Executive Summary of paper for American Statistical Association 2002

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

Author: Cynthia M. Taeuber

- 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.

- As results from the Census 2000 long form and the American Community Survey are released, 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.

- 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.

- 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 not different because of sampling error.

- Nonsampling errors are a major source of difference between survey results and administrative records. Administrative records that generate cash or noncash benefits for program participants are checked for fraud, clerical errors, and management errors, one of the few measurements of error for administrative records. Electronic cross-checking of information has increased in recent years which had reduced inconsistencies among many types of administrative records.

- Data collection and processing methods are poorly documented 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. The difficulty is that this information is rarely documented formally and is generally very difficult to obtain. State documentation systems are most 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 from the Census Bureau's surveys, which complicates comparisons although it is sometimes possible to re-run administrative records to closely approximate the time reference of the census and the American Community

Survey. Geographic disparities in the assignment of residence between surveys and administrative records are a significant barrier in comparisons between data sets. Additionally, some administrative data sets are collected from establishments rather than households, further complicating assignment of residence.

- Coverage problems occur in administrative records and surveys. Examples from administrative records of the homeless population are included.

- From 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. 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.

- Two studies are compared for reported earnings in the American Community Survey profiles 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.

- 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; (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.

- 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.

Rev.09/05/02 [FAQadminrCensACS_082702doc]

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

Introduction

How do data users know the "right" number to use when the results from different sources of data disagree? Surveys such as the American Community Survey and the decennial census long form ask general purpose questions and the results represent the entire population. Administrative records have information about a subset of the population, such as the people enrolled in a particular program. Surveys are designed to respond to policy questions; administrative records are collected to manage programs, not to answer policy questions.

Data producers hear this question often, and often, there is no simple answer. This paper discusses major issues, difficulties, and implications in comparing the American Community Survey and decennial census long form with administrative records collected to manage programs and determine which applicants are eligible for benefits or services. 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.

A community statistical system uses geographically-based statistics for decisionmaking. Some cities have developed such 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, they can be incorporated to produce a picture of the direction 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. A set of comparable statistics could be used 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 the methodology for such models would have to be refined from what has been done thus far.

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 that have been developed thus far are specific to a city and the systems 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

---

[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.

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 might be compared 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.

It isn't that the results from one set of data are "right" and results from the other data set are "wrong." Surveys and administrative records both have strengths and weaknesses, errors, and differences in concepts and data collection methods. 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.

- Every data set has errors and uncertainty about the accuracy of the statistics. Some errors can be measured, some cannot. In sample surveys, as with the long-form of the decennial census and the American Community Survey (hereinafter referred to as "the sample surveys"), total error consists of "sampling error" plus "nonsampling errors." We can measure sampling error. Some nonsampling errors are known, some are not. Of the known errors, some can be measured, some cannot. Of the known errors, some can be measured and reduced as part of quality control procedures. Other known errors cannot be measured.

- Administrative data sets can include recipients of a program's services. These data sets have nonsampling errors. For an administrative records file, questions about the accuracy of the data set are usually related to whether people in the file should be there and whether the data about them is correct, complete, and current. For example, is there an error, or perhaps fraud, in the acceptance of a case into the program?

- There are crucial differences in concepts and data collection methods among data sets. This means there are differences in what is measured even though it seems the concepts are similar. An example of that is the number of poor children and the number of children receiving public assistance. It is comparing the proverbial apples with oranges and ending up with kumquats.

A discussion is below 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 administrative record data sets. We can't completely disentangle the exact contribution of every factor to the difference, 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 published. 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

Sampling error is one measure of a survey's accuracy. It refers to "the variability that occurs by chance because a sample rather than an entire population was surveyed."[4] That is, sampling error is a warning the "the estimates are not exactly equal to the population quantities being estimated."[5]

The standard error is a measure of precision, of how much the survey estimate varies from the true population because of sampling. From the standard error, we can compute the confidence interval, the range of values that describe the uncertainty because of sampling that surrounds the survey estimate. The confidence interval gives us a way to express how "good" an estimate is, how precise it is. The larger the confidence interval, or the range, the more careful you should be about how you use the estimate.

The magnitude of sampling error in a survey could affect the conclusions you draw, or decide you cannot draw, when making comparisons among data sets. It is an especially useful guide in making comparisons between surveys and administrative records. If you want to make comparisons, look at the survey estimate's confidence interval, not just the estimate. The confidence interval is a tool you can use so you won't make too much out

---

[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. Forms change, for example, and copies of outdated ones (which provide information about how questions were asked) are not usually kept. Critical information about differences in the data sets over time may exist only in the memories of long-time employees and is lost once they leave the agency.

[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.

[4] Statistical Policy working paper 31, pg. 1-5, http://www.fcsm.gov/spwptbco.html

[5] Ibid., pg. 3-5.

of small differences between two estimates. It warns us to be careful about interpreting trends or making comparisons when the confidence interval is relatively large.

The difference between the survey estimate and the true value is the result of both sampling and nonsampling errors. Statistics based on a sample, such as from the *decennial census long form and the American Community Survey*, are estimates and may differ somewhat from what would have been obtained if data had been collected from every person. The American Community Survey design allows samples for multiple years to be added together to increase the size of the sample and reduce the variance, an improvement in the estimate. The larger sample improves your odds that your estimate is closer to what you would have gotten if you had counted every person.

A preliminary step in determining whether *apparent* differences between the characteristics of program participants and those estimated in the distributions from the long-form sample of the decennial census and the American Community Survey are *actual* differences is to compute the confidence interval. The sample survey's estimate, the midpoint of the confidence interval is published in the decennial census products and the documentation describes the method of computing the standard error and the confidence interval. The American Community Survey data sets show the survey estimate and the lower and upper bounds of the confidence interval.

- *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[6] was 1,471 to 1,857 poor children in calendar year 1989. The 1,824 children who received welfare fell within that range as expected.

Administrative records are intended to be a full count of clients and thus, sampling error is not a consideration.

Nonsampling Errors

All data sets -- complete censuses, sample surveys, *and* administrative records -- have nonsampling errors that bias the results and affect the accuracy of the statistics.

Nonsampling errors affect the data:

[6] The 90-percent confidence interval can be interpreted roughly as providing 90-percent certainty that the true number falls in the range between the lower and upper bounds.

(a) randomly by increasing the variability and are reflected in the computation of the standard error for sample survey data; or

(b) in a consistent direction by introducing bias not reflected in the standard error.

Nonsampling errors may be introduced during any of the complex operations used to collect, process, and publish statistics. They are referred to as "nonsampling errors" for the obvious reason that they are errors that have nothing to do with the chance errors that occur when part of the population is sampled.

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.

Recording information incorrectly is an example of a measurement error that occurs in both surveys and administrative records when people fill in information themselves or from the consistent errors of survey field workers and caseworkers:

- The person responding to the questions of a survey or a caseworker who will determine their eligibility for a program is a potential source of error, no matter how detailed the instructions are or how clear the questions. This may be the result of confusion. For example, a person may not report in a survey that they received public assistance because they didn't realize the name of the program had changed, such as from "AFDC" to "TANF." Misreporting may be intentional. For example, a person may not report substance abuse, or income from illegal activities, or they may not list all the people living in a household because they are breaking some rule and fear being reported.

- The enumerator or eligibility worker may misinterpret or otherwise incorrectly record information given by a respondent, fail to collect information, or collect information from households not designated as part of the sample. Such miscommunication can create large-scale, consistent errors of bias in all the work the field employee does.

- An evaluation study from the 1950 census showed that error is decreased when individuals fill out a form themselves (self reporting) compared with an enumerator filling it out. The errors that respondents make tend to be random unless there is a significant problem with the question that confuses groups of people. This is usually found and fixed ahead of time by extensive testing of all questions on decennial censuses and the American Community Survey.

- Some administrative data are collected to manage programs and to generate benefit checks for program participants. For these types of files, accuracy of the

data is critical. They are generally carefully checked for fraud, clerical errors, and management errors. The entries on other types of administrative records may not be checked so carefully and are subject to many types of consistent errors. These types of records include, for example, police reports of crime, the incidence of diseases or other health conditions, environmental hazards. The errors could be the result of training and the individual whims of the people writing reports or classifying the information, the political environment (such as willingness to report some types of crimes or diseases), and how well instructions are followed so the information can be processed correctly.

There is a discussion below of the effect of coverage, nonresponse, and processing errors, other sources of nonsampling errors that affect comparisons among data sets.

Differences in Data Collection and Processing Methods

Different data collection methods affect the results. This section discusses and provides examples of different results based on how information is collected, who is part of the data set, how well the intended population is covered, how the data that has been collected is processed, sources of error, and how confidentiality of individual information is protected.

Who Collects the Information and How?

The *American Community Survey and decennial censuses* primarily contain information provided by a household member who often fills out the form for all members of the household ("self reports"), including household members who may be unrelated. These surveys accept the responses provided without checking against other sources. Examples of the types of errors that may occur are discussed in the "Nonsampling errors" section below. There are often differences in distributions of characteristics from different data collection modes. If a form is not provided from an address after multiple follow-up contacts, a Field Representative, as a last resort, may collect a minimum set of information from a neighbor or landlord.

Administrative records provide aggregated data derived from various sources. Sometimes the information comes from the responses to questions on intake forms asked of clients in need of services. Sometimes, data collection occurs because of legislative requirements, such as when the accounting office of a business provides information about individual wages for taxation. Sometimes the data collection is because of events such as an arrest for a crime or the report of a health condition.

Forms differ among states and may change from year to year within a state. Intake forms may be filled out by a caseworker or by the client seeking a service or those living in institutions such as prisons, nursing homes, or mental institutions. The information is often checked and criminal penalties are possible if the applicant provides incorrect information fraudulently. Electronic cross-checking of information has increased in

recent years which reduces inconsistencies and other nonsampling errors in and among administrative records.

Another difference between administrative records and surveys is the data collection cycle. For example, 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 form is filled out. For example, the American Community Survey asks about total earnings from the 12 months before the form is filled. 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, relationships can still be studied.[7]

Additionally, geographic disparities may occur in the assignment of residence between surveys and administrative records. The American Community Survey collects statistics from households. Eventually, the plan is to also include a sample of people living in group quarters. Some data sets, such as Unemployment Insurance records, collect information from business reporting establishments. With complex corporate structures and affiliations, the address of such establishments could be from a corporate headquarters far from the jurisdiction where a person actually works. Even more often, the address of the establishment is in an entirely different place than where a person resides. In short, people working with administrative records must often use some other source to obtain a residence address. This means involving a third data set, such as drivers' license records, with addresses that may or may not be up to date. For example, in Florida, a drivers' license does not have to be renewed for seven years. The address of record may or may not be correct.[8] 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. Sometimes the address is absent from the record and sometimes states fail to assign addresses to a given location. This varies significantly by industry, metro/nonmetro status, and the growth rate of a region. Sweeney suggests an estimation method to "recover unbiased estimates…using biased data." Otherwise, he says:

> …community level analysis based on the data could easily result in spurious conclusions…especially…if time trends are used since perceived trends may simply be the result of improved administrative protocols for collecting and recording address information.[9]

What Are the Differences?

Who is in the data set?

---

[7] 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.
[8] Ibid., pp. 12-17.
[9] 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.

The *decennial census* is an attempt to count every person (see "coverage" below). It has also collected additional information on the "long form" from a sample of households (includes all household occupants) and a sample of the people living in institutional and noninstitutional group quarters. Like the *decennial census long form*, the *American Community Survey* is a sample and it is designed to represent the characteristics of the total population when it is fully implemented (from 1996-2003, it is the household population only and people living in group quarters were not part of the sample).

To maintain *confidentiality* of the decennial census and the American Community Survey responses, as required by law (Title 13 of the U.S. Code), the Bureau of the Census applies a confidentiality edit to the data before publication. This introduces a small amount of uncertainty into the estimates of the characteristics to avoid disclosure of information about any individual person, household, or housing unit. The confidentiality edit is controlled so that the basic structure of the data is preserved.

Administrative records are intended as complete counts of clients receiving services from a program and thus, a subset of the total population of a jurisdiction. A caveat is that when performance measures are involved, there may be incentives for administrative actions that de facto include or exclude potential clients from the final data set.

Many administrative records data sets include people who move in and out of programs over the course of a year. For example, as Phillip Rokicki of Florida Atlantic University points out, the Covered Employment and Wages (ES-202) Program measures the employment of people covered by Unemployment Insurance laws. He notes:

> The UI database captures the number of employer filled jobs, whether full- or part-time. If a person has two jobs, the person 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.[10]

Coverage

Some households and people are missed entirely in *surveys and the census*. This may produce bias in the results to the extent that the people missed have characteristics that are systematically different from those who do respond.

Coverage refers to the proportion of the total population, or eligible universe in the case of administrative records, included in the data set. Coverage error indicates some members of the "target" population, the focus of the data set, are systematically missed, overrepresented, or out of scope. We know, for example, from evaluation studies that infants were more likely to be undercounted in the 1990 census than were older children because of misreporting and enumerator errors in filling out forms. Counting people more than once (overcount) can occur as well. Overcounting occurs, for example, when

[10] Ibid., p. 18.

people move during the census period and incorrectly fill out forms in both places, when housing units are misclassified as occupied when they are vacant, and when an address is listed more than once in the Master Address File (such as both a rural route and a street name) and the household fills out the questionnaires more than once.

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.

Coverage problems may occur in *administrative records* as well. For example, forms can be lost or the data not captured properly, such as keying errors or misreading marks on forms that are optically scanned.

People may be incorrectly included or excluded from administrative records. An example of how incorrectly including or excluding people occurs in administrative records is estimating the number of people without regular housing from management information systems that track services for the homeless. By unduplicating Social Security Numbers (SSN), researchers estimate how many people receive services across various agencies (e.g., shelter, medical, legal assistance) for those without regular housing.

- *Undercount:* Most homeless people eventually receive some type of service although not every homeless person does.
- *Undercount*: Occurs if a service provider is not part of the data set.
- *Overcount*: Occurs in the estimates when people are included in the administrative records who have homes but their limited incomes allow them to receive services labeled as programs for the homeless (for example, urban soup kitchens and rural food banks).
- *Overcount*: Some people without regular housing provide more than one SSN to service providers and so they appear more than once in the data file.

<u>Estimating receipt of program benefits versus eligibility</u>

From administrative records, we know the number of people who receive benefits from a program but we don't know how many are eligible. The American Community Survey and the census long form, because they collect characteristics for the entire population, often have information for estimating the number who would be eligible for programs (potential) to compare with the number actually receiving benefits from programs. For

example, from the American Community Survey, we can estimate the number of people who meet the requirements for eligibility for assistance from programs such as Food Stamps, housing assistance, or medical assistance. Those estimates can be compared with the number who actually received assistance from the programs if the universes for the two data sets can be made essentially the same.

How Are the Data Processed?

Every step in processing data presents a potential source of error. Processing census questionnaires includes the field editing, followup, and transmittal of completed questionnaires; electronic capture of the remaining responses, and manual coding of handwritten responses, such the address of the place where a person works.

The objective of the field and processing operations the Census Bureau uses is to produce a set of data that describe the population as accurately and clearly as possible within the constraints of cost. To meet this objective, questionnaires are reviewed for consistency, completeness, and acceptability.

Surveys differ as to how questions that were not answered, are inconsistent with other information, or unacceptable (such as "don't know") are handled. How and whether follow up is done is described in the documentation[11] for each survey. Any remaining nonresponse to a question is "imputed" by computer edits that use reported data for a person or housing unit with similar characteristics to enhance the usefulness of the data. When imputation is very high for a particular item, however, the analysis of the data should take this source of bias into account. Imputation procedures use information from respondents to represent the characteristics of nonrespondents. The characteristics of those who do not respond may be very different from those who do respond.

The quality of the various choices of technology varies for capturing the information the respondent has put on the form. Data quality assurance operations measure error levels and the Census Bureau takes steps at each stage to ensure a high quality product. Manual data entry in administrative records used to be a significant source of error. Now, it is common to scan or electronically report responses. Errors occur in both methods and it is possible to measure the level of error with standard quality control procedures.

Because errors affect eligibility determination and benefits, administrative records generally include both clerical and computer checks during data processing to reduce errors. Administrative record files may allow "don't know" responses or no response if the information is not critical for determining a person's eligibility for benefits. For administrative records used as statistical files, imputation procedures are rare although it

[11] The 1990 census had an extensive follow-up operation for nonresponse to questions on the long form as does the American Community Survey. Census 2000 did not have this operation and data users should review the imputation rates for nonresponse to individual questions.

could be done more often. Analysts routinely do such procedures now with some files, such as those they use to analyze welfare and employment.

Comparing Information From Different Sources

There are conceptual differences between census data and administrative records in how the population universe is defined, and the time the data are collected as well as the reference period for a question. There are also conceptual differences in the way questions are asked, the order of questions, and definitions.

Are the Universes Similar?

Universes are totals or subgroups of the population, households, or type of housing available from a data set. To make comparisons among data sets, universes need to be as similar as possible. To avoid comparing apples and oranges, a fundamental step is to review the definitions of the universes for the particular year of the data set(s) you are using. As is shown by the examples, below, who is included or excluded can change over time and may differ among data sets, even though the words sound the same.

The decennial census attempts to include the total population (the total excludes missed people and includes those who were double counted). The *American Community Survey*, when fully implemented, *and the decennial census long form* are designed to be representative of the *total population*. Table titles define the specific universes for that table.

The universe for program *administrative records* is a *subset of the total population*. It may be the group of clients who receive services and benefits from the program rather than the total universe of those eligible for assistance**.** Some files are of those who are eligible rather than of service recipients (such as the Selective Service file of those eligible for the draft but not drafted).

- *Example:* Data from the American Community Survey and the decennial census are shown for: (a) housing units; (b) households, families, and "persons"; and (c) specific subgroups of the total population (for example, age groups or race/ethnic groups) or housing stock (e.g., vacant units). The decennial census also includes people living in group quarters.[12] A common error is to compare "family" or "person" information from administrative records with census data for "households" or the total population. Households may include unrelated people living together or living with families. "Total population" includes people living in group quarters.

- *Example:* Census tabulations include citizens and foreign-born people who are not citizens, including an unknown number of undocumented immigrants. The

[12] The universe for the American Community Survey through at least the 2003 data collection is for the household population only. The Census Bureau plans to include the group quarters population starting in 2004, pending Congressional approval of funding.

definitions are found in the survey documentation.[13] It is not always documented in statistical files from administrative records as to whether immigrants are included or excluded.

- *Example:* School enrollment data need to be examined for comparability of definition, residence, and time frame. For example, are students in special programs included as they are in the census long form and American Community Survey? Are students included who are enrolled in school districts outside the one where they live? The American Community Survey and the decennial census long form assign elementary and high schools students to the jurisdiction where their parents live and college students to the jurisdiction where they attend school. Are students in private schools included as they are in the sample surveys? Time frames differ as well. Typically, schools collect enrollment data at a single point in time and when the enrollments are at their highest level for the year.[14]

- *Example:* For several reasons, the unemployment figures from Census Bureau surveys, not only differ among the survey, but also are not comparable with published figures on unemployment compensation claims. One reason is related to universe differences. Figures on unemployment compensation claims exclude people who have exhausted their benefit rights, new workers who have not earned rights to unemployment insurance, and people losing jobs not covered by unemployment insurance systems (including some workers in agriculture, domestic service, and religious organizations, and self-employed and unpaid family workers). People working only a few hours during the week and people with a job but not at work are sometimes eligible for unemployment compensation but are classified as "employed" in the census products. Differences in the geographical distribution of unemployment statistics arise because the place where claims are filed may not necessarily be the same as the place of residence of the unemployed worker. See: http://www.census.gov/acs/www/Methodology/Definitions/Employme.htm

- *Example:* "Earners" in the American Community Survey are people 16 years and older who received wage or salary income and net income from self-employment before deductions such as for personal income taxes, Social Security, bond purchases, union dues, and Medicare deductions. Nationally, about 2 percent of jobs are not covered by unemployment insurance. The Unemployment Insurance database excludes the self-employed, state and local government workers, agricultural workers, unpaid family workers, railroad workers, and some types of nonprofit and religious organizations.[15]

---

[13] A glossary of terms is available at: http://www.census.gov/acs/www/Methodology/Definitions.htm

[14] Barry Edmonston and Sharon M. Lee, "Use of the American Community Survey for Educational Planning in Portland Public Schools," 2001, unpublished study for the U.S. Census Bureau.

[15] Rokicki, op. cit., pp. 10 - 12.

- *Example:* Classification systems vary. For example, the Census Bureau codes occupations and industries according to standard code lists used in federal data systems and in agreement with standards that are also used in Canada and Mexico. These are detailed classifications and expensive to code. Administrative records do not necessarily follow these standards or any other standards. Some are particular to a state, some to an individual researcher. State laws, court cases, and business practices that exempt some groups from the universe affect the classifications used in administrative records. How workers in nonstandard arrangements respond to a survey and how they are classified in administrative records may differ substantially. For example, researchers have noted growth in the number of workers correctly *and* incorrectly classified in administrative records as "Independent Contractors."[16] This classification removes the workers from the Unemployment Information wage reporting system and the requirement that business entities pay Social Security, Medicare, and Workers' Compensation taxes. There are also differences among states in exemptions of specialized occupations and alien non-immigrants.

Are the Questions the Same?

How you ask a question and where it is placed on the questionnaire affects the way people respond. This may be an issue in comparisons between the American Community Survey, the census, and administrative records. There is substantial research, for example, on the effect of the order of the questions on race and Hispanic origin.[17]

The definitions of terms used in the questions and the response choices vary among data sources. Results, therefore, are not always comparable even when the words are the same.

- *Example:* In public assistance records, race and ethnicity, as well as income, are usually defined differently from the definitions used in the decennial census long form and the American Community Survey.

    - In the surveys, people who report they are of "Hispanic origin" may be of any race, whereas in many administrative records, Hispanic origin is treated as a racial group.

    - In the decennial census and American Community Survey, "income" refers to money income only. Noncash benefits are not part of the poverty

[16] Lalith de Silva, Adrian Millett, Dominic Rotondi, and William Sullivan with contributions by Elizabeth Fishcher and Mark Sillings, Planmatics, Inc. for U.S. Department of Labor, Employment and Training Administration, Office of Workforce Security, "Independent Contractors: Prevalence and Implications for Unemployment Insurance Programs," OWS Occasional Papers (http://www.ttrc.doleta.gov/owsdrr/), 2000. Alternative and nonstandard work arrangements include: contingent workers, contract workers, day laborers, independent contractors (both self employed and those who receive wages or salaries), leased employees, on-call workers, and temporary direct hires or temporary workers paid by an agency.

[17] Population Division Working Paper No. 18. Results of the 1996 Race and Ethnic Targeted Test, May 1997. Available at http://www.census.gov/population/www.documentation/twps0018/

definition. Different states have different rules for defining "income" for the receipt of public assistance and rules change within a state over time as programs change. For example, in 1990, the definition of "income" for Maryland's welfare program to determine eligibility for the program included earned income (wages and self-employment earnings) after allowable deductions and disregards, as well as unearned income with some exclusions such as Food Stamps and other means-tested benefits. In addition, a family's resources or assets were taken into account when determining eligibility.

- *Example:* The American Community Survey and the decennial census long form ask about the mode of transportation people use to get to work. Categories include "streetcar or trolley," "subway or elevated train," and "railroad." It isn't clear how people respond in jurisdictions where a mode of transportation is called "light rail" because there is no response choice with that name. This is an example of why it is essential for researchers to review questionnaires and forms.

- *Example:* Profiles are available of "adjusted gross income" from summarized IRS individual tax forms. The results differ from the sample surveys because not everyone files tax returns (legally and illegally). Additionally, the definitions of income are not the same:

    - The IRS concept of income includes inheritances and capital gains, for example, from the sale of stocks and one's home, and allows for some income exemptions such as IRA and thrift savings.

    - In the census and the American Community Survey, "total income" includes wages and salaries, net self-employment income, interest, dividends, net rental income, royalties, Social Security, railroad retirement, Supplemental Security Income (SSI), cash public assistance and welfare payments, retirement, disability, and other cash income received on a regular basis and before deductions. The American Community Survey and the decennial census long form specifically ask for gross receipts before deductions and exclude capital gains, money received from the sale of property unless that is done as a business, withdrawals from bank deposits, money borrowed, tax refunds, gifts, and lump-sum payments such as from inheritances and insurance. See the definitions on the American Community Survey web site: http://www.census.gov/acs/www/Methodology/Definitions.htm

    - Members of some families file separate returns and others file joint returns. Consequently, the income unit is not consistently either a family or a person.

Are the Time Periods Comparable?

The decennial census is a "snapshot" of a point in time. Decennial census questions generally, but not always, refer to April 1 in the year ending in "0." Answers to demographic and many housing questions are supposed to be answered as of April 1, regardless of when the form is actually filled out. Answers to some questions, particularly the economic questions and some of the housing questions, have different reference dates within the question. Some questions ask about one's activities the week or year preceding the census. For example, in the 2000 census, a person's place of residence, age, marital status, family status, and race/Hispanic origin is what it was on April 1, 2000. Income refers to the person's total money income for the calendar year, January 1-December 31, 1999, as discussed below. Look at the survey questionnaire or the form for the administrative record to determine time references.

Data collection for the American Community Survey occurs continuously over a year. The estimates for the summarized characteristics of an area are a 1-, 3-, or 5-year average This is an issue in comparisons with administrative records. Some reference dates are different from those of the decennial census. For example, the income questions in the American Community Survey ask about the 12 months prior to the time of the interview rather than the decennial's calendar year. Continuing evaluation research, available on the Census Bureau's website, tries to determine whether most people follow the instruction literally or actually provide income for the prior calendar year regardless of the instruction. Enrollment in school refers to any enrollment in the three months before the sample survey form is filled out, and the migration question asks about the person's place of residence one year prior to filling out the form. By contrast, for Census 2000, it was 1995, five years prior to April 1, Census day.

Administrative records may refer to the averages for a calendar or fiscal year. The average may refer to a "budget month," a "processing month," or a "payment month." The reference period (fiscal year? calendar year?), and the means of calculating annual averages, is not always documented.

- *Example:* Comparisons that involve income and poverty status are prime examples. In calendar year 1989, the reference year for the census income questions, the economy was growing in many areas and profiles of income and poverty status reflected their particular situation for the year 1989. Shortly after the 1990 census was completed, the economy experienced a recession (July, 1990- March, 1991).[18] The number of welfare cases in Maryland began to rise significantly starting in the second half of 1989, preceding the recession in Maryland by a year. It is incorrect to compare decennial census poverty numbers for 1989 with welfare caseloads in 1990. It is even worse to make the comparison when the decennial poverty data were released in late 1992 and the number of welfare cases had climbed even higher, from about 63,100 families in CY1989 in Maryland to nearly 79,200 in CY1993.

[18] Researchers should check for a similar situation in 2000. The economy was strong at the time of Census 2000 and in 1999, the reference period for the income question. Changes began later in 2000.

- *Example:* Ridership numbers for public transportation, such as busses, may differ between what a locality collects and the census/American Community Survey. The latter ask how the person *usually* got to work *last week*. In Calvert County, MD, the American Community Survey estimated that there were about 37,000 workers in 1999 and that about 600 people commuted to work by public transportation. The 90-percent confidence interval was 267 to 933. The county's Chief of Transportation said that his records indicated that more than 1,000 people rode the bus to work. Why the difference, he asked? There are a number of possibilities to examine. If, for example, 1,000 people ride the bus to work often but in only 60 percent of the weeks, it is their usual form of transportation. Nevertheless, the week before receiving the American Community Survey, some may have driven to work. One would have to look at the local records of ridership to determine, for example, if they count people fractionally by the proportion of the time they use the public transportation. Further, the sample survey questions do not account for complex transportation modes such as people who ride a car to the metro station and then take a bus the rest of the way to their place of work.

Example: Comparison of Earnings Between the American Community Survey and Unemployment Insurance Records in Calvert County, MD and Broward County, FL

Two evaluation studies done for the Census Bureau for Calvert County, MD and Broward County, FL[19] demonstrate some of the issues researchers face in preparing multiple data sets for comparisons with the American Community Survey. The studies compared reported earnings in the American Community Survey profiles with summarized special tabulations from state Unemployment Insurance (UI) records.

In both studies, the researchers spent months obtaining permission from the respective states to do the summary special tabulations of the administrative records. They had to demonstrate that the studies would meet state objectives and establish stringent confidentiality protocols to process the data.

Unlike many administrative records data sets, the definitions and universe for UI data are well documented and many data elements are comparable across states. UI earnings data are collected from business reporting establishments and do not include the home addresses of workers. The American Community Survey (ACS) collects data from households. The researchers used another administrative records set to assign a county of residence for workers. It is impossible for administrative records to match exactly the

[19] 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. 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).

ACS "two-month residence rule" and so an unknown error level is introduced in comparisons between the ACS and the results from the administrative records.

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 unemployment insurance taxes are collected. The UI program does not include self-employed workers, state and local government employees, unpaid family workers, railroad 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 for earnings to be reported 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 and state and local government 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 first and last line of Table 1).

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	41,878 – 44,572
Unemployment Insurance	733,394	42,711
ACS self-employed workers	78,658 (74,728 – 82,588)	3,313 (2,706 – 3,920)
ACS state government workers	12,661 (11,049 – 14,273)	1,191 (821 – 1,561)
ACS local government workers	55,901 (52,583 – 59,239)	3,651 (3,016 – 4,286)
Amer. Community Survey 90% confidence interval for estimated number of self-employed + state/local govt workers	141,818 – 152,622	7,202 – 9,108
UI + ACS self-empl. +ACS state and local government workers	**880,614**	**50,866**
Combined UI/ACS estimated interval	**591,576 – 886,016**	**35,209 – 51,819**

NOTE: Unemployment Insurance records do not include all classes of earners.
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.

While Table 2 shows that the two data sets result in very different estimates of median earnings, the distributions in Chart 1 indicate that the differences between the data sets are most pronounced for those with earnings[20] of less than $10,000. Sampling error in the American Community Survey does not account for the differences. 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.[21] The American Community Survey asks whether earnings were received 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 occasional, temporary work.

Table 2. Estimates of Median Earnings in 1998 in Broward County, FL and Calvert County, MD from the 1999 American Community Survey and Summarized 1998 Unemployment Insurance Records

Data Set	Median earnings	
	Broward County, FL	**Calvert County, MD**
American Community Survey estimate (90-percent confidence interval)	$24,459 ($24,096 – 24,822)	$30,317 ($29,855 – 30,779)
Unemployment Insurance	$16,967	$18,069

NOTE: Unemployment Insurance records do not include all classes of earners.

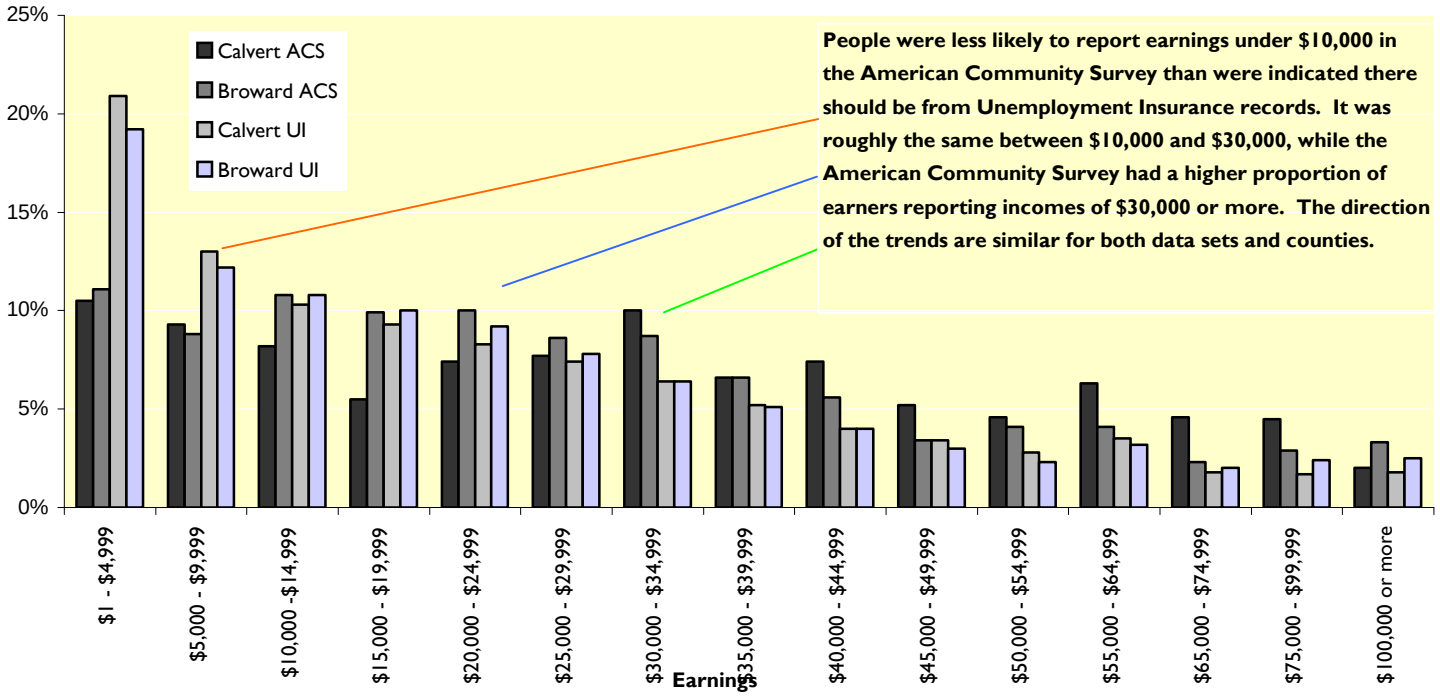
Source: 1999 American Community Survey, Table P67. David Stevens, Jacob France Institute, University of Baltimore, summarized 1998 Unemployment Insurance records from Maryland Department of Labor, Licensing, and Regulation (report forthcoming); and Phillip S. Rokicki, "A Comparison of American Community 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.

[20] Includes bonuses and commissions in addition to wages and salaries.

[21] 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.

Chart 1

Percentage Distribution of Earners Reported in the American Community Survey and Unemployment Insurance Records: 1998

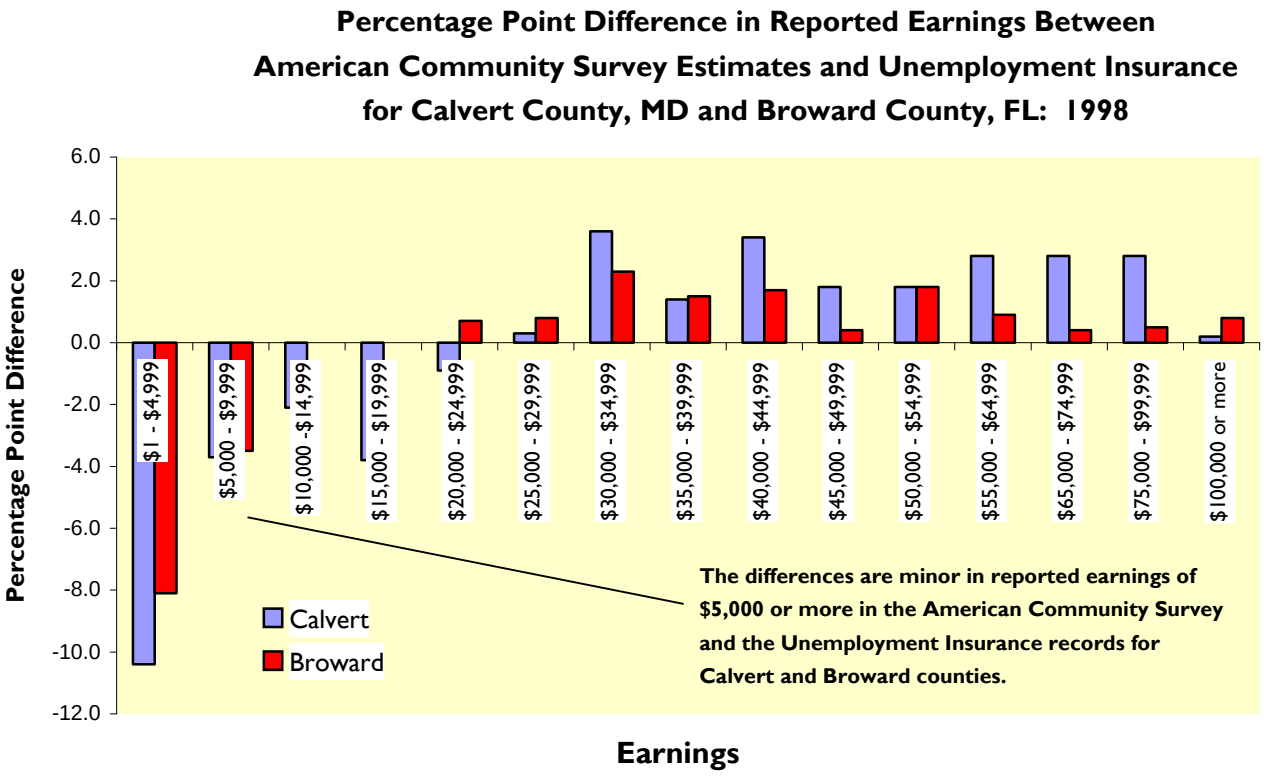


NOTE: Unemployment Insurance records do not include all classes of earners.

Source: U.S. Census Bureau, 1999 American Community Survey, Table P136; 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.

Chart 2 shows us that Calvert and Broward counties, despite their very different demographic characteristics, show similar trends. Chart 2 shows that differences are minor in reported earnings of $5,000 or more between the American Community survey estimates and Unemployment Insurance records.

Chart 2



NOTE: Unemployment Insurance records do not include all classes of earners.
Source: U.S. Census Bureau, 1999 American Community Survey, Table P136; 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.

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.[22]

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

Methodological Research
Needed To Develop Community Statistical Systems

There is much that can be done already to use geographically-based data sets in conjunction with each other. Once we have dynamic views from the updated trends of the American Community Survey, there is an enormously increased potential to inform public policy beyond the traditional uses of the static, historic view from the decennial census. There is research needed, however, to meet the full potential of the American Community Survey in conjunction with other data sets. And, as Charles Alexander has said, the statistical profession must help data users by clearly communicating errors and differences in data, encourage documentation of the methodology and definitions of data, and suggest statistical standards for data collection and processing.[23]

Below, we discuss some research objectives, new opportunities, and research needed to use multiple data sets to inform public policy.

Objective: **Create modern community statistical systems for informed strategic planning**

Primary responsibility for government program strategies and results has shifted from the federal level to state, county, and local governments. For strategic planning, governments need a system of current and comparable statistics. To help meet those needs, states and communities have converted geographically-based management information systems with records of program participants into files they can use for statistical purposes. Software for mapping and data base management has made the analysis of data files relatively fast and cheap. States and local governments use the resulting analyses for improved planning and program evaluation.

One limitation of analyses of administrative records is that they are only for the subset of the population that participates in the program. Information about the total population and subsets of the population, come from the decennial census and the American Community Survey. Until the American Community Survey is fully implemented, the decennial census is the only source of comparable population and housing information about the total community. The decennial census is a static picture that communities have previously had to use for 12-13 years until the next census profiles become available. The American Community Survey eventually will provide comparable statistics every year. By providing updated profiles for the total population every year, communities will be able to track change in the characteristics of their population and housing stock. In addition to providing crucial statistics about changes in the characteristics of people moving in and out of communities, information that is vital for informed strategic planning, the American Community Survey could help areas track movement from, for example, one steady state to

[23] Ibid.

another, the amount of time for which adjustment occurs between states, and correlations among different community characteristics.

A modern state and community statistical system would use multiple data sets that are geographically based to develop a dynamic picture that better informs those who make decisions about program effectiveness and direction.

- ***A research objective is to develop methodology for using administrative records in statistical models in conjunction with the geographic-area profiles from the American Community Survey. Such models can improve estimates, projections, and probability statements of events.*** Individual privacy is maintained by using data sets matched to small geographic levels rather than individual people. Models that use multiple sources of geographically-based information provide the possibility of scenario-based planning for a community's future to inform "what if" questions. We could better explore the likely impact of policy options, such as on community development.

- The American Community Survey has value in traditional statistical analyses such as regressions and mapping. It provides even greater value by enabling development of new research methods and new software that has greatly advanced the uses of maps. ***A research objective is to develop the next generation of Geographic Information Systems (GIS) that displays the American Community Survey statistics appropriately and in spatial interaction models (SIM).*** Because the American Community Survey is a sample and the statistics are updated every year, accounting for sampling error in comparisons is essential. If that is not done, data users are likely to conclude wrongly in some cases that there is change when in fact the apparent change is due to chance. GIS has traditionally not displayed sample data to indicate the range of uncertainty for the estimate. The American Community Survey potentially brings new opportunities in the use of GIS in spatial models[24] that predict "what if" reactions to changes in policies and practices and events.

Objective: **Identify sources of difference between administrative records and the American Community Survey.**

We expect estimates of population and housing characteristics from surveys such as 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. Differences result from sampling error in the survey data and nonsampling errors in both sources, such as definitional differences, response errors, processing errors, and coverage – that is, missing people and double counting. ***A research objective is to identify the impact and sources of differences between the American Community Survey and various administrative records.***

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

Objective: **Address data quality and document administrative records for research purposes.**

- ***A research objective is to address data quality issues ins data sets and to identify what needs to be done to convert program records to files that are useful for statistical analyses of communities.*** To compare multiple data sets, states and communities need to document for analysts essential information about the data sets and make such information readily available such as through the internet. For example, not all administrative records are geocoded to the census tract level. Statistical policy should be coordinated among the multiple data sets to standardize, to the extent possible, definitions of ways to ask demographic questions such as age and race, processing and editing rules such as for missing or inconsistent data, and the coding of characteristics such as occupations and place of work. To increase comparability among areas, it would be advantageous for the standardization to be consistent with the conventions used in the American Community Survey to the extent possible.

- ***A research objective is to identify elements for formal documentation of program records and to maintain them as an historical record to enhance their access and usefulness to analysts.*** Formal documentation includes, for example, intake forms, rules regarding program eligibility, definitions, processing rules such as for blank or "unknown" entries, and reports on the accuracy of the data.

Summary

There is enormous potential for improving estimates, projections, and informing public policy through research that uses multiple data sets. Multiple data sets increase the types and number of public policy questions researchers could address. Ideally, we want the highest quality estimates possible within the constraints of cost for making decisions. High quality statistics are not sufficient, however.

For that potential to be met, data sets need to be as comparable as possible. But, there are always differences among data sets. Analysts need to know about the differences and to account for the differences before coming to any conclusions. The better we understand the types of differences and errors in data sets, and the more we measure the extent of errors, the better we can judge whether we can use estimates from different data sets in conjunction with each other.

You may or may not be able to make comparisons, depending on your purpose and the cost of being wrong about decisions based on the results of the research. As we saw from the studies of earnings in Calvert County, MD and Broward County, FL, the American Community Survey and the Unemployment Insurance records showed, for the most part, similar directions in the patterns from the two data sets. The information about the sources of differences would help researchers develop models that use the two data sets along with current demographic characteristics (including the characteristics of migrants)

to improve current estimates of earnings for counties and to develop projection and "what if" models.

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 such 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).