Forthcoming: *Professional Geographer*, 2024

Using Public Data to Improve Population Estimates with Consistent Boundaries

John R. Logan, Brown University
Wenquan Zhang, University of Wisconsin, Whitewater
Zengwang Xu, University of Wisconsin, Milwaukee

Acknowledgments

This research was supported by research grants from research grants from National Science Foundation (SES-2147919), Russell Sage Foundation (G-2104-31472), and National Institutes of Health (1R01HD075785-01A1) and by the staff of the research initiative on Spatial Structures in the Social Sciences at Brown University. The Population Studies and Training Center at Brown University (R24HD041020) provided general support. All results have been reviewed by the U.S. Census Bureau to ensure that no confidential information is disclosed (approval number CBDRB-FY23-0244). Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau.

Abstract

Studies of neighborhood change rely on interpolated data to cope with inconsistent boundaries of geographic units over time. The standard approach introduces error by assuming, counterfactually, that all kinds of people are distributed in the same manner within tracts as the whole population. This study evaluates estimates of 2000 neighborhood characteristics using 2010 boundaries in the Longitudinal Tract Data Base (LTDB) that uses the standard approach, and an alternative trait-based (TB) method that uses additional small area data to account for spatial heterogeneity. Both are compared to the true (but confidential) original census data. For variables that are available from full-count census data at the block level (including race, age, and some housing characteristics), the TB estimates are much better than the LTDB estimates. However, the same general approach is ineffective when the small area data are subject to sampling variability and published with less spatial granularity.

Geographic research often relies on spatial data from different sources that are based on different geographies that need to be reconciled. Much attention has been given to interpolation methods to develop estimates of population characteristics within consistent geographic boundaries. The problem we focus on here arises from the assumption of spatial stationarity that is routinely part of the interpolation approach (Goodchild, Anselin and Deichmann 1993). Spatial stationarity means that all areas within a geographic unit have the same composition, whether by race, social class, crime rate,or some other local characteristic. In small and socially homogeneous neighborhoods this assumption is reasonable, but it is often counterfactual, and it can result in poor estimates.

Some evidence of this problem was presented by Logan et al (2021), who compared the estimates of a much-used harmonized census tract data file (the Longitudinal Tract Data Base, LTDB) to the actual values in a confidential Federal Statistical Research Data Center (FSRDC). They found that the estimates of total population were quite accurate, but estimates of other characteristics had considerably more error. Their conjecture was that the LTDB methodology takes account of variation in population density within census tracts, but then assumes that other population characteristics are distributed in the same way, In other words, the spatial stationarity assumption proved to be misleading. They explained the problem with a hypothetical example of a census tract Z in 2000 that was split into tracts A and B in 2010. If usual 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 (assuming spatial stationarity) half of Z's affluent residents and half of Z's poor residents would go to each of these 2010 tracts. Yet 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 distorted.

In this study we test an alternative interpolation approach that relies on data for areas smaller than census tracts for every variable, not just population counts. In principle, then, the new estimates might more faithfully reflect the variation in composition within tracts. We re-estimate all of the 100+ variables in the LTDB for 2000 in 2010 tract boundaries, using the published block or block group data on each variable to generate the estimate for that variable. We then compare both the original LTDB estimates and the new ones to the true (but confidential) values in the FSRDC. We find that for estimates based on full-count data that are published at the block level, the new procedure yields markedly better estimates for many key variables. Improving these estimates is essential because the LTDB has become the most widely used source for longitudinal tract data, having been downloaded more than 15,000 times. However, when the best available small area data are at the block group scale and are based on a sample, the original LTDB estimates are more accurate. We will now make publicly available an updated LTDB that incorporates the estimates that rely on full-count data.

Approaches to interpolation

To explain the alternative procedure, it will be helpful first to describe the usual interpolation methods. The general problem is how to estimate data for the same geographic area so that information from different sources or times can be analyzed together. To do this, data from a source using one zonal system (let us refer to the areas in this system as A, B, … Z) needs to be transferred to a target zonal system (areas a, b, …, z). In areal interpolation this estimate is based on the geographic overlap between areas in the two systems (e.g., people are allocated from counts in “A” to “a” in proportion to the share of area of “A” that lies within “a”). Additionally, there are various approaches to account for the fact that population is not evenly dispersed within these areas.

One alternative is to interpolate the source data into an underlying smooth surface that can then be converted to aggregate estimates for target zones. Tobler (1979) pointed out that the characteristics of adjacent areas tend to be correlated, and he proposed a method to estimate the density of any population characteristic within a geographic unit that considers the characteristics of nearby units. The closer a point in the focal unit is to another unit, the more the estimate of its composition is affected by the neighbor’s composition. The key assumption here is that variations across space fit a smooth curve, rather than having discontinuities at borders within or between geographic units. If this criterion is met, there are various ways to estimate the underlying distribution. For example, the surface can be estimated by point-based interpolation methods using centroids as the representatives of zones (Martin 1989, Bracken and Martin 1989, Kyriakidis 2004, Kyriakidis and Yoo 2005).

Another approach (dasymetric interpolation) relies on direct measurement of variation within every geographic unit to improve estimation. Simple areal interpolation (Goodchild and Lam 1980) can be improved by using other sources of data about the distribution of the population in the source zone. The most commonly available data are indicators of population density. Xie (1995) and Reibel and Bufalino (2005) use information about the road network as indirect indicators of how population is distributed within the area, and the LTDB (Logan, Xu, and Stults 2014, described in more detail below) uses block-level population data for this purpose. These dasymetric methods are a combination of areal and population interpolation. The result can be an improved estimate of the population in the target zones because the approach allows for discontinuous spatial distributions that may reflect real social boundaries between adjacent tracts. But its weakness is that the estimates for all other demographic, social, and housing variables from source to target zones are then allocated in the same way as population counts. In other words, if 40% of people in “A” are estimated to live in “a,” then the same 40% of homeowners, of senior citizens, of non-white people are also allocated from “A” to “a.” That is the problem we address here.

Methods

We study all populated census tracts in 2000 and 2010 in the continental United States. These tracts need to be categorized according to how their 2000 and 2010 boundaries may have changed, because the nature of boundary changes affects the challenges in estimation. A majority (69%) can be treated as “unchanged” (Logan, Xu, and Stults 2014). There are three main categories of changes: consolidations, splits, and complex changes. In consolidation (only 1% of cases) 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, representing 17% of tracts) adds difficulty, and some form of interpolation is needed to allocate data from one tract into two or more new ones formed within it. A variation on splits is a “many-to-many” change, where two or more tracts in 2000 are reorganized into two or more entirely different tracts in 2010 (13% of tracts).

For both of these more challenging types, when the changes involve simply a recombination of whole census blocks, the LTDB relies on population interpolation. That is, it determines which whole blocks are in each 2010 tract and uses block population counts to calculate what share of every 2000 tract's population lived within the borders of every 2010 tract that it overlaps with. The LTDB now assumes that this share is a constant, applicable to all census counts, whether characteristics of persons, households, or housing units. Using the example above, if 50% of the population in a 2000 tract lived in the territory of a given 2010 tract, then 50% of its affluent residents are estimated to have lived within that 2010 tract. And the 2000 affluent population of that 2010 tract is the sum of the numbers allocated in this way from every 2000 tract that overlaps with it. The resulting total population calculation should be correct, but the interpolated estimates of all other attributes are susceptible to error.

More than half of the split tracts and many to many tracts involve a further complication because census blocks have been divided between two or more tracts. If only part of a 2000 block is assigned to a given 2010 tract, LTDB must rely on some other information to allocate the block's population. The solution is to allocate population according to the share of the block's land area in the tract. That is, if 20% of the block's area lies within a 2010 tract, then 20% of its population is allocated to that tract. This reliance on areal interpolation may add error to the resulting estimates because it ignores the fact that populations may not be uniformly distributed within blocks, but land area is the only measure of how blocks have been divided that is available from the census.

In this study we investigate a variation of the usual approach. Rather than calculate a constant allocation share for every variable based on the block-level population data, we develop **trait-specific** allocation shares for every variable, using the publicly available information at a scale smaller than the tract. Otherwise the procedure is the same as used in the LTDB. Like the LTDB, it relies on areal interpolation to allocate characteristics from divided blocks. We refer to this as a "trait-based" (TB) approach.

Different methods of interpolation can yield quite different estimates. To illustrate this point and clarify how estimates are derived, we offer a hypothetical example. Figure 1 and Table 1 refer to a case of a single tract (10) at time 1 that was split into two tracts at time 2 (10.1 and 10.2). The goal is to estimate the Hispanic population in 10.1 and 10.2 at time 1. Figure 1 shows that Tract 10.1 contains blocks A and B and part of block C (C_1), which has been divided between the two new tracts. Tract 10.2 includes block D and the remainder of block C (C_2). Table 1 provides the available information that is used by different interpolation approaches. The LTDB uses data the area of each original and each divided block, plus the total population of each block (highlighted in green). In this case blocks A, B, C, and D each have 100 residents. Since tract 10 was known to have 100 Hispanic residents (a 25% share), the LTDB assumes that each block had 25 Hispanic residents. Then area interpolation is used to estimate how the 25 Hispanics in C were divided between C_1 and C_2. C1 contained 30/80 of the area of C, so that proportion of C's Hispanics is allocated to C_1 (30/80 times 25 = 9.375), and the rest to C_2 (15.625).

The trait-based method uses additional data on the Hispanic population in each original block. The table shows, for example, that block A was only 10% Hispanic while block D was 50% Hispanic. Based on this information, the TB approach can directly allocate the 10 Hispanics in block A and 20 in B to 10.1, and 50 Hispanics in block D to 10.2. The 20 Hispanics in C are then divided between C1 and C2 in proportion to the land area of each, 7.5 to C1 and 12.5 to C2. The TB estimate in this case is closer to the true (but unknown) value, because it takes into account the high concentration of Hispanics in block D, and it allocates the Hispanic population accordingly more to 10.2 than to 10.1.

In the best case, the TB approach could estimate every population characteristic as accurately as it could estimate total population. In practice, there are two forms of available small area information for the trait-specific approach. The decennial census in 2000, like previous years, involved two different questionnaires. The "short form" was intended to be completed for a 100% "full count" of the population, and tabulations from the full count data were published at the census block level. These included total population, race (including several Hispanic and Asian categories), age, race by age, vacant and occupied housing units, and owner and rental housing units. These variables are essential to much neighborhood research, representing major dimensions of residential differentiation by age, race, and social class. The other questionnaire, referred to as the "long form," included a wider array of questions such as family composition, labor force participation, income and education. The long form was completed by a one-in-six sample of the population, and the smallest geographic unit for which tables were published was the block group (typically a set of 6-8 blocks). For trait-based interpolation, therefore, we use short form, block data for some variables and long-form, block group data for many others, and it turns out that this difference is consequential.

We evaluate the quality of estimates in two ways that have been reported in prior studies (e.g., Logan, Xu, and Stults [2014]). The first is a standard summary measure of the error in estimation, the "proportional root mean squared error" which is a variant of the often-used root mean squared error (RMSE):

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

Here y_a is the estimated population of tract *i*, y_b is the actual population of tract *i* based on confidential FSRDC data, 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. The second measure is the proportion of tracts where the estimated value varies by more than 10% from the true value. This measure is a more intuitive indicator of accuracy, and researchers may be especially concerned about the effects of these larger errors.

We report these measures separately for tracts that underwent different kinds of boundary changes. In principle we expect to find more accurate estimates for tracts that had no change in their boundaries and those where multiple tracts were consolidated into one. In these cases, the LTDB and TB estimates should be similar. Greater inaccuracy is likely to be found in split tracts and cases of complex reorganization, especially when blocks have been divided.

For both measures we provide results in two forms. The first is graphic illustrations, which provide only two sets of average values (Figures 2-3). One is a set of 14 core demographic variables that are available from 100% census samples at the block level and for which we conclude that the TB estimates are more reliable than the LTDB estimates. We compare these to a set of 54 long-form variables, based on sample data and reported by the census for block groups, not blocks. We find that for these variables the TB estimates are no better and often less accurate than the LTDB estimates. The detailed results for all of the variables in each of these sets, and also for other 100% variables for which the TB and LTDB approach is more similar, are presented in Appendix Tables 1-2. (There are some other LTDB long-form variables, such as measures of poverty. that could not be evaluated because they often have missing or zero values in the confidential FSRDC files, and they are omitted from the study.)

Findings

Consider first results for the 14 core demographic variables, for which the average errors in estimation (RMSE) ae reported in Figure 2 and the proportion of large errors (over 10%) in Figure 3. Results in Figure 2 show that the TB greatly outperforms the LTDB in split tracts with no divide and many-to-many tracts with no divide. It performs modestly better for split tracts with divided blocks, and somewhat worse for many-to-many tracts with divided blocks. Results in Figure 3 show a more uniform advantage of the TB estimates, and by larger margins. In all of the comparisons involving split or many-to-many tracts, the average share of cases with errors as large as 10% is considerably greater in the LTDB estimates. For these variables we can clearly recommend using the trait-specific estimator: counts of the total population and the major categories of race/ethnicity as defined in the LTDB (Hispanics and non-Hispanics who are white, black, Asian/Pacific Islander, and Native American), the number of families, number of housing units and their distribution by occupancy and tenure, and the distribution of the population by age (60+ and 75+).

In contrast, results for the set of 54 long-form variables show an overall advantage for the LTDB estimates. Figure 2 shows that the average RMSE for these variables is much better for all tracts, and also for split or many-to-many tracts with divided blocks, and modestly better for other types of tracts. Figure 3 shows the same pattern, except that the average share with large errors is slightly smaller for the TB estimates for split tracts with no divide.

These are the main findings that support our conclusion that the TB estimation method – despite taking advantage of information about spatial heterogeneity within census tracts – is not always superior to the LDTB. More detailed results are presented in the two appendix tables, and they are consistent for all the core demographic variables in the top panel.

The appendix tables include two additional panels with short-form variables. One is a set of 9 variables showing the main national-origin categories of Hispanics and Asians (e.g,, Mexicans, Chinese). The other is categories of age (15+ and 60+) for specific racial/ethnic groups. In most cases the LTDB and TB estimates perform equally in every type of tract for these population characteristics. In addition, the tables report the mean values for the set of 54 long-form variables, including indicators of education, income, labor force position, marital status, home values and rents, and nativity and ancestry. These are reported as average values here because there are so many variables and their patterns are similar. In a large majority of cases the LTDB estimator is more accurate, often substantially so, while in some cases the two estimators are similarly accurate, and in rare cases the TB estimator outperforms the LTDB.

Conclusion

This study demonstrates both the potential and the limitations of using additional block or block-group data to improve estimates of population characteristics within consistent boundaries. The improvements for 100% data at the block level substantiate the conjecture by Logan et al (2021) that errors in the LTDB stem at least in part from its assumption of spatial stationarity. However, the poor performance of the alternative TB procedure for other variables suggests that sample data at the block group scale are not sufficiently robust to correct for variation in how those characteristics are distributed within tracts.

The data file that will be made public will now provide the TB estimates for the short-form variables studied here for all census tracts in the nation. Based on this research we do not recommend using the TB estimates for long-form variables. For these variables, the available within-tract data are sample data for block groups, which means the estimates are affected by sampling variability and they are not as geographically precise as block data. In contrast, we do recommend using the trait-based estimator for all the short-form variables, even though for many variables the TB and LTDB have similar accuracy, and in some cases the LTDB performs better. In practice, users are likely to operationalize the within-race national origin and age data as rates. For example, the number of Black residents is likely to be converted to Black share of the tract population. In this case, the denominator to calculate the Black share – population – is the same for both estimators. That is because the LTDB used block population counts in the estimation procedure. For the national-origin and group-specific age variables, however, the better denominator is the TB estimator (e.g., number of Hispanics), and it would be more consistent also to use the TB estimator for the numerator (e.g., number of Puerto Ricans, yielding Puerto Rican share of Hispanics). The trait-based interpolation approach has important advantages over the traditional interpolation methods used in systems like the LTDB. This advantage depends, however, on the reliability and geographic specificity of the available within-tract data.

This study deals with estimation of 2000 population characteristics in 2010 census tracts. It highlights the difficulty in estimating long form variables that arises from sampling variability and (related to this) the Census Bureau's decision to use block groups as the smallest areal unit for published tabulations of these variables. The results are directly relevant to estimates of characteristics in prior decennial censuses that used the same methodology. Because block and block-group tabulations are already available for 1990, the TB approach can be applied to that year. The National Historical GIS (NHGIS) Project at the University of Minnesota has begun disseminating these data along with block maps for 1980, and NHGIS plans to do the same for 1970. At this time the prospects for 1960 are poor, because the original printed block maps are readily available only for central cities. For earlier years, fortunately, the 100% microdata including addresses are now available for all decades 1900-1950, and efforts to geocode these data are underway in the Urban Transition GIS Project at Brown University. When that process is completed, interpolation will not be needed for those decades, because it will then be possible to aggregate the original point data to any desired areal units.

On the other hand, looking forward, the decennial long form has been replaced by the American Community Survey (ACS), which introduces new difficulties for the estimation of characteristics in consistent boundaries. The ACS has a much smaller sampling rate than the decennial long form (about 7-8%), and that level is achieved only be cumulating annual ACS survey data across five years. Further, the Census Bureau has introduced new procedures for protection of confidentiality in published data, which it refers to as "differential privacy" estimates In addition to greater sampling variability in the ACS, researchers will now have to deal with random error that is being introduced to both decennial census and ACS tabulations. It is unclear how useful these tabulations will be at areal units smaller than census tracts, and it will be a challenge for researchers to discern how best to use these new data products.

References

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

Eicher, C. L. & C. A. Brewer. 2001. Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science* 28**:**125-138.

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

Goodchild, M. F., L. Anselin & U. Deichmann. 1993. A framework for the areal interpolation of socioeconomic data. *Environment and Planning A* 25: 383-397.

Kyriakidis, P. C. & E.-H. Yoo. 2005. Geostatistical prediction and simulation of point values from areal data. *Geographical Analysis* 37**:**124-151.

Logan, J. R., W. Zhang, B. J. Stults, & T. Gardner. 2021. "Improving Estimates of Neighborhood Change with Constant Tract Boundaries" *Applied Geography* 132:1-11.

Logan, J. R., Z. Xu, & B.J. Stults. 2014. "Interpolating US Decennial Census Tract Data from as Early as 1970 to 2010: A Longitudinal Tract Database" *The Professional Geographer* 66(3):412-420.

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

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

Reibel, M. & A. Agrawal. 2007. Areal interpolation of population counts using pre-classified land cover data. *Population Research and Policy Review* 26**:**619-633.

Tobler, Waldo R. 1979. "Smooth Pycnophylactic Interpolation for Geographical Regions" *Journal of the American Statistical Association* 74: 519-530.

U.S. Census Bureau. 2011. *2010 Relationship Files Technical Documentation*. U.S. Census Bureau. https://www.census.gov/programs-surveys/geography/technical-documentation/complete-technical-documentation/relationship-files-overview.html, accessed 6/16/23.

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

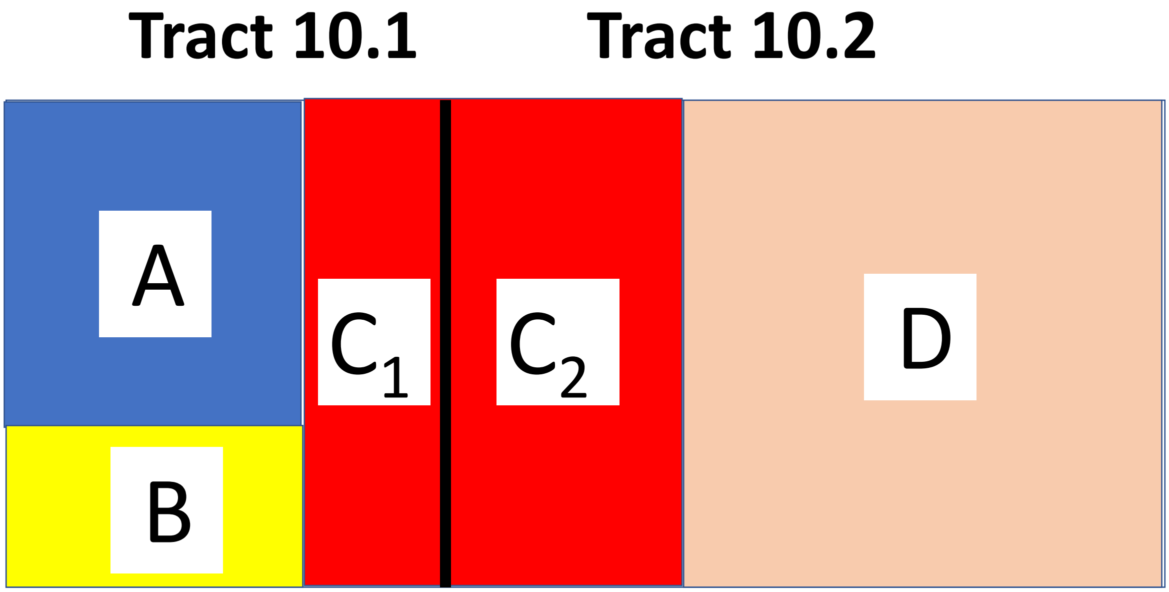

Figure 1. Hypothetical split tract with one divided block

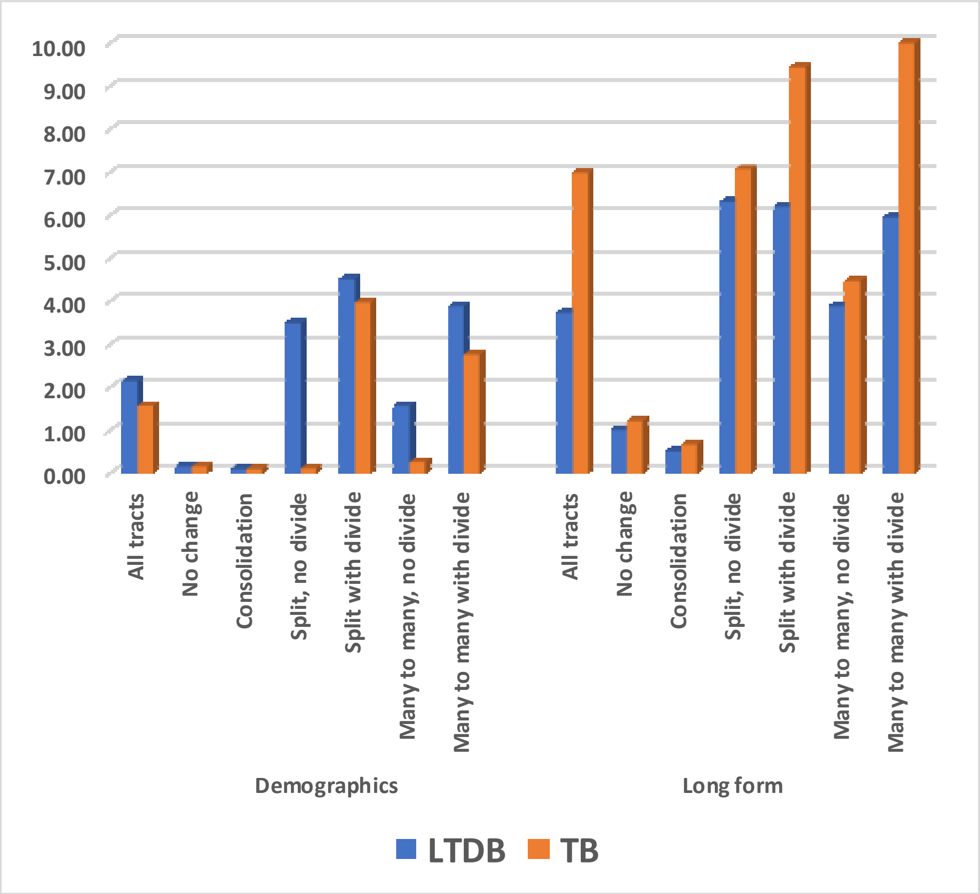

Figure 2. Average errors in estimates (RMSE), LTDB vs TB, comparing 100% demographic variables and sample variables (see Appendix Table 1)

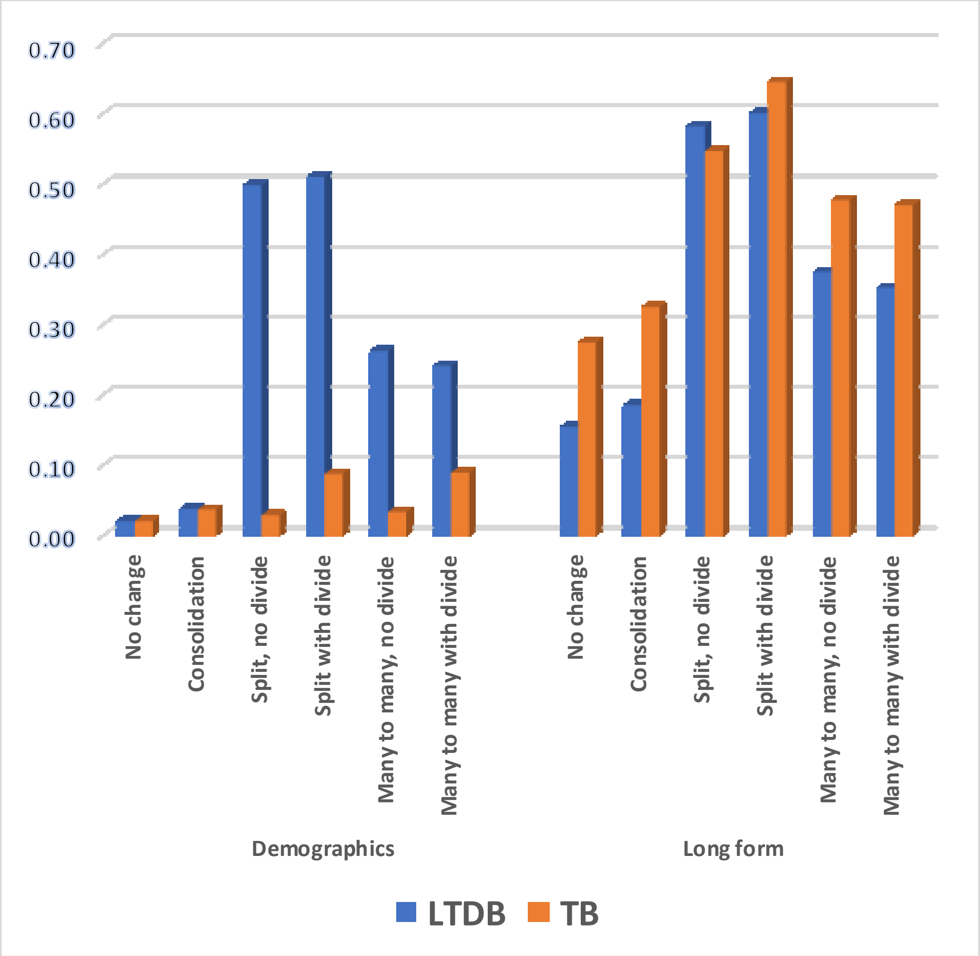

Figure 3. Average share of estimates with errors above 10%, LTDB vs TB, comparing 100% demographic variables and sample variables (see Appendix Table 2)

Table 1. Hypothetical illustration of LTDB and TB interpolation methods

	Known parameters:				Hispanic estimate :	
	Area	% area	Total population	Hispanic population	Block population + area	Hispanic population + area
Tract 10						
Block A	36	15.0%	100	10		
Block B	24	10.0%	100	20		
Block C	80	33.3%	100	20		
Block D	100	41.7%	100	50		
Total	240	100%	400	100		
Tract 10.1						
Block A	36	15.0%	100	10	25	10
Block B	24	10.0%	100	20	25	20
Block C_1	30	12.5%			9.375	7.5
Total	90	37.5%			59.375	37.5
Tract 10.2						
Block C_2	50	20.8%			15.625	12.5
Block D	100	41.7%	100	50	25	50
Total	150	62.5%			40.625	62.5

Appendix Table 1. RMSE, LTDB vs Trait-based (TB) estimates

	All tracts		No change		Consolidation		Split, no divide		Split with divide		Many to many, no divide		Many to many with divide	
	LTDB	TB	LTDB	TB	LTDB	TB	LTDB	TB	LTDB	TB	LTDB	TB	LTDB	TB
Population	1.31	1.31	0.13	0.13	0.05	0.05	0.06	0.06	3.67	3.67	0.08	0.08	2.42	2.42
Non-Hispanic white	1.63	1.63	0.14	0.14	0.17	0.17	0.64	0.06	3.42	3.52	0.52	0.12	4.19	4.17
Non-Hispanic black	2.41	1.86	0.43	0.43	0.17	0.17	2.33	0.21	5.19	1.94	2.06	1.97	5.50	5.63
Hispanic	1.32	1.10	0.53	0.54	0.07	0.07	1.07	0.10	2.91	2.03	0.65	0.31	2.77	2.64
Asian/Pacific Is	1.03	0.80	0.23	0.23	0.20	0.20	1.20	0.17	2.32	1.42	1.19	0.40	2.05	2.15
Native American	0.60	0.30	0.30	0.30	0.16	0.16	0.88	0.17	1.39	0.28	0.65	0.21	0.75	0.44
Families	3.18	2.55	0.04	0.04	0.04	0.04	5.60	0.06	7.18	8.31	2.00	0.05	5.65	2.45
Housing units	3.70	2.69	0.04	0.04	0.05	0.05	8.94	0.06	7.68	8.87	2.18	0.04	4.24	2.25
Owner	3.83	1.73	0.08	0.08	0.07	0.07	8.09	0.09	6.34	2.89	3.29	0.09	7.60	4.99
Rental	3.42	3.00	0.10	0.09	0.09	0.09	5.44	0.26	9.02	9.87	1.72	0.06	4.83	2.62
Occupied	3.99	2.68	0.04	0.04	0.05	0.05	8.78	0.06	7.57	8.78	2.01	0.04	6.88	2.49
Vacant	1.06	0.67	0.08	0.08	0.05	0.05	1.18	0.11	2.45	1.91	2.13	0.11	1.94	1.18
Age 60+	1.13	0.78	0.05	0.05	0.08	0.08	1.87	0.08	2.07	1.37	2.07	0.08	2.30	2.21
Age 75+	1.37	0.94	0.09	0.09	0.11	0.11	2.66	0.13	1.88	0.77	1.39	0.19	3.17	3.02
MEAN	**2.14**	**1.57**	**0.16**	**0.16**	**0.10**	**0.10**	**3.48**	**0.12**	**4.51**	**3.97**	**1.57**	**0.27**	**3.88**	**2.76**
National-origin groups														
Mexican	1.36	1.36	0.50	0.49	0.17	0.17	1.94	1.95	3.54	3.54	1.29	1.30	1.55	1.55
Puerto Rican	1.06	1.06	0.39	0.39	0.35	0.35	1.13	1.13	1.47	1.47	0.84	0.84	2.78	2.78
Cuban	0.97	0.98	0.77	0.77	0.57	0.57	1.61	1.62	1.40	1.41	0.86	0.87	1.14	1.14
Chinese	1.06	1.03	0.47	0.48	0.52	0.54	1.56	1.51	2.49	2.39	1.28	1.21	1.41	1.35
Filipino	0.77	0.78	0.43	0.43	0.45	0.45	1.26	1.27	1.40	1.41	1.01	1.01	1.21	1.22
Indian	1.07	1.08	0.46	0.47	0.51	0.51	2.12	2.13	2.00	2.02	1.56	1.57	1.44	1.45
Japanese	0.73	0.74	0.49	0.49	0.56	0.56	1.20	1.21	1.14	1.16	1.06	1.07	1.04	1.05
Korean	1.17	1.18	0.46	0.46	0.57	0.57	1.36	1.36	1.61	1.62	1.82	1.82	2.85	2.86
Vietnamese	0.99	0.99	0.62	0.62	0.68	0.68	1.61	1.62	1.72	1.73	1.50	1.51	1.34	1.35
MEAN	**1.02**	**1.02**	**0.51**	**0.51**	**0.49**	**0.49**	**1.53**	**1.53**	**1.86**	**1.86**	**1.25**	**1.24**	**1.64**	**1.64**
Age for subgroups														
Age 15+, white	3.62	3.63	0.17	0.16	0.27	0.27	4.41	4.42	8.11	8.13	1.92	1.89	7.88	7.90
Age 15+, black	3.07	3.08	0.33	0.33	0.31	0.31	3.43	3.49	7.45	7.41	2.92	2.90	6.05	6.08
Age 15+, Hispanic	2.49	2.48	0.14	0.14	0.51	0.51	1.86	1.81	6.59	6.57	1.65	1.63	4.75	4.77
Age 15+, Asian	1.09	1.08	0.39	0.40	0.50	0.50	1.78	1.83	1.93	1.91	1.81	1.86	2.04	1.97
Age 15+, Native Ame	0.90	0.96	0.55	0.55	0.60	0.60	1.53	1.77	1.01	1.23	1.50	1.54	1.60	1.63
Age 60+, white	1.77	1.73	0.10	0.10	0.26	0.26	4.04	3.98	2.04	2.02	1.38	1.30	3.79	3.71
Age 60+, black	1.21	1.19	0.46	0.46	0.49	0.49	2.82	2.82	1.95	1.81	1.70	1.66	1.53	1.56
Age 60+, Hispanic	0.93	0.94	0.28	0.28	0.30	0.30	2.72	2.75	0.96	1.04	1.30	1.35	0.65	0.64
Age 60+, Asian	0.63	0.68	0.47	0.47	0.55	0.55	0.97	1.17	0.90	1.08	0.77	0.82	0.89	0.88
Age 60+, Native Ame	0.74	0.79	0.65	0.65	0.65	0.65	0.78	1.02	0.96	1.10	0.87	0.99	0.96	1.00
MEAN	**1.64**	**1.66**	**0.35**	**0.35**	**0.44**	**0.44**	**2.43**	**2.50**	**3.19**	**3.23**	**1.58**	**1.59**	**3.01**	**3.01**
Long Form Variables (n=54)														
MEAN	**3.73**	**6.99**	**1.00**	**1.21**	**0.51**	**0.67**	**6.32**	**7.08**	**6.20**	**9.44**	**3.88**	**4.46**	**5.95**	**18.37**

Appendix Table 2. Share with error greater than 10%, LTDB vs Trait-based (TB) estimates												
	No change		Consolidation		Split, no divide		Split with divide		Many to many, no		Many to many with	
	LTDB	TB	LTDB	TB	LTDB	TB	LTDB	TB	LTDB	TB	LTDB	TB
Population	0.009	0.009	0.026	0.026	0.016	0.016	0.064	0.064	0.023	0.023	0.071	0.071
Non-Hispanic white	0.014	0.013	0.047	0.047	0.285	0.016	0.267	0.065	0.192	0.029	0.150	0.074
Non-Hispanic black	0.045	0.044	0.074	0.074	0.738	0.054	0.770	0.134	0.382	0.050	0.356	0.129
Hispanic	0.029	0.028	0.051	0.049	0.638	0.036	0.678	0.105	0.315	0.036	0.296	0.104
Asian/Pacific Is	0.046	0.045	0.082	0.078	0.725	0.049	0.743	0.115	0.384	0.060	0.335	0.118
Native American	0.044	0.043	0.082	0.078	0.715	0.054	0.716	0.131	0.401	0.074	0.350	0.132
Families	0.006	0.006	0.012	0.012	0.126	0.013	0.160	0.057	0.082	0.020	0.103	0.062
Housing units	0.008	0.008	0.014	0.014	0.257	0.017	0.296	0.064	0.142	0.017	0.152	0.067
Owner	0.007	0.007	0.017	0.017	0.438	0.014	0.390	0.057	0.246	0.018	0.195	0.066
Rental	0.019	0.018	0.026	0.026	0.733	0.027	0.755	0.100	0.352	0.032	0.321	0.093
Occupied	0.007	0.007	0.012	0.012	0.191	0.016	0.210	0.063	0.113	0.019	0.129	0.066
Vacant	0.021	0.021	0.028	0.028	0.756	0.038	0.749	0.099	0.383	0.035	0.328	0.101
Age 60+	0.016	0.016	0.031	0.031	0.631	0.026	0.619	0.078	0.302	0.029	0.275	0.080
Age 75+	0.026	0.026	0.036	0.036	0.733	0.037	0.724	0.099	0.358	0.036	0.323	0.096
MEAN	**0.021**	**0.021**	**0.038**	**0.038**	**0.499**	**0.029**	**0.510**	**0.088**	**0.263**	**0.034**	**0.242**	**0.090**
National-origin groups												
Mexican	0.042	0.042	0.057	0.058	0.717	0.714	0.755	0.748	0.375	0.368	0.341	0.333
Puerto Rican	0.132	0.131	0.162	0.162	0.777	0.743	0.815	0.762	0.448	0.424	0.403	0.378
Cuban	0.305	0.305	0.346	0.344	0.847	0.731	0.867	0.744	0.554	0.491	0.514	0.466
Chinese	0.223	0.300	0.289	0.351	0.834	0.795	0.847	0.787	0.531	0.539	0.470	0.498
Filipino	0.158	0.157	0.235	0.235	0.817	0.768	0.825	0.757	0.499	0.462	0.430	0.401
Indian	0.202	0.202	0.236	0.235	0.853	0.803	0.870	0.792	0.500	0.470	0.465	0.431
Japanese	0.243	0.242	0.327	0.326	0.831	0.741	0.845	0.727	0.547	0.492	0.485	0.438
Korean	0.224	0.223	0.336	0.335	0.844	0.768	0.853	0.748	0.552	0.503	0.483	0.445
Vietnamese	0.371	0.371	0.477	0.475	0.891	0.806	0.901	0.794	0.616	0.568	0.558	0.512
MEAN	**0.211**	**0.219**	**0.274**	**0.280**	**0.823**	**0.763**	**0.842**	**0.762**	**0.513**	**0.479**	**0.461**	**0.434**
Age for subgroups												
Age 15+, white	0.015	0.015	0.045	0.045	0.517	0.525	0.475	0.477	0.292	0.288	0.224	0.220
Age 15+, black	0.096	0.095	0.125	0.123	0.790	0.782	0.822	0.806	0.433	0.423	0.403	0.385
Age 15+, Hispanic	0.039	0.038	0.063	0.064	0.690	0.707	0.723	0.738	0.359	0.363	0.325	0.330
Age 15+, Asian	0.164	0.163	0.271	0.269	0.815	0.819	0.828	0.807	0.494	0.468	0.439	0.425
Age 15+, Native Americ	0.309	0.309	0.350	0.350	0.880	0.832	0.881	0.831	0.627	0.565	0.603	0.570
Age 60+, white	0.023	0.023	0.055	0.056	0.694	0.696	0.665	0.667	0.361	0.364	0.302	0.301
Age 60+, black	0.200	0.199	0.205	0.204	0.846	0.836	0.856	0.836	0.490	0.472	0.461	0.442
Age 60+, Hispanic	0.099	0.099	0.113	0.113	0.748	0.810	0.780	0.824	0.418	0.423	0.400	0.393
Age 60+, Asian	0.231	0.231	0.306	0.304	0.826	0.838	0.841	0.852	0.535	0.512	0.475	0.461
Age 60+, Native Americ	0.434	0.434	0.428	0.428	0.912	0.837	0.913	0.843	0.704	0.633	0.678	0.638
MEAN	**0.161**	**0.160**	**0.196**	**0.196**	**0.772**	**0.768**	**0.778**	**0.768**	**0.471**	**0.451**	**0.431**	**0.417**
Long Form Variables (n=54)												
MEAN	**0.155**	**0.275**	**0.186**	**0.325**	**0.583**	**0.547**	**0.601**	**0.645**	**0.376**	**0.478**	**0.353**	**0.470**