ADMINISTRATIVE RECORDS IN PRIVATE INDUSTRY

Mary Kay Healy, Donnelley Marketing Information Service

For purposes of direct mail applications, Donnelley Marketing has maintained, for the past 50 years, a file containing the names and addresses of all persons listed in the 5,000 telephone and city directories throughout the United States. This data file consists of approximately 58 million records. This basic telephone file is supplemented by information from automobile registration data, available from 35 states and collected by the R. L. Polk Company. The addition of these data increases the file to 70 million nonduplicate records. Since the complete file encompasses about 87 percent of all U.S. households, it can loosely be considered akin to a "pseudo-population" register. It was this expansive coverage of U.S. households that prompted the management of Dun & Bradstreet, Donnelley's parent company, to suggest the development of demographic data products from the telephone and auto records.

While researching a technique to produce population and household estimates, the decision was made to exclude the auto data, in order to eliminate reliance on outside data sources. In addition, the availability of this information changed on a yearly basis. Initial tests of the estimate procedures indicated that the removal of the auto data as a symptomatic data series did not affect the quality of the estimate results.

DATA BASE DEVELOPMENT

The actual compilation of the telephone list is a rather tedious and labor intensive task. Because telephone listings are not currently available to Donnelley Marketing in a computer-readable format, individual directories are reviewed manually for additions or deletions to last years's file. Moves within the same directory area are treated as both a delete and an add, because of the change in address. The date of a person's first appearance at an address is also noted, enabling us to determine a measure of length of residence. The review and updating of the telephone file is an ongoing process, since directories are published at various times throughout the year. The time lag from the date of publication to inclusion in the data base is approximately three months. The timeliness of this information is a distinct strength in the production of estimates. At any point, the data contained are usually no more than a year old and, in many cases, there is only a lag of a few months. This is especially significiant when producing tract-level estimates, where change can be highly volatile in a relatively short time period.

In order to use the telephone list in the production of small area demographic estimates (i.e., tracts and minor civil divisions or MCD's), the information must be geocoded to correct census geography. Donnelley has developed a computerized geocoding system that expands the Census Bureau's GBF/DIME (Geographic Base File/Dual Independent Mapping-Encoding) files to include all geographic areas outside of the DIME coverage. Changes that have occurred since the initial creation of the DIME files have been tracked and introduced to the file. This is of major importance to the accuracy of the final estimates, in view of the fact that most of the DIME files are four or more years old and that the funding for the maintenance of the GBF/DIME system has been greatly curtailed. The geocoding capability outside of DIME areas was generated by the Donnelley staff through the extensive use of city street and census maps.

Although the Donnelley geocoding system allows for geographic assignments to the block group and enumeration district level, demographic updates are produced at the tract and MCD-levels only. Telephone listings without street addresses, such as rural route numbers and post office boxes, cannot be geocoded to a correct residential location. Fortunately, these occur more frequently in rural areas, where the population and household changes are not as rapid, so the loss in coverage is not a serious detriment.

Before any estimates are produced, the geocoded telephone list is compared with the actual number of households enumerated in the census, to determine the rate of coverage for each geographic unit. Evaluations of the 1980 counts are not complete, since the final address-coding guide was not finished until April of this year. The 1970 evaluations, however, revealed an average telephone coverage (of telephone households to census households) in tracted areas of 60 percent. Tracts in suburban areas, those with predominantly single-family homes, had the highest coverage, with rates of 85 percent or better. The lowest coverage levels occur, ironically, at both ends of the income spectrum. This occurence is due to the high proportions of unlisted telephones in wealthy areas and the greater likelihood of no telephone at all in low income households. A review of coverage in test areas for 1980 indicates that coverage will generally follow the same patterns exhibited with the 1970 data, but that the number of misallocated or uncoded addresses will be reduced.

ESTIMATE METHODOLOGY

The estimating method employed by Donnelley Marketing Information Services uses change in the number of telephone households in an individual tract or MCD as a surrogate for the change in actual households. The rate of change is utilized, rather than absolute change, because of incomplete coverage. In areas in which coverage was extremely low a

replacement rate is substituted. This rate is computed from the tracts or MCD's within the same place or county that has acceptable coverage levels.

A basic housing unit technique is applied; however, several of the weaknesses inherent in most housing unit methods have been removed. For example, rather than estimating the total number of housing units, as is necessary when building permits are used, the total number of occupied units or households is estimated directly. This eliminates the problems associated with time lags between the date of issuance and the completion of a unit, as well as estimates of vacancies. While utility data, such as electrical hookups, enable a direct estimate of households, there is a growing problem with master meters and the conversion from master to individual meters or vice versa. Care must be taken so that these changes are not construed as growth trends.

Initial tests of the household estimates were conducted in three census pretest areas: Oakland, California; Richmond, Virginia; and Lower Manhattan, New York City, New York. The results of these comparisons indicated absolute average percent differences of 15.9, 19.0 and 15.4, respectively. When the individual tracts were summed to the larger geographic entity, the level of error was reduced dramatically. In Lower Manhattan, the total number of households estimated was 46,463, while the enumerated households totaled 46,324. An additional review of the household estimates at the county level was made against the residential building permit file maintained by the Census Bureau's Construction Division. Again, the estimate proved to be highly accurate. On the basis of these results, the decision was made to assume that, at larger aggregate levels (place and balance-of-county or county total), the household estimates are virtually correct. It was also assumed that the Census Bureau's place and county population estimates are the most accurate available. Therefore, in order to eliminate the need to estimate household sizes from national level Current Population Survey (CPS) data, an estimate of household size was produced at the place, "balance" and county level, by dividing the Census Bureau's population estimate (extrapolated to 1980 and adjusted for group quarters) by the Donnelley summary level household estimates. The change in household sizes exhibited at these summary levels was applied to all smaller units within the geographic entity. Group quarters figures, as determined by the Census Bureau's college and institutional file or the 1970 census, were added to the household population. While it is weak to assume that all household sizes within an area change at the same rate as a large geographic level, it is far superior to applying national or state averages, as is common in most housing unit method estimates.

1980 TEST RESULTS

Since any methodology is only as valid as the estimates that result, a test of Donnelley's 1980 population and household estimates was undertaken for approximately 43,000 tracts and MCD's, where the geographic boundaries between 1970 and 1980 did not change or could be easily recombined, such as split tracts. This is the first time that a test of this magnitude for small areas has been conducted and presented. Because tests of tract level estimates are not generally available, the only evaluations for comparison purposes are for places and counties. The literature on estimate results, however, does conclude that, the smaller the area being estimated or the greater the change, the higher the expected level of error. These tendencies should have significant impact on the tract and MCD estimates, where the average population size of the areas was 1,500 persons; approximately 50 percent had fewer than 1,000 households; and 60 percent of the areas gained or lost population and households at a rate of over 10 percent in the last ten years. In fact, 30 percent of the tracts and MCDs had a household rate of change in excess of 25 percent. With these kinds of considerations in mind, larger errors were expected than those generally found for more standard levels of estimate (i.e., places and counties).

The absolute average percent difference for the Donnelley estimated tracts and MCD's was 15.8 percent for population and 15.5 percent for households. This compares with Census Bureau test results for 2,000 place estimates in 1975, of 11 percent for population, and a state of Florida study indicating an absolute average percent difference of 14 percent for all cities, when compared with the 1980 population census counts. Almost 50 percent of the Donnelley estimates had error levels less than 10 percent (Table 1). In tracts with less than 1,000 people, the average error of 16.6 percent was lower than the Census Bureau's 18.7 percent level for places in the same size-range (Table 2). When rate of change is taken into account, the Donnelley estimates fare better than the Bureau's place estimates at the extreme ends of population change (Table 3). Also important to note is the fact that both the population and household methods produced results that were relatively unbiased. Population was underestimated 52 percent of the time and overestimated 48 percent of the time. Households were underestimated in 44 percent of the cases, overestimated in 55 percent and exactly equal in 1 percent of the cases.

These test results illustrate that, even in small, rapidly changing areas, the Donnelley estimates are well within an acceptable range of error. The results, however, are more favorable when considering that the estimates were not designed for use on a tract by tract basis, but rather in aggregate forms, such as trade areas and potential site locations. When grouped in this manner, the error levels generally are reduced. A list of some recently requested client-designated sites is found in Table 4. These evaluations indicate the rather dramatic drop in error levels as the estimates are grouped.

This rather vigorous testing of the estimates has shown the quality to be

acceptable when compared to other recently evaluated estimates, despite the differences in geographic levels. Obviously, there are areas where the errors will be larger than the averages depicted, and clients should use the data with the normal cautions taken with any small area estimates. However, the presentation of the estimate methodology and the test results is a very serious effort on the part of Donnelley Marketing Information Services to prepare small area population and household estimates, through the use of sound demographic techniques, that can be improved over time.

ACKNOWLEDGMENTS

A special thanks is extended to Joyce Coleman of the Internal Revenue Service for her typing assistance.

REFERENCES

[1] Smith, Stanley K. and Mandell, Marylou. "A Comparison of Local Population Estimates: the Housing Unit Method vs. Component II, Ratio Correlation and Administrative Records," prepared for presentation at the annual meetings of the Population Association of America, San Diego, California, April 29 - May 1, 1982.

[2] U.S. Bureau of the Census, Current Population Reports, Series P-25, No. 699, "Population and Per Capita Money Income Estimates for Local Areas: Detailed Methodology and Evaluation," U.S. Government Printing Office, Washington, D.C., 1980.

Table 1. Number of Tracts/MCD's By Size of Difference

Difference	Population	Households
Less than 10.0 percent	20,547	20,731
10.0 to 14.9 percent	6,825	6,658
15.0 to 19.9 percent	4,935	4,487
20.0 to 24.9 percent	3,307	3,246
25.0 percent and over	7,625	8,086

Table 2. Number of Tracts/MCD's by Size of Area and Absolute Average Percent Difference

Size of Area	Population	Percent Difference	Households	Percent Difference
Less than 1000	9,616	16.6	21,187	16.9
1,000 to 1,999	5,088	15.9	14,878	14.1
2,000 to 2,999	5,519	15.3	4,690	13.6
3,000 to 3,999	6,186	14.5	1,393	13.5
4,000 and over	16,875	13.7	1,137	13.5

Table 3. Number of Tracts/MCD's by Rate of Change and Absolute Average Percent Differences

Percent Rate of Change 1970 to 1980	Population	Percent Differences	Households	Percent Differences
-50.0 or more	53	35.0	31	58.3
-25.0 to -49.0	1,438	28.4	670	34.8
-10.0 to -24.9	8,180	13.9	3,356	18.2
-0.1 to -9.9	9,295	11.9	5,949	11.8
0.0 to 9.9	9,097	13.4	8,536	12.0
10.0 to 24.9	8,431	14.9	11,369	14.3
25.0 to 49.9	4,519	17.8	8,870	16.2
50.0 and over	1,761	20.9	3,555	20.6

Table 4. Client Designated Sites by Size of Radius and Absolute Average Percent Error

Location	Radius (in miles)	Population	Households
Amherst, NY	2	0.83	11.58
Costa Mesa, CA	5	7.66	3.49
Utica, NY	1	6.38	6.68
East Organge, NJ	5	4.33	2.35
Orlando, FL	3	12.20	6.93
Attleboro, MA	3	3.53	1.37
Northampton, MA	1	10.89	3.93
Natuli, MA	2	8.07	8.46
Tonawanda, NY	5	4.28	3.25