Improving TIGER, Hidden Costs: The Uncertain Correspondence of 1990 and 2010 U.S. Census Geography

Jonathan P. Schroeder

ABSTRACT: The U.S. Census Bureau supplies GIS-compatible definitions of census geographic units via its TIGER (Topologically Integrated Geographic Encoding and Referencing) data product series. Between the 2000 and 2010 censuses, the U.S. Census Bureau completed major improvements to their MAF/TIGER geographic database, from which all TIGER products are derived. The 2010 TIGER products, which supply boundaries for both 2000 and 2010 census units, are therefore significantly more accurate than 2000 TIGER data, which supply boundaries for both 1990 and 2000 census units. The accuracy improvements should be highly beneficial for spatial analyses of recent census data, but for *spatio-temporal* analyses that span the 1990–2010 period (or longer), the improvements impose a cost: in many cases, it is impossible to determine exactly which 1990 units correspond to which 2010 units. Boundaries that are in fact coincident may have *representations* that are not coincident in the separate TIGER versions, and the representational discrepancies are sometimes very large.

The National Historical Geographic Information System (NHGIS – https://nhgis.org) has recently begun releasing geographically standardized time series, which provide U.S. census data from multiple times for a single census's geographic units. To allocate one census's data to another census's geographic units, NHGIS interpolates data from the smallest source units for which the data are available. The first release, in 2015, supplied 2000 data for 2010 census units by interpolating data from 2000 census blocks. The next release will supply 1990 data for 2010 census units, again by interpolating block data, but in this setting, because of the improvements in TIGER data, the interpolation is complicated by the uncertain correspondence between 1990 blocks and 2010 census units.

Fortunately, TIGER data do make it possible to determine correspondences between 1990 and 2000 units, and between 2000 and 2010 units, and from these crosswalks, we can impose certain constraints on possible 1990-2010 unit relationships. Still, not all relationships can be exactly determined. In this paper, I posit three general alternatives for implementing areal interpolation in this setting: simply overlaying 1990 and 2010 boundaries without regard to representational discrepancies; using 2000 units as a bridge between 1990 and 2010 units; or a combined approach, overlaying 1990 and 2010 boundaries, but also using known topological relationships with 2000 units to constrain and refine the interpolation. In order to assess potential relative advantages of these approaches—without yet implementing them—I present here an assessment of how much uncertainty there is in block-based 1990 population estimates for 2010 units, identifying in particular how much uncertainty may be added by inexact correspondence information (i.e., the "hidden costs" of improved TIGER data).

KEYWORDS: areal interpolation, census geography, interoperability, spatio-temporal databases, spatially misaligned data

Introduction

The need for information on census unit correspondence

Spatio-temporal analysis of census summary data is often complicated by changes in the definitions of census geographic units. Where unit boundaries have changed, integrating data across time requires information about the relationships between units from one census and those from another. For example, if a census tract was split in two between two censuses, it is possible to measure population change within the original tract by computing the difference between its population in the earlier census and the sum of the populations of the two corresponding tracts in the later census, but this is possible only if we know how tracts from each census correspond to each other.

It is also possible to estimate—through *areal interpolation* (Goodchild and Lam, 1980)—how a split tract's population is distributed across the resulting smaller tracts and then estimate change within each smaller tract by computing differences between the known later counts and the interpolated earlier counts, but again, this is possible only with correspondence information. Furthermore, the most effective areal interpolation models require more than just basic correspondence information; they generally also require detailed spatial information about the intersections between units. At a minimum, most models restrict interpolated distributions to land areas, not water areas, and more sophisticated *dasymetric models* employ ancillary data about features related to population distributions, such as roads or land cover, to refine distributions within each census unit (Mennis, 2009). For such dasymetric models, it is not only important to have data that represent relationships among census units accurately, but also data that represent unit boundaries accurately relative to the features (e.g., land cover classes or road buffers) that define the dasymetric model.

TIGER data: Characteristics, improvements & limitations

The primary, official source of spatial information about U.S. census geographic units, for the 1980 decennial census through the present, is the U.S. Census Bureau's TIGER (Topologically Integrated Geographic Encoding and Referencing) data product series. TIGER data not only define the spatial extents of U.S. census reporting areas, but they also include representations of most road, railroad, and water features throughout the country, as well as several other types of features that are important for census area delineations and for census operations.

Crucially, for the purposes of determining relationships among census units across time, distinct TIGER data releases have also included definitions of reporting areas for *each consecutive pair* of decennial census years. The earliest version of TIGER data, the 1992 TIGER/Line Files, include both 1980 and 1990 census units; the 2000 TIGER/Line Files include both 1990 and 2000 units; and the 2010 TIGER/Line Shapefiles include both 2000 and 2010 units. Accordingly, TIGER data make it possible to determine exact topological relationships among census units from consecutive census years, and with the additional water and road data that TIGER provides, it is also possible to use TIGER data to refine interpolation models to limit population distributions to land areas near roads.

There are two major caveats, however. First, the principal objective of the initial versions of TIGER data was to describe *topology* accurately—"to show only the relative positions of elements"—which does not "require very high levels of positional accuracy" (U.S. Census Bureau, 2000, p. 5-6). Thus, the topology within early TIGER versions is generally reliable, so it should correctly indicate which units from one census year *intersect* which units from another, but the positional accuracy in early versions varies greatly and is occasionally quite poor, which complicates any attempt to use more accurate spatial information (e.g., high-resolution land cover data) to model distributions within census units.

Between the 2000 and 2010 censuses, the problem of poor positional accuracy was addressed by the MAF/TIGER Accuracy Improvement Project, through which the Census Bureau systematically realigned and updated TIGER features throughout the country (U.S. Census Bureau, 2012). As a result, the 2010 TIGER data are significantly more accurate than 1992 and 2000 TIGER data, which should be highly beneficial for spatial analyses of recent census data, but it has also accentuated another problem with TIGER data—the second major caveat: although TIGER data directly and effectively describe correspondences among census units across *consecutive pairs* of censuses, TIGER data do *not* convey exact correspondences among census units over longer time spans. Boundaries of 1990 and 2010 census units that are in reality coincident may have *representations* that are not coincident in the separate TIGER versions, and given the sweeping changes in TIGER features due to the Accuracy Improvement Project, the representational discrepancies are pervasive and occasionally very large.

Figure 1: Examples of boundary discrepancies in Broomfield County, Colorado. (A) The boundaries of 1990 blocks and block groups, as defined in 2000 TIGER data, and 2010 block group boundaries from the improved 2010 TIGER. (B) 2000 blocks as defined in 2000 and 2010 TIGER, illustrating some severe discrepancies but also offering a bridge between the two TIGER vintages.

Figure 1 provides examples of the discrepancies between 2000 and 2010 TIGER boundary representations. These are clearest in the boundaries of 2000 census blocks in Figure 1B. Each block polygon appears in both TIGER versions, and the topology among the blocks remains the same, but the size, shape, and positions of the blocks vary greatly. In a few cases, the 2000 and 2010 representations of a single 2000 block do not even intersect each other. Figure 1A demonstrates that it would be problematic to overlay 2000 and 2010 TIGER boundaries in order to determine correspondences between 1990 and 2010 census units. The boundaries of the northern two block groups appear to coincide with the same set of roads (topologically) in both 1990 and 2010, so that the two block groups are likely identical *in reality*, but the two TIGER versions suggest significant areas of overlap between these block groups and others between 1990 and 2010.

Figure 1 also illustrates an important means of determining some constraints on relationships between 1990 and 2010 census units. Because 2000 census units are defined in both the 2000 and 2010 TIGER data, it is possible to determine topological relationships between 1990 and 2000 census units *and* between 2000 and 2010 census units. In some cases, this information is enough to determine the exact topological relationship between a given 1990 unit and 2010 unit. E.g., if a 1990 block lies entirely within a 2000 block in 2000 TIGER, and that 2000 block lies entirely within a 2010 block group in 2010 TIGER, we can be sure (to the extent that TIGER topology is reliable) that the 1990 block also lies within the same 2010 block group.

Using 2000 geography as a "bridge" in this way may resolve a great deal of uncertainty about the correspondence between 1990 and 2010 census units. It remains to be determined, however, how effective this bridging can be. A primary aim of this paper is therefore to provide information about the actual scope of what can and cannot be determined about correspondences between 1990 blocks and 2010 census units from their known topological relationships with 2000 blocks.

NHGIS time series data

To simplify studies of trends in census summary data, and to address the challenges posed by changes in both the definitions *and* representations of census units, the National Historical Geographic Information System (NHGIS – https://nhgis.org) has recently begun releasing geographically standardized time series, which provide U.S. census data from multiple times for a single census's geographic units. In producing standardized data, NHGIS aims to achieve the highest practicable accuracy, ideally so that users may rely on NHGIS estimates without concern that gross errors could substantially affect their analyses. One key strategy NHGIS employs to achieve high accuracy is to interpolate data from the smallest source units for which the data are available.

Census blocks are the smallest units for which the Census has published its 100%-count short-form data tables. Accordingly, NHGIS's first release of geographically standardized time series, in 2015, supplied 2000 data for 2010 census units by interpolating data from 2000 census blocks. Being based on block data, this first release necessarily covers only short-form census subjects (e.g., race/ethnicity, sex, age, housing tenure, and household size), but it nevertheless supplies a wide range of counts,

amounting to 1,126 time series organized into 65 tables, with more tables coming soon. Each table provides counts of 2000 and 2010 characteristics for 10 levels of 2010 census geography: states, counties, census tracts, block groups, county subdivisions, places, congressional districts, core based (metropolitan and micropolitan) statistical areas (CBSAs), urban areas, and ZIP Code Tabulation Areas (ZCTAs).

NHGIS researchers are now working to expand the standardized time series to include data interpolated from 1990 blocks, but in this setting, because of the improvements in TIGER data, the interpolation is complicated by the uncertain correspondence between 1990 blocks and 2010 census units.

Research aims

The overriding goal motivating the present research is to identify an effective, practicable approach that NHGIS can use to interpolate census data from 1990 blocks to 2010 census units in the absence of exact correspondence information. In this paper, I do not provide a final, verified solution for that objective. Rather, I begin by positing a few general approaches that *could* be used in this setting, and I identify some potential advantages and limitations of each. Then my primary focus here is to assess the scope of uncertainty in relationships between 1990 blocks and 2010 census units, given the constraints imposed by their topological relationships with 2000 blocks, as determined from 2000 and 2010 TIGER data. Finally, I use this assessment of uncertainty to reconsider the possible interpolation approaches and recommend a way forward that addresses well the forms of uncertainty that have been identified.

The findings are useful not only for specifically assessing possible interpolation strategies, however, but also—and perhaps more importantly—for generally assessing the "hidden costs" of changes in TIGER representations. Also, although I focus here on a specific setting, the general discussion and assessment strategies I present here should be applicable to any other settings involving 2 zonal systems whose topological relationships are determinable only via relationships to a 3rd zonal system.

Methods

Areal interpolation with inexact correspondence information

I posit three general alternatives for implementing areal interpolation in the setting of interest:

1. Disregarding the representational discrepancies between TIGER versions, simply overlay 1990 block and 2010 census unit polygons and allocate 1990 block counts to any "intersecting" 2010 units.

This approach would support most areal interpolation models, including areal weighting, target-density weighting (Schroeder, 2007), or, by overlaying additional ancillary information, dasymetric mapping (Mennis, 2009). The obvious problem would be potentially frequent and large misallocations from 1990 blocks to 2010 units that do not,

in reality, intersect, and in some cases, where discrepancies are especially large, a 1990 block's data could be allocated *wholly* to an incorrect 2010 unit.

2. Use 2000 blocks as a bridge: first interpolate from 1990 blocks to 2000 blocks using 2000 TIGER information; then interpolate from 2000 blocks to 2010 units using 2010 TIGER information.

This approach could reduce significantly—though not eliminate—the risks of misallocation posed by direct overlay. It could also support most areal interpolation models, though a dasymetric mapping approach could perform poorly for some cases of 1990-to-2000 allocation where 2000 TIGER representations have poor positional accuracy and the ancillary information is more accurate. (One solution in that case would be to use only 2000 TIGER roads as the ancillary information, in which case the ancillary information would at least have correct topological relationships with the block boundaries.)

This approach also corresponds to the one used to produce 1970-1990 estimates for 2010 census tracts in the Longitudinal Tract Database (LTDB), another source of geographically standardized census data (Logan *et al*., 2014). The main difference is that for the LTDB, the operational units in the first step (interpolating 1990 and earlier data to 2000 units) are tracts, not blocks.

3. A combined approach: first, overlay 1990 and 2010 boundaries directly; then, eliminate any intersections that are known to be invalid according to relationships with 2000 units, and *add* intersections between units that must intersect according to relationships with 2000 units.

This approach is somewhat more complex and demanding to implement than the prior approaches, so a key question—the focus of the research results I present here—is how useful could it be? It is not possible to answer this question directly, so I identify, report, and discuss some other characteristics of 1990-2010 relationships that are indicative.

Measuring interpolation uncertainty

As a starting point, we may determine the overall uncertainty in block-based estimates of 1990 populations in 2010 census units based on the *known* topological relationships with 2000 blocks given by 2000 and 2010 TIGER data.

In many cases, knowing topological relationships alone can be enough to determine exactly how to allocate block data, with no uncertainty. Census blocks are generally much smaller than other census units, and, for each census since 1990, blocks nest exactly within all larger reporting areas (U.S. Census Bureau, 1994, 2012). We may therefore expect that *most* blocks of one census year will also nest within larger units of another year, and if a block nests wholly within another unit, then we can be certain that its census counts should be allocated wholly to the encompassing unit. Uncertainty arises then only where a source block's area intersects multiple target units. Moreover, given that census-measured features are generally restricted to land areas, we may further

assume that uncertainty will arise only where a source block's *land* area intersects multiple target units.[1]

Following this logic, NHGIS's first release of standardized time series includes lower and upper bounds for each 2000 estimate, indicating the "topologically possible" range for each count. If all of the 2000 blocks that intersect a given 2010 unit lie entirely within that unit, then the lower and upper bounds are equal, and the range is zero, indicating no uncertainty. If instead most of the blocks intersecting a given 2010 unit also intersect another 2010 unit (over land), then it is *possible* that the population of these "split" blocks lies wholly outside the given 2010 unit *or* within the given unit, which results in a large range between the lower and upper bounds, indicating greater uncertainty.

Similar bounds may be computed for 1990 characteristics of 2010 units, but the requirements for determining that a 1990 block lies wholly within a 2010 unit are twofold: first, the 1990 block's land area must lie wholly within a single 2000 block, and second, that 2000 block's land area must lie wholly within a single 2010 unit. If *either* of these conditions are not met, then it is possible that the 1990 block shares land area with multiple 2010 units, and the allocation is uncertain. Applying this logic, using TIGER-based block boundaries and census population counts from NHGIS (Minnesota Population Center, 2011), Table 1 reports the general scope of uncertainty in both 1990 and 2000 block-based population estimates for all 2010 U.S. census units at each of the 10 geographic levels covered by NHGIS's standardized time series data.

Measuring potential reductions in uncertainty

One means of assessing the "cost" of inexact correspondence information is to determine how much less uncertainty there *could* be if we had exact correspondence information. To explain how this is possible, I first proceed through some definitions.

Let us say a relationship between two census units is topologically *indeterminable* if, given available information, it is possible either that the two units share land area or that they share no land area. Likewise, a relationship is topologically *determined* if, given available information, it is necessary that two units share land area.

I posit that, in the setting of interest, a relationship between a 1990 block A and a 2010 unit C is *indeterminable* if and only if *each* 2000 block that shares land with both A and C also shares land with some other 1990 block(s) *and* some other 2010 unit(s).

As proof, first consider the case where there is only a *single* 2000 block B that shares land with A and C. If A is the only 1990 block that shares land with B, then A must also share land with all 2010 units that share land with B, including C, so the A-C relationship is determined. If C is the only 2010 unit that shares land with B, then C must share land with all 1990 blocks that share land with B, including A, so again, the A-C relationship is determined. If, however, B shares land with multiple 1990 blocks *and* multiple 2010

[1] Of the nearly 5 million 1990 blocks with nonzero population or housing unit counts, there are 28 that are entirely over water according to NHGIS's 1990 block shapefile, which is derived from 2000 TIGER data (Minnesota Population Center, 2011). These cases received special handling to produce the results here, generally by allowing for the possibility, in *only* these blocks, that population could be located over water.

units, then it is possible that only a subset of the associated 1990 blocks share land with unit C—possibly including A or not—and likewise, it is possible that only a subset of the associated 2010 units share land with A, so the A-C relationship is indeterminable.

Next, consider the case where there are *multiple* 2000 blocks that share land area with A and C. If *any* of these 2000 blocks share land area with only one 1990 block (A) or one 2010 unit (C), then the A-C relationship is determined. Therefore, for the relationship to be indeterminable, *all* 2000 blocks sharing land area with A and C must share land area with multiple 1990 blocks and multiple 2010 units.

Having established which relationships are indeterminable and which are determined, it is then possible to distinguish three levels of uncertainty in relationships. First, a relationship between a 1990 block A and 2010 unit C is *resolved* if we know either that A's land is entirely within C or that A shares no land with C. In either case, there is no uncertainty about how much of A's count to allocate to C; it is either all or none. Second, a relationship is *unresolvable* if we know it will entail uncertainty even if we had complete correspondence information. This occurs if we know, given available information, that a 1990 block A must share land area with multiple 2010 units, in which case there must be uncertainty in how A's count is distributed among the 2010 units. Third, a relationship is *possibly resolvable* if it might or might not be resolved with complete correspondence information. This occurs if a 1990 block may share land area with a single 2010 unit *or* with multiple 2010 units, and complete correspondence information would tell us which.

I posit that a relationship is *possibly resolvable* if and only if it is indeterminable *or* (it is determined *and* the 1990 block has indeterminable relationships with other 2010 units but no other determined relationships).

As proof, first consider the alternatives. If a 1990 block has two or more determined relationships with 2010 units, then the relationship is unresolvable; there must be uncertainty in how the block's count should be allocated to 2010 units. If a 1990 block has exactly one relationship with a 2010 unit, and it is determined, then the relationship is resolved; we know that the whole block count should be allocated to the one 2010 unit. Then consider the case where a 1990 block has only indeterminable relationships. By definition, the 1990 block may or may not share land in each case, which means it is *possible* that the 1990 block shares land with only one 2010 unit, so the allocation would be certain, but it is also possible that the 1990 block shares land with multiple 2010 units, so the allocation would be uncertain. Either way, complete correspondence information would allow us to determine if there is uncertainty or not. Finally, consider a case where a 1990 block has one determined relationship and one or more indeterminable relationships. In this case, it is again possible that that the 1990 block shares land area with only one 2010 unit (the one with which its relationship is already determined) or with multiple 2010 units.

Applying this logic, for any given 2010 unit that has any "possibly resolvable" relationships with 1990 blocks, it is *possible* that if we knew the exact topological relationships between 1990 blocks and 2010 units, then the uncertainty in the 2010 unit's

block-based 1990 population estimate (measured as the range between the upper and lower bounds) *could* be reduced by an amount equal to—and not exceeding—the population of all 1990 blocks that have possibly resolvable relationships with the unit (*if* the land area of each of those 1990 blocks lay entirely within or outside of the 2010 unit). Table 2 summarizes the scope of such potential reductions across all 2010 census units at 10 geographic levels.

Identifying dubious allocations

Rather than identify only *possible* cases of topologically invalid misallocations, another approach to assessing the costs (or risks) of inexact correspondence information is to identify especially "dubious allocations"—cases where the known relationships make it appear likely that substantial misallocation may occur. For this, I define dubious allocations to be those meeting all of these conditions:

1. According to 2000 TIGER data, more than 50% of the 1990 block's land area lies within one 2000 block (which makes it likely that most of the 1990 block's population lies within the 2000 block).

2. According to 2000 TIGER data, the 1990 block's intersection with the 2000 block includes less than 50% of the 2000 block's land area (which reduces the likelihood that the 1990 block shares land with all of the 2010 census units that intersect the 2000 block).

3. According to the model NHGIS uses to allocate 2000 block counts to 2010 census units (Schroeder, 2016), less than 80% of the 2000 block's population is assigned to a single 2010 census unit (which increases the likelihood that a substantial portion of the population in a 1990 block associated with the 2000 block will be allocated among 2010 units that the 1990 block does not intersect).

The thresholds used here are arbitrary, and, importantly, there may be many cases of substantial misallocation that do not meet all three conditions, but these conditions should nevertheless adequately capture the most dubious cases.

To summarize the scope of these "dubious allocations" in Table 3, I first estimate the 1990 population of each 2010 census unit using a "bridging" approach that interpolates from 1990 blocks to 2000 blocks through simple target-density weighting (Schroeder, 2007; *forthcoming*) and then interpolates from the 2000 blocks to 2010 units using NHGIS's 2000-block-to-2010-unit model, which is a hybrid of target-density weighting and a binary dasymetric model that uses road and imperviousness data (Schroeder, 2016; see also https://nhgis.org/documentation/time-series/2000-blocks-to-2010-geog). I then determine what portion of each estimate is derived from a dubious allocation according to the conditions given above.

Results

General uncertainty

Among 2010 census units, there is pervasive uncertainty in both 1990 and 2000 block-based population estimates (Table 1). At all levels, there are many cases where it is impossible to determine an exact 1990 or 2000 population count by direct allocation of block counts. Even for block groups, the level with the lowest rates of uncertain estimates, about a quarter of 2000 estimates and 40% of 1990 estimates are uncertain. At the extreme, ZCTAs, urban areas, and congressional districts (which each have complex boundaries that often change substantially between censuses) all have very high rates of uncertain estimates for both years, ranging from about 90% to nearly 100%.

Table 1: Frequencies of uncertainty in block-based 2000 and 1990 population estimates for 2010 U.S. census units.

		2000 population is uncertain		1990 population is uncertain		2000 pop. range > 50% of max		1990 pop. range > 50% of max	
Geographic level	**N**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Block groups	217,740	55,792	25.6	85,959	39.5	9,955	4.6	24,180	11.1
Tracts	73,057	24,375	33.4	35,331	48.4	1,171	1.6	4,078	5.6
County subdivisions	35,703	19,136	53.6	22,177	62.1	789	2.2	1,404	3.9
ZCTAs	32,989	29,945	90.8	31,610	95.8	2,134	6.5	4,154	12.6
Places	29,261	19,620	67.1	23,100	78.9	3,341	11.4	5,618	19.2
Urban areas	3,573	3,554	99.5	3,570	99.9	78	2.2	200	5.6
Counties	3,143	1,802	57.3	2,005	63.8	0	0.0	0	0.0
CBSAs	942	621	65.9	670	71.1	0	0.0	0	0.0
Cong. districts	439	389	88.6	429	98.4	0	0.0	0	0.0
States	51	37	72.5	41	80.4	0	0.0	0	0.0

Notes: ZCTAs = ZIP Code Tabulation Areas, CBSAs = core based statistical areas, Cong. districts = 111th Congressional Districts

In most cases of uncertainty, the magnitude of uncertainty is not very large. The frequency of population ranges exceeding 50% of the maximum possible population is much lower than the frequency of cases with *any* uncertainty, and for four levels, there are no cases of such extreme uncertainty. Still, among the other six levels, there are numerous instances, and most notably, the rates of extreme uncertainty for 1990 estimates are generally two or more times the rates for 2000 estimates. It is clear that, first, there are an ample number of cases with a lot of "give" in block-based estimates for both 2000 and 1990, indicating the importance of implementing an effective interpolation model, and second, 1990 population estimates entail considerably more uncertainty than 2000 estimates, which must be due in some part to the lack of exact correspondence information between 1990 blocks and 2010 units.

Potential reductions in uncertainty

The typical magnitudes of uncertainty in block-based 1990 population estimates for 2010 units vary greatly among geographic levels, with means ranging from only 225 for uncertain state estimates to nearly 6,708 for congressional districts (Table 2). In relative terms, block groups have the most severe uncertainty, with an average of 34% of their

potential populations being uncertain (among block groups having an uncertain estimate). In contrast, uncertainty in counties, CBSAs, congressional districts, and states tends to be a small portion of the maximum possible population—1.2% or less on average.

Table 2: Magnitudes of uncertainty in block-based 1990 population estimates for 2010 census units and potential reductions in uncertainty if all 1990-2010 topological relationships were known. Summary limited to units with uncertain 1990 populations.

Geographic level	N with uncertain 1990 pop.	Mean 1990 pop. range	Mean (1990 pop. range % of max pop.)	Mean potential pop. range reduction	Mean potential % reduction in pop. range	Potential pop. range reduction > 50% max pop.	
						N	%
Block groups	85,959	595	34.1	133	22.9	4,119	1.9
Tracts	35,331	630	18.5	155	24.5	594	0.8
County subdivisions	22,177	534	14.6	182	32.6	262	0.7
ZCTAs	31,610	1,036	23.1	346	34.0	996	3.0
Places	23,100	998	31.6	313	30.7	1,187	4.1
Urban areas	3,570	3,290	19.4	1,311	41.5	32	0.9
Counties	2,005	414	0.9	167	39.2	0	0.0
CBSAs	670	384	0.3	179	42.0	0	0.0
Cong. districts	429	6,708	1.2	1,341	20.8	0	0.0
States	41	225	0.0	142	40.5	0	0.0

Notes: ZCTAs = ZIP Code Tabulation Areas, CBSAs = core based statistical areas, Cong. districts = 111th Congressional Districts

The amounts by which uncertainty could potentially be reduced, if all 1990-2010 topological relationships were known, are much more consistent across levels. The mean potential percent reductions all fall between 20 and 42%. Those percentages, of course, describe potential reductions relative to possible population ranges, which may be quite small, so even a high percent potential reduction in this case could indicate a very small potential effect on an actual estimate. The last two columns of Table 2 are perhaps more pertinent, indicating the frequency of very large potential reductions in uncertainty relative to the maximum possible populations of units. Here, as in Table 1, it appears that the problem is negligible for large units (counties and larger), but for smaller units, there are a substantial number of cases where estimates have a high degree of uncertainty that *may* mainly be due to the lack of exact correspondence information.

Dubious allocations

The frequency of estimates that rely on dubious allocations (meeting the three criteria identified above) is, at first glance, reassuringly low (Table 3). There appears to be no problem at all among the largest units, and even among the smaller units, the rates are very low. For the levels with the highest rates, ZCTAs and places, only about 1 in 200 estimates are more than 25% "dubious." Nevertheless, given that the ideal outcome is to produce estimates lacking any gross errors, it is concerning that there could be 56 census tracts and 505 block groups where more than a quarter of the estimated 1990 characteristics are allocated from blocks that do not even intersect the unit of interest.

Table 3: Frequency among 2010 census units of block-based 1990 population estimates that rely on dubious allocations. Summary limited to units with uncertain 1990 populations.

Geographic level	N with uncertain 1990 pop.	Est. 1990 pop. > 5% dubious		Est. 1990 pop. > 25% dubious	
		N	%	N	%
Block groups	85,959	1,642	0.8	505	0.2
Tracts	35,331	165	0.2	56	0.1
County subdivisions	22,177	45	0.1	5	0.0
ZCTAs	31,610	1,183	3.6	150	0.5
Places	23,100	841	2.9	138	0.5
Urban areas	3,570	39	1.1	2	0.1
Counties	2,005	0	0.0	0	0.0
CBSAs	670	0	0.0	0	0.0
Cong. districts	429	0	0.0	0	0.0
States	41	0	0.0	0	0.0

Notes: See text for definition of "dubious allocations." ZCTAs = ZIP Code Tabulation Areas, CBSAs = core based statistical areas, Cong. districts = 111th Congressional Districts

Discussion & Conclusions

The results indicate that the second proposed approach to interpolating from 1990 blocks to 2010 units—using 2000 blocks as a bridge from the source units to the targets—may be a reasonable strategy with adequate accuracy for many applications. The rates of extreme uncertainty are fairly low for most target levels (Table 1); the potential for exact correspondence information to reduce uncertainty greatly is also small in most cases (Table 2); and there are, in relative terms, very few cases where 1990 estimates would be heavily reliant on topologically dubious allocations (Table 3). But for the NHGIS project, with a goal to produce estimates that may be used by a broad range of users who typically will not have any familiarity with the potential errors in areal interpolation, it seems that even "low rates" of extreme uncertainty are best avoided. In other words, the "glass half empty" view of the results presented here is also legitimate: the lack of exact correspondence information for 1990 and 2010 census units is, in fact, a costly source of uncertainty in a substantial number of cases throughout the country.

I therefore conclude that pursuing the third proposed approach—using an overlay of 1990 and 2010 boundary data from the different TIGER vintages and constraining the results to respect the topological relationships that can be determined via relationships with 2000 blocks—is a worthwhile endeavor. Given that there are some areas, most of all in Alaska and Hawaii, where 2000 TIGER features have large systematic positional inaccuracy, it will also be useful to investigate whether there might be a relatively simple strategy to align TIGER features better across vintages. Realigning every individual 1990 block boundary is far beyond scope, but a much simpler approach of "rubber sheeting" some 2000 TIGER data to align better with 2010 TIGER features could by itself address the most severe systematic issues. Then an overlay of adjusted 1990 block boundaries with 2010 boundary data could aid greatly in preventing "dubious allocations" by providing more detailed information about where exactly within each "bridging" 2000 block the intersections with 1990 blocks and 2010 units lie.

Acknowledgements

This work was completed for the NHGIS project with funding from the National Science Foundation [SES-1324875] and Eunice Kennedy Shriver National Institute of Child Health and Human Development [NICHD 2R01HD057929]. The author also gratefully acknowledges support from the Minnesota Population Center [NICHD R24HD041023].

References

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

Logan, J. R., Xu, Z., & Stults, B. J. (2014) Interpolating US decennial census tract data from as early as 1970 to 2010: A longitudinal tract database. *The Professional Geographer*, 66, 3, pp. 412-420.

Mennis, J. (2009) Dasymetric mapping for estimating population in small areas. Geography Compass, 3, 2, pp. 727-745.

Minnesota Population Center. (2011). National Historical Geographic Information System: Version 2.0.

Schroeder, J. P. (2007) Target-density weighting interpolation and uncertainty evaluation for temporal analysis of census data. *Geographical Analysis*, 39, 3, pp. 311-335.

Schroeder, J. P. (2016) Hybrid areal interpolation of census counts from 2000 blocks to 2010 geographies. Manuscript submitted for publication.

U.S. Census Bureau (1994) *Geographic Areas Reference Manual* [PDF]. Retrieved April 21, 2016, from http://www.census.gov/geo/reference/garm.html.

U.S. Census Bureau (2000) *Census 2000 TIGER/Line Files Technical Documentation* [PDF]. Retrieved July 14, 2016, from https://www2.census.gov/geo/pdfs/maps-data/data/tiger/rd_2ktiger/tgrrd2k.pdf.

U.S. Census Bureau (2012) *2010 TIGER/Line Shapefiles Technical Documentation* [PDF]. Retrieved July 14, 2016, from https://www2.census.gov/geo/pdfs/maps-data/data/tiger/tgrshp2010/TGRSHP10SF1.pdf.

Jonathan P. Schroeder, Research Scientist, Minnesota Population Center, University of Minnesota, Minneapolis, MN 55455, jps@umn.edu