person's place of residence, some administrative data sets are collected from establishments. Stuart Sweeney has shown a potential bias in administrative data sets such as ES-202 records (employment and wages) because states vary substantially in the integrity of their address records, a critical factor in achieving comparability of data sets.[5]

Rokicki notes that the Unemployment Insurance database captures the number of jobs, whether full- or part-time. A person with two jobs would be counted twice in the ES-202 database. The ACS shows the number of people with jobs regardless of how many and keeps track of them by place or residence.[6]

Coverage problems may bias the results and occur in administrative records and surveys. For example, when performance measures are involved, there may be incentives for administrative actions that de facto include or exclude potential clients from the final administrative records. . Administrative files of the homeless population have an undercount if a service provider is not part of the data set and an overcount if people with regular housing use services intended for those without homes. These, and many other administrative records data sets include people who move in and out of programs over the course of a year.

In both the American Community Survey and the decennial census, there is field staff follow up at households that do not respond to the initial mailing of the questionnaires, although the steps differ. Census 2000 mailed the questionnaire once, compared with twice for the American Community Survey. The American Community Survey calls first by telephone, and if that fails, sends Field Representatives to make personal visits to a sample of 1 in 3 units. The number of callbacks to a nonresponse unit varies among surveys. Mail response rates to both the decennial census and the American Community Survey are high compared with private surveys, but do differ among specific population groups such as race and ethnic groups, age groups, and owners and renters. Thus far, the final overall response rate for the American Community Survey sites has been about 96 percent.

From some administrative records, we know the numbers of people receiving benefits from programs, but not the number eligible. The American Community Survey and the long form, because they collect characteristics representing the entire population, sometimes have information useful in estimating the potential number eligible for programs to compare with the number actually receiving program benefits.

In making comparisons among data sets, the universes need to be as similar as possible. For example, the American Community Survey includes undocumented immigrants. School enrollment records and unemployment statistics differ because of universe differences. Because of the lack of documentation of administrative records, and the many complicated requirements for program eligibility that differ among states, developing similar universes for analysis are a significant challenge.

The definitions of terms used in the questions and the response choices vary among data sources and results are not comparable even when the words are the same. Classification of race and ethnic groups as well as industry and occupations differ, for example. The composition of "income" differs among data sets.

Two studies[7] are compared for reported earnings in the American Community Survey profiles

[4] Phillip S. Rokicki, "A Comparison of American Community Survey Profiles and Administrative Unemployment Insurance Summaries," a report for the Census Bureau, April 2002, pp. 10-12, 17.

[5] Stuart H. Sweeney, "The Next Generation of Community Statistical Systems: Data Sources Availability and Limitations Panel Session Report," conference in Tampa, FL, 2002, pg. 3.

[6] Rokicki, p. 18.

[7] David Stevens, Jacob France Institute, University of Baltimore, summarized 1998 Unemployment Insurance records from Maryland Department of Labor, Licensing, and Regulation (report forthcoming); and Phillip S. Rokicki, "A Comparison of American Community Survey Profiles and Administrative Unemployment Insurance Summaries

with summarized special tabulations from state Unemployment Insurance records for Calvert, MD and Broward County, FL. Both studies show that the direction of the trends is similar for both counties. People were less likely to report earnings of less than $10,000 in the American Community Survey than were indicated there should be from the Unemployment Insurance records, while the American Community Survey had a somewhat higher proportion of people reporting earnings of $30,000 or more. For one possible explanation of the large differences between the two data sets at the low-end of the earnings continuum, David Stevens points to national statistics of median usual weekly earnings of temporary workers, most of whom make less than $10,000 per year.[8] The American Community Survey asks whether the respondent received earnings in "the last 12 months" before filling out the form. It seems plausible that it could be difficult to accurately report the timing and amount of earnings from temporary work.

The universe for the American Community Survey represents all classes of wage and salary workers who report their earnings, while the UI records include only those classes of workers for whom the state collects unemployment insurance taxes. The UI program does not include self-employed workers, federal government employees, unpaid family workers, railroad workers, out-of-state workers, and certain groups that work for nonprofit organizations. Thus, we expect the total number of earners in the UI records to be lower than the number of earners in the American Community Survey as the survey does ask respondents to report earnings by the classes excluded from the UI records. In Broward and Calvert counties (Table 1), if you add the UI counts to the American Community Survey estimates of self-employed workers, federal government workers, and out-of-state workers, and account for the combined sampling error, we conclude that the two data sets result in about the same number of earners in those two counties (see second and last lines of Table 1). Out-of-state earners are captured in the ACS but not the UI records. This is especially important in Calvert County, MD where many workers commute to Virginia and the District of Columbia.

Charles Alexander has noted that the income distributions at the *national* level from the American Community Survey, Census 2000, and the Current Population Survey are all similar. This suggests that the differences we see in the earnings distributions between the American Community Survey and the Unemployment Insurance records are methodological.[9]

Objectives for methodological research needed to develop community statistical systems include: (1) creating modern community statistical systems for informed strategic planning, including developing the methodology to use multiple data sets in statistical models in conjunction with the trend information the American Community Survey will provide and to develop Geographic Information Systems (GIS) software that displays the American Community Survey statistics appropriately and in spatial interaction models[10]; (2) identifying the impact and sources of differences between administrative records and the American Community Survey; and (3) addressing data quality and documenting administrative records for research purposes.

Summary

There is enormous potential for improving estimates, projections, and informing public policy through research that uses multiple data sets. This greatly multiplies the value of the updated, comparable trend information from the American Community Survey for federal and local governments. We need to understand the extent and type of errors in these data sets to succeed.

for Broward County, FL," Florida Institute for Career and Employment Training of Florida Atlantic University, report to the Census Bureau, April 2002. Both reports use the American Community Survey earnings distributions from the Census Bureau's website (e.g., see Table P136 from the 1999 American Community Survey).

[8] Bureau of Labor Statistics, "Median Usual Weekly Earnings of Full- and Part-Time Contingent Wage and Salary Workers and Those With Alternative Work Arrangements, by Sex, Race, and Hispanic Origin, Table 13, http://www.bls.gov/news.release/conemp.t13.htm.

[9] Charles H. Alexander, unpublished comments at the 2002 American Statistical Association meetings.

[10] Jon Winslow and Anthony Lea, "Customer Relationship Management: Location Maximizes Return on Investment," GeoWorld, April 2002, pp. 33-34.

Table 1. Estimates of Earners in Broward County, FL and Calvert County, MD: 1999
(The 90-percent confidence intervals for the estimates from the American Community Survey are shown in parentheses below the survey estimate.)

Earners	**Broward County, FL**	**Calvert County, MD**
Amer. Community Survey earners	**824,448**	**43,225**
Amer. Community Survey 90% confidence interval for estimated number of earners	(802,343 – 846,553)	(41,974 – 44,476)
Unemployment Insurance*	713,605	29,128
ACS self-employed workers	78,658 (74,728 – 82,588)	3,313 (2,706 – 3,920)
ACS federal government workers	11,591 (10,048 – 13,134)	5,066 4,325 – 5,807)
ACS, worked out of state	6,138 (4,901 – 7,376)	5,591 (4,764 – 6,418)
Amer. Community Survey estimate and 90% confidence interval for people who worked out of state + self-employed + federal government workers	96,387 (93,268 – 99,506)	13,970 (12,959 – 14,981)
UI + ACS self-employed +ACS federal government workers + worked out of state	**809,992**	**43,098**
Combined UI/ACS estimated interval	**806,873 – 813,111**	**42,087 – 44,109**

***NOTE:** Unemployment Insurance records do not include all classes of earners, including those shown above.
Source: U.S. Census Bureau, 1999 American Community Survey, Table P136 for earners, Table P41 for class of workers, and Table P1 to compute the confidence intervals; David Stevens, Jacob France Institute, University of Baltimore, summarized 1998 Unemployment Insurance records from Maryland Department of Labor, Licensing, and Regulation (report forthcoming); and Phillip S. Rokicki, "A Comparison of American Community Survey Profiles and Administrative Unemployment Insurance Summaries for Broward County, FL," Florida Institute for Career and Employment Training of Florida Atlantic University, report to the Census Bureau, April 2002, Table 3.

The point here is not to discourage researchers to use multiple data sets. Our research shows the American Community Survey and the census long forms are reliable and better than most sources because the Census Bureau works hard to reduce errors, to measure errors, and to give data users information about the extent of error. The challenge is to get similar information about administrative records to guide researchers.

There does come a point, however, when you should not push the statistics beyond their limits. Some data sets just can't be compared. As the song says, you've got to know when to fold.

Note of thanks: Valuable comments and information for this paper were provided by Charles H. Alexander, Jr., Sue Love, Charlene Leggieri, Marc Roemer (Census Bureau), Julia Lane (Urban Institute), and David Stevens (Jacob France Institute, University of Baltimore).