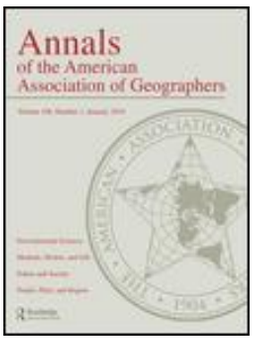

Annals of the American Association of Geographers

ISSN: 2469-4452 (Print) 2469-4460 (Online) Journal homepage: http://www.tandfonline.com/loi/raag21

Validating Population Estimates for Harmonized Census Tract Data, 2000–2010

John R. Logan, Brian J. Stults & Zengwang Xu

To cite this article: John R. Logan, Brian J. Stults & Zengwang Xu (2016) Validating Population Estimates for Harmonized Census Tract Data, 2000–2010, Annals of the American Association of Geographers, 106:5, 1013-1029, DOI: 10.1080/24694452.2016.1187060

To link to this article: https://doi.org/10.1080/24694452.2016.1187060

Validating Population Estimates for Harmonized Census Tract Data, 2000–2010

John R. Logan,* Brian J. Stults,† and Zengwang Xu‡

*Department of Sociology, Brown University
†College of Criminology and Criminal Justice, Florida State University
‡Department of Geography, University of Wisconsin, Milwaukee

Social scientists regularly rely on population estimates when studying change in small areas over time. Census tract data in the United States are a prime example, as there are substantial shifts in tract boundaries from decade to decade. This study compares alternative estimates of the 2000 population living within 2010 tract boundaries to the Census Bureau's own retabulation. All methods of estimation are subject to error; this is the first study to directly quantify the error in alternative interpolation methods for U.S. census tracts. A simple areal weighting method closely approximates the estimates provided by one standard source (the Neighborhood Change Data Base), with some improvement provided by considering only area not covered by water. More information is used by the Longitudinal Tract Data Base (LTDB), which relies on a combination of areal and population interpolation as well as ancillary data about water-covered areas. Another set of estimates provided by the National Historical Geographic Information Systems (NHGIS) uses data about land cover in 2001 and the current road network and distribution of population and housing units at the block level. Areal weighting alone results in a large error in a substantial share of tracts that were divided in complex ways. The LTDB and NHGIS perform much better in all situations but are subject to some error when boundaries of both tracts and their component blocks are redrawn. Users of harmonized tract data should be watchful for potential problems in either of these data sources. *Key Words: boundaries, census data, census tracts, interpolation.*

社会科学家在研究小型地区随着时间的变迁时，习惯仰赖人口估计。美国人口普查单位的数据便是最佳的案例，因为普查单位的边界，每十年皆有着显着的改变。本研究比较相对于人口普查局本身重新列表的居住于2010年普查单位边界内的2000年人口之各种替代式估计。所有的估计方法皆不免有误；本研究则是第一个对美国人口普查单位的替代式内插法之错误直接进行量化的研究。一项简易的地区加权方法，严密地估算由单一标准来源（邻里变迁数据集）所提供的估计，并透过仅考量未被水体覆盖的地区而获得若干改进。纵向普查单位数据集（LTDB）则使用更多的信息，并仰赖面积和人口的内插法之结合，以及有关水体覆盖地区的辅助数据。另一个由全国历史地理信息系统（NHGIS）所提供的估计组，则使用2001年的土地覆盖，以及目前的路网和人口分布与街廓层级的住宅单位数据。仅使用面积加权本身，导致了过去以复杂的方式切割的调查单位中的大部份出现重大错误。LTDB和NHGIS在所有的情况下皆表现较佳，但当两者的调查单位边界及其组成街阈进行重划时，则仍产生若干错误。统一的调查单位数据之使用者，应该注意上述两者中任一个数据来源的潜在问题。 *关键词： 边界, 人口普查数据, 人口普查单位, 内插法。*

Con frecuencia, cuando estudian el cambio en áreas pequeñas a través del tiempo, los científicos sociales tienen que depender de estimativos de la población. Los datos censales de los Estados Unidos por secciones son un buen ejemplo al respecto, en cuanto que se presentan cambios sustanciales en los límites de las secciones censales de década en década. Este estudio compara los estimativos alternativos de la población del 2000 que habita dentro de los límites de las secciones del 2010 de la propia retabulación de la Oficina del Censo. Todos los métodos de cálculo están sujetos a error; este es el primer estudio que cuantifica directamente el error de los métodos alternativos de interpolación en las secciones censales de los EE.UU. Un simple método de ponderación espacial aproxima muy de cerca los estimativos entregados por una fuente estándar (la Base de Datos de Cambio Vecinal), con alguna mejora lograda al considerar tan solo el área no cubierta por agua. Mayor información es usada por la Base de Datos de Secciones Longitudinales (LTDB, por la sigla en inglés), que depende de una combinación de interpolación espacial y poblacional lo mismo que de datos complementarios acerca de las áreas cubiertas con agua. Otro conjunto de estimativos suministrado por los Sistemas de Información Geográfica Históricos Nacionales (NHGIS) usa datos sobre la cubierta del suelo en 2001 y la actual red de carreteras y distribución de la población, y unidades de vivienda a nivel de manzana. La sola ponderación espacial da lugar a error mayor en una parte sustancial de las secciones censales que fueron divididas de maneras complejas. Los LTDB y NHGIS se desempeñan mucho mejor en todas las situaciones, pero están

Annals of the American Association of Geographers, 106(5) 2016, pp. 1013–1029
Initial submission, August 2015; revised submissions, February and April 2016; final acceptance, April 2016
Published by Taylor & Francis, LLC.

sujetos a cierto grado de error cuando se rediseñan los límites de las secciones y de las manzanas que las componen. Los usuarios de datos armonizados de las secciones censales deben estar atentos a problemas potenciales en cualquiera de estas fuentes de datos. *Palabras clave: límites, datos censales, secciones censales, interpolación.*

This study assesses the reliability of population estimates for census tracts where boundaries have been harmonized to control for changes in tract boundaries. It compares alternative methods of interpolation for tracts across the United States in the period from 2000 to 2010. The Longitudinal Tract Data Base (LTDB), which combines areal and population interpolation with ancillary data on water coverage is compared with two simpler approaches using only areal weighting. Results are also compared with two other public sources: the Neighborhood Change Data Base (NCDB), which was first developed for the 1970 to 2000 period and (after being converted to a proprietary system by Geolytics, Inc.) later extended to 2010, and the National Historical Geographic Information Systems (NHGIS) standardized block and tract data, which interpolates 2000 Census block and tract-level data to 2010 Census block group and tract boundaries.[1]

Population data for census tracts are a widely used resource for urban and regional research. Census tracts are small enough to satisfy many needs for very local information. Their sample sizes even in the American Community Survey (ACS) are large enough to meet many researchers' needs for reliable estimates, although concerns are growing about how to deal with the ACS's increased standard errors in comparison to decennial censuses (Spielman, Folch, and Nagle 2014). Yet an obstacle to longitudinal analysis of these data is that the boundaries of tracts are adjusted every decade. 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 in the next one, and their boundaries could be altered in complex ways. Table 1 summarizes the kinds of changes that occurred between 2000 and 2010 for the tracts analyzed here. Nearly 70 percent had no change other than minor adjustment of cartography (defined as changes involving less than 1 percent of their land area in 2000). A small number (1.4 percent) had a consolidation where two or more tracts in 2000 became a single tract in 2010 (so their 2000 populations can be simply summed to yield the value within 2010 boundaries). Other changes affecting about 30 percent of tracts create problems of estimation, because one or more tracts are reorganized to multiple tracts that do not respect original tract boundaries. These include what we refer to as split tracts (one tract divided into many) and cases where two or more tracts were reconfigured into two or more different new tracts (many-to-many). Special problems are created when these reconfigurations also subdivide census blocks and allocate them to multiple 2010 tracts, and this occurred in more than half of the split tracts and about a quarter of the many-to-many tracts.

Table 1 also reveals that the tracts with no change or mergers had relatively stable populations between 2000 and 2010 (increases of 5.2 percent and 2.6 percent, respectively). Split tracts on average grew by 27.0 percent in the decade, and those with many-to-many changes grew 15.2 percent. This finding offers clues about where tract boundaries are changing—in

Table 1. Census tract boundaries over time: Number and population of tracts experiencing various types of changes between 2000 and 2010

			Total population (millions)		Average tract size	
Type of change	Number	Share (%)	2000	2010	2000	2010
No change	49,757	68.9	200.0	210.4	4,020	4,228
Many to one (mergers)	981	1.4	3.5	3.5	3,521	3,614
One to many (splits)	12,445	17.2	43.4	55.1	3,489	4,429
Without divided blocks	6,138	8.5	21.8	26.4	3,544	4,294
With divided blocks	6,307	8.7	21.7	28.8	3,435	4,561
Many to many	9,022	12.5	32.5	37.4	3,597	4,146
Without divided blocks	2,279	3.2	8.1	9.0	3,535	3,968
With divided blocks	6,743	9.3	24.4	28.4	3,618	4,206
Total	72,205	100.0	279.3	306.4	3,869	4,244

faster growing areas that are likely to be found on the outer edge of the metropolis.

One way to deal with these changes (Exeter et al. 2005) is to construct larger areal units that merge together all of the tracts in one year that overlay tracts in another year. *Smart interpolation*, the approach considered here, is more ambitious, seeking to provide estimates of population in one decade within the boundaries of specific census tract areas as defined in another (Martin, Dorling, and Mitchell 2002). The general approaches are well known, dating back at least to the 1980s (Goodchild and Lam 1980; Goodchild, Anselin, and Deichmann 1993). The initial step is based on comparing the tract layers in two years in a GIS framework and allocating population to a tract from other tracts in proportion to their degree of overlap with it (areal weighting). Ancillary data are then used to improve understanding of how populations are distributed within tracts (dasymetric interpolation). The "binary mask"—determining which subareas are populated and which are not—was described by Tapp (2010) as the most basic and commonly used ancillary procedure. For example, forested areas, areas covered by bodies of water, and areas without roads are likely to have little if any population. A binary mask assumes that they have no population and consequently such areas do not contribute population to a tract estimate.

In principle, many forms of ancillary data could be applied for the dasymetric interpolation purpose, as long as they indicate the presence or absence or density of the variables (e.g., population) to be interpolated. Many offer more specific information than a binary mask. They include land cover data (Mennis 2003; Reibel and Agrawal 2007; Buttenfield, Ruther, and Leyk 2015), street networks (Reibel and Bufalino 2005), remote sensing data (Harvey 2002; C. Wu and Murray 2007), and population or other information over time (Schroeder 2007; Mennis 2016). These can also be used in combination. Our purpose here is to provide an evaluation of how closely the estimated counts from these sources and from alternative forms of areal interpolation match the "true" counts. This analysis is now possible for the specific comparison of 2000 and 2010, because the Census Bureau has released tabulations of population in 2000 within tracts as bounded in 2010.

Estimation Procedures

Simple areal weighting can allocate population from the tract as defined in 2000 to a 2010 tract area directly in proportion to the share of its area that lies within that 2010 tract. This areal weighting interpolation can be represented as $\hat{y}_t^1 = \sum_{t_0 \in \phi} \sum_{\rho \in t} (A_\rho / A_{t_0}^0) y_{t_0}^0$ where A_ρ represents the area of the part of 2000 tract t_0 that overlaps with the 2010 tract t, $A_{t_0}^0$ represents the total area of the 2000 tract t_0, $y_{t_0}^0$ is the population in 2000 (or other characteristic) of the 2000 tract t_0, $\hat{y}_t^1$ is the estimated population in 2000 (or other characteristic) of the 2010 tract t, and ϕ is the set of 2000 tracts that contribute to the 2010 tract t. The areas (both A_ρ and $A_{t_0}^0$) can be the total areas that include both land and water or only land areas. Respectively, we term them all-area and land-only versions of areal weighting. Because people reside only in the land area of most tracts, the land-only version areal weighting interpolation is expected to be more accurate. We find that the estimates provided for 2000 to 2010 by Geolytics' NCDB correspond closely but not exactly to the all-area interpolation. The procedures used by NCDB for 2000 to 2010 are not fully documented (see http://www.geolytics.com/USCensus,Neighborhood-Change-Database-1970–2000,Data,Geography,Products.asp), and we cannot exactly replicate them.[2]

The procedures used for LTDB 2000 to 2010 estimates are described in detail by Logan, Xu, and Stults (2014). They involve a combination of area and population interpolation, using a land–water dichotomy as ancillary data. The researchers made use of the Topological Faces layer of the TIGER/Line shapefiles created by the U.S. Census Bureau in 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 *subblock* unit in census geography. Let us refer to it as 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 FIPS code) in both the 2000 and 2010 census. These fragments from the Faces file can be dissolved to the tract and block layers for 2000 and 2010.

The first step is to allocate reported tract-level population counts in 2000 to blocks within the tract. The LTDB bases this allocation on the block's share of the

total tract population in 2000. This procedure avoids having to assume that population was uniformly distributed through the tract. It then estimates what share of the 2000 block population lies in each fragment within that block. This step (land-only areal weighting at the block level) is solely based on the fragment's share of the block's land area, disregarding portions of fragments that are covered with water. It is then straightforward to aggregate populated fragments to the 2010 census tracts.

Formally, the area and population weighting interpolation implemented in LTDB can be represented as $\hat{y}_t^1 = \sum_{t_0 \in \phi} \sum_{b \in t_0} \sum_{frag \in b} (A_{frag}/A_b^0)(y_{t_0}^0 \times p_b^0/p_{t_0}^0)$, where the $\hat{y}_t^1$ is the estimated variable for 2010 tracts, A_{frag} is the land area of the fragments within the 2000 block b, A_b^0 is the land area of the 2000 block b, $y_{t_0}^0$ is the variable at the 2000 tracts, p_b^0 is the 2000 census population at the block b, $p_{t_0}^0$ is the 2000 census population at the tract t_0, and the ϕ is the set of 2000 census tracts contributing to the 2010 tract t. The $y_{t_0}^0$ can be any count variables to be interpolated by the area and population weighting method, for which the LTDB provides interpolation tools through Microsoft Access database and Stata code. If $y_{t_0}^0$ is the 2000 tract population, the interpolation is essentially a block-level areal weighting interpolation using land area only.

The NHGIS estimates expand on procedures used in the 1990 to 2000 NCDB, which used road networks in 1990 as an indicator of population density within census tracts. In principle one would expect considerable improvement from the additional information that NHGIS relies on: land cover from the National Land Cover Database 2001 (NLCD; Homer et al. 2007), as well as road networks, location of water bodies, and population and housing counts in 2010. The estimation procedure is complex, employing a combination of weights. One set of weights is derived from a binary dasymetric interpolation that identifies "inhabited" areas in two ways: (1) whether an area is in a water body (as in the LTDB) and (2) whether it includes at least 5 percent impervious surface (based on the NLCD) and is within 300 feet of a road in 2010. The second set of weights is based on target-density weighting (Schroeder 2007) that is limited to areas classified by the first procedure as inhabited zones (as in Ruther, Leyk, and Buttenfield 2015).

In principle one would expect that the more ancillary information is used, the better the estimate. Hence, an all-area areal weighting should perform least well, a land-only interpolation somewhat better, a population and land-only areal interpolation much better, and an interpolation that also takes into account land cover and road networks should perform best. This is broadly the result of our analysis here. The contribution of this analysis is multifold. It is the first national-level comparison of population estimates to "real" values. It reveals a very large disparity in the accuracy of areal interpolation methods versus approaches that make use of block-level population data. It specifies the conditions under which even these latter approaches are subject to error. Finally, it demonstrates that in many cases the more complex methods employed by NHGIS yield worse results than the simpler LTDB methodology. For most users, either of these latter two sources should be satisfactory, but both should be used with caution. When blocks as defined in 2000 are divided and allocated to more than one 2010 tract, either method can provide erroneous estimates.

Research Design

This study includes almost all populated census tracts in 2000 and 2010 in the continental United States.[3] The validation of estimates from the four approaches is based on the census tract population change file (https://www.census.gov/population/metro/data/c2010sr-01patterns.html) for which the Census Bureau retabulated Census 2000 data using 2010 geography. Discrepancies between this retabulation and estimates from any interpolation method are only partly due to errors of estimation. They can also result from postcensus changes that were made to the 2000 tract populations by the Bureau. These changes include address corrections, geocoding improvements, boundary adjustments, and other enhancements to the Census Bureau's address and spatial database. Such changes were made in many tracts, and they can be large. For example, in the following analysis we find discrepancies of greater than 1 percent between the LTDB and NHGIS population estimates and the revised census counts in about 14 percent of tracts where there was no change in boundaries. In nearly 3 percent of these tracts the discrepancy is 5 percent or larger. The corrected 2000 tract populations are not publicly available, so they cannot be

used to improve the estimates. It would be reasonable to presume that the interpolated estimate in cases with no boundary change is "correct" by definition and that discrepancies in these cases are solely due to the Bureau's revised counts. Because such revisions occurred irrespective of boundary changes (they were completed before the 2010 tracts boundaries were set), we regard the distribution of discrepancies in the no-change tracts as a baseline for what to consider an accurate estimate for other kinds of tracts.

Geographers regularly seek to validate estimates or to compare the performance of alternative procedures through comparisons to true data (Flowerdew, Green, and Kehris 1991; Goodchild, Anselin, and Deichmann 1993). One recent study (Ruther, Leyk, and Buttenfield 2015) sought to validate areal interpolation population estimates for 2000 census tracts within 2010 tract boundaries. Without access to the "real" 2000 populations in the new boundaries, they aggregated data from 2000 for smaller units (2000 blocks) to the larger 2010 tract areas. Where blocks as defined in 2000 did not fall entirely within a single 2010 tract, they allotted portions of the block data to each tract according to the share of land classified as inhabited based on the NLCD. In these cases the resulting tract count could not be considered real because it is estimated by interpolation. The researchers (Ruther, Leyk, and Buttenfield 2015) argued that little bias is introduced because most 2000 census blocks (97 percent in one Ohio county that they studied) lie entirely within a single 2010 census tract. We believe that comparison to actual census counts, even when they have experienced postenumeration corrections, is a superior approach, because it does not rely on any of the interpolation procedures that are being tested.

In validation studies, researchers sometimes report the average, minimum, or maximum discrepancy. We provide a more complete distribution of the size of discrepancies, both in absolute terms and as a proportion of the actual value. We expect minimal error for tracts with no change or merged tracts (many-to-one changes). Errors are likely to be larger for more complex changes: one-to-many and many-to-many tracts. In these latter cases, error is more likely when blocks have been subdivided. Therefore, we report results separately for tracts that experienced different types of boundary changes. We also report a commonly used summary measure of this distribution, the root mean squared error: $RMSE = \sqrt{\sum_i (y_i - \widehat{y}_i)^2 / q}$, where y_i is the actual population of tract i, $\widehat{y}_i$ is the estimated population of tract i, and q is the number of tracts. This statistic sums the disparities between estimated and actual population counts. Because these values are squared before being summed, the RMSE counts large absolute differences disproportionately compared to small ones. An alternative measure treats the discrepancy as a proportion of the actual value, lending more weight to a disparity of a given absolute size in an area with fewer residents than in an area with many (Eicher and Brewer 2001; Gregory 2002; Mennis and Hultgren 2006). We report the proportional error here.

Pitfalls in Areal Interpolation

We begin with two extreme yet representative examples of census tract boundary change to illustrate how reliance on areal weighting can result in poor estimates. These cases clarify how each type of estimate is made, as well as possible pitfalls. In one of these, the 2010 census tract includes the Dallas–Ft. Worth airport (DFW), an urban built-up area with few residents and little water cover. The other is on the island of Oahu, Hawaii, a census tract created in 2010 that is entirely underwater. In both cases, although the Census Bureau has counted almost no residents in these tracts, areal weighting allocates large populations to them.

The DFW Case

Figure 1 displays a satellite image of the area around DFW airport. The yellow-shaded area shows a tract newly created in 2010 (48439980000^4), which includes several terminals, parking areas, runways, and much undeveloped land. The 2000 tract boundaries are shown in red. Much of the area of the new tract lies within two large 2000 tracts (3507 and 3706) that had dense populations to the west of the airport in 2000. The undeveloped area on the north end is part of a 2000 tract (3705) that included densely populated areas to the northwest and northeast.

How does areal weighting work in a case like this? Table 2 shows the components of the calculation: the share of land area in each 2000 tract that overlaps with 980000, their populations in 2000, and the portion of their population that is allocated to 980000. Aggregating the estimated contributions from each tract results in an estimate of 15,826 (slightly more

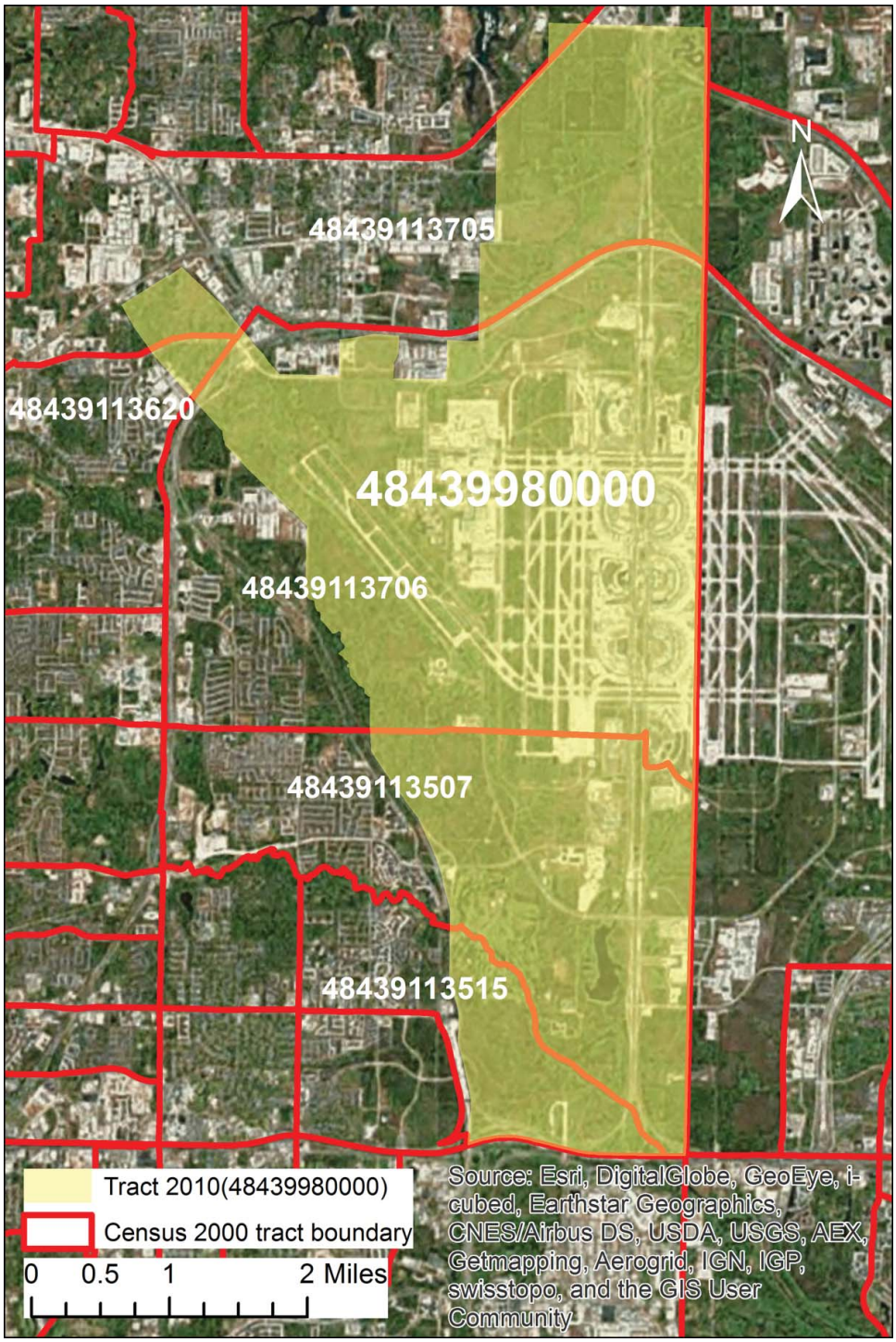

Figure 1. Tract boundary changes in the vicinity of Dallas–Ft. Worth airport. The new Tract 48439980000 is shaded yellow; boundaries of its contributing 2000 census tracts are shown in red. (Color figure available online.)

when water area is included: 15,836), compared to the Census count of 19. The LTDB estimate based on block land area and block populations is 117, also high but much closer to the actual value. The NHGIS estimate of 16 is still more accurate.

The use of block-level population data in both the LTDB and NHGIS greatly improves the estimate because it takes into account that much of the overlapping area was in blocks with few residents. This point can be illustrated by looking more closely at one

Table 2. Contributors to new Dallas Tract 48439980000

	2000 tract number						
	3507	3515	3620	3703	3705	3706	New tract estimate
Total population in 2000	8,224	7,408	11,384	5,240	5,265	4,808	
Share of area allocated to new tract	69.09	52.66	5.31	0.44	36.14	77.43	
Share of land area allocated to new tract	68.8	52.67	5.32	0.66	36.07	77.48	
Interpolation: all area	5,682	3,901	604	23	1,903	3,723	15,836
Interpolation: land area	5,658	3,902	606	35	1,899	3,725	15,824
LTDB	0	0	3	1	109	4	117
NHGIS	0	0	0	1	14	1	16

Note: Real 2000 population = 19. LTDB = Longitudinal Tract Data Base; NHGIS = National Historical Geographic Information Systems.

tract, 113705, which had a population of 5,265 in 2000. It was split in 2010 with a majority of its land area remaining in the original tract. Parts or all of twenty blocks were assigned to 980000. Only two of these were populated: Block 14 with three people and Block 29 with 175. All of Block 14 was placed within the new tract, so its three residents were allocated there. One portion of Block 29 (all water) and a large part of the remainder of Block 29 (60.5 percent of it, all land) became two new blocks in 980000. The former portion contributed nothing, whereas the latter contributed 106 to the estimate for 980000, for a total of 109 from Tract 113705.

In this way the LTDB estimates 117 people in an area where the census counts only 19, but it mostly avoids the misallocation from unpopulated blocks that assigns nearly 16,000 people to the area by areal weighting. The NHGIS estimate is superior because it notices that the portion of Block 29 allocated to the new tract was mostly undeveloped land in 2001, and the developed portion was allocated to a different tract.

Oahu's Doughnut Census Tract

The Census Bureau created a new census tract 15003990001 around the island of Oahu in 2010 in the form of a torus (or doughnut). In 2000 the census tracts along the outer perimeter of the island all extended some distance into the ocean, so that they encompassed both unpopulated water area and populated land area near the shore, including several census tracts in Honolulu. Figure 2 shows the island with the new tract highlighted in green and the boundaries from 2000 of the tracts around the shoreline in red.

This is a case where the invalid assumptions of areal weighting are further aggravated by substantial water coverage. To show how the issues play out, Table 3 displays the calculations for four kinds of interpolations: areal weighting alone, interpolation of land area, and the LTDB and NHGIS interpolations. To simplify, we select a single tract in Honolulu in 2000 for this illustration (008402). This tract was split into four components: three whole new tracts and a small part of 990001.

This is a case where the Census Bureau's postenumeration count for the original 2000 Tract 008402 (8,801) is considerably higher than the published figure (8,087). Still, these counts offer a useful benchmark to judge the accuracy of the estimates made by different methods.

- The allocation from the original 2000 tract to Tract 990001 based on areal weighting (counting water area) is large (7,324) because a large portion of the original Tract 008402 was in 990001. All other allocations only consider land area, and they are 0 because no land area was involved, matching the census retabulation.
- The calculations for allocation to the new Tract 008402 illustrate the advantage of interpolation by land area that also takes into account the block-level populations. The allocation by areal weighting including water is modest (619), because less than 8 percent of the original 008402's area remained in the new 008402. Interpolation by land area provides a more realistic estimate (6,180). The LTDB and NHGIS interpolations (7,980 and 7,996) are closer to the census benchmark (8,041) mostly because the forty-seven blocks that remained in the new 008402 had widely varying population densities that could be taken into account in making the estimate.

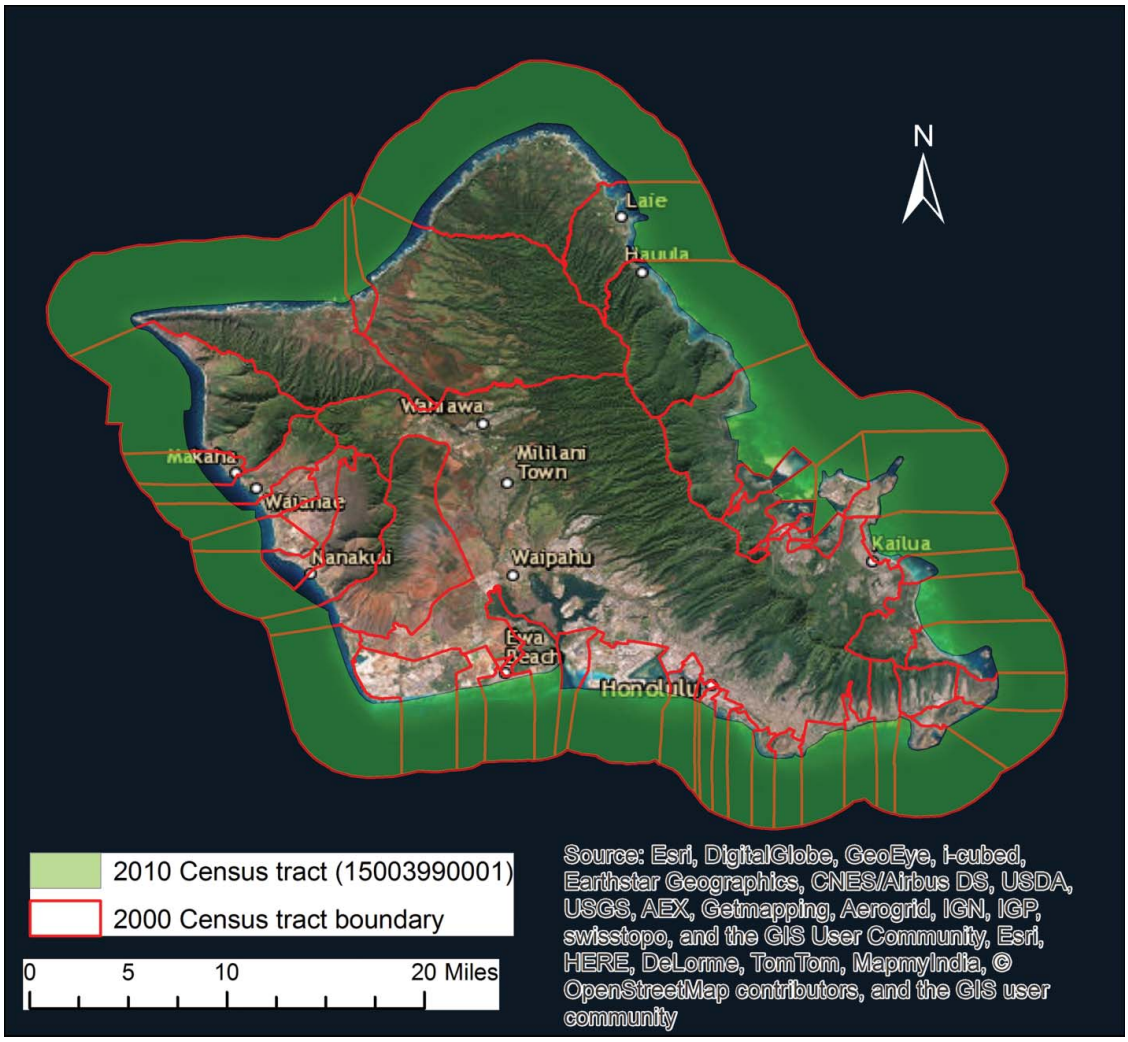

Figure 2. Census tracts in Oahu (Hawaii). 2000 tract boundaries are shown in red and the new Tract 15003990001 on its perimeter is shaded green. (Color figure available online.)

Table 3. Interpolation of 2000 population in the 2010 components of Honolulu Tract 15003008402

	New 2010 tract number			
	8402	8407	8408	990001
Share of total area allocated to the new tract	7.65	0.46	1.33	90.56
Share of land area allocated to the new tract	76.42	6.07	17.52	0
Interpolation: all area	619	37	108	7,324
Interpolation: land area	6,180	491	1,417	0
LTDB	7,980	6	102	0
NHGIS	7,996	16	75	0
Census retabulation	8,041	693	67	0

Note: Real 2000 population = 8,801. LTDB = Longitudinal Tract Data Base; NHGIS = National Historical Geographic Information Systems.

Comparison of Estimates

We now turn to the overall comparison between the Census Bureau count of 2000 population within 2010 tract boundaries with estimates from areal interpolation that either does or does not take into account the water layer, plus estimates from NCDB, LTDB, and NHGIS. The NCDB estimates perform poorly in tracts with boundary changes, similar to both of the areal interpolation estimates. We cannot be sure why, because the methodology behind NCDB estimates is not publicly known. We do know that in many cases the NCDB estimate was the same as the one from areal weighting including water, but there were many other cases where the estimate was quite different. The LTDB and NHGIS estimates are much better than

these, although they are imperfect, and the NHGIS estimates are slightly more accurate overall than the LTDB estimates. We interpret this result as affirmation of the utility of a combined approach (areal and population interpolation using ancillary water layer data) in comparison with a simpler areal weighting, with a small additional improvement from the use of land cover information and 2010 road network and population data.

In Tables 4 through 6 we compare four main kinds of tracts: no change, merged tracts (many to one), split tracts (one to many), and many to many. Because both the LTDB and NHGIS interpolations make use of block-level data for the third and fourth categories, in these categories we also take into account the difference between changes where there was no subdivision of a block versus those where a block was divided across different new tracts. It is in these latter cases where some procedure must be introduced to decide what share of the block's population to allocate to each tract. This is where the LTDB and NHGIS approaches diverge, with NHGIS taking into account much additional ancillary information. The analysis reveals how much was gained in this way.

Tables 4 and 5 summarize the distribution of discrepancies for each estimator. Errors are reported in Table 4 as a proportion of the actual tract population and in Table 5 as absolute values. The tables can be

Table 4. Distribution of tracts by proportional error in estimate

	Exact	<1%	1–2.99%	3–4.99%	5–10%	>10%
Tracts with no change ($n = 49{,}757$)						
Areal interpolation (including water area)	0.140	0.680	0.111	0.024	0.018	0.025
Areal interpolation (land only)	0.145	0.695	0.110	0.023	0.017	0.010
NCDB	0.086	0.596	0.189	0.062	0.038	0.028
LTDB	0.157	0.699	0.098	0.021	0.015	0.010
NHGIS	0.160	0.700	0.095	0.020	0.015	0.010
Many-to-one tracts ($n = 981$)						
Areal interpolation (including water area)	0.115	0.616	0.154	0.041	0.030	0.045
Areal interpolation (land only)	0.120	0.640	0.148	0.038	0.023	0.031
NCDB	0.037	0.412	0.301	0.119	0.080	0.052
LTDB	0.138	0.648	0.127	0.034	0.024	0.029
NHGIS	0.140	0.647	0.123	0.038	0.023	0.029
One-to-many tracts: No divided block ($n = 6{,}138$)						
Areal interpolation (including water area)	0.000	0.022	0.045	0.047	0.115	0.770
Areal interpolation (land only)	0.000	0.020	0.045	0.049	0.119	0.766
NCDB	0.000	0.022	0.046	0.048	0.115	0.770
LTDB	0.196	0.632	0.104	0.027	0.023	0.018
NHGIS	0.196	0.632	0.104	0.027	0.023	0.018
Many-to-many tracts: No divided block ($n = 2{,}279$)						
Areal interpolation (including water area)	0.004	0.129	0.171	0.095	0.130	0.472
Areal interpolation (land only)	0.004	0.132	0.172	0.098	0.134	0.459
NCDB	0.004	0.118	0.172	0.100	0.133	0.472
LTDB	0.221	0.592	0.104	0.032	0.025	0.026
NHGIS	0.221	0.592	0.104	0.032	0.025	0.026
One-to-many tracts with divided block ($n = 6{,}307$)						
Areal interpolation (including water area)	0.001	0.020	0.039	0.038	0.079	0.823
Areal interpolation (land only)	0.001	0.020	0.040	0.039	0.084	0.816
NCDB	0.001	0.021	0.039	0.037	0.081	0.822
LTDB	0.055	0.563	0.183	0.060	0.062	0.076
NHGIS	0.062	0.613	0.162	0.052	0.048	0.063
Many-to-many tracts with divided block ($n = 6{,}743$)						
Areal interpolation (including water area)	0.004	0.194	0.204	0.086	0.101	0.410
Areal interpolation (land only)	0.004	0.193	0.213	0.088	0.104	0.398
NCDB	0.005	0.183	0.198	0.092	0.108	0.415
LTDB	0.057	0.530	0.205	0.064	0.053	0.090
NHGIS	0.066	0.595	0.175	0.050	0.042	0.073

Note: NCDB = Neighborhood Change Data Base; LTDB = Longitudinal Tract Data Base; NHGIS = National Historical Geographic Information Systems.

Table 5. Distribution of tracts by absolute size of error in estimate

	Exact	1–5	6–25	26–100	101–499	500+
Tracts with no change ($n = 49{,}757$)						
Areal interpolation (including water area)	0.140	0.278	0.335	0.172	0.055	0.020
Areal interpolation (land only)	0.145	0.287	0.342	0.170	0.051	0.006
NCDB	0.086	0.212	0.315	0.272	0.094	0.020
LTDB	0.157	0.300	0.337	0.154	0.046	0.006
NHGIS	0.160	0.303	0.337	0.150	0.044	0.006
Many-to-one tracts ($n = 981$)						
Areal interpolation (including water area)	0.115	0.244	0.324	0.203	0.086	0.029
Areal interpolation (land only)	0.120	0.253	0.335	0.201	0.073	0.017
NCDB	0.037	0.119	0.230	0.392	0.192	0.030
LTDB	0.138	0.267	0.329	0.178	0.071	0.016
NHGIS	0.140	0.270	0.324	0.179	0.070	0.016
One-to-many tracts: No divided block ($n = 6{,}138$)						
Areal interpolation (including water area)	0.000	0.003	0.014	0.050	0.266	0.666
Areal interpolation (land only)	0.000	0.002	0.013	0.049	0.275	0.660
NCDB	0.000	0.003	0.014	0.051	0.265	0.667
LTDB	0.196	0.299	0.304	0.140	0.052	0.009
NHGIS	0.196	0.299	0.304	0.140	0.052	0.009
Many to many tracts: No divided block ($n = 2{,}279$)						
Areal interpolation (including water area)	0.004	0.025	0.074	0.208	0.314	0.375
Areal interpolation (land only)	0.004	0.024	0.077	0.215	0.317	0.363
NCDB	0.004	0.026	0.065	0.204	0.325	0.376
LTDB	0.221	0.283	0.285	0.139	0.057	0.014
NHGIS	0.221	0.283	0.285	0.139	0.057	0.014
One-to-many tracts with divided block ($n = 6{,}307$)						
Areal interpolation (including water area)	0.001	0.004	0.016	0.047	0.206	0.726
Areal interpolation (land only)	0.001	0.004	0.015	0.047	0.218	0.716
NCDB	0.001	0.005	0.016	0.046	0.207	0.726
LTDB	0.055	0.201	0.329	0.249	0.138	0.028
NHGIS	0.062	0.233	0.349	0.235	0.103	0.019
Many-to-many tracts with divided block ($n = 6{,}743$)						
Areal interpolation (including water area)	0.004	0.033	0.108	0.248	0.281	0.326
Areal interpolation (land only)	0.004	0.033	0.109	0.252	0.287	0.315
NCDB	0.005	0.030	0.098	0.248	0.292	0.327
LTDB	0.057	0.176	0.303	0.284	0.150	0.030
NHGIS	0.066	0.217	0.339	0.253	0.105	0.020

Note: NCDB = Neighborhood Change Data Base; LTDB = Longitudinal Tract Data Base; NHGIS = National Historical Geographic Information Systems.

read to answer two questions: which type of tract has greater error and which estimate is closer to the actual value. Table 6 reports the RMSE for the proportional error in estimates.

Tracts Requiring No Interpolation

We start with the estimates for the very large number of tracts with no change and the smaller number of merged tracts. In these tracts, in principle, there should be no error because there is no need for interpolation. The error here is due to the postenumeration corrections made by the Census Bureau. The Bureau's corrected counts have not been publicly released, so interpolations by any method must be based on the original published counts. Let us focus on the LTDB and NHGIS estimates for tracts with no change. Table 4 (rows 4 and 5) shows that these estimates had errors of greater than 1 percent in about 15 percent of these tracts. These errors must be due to nontrivial changes in postenumeration population estimates. Users of 2000 tract data from any source, including the files currently being disseminated by the Census Bureau, should be aware of this source of error. For scholars whose primary interest is not in the actual numbers of residents (in total or as tabulated against another attribute like race or education), but in the

Table 6. Root mean squared error for proportional error in estimates by type of tract

	Areal interpolation (including water)	Areal interpolation (land only)	NCDB	LTDB	NHGIS
Tracts with no change ($n = 49{,}757$)	43.7	2.5	43.9	1.0	1.0
Many-to-one tracts ($n = 981$)	0.1	0.1	0.1	0.1	0.1
One-to-many tracts: No divided block ($n = 6{,}138$)	1,229.9	1,104.5	1,229.6	0.2	0.2
Many-to-many tracts: No divided block ($n = 2{,}279$)	2,339.8	2,343.0	2,334.2	1.8	1.8
One-to-many tracts with divided block ($n = 6{,}307$)	2,440.2	2,415.7	2,439.8	97.4	77.6
Many-to-many tracts with divided block ($n = 6{,}743$)	2,939.3	2,902.5	2,938.8	211.1	178.8
All tracts ($N = 72{,}205$)	1,276.6	1,254.4	1,276.1	70.6	59.2

Note: NCDB = Neighborhood Change Data Base; LTDB = Longitudinal Tract Data Base; NHGIS = National Historical Geographic Information Systems.

percentage distribution by other attributes, even a 5 percent or 10 percent discrepancy in total numbers might not be problematic. Residents who were originally credited to one tract have been reassigned to a neighboring tract, and to the degree that neighboring tracts are similar (i.e., that there is a strong spatial structure to population characteristics) the reassignment introduces little error. For studies of population growth and other studies where absolute numbers matter, however, the official public data for tracts in 2000 can be misleading.

The results for these tracts are useful for the evaluation of interpolated estimates, providing a basis for interpreting the importance of interpolation error for scholars conducting longitudinal studies who require data within constant boundaries. For example, we learn from Table 4 that the best estimates of population in tracts with no change diverge from the "real" values in corrected postenumeration data. In 1.0 percent of cases the divergence is larger than 10 percent, and in 4.5 percent of cases the divergence is greater than 3 percent. This amount of error attributable to census corrections can be compared to the additional error that we discover in other types of tracts where interpolation is required. If the latter were large relative to the former, scholars would have stronger reasons to be doubtful of the efficacy of interpolation. If it were relatively small, scholars might be encouraged to use interpolated data. Tables 4 through 6 could be assessed in terms of the difference in errors that we find between tracts not requiring interpolation, tracts requiring interpolation but without divided blocks, and tracts requiring interpolation and involving subdivided blocks. This difference represents the "additional error" induced by interpolation.

In these tracts with no change and merged tracts, results are quite similar across interpolation methods with one exception: The errors for NCDB estimates are noticeably larger. In tracts with no change, for example, only 68.2 percent of NCDB estimates have an error of less than 1 percent, compared to 82 percent to 86 percent for other methods. Only 29.8 percent have errors of five or less, compared to 42 percent to 46 percent for other methods. We have no explanation for this poor performance.[5]

Multiple Destinations Without Divided Blocks

Discrepancies for the more difficult cases—one to many (split tracts) and many to many—are considerably greater. In areal weighting the errors arise when populations are not uniformly spread within the whole tract. In the LTDB and NHGIS these errors arise only when block populations are not uniformly spread across the fragments of blocks that are assigned to different 2010 tract areas. For this reason we find that where tracts in 2000 are divided into multiple destination tracts in 2010 but without block divisions, estimates by LTDB and NHGIS are about as good as in unchanged and merged tracts. For example, compare the discrepancies for NHGIS estimates in unchanged tracts to multiple destination cases with no block divisions. In Table 4, 86.0 percent of NHGIS estimates for unchanged tracts were within 1 percent of the census's corrected count, compared to 82.8 percent of NHGIS estimates for split tracts and 81.3 percent for many-to-many tracts with no block divisions. In Table 5, 46.3 percent of NHGIS estimates for unchanged tracts were within five persons of the census's corrected count, compared to 49.5 percent of NHGIS estimates for split tracts with no block divisions and 50.4 percent for many-to-many tracts with no block divisions. The RMSE (Table 6) was 1.0 for NHGIS estimates for unchanged tracts compared to

0.2 for split tracts and 1.8 for many-to-many tracts with no block divisions. We consider all of the discrepancies across these categories for both the LTDB and NHGIS to be insignificant, reflecting simply that the Census made slightly different degrees of correction of the 2000 population counts in these different sets of tracts.

The situation is different for the areal interpolation methods and the NCDB, with estimates that involve much higher degrees of error. Even when no blocks are subdivided, only 1 to 2 percent of estimates using any of these methods for split tracts or many-to-many tracts is within 1 percent of the census count, whereas more than 75 percent are off by more than 10 percent for split tracts with no divided block, and nearly half are off by this much for many-to-many tracts with no divided blocks (Table 4). Similar results are found using absolute values of discrepancy (Table 5). The RMSE is above 1,000 in these cases for both areal interpolation methods and the NCDB.

Multiple Destinations with Divided Blocks

Finally, we consider tracts that were allocated to multiple 2010 tracts and at least one block was divided between two 2010 tracts. The results for these tracts are even worse for areal weighting and NCDB estimates than we found with undivided blocks. These are the cases, however, where the more sophisticated interpolation methods used by the LTDB and NHGIS begin to have difficulty. We refer to results for NHGIS to illustrate this point, but the situation for LTDB is quite similar. The problem can be seen in Table 4, where in the case of split tracts the NHGIS estimates are within 1 percent of the census count in 82.8 percent of tracts with no divided block but only 67.5 percent of tracts with a divided block. NHGIS estimates are within five for 49.5 percent of split tracts with no divided blocks but only 29.5 percent of split tracts with divided blocks (Table 5). The RMSE shown in Table 6 is only 0.2 for split tracts with no divided blocks, versus 77.6 for split tracts with divided blocks.

How large is the problem of divided blocks? First, there are many of them—about 13,000 tracts (out of 72,205) involved divided blocks. Second, the LTDB and NHGIS estimates are clearly less accurate in these cases, as the RMSE is extremely low in cases without divided blocks but in the range of 75 to 200 in cases with divided blocks. As many as 5 to 10 percent of these estimates are more than 10 percent from the census count, and 10 to 15 percent deviate by more than 100 persons from the census count. Less than half of these deviations could be attributed to errors induced by the Census's postenumeration corrections.

On the other hand, the methods employed by the LTDB and NHGIS to allocate population from divided blocks are much more successful than simpler areal weighting. We also find a small improvement in the NHGIS estimates compared to the LTDB. For example, the RMSE for many-to-many tracts with divided blocks was 211.1 for the LTDB versus 178.8 for the NHGIS. We view this as a small difference in relation to the differences shown here between these estimates and those from areal weighting or between estimates for tracts with divided versus undivided blocks.

Comparing LTDB and NHGIS

A question arising from these comparisons is why the NHGIS estimates, which make use of considerably more information, appear to be only slightly better than the LTDB estimates. The difference between them is difficult to describe in more detail, because our reference point for validation is the postcensus corrected population count. In any given tract, what appears to be a poor estimate might in fact be a very good one—except that it is based on the publicly reported census data. In the aggregate, though, it is reasonable to infer that the estimate that is closer to the census count is better, especially when the difference is large. When the LTDB and NHGIS estimates diverge, the difference is five persons or less in the majority of cases. When the difference is more than five (this is the case in 7,962 tracts), the NHGIS is closer to the census count in about two thirds of the cases, but LTDB is closer in the other third. In more extreme instances where the difference is as much as 100 or more (864 tracts), NHGIS is closer in 86 percent of them, whereas LTDB is closer in only 14 percent. Under what conditions does NHGIS fall further from the mark?

A concern with the NHGIS procedure is that it relies in part on the street grid in 2010 to identify areas that were settled in 2000. In rapidly developing areas on the edge of metropolitan areas, those streets might not have existed in 2000. We illustrate the issue with the example of a tract in Pasco County, Florida, on the northeastern edge of Tampa. A single tract in 2000 (320.02) was split into five tracts in 2010. In two of these tracts, the LTDB estimate for 2000 is very

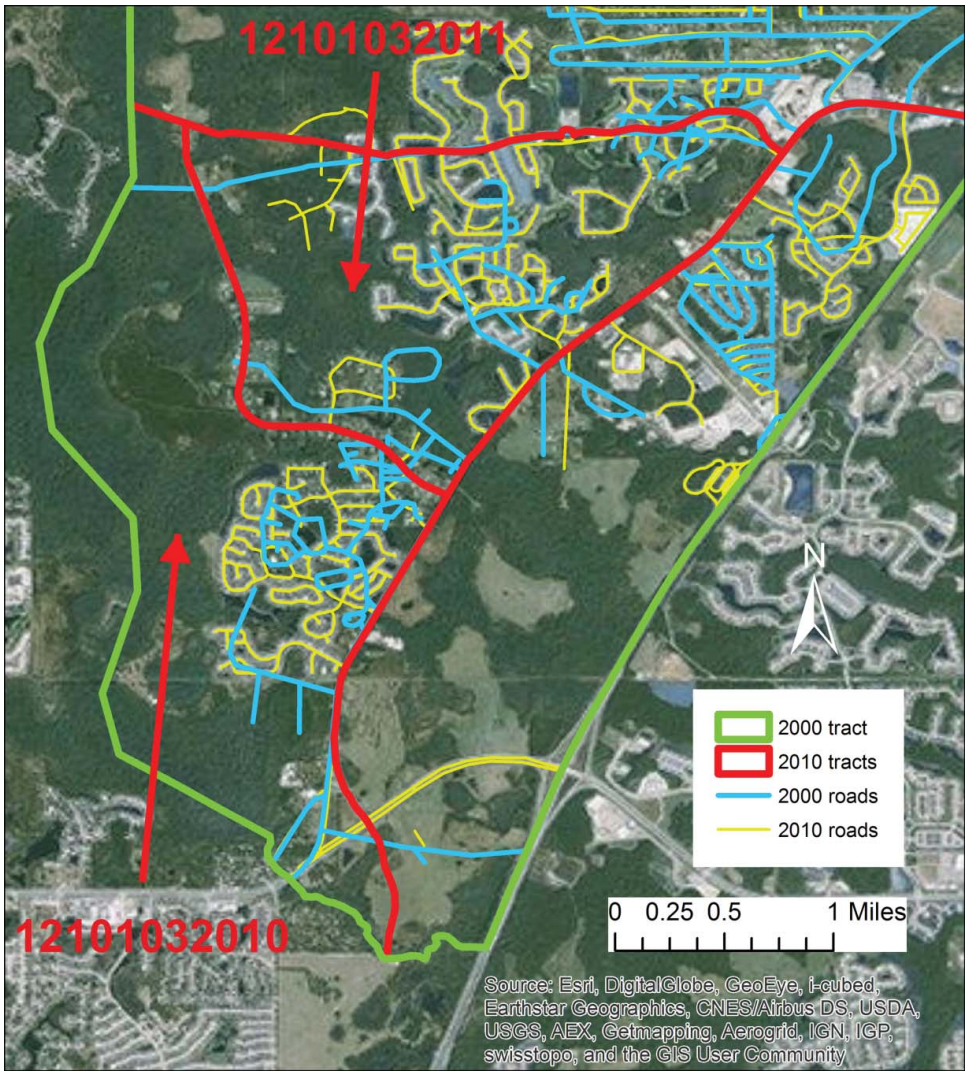

Figure 3. Census Tract 12101032002 in 2000 (Pasco County, Florida), divided into five tracts in 2010. (Color figure available online.)

close to the census count (disparities of two and twenty-six in a total population of over 2,100), but the NHGIS estimate is quite different. For the area in the new Tract 320.10 (population 1,489), NHGIS underestimates at 1,250. For the area in the new Tract 320.11 (population 740), NHGIS overestimates at 1,079.

Figure 3 displays this area, showing the 2010 boundaries of these two new tracts and also the road network as of 2010 and 2000. The position of the 2000 roads is only approximate, clearly offset from the same 2010 roads; this is why they are not used by NHGIS. The 2010 road network aligns much more closely with what appear to be residential structures. The 2000 road network provides clues for an explanation for the error in estimation, however. In Tract 320.10 there is a cluster of roads that were mostly in place in 2000, although some others were built after that time. In Tract 320.11 there is another cluster of roads, but most of these are post-2000. In 2000 both tracts had a very similar total road length (about 3.6 miles). By 2010 Tract 320.10's road length had increased by about 50 percent, and Tract 320.11's road length had nearly doubled.

At the same time, the population of Tract 320.10 (based on census estimates) grew to 2,997 (up 101 percent) and Tract 320.11's population grew to 5,127 (a nearly sevenfold jump). In these circumstances, it is understandable that the NHGIS estimation procedure, which takes into account both the 2010 road network and the 2010 population, will overestimate 320.11's population in 2000 while understating the population of 320.10.

It is beyond the scope of this study to evaluate the NHGIS methodology in greater detail. Very likely the combination of land cover (2001) and road network (2010) is a useful indicator of the settled area in 2000, which is why on average the NHGIS estimates are slightly better than those of the LTDB. It might be that land cover alone in many cases would not provide as much utility as land plus roads. In other cases, such as the Pasco County example, though, the 2010 road network is misleading.

Locating Tracts with Complex Boundary Changes

As already noted, users of cross-sectional census tract data need to be aware of potential effects of errors in population counts that were corrected in postenumeration activities but not published. Where the 2000 population in 2010 boundaries is itself the variable of interest in a longitudinal study, the best source is the census tract population change file referenced earlier. For scholars interested in other variables, the only alternative is to turn to estimates based on interpolation, and these are imperfect. It might be helpful, nevertheless, to understand where errors are most likely to be found. Harmonized data sets intended for general use should include indicators of the type of boundary change that estimates confronted—which tract data are for unchanged tracts, for various categories of restructured tracts, for tracts involving subdivided blocks or not. Clearly the LTDB and NHGIS are the preferred sources, but when using them scholars might find unusual values, and these are most likely in split or many-to-many tracts with divided blocks.

Some locational characteristics offer information about the incidence of such tracts. Unchanged and merged tracts are most prevalent in the New England (79 percent), East North Central (82 percent), and West North Central (83 percent) regions. In contrast, only 55 percent of tracts in the South Atlantic states are unchanged or merged, 62 percent in the Mountain states, and 68 to 70 percent in the East South Central and West South Central regions. Of the most problematic types of restructured tracts, many-to-many tracts with divided blocks are most prevalent in the Middle Atlantic region (16.0 percent) and split tracts with divided blocks in the South Atlantic states (16.2 percent). Another relevant indicator is metropolitan status. Unchanged and merged tracts are least prevalent in suburban areas (66 percent), and suburbs are also where either type of complex change with divided blocks is more likely to be found.

Both of these locational predictors are associated with growth rates, which most likely motivate most changes in tract boundaries. We examined the rate of population change at the county level. In counties with a declining population (loss greater than 2 percent), nearly 85 percent of tracts were unchanged or merged; this compares to only 44 percent of tracts in counties with a growth rate of 25 percent or more. Correspondingly, tracts in counties losing population were least likely to experience complex restructuring. For example, only 3 percent of tracts in declining counties were split into multiple tracts with blocks divided among them, compared to 23 percent of tracts in the fastest growing counties (growth rates of 25 percent or more). These findings reinforce a pattern shown earlier in Table 1, revealing much faster population growth in split and many-to-many tracts.

Conclusion

These results emphasize that estimates are subject to error and that approaches relying solely on areal interpolation (with or without ancillary data on water coverage) are especially error-prone. Because the NCDB is a proprietary system that does not disclose its methods in detail, we are unsure how it handled harmonization to 2010 boundaries. Its estimation errors are of similar magnitude to those found from areal weighting that does not exclude water areas, and surprisingly frequent errors are found even in tracts without boundary changes. In the tracts that involve complex boundary changes (about 30 percent of all tracts), the NCDB yields estimates that are more than 5 percent off in a majority of cases. In contrast, the procedures used by the LTDB (combining areal and population interpolation with exclusion of water-covered areas) and NHGIS (which also incorporates 2001 land cover data and 2010 road networks and population counts) provide good estimates for most of these difficult cases.

This analysis leads to some general recommendations about the use of alternative standard sources for longitudinal tract data. The NHGIS and LTDB are clearly preferable to the NCDB for the 2000 to 2010 decade. Both provide excellent estimates in most cases, even when there have been complex changes in boundaries. The NHGIS estimates are slightly more accurate, although where the estimates from these sources

diverge, the LTDB estimate is sometimes better. The LTDB has two major advantages. First, it offers a tool to harmonize non-Census data from 2000 to 2010 boundaries, which can be important for analysts who have data for variables like crime, foreclosures, or disease that are published at the tract level. Second, it also provides estimates and a crosswalk for census tract data in 1970, 1980, and 1990 (although it relies on simpler interpolation methods for those years).

As applied to census tract data over time, most researchers seek to harmonize data from earlier years to the most recent boundaries. This approach takes advantage of the fact that the number of tracts continues to rise as existing tracts are subdivided in various ways, and the 2010 tract data make possible more precise spatial analysis than the 2000 data. Our results imply, however, that a hybrid approach would produce more reliable estimates. Specifically, use the highly accurate LTDB or NHGIS estimates for tracts with no change, merged tracts, and other cases with no block divisions. For split tracts with block divisions, use a backward interpolation, harmonizing to the original 2000 tract boundaries (a split tract from 2000 to 2010 would be treated as a merged tract from 2010 to 2000).[6] These estimates would also be highly accurate. Finally, in the complex situation of many-to-many tracts with block divisions, the analyst would have two options: (1) use the LTDB or NHGIS estimates for 2010 boundaries, recognizing the potential error or (2) merge together all of the affected tracts in 2000 and 2010 to produce larger areal units with constant boundaries (as in Exeter et al. 2005). It might seem awkward to create a hybrid data set where some tracts are defined by their 2000 boundaries, others by their 2010 boundaries, and others by combinations of tracts in each year, but in some cases it might be advantageous.

Aside from these specific recommendations, a more general recommendation is to be aware of the potential hazards in any census sources. Our work with the "census tract population change file" has alerted us to the fairly extensive revisions of counts that were made after the release of official 2000 census tract data. Another concern applies to the use of the ACS, which replaces the previous decennial long form to provide detailed social and economic data. The ACS is based on smaller samples and has relatively large standard errors for information at the census tract level (National Research Council 2015). The effective sample size for the typical tract for ACS data is less than half as large as was the case for the one-in-six long-form samples from the decennial census. Especially in smaller census tracts, variables like housing tenure and poverty rates are often based on very limited samples with high standard errors. Fortunately, the standard errors for tract data are now routinely disseminated and the research community is beginning to learn how to take unreliability into account.

The same care should be applied to use of estimates of data within constant census tract boundaries. Such data have been available from the NCDB for 1970 to 2000 for over a decade and are now widely used for longitudinal studies and by public and nonprofit organizations. Both the NCDB and LTDB now extend the series through 2010, and NHGIS offers another source for 2000 to 2010. Putting aside the differences between these providers, the results presented here underline the importance of extensive prescreening of data as a first step in analysis. This can be done more readily in studies that are limited to certain geographic areas, such as a sample of cities or counties. In these instances, it is more likely that the analyst will have access to additional ancillary data that can be used to improve estimates or at least to identify large errors. In preliminary data cleaning it is advisable to inspect areas where a higher number of complex boundary changes are clustered. The LTDB identifies the type of tract boundary change for every 2010 tract, which should facilitate such a review. Consistent with findings presented earlier, difficult cases are likely to be found; for example, in fast growing areas where a single large and low-density tract is replaced by multiple, higher density areas created by new housing developments. Although these developments might be relatively homogeneous in the zone of growth, they might also be highly fragmented by price range or housing type. Areas where many people might be located in a specific location (perhaps on only part of a census block), such as group quarters institutions and apartment complexes, also might not be estimated well with interpolation methods.

More specifically and regardless of the source being used, it is advisable to compare the estimated total population in tracts directly with the census retabulations that are now publicly available, identifying tracts that bear closer scrutiny. The retabulated census data are now provided on the Census Bureau Web site.[7] Problems might also be found with other population characteristics, even when the total population estimates are well aligned with the census retabulation. When there are errors in total population estimates,

their effects can be magnified for specific population segments (by personal characteristics such as education or race or by housing characteristics such as single family vs. multifamily). The NCDB, LTDB, and NHGIS presume that all components of the population and housing can be allocated across tracts in the same proportions as the total population (in the case of NHGIS, this involves both the total population and total housing units). This is often a reasonable assumption. To the extent that population and housing characteristics tend to have a spatial pattern that extends beyond a single tract, adjacent tracts will be similar and estimates of their composition will not be much affected by interpolation. Even within relatively small geographic areas (two to four adjacent census tracts that are involved in boundary shifts), however, this "lumpiness" in the data could result in erroneous estimates.

It is not always possible to provide guidance on the nature and magnitude of potential errors in population estimates. The Federal Geographical Data Committee has enacted data accuracy standards for many kinds of geospatial data, but there is no such standard for secondary geodemographic data of the type studied here. The user community is best served by an open-source database with fully documented procedures, allowing future users to refine the approach as new methods or ancillary information become available.

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 National Institute of Child Health and Human Development (5R24HD041020, 5T32HD007338).

Notes

1. Information about the LTDB can be found from Brown University: http://www.s4.brown.edu/us2010/Researcher/Bridging.htm. Information about NCDB is available from Geolytics: http://www.geolytics.com/USCensus,Neighborhood-Change-Database-1970–2000,Products.asp. The NHGIS methodology is summarized at https://www.nhgis.org/documentation/time-series/2000-blocks-to-2010-geog
2. For methodological details, the Geolytics Web page refers users to the documentation of the NCDB's approach to the 1990 to 2000 estimates (http://www.geolytics.com/pdf/Appendix-J.pdf). Many blocks were reconfigured between censuses, and NCDB used ancillary data from the streets coverage from Tiger/Line 1992 to bridge 1990 data to 2000 tract boundaries (Tatian 2003). This is an excellent methodology, but the estimation discrepancies shown here are larger than would be expected if the street network had been used as ancillary data.
3. Data are reported for 72,205 tracts of the 73,057 total tracts in 2010 in the fifty states and District of Columbia. A total of 318 tracts with no land area in 2010 are omitted, although many of these have estimated populations in the NCDB. Of the remaining tracts, 534 were affected by the Census Bureau's Count Question Resolution (CQR) program that resulted in revised population counts of more than 0.1 percent for 2000. In many of these cases, a large group quarters population was shifted from one tract to an adjacent one. The CQR cases are omitted from the analysis because these changes were not available at the time that the LTDB was completed.
4. It takes eleven digits to represent a census tract ID. The first two digits represent the state FIPS code, the next three digits represent the county FIPS code, and the last six digits are the census tract's FIPS code. In the text, after the first mention of a tract, we abbreviate to the last six digits.
5. The comparison of RMSE between the NCDB and where tracts in 2000 methods seems paradoxical. Tables 4 and 5 suggest the discrepancies greater for the NCDB for unchanged tracts, but Table 6 reports values of RMSE for NCDB and areal interpolation including water areas that are identical. This is possible because although NCDB has a higher share of cases with large errors (over 10 percent or larger than 500) the errors are actually larger in magnitude for areal interpolation including water areas (and these deviations are squared in the RMSE). We attach no theoretical or substantive significance to this result.
6. The LTDB provides a "backwards crosswalk" to estimate 2010 population data in 2000 tract boundaries that can be used for this purpose. It is available at http://www.s4.brown.edu/us2010/Researcher/LTDB2.htm.
7. The Census tract population change file is available at https://www.census.gov/population/metro/data/c2010sr-01patterns.html

References

Buttenfield, B. P., M. Ruther, and S. Leyk. 2015. Exploring the impact of dasymetric refinement on spatiotemporal small area estimates. *Cartography and Geographic Information Science* 42 (5): 449–59.

Eicher, C. L., and C. Brewer. 2001. Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science* 28 (2): 125–38.

Exeter, D. J., P. J. Boyle, Z. Feng, R. Flowerdew, and N. Scheirloh. 2005. The creation of "consistent areas

through time" (CATTs) in Scotland, 1981–2001. *Population Trends* 119:28–36.
Fisher, P. F., and M. Langford. 1995. Modelling the errors in areal interpolation between zonal systems by Monte Carlo simulation. *Environment and Planning A* 27:211–24.
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: Standardizing 19th and 20th century census data to allow long-term comparisons. *Computers, Environment and Urban Systems* 26:293–314.
Harvey, J. T. 2002. Estimating census district populations from satellite imagery: Some approaches and limitations. *International Journal of Remote Sensing* 23:2071–95.
Homer, C., J. Dewitz, J. Fry, M. Coan, N. Hossain, C. Larson, N. Herold, A. McKerrow, J. N. VanDriel, and J. Wickham. 2007. Completion of the 2001 National Land Cover Database for the conterminous United States. *Photogrammetric Engineering and Remote Sensing* 73:337–41.
Logan, J. R., Z. Xu, and B. Stults. 2014. Interpolating US decennial census tract data from as early as 1970 to 2010: A longitudinal tract database. *The Professional Geographer* 66 (3): 412–20.
Martin, D., D. Dorling, and R. Mitchell. 2002. Linking censuses through time: Problems and solutions. *Area* 34 (1): 82–91.
Mennis, J. 2003. Generating surface models of population using dasymetric mapping. *The Professional Geographer* 55:31–42.
———. 2016. Dasymetric spatiotemporal interpolation. *The Professional Geographer* 68 (1): 92–102.
Mennis, J., and T. Hultgren. 2006. Intelligent dasymetric mapping and its application to areal interpolation. *Cartography and Geographic Information Science* 33:179–94.
National Research Council. 2015. *Realizing the potential of the American Community Survey: Challenges, tradeoffs, and opportunities*. Washington, DC: National Academies Press.
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. Street weighted interpolation techniques of demographic count estimation in incompatible zone systems. *Environment and Planning A* 37:127–29.
Ruther, M., S. Leyk, and B. P. Buttenfield. 2015. Comparing the effects of an NLCD-derived dasymetric refinement on estimation accuracies for multiple areal interpolation methods. *GIScience & Remote Sensing* 52 (2): 158–78.
Schroeder, J. P. 2007. Target-density weighting interpolation and uncertainty evaluation for temporal analysis of census data. *Geographical Analysis* 39:311–35
Spielman, S., D. Folch, and N. Nagle. 2014. Patterns and causes of uncertainty in the American Community Survey. *Applied Geography* 46:147–57.
Tapp, A. F. 2010. Areal interpolation and dasymetric mapping methods using local ancillary data sources. *Cartography and Geographic Information Science* 37 (3): 215–28.
Tatian, P. A. 2003. *Neighborhood Change Database-NCDB: 1970–2000 tract data. Data users guide*. Washington, DC: Urban Institute.
Wu, C., and A. T. Murray. 2007. Population estimation using Landsat enhanced thematic mapper imagery. *Geographical Analysis* 39 (1): 26–43.

JOHN R. LOGAN is Professor of Sociology at Brown University, Providence, RI 02912. E-mail: john_logan@brown.edu. He directed the US2010 Project through which this research was originally supported, and his research interests include contemporary and historical residential and labor market patterns in U.S. cities, urban change in China, and school segregation.

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. His recent work addresses racial differences in arrest rates and variation in police force size as a result of perceived threat, fear, and prejudice.

ZENGWANG XU is Assistant Professor in the Department of Geography at the University of Wisconsin, Milwaukee, WI 53201. E-mail: xuz@uwm.edu. His primary interests are to investigate the relation between persistent system-level patterns and individual-based processes and the effect of spatiality on the structure and function of evolving complex spatial networks and systems.