DEMOGRAPHIC APPLICATIONS OF ADMINISTRATIVE RECORDS

John F. Long, Bureau of the Census
Room 2019, Federal Building 3, Washington, DC 20233

Key Words: Administrative Records
Demographic Methods

Introduction

Modern societies require current data on population size, characteristics, and geographic distribution. The more recent these data, the better governments, businesses, and individuals can administer their affairs and plan for the future. For many needs, annual data are required. These data may be used to allocate national funds to states or to allocate state funds to lower levels of government. They are also used to administer and plan government and business programs. The data provide controls for national and subnational surveys and serve as denominators for many official data series such as mortality rates and per capita income.

Annual data series for populations and their characteristics require procedures other than population censuses since the expense of censuses prohibits their use more often than every five or ten years. In a few nations (principally in Europe), the needs for current population data are fulfilled by reliable and comprehensive population registers. But for most nations (including the United States, Canada, and Australia), current annual time series of population data are collected using household surveys or estimated using administrative records.

Current household surveys are only precise enough to provide estimates for the nation, regions, and the largest states. For smaller areas of geography, the production of annual data requires adapting administrative data sources designed for non-demographic purposes to provide estimates of population levels, characteristics, or demographic events (such as internal migration). For decades, these administrative records have been the basic resource for independent population estimates. More recently, they have been suggested to supplement the decennial census or to combine with survey techniques to produce small domain estimates.

It may be easier to see the range of current applications for administrative records by organizing these applications by size of geographic unit and method of use. Figure 1 shows some of the current uses of administrative records at the Census Bureau. The 2000 census is proposing to use administrative records directly to assist in counting individuals and households. At the macro-level (blocks and above), the Census Bureau has a number of projects that use administrative records directly or indirectly. The indirect (or modeled) methods use the information from administrative records as controls to surveys(continuous measurement), as methods to determine the rate of demographic components (intercensal demographic estimates), or in an integrated model of survey and administrative data (small area income and poverty estimates).

A Microlevel Model for Administrative Records

Even though administrative records have a long history of practical daily use in demographic techniques, they have never really been integrated into the framework of formal demographic accounting. Perhaps this is because their micro-level and operational nature do not appear to fit well within the basic demographic paradigms of stable population theory or multistate demographic analysis. This presentation will try to overcome this apparent misfit by beginning with a simple model of administrative records that can serve both as a paradigm for current uses of administrative records and as a conceptual framework for comparison with more traditional demographic accounting models.

Let's begin at the micro-level by assuming that we have four data sets composed of individuals, housing units, business establishments, and geographic entities. Each

unit of the four data sets would have its own characteristics. For example, the individuals might have social and demographic characteristics such as age, sex, race, and educational attainment. The housing units could have characteristics relating to housing quality and to vacancy status. The business establishments would have such characteristics as industrial category and gross sales volume. Finally, the geographic units would have characteristics such as region of the country, incorporation status, land area, etc. Notice that each of these characteristics are intrinsic to a given data base and do not require linkage with the other three data sets. Many administrative records data sets have precisely this self contained character.

But the essence of much demographic(as well as economic) accounting is to make connections between these four categories of variable. Thus, in Figure 2 we extend the model to encompass direct linkages between some of the variables and indirect linkages with others. We can link an individual to a housing unit by means of a residence. Continuing counterclockwise, the housing unit can be linked to a geographic unit (county, census tract, etc) by means of the geographic location of the housing unit. In fact, we are so used to thinking of these concepts in the context of a census or household survey that it may be difficult for us to conceive of unlinked variables. In the decennial census process, data on individuals, housing, and geographic entities are collected and linked directly as an integral part of the census process. However, many administrative data sets do not have that linkage. Social Security records, for example, do not have a current address for most non-retired individuals. Instead these records of earnings history need only be kept by social security number to meet the accounting requirements of the administrative system.

A similar situation exists for economic variables on the right side of figure. This time working clockwise from the individual data base, an individual can be linked to a business establishment by means of his employment. For example, the former Continuous Work History Sample of the Social Security Administration and the Bureau of Economic Analysis was based on administrative records that linked an employee's record with the employer's establishment. Continuing clockwise, the establishments in turn are linked to a geographic unit by their physical location which might be obtained by geographic coding of their street address.

In general, linkages between the four variables using administrative data often require rather elaborate devices. The results of such linkages are very dependent on the operational steps of the linkage process. In practice, there are often substantial differences between the operations used to link administrative records with the definitions and operations used in a census. For example, the concept of "usual residence" used in the decennial population and housing censuses may differ substantially from the address (if any) listed on an administrative records system. Yet, linking administrative data on individuals to a housing unit address and then linking that address to a geographic entity is in many ways the key to population estimation methodology. Any system for making subnational population estimates must have a credible method for developing such geographic correspondence. Data are required for legally defined geographic entities such as counties and incorporated places and the estimates methodology must take these requirements into account. The first step is to place each individual in a corresponding housing unit. To do this one must find a data set that associates each individual with a given housing unit address. As an example, to use information on demographic characteristics (age, race, and sex) contained in the social security administration records for all persons with a social security number requires first matching the record with another administrative database (Such as tax return data) that has both social security number and a current filing address.

A further step is needed to link the housing unit address with an exact geographic entity. For this, one could use a master address file linked by TIGER to the current geographic entity file. A system like this is currently under development at the Census Bureau and could eventually entail an annually updated digitized data base that could place most addresses in the United States into the appropriate census block (and thus into places and counties that also have their boundaries in the TIGER system).

Micro-level Methods

Although the operational difficulties are substantial, the micro-level administrative records system illustrated in Figure 1 is conceptually rather simple. It consists of four multiple relational data bases that could link any

individual directly with a corresponding housing unit and establishment, and indirectly place the individual in one geographic unit as place of residence and another as place of work. For a number of proposed applications of administrative records, manipulation of multiple databases using this micro-level formulation may be sufficient. There are a couple of directions such micro-level methods might take -- direct use of individual administrative record data or indirect estimates using models created from individual data.

Direct Methods - Direct methods might consist of keeping detailed individual records that could be tabulated in any combination to provide statistical data on population, housing, businesses, etc. Although the output data would always be presented in aggregate form to assure confidentiality, the actual database management would be maintained at the individual level. While conceptually simple this model would be complex operationally given the size of the data sets and the large number of possible combinations of relationships. Also the potential for loss of confidentiality and privacy with such an individual level multipurpose database weigh heavily against taking such an approach.

Another approach using micro-level administrative data directly is proposed in the current plans of the "reengineered" 2000 census. Under this proposal, administrative data (after editing and checking for consistency) would be used as a substitute for census data when a respondent could not be enumerated by traditional census procedures. This approach is currently being tested in the test census sites.

Modeled Methods - A second set of estimates would use individual data from administrative records in a more indirect way. These methods would use the individual administrative record data to create models of the relationship between certain characteristics of individuals and then apply the results of these models to other data sets where some of the individuals' characteristics are known but others must be estimated. A key example is to be found in current research on producing annual data on income distribution for counties or smaller areas. Here data from administrative records and survey samples would be combined to produce modeled estimates of poverty for small areas. A number of federal programs have used such methods of indirect estimation and the results are currently available in Wesley Schaible(ed.), Indirect Estimators in U. S. Federal Programs, Springer, 1996.

A Macro Level Approach

However, for many applications using administrative records and for comparison to traditional methods of demographic accounting we will have to shift to the macro level. For most demographic methods, we aggregate individual information by geographic areas to produce population values. Thus for any given geographic level (such as counties), we can total the number of housing units in an area and by the residence relationship compute the population of the area. We could just as easily total the number of establishments in an area and thus the population employed in that area. These transactions would provide macro-level cross-sectional data similar to those obtained from population, housing, and economic censuses (Figure 2).

One of the most direct ways to use this macro-level model of administrative data would be to do an administrative record census. A number of European countries now tabulate their population registers (supplemented with household surveys) in place of a complete census enumeration. Proposals for a similar tabulation based on administrative records rather than a population register have been made for the United States and for Canada. To date, the complexity of such an undertaking and the differences in residence and other concepts between censuses and administrative records have forestalled a full-fledged administrative record census.

A more limited operation might substitute administrative records methods for certain variables previously asked as census questions -- either to reduce respondent burden and cost or to increase the frequency of data collection. An example might be an administrative records procedure for estimating "commuting" -- a concept which is currently measured by comparing place of residence on the census with a census question dealing with place of work. Alternatively, using linkages between administrative records shown in Figure 2, one could tabulate the place of employment for all persons residing in a given area. The result would show the commuting flows out of the area of residence. One could then calculate the place of residence for all persons working in a given area giving the commuting flows into the place of employment.

A Longitudinal Model

One of the great advantages of administrative data over censuses and most surveys is their longitudinal character. Since administrative data are generally established to track individuals in a program and to record their entry, exit, and changes of status, the time dimension is an integral part of most administrative data sets. Figure 3 illustrates the time dimension of each of the four variables from Figure 2, ignoring for the moment the relationships between the four variables. If we were interested in changes for the whole universe, we could obtain substantial information from these data sets themselves. For example, the change between population at time 1 and time 2 could be calculated by looking at the net result of entry and exit from the administrative system. If the system covers most of the population over most of its lifetime, then this approximates the vital events (births and deaths) of standard demographic methods. Similarly, the housing database at time 1 is updated by construction and demolition statistics from construction data bases. Establishment data would be updated by information on openings and closings. Geography would change over time as a result of annexations and new incorporations. Such a system would provide considerably more information about the dynamics of each of these variables than cross sectional measures provide.

However, the real gain from multiple administrative records systems comes from the ability to look simultaneously at data across variables and across time. For example, we could calculate such traditional measures as local residential mobility, migration, and job mobility without the need for retrospective questions of a panel study by comparing simple relationships over time:

Local Residential Mobility: $H(t1) \neq H(t2)$ but $G(t1) = G(t2)$

Migration: $H(t1) \neq H(t2)$ and $G(t1) \neq G(t2)$

Job Mobility: $E(t1) \neq E(t2)$

It is this ability to show transitions over time along various dimensions that make administrative records so appealing to demographers. Stable population theory has been extensively used to combine vital statistics on births and deaths with population data from the decennial census. It has been extended to multistate demography by adding information on migration. But a number of suggested advances in multidimensional demography have not been possible because of the lack of data on transitions between other categories over time. Tables of working life suffer from lack of data on job entry and leaving. Household or family life tables require data on household formation and dissolution rather than simply changes in the headship rate over time. Yet, statistics for these and many other transitions could theoretically be obtained from a good series of administrative records data bases.

Macro-level Longitudinal Methods. The Census Bureau postcensal estimates program has used longitudinal administrative records methods at the macro level for over 20 years. The primary method used is the tax return method which uses the basic demographic accounting equation:

$$P(t) = P(0) + B - D + m(P(0)),$$

Where $P(0)$ is the initial population of an area,
$P(t)$ is the estimated population at time t,
B is the number of births during the period,
D is the number of deaths during the period, and
m is the rate of net migration during the period.

While $P(0)$ is available from the census and B and D are available from vital statistics data, m must be estimated from administrative records data. In this case, the filing addresses on tax returns for the initial year and the estimate year are compared. Any differences are assumed to represent migration. The number of exemptions on tax returns with a given pair of geographic units for origin and destination are divided by the total number of exemptions on all returns filed in the origin in the initial year. The resulting "out migration rate" is subtracted from a similarly calculated "in migration rate" to provide the "net migration rate" used in the postcensal estimates.

Conclusions

Using administrative records in demographic accounting requires facing many

practical challenges in addition to those already mentioned. Administrative data usually cover only a portion of the universe of interest; so special care must be taken to allow for missing populations. Also since administrative data are compiled for very specific uses, their generalization to wider statistical uses may raise problems of definitions and concepts as well as the absence of key information. Most importantly, the use of administrative records requires great care in order to protect the privacy and confidentiality of individuals and businesses.

On the other hand, administrative records have great potential to reduce data collection costs, to reduce respondent burden, to increase the frequency and timeliness of data, and to produce new data that was not previously collected by questionnaires. While the direct micro-level approach appears to be the most straight forward, a number of the flaws in administrative data sets may be amenable to correction by using modeled results or by aggregating to the macro-level. Recognizing this potential, the Census Bureau has made a commitment as part of its strategic plan to push the administrative records agenda forward.

Figure 1

Figure 2

Figure 3