

## **New Directions in the Development of Annual Population Data in the United States? \***

### **Abstract**

Small area estimation programs in countries such as the Australia, Canada, and United States, among others, have taken a different direction than those found in the national statistical agencies of many European countries, where population and other registers are used to more effect. In the United States, however, the advent of a continuously updated Master Area File (MAF) following the 2000 census represents an information resource that has not yet been fully tapped for purposes of developing timely, cost-effective, and precise population estimates and projections for even the smallest of geographical units (e.g., census blocks). We argue that the MAF can be enhanced (EMAF) for these purposes. In support of our argument we describe a set of activities needed to develop EMAF, each of which is well within the current capabilities of the US Census Bureau and discuss various costs and benefits of each. We also describe how EMAF data could be directly assessed for statistical uncertainty and its use as a basis for developing population projections containing a wide range of ascribed (e.g., age and sex) and achieved characteristics (e.g., educational attainment and employment). Although such a development would bring the US Census Bureau's small area population estimation programs more in line with its European counterparts, there are several important challenges we describe that must be surmounted, including issues of public trust, confidentiality, and tradition. We conclude by observing that countries currently operating population estimation programs like those found in the United States are likely to be considering similar questions and facing similar challenges.

**Key Words: Registry, Population Estimates, Housing Units, Master Address File**

## **I. Introduction**

In this paper, we explore the needs of researchers in regard to population estimates. We also note that our topic, by necessity, overlaps not only with the population information needs of national and local users in the United States and other countries with strong vital registration and other administrative records, but lacking a population registry system (e.g., Australia, Canada, New Zealand, the United States), but with research that could be done on population estimation programs in these countries. Thus, this paper not only covers the needs of researchers in regard to population estimates in countries lacking registry systems, but also some of the research needs in the area of population estimation for these same countries. However, our paper focuses on these needs in the United States, as identified by the US Census Bureau (2006) and others (Breidt, 2005; Swanson and Pol, 2003).

Who are the researchers that need information in regard to population estimates? In answering this question, we use as a starting point the distinction between applied and basic demography offered by Swanson, Burch, and Tedrow (1996): (1) “applied demography” is primarily concerned with solving exogenously-defined problems by producing the information necessary to effect practical decision-making while minimizing the time and resources needed to produce this information; and (2) “basic demography” is primarily concerned with solving endogenously-defined problems by offering convincing explanations of demographic phenomena while viewing time and resources as barriers to surmount in order to maximize precision and explanatory power. By simply applying the distinction noted above to researchers that need information from the Census Bureau’s estimates program, we have the following: (1) applied

researchers that need information in regard to population estimates are primarily concerned with solving exogenously-defined problems by producing the information necessary to effect practical decision-making while minimizing the time and resources needed to produce this information; and (2) basic researchers that need information in regard to population estimates are primarily concerned with solving endogenously-defined problems by offering convincing explanations of demographic phenomena while viewing time and resources as barriers to surmount in order to maximize precision and explanatory power.

As a means of providing a context for this effort it is important to recall why estimates are done in the United States. As we know, the census is the most complete and reliable source of information on the number of people in the United States – as well as in Australia, Canada, England, and New Zealand, among other countries. In addition to actually conducting census counts, there are three other characteristics that link the United States with these other countries: (1) well-developed administrative records systems (e.g., vital events registration); (2) regular census counts; and (3) no population registration system, such as those found in the Nordic countries. As we know, a census is a time-consuming and costly endeavor. In the United States, a census of the population is done only once every ten years; in Australia, Canada, England and New Zealand, for example, it is once every five years. Because there is the potential for constant and sometimes quite rapid population change, especially at the sub-national level, census statistics for every tenth and even every fifth year are often inadequate for many purposes (Waldrop, 1995). To fill this gap, population estimates are used by government officials, market research analysts, public and private planners and others for determining national and sub-national fund allocations (Murdock and Ellis, 1991; Serow and Rives, 1995; Siegel, 2002),

calculating denominators for vital rates and per capita time series, establishing survey controls, guiding administrative planning, developing marketing, and for descriptive and analytical studies (Long, 1993; Pol and Thomas, 2001: 93-95; Swanson and Pol, 2005). In the United States, the Census Bureau is not the only provider of population estimates (Bryan, 2004b: 524-526), but it is the ultimate source of estimates and the data needed to develop them.

In order to meet the need for current population figures, many estimation methods have been developed, virtually all of which can be categorized into one or the other of two “traditions:” (1) demographic (Bryan, 2004b); and (2) statistical – that is, the methods used by those who do sample surveys (Kordos, 2000; Platek, Rao, Sarndal, and Singh, 1987; and Rao, 2003). Demographic methods are used to develop estimates of a total population as well as the ascribed characteristics – age, race, and sex - of a given population. Statistical methods are largely used to estimate the achieved characteristics of a population – educational attainment, employment status, income, and marital status, for example. As is the case in the national statistical agencies of other countries, the US Census Bureau produces estimates using both of these traditions, demographic and statistical.

In this paper, we focus the discussion on methods that fit within the demographic tradition and only briefly consider those in the statistical tradition. However, we identify links among selected methods in both traditions. This discussion provides a point of departure for our recommendations in regards to the needs of researchers.

Before launching into the main body of our paper, we also want to note that our discussion primarily covers the definition of population used by the US Census Bureau, which is based on place of “usual residence.” This also is known as the “de jure” population (Cook, 1996;

Wilmoth, 2004). We also note that there is the concept of a “de facto” population (Cook, 1996; Wilmoth, 2004). Examples of de facto populations are many. They include vacationers (of interest, for example, to the casino industry in Las Vegas and the Hawaii Visitors Bureau), migratory workers (of interest, for example, to health care, school, and other social service providers), and the people who work in the central business district of a large city each day, but leave it largely vacant in the evenings (of interest to the San Francisco City Planning Office, for example). While estimates of de facto populations are of great interest, they are very difficult to make in the United States because of the lack of census type benchmarks (Cook, 1996, Smith, 1994). We identify this issue as being important, but it is beyond the scope of our mandate to cover research needs for de facto populations in depth. We only suggest here that the US Census Bureau is the logical agency to develop systematic and comprehensive estimates of de facto populations in the United States – as are its sister agencies in other countries currently operating similar population estimation programs.

The remainder of this paper consists of four sections, endnotes, references, and two appendices (one of which has itself references). The following section provides an overview of basic concepts, data sources, and methods used to estimate populations in the U. S. The third section discusses the needs of researchers, both applied and basic, while the fourth section offers a suggestion for meeting the needs of these researchers. The fifth section discusses the obstacles associated with this suggestion and how they might be overcome.

Appendix A is a reproduction of the principles underlying the US Census Bureau’s estimates and projections programs.

## II. Basic Concepts, Data Sources, and Methods

In this section, our intention is not to cover concepts, data sources, and methods related to population estimates in depth. Rather, it is to generally describe them while providing citations to more detailed descriptions and discussions.

**Basic Concepts.** 1. Following Smith, Tayman, and Swanson (2001: 16), we make the following distinctions among the terms “estimate,” “projection,” and “forecast.”

*Estimate* – A calculation of a current or past population, typically based on symptomatic indicators of population change.

*Projection*-- The numerical outcome of a particular set of assumptions regarding future population trends.

*Forecast* – The projection deemed most accurate for the purpose of predicting future population.

In regard to an estimate, there also has been a tradition of distinguishing between “post-censal” and “inter-censal,” where the former refers to an estimate for a date between two censuses that takes the results of these censuses into account and the latter refers to an estimate for a date subsequent to the most recently available census (Bryan, 2004b: 523).<sup>1</sup> These definitions and distinctions fall into the demographic tradition. Among survey statisticians, the demographer’s definition of an estimate is generally termed an “indirect estimate” because unlike a sample survey, the data used to construct a demographic estimate do not directly represent the phenomenon of interest (Swanson and Stephan, 2004: 758 and 763). In this paper we use the demographic tradition’s definitions and distinctions unless specifically noted.<sup>2</sup>

Another useful set of concepts is the notion of “stocks and flows”. As defined by Popoff and Judson (2004: 603), “...stock data are the numbers of persons at a given date, classified by various characteristics...(and) are recorded from censuses....flow data are the collection of or summation of events. At the most basic level this includes births, deaths, and migration flows....” This distinction is useful for purposes of this paper because, as is discussed later in this section, there are population estimations methods that solely rely on “stock” data while others rely on a combination of “stocks” and “flows.”

Finally, it is useful here to define micro data and aggregated data. we take micro data to mean records for individual persons. These records are often linked by relationships to form family and household records and we use the term “micro data” to refer to these linked records as well. The “Public Use Microdata Sample” (PUMS) is such a file (Swanson and Stephan, 2004: 772). Aggregated data are summations of records of individuals (families and households) such as one would find in a table. The aggregations are often done to specific geographic areas, but they can also be done for types of people across different geographies. The life table constructed by Kintner and Swanson (1994) for retirees of General Motors is an example of such an aggregation

**Basic Data Sources.** All estimates, including post-censal ones, rely on one or more censuses and use administrative record systems on which different estimation methods for census-defined populations rely – vital events, tax returns, housing permits, assessor parcel files, utility hookups, licensed drivers, covered employment, K-12 enrollment, medicare, and child support payments, among others ( Bryan, 2004a; Bryan, 2004b). It is important to note that there is some variation in availability and quality of administrative records systems by state and by

local jurisdictions in the US as well as variation among countries. For example in many areas of the United States, Kindergarten through 8<sup>th</sup> grade enrollments are used in the calculations of population estimates to avoid mistaking students who drop out of high school as out migrants from the area. (McKibben, 2006)

It also is important to note that the US Census Bureau maintains as much consistency in data sources and methods as it can because among other desirable features, it wants to have a consistent set of estimates for a given “vintage” year (See Appendix A of this report, U.S. Census Bureau, n. d.). We note here the emergence of an important resource directly collected by the US Census Bureau – a Master Address File (MAF) constructed for the 2000 census that is updated and maintained until the next census. This is a new resource for the Census Bureau’s estimates program because in the previous “mail-out/mail-back censuses, the MAF was constructed from scratch before each census. As observed nearly 25 years ago by Pittenger (1982) and more recently by Wang (1999), this housing unit inventory is serves as a key resource in the Bureau’s ability to construct population estimates and we use his key ideas in discussing how the MAF can so be used later in this paper.<sup>3</sup>

**Methods.** Although it is not used directly in any of the standard population estimation methods used at the sub-national level, the fundamental demographic identity known as the balancing equation forms the conceptual framework for most of these same methods. This identity is defined as  $P_t = P_0 + I - O$ , where  $P_t$  is the given population at time  $0 + t$ ,  $P_0$  is the given population at time 0,  $I$  is the number of persons entering the population through birth and in-migration during the period  $0 - t$ , and  $O$  is the number of persons exiting the population through death and out-migration during the period  $0 - t$  (Swanson and Stephan, 2004: 753).

This identity can be phrased in more detail to separate recognize births, deaths, in-migration, and out-migration and is used as a point of departure to discuss in detail the concept of “stocks and flows” and the measurement thereof encompassed in the following methods.

### **1. Simple Interpolation and Extrapolation Methods**

Although no longer widely used in their own right, interpolation methods (see, .g., Judson and Popoff, 2004) and extrapolation methods (see, e.g., Smith, Tayman, and Swanson, 2001) represent ways to construct, respectively, inter-censal estimates and post-censal estimates. These methods range from being relatively simple (e.g., linear trending) to very complex (ARIMA models). Both interpolation and extrapolation are based on mathematical formulas that are applied to “stock” data to produce “flows” that, in turn, generate estimates. As such, the principles underlying these methods, particularly extrapolation, are often found in other estimation methods (e.g., regression methods).

### **2. Housing Unit Method**

The Housing Unit Method (HUM) is a “stock” method that describes a basic identity in the same way that the balancing equation does. In the case of the HUM, this identity is usually given as  $P = H*O*PPH + GQ$ , where P = Population, H = housing units, O= Proportion occupied, PPH = average number of persons per household, and GQ = the population residing in “group quarters” and the homeless (Bryan, 2004b). Like the balancing equation, the HUM equation can be expressed in less detail (i.e.,  $P= HH*PPH + GQ$ , where  $HH=H*O$ , Smith and Cody, 2004: 2) or more detail - by structure type, for example (Swanson, Baker, and Van Patten, 1983). It also can be used in combination with sample data, which opens the door to developing measures of statistical uncertainty for the estimates so produced (Roe, Carlson, and Swanson, 1992).

Because of how data are collected, the HUM had not been a method that could be used for all sub-national areas and the nation as a whole until recently. However, with the continuous MAF, it has now emerged as a method that can be used by the US Census Bureau for all sub-national areas and the nation as a whole (Wang, 1999).

### **3. Regression Methods**

Regression approaches to population estimation are basically “stock” methods in which measures of change in the ratios of indicators to population are used as “flow” estimates that are extrapolated to generate population estimates (Bryan, 2004b). The flow estimates serve as independent variables in these forms, which means that the dependent variable is a measure of population change. Measures of change can be in the form of ratios, lagged ratios, and differences (Bryan 2004b). These regression methods require a nested set of geographies (e. g., the counties within a given state) and they are inherently embedded in statistical inference (Swanson, 2004). As observed by Prevost and Swanson (1985), the “ratio-correlation” form can be viewed as a regression-based version of the so-called “synthetic” method of estimation.<sup>4</sup>

### **4. Component Methods**

Component methods are directly based on the fundamental demographic identity known as the balancing equation. As such, they are stock and flow methods. Included in this set are “Component Method II,” “Cohort-Component Method,” and the “Tax Return Method,” each of which is described by Bryan (2004b). The stock data are comprised of census counts in each of these methods, which use administrative records (e. g, vital events) to develop flow estimates.

## **5. Administrative Records**

So-called direct estimates can be acquired from selected types of administrative records systems, namely the national population registration systems found in the Nordic countries (Bryan, 2004a: 31-33; Statistics Finland, 2004). Although the United States lacks a national population registration system, it has several national administrative record systems that serve as partial population registers, including those relating to social insurance and welfare and the payment of income taxes (Bryan, 2004a; Judson, 2000).<sup>6</sup> It is worthwhile at this point to again bring up the MAF, which represents a national housing registration system that can be used to generate estimates using the Housing Unit Method.

## **6. Other Methods**

Here, we include the economic-demographic models and urban systems models described by Smith, Tayman, and Swanson (2001: 185-237) as well as the iterative proportional fitting, log-linear, and multiregional methods described by Judson and Popoff (2004). To this list can be added the methods developed for statistically underdeveloped countries and those for estimating wildlife populations (briefly discussed in Endnote # 2) as well as the imputation methods used by the US Census Bureau to compensate for missing data (see, e.g., Swanson and Stephan, 2004: 762).

In concluding this brief overview of methods of population estimation, we note that it is often the case that various data adjustments must be made to effectively operate the preceding methods and that these adjustments serve as “other methods” in themselves (Wang, 1999). For example, the presence of non-household populations, such as found in prisons, school dormitories, and long-term care facilities, can affect the accuracy of virtually all of the methods just described, as

can the presence of seasonal populations, undocumented aliens, and the occurrence of disasters, natural and otherwise.<sup>5</sup>

### **III. Researcher Needs**

In this section, we describe what we believe researchers need and note some of the benefits. Recall that in the following section we propose a solution to meet these needs and describe some of its benefits and postpone a discussion of obstacles to the final section.

Ideally, what we believe is needed by all researchers is a system that provides an historical set of sub-county estimates of populations and their characteristics that can be rolled up to all higher administrative and statistical geographies for a given vintage to produce “one-number” set that is consistent with data both from decennial census counts and sample surveys regularly done by the Census Bureau. This is consistent with the principles underlying the Census Bureau’s estimates program (See Appendix A). Further, the ideal foundation of these estimates would, we believe be comprised of individual data on persons that are linked to households and other living arrangements in specific locations. What we have just described, of course, is something that does not exist for the United States – a national population register, a system that contains micro level data that can be rolled up and linked both across time and with other data, such as the case found in Finland (Statistics Finland, 2004). For discussion purposes, we will refer to this as the US national population file, or more simply, the “population file.”

Although the distinction is not clear-cut, most researchers, applied and basic, tend to use aggregated data. However, as noted later, some types of basic researchers would prefer to use

micro data, given that it is available. Moreover, while both are interested in historical data, most researchers, applied and basic, tend to use cross-sectional data while some types of basic researchers would likely lean toward truly longitudinal data. Both groups, applied and basic, would be served by a population file in these regard to this types of data, however.

Swanson and Pol (2004: 29-30) distinguish public sector and private sector interests in applied demography and observe that estimates and projections serve as an important link between the two sectors. Using a similar public and private sector dichotomy, Siegel (2002: 220-328) describes demographic applications for businesses on the one hand, and government and private nonprofit organizations on the other. Siegel's (2002) list includes: market research, business site location, sales forecasting, legal and regulatory requirements associated with business activities, consumer research, demographic aspects of domestic and foreign investment, the labor force, organizational demography, housing needs, transportation planning, socio-economic and health characteristics, political applications (e.g., apportionment and re-districting), the costs and benefits of insurance, the U.S. social security system, and the allocation of public funds between children and the elderly. Clearly, a population file containing the number of people under consideration for a given study would be useful in all of these areas of applied work, as would information on the characteristics of these people.

Basic researchers can largely be distinguished as being either mathematical demographers or "socio-economic" researchers - those that use demographic perspectives in answering questions in specific disciplines, such as economics, geography, and sociology (Burch, 1993, McNicoll, 1992 and Swanson and Stephan, 2004: 758). Often, those in the latter category are interested in the determinants and consequences of demographic changes.

Like applied demographers, most basic researchers use aggregated data in their work and for most mathematical demographers, this type of data is preferable. (Coale and Demeny, 1966; Dharmalingam, 2004; Li and Tuljapurkar, 2005; Pollard, 1973; Rogers, 1995; and Suchindran, 2004). While it is the case that those interested in using demographic perspectives to answer discipline-specific questions often use aggregated data (Clark, 1986; Rogers, Hummer, and Nam, 2000; Stockwell, Goza, and Balistreri, 2005; Treyz, Rickman, Hunt, and Greenwood, 1993), it is clear from discussions and analyses found in the literature, many would prefer to have micro data, if they were available. This is because many of these basic researchers are interested in hypotheses concerning individuals (Brandon and Hogan, 2004; Livingston, 2006; Mutchler and Baker, 2004; Ryan, Manlove, and Hofferth, 2006) and in using aggregated data to address their hypotheses about individuals, they have to deal with problems such as aggregation bias and the ecological fallacy (Freedman, 2004; and King, Rosen, and Tanner, 2005). As is the case for applied demography, we believe that a population file containing the number of people under consideration for a given study would be useful in both of these areas of basic research, as would information on the characteristics of these people.

We do not believe that there are many who would argue against the utility of a national population file for applied and basic researchers. We do believe that the situation would be similar for national, state and local data users. The issue here, of course, is that it is virtually a certainty that there will be no national population file in our lifetimes, if ever. American traditions and values are not in favor of such a system, given concerns about government intrusion into privacy. So, why have we bothered to discuss this ideal but unachievable data source? The reason is that there is an existing “register” in the US Census Bureau that can yield

something close to a national population register when coupled with the Bureau's record matching, extant data collection, and other capabilities. This register is the Master Address File, or MAF, and to it we now turn.

#### **IV. A Suggestion for Meeting the Needs of Researchers**

Before we offer our suggestion regarding the MAF and its potential for meeting the needs of applied and basic researchers, it is important to note that others have thought along similar lines in regard to other "registers." Here, we are thinking primarily of research into the development of an "administrative records census," which has been going on for at least 20 years (Alvey and Scheuren, 1982; Kliss and Alvey, 1984, Scheuren, 1999). Initially, much of this work was done within the U.S. Internal Revenue Service, but this has broadened to include other US agencies, including the Census Bureau (Prevost, 1996; Judson, 2000; Judson and Bauder, 2002). Research and other activities in the U. S. related to administrative records censuses have also been commented on by researchers outside of the country (Redfern, 1986). However, it is still the case that the US Census Bureau had not attempted to conduct a full-blown administrative records census (Bryan 2004a, Bryan 2004b, Bryan and Heuser 2004).

We also note that our suggestion is largely based on a call by Wang (1999) for greater recognition of the utility of the MAF in regard to population estimates. Wang also provided specific suggestions on how to overcome the problems associated with maintaining and updating the MAF such that the data were of high quality.

Wang's (1999) suggestions, along with the ideas underlying an administrative records census, lead directly to the idea of viewing the MAF as the basis for a national housing register, or more simply the "housing register." What primarily distinguishes the MAF from the housing register is the presence of population data. The first step in turning the MAF into a national housing register is to load it with "estimates" of population and related data for individual housing units. How might this work?

Initial estimates could be provided by matching selected census 2000 short form data to individual housing units in the MAF. On a regular basis (e. g, once each year), the individual housing unit records could be updated using similar "short form" data from the American Community Survey (ACS) in conjunction with demographic methods (e. g, survivorship estimation), direct substitution in housing units appearing in the ACS sample for a given vintage (i.e., a given year), and imputation and related estimation methods for those in the same vintage and area that are not in the ACS. Individual housing unit data from the "old" version would be so identified and remain attached to each record so that measures of change could be computed for individual records (i.e., individual housing units). Thus, the system would be a housing register containing a combination of collected and estimated data centered on demographic characteristics (i.e., age, sex, race, household relationships) distinguished, as appropriate, by year. When a year ending in zero is reached, the data for the preceding decade could be archived and a new file started for the coming decade.

For many applied researchers as well as basic researchers interested in mathematical demography, this short form housing register would serve most of their needs. For others, the housing unit records would need to start each decade with both short and long form data and be

updated accordingly with ACS short and long form data. Because the long form data were collected on a sample basis in the 2000 Census, this would mean that long form data would be imputed for individual housing units not in the sample. Once the annual update cycle starts, all long form data would be imputed for individual housing units not in the ACS sample.

What are some of the specific benefits of a national housing register? Here are some examples. To begin, we believe it would assist the Census Bureau in solving four of the problems facing its estimates program identified by Habermann (2006). First, “short form” data from the housing register would serve well as the population controls for the ACS. This could be particularly important for small pieces of geography. Second, the combination of short and long form data in the housing register would serve to improve estimates of internal migration as well as emigration and immigration. Third, the housing register would allow bringing additional data sources into the sub-national population estimates beyond the ACS, to include administrative data sources on employment and taxes. And, fourth, the housing register would allow for research needed to improve methods to achieve integrated and consistent population estimates at different levels of geography. In this regard, Habermann (2006) observes that the current approach begins at the county level, with the estimates controlled only at the national level.

The US Census Bureau has recently been confronted with the possibility of a reduction of more than \$50 million in the budget proposed by the Executive Branch for its FY 2007 operations (Lowenthal, 2006). This is not a new phenomenon and much of the impetus for reduced and otherwise tight budgets comes from the high costs of collecting data. In this regard, we believe that the housing register would also be of benefit. For example, Statistics Finland (2004: 26) reports that it was pressured by the Ministry of Finance to move to a register-based

system because of the recurring high costs associated with taking a census. After it made the change following its 1980 census, Statistics Finland (2004: 26) reports that in 2003 money terms the cost of its 2000 register-based census was less than one million euros while the traditional 1980 census costs were approximately 35 million euros. This evidence strongly suggests that a housing register would assist the US Census Bureau in containing costs.

A housing register would contribute toward having more timely, comprehensive, and internally consistent demographic, housing, and socio-economic data for the U. S. as a whole and its sub-areas. In regard to geography, we note that register based data are extremely flexible in that they can be geo-coded to a specific location (as opposed to being assigned to an area defined by administrative or statistical boundaries). This also means that the housing register can be overlaid with other features using GIS capabilities. The TIGER street address file comes to mind first in this regard. This would lead to an entirely new way of looking at the concept of a “small area,” in that boundaries could be drawn that are much finer than those allowed by the census defined block. This would allow much higher precision in defining areas for purposes of marketing, site location. Once up and running, this would also allow for greater ease in producing a consistent time series for areas in which administrative boundaries changed over time.

It is also worthwhile to note that if the housing register were assembled into a single register along with similarly geo-coded group quarters locations and commercial establishments, and public buildings (e.g., fire stations), the result would be tremendous data source for applied researchers. Imagine being able to map not only existing, but also historical and potential “future” service areas and their populations using such a system. Here, it is useful to note that is

precisely the situation that exists currently in Finland (Statistics Finland, 2004: 41-44). we also note that this proposal also is in line with recommendations made by the National Research Council's Committee on the Human Dimensions of Global Change (National Research Council, 2005).

To summarize, what we are proposing here is a register – “the national housing register” – in which each individual housing unit contains not only existing MAF variables (e.g., geocode, address, and structure type), but also information on occupancy status; in addition, each occupied housing would include variables that provide “short form” demographic characteristics and, if feasible, variables that provide some degree of “long form” socio-economic characteristics. Occupancy status and the demographic and socio-economic characteristics would be generated using a combination of decennial census and ACS data in conjunction with a combination of record matching and estimation methods, particularly imputation and related forms of modeling.

## **V. Discussion**

Turning now to the obstacles associated with our proposal for a national housing register, we begin with the issue of confidentiality. The National Research Council's Panel on Data Access for Research Purposes (2005) has identified the lack of resources and structural incentives for making data more readily available as major contributors to the difficulty of reconciling access to data by researchers with the need to preserve confidentiality.<sup>7</sup> The issue of confidentiality is not an insignificant problem. As the US Census Bureau recently learned, even the perception of a breach of confidentiality can become a major outcry (Clemetson 2004a, 2004b, 2004c; Lipton,

2004). One can see that the development by the US Census Bureau of any type of register containing information on individuals can run into public and political resistance due to confidentiality concerns. This was noted over twenty years ago by Pittenger (1982). However, we believe that this problem is not insurmountable in regard to our proposal for a national housing register. The National Research (2005) Council has issued recommendations to reconcile access and confidentiality and the US Census Bureau itself has appointed a Chief Privacy Officer and worked to put effective procedures in place regarding this reconciliation. There are recommendations for going even further (El-Badry and Swanson, 2007) as well as the ideas provided by the highly effective laws, rules, and procedures, developed by Statistics Finland (2004) to effect the reconciliation of access to data by researchers and the preservation of confidentiality.<sup>8</sup> Taken altogether, we believe that the US Census Bureau is capable of creating a national housing register that would be useful to researchers while also being subject to strong confidentiality safeguards.

What about the issue of privacy?<sup>7</sup> What may be ideal from a researcher's point of view may not be ideal from the perspective of others. For example, those concerned about the intrusion of the Federal Government into private lives would not be pleased at the prospect of what amounts to a national individual data base even no major outcry has been raised in regard to the three "lightly" regulated, non-mandated, de facto private sector registration systems maintained by Equifax, Experian, and TransUnion for purposes of determining credit worthiness. we believe that this may be a more difficult obstacle for the US Census Bureau to overcome than that represented by concerns over confidentiality. Much of this has to do with privacy being intertwined with the mix of constitutional mandate, case law, executive orders, and general

tradition that calls for an actual count of the population rather than the development of a register (Anderson, 1988; US GAO, 2003; Walashek and Swanson, 2006; Weinjert, 2003). Thus, the US Census Bureau and its allies would have to mount a dedicated effort to build public trust in the idea of a national housing register.

Another obstacle is the financial cost of developing a national housing register. An idea of these costs is given by Redfern (1986) in his discussion of the cost of converting from a traditional census to an administrative records census. However, once developed (or converted, as the case may be), it appears that the costs for a national housing register could be less than the system currently being used in the US for developing post-censal estimates and decennial census counts. We use here the information from Statistics Finland (2004: 26) discussed earlier in regard to the comparative costs of registries and censuses. It also is worth noting here that local officials in Finland update the country's population and housing registries (Statistics Finland, 2004: 21). Thus, we see no major cost obstacle in following Wang's (1999) suggestion that state and local governments be funded to assist in maintaining the MAF under the general supervision of the Census Bureau. Before such a major step is taken, however, it would be wise to research the various forms this could take. El-Badry and Swanson (2007) call for research on such a recommendation in terms of public involvement in administrative oversight of the Census Bureau.

What about accuracy? Can the proposed housing register provide accurate data? In a recent report, the Government Accounting Office (US GAO, 2006) identified MAF/TIGER problems that needed to be solved in order to have a good census in 2010. These problems include: (1) resolving address related issues such as duplication, omission, deletion, and incorrect locations in

the MAF; and (2) implementing GPS-based geo-coding of housing units. These same two problems represent sources of error in the proposed housing register. Consequently, if the US Census Bureau solves these problems in regard to the 2010 census, it will essentially do so in regard to the proposed housing register.

There are problems already known in regard to using the housing unit method of population estimation that would affect the MAF and therefore the accuracy of a housing register. Many of these are known to the US Census Bureau staff already dealing with MAF updates (e.g., tracking new housing units, converted housing units, and deleted housing units). One problem worth mentioning here involves seasonal populations and seasonal housing. In areas with substantial seasonal changes in population, great care must be taken to get an estimate of the de jure population. Since the implementation of the ACS, this problem will be compounded. This is because of differences between the ACS and the decennial census in regard to what constitutes the de jure population (CACPA/PAA, 2005). As such, an accurate housing register will need to deal with the seasonal housing issue and the differences in the definition of the de jure population found in the ACS and the decennial census.

A second issue regarding accuracy is accounting for the populations that don't have a standard permit address, such as the homeless or transient populations. It is true that these types of groups would be missed in any estimate using the MAF and separate methods and practices need to be developed to accurately estimate these populations. However, approximately 97% of the non group-quarters population of the United States lives in housing units. The development of a complete national address registry would insure an improvement in the accuracy of the population estimates for the vast majority of the population.

Judson, Popoff, and Batutis (2001) have pointed out that there is a great deal of evidence to support the idea that administrative records systems have systematic biases and they found support for this in an empirical study they conducted. This means that the MAF and, hence, the proposed housing register will be subject to systematic biases. Fortunately, however, Judson, Popoff, and Batutis (2001) also use their findings to make several recommendations regarding the reduction of these biases. Considering their research in conjunction with the experience being gained by US Census Bureau in regard to the MAF/TIGER system, we believe that the accuracy of a national housing register would be sufficient for purposes of resource allocation, research, and planning.

Another obstacle is the need to have a set of unified identification codes in order to match and merge records from different systems using electronic processing. As noted by Statistics Finland (2004), if there is no unified system of identification codes then it is extremely difficult and laborious, if not impossible, to link records across different systems. In particular, a unique code will be needed for every dwelling in the register, including those in multi-unit structures. In this regard, we point out that Finland has developed such a coding system and that it includes all structures – commercial, residential, and seasonal (Statistics Finland, 2004: 58-60).

With the exception of the issues of confidentiality and privacy, all of the challenges facing the development of a national housing register are in the form of costs, technical problems, or a combination of both. We agree with Wang (1999) that the major technical tasks of the National Accounting of Address and Housing Inventory come down to two areas - Address data collection and MAF/TIGER update. We also agree with Wang (1999) that a feasible way to effect a solution to these problems is to enhance the federal-state-local cooperative programs already part

of US Census Bureau activities such that local entities are compensated for helping to maintain the system. This is how Statistics Finland (2004) maintains its register system and there are data collection activities in the U. S. that already follow this model (Wang, 1999).

The national housing register we are proposing goes beyond what was envisioned by Wang. As such, we believe that his suggestions are necessary but not sufficient for this purpose. As is suggested here, there are many political, administrative, and technical obstacles that would need to be overcome. How exactly would researcher access be reconciled with confidentiality and privacy? What would the housing register cost to build and maintain and what savings elsewhere would be gained, if any? How would ACS data be combined with individual housing units – are they sufficient to provide the household level estimates that we are proposing (e.g., age, race, sex, household relationships, household size, vacancy rates, and socio-economic characteristics) or would that stretch imputation and related modeling techniques, as well as other capabilities too far?<sup>9</sup> If so, could the register function on, say, a block group level? If this were the case, then could the ACS provide block level results of the data? These are questions only further thought and empirical testing are likely to resolve. The question that the US Census Bureau needs to answer is if it appears our recommendation is sufficiently interesting to considering giving it the “thought” test before considering any small “empirical” test (e.g., similar to the Administrative Records Census Experiment reported by Judson and Bauder in 2002) before proceeding further. To give the US Census Bureau some food for thought, as it considers our question, we offer a quote from Ching-Li Wang’s (1999: 15) paper on developing the MAF into a resource for making post-censal population estimates:

“Is the development of the National Accounting of Addresses and Housing Inventory feasible? The ideas presented in the paper may cause many people to say that it is impossible because there are so many problems. This is exactly the same reaction we

saw in the late 80s when the Census Bureau was developing the TIGER to digitize the nation's geography from coast to coast. Now we can see how useful and powerful the TIGER is today.”

In closing, we would like to believe that if Ching-Li Wang were still alive, he would be willing to make a similar statement on behalf of the national housing register and we believe that many of his remarks and ours fit more or less the situation in other countries, which like the United States, while lacking a population registry, have otherwise strong administrative records system that can be used for population estimates, including housing data similar the US Census Bureau's MAF.

## Endnotes

1. The US Census Bureau document distributed at this conference uses the term “inter-censal estimate” in its title, while the document itself clearly makes reference to post-censal estimation work (U.S. Census Bureau, n.d.). we believe, however, that the distinction between inter-censal and post-censal is worth maintaining.
2. For the record, one can also construct estimates for a point in time that predates a census. we have not run across the term “pre-censal,” however and so do not use it here. Here it also is useful to note that there is a large body of literature on how to make estimates of populations and their characteristics for countries that lack censuses and good registration systems (Popoff and Judson, 2004). There are also methods developed for the estimation of wildlife populations that can be used with special populations such as the homeless – “capture-recapture” and “transit surveys,” for example (Williams, Nichols and Conroy, 2002). However, as is the case with the “statistical” tradition, we do not cover the estimation methods associated with “statistically underdeveloped areas” and wildlife populations
3. The MAF is already being used for “direct estimation” because it forms the sample frame for the Census Bureau’s “American Community Survey.”
4. The synthetic method of estimation is defined by Swanson and Stephan (2004: 776) as “a member of the family of ratio estimation methods used to estimate characteristics of a population in a sub-area (e. g., a county) by re-weighting ratios (e.g., prevalence rates or incidence rates) obtained from a survey or other data available at a higher level of geography (e.g., a state) that includes the sub-area in question.” As alluded to in the preceding definition, the synthetic method is usually viewed as belonging to the statistical tradition because of its frequent use with survey data. For a description of the synthetic method see Judson and Popoff (2004: 681-683). we also note that the “composite” method (Bryan, 2004b: 550-551) is a type of synthetic estimation.
5. Although their discussion of such adjustments is in the context of making projections rather than estimates, Smith, Tayman, and Swanson (2001: 239-277) provide a comprehensive description that covers many of the same issues found in developing estimates.

6. While the United States lacks a national population registration system there are, as noted in the body of the report, administrative records in the private sector that contain information on people that is used for commercial purposes (e.g., credit reporting systems such as those operated by Equifax, Experian, and TransUnion). Experian also conducts consumer marketing activities (See endnote # 9). These systems can be used to generate population estimates. However, using them requires money and the accuracy of such estimates is hard to judge because of the proprietary nature of the data.

7. Confidentiality is the idea that there should be restrictions on how information is collected and used and that no data should be disclosed about a respondent that would allow him or her to be either identified or harmed; privacy is the idea that it is the right of an individual to decide whether and to what extent he or she will divulge thoughts, opinions, feelings, and facts to the government (Mayer, 2002).

8. we note here that Statistics Finland (2004) has a measure of oversight over its data users while the US Census Bureau assumes no responsibility for what users do with its data. El-Badry and Swanson (2007) argue that the US Census Bureau's stance serves to decrease public trust in the Census Bureau. This is not a trivial issue because public trust has been identified as a major contributing factor to conflict over census results (El-Badry and Swanson 2007; Walashek and Swanson, 2006), an activity that requires the consumption of Bureau resources

9. In regard to the capabilities of imputation and modeling, we note that Swanson and Knight (1998) developed four model-based procedures for estimating household income using SIPP data statistically matched to Metromail's proprietary database. The procedures were developed with a random sample (n=6,559) from the data base and tested with the remaining "out of sample" portion of it (n= 7,048). The results were found to be sufficiently accurate and the procedures sufficiently tractable for use by the client. Given this personal experience, it is difficult for us to believe that the US Census Bureau is not technically capable of developing accurate and tractable procedures for purposes of developing the demographic and socio-economic information we propose for the national housing register. we also note here that subsequent to the project reported by Swanson and Knight (1998), Metromail was acquired by Experian, a subsidiary of GUS, which holds numerous databases containing public and proprietary information on consumers and also engages in direct mailing lists and other forms of marketing (The Motley Fool, 2000).

## References

- Alvey, W. and Scheuren, F. (1982). "Background for an Administrative Records Census." pp. 47-65 in *Statistics of Income and Related Administrative Record Research*. Washington, DC: U.S. Department of the Treasury, Internal Revenue Service.
- Anderson, M. (1988). *The American Census: A Social History*. New Haven, CT: Yale University Press.
- Brandon, P. and D. Hogan. (2004). "Impediments to Mothers Leaving Welfare: The Role of Maternal and Child Disability." *Population Research and Policy Review* 23 (4): 419-436.
- Breidt, F. J. (2005). "Controlling the American Community Survey to Intercensal Population Estimates." Paper presented at the Annual Meeting of the Southern Demographic Association, Oxford, MS, November 3<sup>rd</sup> to 5<sup>th</sup>.
- Bryan, T. (2004a). "Basic Sources of Statistics." pp. 9-41 in J. Siegel and D. Swanson (Eds.) *The Methods and Materials of Demography, 2<sup>nd</sup> Edition*. New York, NY: Elsevier Academic Press.
- Bryan, T. (2004b). "Population Estimates." pp. 523-560 in J. Siegel and D. Swanson (Eds.) *The Methods and Materials of Demography, 2<sup>nd</sup> Edition*. New York, NY: Elsevier Academic Press.
- Bryan, T. and R. Heuser. (2004). "Collection and Processing of Demographic Data." pp. 43-63 in J. Siegel and D. Swanson (Eds.) *The Methods and Materials of Demography, 2<sup>nd</sup> Edition*. New York, NY: Elsevier Academic Press.
- Burch, T. (1993). "Theory, Computers, and the Parameterization of Demographic Behaviour." *IUSSP International Conference, Montreal*. Liège, Belgium: IUSSP 3: 377-388.
- CACPA/PAA. (2005). "Recommendation 10c" in *Recommendations from the Population Association of America Advisory Committee*. Meeting of the Census Advisory Committee of Professional Associations. Arlington, VA: October 21<sup>st</sup> - 22<sup>nd</sup>.
- Clark, W. A. V. (1986). *Human Migration*. Volume 7, Scientific Geography Series. Beverly Hills, CA: Sage Publications
- Clemetson, L. (2004a). "Homeland Security Given Data on Arab-Americans: Census Bureau Complies with Request." *New York Times*, 30 July. Late Edition – Final: A14
- Clemetson, L. (2004b). "Coalition Seeks Action on Shared Data on Arab-Americans." *New York Times*, 13 August: Late Edition –Final: A12.

Clemetson, L. (2004c). "Census Policy on Providing Sensitive Data is Revised." *New York Times*, 31 August: Late Edition – Final: A14

Coale, A. and P. Demeny. (1966). *Regional Model Life Tables and Stable Populations*. Princeton NJ: Princeton University Press.

Cook, T. (1996). "When ERPs Aren't Enough: A Discussion of Issues Associated with Service Population Estimation." *Working Paper 96/4*. Demography Section, Australian Bureau of Statistics, Belconnen, ACT, Australia.

Dharmalingam, A. (2004). "Reproductivity." pp. 429-453 in J. Siegel and D. Swanson (Eds.) *The Methods and Materials of Demography, 2<sup>nd</sup> Edition*. New York, NY: Elsevier Academic Press.

El-Badry, S. and D. Swanson. (2007). "Providing Special Census Tabulations to Government Security Agencies in the United States: The Case of Arab-Americans." *Government Information Quarterly* 24(2): 470-487.

Freedman, D., (2004). "The Ecological Fallacy." p. 293 in M. Lewis-Beck, A. Bryman, and T. Liao (Eds.) *The Encyclopedia of Social Science Research Methods*. Beverly Hills, CA: Sage Publications.

Habermann, H. (2006). "Research to Improve Population Estimates" Part of a presentation by H. Habermann at the Spring (May 18<sup>th</sup> -19<sup>th</sup>) 2006 meeting of the Census Advisory Committee for Professional Associations.

Judson, D. (2000). "The Statistical Administrative Records System and Administrative Records Experiment 2000." Paper presented at the National Institutes of Statistical Sciences Data Quality Workshop, Morristown, NJ, November 30<sup>th</sup> – December 1<sup>st</sup>.

Judson, D. H. and M. Bauder. (2002). "Evaluating the Ability of Administrative Records Databases to Replicate Census 2000 Results at the Household Level." Paper presented at the Annual Meeting of the American Statistical Association, New York, NY, August 11<sup>th</sup> - 15<sup>th</sup>.

Judson, D. and C. Popoff. (2004). "Selected General Methods." pp. 677-732 in J. Siegel and D. Swanson (Eds.) *The Methods and Materials of Demography, 2<sup>nd</sup> Edition*. New York, NY: Elsevier Academic Press.

King, G., O. Rosen, and M. Tanner (Eds.) (2004). *Ecological Inference*. Cambridge, England: Cambridge University Press.

Kintner, H. and D. Swanson. (1994). "Estimating Vital Rates from Corporate Databases: How Long will GM's Salaried Retirees Live?" pp. 265 – 295 in H. Kintner, T. Merrick, P. Morrison, and P. Voss (Eds.) *Demographics: A Casebook for Business and Government*. Boulder, CO: Westview Press.

Kliss, B. and W. Alvey (Eds). (1984). *Statistical Uses of Administrative Records: Recent Research and Present Prospects, Volumes I and II*. Washington, D.C.: Department of the Treasury, Internal Revenue Division, Statistics of Income Division.

Kordos, J. (editor). (2000). "Special issue on Small Area Estimation." *Statistics in Transition: Journal of the Polish Statistical Association* 4 (4).

Li, N. and S. Tuljapurkar. (2005). "A Formal Model of Age-Structural Transitions." pp. 91 – 105 in S. Tuljapurkar, I. Pool and V. Prachuabmoh (Eds.) *Population Resources and Development: Riding the Age Waves- Volume I*. Dordrecht, The Netherlands: Springer.

Lipton, E. (2004). "Panel Says Census Move on Arab-Americans Recalls World War II Internments." *New York Times*, 10 November.

Livingston, G. (2006). "Gender, Job Searching, and Employment Outcomes Among Mexican Immigrants." *Population Research and Policy Review* 25 (1): 43-66.

Long, J. (1993). *Postcensal Population Estimates: States, Counties and Places*. Technical Working Paper No. 3. Washington DC: U.S. Bureau of the Census.

Lowenthal, T. (2006). "House Cuts \$58.3M from Census Budget; Senate Panel Approved \$50M less than Bush Request." *Census News Brief*, 11 July.

Mayer, T. (2002). *Privacy and Confidentiality Research and the U. S. Census Bureau: Recommendations Based on a Review of the Literature. Research Report Series*. Survey Methodology Report #2002-01. Statistical Research Division, U.S. Census Bureau. Washington, D.C. US Census Bureau (<http://www.census.gov/srd/www/byyear.html>, Last accessed March 2005).

McKibben, J. (2006). "School District Planning and the 2010 Census: Data Uses and Needs". *Journal of Economic and Social Measurement*, 31, (3) 221-232

McNicoll, G. (1992). "The Agenda of Population Studies: A Commentary and Complaint." *Population and Development Review* 18: 399-420.

Murdock, S. and D. Ellis. (1991). *Applied Demography: An Introduction to Basic Concepts, Methods, and Data*. Boulder, CO: Westview Press.

Mutchler, J. and L. Baker. (2004). "A Demographic Examination of Grandparent Caregivers in the Census 2000 Supplemental Survey." *Population Research and Policy Review* 23 (4): 359-377.

National Research Council. (2005). *Population, Land Use, and Environment: Research Directions*. Washington, D.C.: National Academies Press.

National Research Council. (2006). *Expanding Access to Research Data: Reconciling Risks and Opportunities*. Washington, D.C: National Academies Press.

Pittenger, D. (1982). "Critique of Administrative Record Procedures." pp. 39-42 in E. S. Lee and H. F. Goldsmith (Eds.), *Population Estimates: Methods for Small Area Analysis*. Beverly Hill, CA: Sage Press.

Platek, R., J. Rao, C. Sarndal, and M. Singh (Eds.) (1987). *Small Area Statistics: An International Symposium*. New York, NY: John Wiley.

Pol, L. and R. Thomas. (2001). *The Demography of Health and Health Care, 2<sup>nd</sup> Edition*. New York, NY: Kluwer Academic/Plenum Press.

Pollard, J. (1973). *Mathematical Models for the Growth of Human Populations*. Cambridge, England: Cambridge University Press.

Popoff, C. and D. Judson. (2004). "Some Methods of Estimation for Statistically Underdeveloped Areas." pp. 603-641 in J. Siegel and D. Swanson (Eds.) *The Methods and Materials of Demography, 2<sup>nd</sup> Edition*. New York, NY: Elsevier Academic Press

Prevost, R. (1996). "Administrative Records and the New Statistical Era." Paper presented at the 1996 Annual meeting of the Population Association of America. New Orleans, LA, May 9<sup>th</sup> -11<sup>th</sup>.

Prevost, R. and D. Swanson. (1985). "A New Technique for Assessing Error in Ratio-Correlation Estimates of Population: A Preliminary Note." *Applied Demography* 1 (November): 1-4.

Rao, J. (2002). *Small Area Estimation*. San Francisco, CA: Jossey-Bass.

Redfern, P. (1986). "Which Countries Will Follow the Scandinavian Lead in Taking a Register-based Census of Population?" *Journal of Official Statistics* 2 (4): 415-424.

Roe, L, J. Carlson, and D. Swanson. (1992). "A Variation of the Housing Unit Method for Estimating the Population of Small, Rural Areas: A Case Study of the Local Expert Method." *Survey Methodology* 18(1): 155-163.

Rogers, A. (1995). *Introduction to Multiregional Mathematical Demography*. New York, NY: John Wiley & Sons.

Rogers, R., R. Hummer, and C. Nam. (2000). *Living and Dying in the USA: Behavioral, Health, and Social Differentials of Adult Mortality*. New York, NY: Academic Press.

Ryan, S., J. Manlove, and S. Hofferth. (2006). "State-level Welfare Policies and Nonmarital Subsequent Childbearing." *Population Research and Policy Review* 25 (1): 103-126.

Scheuren, F. (1999). "Administrative Records and Census Taking." *Survey Methodology* 25(2): 151–160.

Serow, W. and N. Rives. (1995). "Small Area Analysis: Assessing the State of the Art." pp. 1-9 in N. Rives, W. Serow, A. Lee, H. Goldsmith, and P. Voss (eds.). *Basic Methods for Preparing Small Area Population Estimates*. Madison, WI: Applied Population Laboratory, Department of Rural Sociology, University of Wisconsin.

Siegel, J. (2002). *Applied Demography: Applications to Business, Government, Law, and Public Policy*. San Diego, CA: Academic Press.

Smith, S. (1994). "Estimating Temporary Populations: The Contributions of Robert C. Schmitt." *Applied Demography* 9 (1): 4-7.

Smith, S., and S. Cody. (2004). "An Evaluation of Population Estimates in Florida: April 1, 2000." *Population Research and Policy Review* 23 (1): 1-24.

Smith, S., J. Tayman, and D. Swanson. (2001). *State and Local Population Projections: Methodology and Analysis*. New York, NY: Kluwer Academic/Plenum Publishers.

Statistics Finland. (2004). *Use of Register and Administrative Data Sources for Statistical Purposes: Best Practices of Statistics Finland*. Handbook Series, No. 45. Helsinki, Finland: Statistics Finland.

Stockwell, E., F. Goza, and K. Balistreri. (2005). "Infant Mortality and Socioeconomic Status: New Bottle, Same Old Wine." *Population Research and Policy Review* 24 (4): 387-399.

Suchindran, C. (2004). "Part II: Model Life Tables." pp. 662-675 in J. Siegel and D. Swanson (Eds.) *The Methods and Materials of Demography, 2<sup>nd</sup> Edition*. New York, NY: Elsevier Academic Press.

Swanson, D. (2004). "Advancing Methodological Knowledge within State and Local Demography: A Case Study." *Population Research and Policy Review* 23 (4): 379-398.

Swanson, D. and M. Knight. (1998). *Metromail Wealth Estimation Project Final Report: Recommendations, Summary Findings, and Technical Documentation*. Madison, WI: Third Wave Research Group (a proprietary report).

Swanson, D. and L. Pol. (2005). "Contemporary Developments in Applied Demography within the United States." *Journal of Applied Sociology* 21 (2): 26-56.

Swanson, D. and G. E. Stephan. (2004). "Glossary." pp. 751-778 in J. Siegel and D. Swanson (Eds.) *The Methods and Materials of Demography, 2<sup>nd</sup> Edition*. New York, NY: Elsevier Academic Press.

Swanson, D., T. Burch, and L. Tedrow. (1996). "What is Applied Demography?" *Population Research and Policy Review* 15 (5-6): 403-418.

Swanson, D., B. Baker, and J. Van Patten. (1983). "Municipal Population Estimation: Practical and Conceptual Features of the Housing Unit Method." Paper Presented at the 1983 Annual Meeting of the Population Association of America, Pittsburgh, PA, April 14<sup>th</sup> - 16<sup>th</sup>.

The Motley Fool. (2000). "Extracting Value from Experian" (By Maynard Paton). 16 March 2000 (online at [www.fool.co.uk/aualiport/2000/qualiport000315.htm](http://www.fool.co.uk/aualiport/2000/qualiport000315.htm), last accessed 12 July 2006).

Treyz, G., D. Rickman, G. Hunt, and M. Greenwood. (1993). "The Dynamics of U.S. Internal Migration." *Review of Economics and Statistics* 75: 209-214.

U. S. Census Bureau. (No Date). "The U.S. Census Bureau's Intercensal Population Estimates and Projections Program: Basic Underlying Principles." Unpublished document provided to participants of this conference (reproduced as Appendix A in this paper).

U.S. GAO. (2003). *2000 Census: Coverage Measurement Programs' Results, Costs, and Lessons Learned*. GAO -03-287. Washington, DC: U. S. General Accounting Office (Note: Effective July 7, 2004, the GAO's legal name became the Government Accountability Office).

U.S. GAO. (2006). *2010 Census: Census Bureau Needs to Take Prompt Actions to Resolve Long-standing and Emerging Address and Mapping Challenges*. GAO-06-272. Washington, D.C. U. S. Government Accountability Office (Note: Effective July 7, 2004, the GAO's legal name became the Government Accountability Office).

Walashek, P. and D. Swanson. (2006). "The Roots of Conflict over US Census Counts in the late 20<sup>th</sup> Century and prospects for the 21<sup>st</sup> Century." *Census Counts in the late 20<sup>th</sup> Century.* *Journal of Economic and Social Measurement*. 31(4): 185-206.

Waldrop, J. (1995). "Preface." pp. v-vi in N. Rives, W. Serow, A. Lee, H. Goldsmith, and P. Voss (eds.) *Basic Methods for Preparing Small-Area Population Estimates*. Madison, WI: Applied Population Laboratory, Department of Rural Sociology, University of Wisconsin.

Wang, C. (1999). "Development of National Accounting of Address and Housing Inventory: The Baseline Information for Post-censal Population estimates." Paper prepared for the Estimates Methods Conference, U.S. Bureau of the Census, Federal Office Building #3, Suitland, Maryland, June 8th. (available online at <http://www.census.gov/population/www/coop/popconf/paper.html>, last accessed 28 June 2006).

Wenjert, J. (2003). "Utah v. Evans and Statistical Methodologies in Census Apportionment Calculations." *Jurimetrics* 43: 441-453.

Williams, B., J. Nichols, and M. Conroy. (2002). *Analysis and Management of Wildlife Populations*. San Diego, CA. Academic Press.

Wilmoth, J. (2004). "Population Size." pp. 65-80 in J. Siegel and D. Swanson (Eds.) *The Methods and Materials of Demography, 2<sup>nd</sup> Edition*. New York, NY: Elsevier Academic Press.

## APPENDIX A

### The U.S. Census Bureau's Intercensal Population Estimates and Projections Program Basic Underlying Principles

#### I. Background

The U.S. Census Bureau's Population Estimates and Projections program is designed to fulfill the mandates of Title 13, Section 181, of the U.S. Code.

During the intervals between each census of population required under section 141 of this title, the Secretary, to the extent feasible, shall annually produce and publish for each State, county, and local unit of general purpose government which has a population of fifty thousand or more, current data on total population and population characteristics and, to the extent feasible, shall biennially produce and publish for other local units of general purpose government current data on total population. Such data shall be produced and published for each State, county, and other local unit of general purpose government for which data is compiled in the most recent census of population taken under section 141 of this title. Such data may be produced by means of sampling or other methods, which the Secretary determines will produce current, comprehensive, and reliable data.

- A. To satisfy this mandate, the program of population estimates has grown over the years to produce the following products annually:
1. Monthly estimates of the national population of the United States by age, sex, race, and Hispanic origin
  2. Annual estimates of the population of states by age, sex, race, and Hispanic origin
  3. Annual estimates of the population of counties by age, sex, race, and Hispanic origin
  4. Annual estimates of the total population of functioning governmental units
  5. Annual estimates of the number of housing units for states and counties.
- B. In addition to meeting the mandates of Title 13, these estimate products are used for a variety of purposes, including the following:
1. Controls for federally sponsored surveys, including the Current Population Survey (CPS) and the American Community Survey (ACS)
  2. Allocation of federal dollars totaling over \$200 billion annually
  3. Denominators for various indicators, including vital statistics, per capita income, and cancer incidence rates
  4. Calculation of the number of clerks the Senate hires
  5. Requirements of the Federal Election Commission
  6. Denominators for poverty rate estimation at selected levels of geography
  7. Program planning by federal, state, local, and private entities

## II. **Implicit Assumptions**

Implementation of the annual program of intercensal estimates is guided by several implicit assumptions.

### A. Timely release of the annual products is critical

1. The maximum lag time between estimate date and dissemination of last data product is 12 months.
2. Annual national and state population totals must be released within 6 months of estimate date to meet requirements of IRS Bonding Authority.
3. State estimates of the population aged 18 and older must be available within 6 months of estimate date to satisfy requirements of the Federal Election Commission.
4. National and state population controls to be used for the new calendar year CPS must be available by late January of the new calendar year.
5. Estimates of state and county characteristics must be available within 9 months to meet requirements for use as population controls for the American Community Survey.
6. Estimates of functioning governmental units should be available within 12 months of estimate date for use by HUD in funds allocation.

### B. Each annual production consists of a time series of estimates from the last decennial census date to the estimate date and is produced using the latest available data and the current approved methodology.

1. Current-year data products contain revisions to the prior year's estimates that are caused by incorporating:
  - a. Improved methodology.
  - b. New data inputs.
  - c. Revisions to prior year data inputs.
2. The term "vintage" is used to refer to the reference date of an estimates cycle. Estimates released with a reference date of July 2005 are referred to as the "vintage 2005" set of population estimates and will include a consistent time series back to April 2000.

### C. Within any vintage, all products use the same vintage of input data and must sum to the earlier released products of the same vintage for the same measurement.

1. Since the national and state population totals are the first to be released, all subsequent estimate products must sum to the national and state totals that already appear for that vintage. This insures consistency within any

vintage and means that the sum of the “parts” will always equal the previously released U.S., state, or county total.

2. Since the national population estimates tabulated by characteristics are the first characteristics to be released, the sum of the state and county characteristics must equal the national characteristics of the same vintage.
- D. Only one consistent set of products and related materials is developed within a vintage. That set of products is intended to serve all customers’ needs and uses.
1. The methodology and data inputs used to develop the population estimates used as denominators for vital statistics rates are consistent with those used to develop the population controls for the CPS and ACS.
  2. Custom data products are consistent with the publicly released data products. For example, the annual race estimates for counties use a bridged race algorithm developed by NCHS. However, while the race data conform to the bridging algorithms developed by NCHS, the estimates of total populations and populations by age and sex generally agree with the publicly released data products.
- E. The population estimates begin with the most recent decennial-census enumerated count updated to July 1 of each year, and as such, are based on the usual-residence concept used in the most recent decennial census.
1. The population estimates base for each estimate date is updated to include Count Question Resolution (CQR) changes to the decennial census base as well as geographic updates due to annexation and other geographic program changes.
  2. The components of population change used to update the most recent census will be consistent with the best set of components available. Ongoing evaluation indicates that the coverage and the consistency of vital statistics and other administrative records data differ from those of decennial census data. Therefore, in the annual estimates, the size of the population based mainly on administrative records data differ from the size based mainly on census data.
- F. States, counties, and units of local government have the right to challenge the population estimates prepared by the Census Bureau under the provisions of Title 15, The Code of Federal Regulations, Part 90. The results of accepted challenges will be incorporated into the following year’s population

estimates as long as the challenge is received by October 1 of the year in which the estimate was released.

### III. Current Broad Methodological Assumptions

- A. Prior to incorporating a new methodology or data set, it is desirable to thoroughly evaluate a set of estimates that use this new methodology or data set and compare it with the most recent decennial census results. When this is not possible, the methods are judged by the following criteria.
1. Soundness: The method should be based on solid reasoning – i.e., the formulas that embody the method should be mathematically valid and respect the attributes of the input data as they relate to the estimation task.
  2. Integrity: A strategy that consistently applies the declared method is preferred to one that uses ad-hoc fixes to address particular challenges of the estimation task.
  3. Parsimony: A simpler strategy is preferred to a more complex one.
  4. Robustness: The method that produces the most reasonable estimates (defined below) across the full range of potential input-data values and in the presence of the random variation normally associated with those values while maintaining the orthodoxy and consistency of the estimates (also defined below) is preferred.
  5. Adaptability: A technique that can be applied more broadly (e.g., across geographic summary levels), thus promoting the integration of the Census Bureau's estimates system, is preferred to a more product-specific remedy.
  6. Transparency: A strategy that is more readily understandable and replicable by external parties is preferred. Moreover, a strategy that provides some explanatory information (i.e., how did the size or distribution of the population come to be this way) is preferred over one that is merely predictive.
  7. Usability: The method must be executable along with all other current projects under current staffing levels in a way that allows the Census Bureau to meet current deadlines.
  8. Flexibility: The preferred method will allow the production of estimates when a specific instance of the input data normally required by the method is unavailable or deemed unsuitable.

- B. As a final test, the method should produce output data that have the following qualities.
1. Orthodoxy: The values of the population estimates should be appropriate (e.g., no negative population numbers, all population estimates in whole numbers).
  2. Consistency: The values of the population estimates for all universes (e.g., resident, civilian, civilian non-institutionalized), geographies (e.g., national, state, county), and characteristics (e.g., age, sex, race, Hispanic origin) should not contradict one another.
  3. Reasonableness: The values of the population estimates should approximate the real values as determined by the following assessments.
    - a. Post-Censal Change: The reasonableness of the total change in the population since the last decennial census.
    - b. Time-Series Change: The reasonableness of the annual change in the estimates since the last census.
    - c. Demographic Appropriateness: The values of the estimates and the demographic rates they imply fall within acceptable limits when evaluated by general demographic principles (e.g., the appropriateness of the sex ratios, age progression, implied family size, life expectancies, total fertility rates, etc.).
    - d. Comparability: The estimates appear realistic when compared with other indicators of the size and distribution of the population (e.g., Medicare enrollment, school enrollment, housing unit estimates, etc.).
- C. A consistent method is used for entities at the same level of geographic aggregation.
1. The method adopted for state totals must be used for all states.
  2. The method adopted for counties within a state must be used for all counties within that state.
- D. The Census Bureau develops the basic estimates for the nation, states, and counties by disaggregated race groups in order to meet the various custom race aggregations needed by users.

- E. The cohort-component method is the preferred method for development of the national, state, and county-level total population estimates and population estimates by characteristics.
- F. The distributive housing-unit method is the preferred method for the development of the functioning subcounty governmental-unit-level estimates.
- G. State total population estimates are not developed independently. National population estimates are first developed; then county total population estimates are developed and controlled to the national total population estimates. The state total population estimates are the sum of the “nationally controlled” county total population estimates for the state.
- H. Data on vital statistics and group quarters provided by members of the Federal State Cooperative Program for Population Estimates (FSCPE) are included in the process of developing state and county population estimates.
- I. Although state members of the FSCPE are provided the opportunity to review the state and county population totals prior to final production, they must follow strict criteria and provide objective evidence when requesting modifications.

#### IV. **Current Specified Methodologies**

- A. National level estimates will use the cohort-component technique applied to data from the latest decennial census as the base, data on births and deaths provided by the National Center for Health Statistics, and estimates of net international migration derived from data from the American Community Survey (ACS) See the url

<[http://www.census.gov/popest/topics/methodology/2003\\_nat\\_char\\_meth.htm](http://www.census.gov/popest/topics/methodology/2003_nat_char_meth.htm)  
[http://www.census.gov/popest/topics/methodology/v2005\\_nat\\_char\\_meth.html](http://www.census.gov/popest/topics/methodology/v2005_nat_char_meth.html)>

For a detailed discussion of the methodology used to develop the most recent set of national population estimates by demographic characteristics.

- B. State and county population estimates are developed using a demographic procedure called an "administrative records component of population change" method. A major assumption underlying this approach is that the components of population change are closely tracked by administrative data in a demographic change model. In order to apply the model, Census Bureau demographers estimate each component of population change separately. For the population residing in households, the components of population change are births, deaths, and net migration, including net international migration. For the non-household population, change is represented by the net change in the population living in group-quarters facilities.

Each component in our model represents data that are symptomatic of an aspect of population change. For example, birth certificates indicate additions to the population resulting from births, so we use these data to estimate the birth component for a county. Other components are derived from death certificates, Internal Revenue Service data (IRS), Medicare enrollment records, Armed Forces data, group-quarters population data, and data from the American Community Survey.

For a more detailed discussion of the development of county population totals see

[http://www.census.gov/popest/topics/methodology/2003\\_st\\_co\\_meth.html](http://www.census.gov/popest/topics/methodology/2003_st_co_meth.html)

[http://www.census.gov/popest/topics/methodology/2005\\_st\\_co\\_meth.html](http://www.census.gov/popest/topics/methodology/2005_st_co_meth.html)

- C. State population characteristics are currently developed in a two-stage process. Estimates by age and sex are developed first using a cohort-component procedure whereby estimates of net migration are developed using school enrollment data. These estimates are controlled both to the national-level estimates by age and sex as well as the previously developed state population totals.

---

The second step in the process distributes the state age and sex estimates into race by Hispanic origin categories. This is done by preparing an initial set of state estimates by age, sex, race, and Hispanic origin that are controlled to the state age and sex estimates prepared in the first step and to the previously developed national estimates by age, sex, race, and Hispanic origin.

For a more detailed discussion of the development of the state population characteristics by age, sex, race, and Hispanic origin see

[http://www.census.gov/popest/topics/methodology/2003\\_st\\_char\\_meth.html](http://www.census.gov/popest/topics/methodology/2003_st_char_meth.html)

[http://www.census.gov/popest/topics/methodology/2004\\_st\\_char\\_meth.html](http://www.census.gov/popest/topics/methodology/2004_st_char_meth.html)

- D. County population characteristics are developed using a proportional distribution method beginning with previously developed resident county population estimates by age (0-64 and 65+) and resident state population estimates by age, sex, race, and Hispanic origin. Then county-level estimates of age, sex, race, and Hispanic origin distributions are developed using information about post-censal change in the corresponding populations. Third, these distributions are applied to the original county estimates by age and state characteristics.

A detailed discussion of this method is provided at

<[http://www.census.gov/popest/topics/methodology/2004\\_co\\_char\\_meth.htm](http://www.census.gov/popest/topics/methodology/2004_co_char_meth.htm)>

## **V. Enhancement Priorities**

### **A. Improve estimates of net international migration**

1. Provide up-to-date, useful statistics and methodologies on the size, characteristics, and demographic impact of international migration to and from the United States for use in policy-making decisions and demographic and economic research.
2. Goals of immigration research
  - a. Produce annual estimates of international migration
  - b. Improve current migration-related survey questions on the ACS.
  - c. Conduct extensive evaluations to determine the best method to incorporate ACS data into the population estimates.
3. Activities
  - a. Evaluate reasonableness of estimates of annual change in the foreign-born data from ACS at the national level.
  - b. Produce revised estimates of net international migration at the national level.
  - c. Produce new demographic and geographic distributions for migrants.
  - d. Construct algorithms to estimate the migrant status of the foreign-born populations.
  - e. Produce estimates of international migrants by migrant status (legal migrants, temporary migrants, quasi-legal migrants, unauthorized migrants, and emigrants).

### **E. Improve Estimates of Internal Migration**

1. Improve the accuracy of the annual migration estimates by age, sex, race, and Hispanic origin for counties by maximizing the efficient use of available administrative data files, Census 2000 data, and the American Community Survey (ACS) data.
2. The ultimate goal is to implement a person-based migration model incorporating administrative data from files such as the IRS 1040 and 1099 records, Medicare records, a derived person-characteristic file developed from the Social Security Administrative NUMIDENT file, and

other administrative data that can be merged into the database. The database will enable analysts to match administrative data with Census 2000 (100% and sample data), CPS, and ACS data in order to develop models that correct possible demographic and geographic biases inherent in the use of an administrative records database when estimating migration rates for counties.

F. Develop a new methodology for estimating subnational population characteristics

1. Replace the methodology that develops state estimates by age and sex based on school enrollment data with a method that is consistent with the best set of administrative data available and exploits the power of current computing capacity.
2. Develop a method that addresses current deficiencies in the age distributions of the population in selected states and counties, especially the age distribution of the population aged 18 to 24.
3. Develop a new method to estimate county population by age, sex, race, and Hispanic origin.

G. Develop procedures to systematically incorporate participation by State FSCPE Agencies in the production of state and county population estimates

1. Address issues of consistency
2. Establish criteria for incorporating state participation

**VI. Other Enhancements**

- A. Improve the distributive housing unit approach at the subcounty level.
  - 1. Develop procedures to update Census 2000 measures of vacancy and numbers of people per household (PPH or the Person Per Household measure) used in the estimates process.
  - 2. Improve estimates of housing units.
  - 3. Address inconsistencies between estimates developed using the distributive housing unit approach and those developed using the component approach.
    - a. Develop improved procedures to estimate housing unit loss.
    - b. Integrate enhancements from the Master Address File.
- B. Address inconsistencies between data from the decennial census base and data on components of change from administrative records databases.
  - 1. Address inconsistencies between Census 2000 data and NCHS data on race and Hispanic-origin characteristics.
  - 2. Address unreasonable results from pairing NCHS mortality data with decennial census data and estimate results.

## **VII. Administrative Constraints**

- A. The methods developed must be capable of being implemented with current resources and within the current time frame for estimate production.
- B. Production of the complete set of estimates must continue during any development stages.
- C. Methods must be Transparent and Reproducible