DRAFT ONLY

An Evaluation of Texas Population and Estimates and Projections Program's Population Estimates and Projections for 2000

by

Md. Nazrul Hoque

University of Texas at San Antonio
Department of Demography and Organization Studies
San Antonio, TX 78249-0704

Paper prepared for presentation at the Annual Meetings of the Population Association of America, March 29-31, 2007. New York, New York.

An Evaluation of Texas Population and Estimates and Projections Program's Population Estimates and projections for 2000

Introduction

Population estimates and projections are difficult to complete with accuracy for small areas (Murdock et al., 1991; National Academy of Science, 1980; Ascher, 1978). As a result, it is essential that any ongoing program of population estimation and projection periodically evaluate the results of past estimation and projection efforts against actual counts of the population (Murdock and Ellis, 1991). Only by assessing the accuracy of past efforts it is possible to know the nature of errors made and to take steps to improve future estimates and projections. This paper presents the results of the evaluation of the Texas Population Estimates and Projections Program's population estimates for 2000 compared to the 2000 Census Counts for counties and places in Texas. We evaluated Component Method II, the Ratio-correlation Method, and the Housing Unit Method. Component Method II depends on the use of three characteristics of population that directly determine population change: births, deaths and net migration. Thus, for any period, the population can be determined using the following equation:

$$P_t = P_o + B - D + M$$

where: P_t = Population for the Estimate Period

P_o = Population at the base period

B = Births between P_t and P_o

D = Deaths between P_t and P_o

M = Net Migration between P_t and P_o

Component Method II typically takes direct account of natural increase through actual data on births and deaths, while using symptomatic data for assessing net migration. Component

Method II assumes that the rate of migration of school-age children can be used to assess the migration rate for the population 64 years of age and younger and that medicare data can be used to estimate the migration rate for the population 65 years of age and older.

The Ratio-correlation Method is a multiple regression-based technique which compares change in one areal unit to change occurring in a parent area. Such estimates are developed using the following multiple regression equation:

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + ...\beta_n X_n = e$$

where: Y = the dependent variable to be estimated (e.g., population)

β_o = the intercept to be estimated

β_i = the coefficient to be estimated

X_i = independent variables, such as births, deaths, voter registration, etc.

e = error term

This model uses variables of births, deaths, elementary school enrollment, vehicle registration, and voter registration. The dependent and independent variables are expressed in the form of a ratio. For example, to obtain the estimate of population for a county in 2001, where the State population is known, the following equation could be applied:

$$\frac{\text{County Pop, 2001/State Pop, 2001}}{\text{County Pop, 2000/State Pop, 2000}} = \beta_o + \beta_1 \left(\frac{\text{County Births, 2001/State Births, 2001}}{\text{County Births,2000/State Births, 2000}}\right) + ...$$

In the equation above, all of the indicator values are known except county population. In order to obtain the intercept and coefficients to use in solving the equation, estimates of the values must be obtained. This is done by solving the equation for past periods for which all the values are known. For example, the coefficients obtained by solving the equation for the past periods (e.g., 1980-90) can be used in the above formula for a 2001 estimate.

The Housing Unit Method used is the standard form (i.e., add new units, subtract demolition units from the base units and add group quarters population). For a detailed discussion, see Smith and Lewis, 1980.

$$P_t = (HU_t + BP_t - DU_t) \times OCC_t + GQ_t$$

where: P_t = the total population at time $_t$

HU_t = occupied housing units counted in the most recent census (by type, e.g. single family, multifamily, mobile home)
BP_t = building permits issued by type between the most recent census and time $_t$ (adjusted for time lag)
DU_t = units reported demolished by type between the most recent census and time $_t$
OCC_t = the occupancy rate by type at time $_t$
GQ_t = the group quarters population at time $_t$

The evaluation presented here was used as a major source of information for the modification of population estimates and projection procedures for the 2000s.

In the remainder of this paper we describe several basic principles of population estimation and projection, the historical context of population growth from which the Texas program's estimates and projections were made, and the methods used in the evaluation. We also present an evaluation of county and place-level estimates and projections produced by the Texas program. Finally, we outline potential changes to be implemented in the post-2000 estimates and projections as a result of this evaluation and one which will be completed after the 2000 census.

Principles of Population Estimation and Projection

The history of population estimates and projections suggests certain basic findings from past analyses that merit recognition in an evaluation of any estimation or projection. These basic findings or principles, as outlined by Shryock and Siegel (1980), show that no single method has been found to consistently produce more accurate estimates and projections than any other method, and that population estimates and projections are generally more accurate

1. For large rather than small (in terms of population size) estimation or projection areas and may not be significantly better than past population counts for very small population areas (National Academy of Science, 1980).
2. For total populations rather than for estimates or projections of population characteristics because estimates and/or projections of such characteristics involve additional assumptions that may prove to be in error (Murdock et al., 1991).
3. If completed with data on processes that directly determine population change (such as data on births, deaths and migration) rather than when indirect or symptomatic indicators of population change are used.
4. For short rather than long periods of time past the reference date for the base data used in the estimates or projections.

5. For areas that show consistency in the direction of change during the estimation or projection period compared to the period from which the base data are derived.
6. For areas that experience slow rather than rapid change.

In sum, an estimate or projection is likely to be most accurate if it is based on direct birth, death, and migration data and is for the total population of a large area that is showing slow change that is in the same direction and form as that for the recent past. In general, the greater an area's departure from these conditions, the greater the error in estimates or projections.

Population Change in Texas in the 1990s

Texas population change during the 1980s was such as to make population estimation and projections difficult. After increasing by an annual exponential rate of 2.4 percent from 1970 to 1980 and by nearly 3.5 percent from 1980 to 1982, growth began to slow in Texas during the middle and later parts of the 1980s. Thus, from 1986 to 1988, the state experienced net outmigration and an overall rate of growth of 0.5 percent per year (growth rather than decline occurred because relatively high levels of natural increase offset net outmigration). The State's rate of growth from 1986 to 1988 was only about one-ninth of that from 1980 to 1982 and one-sixth of that of the 1970s. Similarly, patterns among counties and places in the State changed dramatically during the 1980s relative to the 1970s. During the 1970s, only 44 of Texas 254 counties and only 20 percent of the State's nearly 1,200 places showed population decline, but from 1980 to 2000, the number of counties losing population increased to 98 and the proportion of places showing population decline increased to 46 percent.

Despite slow growth during the latter part of the 1980s, the 2000 Census showed the State to have a population of 16,986,510 persons, representing an increase of nearly 2.8 million and a rate of growth of 19.4 percent. This gave Texas the third largest numerical increase in population in the Nation (behind California and Florida).

In sum, during the 1980s the State of Texas and its component areas showed extensive population growth but dramatic changes in rates of growth from 3.5 percent per year in the early 1980s to 0.5 percent per year during the later part of the decade. Many counties and places changed their patterns of population change from growth to decline. The State's patterns were thus ones that generally violated the principles which accurate estimates and projections.

Methods for Evaluation

Given the patterns and principles noted above, several widely used procedures (Murdock et al., 1984; Murdock et al., 1991) were selected to evaluate the population estimates and projections. These methods generally rely on comparisons of values for error measures for the estimates or projections being evaluated relative to expected patterns and relative to those for estimates or projections from other sources. The estimates and projections were evaluated relative to the expected patterns of increased rates of error with decreased population size and increased rates of error with increased rates of population change. They were also evaluated relative to their tendency to underestimate or overestimate the population of different types of areas. The comparison of estimates and projections to those from other sources assists the analyst in identifying which factors may be impacting the accuracy of estimates or projections because the assumptions can be compared to those used by other sources. Such a comparison often helps to determine which of the assumptions are increasing or decreasing the accuracy of the estimates or projections.

Several error measures are used to assess the accuracy of estimates. The error of an estimate or projection is determined by subtracting the estimated or projected population value for an area from the census count (for purposes of this report, the 2000 census count) and dividing the difference by the census count. This proportion is then multiplied by 100 to produce a percentage rate of error.

Three error measures are commonly used in such assessments. The formulas for these measures are shown in Figure 1. They include the Mean Percent Error (MPE), the Mean Absolute Percent Error (MAPE), and the Mean Percent Absolute Difference (MPAD). This later measure is also referred to as the weighted mean absolute percent error.

The Mean Percent Error (MPE) is simply the arithmetic average of the percent errors for each area (county, place, etc.). Although this value is useful, because positive and negative values cancel one another in its computation it may provide one with somewhat deceptive estimates of error. For example, if one were to take two areas and underestimate the population of one by 50 percent and overestimate the population of the other by 50 percent, the MPE would be 0.0 percent suggesting that the estimates were perfect when in fact the two component estimates were quite inaccurate.

The Mean Absolute Percent Error (MAPE) is the mean of the absolute values of the errors (that is, values ignoring the sign of the value). Given that magnitude rather than direction of the error is usually the major concern, MAPE provides a more useful overall estimate of total error and is the most widely used measure of error in evaluations of population estimates and projections. Both MPE and MAPE, however, share a common weakness in that errors for all places contribute equally to the overall error rate computed. Suppose the estimate for an area with 1 million persons fell within two percent of the actual count, and the estimate for an area of 100 persons fell within 18 percent. The MAPE for the two areas would be 10 percent (2 plus 18 divided by 2), although the estimate for the area with most of the population was quite good. The problem is that neither MPE nor MAPE take the size of the areas in the computation into account.

The Mean Percent Absolute Difference (MPAD) or weighted mean absolute percent takes the size of areas into account by weighting areas' values proportionally to their size (proportionately to the population size of the area as a proportion of the sum of the populations of all the areas of interest). The MPAD is thus also widely used in evaluations of estimates and projections.

These three error measures' values are presented for each type of area (i.e., counties and/or places) and for the areas grouped by population size in 2000 and rate of change from 1980 to 2000. Data are also shown for the number of overestimated and underestimated areas to indicate the extent to which the estimates or projections tend to be biased either upward or downward. The number of areas estimated or projected within certain ranges of error is also provided to indicate how many areas are estimated within specified levels of error. Finally, the errors in the estimates are compared to those from other sources.

Evaluation of Texas County and Place Estimates for 1990

Evaluation Procedures

In this section, the results of the evaluation of the Texas Population Estimates and Projections Programs' estimates are presented. The results are presented first for counties and then for places. The evaluation consists of an examination of the accuracy of the base techniques employed by the Texas Population Estimates and Projection Program to produce county and place estimates. It also examines the base techniques compared to the accuracy of estimates made by altering several key features of the base techniques and estimates produced using a methodology not previously used in the Texas Estimates and Projections Program.

During the 1990s, county-level estimates were computed as an average of the Ratio-correlation Method and Component Method II. Place-level estimates were made using only Component Method II.

The Results of the Evaluation of County-Level Estimates

Table 1 presents the three error measures for different estimation methods for counties in Texas. The results in this table show an overall mean percent error of -1.27 percent, a mean absolute percent error of 6.502 percent and a mean percent absolute difference value of 3.65 percent for Component Method II (see Table 1, Panel 1), and an overall mean percent error of -1.25 percent, a mean absolute percent error of 4.79 percent and a mean percent absolute difference of 2.58 percent for the Ratio-correlation Method (see Table 1, Panel 2). For the Housing Unit Method the overall mean percent error was 0.63

percent, the mean percent absolute error was 10.32 percent, and the mean percent absolute difference was 8.31 (see Table 1, Panel 3). Averaging Component Method II and the Ratio-correlation Method produced an overall mean percent error of -1.49 percent, a mean percent absolute error of 4.62 percent, and a mean percent absolute difference of 1.88 percent (see Table 1, Panel 4). Panel 5 shows the mean percent error, mean percent absolute error and mean percent absolute difference for the average of all three methods. In general, the data in Table 1 suggest the expected patterns, with error measures being smaller for counties with larger populations. The data in Table1 also suggest that the average of two or three methods is clearly superior to the use of any single method of estimation. The results reported in Table 1 also suggest that the alternatives to the base method did not markedly improve the accuracy of the estimates, and thus their inclusion in the 1990s estimation methods would not have substantially improved the accuracy of the estimates for counties.

Tables 2 through 4 examine the rates of error for counties for each of the same procedures described above. However, these are shown alternatively for the rate of population change from 1990 to 1990 (Table 2), the range of error for the estimate (Table 3), and by the number of counties estimated above or below the 1990 Census value (Table 4). It should be noted that in Table 2 the values for counties whose population is less than 1,000 are the same in the base, Housing Unit and average of Housing Unit and base methods panels of the table. This is because no housing permit data are available for counties with populations of less than 1,000 and so base method data were substituted to provide an overall error value that included all counties.

In general, the data in Tables 2 and 3 show the expected patterns with the mean error measures being smaller for counties with larger populations and for counties with moderate levels of population change. The data in these two tables, like those in Table 1, show that the overall levels of error are comparable to those expected, and indicate that the alternative of using multiple methods (as opposed to the average of Component Method II, the Ratio-correlation Method, and the Housing Unit Method) shows little improvement in accuracy when compared to the average of Component Method II and the Ratio-correlation Method. In fact, in the largest and fastest growing counties, the average of Component Method II and the Ratio-correlation Method is clearly superior to the other methods. The data in Table 3 also provide general support for the relative accuracy of the methods with more than 91 percent of all counties in the base methods (Component Method II and Ratio-correlation) being estimated within 10 percent of the actual 1990 Census Counts. Only 4 of the 254 counties have a 20 percent or more error from the actual 1990 Census Counts.

Finally, the results in Table 4 show that the base method (an average of Component Method II and the Ratio-correlation Method) tended to be biased downward with 63.24 percent of the counties being underestimated, while an average of three methods produced estimates that tended to overestimate the populations of counties (56.52 percent) in 1990. Table 5 presents the summary error measures for the base estimation methods, for the Housing Unit Method and for an average of Component Method II, the Ratio-correlation Method, and the Housing Unit Method.

Overall, the results of the evaluation of the county estimates indicate that the base method produced estimates that were clearly within acceptable ranges of error and that the alternative method assessed did not markedly improve the overall accuracy of the estimates relative to those prepared using the base method.

The Results of the Evaluation of Place-Level Estimates

Table 6 presents error measures for place estimates for Texas for 1990 for Component Method II, the Component Method II without the migration adjustment factor, the Housing Unit Method, and the average of Component Method II and the Housing Unit Method. The overall levels of error shown in this table are within the high end of the expected range of 10 to 15 percent error when estimates for a relatively large number of small population areas are involved; however, they are higher than is desirable. In addition, the data in this table do not show substantial improvements when using Component Method II without the migration adjustment factors. The Housing Unit Method shows some improvement in terms of the error measures. However, the Housing Unit Method could be used for only 726 of the 1,208 places because housing permit data were not available for the rest of the places. As a result, the error measures for the Housing Unit Method in part reflect a different sample of places than those for the base methods and its altered versions. Despite this fact, the use of the Housing Unit Method in conjunction with Component Method II, as shown in the last part of Table 7, suggests that this combined method may improve place-level estimates over those made using the base method alone. As a result, the Texas Population Estimates and Projections Program added the Housing Unit Method to its base methods for the 1990s.

Tables 6 through10 show results of the evaluation of estimates in terms of size of place (Table 6), percent population change (Table 7), range of error (Table 8), and relative direction of the estimation error (Table 9) for places in Texas. Table 10 presents the summary measures for all of the methods. The data in these tables show the expected patterns of higher levels of errors for smaller and more rapidly changing areas (Tables 6 and 7), but also show much higher levels of error than those of counties. Similarly, as shown in Table 7, the percentage of places with extreme errors is larger than the same percentage for counties. Only about 40 percent of places (compared to 91 percent of counties) were estimated within 10 percent, and less than 60 percent of the places were estimated within 15 percent of the 1990 population count. Finally, the results in Table 10 point to a tendency for the population of places to be overestimated. Given the tendency for counties to be underestimated, it is clear that the base estimation procedure had difficulty in estimating the relative population of place and nonplace areas within the counties.

The data in Tables 6 through 10 also suggest that in nearly all cases, the use of the Housing Unit Method improves the accuracy of the estimates obtained. Thus the results of the Housing Unit Method averaged with Component Method II appear to lead to a reduction in the error of estimate for places in Texas.

In sum, the data for places suggest that the estimates for places are less accurate than those for counties. They suggest that there was substantial potential for improvement in the place estimates. Finally, they suggest that the use of the Housing Unit Method averaged with estimates from the base (Component Method II) method may provide one means of moving towards such improvement. As a result, the Housing Unit Method was added to the base procedures for the Texas Population Estimates and Projections Program for the 1990s.

Conclusion

The evaluation of the Texas Population Estimates and Projections Program's population estimates presented here suggests that the estimates are generally adequate and show levels of error that, when compared to the 2000 Census Counts, are within generally accepted ranges. They also show the expected patterns by population size and population change. Of the several methods tested, no single one produced more accurate estimates than the average of two or three methods. The assessment of the accuracy of the place-level estimates showed substantially higher levels of errors than those found for counties, and provided the basis for the revision of our estimates procedures for the 2000s.