Evaluation of 2000 Subcounty Population Estimates

Greg Harper, Chuck Coleman and Jason Devine
Population Estimates Branch, Population Division
U.S. Census Bureau

This paper reports the results of research and analysis undertaken by U.S. Census Bureau staff. It has undergone a more limited review than official U.S. Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion. This paper was presented at the 2002 Population Association of America meeting.

ABSTRACT

Since 1996 the Population Estimates Branch (PEB) of the Census Bureau has used a housing unit method to produce population estimates for subcounty areas. The distributive housing unit method uses data on building permits, mobile home shipments and housing unit loss to distribute county population to incorporated places and minor civil divisions within counties. The availability of Census 2000 data gives PEB the opportunity to evaluate the accuracy of these estimates.

In addition to comparing Census 2000 results with April 1, 2000 population estimates using Mean Absolute Percent Error and Mean Algebraic Percent Error, this paper presents regression analyses to explain the causes of estimate errors. The results of this analysis will be used to inform Census Bureau analysts on ways in which the current subcounty estimates method can be improved.

Contents

Charts

Tables

Map

BACKGROUND

This report presents an evaluation of estimates of the total population for 40,630 subcounty areas. These subcounty areas, all of which are governmental units, consist of both incorporated places, such as cities, boroughs, and villages, and minor civil divisions, such as towns and townships. Estimates for these areas were first produced in the early 1970s in response to the State and Local Fiscal Assistance Act of 1972 (Public Law 92-512) that created the Federal General Revenue Sharing program. Although this program ended in 1986, these estimates are mandated by Title 13 and remain one of the ongoing programs in the Population Division of the U.S. Census Bureau. These estimates are a major part of the entire set of estimates used as the basis for the distribution of over $200 billion in annual funding and to determine eligibility for a variety of government programs at the Federal, State, and local levels.

During the 1990s the Census Bureau produced July 1 estimates for 1992, 1994, 1996, 1998 and 1999. An additional set of April 1, 2000 estimates was produced for comparison with the decennial census. Although the 2000 test estimates were produced using the same method as in earlier years, they did not undergo the same review procedures as a typical set of published estimates. Specifically, the 2000 test estimates were not reviewed by members of the Federal State Cooperative Program for Population Estimates (FSCPE), nor were they put through the same rigorous internal Census Bureau review as the estimates that are used for the purposes of fund allocation. The comparison of the April 1, 2000 estimates to the April 1, 2000, decennial census counts forms the basis for this report.

The 1992 and 1994 subcounty estimates were produced using a component method known as the Administrative Records method. This method used address information on IRS tax returns to measure domestic migration. The method was discontinued at the subcounty level when it was determined that addresses on tax returns could no longer reliably be coded to subcounty areas. Beginning with the 1996 round of estimates, the Administrative Records method was replaced by the Distributive Housing Unit method. The Distributive Housing Unit method uses housing unit data at the subcounty level to distribute the county population to subcounty areas. See Appendix A for a more detailed explanation of the methodology.

ACCURACY OF 2000 ESTIMATES

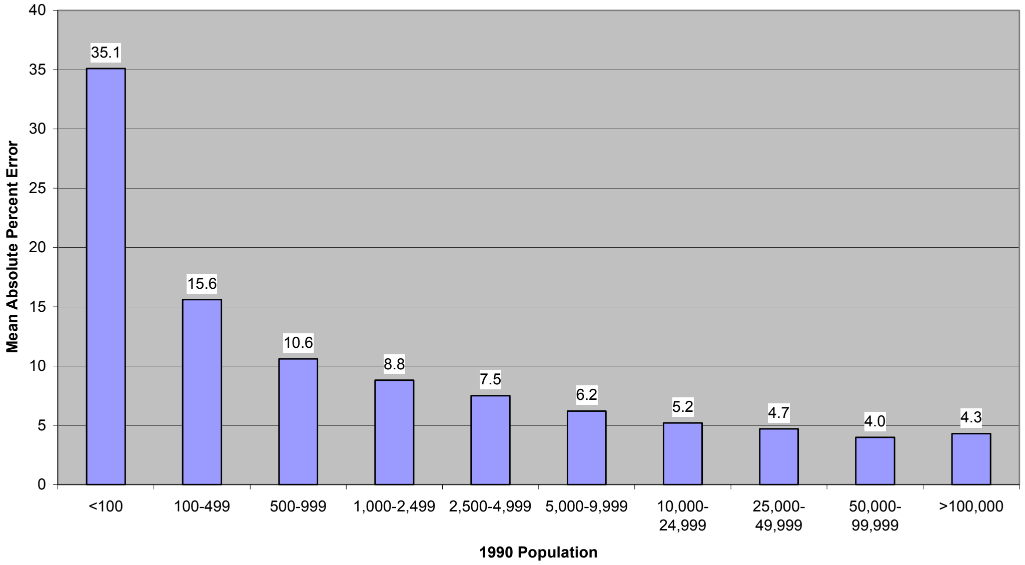
For the purposes of this evaluation, the differences between the estimates and the census counts are assumed to be due to errors in the estimates. Since these estimates begin with the 1990 Census as enumerated, differences in census coverage between 1990 and 2000 also account for some differences. The results obtained from the evaluation of the 2000 estimates are similar to those obtained from an evaluation of the 1980 estimates (Galdi, 1985). No evaluation results for the 1990 subcounty estimates have been published.

Table 1 presents a comparison of the April 1, 2000 estimates and the April 1, 2000 census counts by 1990 population size. The mean absolute percent error (MAPE) for all areas was 12.4 percent, 2.8 percentage points lower than the MAPE in 1980. The MAPE varied by area size, from a low of 4.0 percent for areas with a population of 50,000 to 100,000 to a high of 35.1 percent for areas with population less than 100 (Figure 1).

Table 1. Selected Measures of the Accuracy of Subcounty Population Estimates by 1990 Population Size: April 1, 2000

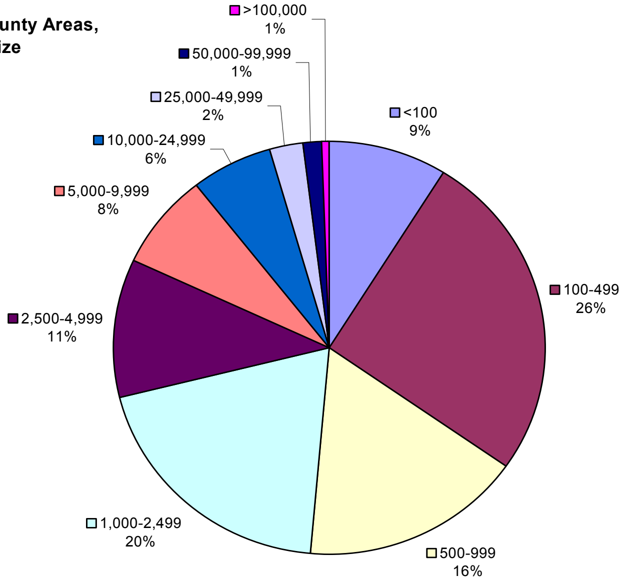
1990 Population Size	Number of Areas	Mean Absolute Percent Error	Percent Positive Errors
All	40,630	12.4	50.0
<100	3,572	35.1	55.5
100-499	10,590	15.6	53.5
500-999	6,661	10.6	50.8
1,000-2,499	8,085	8.8	49.5
2,500-4,999	4,375	7.5	48.3
5,000-9,999	3,083	6.2	49.4
10,000-24,999	2,493	5.2	42.1
25,000-49,999	987	4.7	36.2
50,000-99,999	505	4.0	30.5
>100,000	279	4.3	29.4

FIGURE 1
Mean Absolute Percent Error,
By Size Class in Population Estimates for 2000



The relatively high overall mean error (12.4 percent) was partially due to the large number of small areas estimated. Of the 40,630 areas, 71.2 percent had less than 2,500 people (Figure 2). However, when larger areas (population greater than 2,500) were independently examined, the mean absolute error was 6.1 percent.

FIGURE 2.
Distribution of Subcounty Areas, by 1990 Population Size



Eighty-one percent of errors of 20.0 percent or more occurred in areas with populations under 1,000. Of the 20,823 areas with populations less than 1,000, 22.8 percent had errors of 20 percent or more, while there were no errors of this magnitude for the 279 areas with populations of 100,000 and over.

Overall, 50.0 percent of the estimation errors were positive, indicating no median bias.[1] However, indications of bias were apparent for certain size classes. Specifically, only 29.4 percent of estimation errors for the 100,000 and over size class were positive. Some of this negative bias may be attributed to the relatively large underestimation of the national population in 2000 (2.4 percent).

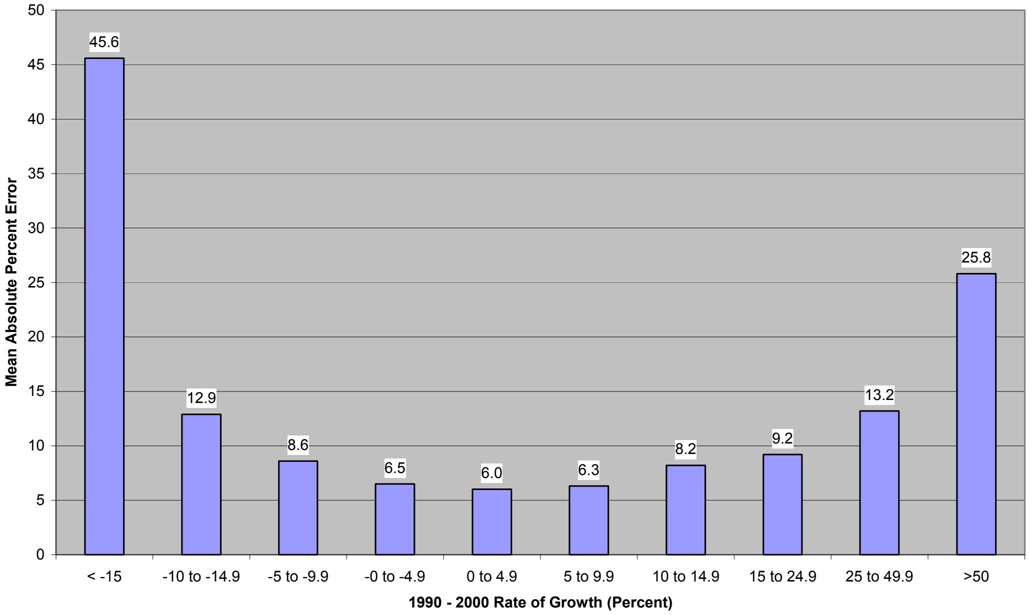
Table 2 presents the relationship between the 1990 and 2000 population change and the error rate. The degree of error in an estimate is related to the area's rate of change in population between 1990 and 2000 (Figure 3). Areas that grew by 50.0 percent or more had a mean error of 25.8 percent while areas that declined by 15.0 percent or more had a mean error of 45.6 percent. Areas that grew by less than 25.0 percent or declined by less than 10.0 percent had average errors of less than 10.0 percent.

[1] For an explanation of the distinction between median and mean bias, see Coleman (1999). The discussion in this section focuses on median bias.

Table 2. Selected Measures of the Accuracy of 2000 Subcounty Population Estimates by 1990-2000 Population Change

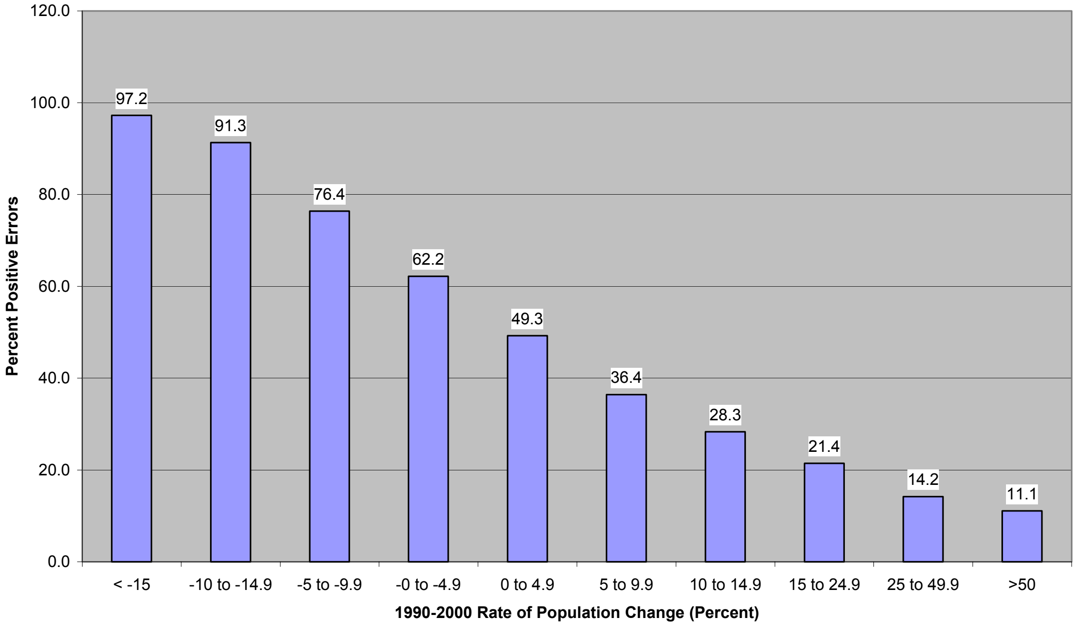
1990 to 2000 Population Change	Number of Areas	Mean Absolute Percent Error	Percent Positive Errors
All	40,630	12.4	50.0
< -15	3,769	45.6	97.3
-14.9 to -10	2,470	12.9	91.3
-9.9 to -5	4,243	8.6	76.4
-4.9 to 0	5,852	6.5	62.2
0 to 4.9	6,217	6.0	49.3
5 to 9.9	5,024	6.3	36.4
10 to 14.9	3,565	8.2	28.3
15 to 24.9	4,314	9.2	21.4
25 to 49.9	3,490	13.2	14.2
> 50	1,686	25.8	11.1

FIGURE 3
Mean Absolute Percent Error for Subcounty Population Estimates, by 1990-200 Population Change



There is a strong relationship between rate of growth and percent positive errors. Areas with high rates of population growth had estimates that tended to be lower than the census results. For areas that grew in population by 50 percent or more, only 11.1 percent had estimates that were higher than the census results. Similarly, areas that declined in population tended to have estimates that were biased high (Figure 4). Of the 16,334 areas that declined in population between 1990 and 2000, 78.4 percent had estimates that were higher than the census results.

FIGURE 4
Direction of Error for Subcounty Population Estmates, by 1990-2000 Population Change



Figures 5 and 6 respectively depict MAPE and MALPE by the size and growth classes of Figures 2 and 3. Figure 5 shows the usual U-shaped curves of MAPE for almost all size classes. These U-shaped curves indicate that MAPE is not monotonic in the growth rate, so that regression analyses of absolute percentage errors is impossible. Figure 6 illustrates the phenomena underlying the U-shaped curves. MALPE converges to zero for all growth classes as population increases and declines in the growth rate for all population classes. Thus, MALPE is monotonic in size and growth, holding either constant. An important interaction effect is identified: MALPE reaches its most extreme values for the lowest population class. Thus, the combination of size and growth is an important determinant of MALPE.

Figure 5.

MAPE by 1990 Population and 1990-2000 Population Growth Rate

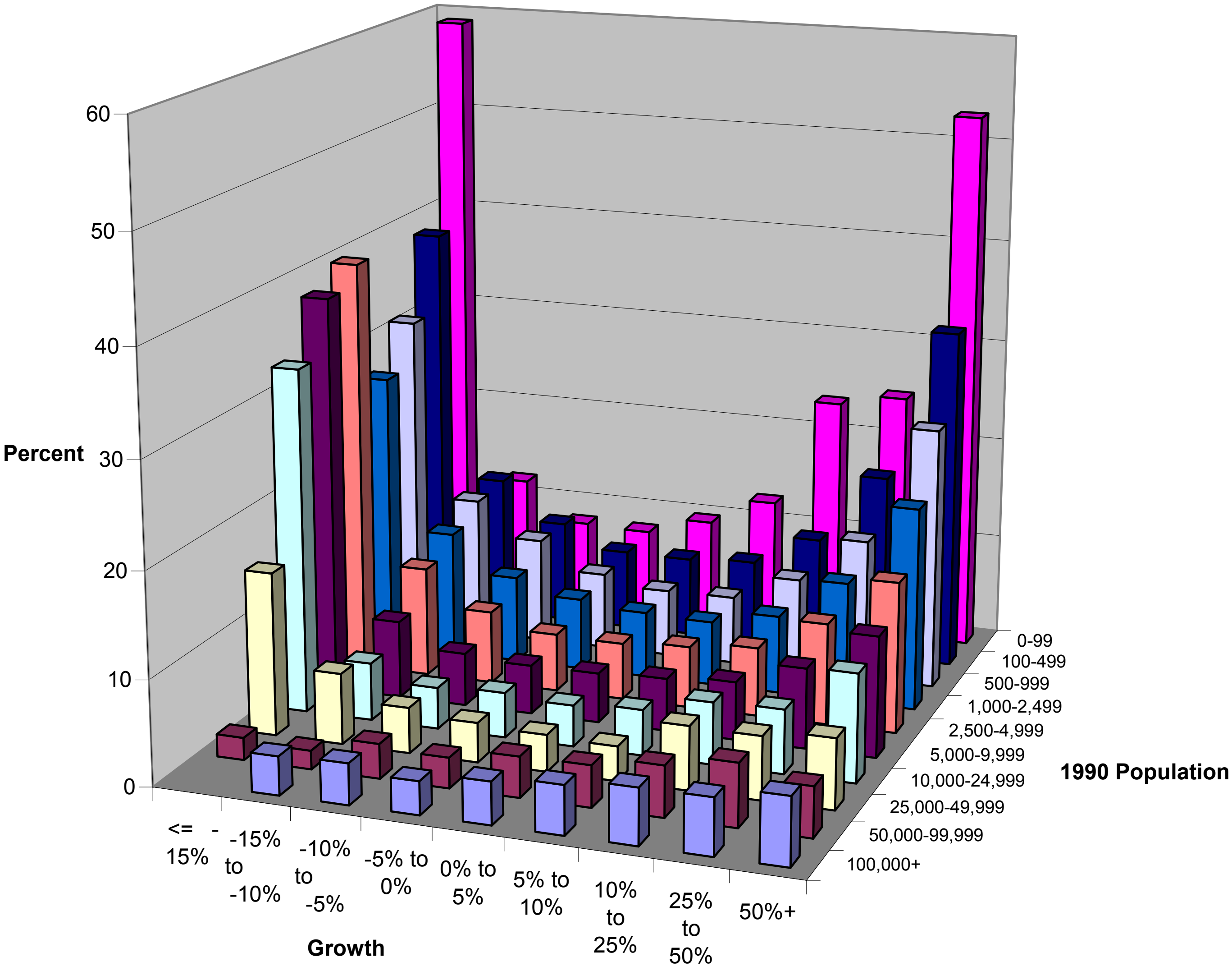
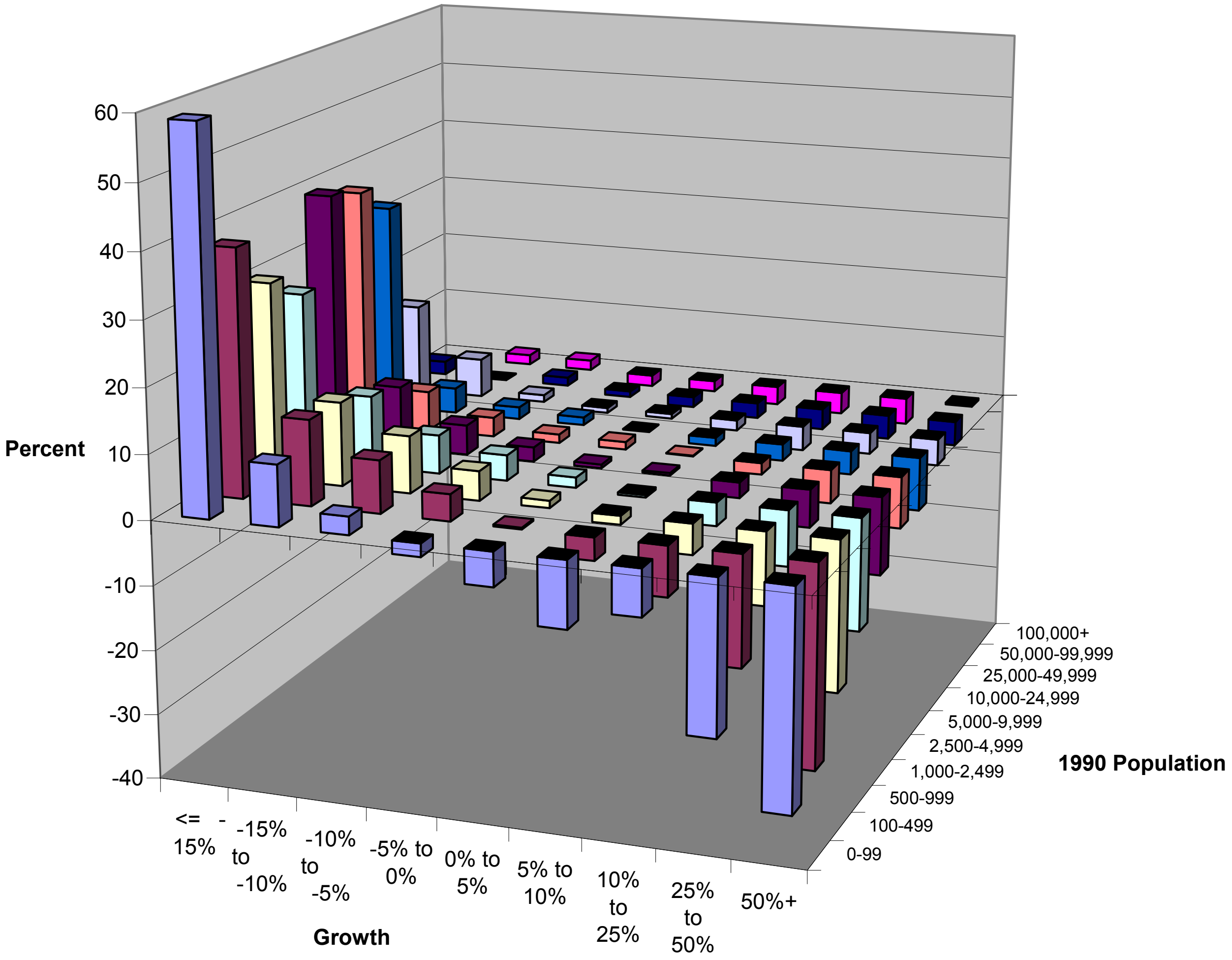


Figure 6.

MALPE by 1990 Population and 1990-2000 Population Growth Rate



GEOGRAPHIC DIFFERENCES

Table 3 depicts a breakdown of estimate accuracy by state. The range of mean errors is quite wide, from a low of 4.9 percent for Massachusetts and Connecticut to a high of 19.5 percent for North Dakota. Because small areas and areas with large population changes are harder to estimate, these results are not surprising. Massachusetts and Connecticut, as well as the rest of New England, contains relatively few small or rapidly changing places, while North Dakota has a very high concentration of small areas. Over 94 percent of the areas in North Dakota have a population less than 1,000.

Table 3. Selected Measures of the Accuracy of Subcounty Population Estimates by State: April 1, 2000

State	Number of Areas	MALPE	MAPE
United States	40,630	3.4	12.4
Alabama	452	6.1	13.9
Alaska	148	0.2	17.8
Arizona	87	0.3	11.6
Arkansas	500	4.0	13.7
California	474	-0.1	7.1
Colorado	269	-1.8	15.3
Connecticut	199	-2.1	4.9
Delaware	57	0.7	17.5
District of Columbia	1	-8.6	8.6
Florida	402	5.5	17.6
Georgia	535	4.9	16.6
Hawaii	1	7.2	7.2
Idaho	201	3.5	16.6
Illinois	2,720	2.7	9.9
Indiana	1,576	6.2	13.0
Iowa	949	1.5	10.0
Kansas	2,119	10.5	18.8
Kentucky	430	6.3	13.6
Louisiana	302	2.2	9.5
Maine	544	0.4	13.8
Maryland	157	1.7	14.0
Massachusetts	396	-1.7	4.9
Michigan	2,064	-0.3	8.4
Minnesota	3,570	2.0	12.2
Mississippi	296	0.7	12.1
Missouri	1,267	4.5	13.5
Montana	129	7.6	12.9
Nebraska	1,028	0.9	11.3
Nevada	19	4.2	13.7
New Hampshire	250	-2.9	7.2
New Jersey	890	0.7	7.3
New Mexico	101	3.1	18.1
New York	1,622	1.4	8.5
North Carolina	540	5.4	16.8
North Dakota	2,114	7.8	19.5
Ohio	2,515	2.7	9.8
Oklahoma	591	2.9	12.7
Oregon	239	-1.6	11.1
Pennsylvania	3,592	0.2	7.8
Rhode Island	47	-4.0	5.0
South Carolina	268	6.6	14.8
South Dakota	1,596	5.0	17.5
Tennessee	349	4.6	13.0
Texas	1,192	3.6	13.0
Utah	235	-1.2	13.2
Vermont	299	-1.0	12.1
Virginia	230	2.5	11.8
Washington	279	-3.6	10.5
West Virginia	234	8.0	13.2
Wisconsin	2,458	8.4	16.8
Wyoming	97	5.4	16.8

Table 4 presents a breakdown of estimate accuracy by size class and type of geography. The MAPE for minor civil divisions (MCDs) was 12.7 percent, compared with 12.1 percent for places. The slightly higher error rate for MCDs was primarily due to the smaller population size of MCDs. When examined by size class, the estimates for MCDs were more accurate for all categories. The accuracy of the MCD estimates was greatly enhanced by inclusion of the towns in New England, as is evident when the MCD estimates are examined by region. MCDs in the Northeast had an average error of only 7.9 percent, compared with 14.5 percent for MCDs in the Midwest.

Table 4. Mean Absolute Percent Error of Subcounty Population Estimates by Type of Error: April 1, 2000

	Incorporated Places		Minor Civil Divisions	
1990 Population	Number of Areas	Mean Absolute Percent Error	Number of Areas	Mean Absolute Percent Error
All	19,426	12.1	21,204	12.7
<100	982	36.4	2,590	34.7
100-499	5,337	16.0	5,253	15.3
500-999	3,270	11.5	3,391	9.7
1,000-2,499	3,679	9.8	4,406	8.0
2,500-4,999	2,067	8.4	2,308	6.7
5,000-9,999	1,619	6.5	1,464	5.8
10,000-24,999	1,359	5.6	1,134	4.8
25,000-49,999	583	4.9	404	4.4
50,000-99,999	327	4.1	178	3.8
>100,000	293	4.6	76	3.5

USE OF ESTIMATES VERSUS DECENNIAL CENSUS COUNTS

Because of the difficulties in producing estimates for very small areas, it is worth asking whether these estimates offer any improvement over the alternative of using population counts from the previous census. The results of this comparison appear in Table 5. For each size class, the 2000 estimates were closer than the 1990 census counts[2] to the Census 2000 results for a majority of cases. This measure ranged from a low of 51.6 percent for areas with populations between 100 and 499 to 75.0 percent for areas with populations between 50,000 and 99,999.

Another approach to the comparison between the 2000 estimate and the alternative 1990 census count is to calculate the mean percent errors for each area, that is, the average percent difference from the Census 2000 count (Table 5, Figure 7). Galdi's evaluation of the 1980 estimates showed that for areas below 500, 1970 census counts were actually preferable to 1980 estimates. The same evaluation was conducted for the 2000 estimates with similar results. For all areas, the MAPE resulting from use of

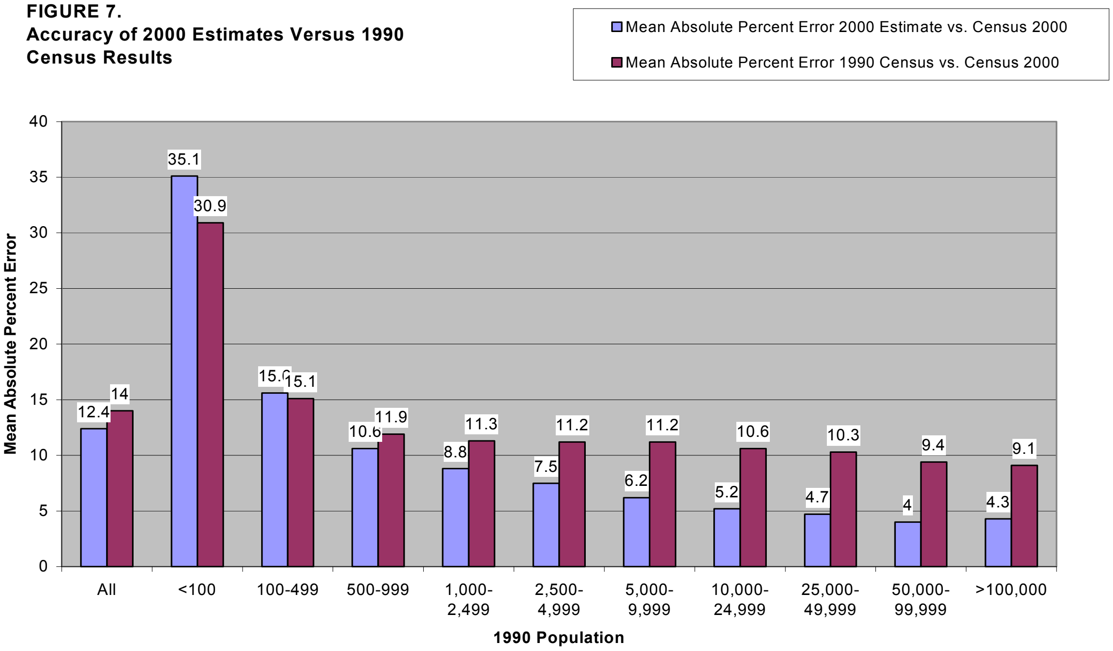
[2] The 1990 census results used in this analysis are the tabulated 1990 census counts revised to reflect legal boundary changes reported nationwide during the 1999 Boundary and Annexation Survey.

the 1990 census counts was 14.0 percent, compared with 12.4 percent for the 1990 estimates. The MAPE for areas of 100,000 and over was 9.1 percent when using the 1990 counts, compared with only 4.3 percent for the 2000 estimates. In fact, the 2000 estimates had an error rate of at least 2.5 percentage points below that of the 1990 census counts for all size classes over 1,000. However, for areas under 500 the mean error of the 1990 census counts was less than the 2000 estimates, implying that holding the 1990 census results constant would have produced a more accurate result. The mean error for the 2000 estimates was 20.6 percent for areas with less than 500, compared with 19.1 percent for the 1990 census counts.

Table 5. Accuracy of 2000 Subcounty Population Estimates versus 1990 Census Results

1990 Population	Number of Areas	Mean Absolute Percent Error (2000 Estimates)	Mean Absolute Percent Error (1990 Census)	Percent of Estimates Preferable to 1990 Census
All	40,630	12.4	14.0	59.2
<100	3,572	35.1	30.9	56.0
100-499	10,590	15.6	15.1	51.6
500-999	6,661	10.6	11.9	55.8
1,000-2,499	8,085	8.8	11.3	60.9
2,500-4,999	4,375	7.5	11.2	64.2
5,000-9,999	3,083	6.2	11.2	67.4
10,000-24,999	2,493	5.2	10.6	71.1
25,000-49,999	987	4.7	10.3	72.0
50,0000-99,000	505	4.0	9.4	75.0
>100,000	279	4.3	9.1	72.8

FIGURE 7.
Accuracy of 2000 Estimates Versus 1990 Census Results



WITHIN-COUNTY DISTRIBUTION

The subcounty estimates are produced by distributing total county household population among constituent parts, then adding in the group quarters population. It is thus appropriate to measure distributive accuracy. The metric used is the Index of Dissimilarity (ID). ID is a simple statistic that removes the effect of the county control population, focusing strictly on the distribution. It also avoids some pitfalls associated with MAPE that are illustrated in Appendix B. Formally, it is as follows:

$$\mathrm{ID} = 50\% \times \sum_{i=1}^{n} \left|\hat{s}_i - s_i\right|,$$

Where :

i indexes the *n* places or MCDs in a county

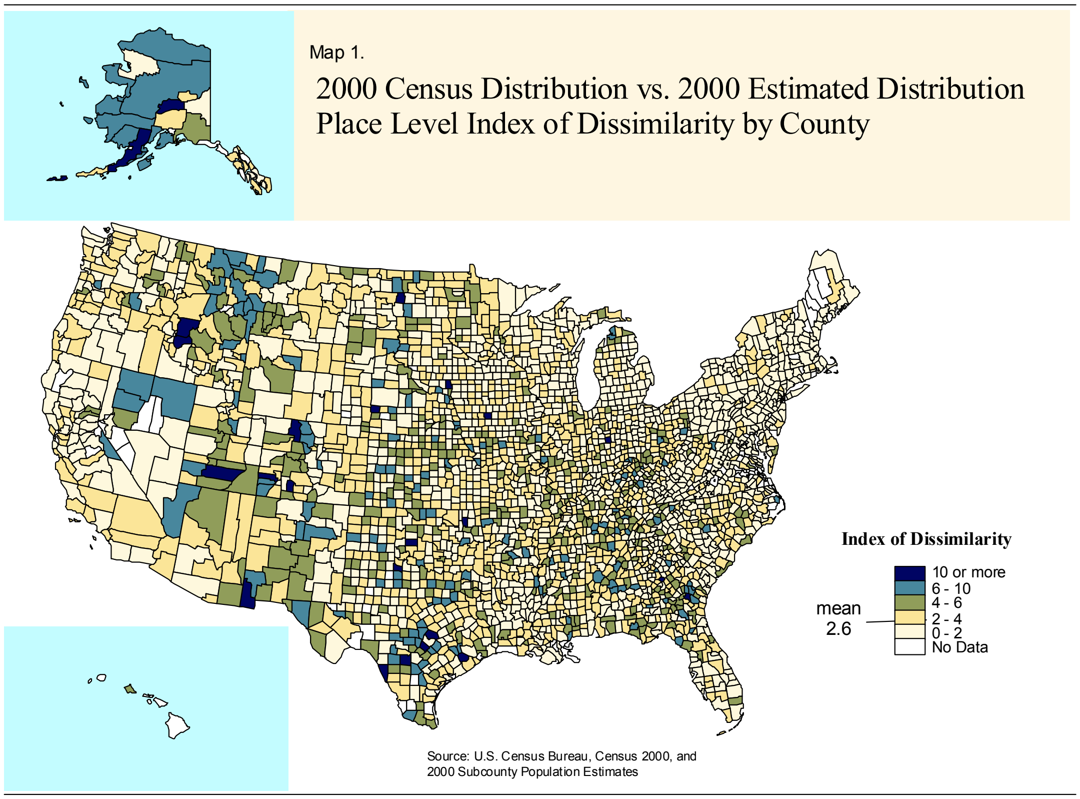
s_i = actual population share

$\hat{s}_i$ = estimated population shares

Map 1 shows that counties with high ID's (greater than 6%) tend to be concentrated in the sparsely populated areas of the West's Mountain division or in several states (Texas, Alabama, Georgia) in the South[3]. Many of the places in counties with high ID's either do not collect or do not report building permit data for new residential construction. Because the Distributive Housing Unit Method uses building permit data as a primary input to produce population estimates, the intra-county household population distribution where building permit data are unavailable is likely to be relatively inaccurate.

[3] The Mountain division consists of Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah and Wyoming. The South region consists of Alabama, Arkansas, Delaware, the District of Columbia, Florida, Georgia, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Texas, Virginia and West Virginia

A similar analysis was conducted for the 20 states that have functioning MCDs and the results are similar to the place level map. The counties that have a large percentage of their population misallocated are concentrated in the rural areas of northern Maine, northern Michigan, Minnesota and the Dakotas.



Map 1.

2000 Census Distribution vs. 2000 Estimated Distribution
Place Level Index of Dissimilarity by County

Source: U.S. Census Bureau, Census 2000, and
2000 Subcounty Population Estimates

58 Counties show no data because there are no incorporated places within their boundaries

REGRESSION ANALYSES

To better understand the causes of estimate errors, we performed regression analyses. As Figure 5 shows, the nonmonotonic U-shaped curves generated by MAPE make regression analysis of accuracy (i.e., the "closeness" of errors to zero) impossible. Furthermore, Figure 6 shows monotonic relationships between bias and 1990 population and the 1990-2000 growth rate. In order to avoid these problems, we use a measure of bias as the dependent variable in our regressions. Four regressions are run for four different sets of geographic areas: all incorporated areas, all MCDs, all intersections of places with counties and all "primitive" areas. The primitive areas form a complete, mutually exclusive partition of the U.S. such that all geographic units for which the Census Bureau produces estimates can be aggregated from them. For the purposes of this analysis, only primitive subcounty areas are used. The independent variables were selected based on *a priori* beliefs about their possibilities of contributing to bias. These included subcounty-specific variables and the county bias measure, when applicable.

To measure bias, we used the natural logarithm of the prediction ratio (LOGQ). The prediction ratio is simply the ratio of the estimated value to the actual value. Its logarithm is zero when the estimate equals the actual value. Not only did these regressions have greater explanatory power than those using the Algebraic Percentage Error (ALPE) as measured by R^2 and lower significance levels of the independent variables, but if the logarithm of the error has zero expectation, then ALPE is biased (Coleman, 2002). A natural distribution for a prediction ratio is the lognormal distribution, which results in an upwardly biased ALPE (Coleman, 2002).

The independent variables in all regressions include the subcounty area's 1990 population (SIZE), 1990-2000 population growth ratio (GROWTH), the product (or interaction) of the previous two variables (SIZE*GROWTH), the 1990 population per household (PPH), the 1990 vacancy rate in percent (VR), the proportion of mobile homes in 1990 (MOBPROP) and the geographic coordinates of the interior point[4] in degrees of latitude North (LATITUDE) and longitude West (LONGITUDE). The last two variables control for long-range spatial effects. In all regressions except those for the incorporated places, the logarithm of the county's prediction ratio (CLOGQ) is included to account for bias caused by bias in the county control. Also, an estimate of the county's 1990 population covered by building permits systems (BPCOV) is included to account for effects of building permit coverage on subcounty estimates.

The regressions must be viewed with some caveats. First, the functional specification may not be correct. While we have strong evidence that the choice of dependent variable is superior to the conventionally used ALPE, the evidence for choice of the independent variables is weaker. From Figure 6 and the regression analyses, the interaction between an area's size and growth rate is an important predictor. However, we did not attempt to include other interactions, which are difficult to diagnose when a large number of predictor variables are present. Finally, we have made no attempt to measure or account for spatial correlations, that is, short-range spatial effects, as we lack the software to do this.

[4] The internal point is a point, not necessarily the centroid, defined by the U.S. Census Bureau in the center of the area's land.

Table 6 contains the regression results. Except for the MCD regression, SIZE and GROWTH have the familiar negative and positive signs, respectively.[5] Their interaction is negative and significant in these cases. In the MCD regression, both SIZE and SIZE*GROWTH are positive and significant. This difference appears to be caused by the magnitude of the coefficient of GROWTH in the MCD regression's being an order of magnitude larger than those of the other regressions. Thus, we conclude that GROWTH is more important to MCD estimates relative to other estimates and that this contribution accounts for the large $\overline{R}^2$ in the MCD regression, about 3 times the size of $\overline{R}^2$s in the other regressions. Of all other variables, only LATITUDE has similar coefficients across regressions. PPH and MOBPROP have positive coefficients, while those of VR and LATITUDE are negative. The coefficient of PPH for primitive areas is about 10 times those of the other regressions. The coefficients of VR of MOBPROP in the MCD regression are about one-half and one-third, respectively, of those in the other regressions. LONGITUDE is interesting: it is negative in the regressions for places, indicating that the prediction ratio falls in the west, while it is positive in the other regressions. This appears due to differences in the place and MCD estimates. The primitive geography contains a large number of MCDs and their intersections with places, so that it behaves more like the MCD regression. CLOGQ is significant and its coefficient ranges from 0.69 to 0.86, indicating, unsurprisingly, that county error is the dominant contributor to subcounty error. Interestingly, the equation for places, which excludes county variables, has the smallest $\overline{R}^2$ (0.132), but one which is not much smaller than those of the other non-MCD regressions (0.147 and 0.154). This contradictory result is difficult to interpret and requires further investigation.

[5] For example, see Smith (1987), Davis (1994), and Tayman, Schafer and Carter (1998) for empirical evidence and Beaumont and Isserman (1987) for theoretical explanations.

Table 6. Regression Results for Subcounty Population Estimates: April 1, 2000

	Universe			
Variable	Incorporated Places	Intersections of Incorporated Places and Counties	Minor Civil Divisions	Primitive Areas
Intercept	0.07104 (4.75)***	0.05189 (2.25)**	0.28289 (14.85)***	0.01937 (1.03)
SIZE	4.6144E-7 (1.75)*	1.34E-6 (3.63)***	-3.37E-6 (5.84)***	2.48E-6 (6.71)***
GROWTH	-0.08926 (49.12)***	-0.0341 (52.91)***	-0.35308 (108.46)***	-0.02384 (71.00)***
SIZE*GROWTH	-4.5655E-7 (1.88)*	-1.29E-6 (3.77)***	3.25E-6 (5.78)***	-2.35E-6 (6.86)***
PPH	0.0492 (12.80)***	0.05315 (10.28)***	0.05777 (17.76)***	0.4825 (12.57)***
VR	-9.3291E-4 (7.22)***	-0.00101 (5.67)***	-4.678E-4 (6.99)***	-0.00132 (13.83)***
MOBPROP	0.40343 (11.84)***	0.38413 (8.32)***	0.13739 (3.08)***	0.32128 (7.71)***
LATITUDE	-7.3826E-4 (2.72)***	-0.001 (2.61)***	-0.00111 (2.76)***	-0.00142 (4.52)***
LONGITUDE	-7.0563E-4 (6.47)***	-6.6464E-4 (4.33)***	2.8725E-4 (2.45)**	2.9807E-4 (2.48)**
BPCOV		-2.7032E-4 (2.11)**	-3.1964E-4 (3.67)***	-6.63121E-4 (6.27)***
CLOGQ		0.86615 (16.80)***	0.69463 (18.61)***	0.85901 (19.82)***
Number	19,423	20,229	21,202	34,026
$\bar{R}^2$	0.132	0.147	0.398	0.154

* Significant at 10%
** Significant at 5%
*** Significant at 1%

Note: See text for variable definitions.

CONCLUSIONS

This evaluation found that the 2000 estimates had improved accuracy when compared with the 1980 estimates. The accuracy of the estimates improves for larger places and places with small amounts of change. For places larger than 5,000, the 2000 estimates are clearly superior to the use of the 1990 decennial census counts.

For places under 1,000 the estimates do not fare as well. Places below 1,000 in population represent over half of all subcounty areas in the country, yet contain only 1.3 percent of the total population. The mean error for the 2000 estimates for places under 1,000 was 17.4 percent compared to 16.8 percent for the 1990 census. These results raise the question whether the continued production of estimates for small places can be justified.

APPENDIX A: ESTIMATES METHODOLOGY

The following equation describes how housing unit estimates for 2000 were calculated.

$$\mathrm{HU00} = \mathrm{HU90} + \mathrm{NC00} + \mathrm{NPC00} + \mathrm{NM00} - (\mathrm{PL00} + \mathrm{NPL00})$$

Where:

HU00	=	Estimated 2000 Housing Units
HU90	=	1990 Census Housing Units
NC00	=	Estimated New Permitted Construction: 4/1/90 – 4/1/00
NPC00	=	Estimated Nonpermitted Construction: 4/1/90 – 4/1/00
NM00	=	Estimated New Mobile Home Placements: 4/1/90 – 4/1/00
PL00	=	Estimated Permitted Housing Unit Loss: 4/1/90 – 4/1/00
NPL00	=	Estimated Nonpermitted Housing Unit Loss: 4/1/90 – 4/1/00

Estimated Permitted Construction (NC00) -- Building permits are compiled from internal data files developed by Manufacturing and Construction Division (MCD). These files include imputed permits where a local jurisdiction did not report permit issuance for the entire year. Housing growth calculated from building permits employs a census region-specific lag time (weighted average by number of units in structure) between the issuance of permits and completion of construction. Two percent of all building permits never result in the actual construction of a housing unit (as derived from the U.S. Census Bureau, Current Construction Reports, Series C20-9103 and Series C22-9107). Therefore, a factor of 0.98 is used to estimate completed new units.

Estimated Nonpermitted Construction (NPC00) -- Not all subcounty areas are permit issuing jurisdictons. Estimates of nonpermitted residential construction are calculated by applying 1990 Census data on units in jurisdictions that do not issue building permits to an updated national estimate of nonpermitted construction taken from the Survey of Construction.

Estimated New Mobile Home Placements (NM) -- The Population Estimates Branch receives state mobile home shipment data from the MCD. The state mobile home shipments are allocated to subcounty areas based on the subcounty area's share of state mobile homes in the 1990 Census.

Estimated Permitted Housing Loss (PL00) -- Demolition permits were compiled from internal data files developed by MCD. These files include imputed permits where a permit-issuing jurisdiction did not report permit issuance for the entire year. No lag time is assumed for demolition permits.
MCD stopped collecting demolition permit data in 1995.

Estimated Nonpermitted Housing Unit Loss (NPL00) -- Estimates of nonpermitted housing unit loss (NPHL00) were calculated by applying data taken from the 1993 Components of Inventory Change

Survey (CINCH) to 1990 Census data. Data from the CINCH survey indicated that the following types of structure were at a greater risk of loss:

1. Mobile homes and other
2. Older units (pre-1939 construction)
3. Vacant for seasonal and recreational use
4. Boarded up

Note: In essence, this process operates by applying a crude death rate to the housing inventory to develop a baseline estimate of housing unit loss. This crude death rate is refined by the type of housing stock contained within each subcounty area.

The following equation describes the calculation of an uncontrolled household population estimate given a new estimate of housing units.

$$UCHHP00 = HU00 * OCR90 * PPH90$$

Where:

UCHHP00	= Estimated 2000 Uncontrolled Household Population Estimate
HU00	= Estimated 2000 Housing Units
OCR90	= 1990 Occupancy Rate
PPH90	= 1990 Population per Household

1990 Occupancy Rate (OCR90) -- The occupancy rate is calculated by dividing the number of occupied units in the locality by the total number of units as reported in the 1990 Census.

1990 Persons per Household (PPH90) -- The number of persons per household is obtained by dividing the household population as reported in the 1990 Census by the number of occupied housing units in 1990.

The final step in producing a population estimate using the Distributed Housing Unit Method is controlling the uncontrolled subcounty estimates to the published county totals. The following equation describes the calculation of a controlled estimate:

$$SCEST00 = [UCHHP*(CHP00/SUCHHP00)] + GQ00$$

Where:

SCEST00	= Final 2000 Subcounty Population Estimate
UCHHP00	= Uncontrolled 2000 Household Population Estimate
CHP00	= Published County 2000 Household Population Estimate
SUCHHP00	= County Sum of UCHHP00 for all Subcounty Areas
GQ00	= 2000 Group Quarters Population Estimate

Published County Estimate (CHP00) -- The published county population estimate as calculated by the Tax Return Method for the current estimate year.

County Sum of Uncontrolled Household Population Estimates (SUCHHP00)—The county sum of the uncontrolled county population is obtained by summing the estimates for all subcounty areas within a county.

Group Quarters Population (GQ00) -- This component is primarily a combination of military personnel living in barracks, college students living in dormitories and persons residing in institutions. Inmates of correctional and juvenile facilities and persons in health care facilities and Job Corps Centers are also included in this category.

At various stages of the production process data are made available to each state's FSCPE representative for review. In a limited number of cases the FSCPE's data were used to distribute the county population instead of the housing unit estimates calculated by the Census Bureau.

APPENDIX B: The Inappropriateness of MAPE and Similar Measures Within Counties

MAPE is well known as favoring underestimates over overestimates.[6] (Armstrong, 1985, p. 348) The absolute percentage error (APE) associated with an underestimate is at most 100 percent, while the APE associated with an overestimate is unbounded. In the subcounty context, this problem is amplified by the leverage possessed by small areas. Table B-1 illustrates this problem. Table B-1 contains data for a hypothetical two-county example, in which each county has two areas and the same total population: 10,000.

TABLE B-1

Data for Two Hypothetical Two-Area Counties

	County 1		County 2	
	Area 1	Area 2	Area 1	Area 2
Actual Value	1	9999	5000	5000
Share of Total	0.0010	0.9999	0.5000	0.5000
Estimate	2	9999	10000	10000
Share of Total	0.0020	0.9999	0.5000	0.5000
Absolute Percentage Error	100%	0%	100%	100%
MAPE	50%		100%	
Index of Dissimilarity	0.05		0	

In this example, County 2's distribution is estimated perfectly. Thus, its Index of Dissimilarity (ID) is zero. However, because the total population is grossly overestimated (100 percent), its MAPE is twice that of County 1. Thus, MAPE picks up only the error in estimating the county total, instead of errors in distributing that total to subcounty areas. Table B-1 also shows the leverage associated with small counties. In County 1, Area 1 is overestimated by 1 person, resulting in an estimate that is double its actual population. This error dominates the perfect estimate of Area 2's population. It should be noted that ID is not invulnerable to this problem. A different statistic, such as φ^2 could be used to remedy the leverage problem, at the cost of interpretability and a more difficult statistic to compute.[7]

In this example, MAPE is identical to the Mean Algebraic Percentage Error (MALPE), a measure of mean bias. Thus, the same criticisms that apply to MAPE apply to MALPE.[8]

[6] It is, however, inappropriate to speak of these measures as being biased without understanding the probability distributions of the variables being measured and how the values of the measures relate to the variables.

[7] The formula for φ^2 is $\varphi^2 = \sum_{i=1}^{n} \left(\hat{s}_i - s_i\right)^2 / s_i$. It is closely related to the χ^2 goodness-of-fit statistic.

[8] Moreover, we would expect the standard deviation of the algebraic percentage errors to increase in the same manner as MALPE, reducing the significance level of the *t*-ratio. Thus, one would not expect the increase in MALPE to have any effect on tests for mean bias per Coleman (1999).

REFERENCES

Armstrong, J. Scott (1985), *Long-Range Forecasting: From Crystal Ball to Computer*, New York: Wiley.

Coleman, Charles D. (1999), “Nonparametric Tests for Bias in Estimates and Forecasts,” in *American Statistical Association: 1999 Proceedings of the Business and Economic Statistics Sections*, Alexandria, VA: American Statistical Association, 251-256.

Coleman, Charles D. (2002), “New Diagnostic Statistics for Mean Bias in Positive Predictions,” manuscript, U.S. Census Bureau.

Galdi, David (1985), Evaluation of 1980 Subcounty Population Estimates, U.S. Census Bureau, Current Population Reports, Series P-25, No. 963, U.S. Government Printing Office

Davis, Sam T., (1994) “Evaluation of Postcensal County Estimates for the 1980s,” Population Division Working Paper No. 5, U.S. Census Bureau, Washington, DC.

Beaumont, Paul M. and Andrew M. Isserman (1987) “Tests of Accuracy and Bias for County Population Projections: Comment” *JASA* **82**, 1004-1009.

Smith, Stan (1987) “Tests of Accuracy and Bias for County Population Projections” *Journal of the American Statistical Association* **82**, 991-1003.

Tayman, Jeff, Edward Schafer and Lawrence Carter (1998), “The Role of Population Size in the Determination and Prediction of Population Forecast Errors: An Evaluation using Confidence Intervals for Subcounty Areas,” *Population Research and Policy Review* **17**, 1-20.