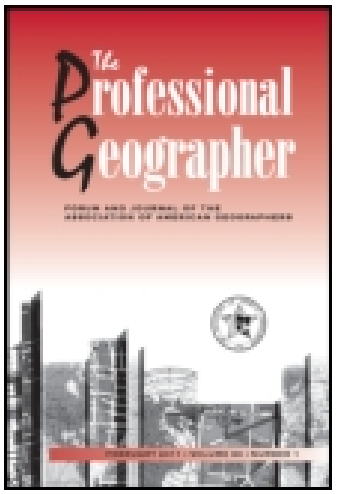


The Professional Geographer
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rtpg20

# Interpolating U.S. Decennial Census Tract Data from as Early as 1970 to 2010: A Longitudinal Tract Database

John R. Logan[a], Zengwang Xu[b] & Brian J. Stults[c]
[a] Brown University
[b] University of Wisconsin, Milwaukee
[c] Florida State University
Published online: 13 May 2014.



To cite this article: John R. Logan, Zengwang Xu & Brian J. Stults (2014) Interpolating U.S. Decennial Census Tract Data from as Early as 1970 to 2010: A Longitudinal Tract Database, The Professional Geographer, 66:3, 412-420, DOI: 10.1080/00330124.2014.905156

To link to this article: http://dx.doi.org/10.1080/00330124.2014.905156

# Interpolating U.S. Decennial Census Tract Data from as Early as 1970 to 2010: A Longitudinal Tract Database

**John R. Logan**
*Brown University*

**Zengwang Xu**
*University of Wisconsin, Milwaukee*

**Brian J. Stults**
*Florida State University*



Differences in the reporting units of data from diverse sources and changes in units over time are common obstacles to analysis of areal data. We compare common approaches to this problem in the context of changes over time in the boundaries of U.S. census tracts. In every decennial census, many tracts are split, consolidated, or changed in other ways from the previous boundaries to reflect population growth or decline. We examine two interpolation methods to create a bridge between years, one that relies only on areal weighting and another that also introduces population weights. Results demonstrate that these approaches produce substantially different estimates for variables that involve population counts, but they have a high degree of convergence for variables defined as rates or averages. Finally, the article describes the Longitudinal Tract Database (LTDB), through which we are making available public-use tools to implement these methods to create estimates within 2010 tract boundaries for any tract-level data (from the census or other sources) that are available for prior years as early as 1970. **Key Words: 2010 Census, areal interpolation, census geography, census tract, population interpolation.**

各种来源与随着时间变化的单位中, 数据报告单位的差异, 是地区数据分析中的普遍障碍。我们在美国人口普查区的边界随着时间改变的脉络中, 比较解决此一问题的常见办法。在每十年的人口普查中, 诸多普查区原本的边界被分离、合併或改变, 以反映人口的成长或减少。我们检视两种在不同年度之间建立连结的内插法, 其中一个仅仅依赖面积加权, 另一个则同时引进人口加权。研究结果显示, 这些方法, 对于涉及人口计算的变异数而言, 产生了显着的差异评估, 但它们对于定义为比率或平均数的变异数而言, 则具有高度的收敛。本文最后描绘纵向普查区数据集 (LTDB), 我们藉此创造可提供执行这些方法的公共工具, 以在 2010 年的普查区边界之中, 评估任何可取得早先年度、且最早可溯及 1970 年 (来自普查或其他来源) 的普查区层级数据。 **关键词: 2010 年人口普查, 面积内插, 人口普查地理, 人口普查区, 人口内插**。

Las diferencias en las unidades que reportan datos de diversas fuentes y los cambios en las unidades a través del tiempo son obstáculos comunes en el análisis de datos espaciales. Comparamos los enfoques corrientes que se usan para enfrentar este problema dentro del contexto de cambios en los límites de los distritos censales de los EE.UU. a través del tiempo. En cada censo decenal aparecen subdivididos muchos de esos distritos, o consolidados, o de otro modo se cambian las delimitaciones anteriores para reflejar el crecimiento de la población o su declinación. Examinamos dos métodos de interpolación para hacer puente entre diferentes años, uno que se basa solamente en el peso espacial, y otro método que también incluye pesos de población. Los resultados demuestran que estos enfoques generan estimativos sustancialmente diferentes para variables que involucran conteos de población, pero que tienen un alto grado de convergencia para variables definidas como tasas o promedios. Por último, el artículo describe la Base de Datos Longitudinal de los Distritos (LTDB), a través de la cual habilitamos herramientas de uso público con las cuales implementar estos métodos para crear en 2010 estimativos, dentro de los límites de distrito, para cualquier tipo de datos disponibles para años anteriores hasta 1970, a nivel de distrito (a partir de los censos o de otras fuentes). **Palabras clave: Censo de 2010, interpolación espacial, geografía censal, distrito censal, interpolación de población.**

A common situation faced by researchers using areal data is discrepancies in the boundaries of reporting units. For example, population data might be reported in census tracts, whereas crime data might be reported in police precincts, or election data in voting districts, or school data in school attendance zones. Another example is when there are changes over time in the boundaries of the same units (Martin, Dorling, and Mitchell 2002). In either case, the general problem is how to harmonize data to the same geographic unit so that information from different sources and times can be analyzed together. Social scientists sometimes avoid the issue of boundary changes by simply comparing the cross-sectional pattern of results in one year with another year. This is not possible where the purpose is to study the changes in the characteristics of specific places (however these are defined); shifting boundaries introduce greater likelihood of

The Professional Geographer, 66(3) 2014, pages 412–420 © Copyright 2014 by Association of American Geographers.
Initial submission, October 2011; revised submission, January 2012; final acceptance, March 2012.
Published by Taylor & Francis Group, LLC.

drawing the wrong conclusions. At the least, one would want to know what changes are due to new boundaries and what changes have occurred within the places as previously bounded and to be able to identify tracts where the estimates of change are susceptible to greater error.

## Dealing with Boundary Changes Using Interpolation Methods

We deal here with boundary changes, although similar principles should apply to interpolation of data from different sources. To transfer data from a source to a target zonal system, sophisticated areal interpolation methods usually use ancillary information or statistical methods to refine the source data to a more detailed or finer spatial scale and then reaggregate these data to the target zones. Surface modeling techniques interpolate the source data into an underlying smooth surface that can then be aggregated to target zones (Bracken and Martin 1995). The surface can be estimated by point-based interpolation methods using centroids as the representatives of zones (Bracken and Martin 1989; Martin 1989) or other statistical methods (Kyriakidis 2004; Kyriakidis and Yoo 2005). Data from the surface can be aggregated to any desired areal unit. A criticism of this approach is that population characteristics are not likely to fit a smooth surface. It is common to find discrete boundaries, such as certain major streets or nonresidential zones, where a population variable is discontinuous. In the evolution of minority neighborhoods in the United States, for instance, observed processes of invasion and succession often were associated with specific locations, whose street boundaries were well known but tended to expand over time.

Another current approach is to apply dasymetric (or intelligent) interpolation methods (Wright 1936; Mennis 2003; Maantay, Maroko, and Herrmann 2007; Reibel and Agrawal 2007; Sleeter and Gould 2007; Tapp 2010; Zandbergen and Ignizio 2010). The idea is that simple areal interpolation (Goodchild and Lam 1980) can be improved by using other sources of data about the distribution of the population in the source zone. One type of ancillary data is land use information from remote sensing that can identify areas with no population (Eicher and Brewer 2001). Xie (1995) and Reibel and Bufalino (2005) used information about the road network as indirect indicators of population density. Gregory and Ell (2005) discussed the use of parish population records as the ancillary data for historical interpolation in Britain. Ancillary data can also be zero-dimensional point data. Zhang and Qiu (2011) used schools to estimate a density surface as ancillary data in areal interpolation of population from census tracts to postal zones in Texas. More generally, Goodchild, Anselin, and Deichmann (1993) suggest the use of control zones, areas that are known based on external information to be internally homogeneous on the attribute in question to improve areal interpolation.

## Applications of Interpolation to Boundary Changes in U.S. Census Tracts

Prior to every U.S. census it is the prerogative of state and local officials to identify small areas for which they wish to receive census population totals for electoral redistricting purposes and for other planning and policy functions. As a result, the fundamental units (census blocks and tracts) defined in the previous census could be split or consolidated, and their boundaries could be altered in complex ways. We use areal interpolation to estimate population characteristics of U.S. census tracts from prior years within 2010 boundaries. Depending on what information is available at a very local scale, the interpolation is based on a combination of area and population weighting (2000) or only on area weighting (1970–1990). The former approach is the current standard in the field, but where appropriate small unit population data are not available, areal interpolation is the fallback option (as in Gregory's [2002] harmonization of nineteenth-century British data to contemporary boundaries). Additionally, we take advantage of ancillary data, using a water layer to identify locations with no land area (and therefore no population).

The following sections offer an overview of boundary changes between census years in the United States, outline two approaches to interpolating data to adjust for these changes and compare it to the methods used in the commercially available Neighborhood Change Database (NCDB; Tatian 2003) for 1990–2000, assess the differences in estimates from these different approaches, and introduce our own Longitudinal Tract Database (LTDB) tool for researchers who work with census data.

We begin by describing the changes in census geography that need to be considered in any intercensal bridging system. Examples are demonstrated by tract boundary changes between 2000 and 2010. There are three main categories of changes: consolidations, splits, and complex changes. These are illustrated in Figure 1 for several tracts in the Kansas City metropolis. Consolidation creates no difficulties for analysis; in this example, data for three tracts in 2000 can simply be combined into a single tract as defined in 2010. A split adds difficulty. In this example, some rationale is needed to allocate data from one tract (053106) into three new tracts formed within it.

More complex changes are shown in the right panel of Figure 1. First, the western and southern boundaries of tract 013401 have been adjusted, which means that some population needs to be exchanged between adjacent tracts. Note that some of these changes appear to be very small, and these likely reflect routine technical improvements in the geographic information system (GIS) file. The Census Bureau makes many minor corrections to the digitizing of tract boundaries between decennial census years. The section removed from this tract's southwest corner could be more significant, however. In addition, what used to be two tracts to the east of 013301 have been reorganized

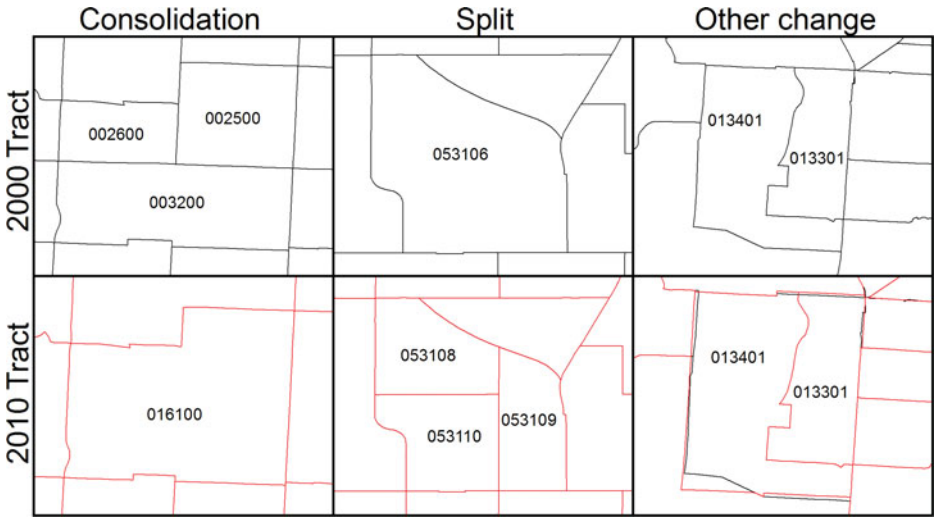


**Figure 1** *Three types of boundary changes in the Kansas City metropolis from 2000 (in black) to 2010 (in red). (Color figure available online.)*

into three, retaining the outer boundaries of the original two but entirely disregarding the prior boundary between them. Nationally for 1990–2000, Tatian (2003) reported that about 80 percent of tract boundary changes were of this latter type (which he described as "many to many" changes), and splits were most of the remaining cases. Our own estimate (see later) is that these two types were about equally prevalent in 2000–2010, if we remove tiny boundary shifts from consideration.

The distribution of tract changes is reviewed in detail for 1990–2000 and 2000–2010 in Table 1, which shows the number of tracts that did not change, tracts that were consolidated from many tracts to one, tracts that were split from one into more than one, and complex types of change that involved multiple tracts in both years. For the purpose of Table 1, we treat as "no change" those cases where the difference in boundaries between a tract in Year 1 and Year 2 involves less than 1 percent of the land area of the Year 2 tract. Of the 72,739 tracts with land area in 2010, we classify 50,062 tracts as unchanged in 2010, though about one third of these (17,898) experienced slight boundary corrections. Of tracts with changes, only a small number of cases (less than 1,000) are consolidations, which pose no problem for interpolation. The most common types of change are those where some form of estimation is required. In over 17 percent of cases a single tract in 2000 was split into more than one tract in 2010, and most of these were the result of one-to-two splits. A nearly equal number of 2010 tracts fall into the "many to many" category, where multiple tracts in 2000 were reconfigured to produce a different set of tracts in 2010. The distribution is similar in 1990–2000.

***Table 1*** *Census tract boundaries over time: Number of tracts experiencing various types of changes between 1990–2000 and 2000–2010*

| Type of change | From 1990 to 2000 | | From 2000 to 2010 | |
|---|---|---|---|---|
| No change | 43,507 | 66.6% | 50,062 | 68.8% |
| Many to one | 969 | 1.5% | 999 | 1.4% |
| One to two | 5,962 | 9.1% | 9,288 | 12.8% |
| One to three | 1,722 | 2.6% | 2,013 | 2.8% |
| One to four or more | 1,005 | 1.5% | 1,267 | 1.7% |
| Many to many | 12,144 | 18.6% | 9,110 | 12.5% |
| Total | 65,309 | 100.0% | 72,739 | 100.0% |

We illustrate the extent and location of these changes in Figure 2, which presents an overlay of 2000 and 2010 tract boundaries in the Kansas City metropolis. There were a number of consolidations, particularly in central city areas of Kansas City, Missouri, that were losing population and a larger number of splits located mainly in outer suburban areas.

### *Combining Areal and Population Interpolation*

Bridging between 2000 and 2010 is greatly facilitated by the Topological Faces layer of the TIGER/Line shapefiles created by the U.S. Census Bureau (2011), which shows the intersection between blocks and tracts (and many other geographic layers) as defined in the 2000 and 2010 censuses. This file is available to be downloaded (http://www.census.gov/geo/www/tiger/tgrshp2010/documentation.html). U.S. census geography includes several nested scales, of which the most commonly used are the state, county, census tract, block group, and block. The face polygons created by the intersection of these multiple geographic boundaries are in effect the smallest possible sub-block unit in census geography, which we term a *fragment*. Each one is uniquely identified by a topological face ID (TFID), and it includes several useful attributes: total area, an indicator of whether the face polygon is water or land, and all geocodes (from block ID to state federal information processing standards code) in both the 2000 and 2010 census. We work with the

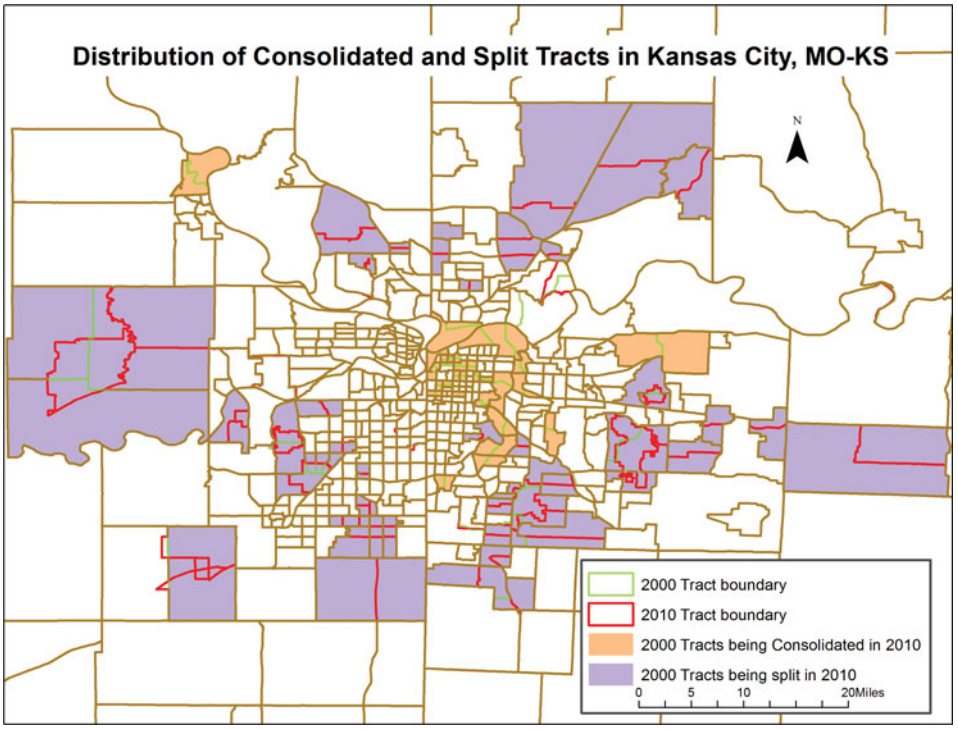


**Figure 2** *Overlay of tract boundaries in 2000 and 2010 in the Kansas City, Missouri–Kansas Metropolitan Statistical Area. (Color figure available online.)*



fragments from the Faces file that can be dissolved to the tract and block layers for 2000 and 2010.

The next step is to allocate reported tract-level population characteristics from 2000 (e.g., counts by race and age) to blocks within the tract. Our LTDB bases this allocation on the block's share of the total tract population in 2000. It then estimates what share of the 2000 block population and of every population subgroup (estimated in the previous step) lies in each fragment within that block. It does this through simple areal interpolation based on the fragment's share of the block area. This estimate is refined with ancillary data provided in the Faces file that identifies water fragments with no population that should be disregarded.

It is straightforward to aggregate fragments to the 2010 census tracts. The assumption that all population characteristics have the same distribution as the total population across blocks within a tract, and across fragments within a block, is the main source of error in the estimate. It would be desirable to use additional information sources to refine the allocation of block populations to the fragments of a block that are in different 2010 tracts. This can be important because blocks are often reconfigured between censuses. NCDB used ancillary data from the streets coverage from Tiger/Line 1992 to bridge 1990 data to 2000 tract boundaries. Every 1990 block was linked to census tracts in 2000. When the block was fully within the boundaries of the 2000 tract, its 1990 population was used as the population weight. When the block was located in more than one 2000 tract, the length of streets within each fragment was used to determine what share of the block population to allocate to each tract. The assumption is that population is highly correlated with the extent of local roads, although it was not known whether there were homes on these roads. To the extent that roads indicate population, this procedure is superior to weighting block fragments by their area. NCDB created a 1990–2000 proprietary Block Weighting File (BWF) to represent what share of a given 1990 block's population should be estimated to fall within each 2000 tract. As in the LTDB, these same weights were used to estimate all census variables.

### *Interpolation with Area Weights*

Areal interpolation requires only that we have an accurate overlay of the tract boundaries in two years. The LTDB estimates for 1970–1990 are based on tract boundaries from the National Historic Geographic Information System (NHGIS). With these we created a tract-level equivalent of a Topological Faces relationship table for 1970–2000. The first step is to overlay the 2000 tract boundary file onto the 1990 boundary file and merge these into a single layer. For each tract that did not change between 1990 and 2000, the result is a single polygon and data record. For tracts that changed, multiple records exist in the new layer. We then merge 1990 census data with this new layer using 1990 state, county, and tract codes, and we apportion the 1990 counts to each fragment of the split tract using the area proportions as weights.

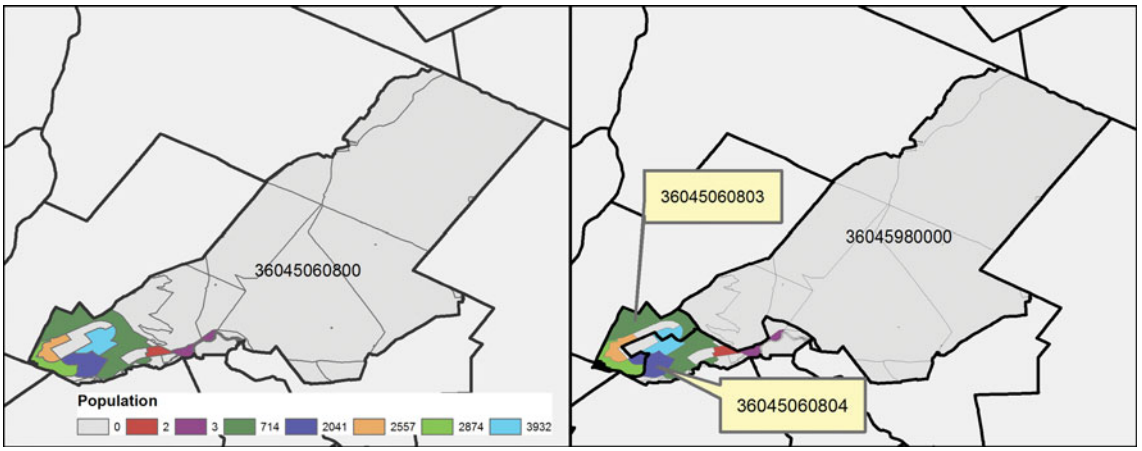


**Figure 3** *Example of a split tract in 2000–2010, showing the block populations (color-coded) in 2000 in each panel. On the left are the 2000 tract boundaries; on the right are the 2010 boundaries. (Color figure available online.)*

We repeat the same process for 1970 and 1980, again using the 2000 tract file as the overlay. We then use the population and area-based interpolation method described previously to adjust the data from 2000 tract boundaries to 2010 tract boundaries.

NCDB used a similar approach for 1980, first linking source year tracts to 1990 blocks and then interpolating from those blocks to 2000 tracts. NCDB used area-weighted interpolation using spatial data from Tiger/Line 1992. A less precise area weighting was used for 1970 that relied on the Census Bureau's tract correspondence file between 1970 and 1980. Every 1970 tract contributing to a 1980 tract was weighted equally. Then 1980 tracts were linked to 1990 blocks and, in a final step, to 2000 tracts.

Researchers should be aware of the potential for error in interpolation that is based only on area weights. Figure 3 presents an extreme example of what can happen. Here a single tract in 2000 was split into three tracts in 2010. The block populations in 2000 show that the very large area that became tract 36045980000 was almost unpopulated. The LTDB population estimate for this tract based on area + population weighting is only 11. Yet areal interpolation alone suggests that most population in the source tract should be estimated to be in 36045980000. Note also that some populous blocks in 2000 have been divided in 2010 between two tracts. An area + population weighting would yield reasonable estimates if population within each of these is not greatly skewed to one portion of its area.

## Assessing Alternative Approaches

Social scientists have used both area- and population-weighted approaches in other similar situations. We presume that inclusion of population weighting yields improved estimates, but it would be useful to have more information on how different the estimates are from alternative methods and for what kinds of variables one would expect to find the largest discrepancies. Researchers often have to use less than optimal data, and in those cases it is helpful to understand better the amount and sources of error.

To assess differences in the results from these estimation procedures, we present a series of comparisons for 2000–2010 (comparing our combined area and population interpolation with an alternative in which we only take into account area).[1]

These comparisons involve a selection of variables. Some of these are population counts: total population, non-Hispanic white population, Asian population, college graduates, and homeowners. Others are rates or medians: population density, percentage non-Hispanic white, percentage Asian, percentage college graduates, percentage homeowners, and median household income. It is more difficult to estimate absolute numbers (because these depend on how fully the area of a census tract has been settled) than to estimate compositional characteristics such as percentages and rates (which tend to be similar across adjacent tracts).

Table 2 provides comparisons for split tracts, many-to-many tracts, and (for reference) all tracts, including those with no changes. For each variable, Table 2 lists the mean and standard deviation in the initial year, based on the combined area/population interpolation estimates. These values are useful points of reference for evaluating in absolute terms how large the discrepancies are between the two estimation methods. The next column shows the correlation between the two estimates for a given set of tracts. Then four columns show the distribution of cases by how large the discrepancy is between the estimates, from less than 0.1 standard deviation (which we take to be a minor difference) to over 1.0 standard deviation.

We notice that split tracts yield more disparate estimates than do tracts with many-to-many changes. A careful analysis of change over time should take into account which tracts had no change in boundaries, which had simple consolidations, which were split, and which had many-to-many changes. The latter two types of tracts should be inspected separately for

***Table 2*** *Comparison of tract estimates between area + population and area-only interpolation*

| | | | | Size of discrepancy | | | |
|---|---|---|---|---|---|---|---|
| | ***M*** | ***SD*** | ***r*** | **< 0.1 *SD* (%)** | **0.1–0.5 *SD* (%)** | **0.5–1.0 *SD* (%)** | **> 1.0 *SD* (%)** |
| Splits: 2000–2010 ($n = 12{,}567$) | | | | | | | |
| Population count | 3,489 | 1,548 | 0.511 | 10.7 | 34.3 | 25.3 | 29.7 |
| Non-Hispanic white count | 2,426 | 1,420 | 0.681 | 18.8 | 37.5 | 22.5 | 21.2 |
| Non-Hispanic Asian count | 151 | 295 | 0.893 | 66.6 | 25.5 | 5.0 | 2.9 |
| College graduate count | 600 | 485 | 0.783 | 26.4 | 44.3 | 18.0 | 11.3 |
| Homeowner count | 880 | 479 | 0.634 | 16.3 | 37.3 | 23.2 | 23.2 |
| Population density | 1,689 | 4,722 | 0.926 | 79.2 | 18.0 | 1.9 | 1.0 |
| Non-Hispanic white % | 71.1 | 26.1 | 0.996 | 99.3 | 0.4 | 0.1 | 0.1 |
| Non-Hispanic Asian % | 4.1 | 7.1 | 0.987 | 99.2 | 0.6 | 0.2 | 0.1 |
| College graduate % | 26.6 | 16.3 | 0.997 | 99.0 | 0.6 | 0.2 | 0.1 |
| Homeowner % | 69.4 | 22.2 | 0.996 | 99.0 | 0.6 | 0.2 | 0.2 |
| Median household income | $48,857 | $18,736 | 0.999 | 99.1 | 0.6 | 0.2 | 0.1 |
| Many to many 2000–2010 ($n = 9{,}106$) | | | | | | | |
| Population count | 3,601 | 1,819 | 0.788 | 53.1 | 24.8 | 10.8 | 11.3 |
| Non-Hispanic white count | 2,382 | 1,680 | 0.874 | 61.3 | 22.3 | 9.4 | 7.0 |
| Non-Hispanic Asian count | 201 | 494 | 0.918 | 85.1 | 12.5 | 1.6 | 0.9 |
| College graduate count | 609 | 577 | 0.882 | 65.8 | 22.3 | 7.4 | 4.5 |
| Homeowner count | 850 | 544 | 0.860 | 59.5 | 22.8 | 10.0 | 7.7 |
| Population density | 2,078 | 4,514 | 0.949 | 85.8 | 11.6 | 1.8 | 0.8 |
| Non-Hispanic white % | 66.7 | 30.4 | 0.993 | 96.9 | 2.6 | 0.3 | 0.2 |
| Non-Hispanic Asian % | 5.2 | 10.4 | 0.979 | 97.4 | 2.2 | 0.2 | 0.2 |
| College graduate % | 25.8 | 17.6 | 0.991 | 95.6 | 3.4 | 0.6 | 0.3 |
| Homeowner % | 64.5 | 24.8 | 0.992 | 94.7 | 4.3 | 0.6 | 0.4 |
| Median household income | $47,500 | $21,605 | 0.996 | 95.8 | 3.4 | 0.6 | 0.2 |
| All tracts 2000–2010 ($n = 72{,}739$) | | | | | | | |
| Population count | 3,871 | 1,602 | 0.883 | 78.3 | 9.3 | 5.8 | 6.6 |
| Non-Hispanic white count | 2,676 | 1,618 | 0.932 | 81.3 | 9.8 | 5.0 | 3.9 |
| Non-Hispanic Asian count | 164 | 375 | 0.971 | 92.8 | 5.7 | 1.0 | 0.5 |
| College graduate count | 612 | 556 | 0.947 | 83.4 | 10.6 | 3.9 | 2.1 |
| Homeowner count | 960 | 514 | 0.917 | 80.4 | 9.6 | 5.2 | 4.8 |
| Population density | 1,988 | 4,549 | 0.983 | 94.5 | 4.7 | 0.6 | 0.3 |
| Non-Hispanic white % | 69.4 | 29.6 | 0.998 | 99.5 | 0.4 | 0.1 | 0.0 |
| Non-Hispanic Asian % | 4.0 | 8.1 | 0.994 | 99.5 | 0.4 | 0.1 | 0.0 |
| College graduate % | 23.8 | 16.9 | 0.998 | 99.3 | 0.6 | 0.1 | 0.1 |
| Homeowner % | 66.5 | 22.7 | 0.998 | 99.1 | 0.7 | 0.1 | 0.1 |
| Median household income | $45,158 | $20,492 | 0.999 | 99.3 | 0.6 | 0.1 | 0.0 |



unusual patterns of change that might be an artifact of the interpolation method.

We also notice that the estimates of absolute numbers (counts) have much greater discrepancies than the estimates of rates or averages. As an example, consider the estimates of non-Hispanic white residents and non-Hispanic white percentage for split tracts. The correlations for the number of whites is 0.681, and close to 20 percent of cases have discrepancies of less than 0.1 standard deviation. But estimates of the white percentage have near-perfect correlations. Close to 100 percent of estimates are within 0.1 standard deviation of each other.

As expected, when all tracts are included in the comparison, the correlations are higher and discrepancies are smaller. For example the two estimates of total population are correlated at 0.88, white count at 0.93, and Asian count at 0.97. Hence the potential errors resulting from reliance on area-weighted interpolation are moderated by the many tracts that require no estimation.

## Dissemination: The LTDB

Here we describe a new resource that we have created and made freely available for public use. The LTDB provides tools that can be used by scholars who have data reported within census tracts in the period from 1970 to 2000 (regardless of the source) and wish to estimate the same data using 2010 tract geography.

The LTDB (http://www.s4.brown.edu/us2010/Researcher/Bridging.htm) provides estimates using 2010 boundaries for a standard set of variables from 1970 through the 2006 through 2010 American Community Survey and Census 2010 (the 2006–2010 tract data were reported for 2010 tract boundaries). These data might meet the needs of many users. More versatile is the set of tools that allows users to input their own data. Key to this system are crosswalks for each prior year, similar to the Geographic Conversion Tables developed by Simpson (2002) and made available for public use and the proprietary BWF developed by NCDB. For every decennial year from 1970 to 2000, a crosswalk file is provided in which every row lists a 2010 tract ID, the ID of a tract in the source year that contributes to it, and the share of the source tract's population attributes that should be allocated to the 2010 tract. In cases where there is an exact correspondence between the source tract and the 2010 tract, there is only one row of data for the 2010 tract. Otherwise there are as many rows as there are contributing tracts. For completeness, the crosswalk file includes

every contributing tract, regardless of how small a fraction of its population should be allocated to the 2010 tract.

Supplementary information includes the 2010 metropolitan area (formally the Core Based Statistical Area or CBSA) code, flags to identify central city, tracts in 2010,[2] and the 2010 population and land area of the tract. For the 2000–2010 crosswalk we provide one additional indicator that we believe will assist users of the interpolated data: whether there was a boundary change involving this tract and, if so, what type of change occurred between 2000 and 2010.

The LTDB offers code in Microsoft Access and STATA that can be used in conjunction with the crosswalk file and an input data file prepared by the researcher. Input variable names need to be added to the code. Some variables, such as a median income, should be aggregated as a weighted average, and the user must identify the variable (e.g., number of households) to be used in weighting. The output file from Access or STATA lists all of the 2010 information about the tract from the crosswalk file and values of the input variables converted to 2010 boundaries.

## Summary and Discussion

Many 2000 census tracts are split, consolidated, or otherwise redrawn in Census 2010, and similar changes have occurred in prior years. These changes obstruct longitudinal analysis at the tract level and require the use of estimation procedures to harmonize data over time. We have focused on two approaches to interpolation that are practical at a national scale. The simpler approach is based on area weighting; a more desirable method also takes into account the distribution of population by blocks within the source tracts. We have shown that the differences between these two estimates can be very substantial.

Some kinds of analysis are especially sensitive to the actual counts (counts of the number of people, the number of members of particular population subgroups, and the number of housing units, etc.). A prime example would be a study of population growth at the tract level or a study of a phenomenon like rate of crime or disease that uses a population count in the denominator. For such variables, the correlations between the estimates from the two interpolation methods are mostly in the range of 0.50 to 0.85 in Table 2. For some counts, particularly for split tracts, the absolute value of discrepancies can be over 0.5 standard deviation for as many as half of these tracts. Of course, these results are for tracts that required interpolation. In the full data set, including the approximately 70 percent of tracts that did not change boundaries or experience consolidation, we saw that the correlations are much higher.

On the other hand, our results for variables calculated as percentages or averages suggest that area-weighted estimates for such variables can be used with a high degree of confidence when the analysis is based on correlations. Although the absolute values of these variables might diverge somewhat from those that would be estimated with population weighting, these two sorts of estimates are so highly correlated that their relationships with other variables are indistinguishable.

Nevertheless, both types of interpolation introduce error. Although one cannot assess how close estimates from either approach come to the "real" values (which would require access to the original point or block-level data), we provide an indicator of whether a tract's data have been interpolated and what kind of boundary changes are involved. Researchers might wish to check whether the same results are found when data for interpolated tracts are excluded or weighted less heavily than other cases.

The LTDB offers researchers a versatile, open-source approach to study census tract data in a longitudinal framework. For 2000–2010 the estimation methods are similar to those that have proved useful in the past, and they can be combined with input data from NCDB from 1970 through 1990 to update those estimates to 2010 boundaries. For some users it might be preferable to rely on the LTDB's area-only interpolation estimates for these prior years, especially for variables not included in the NCDB standard data set. For all users, the supplementary information from LTDB on types of boundary changes experienced in 2000–2010 offers new methods of assessing how errors in estimation affect their research results.

To clarify the contribution made by this research, we review the options it makes available to researchers to study tract-level changes between some earlier time and 2010.

1. For researchers wishing to harmonize data for pre-2000 census tracts, the LTDB uses areal interpolation to create a bridge. An alternative option is to acquire the NCDB files for 1970 through 1990 adjusted to 2000 boundaries and then apply the LTDB to bridge these data to 2010.
2. There are conditions in which using the NCDB in this fashion is a less satisfactory solution, two of which deserve emphasis here. Most important, NCDB does not provide linked files for all census variables, but only for a selection of variables from the sample count files.[3] Some researchers will need other census variables. In addition, researchers are increasingly working with information aggregated to the tract level from noncensus sources, such as criminal justice, public health, and voting records. The LTDB is well suited to these needs.

These harmonized data will facilitate studies of neighborhood change, such as population growth and decline, shifts in racial and ethnic composition, homeownership, and socioeconomic status. The long time series, extending over four decades, might make possible estimation of more complex models, such as

reciprocal causation or varying time lags. For researchers working with data from other countries and time periods, the interpolation methods used here could prove to be useful. The comparison of areal only and area + population interpolation might not prove to be the same in other contexts, but the more general finding—that spatial dependence of characteristics measured as rates or percentages tends to minimize errors in interpolation even when actual counts are over- or underestimated—might be widely applicable. ■

## Funding

We gratefully acknowledge funding support from the Russell Sage Foundation's US2010 Project and from the Population Studies and Training Center at Brown University, which receives core support from the NICHD (5R24HD041020, 5T32HD007338).



## Notes

[1] We also compared our area-weighted estimates for 1990–2000 with NCDB's population + area estimates. Results are similar except that we find a much lower correlation of estimated values for median household income than in 2000–2010. Our approach with this variable was to calculate an area-weighted average of the median incomes of source tracts. There is no documentation of NCDB's method, but we find a much higher correlation if we take a simple unweighted average of medians from the source tracts.

[2] In longitudinal research on metropolitan areas, it is desirable to hold constant the boundary between the central city and suburbia. The NCDB provides the place code for the place in which the largest area of the tract is located. We base the location flag on population share for 2000 and on area share for 1970 through 1990. The central city variable identifies tracts located in a principal city of the CBSA in 2010.

[3] NCDB provides sample data (Summary Files 3 and 4 in 2000, and its equivalents in prior years) even for variables that are available from full count tabulations in Summary Files 1 and 2. Not all users are aware that in the files based on sample count data, the Census does not adjust population totals to match the full count information that is available at the tract level. The correlations between values reported by the Census Bureau in 2000 Summary File 1 and Summary File 3 for variables like the total population and number and share of white and Asian residents are 0.98 or higher. In some tracts, however, there are larger discrepancies. For example, the average Asian count was 160 with a standard deviation of 384 in Summary File 1. In about 21 percent of tracts, the Summary File 3 value was different from the Summary File 1 value by more than 0.1 standard deviation (i.e., more than 38).

## Literature Cited

Bracken, I., and D. Martin. 1989. The generation of spatial population distributions from census centroid data. *Environment and Planning A* 21:537–43.

———. 1995. Linkage of the 1981 and 1991 UK censuses using surface modelling concepts. *Environment and Planning A* 27:379–90.

Eicher, C. L., and C. A. Brewer. 2001. Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science* 28: 125–38.

Goodchild, M. F., L. Anselin, and U. Deichmann. 1993. A framework for the areal interpolation of socioeconomic data. *Environment and Planning A* 25:383–97.

Goodchild, M. F., and N. Lam. 1980. Areal interpolation: A variant of the traditional spatial problem. *Geo-Processing* 1:297–312.

Gregory, I. N. 2002. The accuracy of areal interpolation techniques: Standardising 19th and 20th century census data to allow long-term comparisons. *Computers, Environment and Urban Systems* 26:293–314.

Gregory, I. N., and P. S. Ell. 2005. Breaking the boundaries: Geographical approaches to integrating 200 years of the census. *Journal of the Royal Statistical Society: Series A. Statistics in Society* 168:419–37.

Kyriakidis, P. C. 2004. A geostatistical framework for area-to-point spatial interpolation. *Geographical Analysis* 36:259–89.

Kyriakidis, P. C., and E.-H. Yoo. 2005. Geostatistical prediction and simulation of point values from areal data. *Geographical Analysis* 37:124–51.

Maantay, J. A., A. R. Maroko, and C. Herrmann. 2007. Mapping population distribution in the urban environment: The cadastral-based expert dasymetric system CEDS. *Cartography and Geographic Information Science* 34:77–102.

Martin, D. 1989. Mapping population data from zone centroid locations. *Transactions of the Institute of British Geographers* 14:90–97.

Martin, D., D. Dorling, and R. Mitchell. 2002. Linking censuses through time: Problems and solutions. *Area* 34:82–91.

Mennis, J. 2003. Generating surface models of population using dasymetric mapping. *The Professional Geographer* 55:31–42.

Reibel, M., and A. Agrawal. 2007. Areal interpolation of population counts using pre-classified land cover data. *Population Research and Policy Review* 26:619–33.

Reibel, M., and M. E. Bufalino. 2005. Steet-weighted interpolation techniques for demographic count estimation in incompatible zone systems. *Environment and Planning A* 37:127–39.

Simpson, L. 2002. Geography conversion tables: A framework for conversion of data between geographical units. *International Journal of Population Geography* 8:69–82.

Sleeter, R., and M. Gould. 2007. Geographic information system software to remodel population data using dasymetric mapping methods. In *Techniques and methods 11-C2.* Reston, VA: U.S. Department of the Interior. http://pubs.usgs.gov/tm/tm11c2 (last accessed 2 April 2014).

Tapp, A. F. 2010. Areal interpolation and dasymetric mapping methods using local ancillary data sources. *Cartography and Geographic Information Science* 37:215–28.

Tatian, P. A. 2003. *Neighborhood Change Database. NCDB. 1970–2000 Tract data: Data users guide*. Washington, DC: Urban Institute.

U.S. Census Bureau. 2011. *2010 Census tract relationship file overview*. Washington, DC: U.S. Census Bureau.

Wright, J. K. 1936. A method of mapping densities of population: With Cape Cod as an example. *Geographical Review* 26:103–10.

Xie, Y. 1995. The overlaid network algorithms for areal interpolation problem. *Computers, Environment and Urban Systems* 19:287–306.

Zandbergen, P. A., and D. A. Ignizio. 2010. Comparison of dasymetric mapping techniques for small-area population estimates. *Cartography and Geographic Information Science* 37:199–214.

Zhang, C., and F. Qiu. 2011. A point-based intelligent approach to areal interpolation. *The Professional Geographer* 63:262–76.

JOHN R. LOGAN is Professor in the Department of Sociology and Director of the Initiative on Spatial Structures in the Social Sciences at Brown University, Providence, RI 02912. E-mail:john_logan@brown.edu. His research focuses on urban development in the United States and China, incorporation of immigrants and minorities, and spatial inequalities.

ZENGWANG XU is Assistant Professor in the Department of Geography, University of Wisconsin, Milwaukee, WI 53211. E-mail: xuz@uwm.edu. His research integrates GIS, complex networks and systems science, and spatial and statistical analyses.

BRIAN J. STULTS is Associate Professor in the College of Criminology and Criminal Justice at Florida State University, Tallahassee, FL 32306. E-mail: bstults@fsu.edu. He studies the relationships among neighborhoods, race, and crime, with particular attention to impacts of racial segregation.