The Effects of Changes in Geographic Coding Methods on Estimated Administrative Record Coverage, Migration Rates, and Estimates

Rodger V. Johnson, Esther Miller, Hyo Park, Barbara van der Vate
Administrative Records and Methodology Research Branch
Population Division
U.S. Census Bureau

Introduction

The U.S. Census Bureau annually estimates domestic migration change among the Nation's states and counties[1] as a component to the demographic change model. The Census Bureau acquires selected administrative data from the Internal Revenue Service (IRS)[2] to estimate migration for the under-65-age population, which is the largest component of change in most counties. The assumption underlying the use of the data is that they are representative of the under-65-age population. Two consecutive years of matched IRS returns are used to calculate both gross and net migration rates for each state and county.

Improved estimates of coverage and more accurate migration rates are influenced by the quality of state and county geographic (geo-) codes assigned to each matched set of tax returns. The Census Bureau continues pursuing better ways to improve the methodology to assign these geographic codes. For example, the Census Bureau's method from 1995 to 2005 relied upon a system that matched the IRS returns via their associated ZIP+2 postal codes to the appropriate state and county. The file was updated annually and provided a comfortable level of accuracy, however, it also required extensive and time-consuming preparation by the Population Division. While this file continued in use, Census staff were planning a new method to increase the geo-coding precision, reduce processing time, and rely on the capabilities of the Census Bureau's (Topologically Integrated Geographic Encoding and Referencing) TIGER® System, the Census Bureau's geographic base.

This paper presents a brief summary of selected findings from the testing of the new geo-coding system and its positive effects upon estimated coverage and migration rates at the state and county levels. The IRS returns filed in 2004 were geo-coded with both the original (ZIP+2 based) system and the new (ZIP+4 based) system and the results compared to assess impacts upon estimated coverage, migration rates, and population estimates. Testing was conducted with data from IRS returns filed in 2003 and 2004 and matched for use in the 2004 population estimates.

Building the ZIP+4 Based System

The Census Bureau's Geography Division produced a ZIP+4 coding file to meet migration process requirements. The coding file mapped ZIP+4 postal codes to official

[1] There are 3,141 counties (including statistical equivalents) in the 2004 Estimates universe.
[2] Includes street address, including ZIP+4, and number of exemptions on each 1040 return.

state and county FIPS (Federal Information Processing) geo-codes that were controlled by the geo-relationships in the TIGER® System and associated Master Address File (MAF). Expanding the coding file to include the last two digits of the postal code was anticipated to increase the level of precision in assigning the geo-codes to each IRS return.

Assigning State and County Geo-codes

The new geo-coding methodology uses the entire ZIP+4 code on each tax return in order to assign it to a state and county. Using the ZIP+4-based system, two sequential years of matched IRS returns are individually coded to a county consistent with their addresses. Any matched returns whose county codes have changed are defined as *migrants* and become part of the migration rates calculated with the exemptions. The exemptions on those returns whose county codes are unchanged are defined as *non-migrants* and remain in the migration base.

Assessing Impact upon Demographic Coverage Estimates

Under both the new (ZIP+4) and old (ZIP+2) geo-coding methods, the estimated average national IRS coverage for the under-65-age population is 81 percent[3] and there have been improvements noted at the county level under the new method. There is a reduction in the number of counties with extreme over-coverage and under-coverage estimates, bringing county coverage rates closer to the national average.

The median coverage rate for both the new and old geo-coding methods is 84.0 percent. The number of counties whose estimated coverage of the population deviated by more than 10 percentage points from the median decreased from 659 counties to 317 counties. Under the ZIP+4 geo-coding system, 89.9 percent of all counties (2,824) have a coverage rate between 74 and 94 percent of the county's population.

These facts demonstrate that the new system is providing better-balanced coverage estimates for more counties than ever before while also reducing the variance among them. This is turn can be interpreted to mean that on average, *the under-65-age population in more counties is achieving the important attribute of uniform representation by the IRS return universe*.

Assessing Impact on the Migration Estimates

The ZIP+4 based system produces migration rates, and subsequent numbers of net migrants that are comparable to those of the ZIP+2 based system. We observed that the migration rates of less than 100 counties had a difference of one percentage point or more, while only eight counties had a difference of two percentage points or more. Stability in the results is a strong sign that we achieved one of our major goals - - that of reducing nonstandard revisions in the migration rates under the new system.

---

[3] Demographic coverage estimates are based upon comparison with the 2004 estimates.

Assessing Impact on the Population Estimates for the Under-65-age Population

We compared a set of 2004 estimates derived from the ZIP+4 based migration rates to the official estimates (produced by the ZIP+2 system). In order to ensure that estimates of domestic migration remained stable, the results had to be demonstrably similar to those produced earlier in the decade with the ZIP+2 system. The effect upon state level estimates of the population under age 65 with the ZIP+4 based migration rates was observed to be negligible, as expected.

At the county level more changes were evident. This outcome also was anticipated, given the changes in the estimates of coverage and the impacts of increasing or decreasing denominators used as the basis of calculation of the migration rates. We determined that the population estimates differed by less than 100 people in 87 percent of all counties and that only two percent (63) of all counties differed by more than 500 people under age 65.

Summary

The Census Bureau had determined that moving to a new geo-coding system should increase efficiency and reliability, while at a minimum, demonstrating comparable if not more accurate results. The test results have met those goals, while substantially improving the distribution of estimated under-age-65 coverage among counties. Relying on the capabilities introduced by updates to TIGER® and the MAF, in combination with U.S. Postal Service data, demonstrates that the new system will sustain the requirements of the Population Estimates Program (PEP). Commencing with the 2005 population estimates, the PEP is using the ZIP+4 based system to support its migration processing. As improvements continue in Census Bureau geo-processing and address-matching capabilities, it is anticipated that the ZIP+4 based system will serve the intended purpose well into the next decade.