

Improving estimates of neighborhood change with constant tract boundaries

John R. Logan^{a,*}, Wenquan Zhang^b, Brian J. Stults^c, Todd Gardner^d

^a Department of Sociology, Box 1916, Brown University, Providence RI 02912, USA

^b Department of Sociology, University of Wisconsin, Whitewater WI 53190, USA

^c College of Criminology and Criminal Justice, Florida State University, Tallahassee, FL 32301, USA

^d U.S. Census Bureau, 4600 Silver Hill Road, Washington DC 20746, USA

ARTICLE INFO

Keywords:

Census tract
Differential privacy
Interpolation
Harmonization

ABSTRACT

Social scientists routinely rely on methods of interpolation to adjust available data to their research needs. Spatial data from different sources often are based on different geographies that need to be reconciled, and some boundaries (e.g., administrative or political boundaries) change frequently. This study calls attention to the potential for substantial error in efforts to harmonize data to constant boundaries using standard approaches to areal and population interpolation. The case in point is census tract boundaries in the United States, which are redefined before every decennial census. Research on neighborhood effects and neighborhood change rely heavily on estimates of local area characteristics for a consistent area of time, for which they now routinely use estimates based on interpolation offered by sources such as the Neighborhood Change Data Base (NCDB) and Longitudinal Tract Data Base (LTDB). We identify a fundamental problem with how these estimates are created, and we reveal an alarming level of error in estimates of population characteristics in 2000 within 2010 boundaries. We do this by comparing estimates from one of these sources (the LTDB) to true values calculated by re-aggregating original 2000 census microdata to 2010 tract areas. We then demonstrate an alternative approach that allows the re-aggregated values to be publicly disclosed, using “differential privacy” (DP) methods to inject random noise that meets Census Bureau standards for protecting confidentiality of the raw data. We show that the DP estimates are considerably more accurate than the LTDB estimates based on interpolation, and we examine conditions under which interpolation is more susceptible to error. This study reveals cause for greater caution in the use of interpolated estimates from any source. Until and unless DP estimates can be publicly disclosed for a wide range of variables and years, research on neighborhood change should routinely examine data for signs of estimation error that may be substantial in a large share of tracts that experienced complex boundary changes.

1. Introduction

Social scientists routinely rely on methods of interpolation to adjust available data to their research needs. Spatial data from different sources often are based on different geographies that need to be reconciled, and some boundaries (e.g., administrative or political boundaries) change frequently. This study calls attention to the potential for substantial error in efforts to harmonize data to constant boundaries using standard approaches of areal and population interpolation. The case in point is census tract boundaries in the United States, which are redefined before every decennial census. We study the accuracy of standard methods of harmonizing such data over time to deal with changing

boundaries. Previous research by Logan et al. (2016) took advantage of the public release of population counts from the 2000 Census using 2010 boundaries, showing that estimates using current methods from the Geographic Information Systems (GIS) toolkit were close to the true values in most tracts. We confirm that finding here, drawing on the original confidential 2000 data in a Federal Statistical Research Data Center (FSRDC). We also compare the “true” and estimated values of a selected set of other population characteristics. We find first that there is much more error in these estimates than in estimates of simple population counts. Second, an alternative approach that injects random noise into the true values, so that they can be publicly disclosed, provides considerably more accurate estimates. And third, there are identifiable

* Corresponding author.

E-mail addresses: john_logan@brown.edu (J.R. Logan), zhangw@uww.edu (W. Zhang), bstults@fsu.edu (B.J. Stults), Todd.K.Gardner@census.gov (T. Gardner).

<https://doi.org/10.1016/j.apgeog.2021.102476>

Received 12 September 2020; Received in revised form 15 April 2021; Accepted 3 May 2021

0143-6228/© 2021 Elsevier Ltd. All rights reserved.

conditions under which the interpolation-based tract estimate is more prone to error and therefore requires closer attention by researchers.

2. Geographical approaches to adjusting boundaries

Geographers have devoted much attention to the effect of discrepancies in the boundaries of the areal units used for the analysis of spatial data. A typical situation is when data are being drawn from different sources. For example, population data may be reported in census tracts, while crime data may be reported in police precincts, or election data in voting districts, or school data in school attendance zones. Another situation, the one we tackle here, is when there are changes over time in the boundaries of the same units (Martin et al., 2002). In either case the general problem is how to estimate data for the same geographic unit so that information from different sources or times can be analyzed together. To do this, data from a source using one zonal system needs to be transferred to a target zonal system for which other data are reported.

A sophisticated areal interpolation approach is to refine the source data to a more detailed or finer spatial scale and then re-aggregate these data to the target zones. Bracken and Martin (1995) describes a surface modeling technique to interpolate the source data into an underlying smooth surface that can then be aggregated to the target zones. The surface can be estimated by point-based interpolation methods using centroids as the representatives of zones (Bracken & Martin, 1989; Martin, 1989) or other statistical methods (Kyriakidis, 2004; Kyriakidis & Yoo, 2005). If the surface has been accurately depicted, data from the surface can be aggregated to any desired areal unit. Unfortunately, this approach works best with variables such as air pollution that are likely to fit well to a smooth surface. With much population data, however, it is common to find discrete boundaries, such as certain major streets or highways that create discontinuities in the spatial distribution of different kinds of people. The “other side of the railroad track” is a prototypical example where there is a clear dividing line between communities based on social class or race.

Another approach to improve areal interpolation is to use other sources of data about the distribution of the population in the source zone (Goodchild & Lam, 1980). This is referred to as dasymetric (or intelligent) interpolation (Maantay et al., 2007; Mennis, 2003; Reibel & Agrawal, 2007; Tapp, 2010; Sleeter & Gould 2007; Wright, 1936; Zandbergen & Ignizio, 2010; Eicher & Brewer, 2001). One such source is the identification of areas covered by water, which can reasonably be presumed to be unpopulated and discounted in areal interpolation. Remote sensing of land cover can identify other areas with no population (Eicher & Brewer, 2001). Data on the road network can also be used as evidence of population density (Reibel & Bufalino, 2005; Xie, 1995), although the presence of roads does not have a one-to-one association with the density of housing. There are two main problems with these sources. First, the computing demands are still very high for all but local studies. Second, remote sensing can more readily distinguish undeveloped from developed areas, but it can only approximate the populations in inhabited areas and it provides no information on other characteristics such as household composition or socioeconomic status that many researchers rely on.

3. Interpolation to harmonize census tract boundaries over time

We consider here the specific case of harmonizing data for census tracts in the United States over time, and we point to a critical source of error in the available interpolated estimates. The earliest national source was the Neighborhood Change Data Base (NCDB, originally developed by the Urban Institute) that first became available in 2002 and was quickly adopted by most social scientists studying tract data in the 1970–2000 period. Its great attraction was that for the first time it promised to allow researchers to reliably study neighborhood change between census years, unaffected by the Census Bureau’s routine revision of census tract boundaries that could either combine two or more

tracts into one or divide a tract into two or more new ones. A study that made use of a special tabulation by the Census Bureau of tract population counts in 2000 within 2010 boundaries (Logan et al., 2016) showed that there was considerable error in the NCDB estimates, but it reinforced confidence in the interpolation methods used by two other research teams. These are the Longitudinal Tract Data Base (LTDB) and an alternative developed by the National Historical Geographic Information System (NHGIS). The analysis showed that the estimates from these latter two sources mostly had a very small margin of error.

Yet for the first time we can now evaluate estimates for other variables, and these are problematic. To understand how interpolation works and where it may fail, we focus on the dasymetric interpolation approach used by the LTDB (for greater detail see Logan et al., 2014). The LTDB relies on 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 Census Bureau (2011), which shows the intersection between blocks and tracts (and many other geographic layers) as defined in the 2000 and 2010 censuses. The first step is to allocate reported tract level population counts in 2000 to blocks within the 2010 tract. A special situation is when part of a 2000 block was reassigned to one 2010 tract and another part to a different tract. Without information on the population in such block subdivisions (which are called fragments), the LTDB estimates the population in each fragment using areal weighting that is solely based on each fragment’s share of the block’s land area. It is then straightforward to aggregate blocks and fragments of blocks to the 2010 census tracts.

The next step is more susceptible to error. Having determined what share of each 2000 tract’s population (the source zone) should be allocated to each 2010 tract area (the target zone), counts of all other population characteristics such as race, income, or educational attainment are allocated according to those same population proportions. These are second-order estimates because they depend on first determining how to allocate total population, and then allocate subcategories of the population (e.g., by race or income) in the same proportions. This procedure assumes that all population subgroups are distributed evenly within the 2000 tract. This may be called the “uniform spatial distribution” (or “spatial stationarity”) assumption. Goodchild et al. (1993, p. 386) discussed this assumption as applied to interpolating population counts in terms of the degree of “homogeneous densities within reporting zones,” which are used as the basis for allocating population to target zones. Where “one can reasonably assume that the source-zone density is constant ... [t]he derivation of the target-zone estimates is straightforward.” They acknowledge that “In general, however, homogeneous densities within reporting zones are exceptions rather than the rule.”

For example, if interpolation methods call for half of the residents of tract Z in 2000 to be allocated to 2010 tract A and half to 2010 tract B, then equally half of Z’s affluent residents and half of Z’s poor residents would go to one of these 2010 tracts. The problem is that residents may actually have been quite segregated by income in 2000 within tract Z. And possibly the more affluent section of the tract was then incorporated into tract A and the poorer section into tract B in 2010. In that case, even if total population were allocated properly to A and B, the allocation by income would be severely distorted. Let us offer a concrete example that draws on analyses that we will present below, where we can compare the LTDB interpolated estimates to what we find to be closer to the “true” values.

Fig. 1 depicts the case of Randall’s Island in New York City, an area that has been undergoing redevelopment since the 1970s. The island was enumerated as a single census tract in 2000, but split into two tracts in 2010. The problem then is to use data from the 2000 census to estimate characteristics of each 2010 tract area in 2000. We present the LTDB estimates of each 2010 tract area’s characteristics in terms of the share with a college degree and the median income of residents in 2000. We also present alternative estimates based on aggregating the original

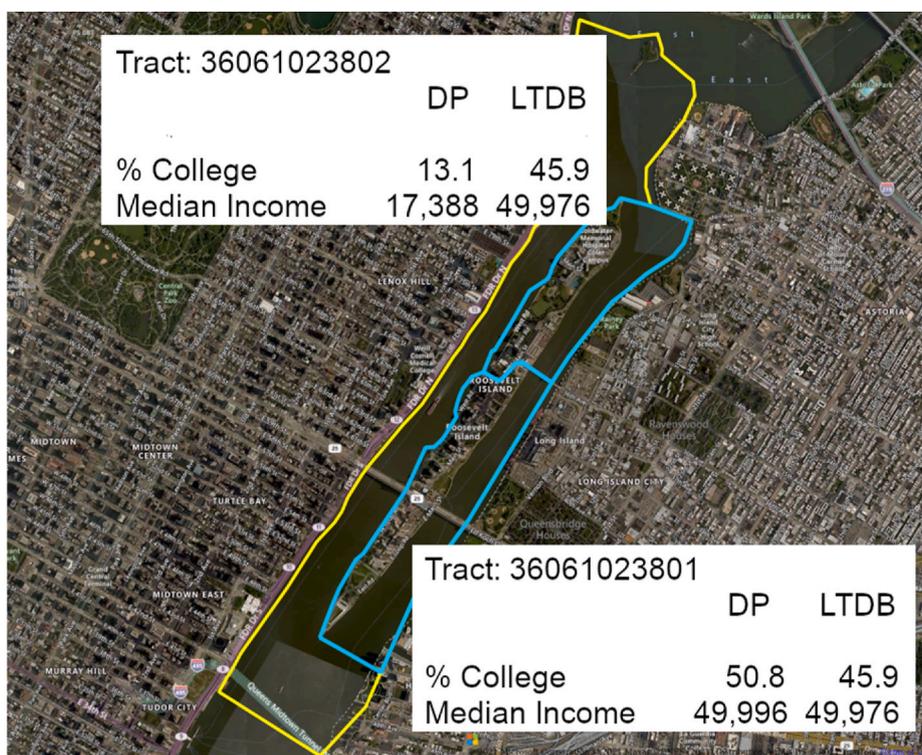


Fig. 1. Population estimates for tracts in Randall’s Island, NY, in 2000, based on 2010 tract boundaries (2000 boundary in yellow, 2010 boundaries in green). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

census microdata to the 2010 tract boundaries, with a small amount of random noise added to protect confidentiality. (This procedure is explained in detail below.) We refer to these as the “true” values.

Aerial photography provided by Google Maps shows that one tract (238.02) includes a nursing home and a new apartment complex that we determined was constructed after 2000 and is otherwise uninhabited. The other tract (238.01) includes a variety of institutions, residential complexes, and commercial developments, and also subway and aerial tramway connections to Manhattan. Large errors arise in the LTDB estimates of income and education levels. Because of the assumption of spatial stationarity, the LTDB assigns both 2010 tract areas the same median income (\$49,976) and college share (45.9%). The true values reveal that tract 238.01’s median income was \$49,996 and its college share was 50.8%. But Tract 238.02, reflecting its nursing home population, had a median income of only \$17,388 and a college share of only 13.1%.

The stationarity assumption is also routinely applied in other applications of interpolation. Working with data over time, if we wish to estimate unobserved neighborhood characteristics in 2005, we may look for the midpoint between the observed data in 2000 and 2010. But if something unusual occurred during the decade – if the smooth curve from 2000 to 2010 were disrupted by a recession and foreclosure crisis – that estimate could be unreliable. When we assess the quality of public schools available to residents of individual census tracts, we may rely on measures for the school district that the tract is within. But if there is much variation among schools in that district, the district-wide measure may be a poor indicator. Nevertheless, in the absence of better alternatives, researchers do their best to accommodate to the constraints of available data sources, trusting that the unknown (and unnoticed) measurement error will not be misleading. Seldom can we test the accuracy of these estimates.

4. Assessing interpolation estimates and a new alternative

We now turn to an effort to gauge the quality of the estimates

provided by the LTDB. Records held in a Federal Statistical Research Data Center (RDC) allow us to determine the 2010 tract area where persons and households lived when enumerated in the 2000 Census, either for short form (intending to cover the full population) or long form (covering one in six households) samples. We can then aggregate these 2000 census records within 2010 geography to provide the best, unbiased estimate of the “true” tract characteristics. The analysis below reports the results of comparing these true values to the estimates from one widely used harmonized data source, the Longitudinal Tract Data Base (LTDB). These statistical summaries have been approved for release by the Bureau’s Disclosure Review Board and other layers of administrative review.

We also evaluate an alternative approach. The Census Bureau has begun relying on noise injection to protect the confidentiality of census data that are released to the public domain. This is a controversial step with regard to the 2020 Census, due to the possibility that noise injection will also have negative effects on the use of census data for political redistricting and other applications where accurate counts of minorities in small areas are essential (e.g., Santos-Lozado et al., 2020). In our application we will show that DP has only small effects on average on estimates at the tract level. We apply noise injection to the true tract characteristics that we compiled in the RDC, and we refer to the resulting estimates as the “differential privacy” (DP) estimates. In the following tables we compare the DP estimates to the true values for a selected set of tract characteristics, parallel to our comparison with LTDB estimates.

DP is described in some detail in a recent working paper by Chetty and Friedman (2019), which explains how their project could reveal results of mobility models for individual census tracts in the U.S. (see also earlier work by Dwork et al., 2006). Differential privacy is intended to ensure that virtually nothing more can be learned about an individual from a dataset than if that person’s record were absent from the dataset. The method involves introducing a fixed level of noise into the tract-level data (which can be counts, means, or medians), meeting a known “privacy threshold” that is represented by the statistic epsilon (ε).

The lower the value of ϵ , the more noise is introduced and therefore the lower the accuracy of the noise-infused estimate, but the greater the protection of privacy.

For counts the procedure is straightforward. Let N_t be the actual count (of white, young, college educated, etc.) in a tract and N_n the noisy version, and let L stand for the LaPlace distribution which is defined in stata. Then.

$$N_n = N_t + L(0, 1 / \epsilon)$$

This transformation is easily implemented in most statistical applications using a single command. The extra noise added or subtracted from the count has a mean of 0, and its “diversity” or variation is equal to the inverse of ϵ – the larger ϵ , the lower the range of the extra noise. For variables that are defined as a median (such as median income, which is included in our analysis) it turns out that the procedure is much more involved. The code is available on request.

One of our goals is to make the DP estimates for all census tracts public so that researchers can evaluate them independently. The Census Bureau reviews the disclosure request, seeking to balance the accuracy of the estimates against the potential loss of privacy, both of which depend on the value of epsilon. We submitted calculations of the level of error (the Root Mean Squared Error, RMSE) of the true values in comparison to the DP estimates for various alternative levels of epsilon ranging from 1 (least accurate and most protected) to 9 (most accurate and least protected). Fig. 2 reports the mean RMSE from 200 separate runs (50 for median income), each beginning with a different random seed and therefore yielding different estimates. The figure includes results for the share of resident who were non-Hispanic white, college educated, under 18, and homeowner. Two other variables, total population and median income, had very low values of RMSE even at an epsilon value of 1. Note that every increment to epsilon from 1 to 9 improved accuracy, but to varying degrees. The error in college education declined noticeably between 1 and 2, but then remained fairly high. The error in other variables declined sharply between 1 and 2, and continued declining up to 4 before leveling off. Based on these plots, the Census Bureau approved release of DP tract estimates based on $\epsilon = 3$.

The error in DP estimates is random, and we show below that it is relatively small. The error in LTDB estimates is larger and not random. But unless and until it is possible to calculate and receive approval for public disclosure of DP estimates for a much larger set of tract characteristics, researchers must rely on interpolated estimates such as those in the LTDB for most purposes. Further, there are other situations where DP estimates are not possible, including estimates in 2010 boundaries for

census data from 1990 and before, and data from non-census sources (e.g., health or crime statistics) that use pre-2010 boundaries (where the interpolation crosswalks provided by the LTDB allow researchers to harmonize other data). Hence it is important to be able to identify census tracts with greater likelihood of error in the estimates.

Because the true values cannot be disclosed and in order to make this analysis as transparent and replicable as possible, we carry it out as a comparison between the LTDB and DP estimates (which are equivalent to the true values with a small random measurement error). The multivariate analyses reported below offer guidance about conditions associated with error in the LTDB estimates. Researchers can identify specific tracts with a probability of higher error by comparing the LTDB and DP estimates that we will disseminate for every tract.

5. Research design

This study includes all populated census tracts in 2000 and 2010 in the continental United States. These tracts can be categorized according to how their 2000 and 2010 boundaries compare. We treat as “unchanged” 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. There are three main categories of changes: consolidations, splits, and complex changes. Consolidation is when several 2000 tracts are combined into one 2010 tract. This creates no difficulties for analysis; the multiple tracts in 2000 can simply be combined into a single tract as defined in 2010. A split (one tract is divided into two or more) adds difficulty. Some rationale is needed to allocate data from one tract into two or more new ones formed within it. More complex changes, where two or more tracts in 2000 are reorganized into entirely different tracts in 2010, are more difficult to deal with. The incidence of these types of change is summarized in Table 1.

A further complication occurs when blocks within tracts are subdivided. As noted above, the LTDB typically allocates population according to the number of residents in each census block that is found in the 2010 tract. But if only part of a block is assigned to that tract, LTDB must instead allocate people according to the share of the block’s land area in the tract. This fallback approach ignores the fact that populations may not be uniformly distributed within blocks.

The type of tract change is a principal predictor of the discrepancy between the LTDB and DP estimates. Note that in the case of unchanged boundaries, the publicly reported data in 2000 boundaries can be used without change for 2010 boundaries. In the case of consolidations, no interpolation is needed for the LTDB estimate; the tract characteristics in

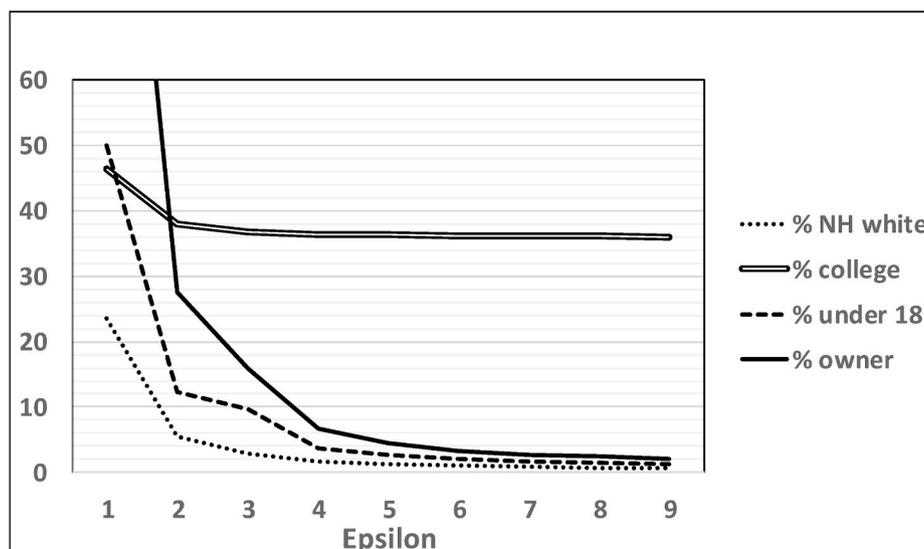


Fig. 2. Estimation error (RMSE) at epsilon values of 1–9.

Table 1

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

| Type of Change | Number | Share | Population (millions) |
|------------------------|--------|--------|-----------------------|
| No change | 49,757 | 68.9% | 200.0 |
| Consolidation | 981 | 1.4% | 3.5 |
| Splits | 12,445 | 17.2% | 43.4 |
| Without divided blocks | 6138 | 8.5% | 21.8 |
| With divided blocks | 6307 | 8.7% | 21.7 |
| Many to many | 9022 | 12.5% | 32.5 |
| Without divided blocks | 2279 | 3.2% | 8.1 |
| With divided blocks | 6743 | 9.3% | 24.4 |
| Total | 72,205 | 100.0% | 279.3 |

Source: Logan et al., 2016.

2010 boundaries can be calculated by combining two or more 2000 tracts. Hence any discrepancy between LTDB and DP estimates is a result of random noise injection in the latter. Researchers may prefer to use the LTDB estimates in these cases, which account for a majority of tracts.

Variable selection. For the purpose of this study we selected several tract-level variables from the 2000 census. Some are short form items (collected for the full population). These are population, the number of non-Hispanic white residents, the number of persons under age 18, number of homeowner households, and number of occupied housing units. From these we calculated the percentage of non-Hispanic white and under 18 persons and percent of homeowners as tract-level variables. Among long form items (based on a one-in-six random sample) we selected the number of college-educated persons along with the number of persons age 25 and above to serve as the denominator for percentage with college education. Finally, to represent a variable that is not based simply on counts, we include the median family income.

Measuring error: Geographers regularly seek to validate estimates or to compare the performance of alternative procedures through comparisons to true data (Flowerdew, Green, and Keris, 1991; Goodchild et al., 1993). We follow the validation procedure in Logan et al. (2016). First we report the distribution of the size of discrepancies between the true value and alternative estimates as a proportion of the actual value. Errors are more likely for more complex changes: split tracts and many to many tracts. Further, in these latter cases error is more likely when blocks have been subdivided. Therefore we report results separately for tracts that experienced these different types of boundary changes.

Table 1 reports on the distribution of proportional error for each variable (comparing each type of estimate to the true value). Table 2 reports a summary measure of this error, the “proportional root mean squared error” which is a variant of the often-used root mean squared error (RMSE):

$$Proportional\ RMSE = \sqrt{\frac{\sum_i ((y_a - y_b)/y_b)^2}{q}}$$

Here y_a is the estimated population of tract i , y_b is the actual population of tract i , and q is the number of tracts. This statistic sums the proportional differences between estimated and actual population counts. Because these values are squared before being summed, the proportional RMSE counts large percentage differences disproportionately compared to small ones. In the following text we refer to this measure simply as RMSE.

Modeling sources of error. The final step in this analysis is to examine what measurable conditions are associated with greater or lesser error in estimation. For this purpose and in order to make this analysis replicable by other researchers, we treat the DP estimate as a proxy for the true value, and we model the size of the discrepancy between the DP and LTDB estimates. We consider only tracts with complex changes, because errors are modest for tracts that experienced no boundary change or simply a consolidation (many merged into one). Such tracts account for about 21,500 or 29.7% of the 72,000+ census tracts in this study.

We have identified three kinds of predictors that in principle could complicate estimation. These are all cases where the “stationarity” assumption of interpolation – that different population subgroups are evenly distributed among portions of tracts that are reallocated to new tract boundaries – may be problematic.

- **Heterogeneity.** The potential for people with different backgrounds to be segregated from one another in different parts of an existing tract is greater if the population is more heterogeneous. Conversely, if the area is home only to a very specific category of people, it is more likely that similar people will be found in all parts of the tract. In that case the spatial stationarity assumption will not be violated. We measure heterogeneity by treating every population characteristic as a dichotomy and calculating the standardized Simpson Index that indicates the probability of randomly drawing two people who are in different groups. For the analysis of error in income, we divided household incomes into categories above and below \$60,000.
- **Size.** When newly created tracts are smaller, they are more susceptible to having a population composition that is different from adjacent areas. The larger the new tract, the more likely that it will be representative of the residents in the local community. Population size is logged in order to give more weight to variation among tracts with smaller populations, which we found have the highest estimation error. Heterogeneity is based on the residents of a 2000 tract that was split in 2010 or, in the many-to-many case, on the aggregate of the initial 2000 tracts that are involved in the new tract boundary.

Table 2

Error (RMSE) in LTDB and DP estimates by type of tract change between 2000 and 2010.

| | N of tracts | Total population | % under 18 | % NH white | % college | % owner | Median income |
|----------------------------------|-------------|------------------|------------|------------|-----------|---------|---------------|
| LTDB vs. true | | | | | | | |
| All tracts | 71,628 | 1.305 | 5.743 | 6.742 | 12.160 | 11.230 | 0.169 |
| Unchanged | 49,639 | 0.132 | 0.163 | 0.069 | 0.172 | 1.232 | 0.042 |
| Consolidated | 979 | 0.047 | 0.074 | 0.152 | 0.125 | 0.069 | 0.098 |
| Split, no divide | 6087 | 0.063 | 7.025 | 11.910 | 16.340 | 21.740 | 0.202 |
| Split with divided blocks | 6159 | 3.669 | 11.100 | 11.520 | 19.250 | 15.490 | 0.379 |
| Many to many no divide | 2243 | 0.079 | 1.072 | 25.250 | 12.410 | 15.950 | 0.520 |
| Many to many with divided blocks | 6521 | 2.419 | 14.120 | 4.704 | 31.190 | 24.880 | 0.184 |
| DP vs. true | | | | | | | |
| All tracts | 71,628 | 0.007 | 1.289 | 0.803 | 1.972 | 3.727 | 0.019 |
| Unchanged | 49,639 | 0.002 | 0.493 | 0.399 | 1.205 | 2.378 | 0.014 |
| Consolidated | 979 | 0.000 | 0.006 | 0.015 | 0.072 | 0.018 | 0.098 |
| Split, no divide | 6087 | 0.003 | 1.976 | 1.887 | 1.411 | 7.877 | 0.017 |
| Split with divided blocks | 6159 | 0.016 | 3.097 | 1.161 | 2.850 | 2.669 | 0.022 |
| Many to many no divide | 2243 | 0.001 | 0.014 | 1.027 | 1.992 | 8.102 | 0.030 |
| Many to many with divided blocks | 6521 | 0.018 | 1.925 | 0.958 | 4.558 | 4.728 | 0.031 |

- **Growth.** Rapid population growth often involves a change in the composition of residents, because new residents may be unlike established residents. We cannot directly measure how fast each newly established tract area is growing, because we are only estimating its population at one time point. However, communities in rapidly growing counties may be more susceptible to changes in composition that could result in different kinds of people being allocated to different new tract areas.

It would be desirable also to be able to take into account information about the spatial distribution of different categories of residents within the origin tract(s). The LTDB does this in the population interpolation component of estimation, relying on block-level population counts. However, this is not feasible for most of the dimensions of neighborhood change that researchers wish to study, because they draw on sample data that are not reported at the block level.

6. Results

6.1. Errors in LTDB and DP estimates

Table 2 summarizes the level of errors in the LTDB and DP estimates in terms of RMSE for every type of tract and for all six population variables. The key finding with respect to the purpose of this study is that the LTDB estimates of total population are much better than estimates of the under 18, non-Hispanic white, college-educated, and homeowner populations. This differential is small for unchanged and consolidated tracts. Here no interpolation was conducted for the LTDB; data were taken directly from the published tabulations. The small discrepancies between the true values and LTDB estimates likely stems from small errors made by the Census Bureau when assigning the 2000 respondents to 2010 tract areas. Larger discrepancies arise for tracts with more complex boundary changes.

The pattern for LTDB estimates is not uniform. First, we note that estimates for median income are generally as good, often better (reflected in a lower RMSE), than for total population. We might infer that tracts where boundary changes are made tend to be more internally homogeneous with respect to income than with respect to other social characteristics; unfortunately, we cannot test this interpretation. Second and surprisingly, in some cases the RMSE for tracts with subdivided blocks is lower than for those without divides. The best example of this is for estimates of non-Hispanic whites, where the RMSE for many to many tracts is six times higher for those with no divides than those with divides.

Now consider the DP estimates. The motivating question is whether noise injection, despite introducing random error into estimates, offers a useful alternative to the interpolation methods underlying systems such as the LTDB. Table 2 reveals that it is not only a useful alternative, but in fact it is clearly superior. The RMSE for DP estimates for population and median income are all below 0.10, a small improvement even upon the very accurate LTDB estimates. It performs least well for percent homeowner, but even here the overall RMSE is 3.7 for DP estimates compared to 11.2 for LTDB estimates. In many comparisons for unchanged and consolidation tracts the error in DP estimates is greater than the error in LTDB estimates, although both are very accurate. This is to be expected, because random noise was introduced into the DP estimates while no interpolation was employed for the LTDB estimates. But in all comparisons of estimates for the tracts with more complex changes, for every variable, the DP estimates have less error.

The RMSE is a summary statistic of how well an estimator performs, on average. It does not reveal how much variation there is in the estimates for different census tracts. For many researchers it may be hard to interpret. How good is an RMSE of under 1.0, and how bad is an RMSE of over 10? To deal with these concerns we have also calculated how the disparity between LTDB or DP estimates with the true value is distributed across census tracts.

Table 3 reports these distributions for a variable on which we conclude that both estimators perform well. Here tracts are categorized by the proportional error in the estimates, ranging from exactly correct to an error of 10% or more (i.e., the ratio of the estimated value to the true value is greater than or equal to 1.10 or less than or equal to 0.90). A typical tract with a true population of 3500 and in the highest category of disparity would have an estimate that is 350 or more above or below that value. Consider first the DP estimates. Very few exactly equal the true value (because error has been intentionally inserted), while almost all are within 1% of the true value. This distribution yields an RMSE of less than 0.02 for every type of tract.

The distribution of disparities is quite different for the LTDB estimates of population. First, in a substantial share of cases (around 10–15% for unchanged and consolidated tracts (though much lower for tracts where boundary changes involved subdivided blocks), the LTDB estimate is exact. Another very large share (in the 55–70% range) are within 1% of the true value. This is why the overall RMSE is just over 1.0. However, in this case there are also many outliers. At the extreme end of the distribution, looking at tracts with block subdivisions, as many as six or seven percent of tracts have an error of 10% or more in the LTDB estimate. This variation indicates that even when an estimator has a high level of accuracy overall, there may be a considerable number of tracts with poorer estimates. A further cause for concern is that error in LTDB estimates is not random, unlike error in DP estimates.

Now let us consider a variable where the LTDB was less successful, the estimates of college educated residents (see Table 4). In this case no DP estimate is exactly correct, but upwards of 97% of estimates are within 1% of the true value. At the upper end of error, for many to many tracts with divided blocks, 1.1% of estimates have an error as high as 10%. For these tracts the corresponding RMSE is 4.6, although the overall RMSE is under 2.0.

The results for LTDB estimates give more cause for concern. On the positive side, the LTDB estimate for about a third of unchanged and consolidation tracts is exact. About 40% more are within 1%. Even for these “uncomplicated” tracts there are outliers, but the distribution corresponds to RMSE of less than 0.2. On the negative side, large shares of estimates for tracts with complicated boundary changes have errors of more than 10% – as high as 60% for split tracts. In a set of split tracts with a true value of 40% college educated, a majority of them would have estimates of 44% or more. This situation corresponds to RMSE in the 15–20 range.

Comparable tables are presented in Appendix A for the under 18, non-Hispanic white, and homeowner shares and for median household income. All these tables reveal the same sharp contrast in performance of the LTDB and DP estimates shown here for the college educated population.

6.2. Predicting estimation error

Although the DP estimates analyzed here are now publicly disclosed and available for use, research on neighborhood change remains dependent on interpolation-based estimates for other neighborhood characteristics. We have shown that such estimates should be viewed with caution. We now consider what are the conditions associated with smaller or greater error in the estimates. It is prudent to use extra vigilance with interpolated data, searching especially for outliers that could represent poor estimates. Data providers should clearly identify which tracts are more subject to error from boundary changes because they involved splits or recombinations. Beyond that there are situations where it seems more likely that local communities are susceptible to distorted estimates from interpolation.

To evaluate the importance of these predictors, we treat the DP estimate as though it were the true value, and measure the absolute value of the discrepancy between it and the LTDB estimate. The same analysis could be carried out within the FSRDC with the original data, but there could be difficulties in disclosure review, because the resulting models

Table 3
Distribution of errors in LTDB and DP estimates: total population.

| | exact | <1% | 1–2.99% | 3–4.99% | 5–10% | >10% | Total |
|----------------------------------|-------|--------|---------|---------|-------|------|--------|
| LTDB vs. true | | | | | | | |
| All tracts | 12.3% | 68.3% | 11.9% | 3.0% | 2.4% | 2.1% | 100.0% |
| Unchanged | 14.0% | 71.4% | 10.0% | 2.1% | 1.5% | 0.9% | 100.0% |
| Consolidated | 11.2% | 66.6% | 13.3% | 3.7% | 2.6% | 2.7% | 100.0% |
| Split, no divide | 16.9% | 65.9% | 10.7% | 2.7% | 2.3% | 1.6% | 100.0% |
| Split with divided blocks | 1.9% | 60.6% | 18.7% | 6.2% | 6.3% | 6.4% | 100.0% |
| Many to many no divide | 20.0% | 61.4% | 10.6% | 3.3% | 2.5% | 2.3% | 100.0% |
| Many to many with divided blocks | 2.8% | 56.7% | 21.4% | 6.7% | 5.4% | 7.1% | 100.0% |
| DP vs. true | | | | | | | |
| All tracts | 0.1% | 99.8% | 0.1% | 0.0% | 0.0% | 0.0% | 100.0% |
| Unchanged | 0.1% | 99.9% | 0.0% | 0.0% | 0.0% | 0.0% | 100.0% |
| Consolidated | 0.0% | 100.0% | 0.0% | 0.0% | 0.0% | 0.0% | 100.0% |
| Split, no divide | 0.1% | 99.7% | 0.1% | 0.1% | 0.0% | 0.0% | 100.0% |
| Split with divided blocks | 0.1% | 99.7% | 0.1% | 0.1% | 0.0% | 0.0% | 100.0% |
| Many to many no divide | 0.0% | 99.9% | 0.1% | 0.0% | 0.0% | 0.0% | 100.0% |
| Many to many with divided blocks | 0.1% | 99.3% | 0.4% | 0.1% | 0.0% | 0.0% | 100.0% |

Table 4
Distribution of errors in LTDB and DP estimates: percent college educated.

| | exact | <1% | 1–2.99% | 3–4.99% | 5–10% | >10% | Total |
|----------------------------------|-------|-------|---------|---------|-------|-------|--------|
| LTDB vs. true | | | | | | | |
| All tracts | 26.5% | 34.9% | 13.5% | 4.8% | 6.2% | 14.1% | 100.0% |
| Unchanged | 36.5% | 43.0% | 13.8% | 3.3% | 2.3% | 1.2% | 100.0% |
| Consolidated | 30.6% | 38.8% | 16.5% | 5.4% | 4.8% | 3.8% | 100.0% |
| Split, no divide | 0.1% | 4.8% | 8.2% | 8.7% | 18.9% | 59.2% | 100.0% |
| Split with divided blocks | 0.0% | 5.1% | 8.5% | 8.4% | 17.8% | 60.0% | 100.0% |
| Many to many no divide | 12.3% | 19.6% | 16.5% | 9.2% | 13.0% | 29.4% | 100.0% |
| Many to many with divided blocks | 4.3% | 34.5% | 19.1% | 7.6% | 10.8% | 23.7% | 100.0% |
| DP vs. true | | | | | | | |
| All tracts | | 98.9% | 0.7% | 0.1% | 0.0% | 0.3% | 100.0% |
| Unchanged | | 99.2% | 0.6% | 0.1% | 0.0% | 0.1% | 100.0% |
| Consolidated | | 98.2% | 1.3% | 0.1% | 0.0% | 0.4% | 100.0% |
| Split, no divide | | 98.6% | 0.7% | 0.1% | 0.1% | 0.4% | 100.0% |
| Split with divided blocks | | 98.1% | 0.9% | 0.1% | 0.1% | 0.8% | 100.0% |
| Many to many no divide | | 97.8% | 1.3% | 0.2% | 0.0% | 0.6% | 100.0% |
| Many to many with divided blocks | | 97.4% | 1.2% | 0.2% | 0.1% | 1.1% | 100.0% |

would provide an indirect way to determine the true values that the Census Bureau will not release. The DP estimate is a useful proxy for the true value, since we showed above that it is typically close to the true value and the error in the estimate is random by design. For variables expressed as a percentage, the discrepancy between the DP and LTDB estimate is the absolute value of the proportional error (the difference in the two estimates divided by the DP estimate). For income, the discrepancy is simply the absolute value of the difference in the two estimates.

Table 5 presents the results of OLS regressions predicting the size of the discrepancy for all tracts that experienced a complex boundary change. The analysis is presented in two steps. The first step (Model 1) includes only the type of change, using the categories applied above. The second step adds heterogeneity in 2000 on the population characteristic that is being predicted, tract population size in 2010 (including all tracts that contribute to the 2000 estimate), and county population growth. Measures of these population predictors are mean-centered.

The intercept in every model is the average estimation error for a split tract with no divided blocks, with the average heterogeneity on the predicted characteristic, average size, and in a county with average population growth. The intercepts are similar in Model 1 and Model 2, and they show that in the average reference tract the LTDB estimate is about 2% different from the DP estimate for under 18 population share, above 4% different for white and college share, and nearly 8% different for homeowner share. The average estimate of median household income has a difference of about \$6000.

There are significant differences among types of tracts. The effect of divided blocks (directly assessed for split tracts) is significant only for

homeowner share and income. For all estimates, errors are smaller in many-to-many tracts than in split tracts. In tracts that experienced these recombinations the negative coefficients reduce the error by a third or more compared to split tracts. This finding is not directly comparable with results from previous tables, but it brings out a different pattern, giving more emphasis to splits vs recombination than to divided blocks within tracts.

In further analysis not presented here, we found that many of the recombinations involved very small adjustments. If we treated tracts as “changed” only if more than 5% of land area was exchanged (vs the 1% in prior studies), the result would be reclassifying a substantial share of “many-to-many” tracts to unchanged. Among the remaining tracts in this category the estimation errors were higher than we found before, but they were still smaller than in split tracts. In the course of this reexamination, we inspected examples of boundary changes in greater detail, even drawing on a database of non-residential land uses (parks, golf courses, and universities) to identify the rationale for redrawing boundaries and the implications for estimation by interpolation. For example, when one new tract derived from a split is devoted primarily to a golf course, it is likely to have fewer residents than one would have estimated; if it houses a university, it is likely to have an unusual population mix. Our impression is that in many cases an estimate could be improved if more information of this type were taken into account, and that is the basis for what geographers refer to as “dasymmetric” methods that rely on ancillary data. Unfortunately, there is not a standard source or guide to use such data at a national scale.

Other predictors in Table 5, especially tract population size, are also useful. Although the type of boundary change is statistically significant,

Table 5
Predictors of the discrepancy¹ between LTDB and DP estimates for tracts with complex boundary changes.

| Change type | % Under 18 | | % Non-Hispanic white | | % College | | % Owner | | Median HH income | |
|------------------------------|------------|---------|----------------------|---------|-----------|---------|---------|---------|------------------|----------|
| | Model 1 | Model 2 | Model 1 | Model 2 | Model 1 | Model 2 | Model 1 | Model 2 | Model 1 | Model 2 |
| Split, no divide (ref) | 0.07 | -0.07 | -0.07 | -0.08 | 0.16 | -0.10 | -0.57 | -0.65 | \$95 | -\$324 |
| Split, divided blocks | -0.98 | -0.94 | -1.75 | -1.87 | -2.00 | -1.73 | -3.78 | -4.00 | -\$2128 | -\$1761 |
| Many to many, no divide | -1.04 | -1.03 | -2.19 | -2.00 | -1.96 | -1.90 | -4.37 | -4.66 | -\$2376 | -\$2318 |
| Many to many, divided blocks | | -1.63 | | -2.80 | | -3.51 | | -3.86 | | -\$4355 |
| Population size (ln) | | -1.37 | | 12.41 | | 9.68 | | 21.66 | | \$15,457 |
| Heterogeneity index | | 0.01 | | 0.01 | | 0.02 | | 0.04 | | \$18 |
| County population growth (%) | 2.31 | 2.34 | 4.24 | 4.19 | 4.35 | 4.38 | 7.80 | 7.94 | \$5991 | \$6179 |
| Intercept | 0.032 | 0.225 | 0.031 | 0.278 | 0.032 | 0.317 | 0.052 | 0.245 | 0.019 | 0.196 |
| R ² | | | | | | | | | | |

Note: The discrepancy¹ is the absolute value of the difference between the two estimates, measured as a percentage of the DP estimate or in dollars.

**p < 0.01; *p < 0.05

this set of dummy variables explains only a modest share (1%–3%) of variance in estimation error. The explained variance in Model 2 rises to a more appreciable 20–30%. Most coefficients are consistent with our expectations, but they represent very small effects.

- Smaller tracts have larger estimation error for all estimates, and this effect is much greater for tracts that are far from the average size of around 3700 persons. As an example, we have calculated the predicted discrepancy for the under-18 share, where the coefficient is smaller than for other population estimates. For a tract with only 100 residents, the net discrepancy (taking into account the intercept), is more than 8%, with 1000 residents it is around 4%, and with average size of 3700 it is only 2%.
- Greater heterogeneity within tracts is associated with greater estimation error for all estimates except under-18 share of residents. Heterogeneity ranges from 0 to a maximum of 0.50, average values for these variables tend to be around 0.30 with a standard deviation of around 0.15, and value that is 0.10 above average would represent a substantial difference. For such a case, the estimation error would decline slightly for under-18 share (0.1 * 1.37) or 0.14%. That is a negligible difference in an unexpected direction. But it would rise more for other outcomes, about 1% for non-Hispanic white and college-educated share, 2% for owner share, and \$1500 for median income.
- Faster population growth at the county level is associated with larger errors (with the exception that this coefficient is not significant for non-Hispanic white share). The average county grew by 15% over the decade, and a growth rate of ten points above that would be substantial. That rate of growth would be associated with less than an 0.5% increase in the estimation error for the four estimates expressed in percentages and a \$180 greater error in median income.

7. Conclusion

The results are clear. Although interpolated estimates of tract “total population” are very reliable, there is even less error in the DP estimates. For other demographic characteristics, interpolation introduces considerable error, while the DP estimates are generally very close to the true values. How great is the problem? In a substantial share of cases for tracts with complex boundary changes, the LTDB estimates differ from the true value by five or ten percent or more.

Fortunately, though they are not random, the LTDB estimates are not systematically biased. Also, despite error in estimation, in a standard multivariate cross-sectional analysis the interpolated measures serve approximately as well as DP estimates, because the two are very highly correlated. In data covering all tracts in the nation, the correlation between these estimates is in the range of .95 or higher. Unfortunately, interpolated measures are more problematic for studies of neighborhood change, which is the situation in which they are needed. Of the variables studied here, the most problematic example is the estimate of owner-occupied housing share. The cross-sectional correlation between the LTDB and DP estimates in 2000 is near-perfect, 0.97. However, if we calculate the change in the percentage of owners between 2000 and 2010 (using the published data for 2010), the correlation between the change based on the LTDB and DP estimates is only 0.66. By implication, in models where change in one characteristic is modeled in relation to change in others, the estimated coefficients can potentially vary substantially depending on which set of estimates is used.

Aside from problems with measuring change, faulty point estimates can also affect other aspects of the cross-sectional distribution of values. Like sampling variability, random error from interpolation will increase the variance of observed characteristics across census tracts. It will also introduce noise into the spatial distribution of population variables, weakening measures of spatial clustering. Further research is needed to understand these effects.

On the basis of our findings, we offer two recommendations. First,

researchers should exercise caution in the use of interpolated data, especially for census tracts that have undergone complex boundary changes over time. Second, to the extent possible, future standard harmonized databases should rely on DP estimates that are feasible now for 2000–2010 and seek to create comparable estimates for other years. It is not yet known whether the Bureau of the Census will allow DP estimates to be disclosed for a broader range of variables. Short of that, there are steps that could be taken to improve interpolated estimates. One change is to apply the spatial stationarity assumption differently for all full-count decennial census variables. Instead of allocating all subgroups of the population to new tract areas in the same proportion as the total population, their allocation could be based on the published block-level counts for each subgroup. A more complex procedure would be needed for other census characteristics such as education level or occupation that are based on sample data, and for which no data are published at the block level. More research is needed on how the full count information for a few variables can be leveraged to improve estimates of others that are known to be correlated with them.

In the short term most researchers will need to rely on the existing interpolated estimates such as those provided by the LTDB and NCDB. The key recommendation is to be aware that these data are estimates, as are all the sample-based population counts that are provided by the decennial census and ACS. As such they are subject to two main sources of error. The first is sampling variation. The Census Bureau makes extensive efforts to improve the precision of estimates through sampling procedures and complex weighting to correct for known bias in sample composition. Yet these steps do not overcome the inherent variability of sample results, which has been exacerbated by the use of smaller samples in the ACS (even when pooled over five years) than were drawn for the decennial census long-form questions through 2000. This source of unreliability in sample-based estimates is expressed in the margins of error that are now more visibly published by the Census Bureau. For studies involving large samples of census tracts, the problem is limited by the tendency for errors to even out. The problem is greater for studies that focus on findings for smaller areas (with relatively few tracts) or for local studies of characteristics of specific tracts. In such research, standard advice is to be aware of outliers that may be due to poor estimation, to be especially attentive to data for tracts with small populations (even to omit such tracts from the analysis), or if possible to shift research to a

larger spatial scale.

In the cases of estimates for harmonized tract boundaries, as we studied here, there are additional considerations for the large minority of census tracts that involve boundary changes. Estimates for these tracts are imperfect, especially for split tracts, and even more so where census blocks have been divided between two new tracts. Error from interpolation is greater for smaller census tracts, which compounds the problem of sampling variability in small tracts. Error is greater in more heterogeneous tracts, and in faster growing counties. It should be standard practice to scan data for outliers, such as tracts where estimates at time 1 differ greatly from those at time 2, or where the estimate for a given characteristic appears out of line with the estimate for another characteristic that is expected to be highly correlated with it. Knowing which tracts have greater potential for estimation error can improve identification of aberrant cases.

Declaration of interest

No potential conflict of interest was reported by the authors.

Data availability statement

The publicly available data analyzed here are available at this website: "<https://s4.ad.brown.edu/Projects/Diversity/Researcher/Bridging.htm>" \o "<https://s4.ad.brown.edu/Projects/Diversity/Researcher/Bridging.htm>".

Acknowledgments

This research was supported by the Sociology Program of the National Science Foundation (grant 1756567). The Population Studies and Training Center at Brown University (P2CHD041020) provided general support. We thank John Friedman (Brown University) and Adam Smith (Boston University) for assistance with differential privacy methods. All results have been reviewed by the U.S. Census Bureau to ensure that no confidential information is disclosed (approval number CBDRB-FY20-208). Any opinions and conclusions expressed herein are those of the author(s) and do not necessarily represent the views of the U.S. Census Bureau.

Appendix A

These tables present the distribution of errors in the LTDB and DP estimates for four variables.

Appendix Table 1

Distribution of errors in LTDB and DP estimates: percent under 18

| | exact | <1% | 1–2.99% | 3–4.99% | 5–10% | >10% | Total |
|----------------------------------|-------|-------|---------|---------|-------|-------|--------|
| LTDB vs. true | | | | | | | |
| All tracts | 9.5% | 63.2% | 9.2% | 4.4% | 5.8% | 7.9% | 100.0% |
| Unchanged | 13.0% | 78.5% | 5.8% | 1.2% | 0.8% | 0.7% | 100.0% |
| Consolidated | 10.0% | 73.1% | 9.6% | 2.6% | 2.8% | 1.9% | 100.0% |
| Split, no divide | 0.0% | 10.0% | 18.2% | 14.9% | 23.2% | 33.7% | 100.0% |
| Split with divided blocks | 0.0% | 11.7% | 18.1% | 14.7% | 23.4% | 32.0% | 100.0% |
| Many to many no divide | 5.5% | 40.6% | 17.3% | 8.6% | 10.6% | 17.3% | 100.0% |
| Many to many with divided blocks | 1.6% | 51.5% | 14.8% | 8.1% | 10.0% | 14.0% | 100.0% |
| DP vs. true | | | | | | | |
| All tracts | | 99.4% | 0.3% | 0.1% | 0.1% | 0.1% | 100.0% |
| Unchanged | | 99.8% | 0.1% | 0.0% | 0.0% | 0.1% | 100.0% |
| Consolidated | | 99.5% | 0.3% | 0.1% | 0.1% | 0.0% | 100.0% |
| Split, no divide | | 98.9% | 0.5% | 0.2% | 0.1% | 0.3% | 100.0% |
| Split with divided blocks | | 98.6% | 0.7% | 0.2% | 0.2% | 0.3% | 100.0% |
| Many to many no divide | | 99.2% | 0.4% | 0.2% | 0.1% | 0.1% | 100.0% |
| Many to many with divided blocks | | 98.0% | 0.8% | 0.2% | 0.3% | 0.6% | 100.0% |

Appendix Table 2
Distribution of errors in LTDB and DP estimates: percent non-Hispanic white

| | exact | <1% | 1–2.99% | 3–4.99% | 5–10% | >10% | Total |
|----------------------------------|-------|-------|---------|---------|-------|-------|--------|
| LTDB vs. true | | | | | | | |
| All tracts | 9.5% | 67.9% | 8.3% | 3.5% | 4.3% | 6.5% | 100.0% |
| Unchanged | 13.1% | 80.6% | 3.8% | 0.9% | 0.8% | 0.7% | 100.0% |
| Consolidated | 9.9% | 69.9% | 10.4% | 3.3% | 3.2% | 3.4% | 100.0% |
| Split, no divide | 0.0% | 22.0% | 21.9% | 12.2% | 16.1% | 27.8% | 100.0% |
| Split with divided blocks | 0.0% | 26.9% | 22.8% | 11.7% | 15.7% | 22.9% | 100.0% |
| Many to many no divide | 5.5% | 45.7% | 14.8% | 7.1% | 8.7% | 18.2% | 100.0% |
| Many to many with divided blocks | 1.6% | 60.2% | 13.5% | 5.9% | 7.4% | 11.5% | 100.0% |
| DP vs. true | | | | | | | |
| All tracts | | 99.0% | 0.7% | 0.2% | 0.1% | 0.1% | 100.0% |
| Unchanged | | 99.0% | 0.7% | 0.2% | 0.1% | 0.0% | 100.0% |
| Consolidated | | 96.6% | 2.9% | 0.1% | 0.1% | 0.3% | 100.0% |
| Split, no divide | | 99.4% | 0.4% | 0.0% | 0.1% | 0.1% | 100.0% |
| Split with divided blocks | | 99.2% | 0.5% | 0.1% | 0.1% | 0.1% | 100.0% |
| Many to many no divide | | 98.5% | 1.1% | 0.2% | 0.1% | 0.1% | 100.0% |
| Many to many with divided blocks | | 98.5% | 0.9% | 0.3% | 0.1% | 0.1% | 100.0% |

Appendix Table 3
Distribution of errors in LTDB and DP estimates: percent homeowner

| | exact | <1% | 1–2.99% | 3–4.99% | 5–10% | >10% | Total |
|----------------------------------|-------|-------|---------|---------|-------|-------|--------|
| LTDB vs. true | | | | | | | |
| All tracts | 12.1% | 62.8% | 7.3% | 3.5% | 4.8% | 9.6% | 100.0% |
| Unchanged | 16.6% | 77.4% | 3.6% | 0.9% | 0.8% | 0.7% | 100.0% |
| Consolidated | 14.7% | 74.4% | 5.7% | 1.6% | 1.7% | 1.8% | 100.0% |
| Split, no divide | 0.1% | 10.4% | 16.0% | 11.1% | 18.4% | 44.1% | 100.0% |
| Split with divided blocks | 0.0% | 13.5% | 19.1% | 12.9% | 19.1% | 35.4% | 100.0% |
| Many to many no divide | 6.8% | 38.7% | 14.9% | 7.0% | 9.0% | 23.7% | 100.0% |
| Many to many with divided blocks | 1.9% | 53.4% | 13.4% | 6.4% | 7.9% | 17.0% | 100.0% |
| DP vs. true | | | | | | | |
| All tracts | | 98.9% | 0.6% | 0.2% | 0.2% | 0.2% | 100.0% |
| Unchanged | | 99.3% | 0.4% | 0.1% | 0.1% | 0.1% | 100.0% |
| Consolidated | | 98.4% | 0.6% | 0.3% | 0.5% | 0.2% | 100.0% |
| Split, no divide | | 98.4% | 0.7% | 0.3% | 0.2% | 0.4% | 100.0% |
| Split with divided blocks | | 98.4% | 0.8% | 0.2% | 0.3% | 0.4% | 100.0% |
| Many to many no divide | | 97.4% | 1.3% | 0.5% | 0.4% | 0.3% | 100.0% |
| Many to many with divided blocks | | 97.1% | 1.3% | 0.3% | 0.5% | 0.8% | 100.0% |

Appendix Table 4
Distribution of errors in LTDB and DP estimates: median household income

| | exact | <1% | 1–2.99% | 3–4.99% | 5–10% | >10% | Total |
|----------------------------------|-------|-------|---------|---------|-------|-------|--------|
| LTDB vs. true | | | | | | | |
| All tracts | 0.1% | 26.7% | 34.6% | 14.9% | 12.3% | 11.4% | 100.0% |
| Unchanged | 0.2% | 32.7% | 41.1% | 15.4% | 8.8% | 1.8% | 100.0% |
| Consolidated | 0.0% | 20.2% | 32.9% | 17.1% | 16.8% | 13.1% | 100.0% |
| Split, no divide | 0.0% | 7.2% | 12.4% | 11.9% | 23.9% | 44.7% | 100.0% |
| Split with divided blocks | 0.0% | 7.5% | 13.6% | 12.0% | 23.6% | 43.3% | 100.0% |
| Many to many no divide | 0.1% | 18.3% | 25.4% | 14.3% | 17.5% | 24.6% | 100.0% |
| Many to many with divided blocks | 0.0% | 21.4% | 29.2% | 15.8% | 15.2% | 18.4% | 100.0% |
| DP vs. true | | | | | | | |
| All tracts | 0.0% | 81.8% | 14.7% | 2.2% | 0.8% | 0.4% | 100.0% |
| Unchanged | | 83.3% | 14.1% | 1.9% | 0.6% | 0.2% | 100.0% |
| Consolidated | 0.0% | 77.0% | 17.2% | 3.9% | 1.2% | 0.7% | 100.0% |
| Split, no divide | 0.0% | 78.0% | 17.1% | 2.9% | 1.4% | 0.6% | 100.0% |
| Split with divided blocks | 0.0% | 77.8% | 16.9% | 2.9% | 1.3% | 1.0% | 100.0% |
| Many to many no divide | 0.0% | 79.9% | 14.7% | 2.8% | 1.7% | 0.9% | 100.0% |
| Many to many with divided blocks | 0.0% | 79.5% | 15.2% | 2.6% | 1.3% | 1.3% | 100.0% |

References

Bracken, I., & Martin, D. (1989). The generation of spatial population distributions from census centroid data. *Environment & Planning A*, 21, 537–543.

Bracken, I., & Martin, D. (1995). Linkage of the 1981 and 1991 UK censuses using surface modelling concepts. *Environment & Planning A*, 27, 379–390.

Chetty, R., & Friedman, J. N. (2019). A practical method to reduce privacy loss when disclosing statistics based on small samples” *NBER Working Paper 25626*. Retrieved from <https://www.nber.org/papers/w25626>.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In S. Halevi, & T. Rabin (Eds.), *Theory of cryptography* (pp. 265–284). Springer Berlin Heidelberg.

- Eicher, C., & Brewer, C. (2001). Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science*, 28, 125–138.
- Flowerdew, R., Green, M., & Kehris, E. (1991). Using areal interpolation methods in geographic information systems. *Papers in Regional Science*, 70, 303–315.
- Goodchild, M. F., Anselin, L., & Deichmann, U. (1993). A framework for the areal interpolation of socioeconomic data. *Environment & Planning A*, 25, 383–397.
- Goodchild, M. F., & Lam, M. (1980). Areal interpolation, A variant of the traditional spatial problem. *Geo-Processing*, 1, 297–312.
- Kyriakidis, P. C. (2004). A geostatistical framework for area-to-point spatial interpolation. *Geographical Analysis*, 36, 259–289.
- Kyriakidis, P. C., & Yoo, E.-H. (2005). Geostatistical prediction and simulation of point values from areal data. *Geographical Analysis*, 37, 124–151.
- Logan, J. R., Stults, B., & Xu, Z. (2016). Validating population estimates for harmonized census tract data, 2000–2010. *Annals of the Association of American Geographers*, 106 (5), 1013–1029.
- Logan, J. R., Xu, Z., & Stults, B. (2014). Interpolating US decennial census tract data from as early as 1970 to 2010. A longitudinal tract database” *The Professional Geographer*, 66(3), 412–420.
- Maantay, J. A., Maroko, A. R., & Herrmann, C. (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., Dorling, D., & Mitchell, R. (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., & Agrawal, A. (2007). Areal interpolation of population counts using pre-classified land cover data. *Population Research and Policy Review*, 26, 619–633.
- Reibel, M., & Bufalino, M. E. (2005). Street weighted interpolation techniques of demographic count estimation in incompatible zone systems. *Environment & Planning A*, 37, 127–129.
- Santos-Lozado, A. R., Howard, J. T., & Verdery, A. M. (2020). How differential privacy will affect our understanding of health disparities in the United States. *Proceedings of the National Academy of Sciences*, 117, 13405–13412.
- Sleeter, R., & Gould, M. (2007). Geographic information system software to remodel population data using dasymetric mapping methods. In United States Geological Survey (Ed.), *Techniques and methods 11-C2*. Reston, VA: U.S. Department of the Interior. <https://pubs.usgs.gov/tm/tm11c2/>. (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, 215–228.
- Wright, J. K. (1936). A method of mapping densities of population, with Cape Cod as an example. *Geographical Review*, 26, 103–110.
- Xie, Y. (1995). The overlaid network algorithms for areal interpolation problem. *Computers, Environment and Urban Systems*, 19, 287–306.
- Zandbergen, P. A., & Ignizio, D. A. (2010). Comparison of dasymetric mapping techniques for small-area population estimates. *Cartography and Geographic Information Science*, 37, 199–214.