Public Opinion Quarterly, Vol. 81, Special Issue, 2017, pp. 385–400

WHEN BOUNDARIES COLLIDE
CONSTRUCTING A NATIONAL DATABASE OF DEMOGRAPHIC AND VOTING STATISTICS

BRIAN AMOS*
MICHAEL P. McDONALD
RUSSELL WATKINS

Abstract Scholars have long merged election and census geography to correlate census demographics and election results to infer political behavior (Ogburn and Goltra 1919; Gosnell and Gill 1935; Key 1949). Increasing accessibility of geospatially defined election data provides a valuable tool to understanding voting behavior in the United States at geographic levels unavailable to previous scholars. Here, we describe these data and examine four methods to merge spatial data when precinct and census boundaries are non-conforming: areal weighting, dasymetric mapping, point kriging, and kriging-based areal interpolation. Through a case study of sixteen states and the District of Columbia, we find that dasymetric mapping—a method that uses external data to construct more accurate and realistic weights for areal weighting, in this case the National Land Cover Database—is the best method to estimate demographic characteristics of precincts when census block boundaries do not conform to precinct boundaries. We apply dasymetric mapping to generate a publicly available national database of merged election results and census data for precincts.

Election surveys provide much of what news organizations, campaigns, and scholars know about the electorate's attitudes and behaviors. However, circumstances exist where election surveys cannot provide information about the electorate. Before the advent of modern surveys, scholars correlated aggregate, geographically bound data to infer individual voting behavior

BRIAN AMOS is a PhD candidate in the Political Science Department, University of Florida, Gainesville, FL, USA. MICHAEL P. McDONALD is an associate professor in the Political Science Department, University of Florida, Gainesville, FL, USA. RUSSELL WATKINS is a geographer in the Shimberg Center for Housing Studies, University of Florida, Gainesville, FL, USA. *Address correspondence to Brian Amos, University of Florida, Department of Political Science, 234 Anderson Hall, PO Box 117325, Gainesville, FL 32611, USA; e-mail: bamos@ufl.edu

doi:10.1093/poq/nfx001

(e.g., Ogburn and Goltra 1919; Gosnell and Gill 1935; Key 1949). In modern voting rights litigation, surveys might not exist for an election of interest, particularly for local offices. Here, too, scholars frequently estimate racial voting patterns from aggregate data (Grofman, Handley, and Lublin 2000–2001). These techniques have also been applied in scholarship (e.g., Ansolabehere, Persily, and Stewart 2013; Hirsch and Nall 2016). The method to estimate individual behavior from aggregate data is known as *ecological inference*, a method that has been refined since the initial application of simple correlation (e.g., Goodman 1953; King 1997; Rosen, King, and Tanner 2001).

Any inferential method is laden with assumptions; our purpose here is not to contrast the strengths and weaknesses of ecological inference and survey methods. We wish to describe how the recent increase of open and accessible data enables the creation of databases from which scholars can conduct ecological inference at a heretofore prohibitively costly scale and scope. We describe the collection of the relevant census and election data and evaluate the methods to merge these data together, and use our recommended approach to construct a publicly available national database. Our application merges 2008 election data and 2010 census data; scholars can use the methods we explore in this paper in other analyses of geographically bound data.

The Relationship between Census and Electoral Cartography

Ecological inference concerning demographic patterns of voting using aggregate electoral data requires merging census and electoral cartography. Two government entities create these delineations—the US Census Bureau and local elections offices—for different purposes. The Census Bureau is a national agency that maps the United States' spatial framework to report aggregate population statistics collected through the decennial enumeration and other survey projects, such as the American Community Survey. In these surveys, aggregating individuals' responses within geographic areas serves to protect respondents' confidentiality.

Local election officials across the country create election areas to associate voters with polling locations and to report aggregate election results, which similarly maintains the confidentiality of voters' ballots. Because different bureaucracies with separate needs and intentions create these census and election cartographies, the boundaries do not necessarily conform to one another. However, there are important procedural and policy reasons why their boundaries may be especially congruent in an election preceding a decennial census.

The Census Bureau tabulates, within geographies, aggregate statistics drawn from the decennial population enumeration, such as the racial composition of the total population and the voting-age population. The *census block* is the smallest unit of census geography. Census blocks are roughly analogous

to city blocks in urban areas, and they generally closely follow geographic and man-made features, such as streams and roads. Census blocks tile the entire country and may contain no population, such as those that describe rural expanses, road median segments, or water features. The Census Bureau also reports statistics for their survey products in higher levels of census geography, which are collections of census blocks, in order of ascending size—block groups, tracts, counties, and states. There are other geographies that are not nested within this hierarchy, or do not necessarily tile a state, such as municipal boundaries. Census blocks are not static. The Census Bureau continually makes adjustments throughout the decade between censuses, for example, to create blocks to accommodate new housing development. Furthermore, the Bureau may shift block boundaries over the course of the decade. For example, the 2000 census blocks for St. Croix County, Wisconsin, have the same number of units and are located in roughly the same space when taken from the 2007 or the 2009 geography release, but their boundaries are not identical.[1]

Election officials call the administrative boundaries they use for conducting elections by various names: election districts, wards, and most commonly—the name we employ—*precincts*. The average precinct is about sixty times larger than the average census block, although like census blocks, they vary considerably in size based on population density. Furthermore, precinct boundaries may undergo change each election. Election officials may consolidate two or more precincts into a single precinct, particularly to reduce the number of costly polling places for anticipated low-turnout elections. A precinct may be split into two or more precincts when the number of voters grows beyond the capacity of a precinct's polling place. Precincts may be entirely redrawn, or re-precincted, when precincts are reorganized to fit within new legislative district boundaries. This change may also occur if a polling place is no longer available, causing a shuffling of voters across polling places.

Since precincts are geographically larger than census blocks, many blocks can be associated with a unique precinct. In the 2008 general election, 95.2 percent of census blocks were associated with a unique precinct; of the 4.8 percent that were not, 43 percent had no population. There is an important policy reason why census blocks often align with precincts, too. Legislative district lines are typically drawn from census blocks to conform to federal courts' precise population standards (Levitt and McDonald 2007). Since precincts often adhere to legislative district boundaries, precincts mostly conform to census blocks. Moreover, in the three years prior to the decennial census, the Census Bureau conducts what is known as the Redistricting Data Program, in which the federal government invites their state counterparts to submit their political delineations for inclusion as geographic units in the next

1. See http://www2.census.gov/geo/pdfs/maps-data/data/tiger/tgrshp2009/TGRSHP09C3.pdf, accessed April 3, 2016.

census. Census blocks are updated to conform to the submitted political geography, including the precincts, or what the Census Bureau generically calls Voting Tabulation Districts (VTDs). This collaboration assists states in merging political and census data to forecast potential electoral outcomes during the highly politicized redistricting process. Some states freeze their precincts once transmitted to the Census Bureau, or even throughout the decade following their decennial redistricting. For these states or localities, census blocks often perfectly conform to precincts for at least some elections.

Elsewhere, VTDs may not be entirely accurate representations of the precincts due to ongoing re-precincting that occurs between elections, and even between the transmission of precinct boundaries to the Census Bureau and the next election. Moreover, while nearly all states participate in the Redistricting Data Program, some do not provide VTDs, or the quality of their participation may be less than ideal given time constraints (McCully 2014, 12). A few states may, on their own, conform their precincts to census blocks, such as work performed by California's Statewide Database or Ohio's Apportionment Board. There thus exists variation across states and localities, as well as across time, as to how well census blocks conform to precinct boundaries.

Collecting these data presents a further challenge. Census Bureau data are relatively easy to obtain, as the federal government provides electronic representations of all states' census geography with keys relatable to statistical data. Data dissemination by election officials varies because of the nation's decentralized and localized election administration. Some states centralize their election administration, to a degree such that much information, including precinct boundaries and the associated election results, is readily available in electronic format from state agencies. In others, these data are available only from local election officials, and may not be in electronic formats. Even when electronically available, interested persons may not easily spatially overlay precinct maps on census blocks, when such maps are scanned or are otherwise generated images. There is no standard data schema across states and sometimes across local election offices within the same state. Cross-walking precinct boundaries and election data can be further challenging when unique identifiers are not the same in the boundary and election results files.

There are additional issues that deserve brief mention. First, election officials report statewide elections within precincts, but some localities allow lower office districts to split precincts. Here, a scholar must develop an assignment algorithm to link the election results to the split portions of the precinct; an extended discussion is beyond our scope, since we are interested here in statewide elections. Second, some local election officials report "nonprecinct" votes, such as mail, in-person early, and provisional votes in the voters' home precinct, while others use jurisdiction-wide precincts to report results. A typical approach for the latter reporting method is to apportion jurisdiction-wide votes into precincts, proportional to the candidates' votes within

precincts (McDonald 2014). Some of the methods we describe below may be applied to either case.

Despite these challenges, there exist efforts to collect comprehensive election data. The primary data source for our case study is a 2008 precinct boundary and election results collection effort by the Stanford Election Atlas (Rodden 2014). A scholar interested in analyzing voting patterns for demographic groups must overlay these data onto census geography. In table 1, we describe the congruence of the 2010 census blocks and precincts within the states in the Stanford Election Atlas by reporting the percentage of census blocks that either are split by one or more precincts or are assigned to no precinct. We report statistics for census blocks that contain population, since we wish to aggregate only blocks containing population data. The general pattern is one of concordance: sixteen states and the District of Columbia do not have any blocks split by precincts, while another twenty-seven states do so only rarely. Six states have 1 percent or more census blocks with indeterminate matches with precincts. It is among these states that the method to assign blocks' populations to precincts will potentially have the greatest effect on the construction of precincts' demographic statistics.

We classify the six outlier states in terms of the congruence of election and census geography into three groups: Rhode Island, California, and the remaining four of Florida, Kentucky, Montana, and Wisconsin. For Rhode Island and two other states—Arkansas and Oregon—for data accuracy reasons we conclude that we must abandon attempts to merge precinct-level election data with census geography, and instead look to higher geographic levels for election data, such as counties or townships. For the others, we need to implement an estimation method to construct demographic statistics for precincts.

In the Stanford Election Atlas, our investigation revealed that Rhode Island's precincts are apparently not actual precinct boundaries, but instead that the

Table 1. Percentage of Blocks with Population That Are Split or Unassigned, by State

State(s)	Split or unassigned blocks with population
AZ, CO, DE, DC, ID, IN, IA, KS, MS, MO, NE, NH, NM, ND, OK, UT, WY	0.00%
AL, AK, AR, CT, GA, HI, IL, LA, ME, MD, MS, MI, MN, NV, NJ, NY, NC, OH, PA, SC, SD, TN, TX, VT, VA, WA, WV	< 0.03%
Montana	1.63%
Kentucky	4.08%
Florida	14.55%
Rhode Island	18.47%
California	20.05%
Wisconsin	27.93%

researchers constructed them from polling place addresses using a method known as a Voronoi diagram. The Stanford Election Atlas excluded Oregon completely from their results (as we do from table 1) because Oregon officials did not participate in the Census Bureau's Redistricting Data Program, and local election officials maintain difficult-to-collect paper copies of precinct boundaries. Arkansas presents a different problem: the vote counts for the state's precincts had an element of estimation involved—candidates' vote totals are not whole numbers—and the imprecision ultimately led to the Atlas reporting more votes than the 2010 census reported voting-age residents for 25 percent of the precincts. This is not only an obvious inaccuracy; it also violates an assumption of ecological inference analysis that the total candidate votes must be smaller than the population (King 1997). For these three states, we believe the best available method is to use higher levels of geography, where election boundaries and census boundaries align perfectly and where election results are available. For Arkansas and Oregon this geographic layer is counties, and for Rhode Island it is townships.

For Florida, Kentucky, Montana, and Wisconsin, misalignments between precincts and census blocks appear to be issues with the geospatial representations of the precincts or census blocks. Visual inspection suggests that Montana precincts conform to 2010 census geography, but with minor errors: 98 percent of the split blocks clearly favor one precinct, with 5 percent or less of the block's area split into a second precinct. Florida, Kentucky, and Wisconsin precincts, on the other hand, have clear signs that they are based on 2000 census geography, rather than 2010 census geography. Our examination suggests that the Census Bureau's mid-decade improvements to census geography in consultation with state and local governments greatly affected Florida and Wisconsin. As a result, few precinct lines coincide with the post-mid-decade shift of the census block map, and by extension, the 2010 geography. Kentucky is similarly affected, but the shifting and reshaping process appears less dramatic for that state. California has a different problem: the state simply does not use census geography in defining its precincts. Instead, the California Statewide Database uses internal methods to apportion election results to census geography.[2]

The Polygon Overlay Problem

We need to consider the best approach of constructing census statistics for misaligned precinct and census block boundaries, with the highest priorities being California, Florida, Kentucky, Montana, and Wisconsin. To inform these decisions, we need to understand in detail the polygon overlay problem, which occurs when computing statistics for different geographies when their

2. See http://statewidedatabase.org/.

boundaries do not conform. Interestingly, our use case of the 2008 election presents us with an opportunity to explore the accuracy of the techniques to approach the polygon overlay problem in states where precinct and census block boundaries precisely conform by constructing groups of census blocks that are intentionally misaligned with precincts.

Our goal is to compute statistics at one of the available geographic levels. Since precincts are generally in a hierarchical relationship with smaller census blocks, many of which entirely nest within a single precinct, the larger precincts are the preferable geographic *target unit* that we may wish to compute statistics for, while the smaller census blocks are the *source unit* of demographic data. Some census blocks intersect or, in election administration terminology, "split" two or more precincts. The situation where source units and target units do not perfectly correspond with one another is generally known as the polygon overlay problem (for a review, see Gotway and Young [2002]). There are several proposed approaches to resolve the polygon overlay problem; the ones we examine here are known as areal weighting, dasymetric mapping, point kriging, and kriging-based areal interpolation.

A simple approach to resolving the polygon overlay problem is areal weighting (Sadahiro 2000). A researcher assumes the source unit data are uniformly distributed spatially and apportions the source-unit data to the target units by the proportion of the source units' area that is contained within each target unit. This assumption is incorrect in virtually all cases, but carries the benefit of the computations being relatively easy and fast to calculate.

A scholar can improve upon areal weighting with knowledge of the spatial distribution of the data in the source targets, using a procedure known as dasymetric mapping (McCleary 1969). Recall that we are working with individual-level population data that administrators have intentionally aggregated into geographic areas to protect individuals' confidentiality. We can use other data to estimate where population is geographically located, such as a land-cover database that distinguishes between cities, forests, rural areas, and so on. One such database is readily available: the 2011 National Land Cover Database, or NLCD (Homer et al. 2015), produced by a cooperative effort of a number of federal agencies working under the umbrella of the Multi-Resolution Land Characteristics Consortium. The NLCD assigns land to one of sixteen usage classes at a resolution of thirty meters; a small state like Delaware has over seven million data points in the NLCD.

Another popular method of apportioning geographic data is kriging. Kriging is often used in situations where known data at given points or areas are used to infer, through statistical methods, the distribution of the data throughout the area of interest (Gotway and Young 2007; Krivoruchko, Gribov, and Krause 2011). For example, a meteorologist may map the distribution of rainfall over an area by interpolating rainfall amounts measured at rain gauges, with the assumption that rainfall follows a continuous distribution between gauges.

This method is suited for continuous data provided certain assumptions are met; the paramount assumption is that the data follow a known spatial distribution. Meeting this assumption can be problematic for most socioeconomic data, and especially for population characteristics, as population distributions often do not follow continuous functions (Qiu, Zhang, and Zhou 2012).

Point kriging has been improved upon to better serve data representing areas (Krivoruchko, Gribov, and Krause 2011), and these kriging-based areal interpolation methods have been incorporated into the popular software suite ArcGIS since 2012 through its Geostatistical Analyst extension, making the analysis relatively accessible to the average user. However, even this method can suffer from similar issues as point kriging, namely the assumption that the populations of census blocks near each other will be similar, when in fact racial and ethnic populations tend to highly segregate into communities (Schelling 1971).[3] Still, despite violating the assumptions of kriging, it is illustrative to our use case to demonstrate the accuracy of point and area-based kriging methods in comparison to areal weighting and dasymetric mapping.

Comparing Aggregation Methods: A Case Study

An obvious problem in judging the success of the compilation of data from one polygon type to another is that there is no baseline of true values to compare against; if there were, there would be no polygon overlay problem. We wish to emphasize again, most 2010 census blocks perfectly or near-perfectly conform to the 2008 precincts. Still, there are blocks split by precincts where an estimation method is necessary, and these misalignments can occur systematically across certain states. We thus wish to judge the relative accuracy of the four methods of areal weighting, dasymetric mapping, point kriging, and kriging-based areal interpolation in our use case so we can proceed with the best estimation method for overlaying demographic and voting data.

We examine the accuracy of proposed methods to address the polygon overlap problem by constructing misaligned geography in states where census blocks align perfectly with precincts. In this case, we know the "truth" to which we can compare the relative accuracy of the estimation methods. Our case study is the sixteen states and the District of Columbia where census blocks align perfectly with precincts. We apply the four estimation methods to two geographic levels. The first level is census blocks, where the truth

3. In this use case, census block data are disaggregated by creating a continuous surface of predicted population based on the variation in space of the total population value assigned to each census block. The continuous surface is then reaggregated into precinct polygons. While the results in table 2 appear promising, applying this method to population data is problematic due to violation of the inherent model assumption of data stationarity (i.e., that population smoothly varies in density across a given landscape).

is known, since census blocks perfectly conform to precincts.[4] The second geographic level is census block groups, which are the next highest level of geography the census uses above blocks. Nationwide, there is an average of fifty-one census blocks for every block group. Within our deliberate subsample of states, census blocks nest perfectly into precincts and block groups, but block groups do not perfectly align with precincts. Since we have the true results available to us from the census block level, we can compare the success of four methods against the truth, with the hope that it will provide guidance in which to choose when the truth is unknown. This exercise also applies to interesting demographics supplied by the American Community Survey, like citizenship status or Hispanic ethnicity, which the Census Bureau provides at only the block group or higher geographic levels.

We implement the estimation methods for areal weighting, dasymetric mapping, point kriging, and kriging-based areal interpolation as follows, performing the procedures separately for each state and combining the results afterward.

For areal weighting, we

1. Trim the block group shapefile to remove portions that are not covered by precincts.[5]
2. Calculate the area of each block group.
3. Perform a union between the block group shapefile and the precinct shapefile to create polygons for each overlapping combination of block group and precinct.
4. Calculate the area of each polygon in the union shapefile.
5. Divide each union shapefile area by the total area of its parent block group.[6]
6. Multiply this proportion by the attribute of interest for each block group.
7. Sum the quantities calculated in step 6 by precinct.

4. At the risk of being pedantic, the precincts may not cover the full extent of the legal boundaries of the state as census geography does. For example, in Delaware, Delaware Bay and the several smaller bays in the southeast are not assigned to precincts. However, they do cover the full extent of the *population* of the state, meaning that we can derive the true population figures for each precinct.

5. This is not necessary for every state, but states with ocean or Great Lakes borders often do not assign all water area to precincts; this step prevents people from being removed for being assigned to a portion of a block group without coverage.

6. One potential option not reported here is to stop at this stage, and instead of breaking source units across multiple target units, assign each source unit to exactly one target unit based on the largest overlap. In our case study, this is obviously and extremely wrong; given that there are a similar number of block groups and precincts, some precincts would end up being assigned no block groups, and others would be assigned much more population than could be reasonably expected. However, if the expectation is that the source units *should* nest within target units but erroneously do not in the shapefiles available, it may be worth considering.

To apply dasymetric mapping, we must generate weights for each of the sixteen NLCD land use classifications. Previous studies on the topic have assigned weights that predominantly (Eicher and Brewer 2001) or entirely (Zandbergen and Ignizio 2010) place population in land classified as developed. There are four classifications of developed land in the NLCD: open space, low intensity, medium intensity, and high intensity. We test every possible combination of weights using a grid search from 0 to 1 in 0.05 steps for each land type. We find that the best fit for our study occurs by assigning a weight of 1.00 to low intensity, 0.85 to medium intensity, and 0.05 to open space, while high-intensity development and non-developed land types are all weighted 0. It may seem counterintuitive that we found a zero weight ideal to high-intensity development. Keep in mind that our use case is to correctly apportion population in localized geography split by precincts. In this context, high-intensity development may identify businesses situated along main thoroughfares with residential neighborhoods found on side streets. For dasymetric mapping, we

1. Select weights for each land use type in the dataset, as described above.
2. Follow the same process as areal weighting, but use our weights to carry out the area calculation in steps 2 and 4.

For point kriging, we must create fictitious data. Point kriging assumes that a block group's population data are clustered at a single point. We use the centroid of the block group polygon as this point. For point kriging, we

1. Create a point file from block group centroids.
2. Apply a simple kriging procedure that uses untransformed centroid summary values, a covariance model, and an error term to predict a continuous surface of variable values.
3. Convert the kriged surface to a raster, with predicted values assigned to each cell.[7]
4. Intersect cell values with precinct boundaries to select all cells that fall within the boundaries.
5. Summarize selected cell values by precinct.

Kriging-based areal interpolation is a relatively complicated procedure, and building a system from scratch to carry it out is outside the scope of this paper (for those interested, see Krivoruchko, Gribov, and Krause [2011]). However, ArcMap, part of the common software suite ArcGIS, has a preprogrammed algorithm that aids the process by estimating best values for certain parameters

7. The resolution of this raster is important. Our testing chose a very fine resolution, but still managed not to create a sample point in 1.1 percent of the precincts, mainly in those that were especially small or oddly shaped (i.e., long and thin blocks representing interstate highways). These precincts were modeled as missing data in our tests. They illustrate problems in the method beyond its already poor performance.

and providing goodness-of-fit diagnostics to aid in the choice of other parameters. For kriging-based areal interpolation, we

1. Load the census block group shapefile into the ArcMap Geostatistical Wizard, and generate an Areal Interpolation Layer.[8]
2. Run the layer from step 1 through the Areal Interpolation Layer to Polygon tool, using the precinct shapefile as the target.

Results

Table 2 presents statistics for the true distribution of voting-age population among precincts and the predictions generated by our four methods: areal weighting, dasymetric mapping, point kriging, and kriging-based areal interpolation. We report descriptive statistics of the standard deviation, skewness, and kurtosis of the precinct estimates. We also report goodness-of-fit metrics between the true precinct values and the estimated values: the average absolute error and Pearson's correlation between the true values and the estimates.

As expected, the two kriging methods perform the worst. Point kriging performs the worst of the four methods in that its descriptive statistics are most dissimilar to the truth and it has the worst goodness-of-fit metrics. Kriging-based areal interpolation is a considerable improvement over point kriging: the descriptive statistics are more similar to the truth than point kriging, and the goodness-of-fit metrics suggest improved estimates. Still, there are problems. We are unable to find parameters in the creation of the interpolation layer that fit the model into the ArcGIS-recommended error bounds or that perform well in the diagnostic test statistics. These diagnostic issues are likely a symptom of our data violating the assumptions of the method. Furthermore, ArcGIS generates

Table 2. Summary Statistics of the Precinct Voting-Age Populations Predicted by the Four Methods (average absolute error and Pearson's correlation are in comparison to the true values)

	Std. dev.	Skewness	Kurtosis	Average absolute error	Pearson's correlation
True population	1014.9	3.754	38.324	–	–
Areal weighting	1026.5	3.501	34.167	264.1	0.898
Dasymetric mapping	1009.1	3.656	36.128	146.7	0.971
Point kriging	210.4	1.226	6.548	626.5	0.285
Kriging-based areal interpolation	945.1	3.294	27.339	384.0	0.823

8. We adjust parameters to make the model fit acceptable. ArcGIS documentation suggests adjusting the lag size and count to remove negative covariances and to set the model type to K-Bessel.

measures of uncertainty for the precinct point estimates, and these standard errors are quite large, the majority being even larger than the prediction itself.

The two areal-based methods perform better than the kriging methods. The simple areal weighting method performs only slightly better than the kriging-based areal interpolation in correlation and average error. The addition of ancillary data using the dasymetric mapping method improves the error and fit: the error is nearly half that of areal weighting, and the correlation is the highest of any methods, at 97.1 percent.

A complication of dasymetric mapping is that in a real-world use case, exploratory testing will not be possible to deduce the optimal weights for the different land cover types. To get a sense of the potential variation possible, we move beyond the pooled weight optimization, the statistics for which we present in table 2, and optimize each state separately. We present the ideal weights derived from this process in table 3, along with a comparison of the average absolute error created by using the state-ideal weights versus the pooled weights.

Table 3. Optimal Weights for States When Derived Individually, and a Comparison of the Average Absolute Error Using These State-Ideal Weights versus Pool-Derived Weights with Dasymetric Mapping, as Well as the Average Absolute Error for Area Weighting

	State-ideal weights, developed land				Average absolute error		
	High	Medium	Low	Open	Dasymetric state-ideal	Dasymetric pooled	Areal weighting
Arizona	0.00	1.00	0.40	0.05	174.6	177.7	294.7
Colorado	0.00	1.00	0.90	0.05	169.5	170.0	276.5
Delaware	0.05	0.35	1.00	0.10	175.5	179.9	318.8
District of Columbia	0.00	1.00	0.45	0.00	183.7	193.0	260.6
Idaho	0.00	0.40	1.00	0.05	173.6	174.9	351.3
Indiana	0.00	0.45	1.00	0.10	145.8	148.6	268.9
Iowa	0.00	1.00	0.80	0.00	133.7	139.8	343.4
Kansas	0.00	0.70	1.00	0.00	93.5	96.1	203.2
Mississippi	0.00	0.00	1.00	0.20	165.8	178.6	243.6
Missouri	0.00	0.65	1.00	0.15	129.5	131.3	216.3
Nebraska	0.00	1.00	0.70	0.00	145.1	149.7	301.6
New Hampshire	0.00	1.00	0.40	0.05	70.5	72.5	133.5
New Mexico	0.00	1.00	1.00	0.05	165.3	165.3	256.9
North Dakota	0.00	1.00	0.60	0.00	155.4	158.0	325.8
Oklahoma	0.00	0.00	1.00	0.05	139.0	144.6	239.6
Utah	0.00	1.00	0.55	0.00	136.1	139.2	256.8
Wyoming	0.00	1.00	0.15	0.05	202.2	213.8	463.5

Several patterns stand out in looking at the state-ideal weights. First, high-intensity development is almost uniformly weighted zero, with the exception of Delaware, at a meager 0.05. Second, open space development is generally weighted low, ranging from zero to a maximum of 0.2. The most variation comes from the interaction between the weighting of medium-intensity and low-intensity development. Cases like Mississippi and Oklahoma place minimal value on the predictive power of medium-intensity development, while the District of Columbia and New Hampshire are at the other extreme, placing only moderate value on low-intensity development. This variation may raise concerns about the generalizability of the weights derived from the pooled set, but as the comparison of errors in table 3 shows, the increase in error from using the less optimal weights is minor, peaking at 8 percent for Mississippi. Furthermore, the final column shows the error produced by the next-best method, areal weighting, and using the pooled weights is still a vast improvement over the more naïve method. Indeed, virtually any weighting scheme that weighs either medium- or low-intensity development over high-intensity and open space development outperforms the areal weighting method.

The success in testing, the simplicity in calculation, and the national coverage of the ancillary NLCD dataset lead us to recommend dasymetric mapping as the best estimation method among the four we examined. We observe what appear to be state-specific optimal weights in the states where the truth is known. This variation may be a consequence of Census Bureau policies and procedures, the geographies of the states, or state-specific election administration rules. These state-specific weights may be used with data reported only at the block group level within these states, such as American Community Survey data. However, without a clear pattern of weights from the sixteen states plus DC, we are not confident in applying state-specific weights to other states. We thus recommend and have used the weights we derive from our pooled analysis to construct a national database containing demographic information for voting districts.

Conclusion: A National Database of Demographic and Voting Statistics

The increasing accessibility of geospatially defined election data holds the potential to provide a valuable tool to further understand voting behavior in the United States, at heretofore prohibitively costly geographic levels to collect data. We describe four methods to merge spatial data when boundaries are non-conforming: areal weighting, dasymetric mapping, point kriging, and kriging-based areal interpolation. Our case study of these methods, where we compare the known values to constructed misalignments of boundaries, reveals dasymetric mapping using the National Land Cover Database to be

the best method to estimate demographic characteristics of precincts where census block boundaries do not conform to precinct boundaries.

Following this testing exercise, we have constructed a nationwide database containing merged demographic information and voting statistics (online appendix 1; Amos 2016). We provide 2008 presidential election results (Rodden 2014) and census voting-age population totals by race and Hispanic ethnicity for precincts in all states and DC, with the exception of Arkansas, Oregon, and Rhode Island. We provide county data for Arkansas and Oregon and township data for Rhode Island due to issues with these states' precinct-level data, described above. For all states with precinct-level data, precinct boundaries closely or perfectly align with census block boundaries due to collaboration between election administrators and census cartographers (see table 1). Where precinct boundaries and census blocks do not perfectly correspond, we apply dasymetric mapping using the National Land Cover Database using the pooled weights, described above. The data are freely available in online appendix 1 for researchers who wish to pursue aggregate analysis. In online appendix 2, we provide the data employed in the geospatial testing exercise that led to construction of the national database.

We expect researchers will create similar datasets for future elections. The Census Bureau will again collect precinct boundaries from participating states in preparation for the 2020 census, as was done last decade and which forms the basis of the Stanford Election Atlas. Emerging sources for national precinct boundaries and election results include the Voting Information Project and the Open Elections Project.[9] Many states and localities make these data publicly available in electronic formats. The US Electoral Assistance Commission is working with the National Institute for Standards and Technology to develop data schema standards for adoption by election technology providers.[10] The trends are toward more data standardization, transparency, and interoperability, so we expect that the continued Big Data revolution in the elections sphere will expand opportunities to create merged election and census databases. Exciting possibilities will flow from these data, as researchers will be able to analyze both cross-sectional and longitudinal data by ecological inference techniques, which will provide additional measures of local context for voting behavior in the United States at levels previously prohibitively costly to explore with traditional survey research.

Supplementary Data

Supplementary data are freely available at *Public Opinion Quarterly* online.

9. See the Voting Information Project, https://votinginfoproject.org/, and the Open Elections Project, http://www.openelections.net/.
10. See http://www.nist.gov/itl/vote/public-working-groups.cfm.

References

Amos, Brian. 2016. "Replication Data for: When Boundaries Collide: Constructing a National Database of Demographic and Voting Statistics." *Harvard Dataverse*, doi:10.7910/DVN/1LXHIX.

Ansolabehere, Stephen, Nathaniel Persily, and Charles Stewart III. 2013. "Regional Differences in Racial Polarization in the 2012 Presidential Election: Implications for Constitutionality of Section 5 of the Voting Rights Act." *Harvard Law Review* 126:205–20.

Eicher, Cory L., and Cynthia A. Brewer. 2001. "Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation." *Cartography and Geographic Information Science* 28:125–38.

Goodman, Leo A. 1953. "Ecological Regressions and the Behavior of Individuals." *American Sociological Review* 18:663–64.

Gosnell, Harold F., and Norman N. Gill. 1935. "An Analysis of the 1932 Presidential Vote in Chicago." *American Political Science Review* 29:967–84.

Gotway, Carol A., and Linda J. Young. 2002. "Combining Incompatible Spatial Data." *Journal of the American Statistical Association* 97:632–48.

———. 2007. "A Geostatistical Approach to Linking Geographically Aggregated Data from Different Sources." *Journal of Computational and Graphical Statistics* 16:115–35.

Grofman, Bernard, Lisa Handley, and David Lublin. 2000–2001. "Drawing Effective Minority Districts: A Conceptual Framework and Some Empirical Evidence." *North Carolina Law Review* 79:1383–430.

Hirsch, Eiten, and Clayton Nall. 2016. "The Primacy of Race in the Geography of Income-Based Voting: New Evidence from Public Voting Records." *American Journal of Political Science* 60:289–303.

Homer, Collin, Jon Dewitz, Limin Yang, Suming Jin, Patrick Danielson, George Xian, John Coulston, Nathaniel Herold, James Wickham, and Kevin Megown. 2015. "Completion of the 2011 National Land Cover Database for the Conterminous United States." *Photogrammetric Engineering and Remote Sensing* 81:345–54.

Key, V. O. Jr. 1949. *Southern Politics in State and Nation*. New York: Knopf.

King, Gary. 1997. *A Solution to the Ecological Inference Problem*. Princeton, NJ: Princeton University Press.

Krivoruchko, Konstantin, Alexander Gribov, and Eric Krause. 2011. "Multivariate Areal Interpolation for Continuous and Count Data." *Procedia Environmental Sciences* 3:14–19.

Levitt, Justin, and Michael P. McDonald. 2007. "Taking the 'Re' out of Redistricting: State Constitutional Provisions on Redistricting Timing." *Georgetown Law Review* 95:1247–86.

McCleary, George F. 1969. *The Dasymetric Method in Thematic Cartography*. PhD dissertation, University of Wisconsin.

McCully, Catherine. 2014. *Designing P.L. 94-171 Redistricting Data for the Year 2020 Census: The View from the States*. Washington, DC: US Census Bureau.

McDonald, Michael P. 2014. "Calculating Presidential Vote in Legislative Districts." *State Politics and Policy Quarterly* 14:196–204.

Ogburn, William F., and Inez Goltra. 1919. "How Women Vote." *Political Science Quarterly* 34:413–33.

Rodden, Jonathan. 2014. *"Stanford Election Atlas."* Data provided through private communication.

Rosen, Ori, Wenxin Jiang, Gary King, and Martin A. Tanner. 2001. "Bayesian and Frequentist Inference for Ecological Inference: The R x (C-1) Case." *Statistica Neerlandica* 55:134–56.

Qiu, Fang, Caiyun Zhang, and Yuhong Zhou. 2012. "The Development of an Areal Interpolation ArcGIS Extension and a Comparative Study." *GIScience and Remote Sensing* 49:644–63.

Sadahiro, Yukio. 2000. "Accuracy of Count Data Transferred through the Areal Weighting Interpolation Method." *International Journal of Geographical Information Science* 14:25–50.

Schelling, Thomas C. 1971. "Dynamic Models of Segregation." *Journal of Mathematical Sociology* 1:143–86.

Zandbergen, Paul A., and Drew A. Ignizio. 2010. "Comparison of Dasymetric Mapping Techniques for Small-Area Population Estimates." *Cartography and Geographic Information Science* 37:199–214.