Accepted Manuscript

Title: Hybrid Areal Interpolation of Census Counts from 2000 Blocks to 2010 Geographies

Author: Jonathan P. Schroeder

Affiliation: Minnesota Population Center
University of Minnesota
Minneapolis, Minnesota

Correspondence: Jonathan P. Schroeder
Minnesota Population Center
50 Willey Hall
225 19th Avenue South
Minneapolis, MN 55455
USA
Email: jps@umn.edu

Citation: Schroeder, J. P. (2017). Hybrid areal interpolation of census counts from 2000 blocks to 2010 geographies. Computers, Environment and Urban Systems, 62, 53-63.
http://dx.doi.org/10.1016/j.compenvurbsys.2016.10.001

Highlights

- Key product: block-based estimates of numerous 2000 population and housing characteristics for 2010 census units.
- Data cover entire U.S. at 10 geographic levels, including block groups, tracts, places, and ZIP Code Tabulation Areas.
- Even using a block basis, uncertainty in 2000 counts for 2010 units is pervasive and occasionally large.
- 18 interpolation models—employing water, roads, imperviousness, and/or 2010 block data—are defined and assessed.
- The final model, a hybrid of dasymetric and density weighting approaches, improves substantially on other tested models.

Hybrid Areal Interpolation of Census Counts from 2000 Blocks to 2010 Geographies

Abstract

To measure population changes in areas where census unit boundaries do not align across time, a common approach is to interpolate data from one census's units to another's. This article presents a broad assessment of areal interpolation models for estimating counts of 2000 characteristics in 2010 census units throughout the United States. We interpolate from 2000 census block data using 4 types of ancillary data to guide interpolation: 2010 block densities, imperviousness data, road buffers, and water body polygons. We test 8 binary dasymetric (BD) models and 8 target-density weighting (TDW) models, each using a unique combination of the 4 ancillary data types, and derive 2 hybrid models that blend the best-performing BD and TDW models. The most accurate model is a hybrid that generally gives high weight to TDW (allocating 2000 data in proportion to 2010 densities) but gives increasing weight to a BD model (allocating data uniformly within developed land near roads) in proportion to the estimated 2000-2010 rate of change within each block. Although for most 2010 census units, this hybrid model's estimates differ little from the simplest model's estimates, there are still many areas where the estimates differ considerably. Estimates from the final model, along with lower and upper bounds for each estimate, are publicly available for over 1,000 population and housing characteristics at 10 geographic levels via the National Historical Geographic Information System (NHGIS – http://nhgis.org).

Keywords: Areal interpolation; Census geography; Spatio-temporal analysis; Population estimation

1. Introduction

Summary data from national censuses are a vital resource for studies of local and regional trends in population or housing characteristics, but measuring changes in summary data is frequently complicated by boundary changes in census geographic units (Martin, Dorling, & Mitchell, 2002;

Gregory, 2002; Schroeder, 2007). When, for example, a city annexes land or a census tract boundary is revised, census agencies typically publish new data only for the updated units. In such areas, determining exactly how characteristics changed *within a fixed extent* is not generally feasible.

One may, however, *estimate* the characteristics of fixed extents by applying some form of areal interpolation. "Areal interpolation" takes aggregate data describing a feature's distribution over a set of *source zones* and transforms the data to produce estimates of how the same feature is distributed over a set of *target zones* (Goodchild & Lam, 1980). To study changes in census summary data, we may interpolate data from one census's geographic units (the source zones) to another census's (the target zones), thereby enabling estimations of change within the target zones.

In ongoing work for the National Historical Geographic Information System (NHGIS), we are applying areal interpolation to U.S. census data to produce geographically standardized time series for a range of population and housing characteristics at several geographic levels for years ranging back at least to 1980. This plan entails several interpolation settings, each with unique data constraints, and we aim to refine our areal interpolation model for each setting to achieve the highest practicable accuracy. As we establish a model for each setting, we will continually extend NHGIS's standardized time series to cover more years and more characteristics.

This article provides an explanation and assessment of the interpolation model used to produce NHGIS's first release of standardized time series (now available at http://nhgis.org). The release consists of 1,126 time series organized into 65 tables,[1] each providing counts of 2000 and 2010 characteristics for 2010 census units at 10 geographic levels: 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). By design, the time series cover only characteristics that were tabulated for 2000 blocks, which limits the scope of this release to one areal interpolation setting: allocating count data from 2000 blocks to larger 2010 units.

The NHGIS model for this setting uses four types of ancillary data to guide interpolation: census counts from the target zone year (2010), imperviousness data, road buffers, and water body polygons. Each of these data types have been used in previously reported models, sometimes in combination with one or two others, but the NHGIS model is novel in its use of all four types and in the way it combines the types in a hybrid model. The findings presented here are therefore relevant for other settings as well—particularly where the source and target zones are census units of different vintages, where the source zones are relatively small (as are blocks), or where multiple types of ancillary data are available for model refinement.

2. Background

2.1. Areal interpolation from census blocks

Among U.S. census units, blocks are the smallest and most numerous unit by a large margin. In 2000 census data, there are 8.2 million blocks (excluding Puerto Rico and other territories), 39 times more than the next most numerous unit, block groups. Additionally, for each census since 1990, blocks nest exactly within all larger reporting areas (U.S. Census Bureau, 1994, 2012). These conditions *suggest* that most 2010 census units, at all levels higher than blocks, should intersect a large number of 2000 blocks, and most 2000 blocks should lie entirely within a single 2010 unit (at all higher levels). If so, any interpolation model that respects basic zone relationships will allocate most block counts wholly to the encompassing 2010 units with exact accuracy, resulting in fewer and smaller errors overall than in settings with larger source zones.

It is therefore understandable that prior applications of block-based areal interpolation have employed relatively simple models without investigating more sophisticated approaches. For example, the Longitudinal Tract Database (LTDB), which supplies 1970-2010 census data for 2010 tracts (Logan, Xu, & Stults, 2014), estimates 2000 population totals by interpolating from blocks using *areal weighting*

(AW), which allocates source zone counts in proportion to the area of intersection with each target zone (Goodchild & Lam, 1980). Other research has applied AW to block data as a benchmark against which to assess the interpolation of tract data (Buttenfield, Ruther, & Leyk, 2015; Ruther, Leyk, & Buttenfield, 2015; Zoraghein et al., 2016). AW's basic assumption—that characteristics are uniformly distributed within each source zone—may often be inaccurate, and numerous studies have shown that, in settings with larger source zones, more sophisticated models are more effective (e.g., Goodchild, Anselin, & Deichmann, 1993; Fisher & Langford, 1995; Mrozinski & Cromley, 1999; Gregory, 2002; Reibel & Bufalino, 2005; Langford, 2006; Reibel & Agrawal, 2007; Schroeder, 2007; Zandbergen & Ignizio, 2010; etc.). It is nevertheless possible that *block-based* AW may be acceptably accurate for many applications, and if so, AW's simplicity makes it a sensible choice.

Another provider of geographically standardized census data, Geolytics, also interpolates from block data but weights by road lengths rather than by areas (Tatian, 2003).[2] Schroeder (2007) uses a similar street-side weighting model to interpolate block data for an assessment of tract-based interpolation, and the Census Bureau also uses street-side weighting for block-based population estimates in the 1990–2000 tract relationship files (U.S. Census Bureau, 2000). It is intuitive that population distributions should generally concentrate along road networks, and road-based interpolation has been more effective than AW in several test settings (Xie, 1995; Mrozinski & Cromley, 1999; Reibel & Bufalino, 2005; Zandbergen & Ignizio, 2010), but no prior research has specifically assessed road-based interpolation of *block* data.

One prior assessment does come close, however. The researchers who produced the LTDB recently assessed several versions of 2000 population estimates for 2010 tracts (Logan, Stults, & Xu, 2016), including NHGIS's new estimates, which, as detailed below, make use of road data along with other ancillary data types. Their assessment strategy is to compare estimates with a set of retabulated 2000 populations for 2010 tracts that the Census Bureau produced for a special report (Wilson et al.,

2012). They find that for most tracts, NHGIS's estimates and LTDB's block-based AW estimates are very similar, and the errors are mostly small, especially relative to the much greater errors produced by tract-based interpolation models. Still, in the 864 tracts where the LTDB and NHGIS population estimates differ by 100 or more, they find that NHGIS is closer to the benchmark populations in 86% of cases.

The research presented here expands on Logan, Stults, & Xu's findings by assessing 18 distinct block-based interpolation models, including LTDB's and NHGIS's, making it possible to distinguish the relative advantages of several ancillary data types and model innovations. The assessment strategy used here (disaggregating from block pairs, as detailed in Section 3.5) does not enable measuring errors for actual target census units, as does the prior assessment, but it does avoid a problem of using the Census's retabulated tract data, which, as noted by Logan, Stults, & Xu, include many post-census corrections that artificially inflate measured interpolation errors, suggesting poor model performance in cases where the actual problem is erroneous counts in the Census's 2000 block data. The prior assessment is also restricted to tract-level estimates, whereas the present work provides summaries for 10 geographic levels.

2.2. Uncertainty in block-based estimates

Before detailing the interpolation models examined in this research, it makes sense first to consider how much effect model selection could have on block-based estimates. Logan, Stults, & Xu's (2016) findings demonstrate that, for some census tracts, it is possible for two block-based interpolation models to produce significantly different results. More generally, we can determine the full extent to which any two block-based models could *possibly* differ for any target geographic level by measuring the total uncertainty inherent in block-based estimates.

In practice, most areal interpolation models assume that census-measured features are located only in *land* areas (not water), but beyond that, we cannot know the exact locations of features within a zone when precise address-level information is not publicly available. Accordingly, in the setting of

interest, the minimum possible 2000 count for a 2010 unit is the sum of counts for 2000 blocks that share *all* of their land area with the 2010 unit, and the maximum is the sum of counts for 2000 blocks that share *any* land area with the 2010 unit. Wherever the minimum and maximum are not equal, there is some uncertainty in the 2000 count, and the range between the two limits indicates the magnitude of uncertainty.

Following this rubric, Table 1 summarizes the uncertainty in 2000 population counts for all 2010 census units in the U.S. at each of the 10 geographic levels covered in NHGIS's first release of standardized time series. To compute these numbers, land areas for intersections between blocks are drawn from the 2000-2010 block relationship files (U.S. Census Bureau, 2010), and 2000 block populations and 2010 geographic codes are drawn from the 2000 and 2010 Summary Files via NHGIS (Minnesota Population Center, 2011).

Table 1. Frequency and magnitude of uncertainty in block-based 2000 population estimates for 2010 U.S. census units.

		2000 population is uncertain		2000 pop. range > 10% of max		2000 pop. range > 50% of max		Mean 2000 pop.	Mean (2000 pop. range % of
Geographic level	**N**	**N**	**%**	**N**	**%**	**N**	**%**	**range**	**max)**
Block groups	217,740	55,792	25.6	33,361	15.3	9,955	4.6	122	6.6
Tracts	73,057	24,375	33.4	8,542	11.7	1,171	1.6	153	4.1
County subdivisions	35,703	19,136	53.6	6,718	18.8	789	2.2	228	6.4
ZCTAs	32,989	29,945	90.8	13,448	40.8	2,134	6.5	721	14.4
Places	29,261	19,620	67.1	11,621	39.7	3,341	11.4	546	16.6
Urban areas	3,573	3,554	99.5	1,717	48.1	78	2.2	2,492	13.2
Counties	3,143	1,802	57.3	5	0.2	0	0.0	185	0.3
CBSAs	942	621	65.9	0	0.0	0	0.0	198	0.2
Cong. districts	436	389	89.2	0	0.0	0	0.0	2,890	0.4
States	51	37	72.5	0	0.0	0	0.0	139	0.0

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

The table reveals that uncertainty is pervasive. At all levels, there are large numbers of 2010 units where the 2000 population cannot be exactly determined from 2000 blocks. At the extreme, 99.5% of 2010 urban areas and 90.8% of 2010 ZCTAs have boundaries that cut through the land area of a populated 2000 block. Even for the level with the lowest uncertainty rate, block groups, uncertainty

exists in about 1 in 4 cases. Most surprisingly, there are 37 *states* where the 2000 population cannot be exactly determined for the 2010 extent, according to the Census Bureau's definitions of state boundaries.

One may wonder whether all of these misalignments are legitimate. Did the boundaries of 37 states *really* change? In fact, state boundaries are sometimes officially adjusted (e.g., Greenhouse, 1998; Beam, 2012) as are boundaries for other administrative units, but misalignments may also reflect corrections the Census made to their own boundary definitions. All evidence suggests, however, that all misalignments indicate *de facto* changes in census tabulation units even if they do not correspond to "real" *de jure* changes in administrative units. The 2000-2010 block relationship files correspond exactly to the relationships in the Census's 2010 TIGER/Line Shapefiles, and the Census provides no more detailed information about relationships between its 2000 and 2010 boundaries. Therefore, if the 2010 TIGER/Line data indicate that a parcel of land was included in different states in 2000 and 2010, then we can only assume that 2000 and 2010 census tabulations honor that change *regardless* of whether the state boundary officially changed.

Even then, one may also wonder how many misalignments are really *significant*. The method used here to specify absolute limits assumes that any nonzero portion of a block's land area could possibly contain *all* of the block's population and housing though that may be practically impossible in cases of very small misalignments. For example, of the 24,375 2010 tracts identified to have uncertain 2000 populations, 6,009 (24.7%) involve only misalignments that encompass less than 1% of any 2000 block's land area *and* less than 1 hectare (2.47 acres) of land. It seems very unlikely that the entire population of a block would reside in such a small part, but importantly, it is *possible.* Even a minuscule misalignment may only appear to have a small area due to imprecise TIGER/Line definitions, or the Census may have located addresses there for housing that lies outside of the area. It therefore seems appropriate to consider *any* misalignment over land to be a source of uncertainty.

Still, for many applications, frequently occurring uncertainty is unimportant if the uncertainties are consistently small. The remaining columns in Table 1 demonstrate, however, that uncertainty is often quite large. For many thousands of units, including 5 counties, the difference between the minimum and maximum possible 2000 population exceeds 10% of the maximum, and for six levels—block groups, tracts, county subdivisions, ZCTAs, places, and urban areas—there are also many units where the possible population range exceeds 50% of the maximum. For these same levels, the rightmost two columns show that the *average* uncertainty is also large. In the worst case, about 1/6th of each 2010 place's potential 2000 population is uncertain, on average. The average proportional uncertainty is similarly high for ZCTAs and urban areas, and even among tracts, the 2000 populations are *on average* about 4% uncertain.

These findings demonstrate the importance of assessing alternative interpolation models for the setting of interest. In general—and especially for units smaller than counties—2000 blocks do not nest well enough within 2010 units to ensure accurate allocations of block counts, so the choice of model—determining how counts are to be allocated within split blocks—*could* significantly affect the accuracy of block-based estimates.

These findings also indicate that researchers using block-based estimates should be aware of the uncertainties. To that end, all of NHGIS's standardized time series come with lower and upper bounds for interpolated estimates, following the specification for absolute limits used here, derived separately for each interpolated population or housing characteristic.

2.3. Example case

Fig. 1 presents a somewhat extreme case of uncertainty where, according to NHGIS boundary files (based on the U.S. Census's 2010 TIGER/Line Shapefiles), a single 2000 block (ID: 484910205022002) shares land area with three 2010 places (Austin, Brushy Creek, and—in a small sliver—Cedar Park) and with two 2010 tracts (205.09 and 205.10). The uncertainty here is large because the block's 2000

population, 1,624, is large. In comparison, among all populated 2000 blocks that share land area with multiple units in one of the 10 target levels, the mean population is 102, but 1,391 of these blocks have populations larger than 1,624, so the example is not altogether exceptional.

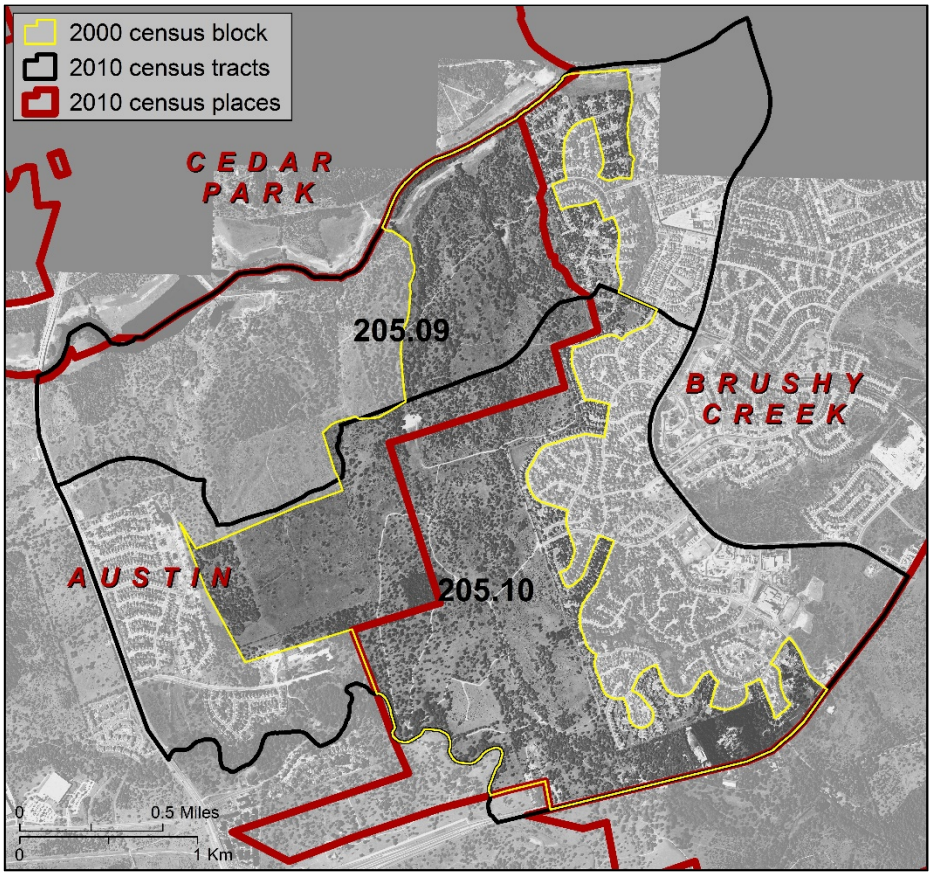

Fig. 1. A 2000 census block in Williamson County, Texas, that intersects two 2010 census tracts and three 2010 places. Background image source: 2000 ASI (CAPCOG) imagery, accessed in January 2016 at ftp://ftp.ci.austin.tx.us/GIS-Data/Regional/javascript/coa_gis.html.

The aerial imagery in Fig. 1 shows that most of the block's area was undeveloped in 2000, and most of the housing was located along the block's eastern boundary, within Brushy Creek. Interpolation by AW would allocate too much population and housing to Austin and possibly too much to tract 205.10 as well. This example demonstrates that it is possible for AW to produce gross errors even when the source units are census blocks, and, as in previously studied settings with larger source zones, we should expect that using ancillary data to model distributions will yield more accurate estimates.

2.4. Alternative approaches

To integrate ancillary data into our interpolation model, we consider two general modeling approaches—*binary dasymetric* modeling (BD) and *target-density weighting* (TDW)—separately and in combination. As detailed in this section, these models are appealing because they are relatively simple in terms of data requirements and ease of implementation, and in past research, they have often performed well relative to more sophisticated models.

2.4.1. Binary dasymetric interpolation

The BD approach is the simplest of a wide range of dasymetric mapping techniques. The general aim of dasymetric mapping is to improve the representation of a spatial distribution by disaggregating counts of a feature of interest from one set of internally heterogeneous zones (typically census units) to another set of zones expected to have relatively uniform densities (Mennis, 2009; Holt & Lu, 2011). In a BD model, there are only two control zones, an inhabited and an uninhabited zone. To apply areal interpolation using a BD model, the implementation mirrors AW but with the measured areas restricted to the inhabited zone.

Many types of ancillary data may be used to delineate zones for BD models. In an early example, Wright (1936) identifies zones through interpretation of topographic maps. More recently, the most common ancillary data type has been land cover and land use data (e.g., Eicher & Brewer, 2001; Holt, Lo, & Hodler, 2004; Langford, 2006; Mennis & Hultgren, 2006; Lin, Cromley, & Zhang, 2011; Cromley, Hanink, & Bentley, 2012; Lin et al., 2013; Schroeder & Van Riper, 2013; Buttenfield, Ruther, & Leyk, 2015; Lin & Cromley, 2015; Ruther, Leyk, & Buttenfield, 2015; Zoraghein et al., 2016). Under a strict definition of dasymetric mapping, the ancillary data must represent zones, which excludes models based on road lengths (as discussed in Section 2.1) or counts of address points (e.g., Tapp, 2010), but line and point data can still be used to define BD models through the use of buffering. For example, one may define the inhabited zone as a 100-foot buffer around roads (Mrozinski & Cromley, 1999).

Another data type that may be effective for BD models is imperviousness data, which describe how much of the land surface is impenetrable by water, typically in raster format with cell values ranging from 0 to 100%. Past research has effectively modeled population densities as a continuous function of imperviousness (Wu & Murray, 2005; Lu, Weng, & Guiying, 2006; Morton & Yuan, 2009; Zandbergen & Ignizio, 2010), but a simpler option, which we consider here, is to reclassify the data, using a single imperviousness threshold to distinguish an inhabited zone for a BD model.

Much research on dasymetric models has focused on settings where there are multiple classes of inhabited zones with different expected density levels. For example, land cover data sets often distinguish different classes of developed land, but making use of such distinctions requires an additional calibration step in order to determine an expected density for each zone class. Existing calibration approaches range in complexity from simple "controlled guesswork" (Wright, 1936) and subjective decisions (Eicher & Brewer, 2001) to systematic sampling from source zones that are representative of each control zone class (Mennis, 2003; Mennis & Hultgren, 2006), and finally to a wide array of statistical modeling techniques (e.g., Langford, Maguire, & Unwin, 1991; Goodchild, Anselin, & Deichmann, 1993; Langford, 2006; Reibel & Agrawal, 2007; Lin, Cromley, & Zhang, 2011; Cromley, Hanink, & Bentley, 2012; Schroeder and Van Riper, 2013; Lin & Cromley, 2015).

Importantly, however, in studies that compare BD and multi-class dasymetric models, BD models are often nearly as accurate as, and sometimes more accurate than, multi-class models (Eicher & Brewer, 2001; Fisher & Langford, 1995; Langford, 2006; Lin, Cromley, & Zhang, 2011; Cromley, Hanink, & Bentley, 2012; Lin et al., 2013; Schroeder and Van Riper, 2013; Lin & Cromley, 2015). Based on these findings, we acknowledge that a well-designed multi-class dasymetric model would likely outperform BD models in the setting of interest, but we also expect the potential accuracy gain of a multi-class model to be small, especially relative to the gain achieved by blending any dasymetric model with a TDW model through a hybrid approach, as indicated by the findings of Schroeder & Van Riper (2013). We therefore

leave the assessment of multi-class models for block-based interpolation to future research and focus here on simpler BD models.

2.4.2. Target-density weighting

The basic assumption of a TDW model is that, within each source zone, the density distribution of the feature of interest is proportional to the density distribution of another, related feature among intersecting target units (Schroeder 2007). For the setting of interest, the TDW assumption is that within each 2000 block, the distribution of 2000 densities is proportional to the densities of a related 2010 characteristic. For example, if a 2000 block intersects two 2010 target units, and one of the target units is twice as dense as the other in 2010, then TDW allocates the 2000 block's characteristics as needed to produce a 2:1 ratio in the estimated densities for the two areas of intersection with the target units.

In the original TDW specification, Schroeder (2007) measures areas and densities using total land areas. An alternative is to restrict TDW's area measures, as in a BD model, to an "inhabited zone" based on some other ancillary data. This approach of dasymetrically refining TDW improves on standard land-area TDW in several settings of tract-based interpolation (Ruther, Leyk, & Buttenfield, 2015; Buttenfield, Ruther, & Leyk, 2015; Zoraghein et al., 2016). We also expect this approach to perform well for block-based interpolation. Given a 2000 block that intersects a small part of a large 2010 unit where most of the land is a roadless forest, standard TDW would assign a low density to the area of intersection, but dasymetric refinement using land cover data or road buffers could indicate that the intersection lay within a developed part of the 2010 unit, resulting in a much higher—and probably more accurate—modeled density within the intersection. Therefore, for each of the BD models we test, we also specify and test a corresponding dasymetrically refined TDW model.

2.4.3. Hybrid approach

Compared directly to BD models, TDW (with or without dasymetric refinement) has generally performed similarly well or better (Schroeder & Van Riper, 2013; Ruther, Leyk, & Buttenfield, 2015; Buttenfield, Ruther, & Leyk, 2015; Zoraghein et al., 2016). There is no need, however, to employ only one of the two. A hybrid approach can leverage the complementary advantages of each individual model. For settings like ours, TDW should generally be effective where distributions remain stable over time (e.g., where the 2010 population distribution is proportionally similar to the 2000 distribution). However, where distributions changed significantly between censuses, or where the density within a source-target intersection is very different from the density of the whole target zone, we may expect a BD model to be more effective. Therefore, as detailed in Section 3.8, we investigate ways to combine the best-performing BD and TDW models through a weighted average, following the approach used by Schroeder and Van Riper (2013) to construct a TDW-dasymetric hybrid in another setting.

3. Data & methods

3.1. General framework

To derive block-based 2000 counts for several levels of 2010 geography, NHGIS's process consists of two phases. In the first, we derive interpolation weights for all intersections between 2000 and 2010 census blocks, where each weight represents the estimated proportion of a 2000 block's 2000 residents and housing units that are located in its intersection with a 2010 block. In the second phase, we aggregate the block-to-block interpolation weights to construct block-to-target-unit weights for each of the target geographic levels. We then apply the aggregated weights to allocate 2000 block counts among target units, and we sum the allocated counts for each target unit to produce the final estimates.

In this process, the critical interpolation setting is from 2000 blocks to 2010 blocks. Therefore, our interpolation model development and testing is focused on the block-to-block setting.

3.2. Base data

We obtain spatial definitions of 2000 and 2010 blocks from NHGIS boundary files (Minnesota Population Center, 2011) based on the U.S. Census Bureau's 2010 TIGER/Line files, and we obtain 2000 block counts from NHGIS tables drawn from Census 2000 Summary File 1. We determine the relationships between 2010 blocks and other 2010 census units from geographic codes in NHGIS block files drawn from 2010 Census Summary File 1.

At this time, NHGIS does not include 2000 data for Puerto Rico or other territories, so the time series are limited to the U.S. proper, as are all results presented here.

3.3. Ancillary data

We use four types of ancillary data from three sources:

- Transportation lines *and* water body polygons from the 2010 TIGER/Line Shapefiles
- Percent imperviousness from the 2001 National Land Cover Database (NLCD 2001, 2011 Edition; Homer et al., 2007)
- 2010 census block population and housing unit counts from 2010 Census Summary File 1

The three sources each provide free, publicly available data with nationwide coverage (or in the case of NLCD data, *nearly* nationwide coverage, as discussed in Section 3.5), and the four data types have each been used effectively in past areal interpolation research. We set aside several other types of ancillary data that would likely be useful for block-based interpolation but are not readily available for the entire U.S., including parcel data and address points.

NLCD 2001 is provided in a raster format at 30-meter resolution, which is unfortunately too coarse to consistently distinguish isolated housing in rural areas (Zandbergen & Ignizio, 2010), but it is still fine enough to distinguish land cover types within most census blocks. Of the 352,082 2000 blocks that require interpolation and are covered by NLCD data, the median land area is 0.56 km^2, and only 9,896 cases (2.8%) have land areas under 9,000 m^2 (the size of 10 NLCD cells).

NLCD 2001 also includes classified land cover data that distinguishes four classes of developed land (open space, and low, medium, and high intensity), which have been used to define BD models in past research (Buttenfield, Ruther, & Leyk, 2015; Ruther, Leyk, & Buttenfield, 2015; Zoraghein et al., 2016). We examined the content of each of these classes by comparison with satellite imagery in six counties with distinct environments and development patterns (Allegheny County, Pennsylvania; Bell County, Texas; Hennepin County, Minnesota; Kootenai County, Idaho; Richland County, South Carolina; and Santa Barbara County, California). We find that the open space class includes many large uninhabited areas (e.g., parks, golf courses, roadways, etc.) but also many low-density residential areas, so it would be problematic either to include the whole class in the inhabited zone or to omit it. Using the imperviousness data offers more flexibility for fine-tuning the inhabited zone's extent.

In our TDW models, the guiding target zone data are the *summed* population and housing unit densities for each 2010 block: (N persons + N housing units) / area. We opt to use this sum because our aim is to use a single TDW model to interpolate census counts of both population and housing characteristics. In some areas, residents may greatly outnumber housing units (e.g., where most residents live in group quarters) or vice versa (e.g., where most of the housing is vacant). Therefore, local spatial distributions of population and housing may differ considerably, and using the densities of one to guide the interpolation of the other could sometimes be grossly inaccurate. Using the summed densities appears to be a suitable compromise, whereby all interpolation is guided by *both* the population *and* housing distributions. It would also be possible to use different 2010 characteristics to guide the interpolation of different 2000 characteristics, but using a single set of weights simplifies the model and ensures that sums of interpolated subtotals will consistently match interpolated totals.[3]

3.4. Inhabited zone definitions

We test eight BD models and eight TDW models corresponding to eight definitions of inhabited zones (Table 2). The first zone definition (L) covers all land area, which we delineate by erasing water

polygons from block polygons. The BD model using the L zone (BD-L) is effectively identical to many past applications of AW, which are also often limited to land area, and the TDW-L model is effectively equivalent to standard TDW without dasymetric refinement. To construct the remaining seven zone definitions, we begin with the L zone (water erased) and then apply one or more of three additional restrictions, identified as D, R, and N.

Table 2. Eight definitions of inhabited zones.

Zone ID	Description
L	All land
D	Developed land: imperviousness ≥ 5%
R	Land within 300 feet of a residential road
N	Land not in transportation use
DR, DN, RN, DRN	Intersections of D, R, & N zones

The D zone is comprised of land within any NLCD 2001 30-meter square cell that is at least 5% impervious. We selected the 5% threshold after inspecting various imperviousness levels compared to satellite imagery in the six counties noted in 3.3. This threshold omits some residential land (mainly in low-density or heavily wooded areas) and includes developed land of all types (commercial, industrial, transportation, institutional, recreational, *etc*.), but overall, the 5% threshold appears to achieve an effective balance between errors of omission and commission. Some past research has also omitted high-intensity developed or highly impervious land from population distribution models because such areas are predominantly industrial or commercial (Holt, Lo, & Hodler, 2004; Morton & Yuan, 2009; Zandbergen & Ignizio, 2010; Buttenfield, Ruther, & Leyk, 2015; Ruther, Leyk, & Buttenfield, 2015), but we found many locales, especially in urban cores, where residential land is highly impervious, so we impose no upper bound.

The R zone is comprised of land within 300 feet of a residential road according to 2010 TIGER/Line road definitions. We inspected various examples of each TIGER/Line road class and identified a class as

residential if it, at least occasionally, provides direct access to housing.[4] Similarly, we considered a range of possible buffer widths around roads, beginning with 100 feet, as used in prior research (Mrozinski & Cromley, 1999; Lin *et al.*, 2013; Lin & Cromley, 2015). We determined that a distance of 300 feet, though it extends beyond most housing in typical urban settings, is more suitable for exurban and rural areas, where long driveways are common, and for large housing complexes, which often contain access roads that are not represented in TIGER/Line files. Using 2010 road definitions to model 2000 distributions is not ideal, but due to major accuracy improvements between TIGER/Line versions, the 2000 TIGER/Line roads can deviate greatly from the more accurate block boundaries in the 2010 TIGER/Line files, so we have opted to construct the R zone from 2010 roads. It may be possible to leverage 2000 TIGER/Line information to distinguish 2010 TIGER/Line roads that likely existed in 2000, but we achieve a similar outcome by intersecting the R and D zones, as discussed below.

The N zone is comprised of all land not covered by transportation features, as determined by applying a small buffer around 2010 TIGER/Line roads and railroads. We use different buffer widths to represent an expected minimum distance from center line to housing for each feature class.[5] This type of zone has not often been used in dasymetric models. Morton & Yuan (2009) omit pixels that intersect major highways from one of their models, and Zandbergen & Ignizio (2010) investigate a model using "cleaned" imperviousness data that effectively masks out isolated roads. Zandbergen & Ignizio's findings suggest that excluding road areas is generally unhelpful, but their approach and setting differ from ours in several ways. It seems that excluding land in transportation use should at least reduce errors in the many instances where block boundaries that follow transportation features were adjusted by a small distance. Such changes may comprise a large portion of a block's area within the D or R zones, but if the adjustment occurred mainly over transportation features, the N zone should properly indicate that little or no population was involved.

The last four zone definitions (DR, DN, RN, and DRN) are the intersections of each possible combination of the D, R, and N zones. For example, the DR zone is comprised of land that is at least 5% impervious *and* within 300 feet of a residential road. Prior research has shown that dasymetric models based on land cover data yield more accurate results if they are further restricted to areas within road buffers (Lin *et al.*, 2013; Lin & Cromley, 2015), and in our case, we believe intersecting the D and R zones should mitigate key problems for each individual zone definition. The R zone undesirably includes areas where roads were built between 2000 and 2010, but if there was no impervious surface in these areas in 2001, the DR zone properly excludes them. The D zone includes areas of impervious surface that are far removed from residential roads, but such areas are typically uninhabited industrial or commercial complexes, quarries, airports, expressways, golf courses, *etc.*, and the DR zone excludes them. As for the N zone, it may be a poor model on its own given that it assigns *lower* weights to areas with more roads, but by intersecting the N zone with the D or R zone, the D and R zones' more restrictive extents might alleviate the N zone's inverse weighting problem, and the N zone's key benefit—preventing population from being allocated to road areas—might then outweigh its costs.

Comparing the modeled distributions in Fig. 2 with the aerial imagery in Fig. 1 reveals several potential pros and cons for the tested models. The BD-L (AW) model, as expected, assigns undesirably large weights to the block's western parts, which were generally undeveloped in 2000. In contrast, the BD-D and BD-DR models properly assign more weight to the east side of the block. Red circles highlight the main differences between the BD-D and BD-DR models, areas that today encompass a plant nursery and a storage facility: nonresidential uses that the BD-DR inhabited zone effectively excludes.

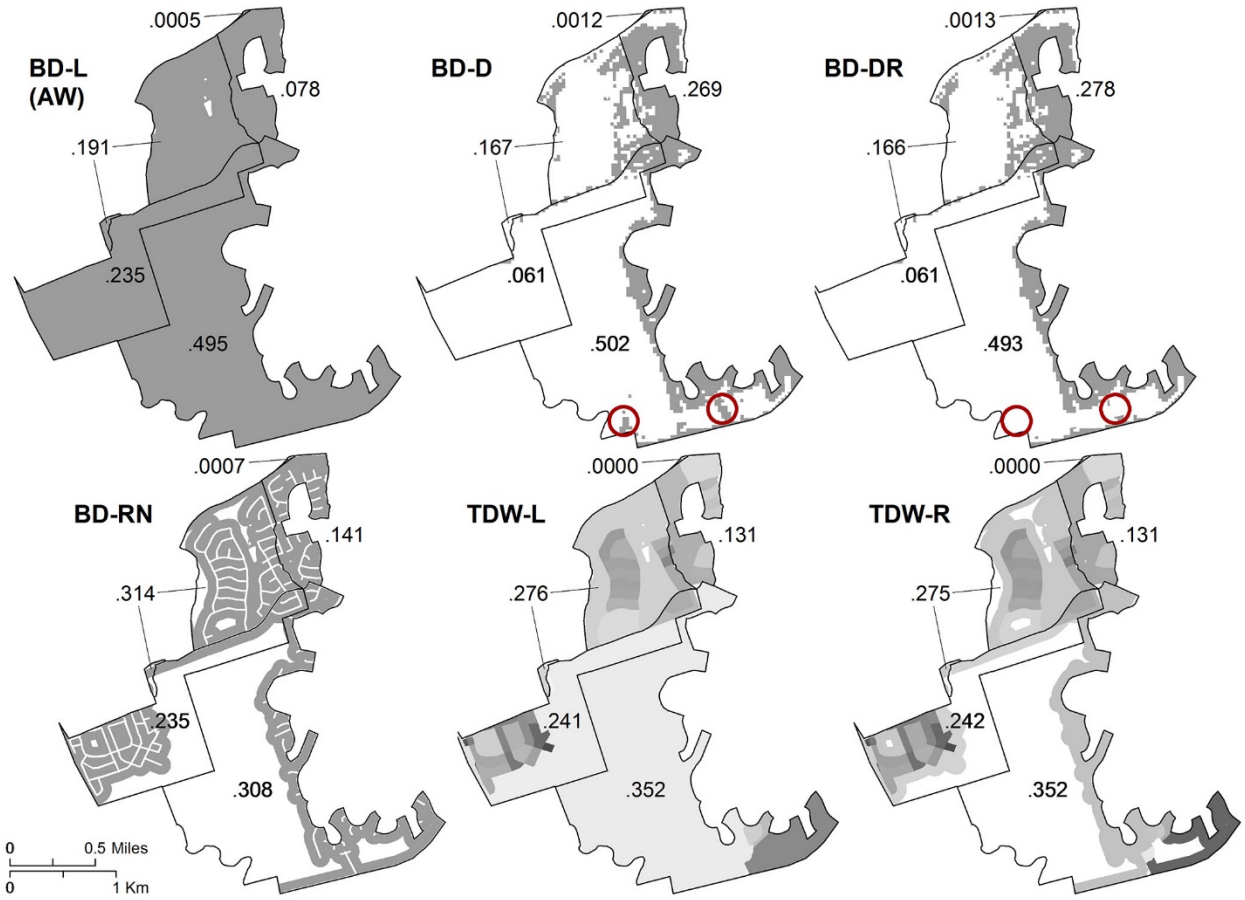

Fig. 2. Density distributions (in shades of gray) and interpolation weights given by six of the tested models for the example case from Fig. 1.

The lower examples in Fig. 2 all indicate how using ancillary data from 2010 can be problematic. Because there are many new 2010 roads in the western parts of the block, the BD-RN model assigns even more weight to those parts than does the BD-L model. Similarly, because the 2010 population and housing unit density was relatively high in the western parts, the TDW models also assign high weights there. Still, some potential advantages of TDW are also apparent. Because there is no 2010 population or housing in the 2010 block that intersects the northern tip of this 2000 block, the TDW models appropriately assign no weight to the small sliver there. Also, importantly, the rapid growth that occurred in this area is not typical of all split blocks. If this area's population had been more stable, we

might reasonably expect the continuously varying densities of the TDW models to be more accurate than the simpler uniform densities of the BD models.

3.5. Assessment setting

The assessment approach used here, mirroring that of Schroeder & Van Riper (2013), is to simulate the setting of interest by interpolating from *pairs* of neighboring source zones back to individual source zones, for which the actual counts are known. Specifically, we first select all 2000 blocks that require interpolation—those that have nonzero population or housing counts and share land with multiple 2010 units in any of 14 census summary levels (the 10 levels of NHGIS's first standardized data release plus 4 levels that NHGIS may include in a future release: 113th Congressional Districts and elementary, secondary, and unified school districts). For our primary assessment work, we also limit the selection to areas that are completely covered by NLCD data, which omits 303 blocks of interest in Alaska. For each of the remaining 352,082 blocks of interest, we identify the nearest neighboring block (whether it is a block of interest or not) according to distance between centroids, sum the population and housing unit counts for the block pair, estimate the sum of population and housing units within the block of interest by interpolating from the pair total using each of the tested models, and finally compute each model's errors for all blocks of interest.[6]

In the proxy setting, the target zones are effectively 2000 blocks, so true "target-density" weighting would entail using 2000 block data to guide the interpolation. To better simulate the setting of interest, our TDW applications instead use 2010 block data, again mirroring the approach of Schroeder & Van Riper (2013).

We also undertake a separate series of tests to optimize models for the areas in Alaska not covered by NLCD 2001 data, omitting all models that use the D zone.

3.6. Error measures

We compute (and aim to minimize) errors in the *sums* of 2000 population and housing unit counts for the same reason that we use the corresponding 2010 sums as ancillary data in the TDW models: our aim is to use a single model to interpolate all types of census counts, and we believe this sum is a suitable general proxy for all population and housing characteristics.

To summarize errors for each model, we first consider the mean absolute error (MAE) and root mean square error (RMSE) in estimated counts. Block counts, however, have an extremely skewed distribution. Among the tested block pairs, the median sum of 2000 population and housing units is 72, the 99th percentile is 1,621, and the maximum is 34,586. Given this distribution, a small number of cases can have an outsize influence on count error summary statistics. For example, using the BD-DR model, the largest 0.1% of errors account for 48% of the sum of squared errors.

We therefore also consider the mean absolute proportion error (MAPE) and root mean square proportion error (RMSPE), defining a "proportion error" to be the difference between the actual and estimated proportion of each block pair's count in the corresponding block of interest. Proportion errors are constrained between -1 and 1, and the outliers are much less extreme than among count errors. Proportion errors, however, tend to be larger for block pairs with *small* counts. The MAPE in BD-DR estimates for the 12% of block pairs with counts less than 10 is 0.231, which is nearly twice the MAPE for the 59% of pairs with counts greater than 50 (0.120). This makes sense given that in block pairs with small counts, it is common for all population and housing to be located in a single block, resulting in actual proportions of 0 and 1, which maximizes the potential deviation from estimated proportions.

In short, a refinement strategy that minimizes MAE or RMSE may be too sensitive to cases with very large counts, and a strategy that minimizes MAPE or RMSPE may be too sensitive to cases with very small counts. We base our final model selection on a compromise approach: we compute *weighted* MAPE and RMSPE statistics, where each case is weighted by the *log* of the source zone count (the block

pair's sum of population and housing). In these log-weighted error summaries (LW-MAPE and LW-RMSPE), large-count cases are given only moderately higher weights than small-count cases, striking a balance between the extremes of the unweighted error summaries.

3.7. Zero denominators

Both the BD and TDW models assume that each 2000 block contains some area classified as inhabited, and the TDW model further assumes that there is some 2010 population and housing in the intersecting 2010 blocks. In cases where these assumptions do not hold, the standard model specifications produce divide-by-zero errors. To avoid this problem, our implementation for each model "cascades" through less restrictive models until it reaches a model with a nonzero denominator. The final ordering is based on model performance according to LW-RMSPE in the proxy setting, as reported in Section 4.1. For example, if the TDW-DR implementation encounters a zero denominator, it defaults to the TDW-R model, which in turn defaults to TDW-L, to BD-DR, to BD-R, and finally to BD-L (AW).

3.8. Hybrid model definitions

The general formula we use for a hybrid TDW-BD model is

$$\hat{p}_{\mathrm{H}} = w_{\mathrm{T}}\hat{p}_{\mathrm{T}} + (1 - w_{\mathrm{T}})\hat{p}_{\mathrm{B}} \quad (1)$$

where $\hat{p}_{\mathrm{T}}$ and $\hat{p}_{\mathrm{B}}$ are interpolation weights (i.e., estimated proportions) for a given source-target zone intersection as given by a TDW and a BD model, respectively; $\hat{p}_{\mathrm{H}}$ is the "hybrid" interpolation weight; w_{T} is the TDW model weight; and the BD model weight is set to $(1 - w_{\mathrm{T}})$.

Following Schroeder & Van Riper (2013), we construct two hybrid models: one that uses constant model weights (Hybrid-CW) and one in which the model weights vary among source zones (Hybrid-VW). To select a constant value for w_{T}, we apply weighted least-squares (WLS) regression to fit the model

$$(\hat{p}_{\mathrm{H}} - \hat{p}_{\mathrm{B}}) = w_{\mathrm{T}}(\hat{p}_{\mathrm{T}} - \hat{p}_{\mathrm{B}}) \quad (2)$$

using data from the proxy setting described in Section 3.5. The observation weights are the same as for the LW-MAPE and LW-RMSPE statistics (the log of each block pair's sum of population and housing units). The least-squares fitted value of w_{T} is therefore the value that minimizes the LW-RMSPE for the proxy setting.

The design of the Hybrid-VW model is based on the expectation that the relative accuracy of BD and TDW estimates is proportional to the degree of change in distributions between the source and target zone years. Where distributions are stable, TDW should be more effective; where distributions have changed greatly, BD should be more effective. Accordingly, the Hybrid-VW model sets w_{T} to be a linear function of the absolute rate of change in each source zone:

$$w_{\mathrm{T},s} = \alpha_0 + \alpha_1 \left| \widehat{\Delta y_s} \right| \quad (3)$$

where $\widehat{\Delta y_s}$ is an estimate of the normalized rate of change in the feature of interest within each source zone *s* between the source and target years (in our setting, between 2000 and 2010 sums of population and housing units). A "normalized change rate" divides the difference in values by their sum, which helpfully prevents the extreme positive skew that standard rate measures often produce (Schroeder & Van Riper, 2013). To estimate 2010 counts in source zones for the change measure, we use the counts produced during the TDW model implementation, which are effectively a BD interpolation of 2010 block counts using the same inhabited zone definition as the given TDW model. To obtain fitted values for the α coefficients in equation (3), we substitute its right-hand side for w_{T} in equation (2) and fit the model using WLS regression as for the Hybrid-CW model.

4. Results & discussion

4.1. Assessment of BD & TDW models

In the proxy setting, all of the TDW models outperform all of the BD models on all error measures (Table 3). This is consistent with past findings, where TDW models have outperformed BD models on

most—though sometimes not all—error measures (Schroeder & Van Riper, 2013; Ruther, Leyk, & Buttenfield, 2015; Buttenfield, Ruther, & Leyk, 2015; Zoraghein et al., 2016). We conclude that, in our setting of interest, 2010 block data are the most informative ancillary data type of those we examine (as used in the relatively simple models we consider here). Even for the simplest model using 2010 block data (TDW-L), the LW-RMSPE is 19% less than for the best-performing BD model (BD-DR), which is in turn 10% less than the LW-RMSPE for BD-L (AW), the simplest model.

Table 3. Summary of errors for the tested models in the proxy setting.

Model	MAE	RMSE	MAPE	RMSPE	LW-MAPE	LW-RMSPE
BD-L (AW)	20.55	61.49	.1626	.2451	.1519	.2316
BD-N	20.87	62.52	.1637	.2501	.1537	.2365
BD-R	18.53	56.91	.1525	**.2272**	.1396	.2114
BD-RN	18.72	57.42	**.1524**	.2310	.1403	.2152
BD-D	17.48	53.98	.1565	.2358	.1396	.2136
BD-DN	17.57	54.53	.1565	.2412	.1402	.2183
BD-DR	**17.07**	**53.11**	.1541	.2323	**.1368**	**.2093**
BD-DRN	17.10	53.27	.1540	.2376	.1372	.2138
TDW-L	10.82	45.89	.1047	.1947	.0909	.1695
TDW-N	10.90	46.12	.1051	.1963	.0914	.1712
TDW-R	10.55	45.60	***.1027***	**.1915**	**.0886**	**.1659**
TDW-RN	10.60	45.80	.1029	.1928	.0890	.1673
TDW-D	10.50	45.43	.1036	.1940	.0890	.1672
TDW-DN	10.57	45.70	.1040	.1960	.0897	.1692
TDW-DR	**10.45**	**45.36**	.1033	.1938	.0887	.1668
TDW-DRN	10.52	45.61	.1037	.1956	.0893	.1687
Hybrid-CW	11.12	42.67	.1067	.1801	.0920	.1574
Hybrid-VW	***10.37***	***40.52***	**.1028**	***.1752***	***.0877***	***.1525***

Notes: N = 352,082. **Bold** numbers indicate the lowest value for each measure among each class of models. ***Bold italic*** numbers indicate the lowest values overall. MAE = mean absolute error; RMSE = root mean square error; MAPE = mean absolute proportion error; RMSPE = root mean square proportion error; LW = log-weighted; BD = binary dasymetric; AW = areal weighting; TDW = target-density weighting; Hybrid-CW = constant-weight hybrid; Hybrid-VW = variable-weight hybrid; see Table 2 for zone identifiers.

According to both LW-MAPE and LW-RMSPE, BD-DR is the most accurate BD model, and TDW-R is the most accurate TDW model, so these are the two models we use to construct hybrid models (Section 4.2). These two models do not, however, yield the lowest values for all error summary statistics. Among the BD models, according to MAPE, BD-RN is most effective, and according to RMSPE, BD-R is most effective. Among the TDW models, according to both MAE and RMSE, TDW-DR is most effective. Given that the MAE and RMSE are more sensitive to large-count cases, and the MAPE and RMSPE are more

sensitive to small-count cases, these outcomes suggest that, among small-count cases, it is more effective to limit the inhabited zone only to 2010 road buffers (with or without excluding land in transportation use) than to limit the zone to 2001 developed (≥ 5% impervious) land in any way. This may indicate that in rural areas, where low block counts are more common, road buffers generally correspond better to distributions of population and housing than does imperviousness even though the road buffers in this case describe conditions 10 years removed from the census. Meanwhile, among blocks with very large counts, it appears that the most effective inhabited zone definition, for both BD and TDW models, is the intersection of the road buffer and developed land (DR).

Somewhat surprisingly, models with the N restriction are less effective than corresponding models without the N restriction in nearly all cases. The only exceptions are where the MAPEs for BD-RN and BD-DRN are slightly lower than for BD-R and BD-DR, respectively. Multiple factors may contribute to this result. Most importantly, in block parts that are relatively large, larger quantities of road surface may generally correspond to higher—not lower—population and housing densities. In residential developments, parcel sizes are generally smaller and cul-de-sacs are more common than in non-residential areas, so road density tends to be higher in residential areas, and erasing land in transportation use will often diminish the weight given to residential parts of a block. In addition, upon closer inspection, we discovered that the N models occasionally perform poorly even where we expect them to be best—namely, where one block in a pair is comprised mainly of land in transportation use, leaving little space for housing. Models without the N restriction can assign unrealistically high counts to such areas, but in fact, official counts for such areas occasionally *are* unrealistically high because of mismatches between where the Census locates people and where people actually live. We have identified instances where large housing complexes have had all of their units counted in a small neighboring block comprised mainly of road surface. Such cases, though uncommon, occur often enough to diminish what could otherwise be a key advantage of the N models.

Overall, the choice of the inhabited zone definition matters more for BD models than for TDW. Among TDW models, the difference between the highest and lowest values for each error measure is small; the largest difference is in MAE, for which the lowest value is only 4% less than the highest value. In contrast, the differences for BD models are larger across all measures, with lowest values that are 7 to 18% less than the highest values. This finding differs from past research where the impact of dasymetric refinement on TDW outcomes has been greater (Ruther, Leyk, & Buttenfield, 2015; Buttenfield, Ruther, & Leyk, 2015). One explanation may be that, in our setting, the target zones (2010 blocks) tend to nest within source zones (2000 blocks or block pairs) more often than in the previously tested settings where the zones are tracts. Where target zones nest completely within source zones, the inhabited zone definition has essentially no effect on TDW outcomes. The new results suggest that, in our setting, undertaking further refinement of the zone definitions (e.g., optimizing the width of the road buffer or the imperviousness threshold) would likely have little impact on the TDW results, though it could be important for the BD results, and by extension, for the hybrid models as well.

4.2. Fitted hybrid models

The fitted Hybrid-CW model assigns a weight of 0.7247 to TDW-R and 0.2753 to BD-DR, which yields an RMSE, RMSPE, and LW-RMSPE lower than any of the non-hybrid models (Table 3). However, Hybrid-CW's MAE, MAPE, and LW-MAPE are *higher* than for any of the TDW models. Because the root mean square statistics are more sensitive to large errors than the mean statistics, the results suggest that in most cases using TDW alone is more effective than a simple constant-weight hybrid with BD (as indicated by the mean values), but the Hybrid-CW model yields fewer very large errors than using TDW alone (as indicated by the root mean square values).

The fitted formula for TDW-R weights in the Hybrid-VW model is:

$$w_{\mathrm{T},s} = 0.9192 - 0.8057\left|\widehat{\Delta y_s}\right| \quad (4)$$

As expected, $w_{T,s}$ is high where population and housing are stable (0.9192 in a source zone with no estimated change), and it declines as the magnitude of change increases (to a lower limit of 0.9192 – 0.8057 = 0.1135). In the example block from Fig. 1, the estimated growth is large, from 2,111 in 2000 to 8,137 in 2010, which gives a normalized change rate of 0.5880 and a TDW-R weight of 0.4454, well below the Hybrid-CW model's 0.7247. Fig. 3 illustrates the effect of this difference on the two hybrid models' outcomes. By giving a lower model weight to TDW-R, the Hybrid-VW model appropriately assigns higher interpolation weights to the eastern parts of the block.

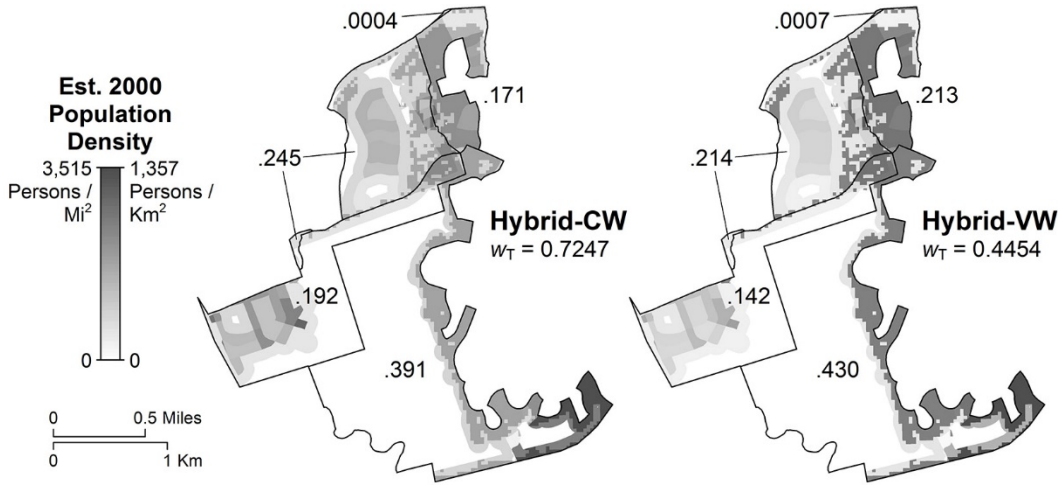

Fig. 3. Density distributions and interpolation weights given by the constant-weight (Hybrid-CW) and variable-weight (Hybrid-VW) hybrid models for the example case from Fig. 1.

Table 3 shows that Hybrid-VW is the most effective tested model by nearly all measures. (TDW-R's MAPE is lower than Hybrid-VW's by only 0.0001.) In terms of LW-RMSPE, Hybrid-VW errors are 3% lower than Hybrid-CW's, 8% lower than TDW-R's, 27% lower than BD-DR's, and 34% lower than AW's. Therefore, we use the Hybrid-VW model for NHGIS's first release of geographically standardized time series tables in all areas covered by NLCD 2001 data.

Despite Hybrid-VW's overall advantages, there are still 112,417 cases (31.9% of the proxy setting) where AW yields a smaller absolute error, but—as among tract population estimates (Logan, Stults, & Xu, 2016)—in most cases where AW outperforms Hybrid-VW, the difference between the two models' estimates is small. Where the estimates differ by more than 50 in the proxy setting (N = 24,560), AW is

more accurate in only 9.2% of cases. There are also 1,905 cases where AW's absolute error is greater than 100 and Hybrid-VW's is less than 10, but there are only 69 cases where the opposite is true. In the extreme, there are 65 cases where AW's absolute error exceeds Hybrid-VW's by more than 1,000 and only 7 cases where the opposite is true.[7] In short, although Hybrid-VW is not always more accurate than AW, it is rarely much worse, and it is commonly much better.

4.3. Secondary Alaska model

For the 303 blocks of interest in parts of Alaska not covered by NLCD 2001 data (everywhere outside of a section centered on Anchorage, from Denali in the north to the southern reaches of the Kenai Peninsula), the most effective BD model (using the L, N, R, or RN zones) is BD-R, with an LW-RSMPE of 0.2665. The lowest LW-RMSPE among TDW models is, somewhat surprisingly, for TDW-N (0.16032), but TDW-L's is only slightly higher (0.16033), and TDW-L's LW-MAPE (0.0751) is less than TDW-N's (0.0753), so we opt to use the simpler TDW-L model for the hybrid models. The Hybrid-CW model sets w_T = 0.8773 and yields an LW-RMSPE of 0.1575. The Hybrid-VW model yields an only slightly lower LW-RMSPE (0.1574), and its α_1 coefficient is not statistically significant, so we opt to use the simpler Hybrid-CW model for the final secondary Alaska model.

4.4. Effects of refinement

Although Table 3 reveals some major differences among the tested models in the proxy setting, the relative differences will necessarily be smaller, on average, in the actual setting of interest. After all, the uncertainty in block-based 2000 counts is small for most target 2010 units (Table 1), and interpolation refinements may affect only the "uncertain portion" of block count assignments.

Table 4 demonstrates that the effects of model refinement are nevertheless still substantial in many areas. For example, using the Hybrid-VW model instead of BD-L (AW) alters 2000 population estimates for 2010 urban areas by an average of 425 persons (among *all* urban areas). Among the nearly

20,000 places where the 2000 population is uncertain, Hybrid-VW estimates differ from AW estimates by about 6.7% on average. Among tracts, the average difference is quite small, but there are still 955 tracts where the difference is greater than 5% (relative to the mean of the BD-L and Hybrid-VW estimates), and for 317 tracts, the difference is greater than 25%. Not surprisingly, the four highest summary levels (counties, CBSAs, congressional districts, and states) exhibit no major effects, which confirms that for those levels, the exact model design is generally unimportant. At all other levels, model refinement significantly alters some estimates, and the results in the proxy setting—as in Logan et al.'s (2016) research—indicate that the refined estimates are generally more reliable.

Table 4. Summary of absolute differences in block-based estimates of 2000 population for 2010 census units, comparing results given by the BD-L (AW) and Hybrid-VW models.

	All units			2000 population is uncertain			AD > 5% of 2 estimates' mean			AD > 25% of 2 estimates' mean		
Geographic level	N	Mean AD	Mean (AD as % of 2 estimates' mean)	N	Mean AD	Mean (AD as % of 2 estimates' mean)	N	% of all units	% of uncer-tain units	N	% of all units	% of uncer-tain units
Block groups	217,740	9	1.0	55,792	35	3.9	7,366	3.4	13.2	2,020	0.9	3.6
Tracts	73,057	7	0.6	24,375	20	1.8	955	1.3	3.9	317	0.4	1.3
County subdivisions	35,703	6	0.4	19,136	12	0.8	426	1.2	2.2	71	0.2	0.4
ZCTAs	32,989	45	2.2	29,945	50	2.4	2,554	7.7	8.5	532	1.6	1.8
Places	29,261	34	4.5	19,620	50	6.7	4,180	14.3	21.3	1,462	5.0	7.5
Urban areas	3,573	425	2.8	3,554	428	2.8	448	12.5	12.6	24	0.7	0.7
Counties	3,143	3	0.0	1,802	5	0.0	0	0.0	0.0	0	0.0	0.0
CBSAs	942	3	0.0	621	4	0.0	0	0.0	0.0	0	0.0	0.0
Cong. districts	436	12	0.0	389	14	0.0	0	0.0	0.0	0	0.0	0.0
States	51	1	0.0	37	2	0.0	0	0.0	0.0	0	0.0	0.0

AD = absolute difference, ZCTAs = ZIP Code Tabulation Areas, CBSAs = core based statistical areas, Cong. districts = 111th Congressional Districts

5. Conclusions

To produce high-quality block-based estimates of 2000 census counts for 2010 units, we define and test several areal interpolation models, integrating several types of ancillary data. We arrive at a hybrid model that blends a TDW and BD model through a variably weighted average, making use of 2010 block data, imperviousness data, road buffers, and water polygons. Although the final estimates from this hybrid model are generally similar to areal weighting estimates, there are many cases where the estimates differ greatly (Table 4), and given the hybrid model's relatively strong performance in a

proxy test setting (Table 3), we may assume that it yields substantial accuracy gains among the target census units as well.

This work has uncovered several additional findings that should be relevant elsewhere. Most notably: when allocating 2000 block counts to 2010 units, uncertainty due to boundary misalignment is pervasive and occasionally large; using an imperviousness threshold together with a road buffer can be an effective way to construct a BD model; omitting land in transportation use from a BD inhabited zone is generally unhelpful; in our test setting, TDW is considerably more accurate than any BD model, and the exact definition of inhabited area used in the TDW model appears to be relatively unimportant; and lastly, in accord with a previous finding (Schroeder and Van Riper, 2013), a variable-weight hybrid of TDW and BD is an effective way to exploit context-dependent advantages of each approach.

Estimates using the final hybrid model are now publicly available via NHGIS along with lower and upper bounds for each estimate. A primary goal of our future work will be to extend NHGIS's geographically standardized time series to cover more years and characteristics. To cover more years, we must address new challenges due to the varying availability and quality of data across time. To cover more characteristics, particularly those that the Census does not report at the block level (*e.g.*, income, educational attainment, nativity, *etc.*), we must address the challenge of interpolating from larger units, potentially using related block characteristics as a guide. Geolytics and the LTDB both interpolate tract data using block populations as a weighting factor, but weighting by population alone for all interpolated characteristics results in uniform rate estimates among all parts of each source tract. It remains to be determined what effect this condition may have in application settings and whether there is a practical means of using multiple block-level characteristics to model distributions more effectively.

Geolytics and LTDB also both use tract data as the basis for all counts other than total population. Tracts are the smallest units for which some counts are tabulated, but there are also many counts available for block groups and blocks, and for such counts, tracts are an unnecessarily coarse basis. The

comparisons of block-based estimates provided here and by Logan, Stults, & Xu (2016) therefore directly pertain only to LTDB's and NHGIS's *total population* estimates. For all other standardized counts, we may expect that the differences between NHGIS's block-based estimates and LTDB's tract-based estimates will generally be larger, and that the block-based estimates could occasionally be much more accurate. E.g., where 2000 blocks—but *not* 2000 tracts—nest exactly within 2010 tracts, only the tract-based counts require estimation; the block-based counts are exact. We leave to future research the assessment of these different approaches for counts (and rates) other than total population.

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]. David Van Riper advised on the research and data product design, and Kevin Horne and Jacob Wellington implemented the software that generates the final standardized data. The author also gratefully acknowledges support from the Minnesota Population Center [NICHD R24HD041023].

Notes

[1] The initial 65 tables cover: Total Population and Persons by Sex, by Age, by Race, by Hispanic or Latino Origin, by Household and Group Quarters Type, by Relationship to Householder, and by Housing Tenure (including some cross-tabulations of these subjects); Total Households and Households by Type and Size; Total Families, Persons in Families, and Families by Type and Presence and Age of Children; and Total Housing Units and Housing Units by Occupancy and Vacancy Status, by Tenure, and by Race of Householder. Follow-up releases will continue to broaden the subject coverage.

[2] It appears that Geolytics employed road-based interpolation of block data only when producing its original estimates for 2000 census units. Geolytics' newer estimates for 2010 census tracts are considerably less accurate than either LTDB's or NHGIS's block-based estimates (Logan, Stults, & Xu, 2016).

[3] For example, if we used total household counts to guide interpolation of population in households but used total population counts to guide interpolation of other population groups, then in some cases, due to disproportionate

changes in household and population distributions between censuses, an area's estimated count of population in households could *exceed* its estimated total population. A more sophisticated model might impose constraints to prevent such outcomes, but that is beyond the scope of the present work.

[4] Our residential road classes are S1200: secondary road; S1400: local neighborhood road, rural road, city street; S1640: service drive; S1730: alley; S1740: private road for service vehicles; S1750: internal U.S. Census Bureau use; and S1780: parking lot road. We identify all other road classes as nonresidential, including primary roads, ramps, trails, and stairways. We also omit any segments classified as bridges, tunnels, or fords.

[5] Our transportation-use buffer widths are 60 feet for primary roads (S1100), 45 feet for secondary roads (S1200), 30 feet for most other roads (S1400, S1630, S1640, S1780, S2000) and for railroads and transit lines (R1011, R1051, R1052), and 15 feet for trails, alleys, private service roads, and "internal use" roads (all other "S" classes).

[6] Among the 352,082 blocks of interest, there are 24,432 blocks (6.9%) involved in "duplicate pairs," where two blocks of interest are each nearest the other. In order to maintain exactly one observation for each block of interest, we include each duplicate pair as a unique observation, even though the measures of error used here are identical for each block in a pair. A sensitivity analysis, wherein one of each of the duplicate pairs was omitted, yielded results very similar to those produced when all duplicate pairs were included, with no model coefficient or error summary statistic differing by more than a few percentage points.

[7] A comparison of block boundaries with satellite imagery reveals that in 5 of these 7 cases where Hybrid-VW's error is much larger than AW's, the block pair consists of a large block and a small block where the Census apparently allocated a large prison population *incorrectly* to the large block in 2000 and *correctly* to the small block in 2010. AW thus "outperforms" Hybrid-VW in these cases only because Hybrid-VW assumes that the 2000 distribution resembles the 2010 distribution, which agrees with satellite imagery if not with official census counts. The other 2 cases of very poor Hybrid-VW estimates are for 2 heavily populated neighboring blocks in Fort Pendleton, California, where it appears there was a large shift in population between 2000 and 2010, possibly due to real relocations of military personnel or perhaps, as in the other 5 cases, due to data misallocations in one year or the other.

References

Beam, A. (2012). New SC-NC border will affect some residents. *The State*. Retrieved from http://www.thestate.com/news/article14398952.html

Buttenfield, B. P., Ruther, M., & Leyk, S. (2015). Exploring the impact of dasymetric refinement on spatiotemporal small area estimates. Cartography and Geographic Information Science, 42(5), 449-459.

Cromley, R. G., Hanink, D. M., & Bentley, G. C. (2012). A quantile regression approach to areal interpolation. Annals of the Association of American Geographers, 102(4), 763-777.

Eicher, C. L., & Brewer, C. A. (2001). Dasymetric mapping and areal interpolation: Implementation and evaluation. Cartography and Geographic Information Science, 28(2), 125-138.

Fisher, P. F., & Langford, M. (1995). Modelling the errors in areal interpolation between zonal systems by Monte Carlo simulation. Environment and Planning A, 27(2), 211-224.

Goodchild, M. F., Anselin, L., & Deichmann, U. (1993). A framework for the areal interpolation of socioeconomic data. Environment and Planning A, 25(3), 383-397.

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

Greenhouse, L. (1998). THE ELLIS ISLAND VERDICT: THE RULING; High Court Gives New Jersey Most of Ellis Island. *The New York Times*. Retrieved from http://www.nytimes.com/1998/05/27/nyregion/ellis-island-verdict-ruling-high-court-gives-new-jersey-most-ellis-island.html

Gregory, I. N. (2002). The accuracy of areal interpolation techniques: Standardising 19th and 20th century census data to allow long-term comparisons. Computers, environment and urban systems, 26(4), 293-314.

Holt, J. B., Lo, C. P., & Hodler, T. W. (2004). Dasymetric estimation of population density and areal interpolation of census data. Cartography and Geographic Information Science, 31(2), 103-121.

Holt, J. B., & Lu, H. (2011). Dasymetric mapping for population and sociodemographic data redistribution. In X. Yang (Ed.), Urban remote sensing: Monitoring, synthesis and modeling in the urban environment (pp. 195-210). Chichester, UK: John Wiley & Sons.

Homer, C., Dewitz, J., Fry, J., Coan, M., Hossain, N., Larson, C., … Wickham, J. (2007). Completion of the 2001 National Land Cover Database for the conterminous United States. Photogrammetric Engineering and Remote Sensing, 73(4), 337-341.

Langford, M. (2006). Obtaining population estimates in non-census reporting zones: An evaluation of the 3-class dasymetric method. Computers, Environment and Urban Systems, 30(2), 161-180.

Langford, M., Maguire, D. J., & Unwin, D. J. (1991). The areal interpolation problem: Estimating population using remote sensing in a GIS framework. In I. Masser, & M. Blakemore (Eds.), Handling geographical information: Methodology and potential applications (pp. 55-77). London: Longman.

Lin, J., & Cromley, R. G. (2015). A local polycategorical approach to areal interpolation. Computers, Environment and Urban Systems, 54, 23-31.

Lin, J., Cromley, R. G., Civco, D. L., Hanink, D. M., & Zhang, C. (2013). Evaluating the use of publicly available remotely sensed land cover data for areal interpolation. GIScience & Remote Sensing, 50(2), 212-230.

Lin, J., Cromley, R., & Zhang, C. (2011). Using geographically weighted regression to solve the areal interpolation problem. Annals of GIS, 17(1), 1-14.

Logan, J. R., Stults, B. J., & Xu, Z. (2016). Validating population estimates for harmonized census tract data, 2000–2010. Annals of the American Association of Geographers, 106(5), 1013-1029.

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), 412-420.

Lu, D., Weng, Q., & Li, G. (2006). Residential population estimation using a remote sensing derived impervious surface approach. International Journal of Remote Sensing, 27(16), 3553-3570.

Martin, D., Dorling, D., & Mitchell, R. (2002). Linking censuses through time: problems and solutions. Area, 34(1), 82-91.

Mennis, J. (2003). Generating surface models of population using dasymetric mapping. The Professional Geographer, 55(1), 31-42.

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

Mennis, J., & Hultgren, T. (2006). Intelligent dasymetric mapping and its application to areal interpolation. Cartography and Geographic Information Science, 33(3), 179-194.

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

Morton, T. A., & Yuan, F. (2009). Analysis of population dynamics using satellite remote sensing and US census data. Geocarto International, 24(2), 143-163.

Mrozinski, R. D., & Cromley, R. G. (1999). Singly-and Doubly-Constrained Methods of Areal Interpolation for Vector-based GIS. Transactions in GIS,3(3), 285-301.

Reibel, M., & Agrawal, A. (2007). Areal interpolation of population counts using pre-classified land cover data. Population Research and Policy Review, 26(5-6), 619-633.

Reibel, M., & Bufalino, M. E. (2005). Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems. Environment and Planning A, 37(1), 127-139.

Ruther, M., Leyk, S., & Buttenfield, B. P. (2015). Comparing the effects of an NLCD-derived dasymetric refinement on estimation accuracies for multiple areal interpolation methods. GIScience & Remote Sensing, 52(2), 158-178.

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

Schroeder, J. P., & Van Riper, D. C. (2013). Because Muncie's densities are not Manhattan's: Using geographical weighting in the expectation–maximization algorithm for areal interpolation. Geographical Analysis, 45(3), 216-237.

Tapp, A. F. (2010). Areal interpolation and dasymetric mapping methods using local ancillary data sources. Cartography and Geographic Information Science, 37(3), 215-228.

Tatian, P. A. (2003). CensusCD Neighborhood Change Database (NCDB): 1970-2000 Tract Data: Data Users' Guide. Washington, DC: The Urban Institute, in collaboration with GeoLytics.

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 - Census Tract Relationship Files. Retrieved April 21, 2016, from https://www.census.gov/geo/maps-data/data/2000tract_rel.html.

U.S. Census Bureau. (2010). Census 2000 Tabulation Block to 2010 Census Tabulation Block Relationship Files. Retrieved November 4, 2014, from https://www.census.gov/geo/maps-data/data/rel_blk_download.html.

U.S. Census Bureau. (2012). 2010 Census Summary File 1—Technical Documentation [PDF]. Washington, DC. Retrieved April 21, 2016, from http://www.census.gov/prod/cen2010/doc/sf1.pdf.

Wilson, S. G., Plane, D. A., Mackun, P. J., Fischetti, T. R., & Goworowska, J. (2012) Patterns of metropolitan and micropolitan population change: 2000 to 2010. Census 2010 Special Report: C2010SR-01. U.S. Census Bureau, Washington, DC.

Wright, J. K. (1936). A method of mapping densities of population: With Cape Cod as an example. Geographical Review, 26(1), 103-110.

Wu, C., & Murray, A. T. (2005). A cokriging method for estimating population density in urban areas. Computers, Environment and Urban Systems, 29(5), 558-579.

Xie, Y. (1995). The overlaid network algorithms for areal interpolation problem. Computers, Environment and Urban Systems, 19(4), 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(3), 199-214.

Zoraghein, H., Leyk, S., Ruther, M., & Buttenfield, B. P. (2016). Exploiting temporal information in parcel data to refine small area population estimates. Computers, Environment and Urban Systems, 58, 19-28.