



ORIGINAL ARTICLE

The Accuracy of Identifying Constituencies with Geographic Assignment Within State Legislative Districts

Tyler Steelman¹  and John A. Curiel² 

¹Office of Institutional Research and Assessment, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

²Political Science, Ohio Northern University, Ada, OH, USA

Corresponding author: Tyler Steelman, email: tsteelman@unc.edu

(Received 11 December 2020; revised 08 March 2022; accepted 18 May 2022)

Abstract

Identifying the geographic constituencies of representatives is among the most crucial, yet challenging, aspects of state and local politics research. Regularly changing district lines, incomplete data, and computational obstacles can present barriers to matching individuals to their respective districts. Geocoding residential addresses is the ideal method for matching purposes. However, cost constraints can limit its applicability for many researchers, leading to geographic assignment methods that use polygonal units, such as ZIP codes, to estimate constituency membership. In this study, we quantify the trade-offs between three geographic assignment matching methods – centroid, geographic overlap, and population overlap matching – on the assignment of individual voters to state legislative districts. We confirm that population overlap matching produces the highest accuracy in assigning voters to their state legislative districts when polygonal location data are all that is available. We validate this finding by improving model estimates of lobbying influence through a replication analysis of Bishop and Dudley (2017), “The Role of Constituency, Party, and Industry in Pennsylvania’s Act 13,” *State Politics and Policy Quarterly* 17 (2): 154–79. Our replication suggests that distinguishing between out-of-district and in-district donations reveals a greater impact for in-district lobbying efforts. We make evident that population overlap assignment can confidently be used to identify constituencies when precise location data is not available.

Keywords: GIS/Spatial Analysis; Lobbying; Redistricting; Roll Call Voting

Introduction

The institutional design of representation in the US necessitates matching individuals to their respective legislative districts amidst an array of geographic boundaries that vary in both size and shape. Understanding a representative’s geographic constituency is central to American state and local politics research (Fenno 1978), from progressive ambition (Rohde 1979) to policy responsiveness and lobbying influence

(Caughey and Warshaw 2018; Lewis 2013). However, the process of matching individuals to their representatives is a major constraint within the study of state and local politics; one that is further complicated by regularly changing district boundaries and the lack of perfectly nested sub-geographies inside politically relevant boundaries like legislative districts. The challenges arising from the need to match individual data points (like voters) to geographic units (like legislative districts) are experienced by numerous groups including both election administrators¹ and researchers.²

Although there have been significant advances in geocoding and database management to improve data sources like voter registration files (Amos and McDonald 2020; Ghitzza and Gelman 2020), modern techniques can still be out of reach. If a researcher has access to precise location data like the full address of individuals – which is often not the case – the costs of locating those addresses inside a geographic unit can be prohibitively expensive in regard to time and money. Even with appropriately powerful software and hardware, large-scale geocoding of voter files, as employed by Amos and McDonald (2020), can take dozens of hours to complete.³ Pay-as-you-go geocoding tools like Google API can be financially demanding when the \$5.00/1,000 addresses rate is applied to state voter files with tens of millions of registrants.⁴ Although universities can help bridge the resource gap needed to conduct large-scale spatial audits, less-resourced individuals may find the steps outlined by researchers like Amos and McDonald (2020) inaccessible. Fortunately, these costs can be overcome through geographic assignment matching – the process of assigning individuals to a higher level geography based on their inclusion in a nested lower level geography. Crucially for state politics scholars making use of legislative districts, there is no single geography that can be used to match individuals to districts. During the 2011 redistricting cycle, state lower and upper chambers split approximately 48% and 32% of the smallest unit of publicly known geographic units within the US – ZIP codes – respectively.⁵ Failing to account for these geographic nuances can lead to error, yet the costs involved in addressing these sources of error can be intimidating. To date, the trade-offs to geographic assignment matching over more computationally intensive methods are presently unclear.

In this study, we first quantify these trade-offs by testing three geographic assignment matching methods – centroid, geographic overlap, and population overlap matching – on the assignment of individual voters to their state legislative districts. In doing so, we confirm that population overlap matching produces the most accuracy in assigning voters to their legal state legislative districts when

¹In November 2017, some residents of Virginia's 94th state house district were inadvertently assigned to a neighboring district, and subsequently given the wrong ballot which were subsequently thrown out. The number of misassigned voters exceeded the margin of their legal state house race and could have changed the partisan control of the chamber.

²Applicable research includes attempts to impute individual level race data (Imai and Khanna 2016), estimate exposure on a geographic unit of interest (Marigalt 2011; Naman and Gibson 2015), or study the responsiveness of a politician to their donors (Gimpel, Lee, and Pearson-Merkowitz 2008).

³Amos and McDonald (2020) cataloged a duration of 5.5 hours to geocode the Florida.

⁴See Geocoding API Usage and Billing. Google Maps Platform. <https://developers.google.com/maps/documentation/geocoding/usage-and-billing> (accessed June 1, 2020).

⁵Estimated from Missouri Census Data Center (2018).

polygonal location data (e.g., ZIP codes) are all that is available.⁶ Additionally, so long as the effective number of districts within a lower level polygonal unit is under 1.3, population overlap matching can be used to locate individuals with confidence. We illustrate the applicability of geographic assignment in improving research by replicating and extending the effect of lobbying on legislator behavior by Bishop and Dudley (2017). We show that the burden associated with geocoding individual cases can be significantly reduced by first identifying which lower level geographies are not split between multiple higher levels (like state legislative districts). By better distinguishing between these areas, we discover that an average of 3% of the data is in question and requires use of a geographic assignment method for allocation. Due to the nature of the data, all three methods of controlling for district residency improve model estimates. Following these results, we conclude this study with a set of suggestions by which users can determine whether and how to implement geographic assignment methods within the US context.

The problem

Identifying the geographic constituency of an individual is a multi-stage problem. First, is there an existing data source that matches individuals to their legal representatives? Second, when such data is not present, do the necessary district boundaries in the form of geographic shapefiles or individual coordinates for constituents exist? Third, if coordinate data for individuals is not available, are the addresses in a format conducive to geocoding? And, finally, does the researcher have the means to pay the cost – in both time and money – to perform these computations?

Regarding these problems, the first issue tends to afflict any data that is not a state voter file or a proprietary equivalent. While there is much that can be done with voter files, as demonstrated by Ghitza and Gelman (2020) in improving upon multilevel regression with post-stratification, most research cannot use voter files – even when they are publicly available.⁷ Furthermore, the second issue in identifying the geographic constituency of an individual lies in the quality (and existence) of residential address data or legislative district shapefiles for the chamber(s) being studied.⁸

Scholars addressing issue three – matching residential addresses to coordinates – have made many advances. Amos and McDonald (2020) demonstrate the process by which to employ ESRI and Google geocoders, which engage in fuzzy string matching, to geocode millions of addresses from state voter files. The process of hierarchical geocoding devised by Amos and McDonald (2020) improves upon the strengths of each geocoder while mitigating their shortcomings (Swift, Goldberg, and Wilson 2008) to the point of even identifying thousands of errors in misassigned voters

⁶In this study, we use the term ZIP code to describe the geographic unit of a ZIP Code Tabulation Area (ZCTA). ZIP codes are mail routes created by the U.S. Postal Service for efficient mail delivery. ZCTAs are a geographic approximation of those routes maintained by the U.S. Census Bureau. Our calculations use ZCTAs.

⁷States vary in regard to the accessibility of voter files. Although supposed to be free, Wisconsin, for example, charges \$12,000, even when for research. Maine only makes voter files available to Maine residents or political action committees.

⁸State legislative districts date back to the 1990s from the US Census. The 2000s see better coverage, though see some gaps whenever a state redistricts mid-decade. Post-2010 data see legislative district boundaries as not as much of an issue. Local boundaries, such as electoral wards, varies in availability.

within official Colorado and Florida state voter files.⁹ Therefore, their process demonstrates that it is possible to locate constituents when address and boundary data are present.

The final issue is related to cost. Even when precise coordinate data is available, the issue of cost – in both time and money – often prohibits research from approaching the standards set by researchers like Amos and McDonald (2020). As Goplerud (2015) notes, addressing these issues related to geographic assignment and matching individuals to different levels of geography is consistently expensive in regard to programming skills or the purchase of proprietary software. For example, the hierarchical geocoding process employed by Amos and McDonald (2020) requires access to ESRI proprietary software, ArcGIS, and required 5.5 hours for the Florida voter file alone. If accuracy and non-missingness are a concern, the set of backup checks necessary with a suite of several geolocators can significantly increase time costs.¹⁰ Furthermore, those interested in relying upon proprietary software, such as the Google API, will spend upward of tens of thousands of dollars to geocode a state as populous as Florida. Shepherd et al. (2021), in their recent work analyzing polling place access in North Carolina, relied on a service that provides unlimited geocoding, but at the cost of a \$1,000 per month subscription.¹¹ Absent the resources of a larger research university – or one of the 175 universities with the infrastructure necessary to run graduate programs in geographic information systems (GIS)¹² – these costs can be prohibitively expensive. Therefore, the question naturally arises, are there any methodological shortcuts that might decrease the burden of relying upon hierarchical geocoding that does not sacrifice the quality of the research?

Within the US context, it is possible to avoid both geocoding and more advanced geographic assignment methods in cases where a small enough unit of geography can be identified that is fully nested within a larger geography. ZIP codes provide the smallest unit of publicly known geography and, in many cases, can fulfill this purpose. Figure 1 graphically shows the percentage of a state's population that lives in effectively wholly nested ZIP codes within in relation to state legislative districts. There are apparent differences between chambers, with the results ranging from a low of 5.6% in Rhode Island's state house to a high of 92% in Vermont's state senate. We additionally see that heavily populated states, such as California, Florida, and Michigan, have populations where well over 50% of the state's residents live in ZIP codes nested fully within both the lower and upper state legislative districts. Nationwide, 43% of the population lives in ZIP codes fully nested within state house districts and 61% in state senate districts. It is therefore possible to reduce the need and burden associated with geocoding and more complex methods of geographic assignment, though some of either method will still be necessary when locating individuals that live in non-nested ZIP codes.

⁹Their identification strategy returns to the issue of error in “correctly” geocoded voters.

¹⁰With a computer with 32 GB of RAM, it took approximately 137 hours to code several snapshots of the North Carolina voter file and its approximately 4.3 million unique addresses. The geocoding made use of ESRI's USA point address locator, street address locator, street name centroid locator, and five-digit ZIP code locator. Of these, 8.1% relied upon ZIP code centroids, which we will go into later in this study.

¹¹“Straightforward, Affordable Pricing.” Geocodio. <https://www.geocodio.io/pricing/> (accessed September 1, 2020).

¹²See AAG Guide to Geography Programs in the Americas. <https://www.arcgis.com/apps/webappviewer/index.html?id=2f115c9f7ff74723a07aacb6e266b2af> (accessed September 25, 2020).

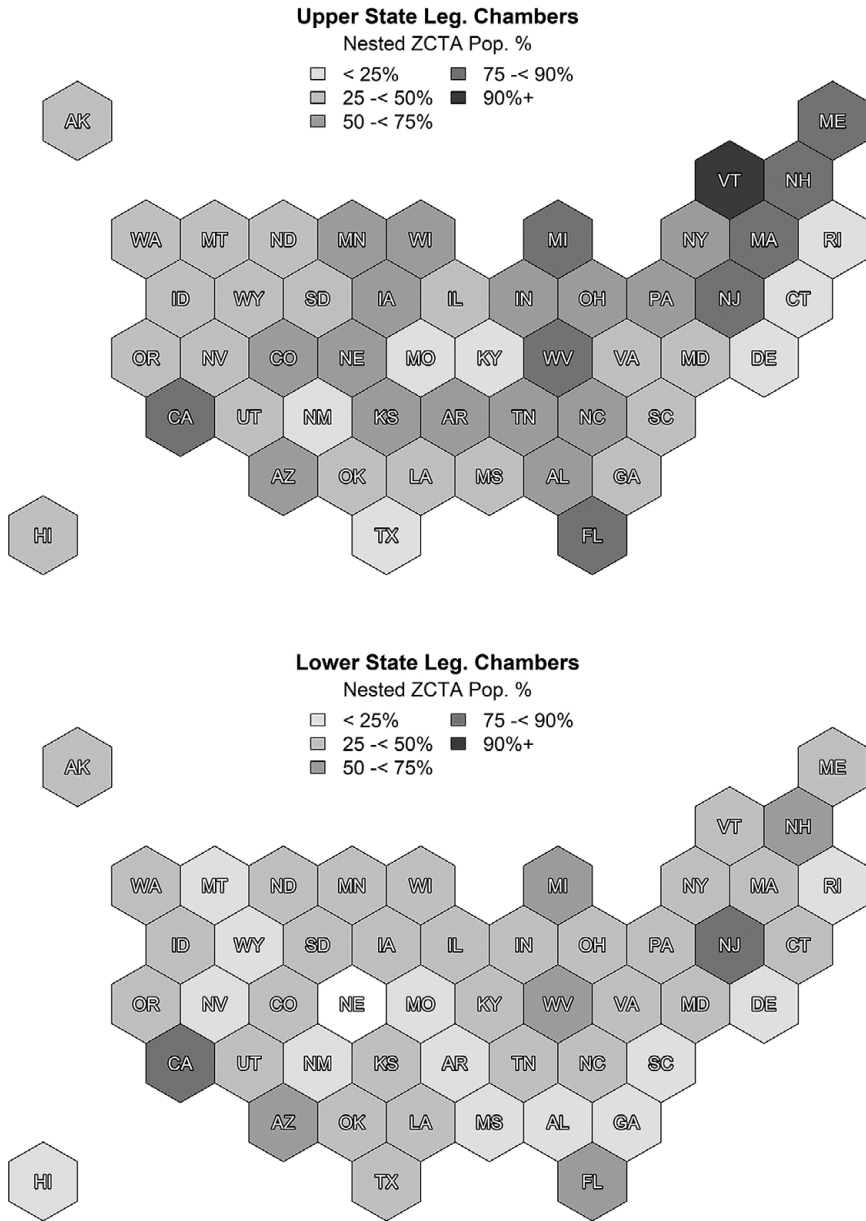


Figure 1. Percentage of state's population living within fully nested ZIP codes, by state legislative chamber.

Assignment of lower levels of geography to higher levels when the two are not perfectly nested has historically used one of three geographic assignment methods: centroid matching, geographic overlap, and population overlap. The centroid matching technique assigns an observation from a lower-level geography when its geographic center, or centroid, falls within the boundaries of a higher-level geographic

unit. Geographic overlap matching assigns or weights a lower-level geography according to the shared area between it and its higher-level overlapping units. Population overlap matching assigns or weights a lower-level geography by using a third level of atomic tabulation units to estimate the population distribution for the overlap between the lower-level and higher-level geographic units.¹³ While each of these methods is prone to some assignment error, Amos, McDonald, and Watkins (2017) determine the assignment accuracy of these three techniques is highest for population overlap and lowest for centroid matching. However, their analysis was conducted in aggregate. It did not distinguish between where these methods were most useful and when researchers could expect their results to be biased depending on the geographies employed. Some lower-level geographic units are heavily split between several higher-level geographic units, other lower-level units are wholly nested inside a higher-level unit. Knowing where and when these matching methods can accurately assign individuals without geocoding is necessary for its application to be used with confidence.

Figure 2 illustrates how these geographic matching methods are computed and the challenges that arise from their use. Pictured is the 27713 ZIP code in Durham, North Carolina. For individual data points originating from this ZIP code, there are three overlapping legislative districts.¹⁴ For a ZIP code like 27713, assigning a voter to a legislative district using only this identifier is challenging. Using the centroid method, a researcher would place all individual data points from this ZIP code in the fourth legislative district (as denoted by the star in the center of the figure). Researchers using geographic overlap would also assign this ZIP code's data points to the fourth legislative district given the approximately 40% of geographic space that is shared between the ZIP code and the legislative district. Researchers using population overlap, though, would assign this ZIP code to the first legislative district because the majority of the ZIP code's population resides to the north. This divergence plagues research attempting to allocate individual data points to one geography based on the point's membership in a smaller level of geography. This study quantifies the trade-offs when using centroid matching, geographic overlap matching, or population overlap matching in these situations and confirms that population overlap matching is consistently more accurate than alternative geographic matching methods.

Validation

We probe the accuracy of geographic matching techniques against the validated and audited geocoded voter file used by Amos (2019). Their data provide the correct and known district residency for each voter within their voter file data (Amos and McDonald 2020). Using their data on correct legislative districts, we predict the correct assignment of voters to their (upper and lower) state legislative districts using only their ZIP codes in Colorado, Florida, Louisiana, New York, North Carolina, and Ohio. ZIP codes are the smallest publicly known geographies within state voter files.

¹³For more, see Amos, McDonald, and Watkins (2017), Duque, Laniado, and Polo (2018), Eicher and Brewer (2001), and Rao (2003).

¹⁴The main focus of this paper is assignment of individual data points to state legislative districts using ZIP codes. For illustrative purposes, this figure uses congressional districts.

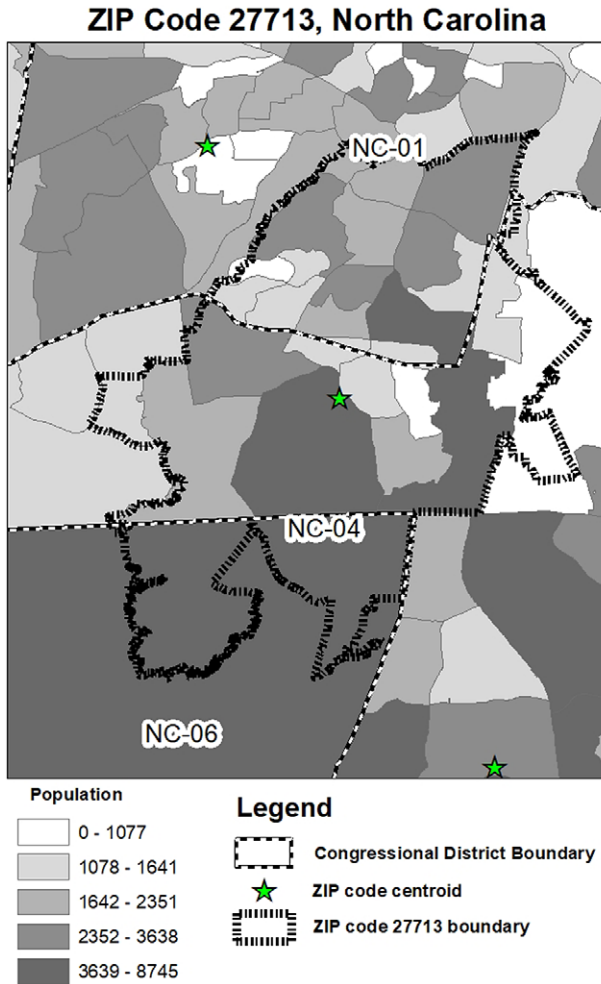


Figure 2. Example of difficulties matching ZIP codes to legislative districts, NC 27713.

Therefore, their use does not require geocoding for matching purposes and can significantly reduce the costs associated with geographic matching.¹⁵

The dichotomous dependent variable for this validation captures whether an individual within a voter file is assigned to the correct district (1) or not (0) for each matching method. To match each voter to their legal state legislative districts using centroid matching, we employed the ArcGIS feature-to-point tool to calculate the geographic center of each ZIP code (constrained to fit within the boundary of the ZIP code) and then overlaid these onto state legislative maps.¹⁶ Geographic overlap

¹⁵Curiel and Steelman (2018) note that the population distribution of ZIP codes is on par with Census tracts, with a median population of approximately 3,000 people.

¹⁶This took under 1 minute to complete for all states using 16 GB of memory.

matching was accomplished using Missouri Census Data Center (2018), which produced dyads of each ZIP code/legislative district pairing as well as the degree of geographic overlap ranging from 0 to 1.¹⁷ For the dichotomous assignment, we assign each ZIP code to the legislative district that it shares the most geographic overlap with.¹⁸ Likewise, we make use of Missouri Census Data Center (2018) to produce dyads of each ZIP code/legislative district pairing as well as the degree of population overlap ranging from 0 to 1, with Census blocks as the atomic tabulation unit. We then follow the same dichotomous assignment of ZIP codes to state legislative districts as used in geographic overlap matching. This resulted in three models, one for each matching method.

We identify the context of where a matching method is most appropriate through a continuous measure of the degree of nestedness between ZIP codes and legislative districts. To do so, we employ the recommended measure as posited by Curiel and Steelman (2020) – the Herfindahl index. It is calculated on a 0 to 1 scale by taking the sum of squared proportions for all of the ZIP code-district dyadic population overlap scores to the ZIP code level. When a ZIP code is fully nested inside a legislative district, its Herfindahl index score is 1. As the effective number of districts inside a single ZIP code reaches infinity, the score approaches 0.¹⁹ The scores are calculated from the GeoCorr output and represent every ZIP code's overlap for the state house and state senate district maps.²⁰

Of the three methods, population overlap performs best in aggregate, ranging from 80% to 90% accuracy in predicted legislative district membership for the six states – in line with expectations from Amos, McDonald, and Watkins (2017). As evident in Table 1, geographic overlap performs on par or slightly better than centroid matching – both of which are less accurate than population overlap in assigning voters based solely on a ZIP code. Increased accuracy when using population overlap matching varies across states from a minimum of a single percentage point in Ohio to eight percentage points in Colorado. These results fall short of what is necessary for a full spatial audit consistent with the recommendations from Amos and McDonald (2020) for election administration. However, these findings support the notion that geographic assignment methods are generally helpful for research applications.

Table 1. Accuracy of geographic matching methods in six US states by matching method

State	Population overlap	Centroid	Geographic overlap
CO	0.83	0.75	0.75
FL	0.90	0.88	0.88
LA	0.80	0.74	0.75
NC	0.81	0.77	0.77
NY	0.84	0.81	0.82
OH	0.88	0.87	0.87

¹⁷This method, which uses a web service, took approximately 10 minutes for all states.

¹⁸It is possible to weight a given observation instead. However, for comparison to prior work (i.e., Winburn and Wagner 2010), we are employing simple dichotomous assignment.

¹⁹The inverse of the Herfindahl index provides the effective number of districts with a ZIP code.

²⁰Data can be found on the SPPQ dataverse repository (Steelman and Curiel 2022).

In order to determine *where* these methods differ in accuracy, we predict the probability of correct assignment given the degree of nestedness between a ZIP code and its overlapping legislative districts. We conducted the analysis stratified by state, with the results not substantively different by state. As an example, we present the predicted probability plot for North Carolina in Figure 2. The *x*-axis presents the Herfindahl index to measure the degree of nestedness, with (1) equating to a ZIP code wholly within a legislative district and (0) representing a ZIP code that is infinitely split.

Looking at the left panel of Figure 3, we see that with a Herfindahl score of 0.90, the probability of correct matching exceeds 95% for all three methods. A score of 0.95 on the Herfindahl index corresponds to effectively 100% accuracy in assignment. Insofar as the accuracy starts to dip below 90% accuracy, it will occur for Herfindahl scores around the 0.75 to 0.79 range. Such a score is equivalent to an effective number of districts within a ZIP code being approximately 1.3. It is around this range that we also start to see the differences in the accuracy of each matching method diverge.

The right panel of Figure 3 plots the difference between the accuracy of population overlap matching compared to the centroid and geographic overlap methods. We see that population overlap reaches its maximum advantage over geographic overlap at a Herfindahl score of 0.41 by approximately 10% points. However, at such a score, the population overlap method is only accurate in 46% of cases. Compared to the centroid approach, population overlap performs its best at a Herfindahl score of 0.47, increasing 8.5% points in accuracy. At such a score, population overlap is estimated to have approximately 56% accuracy in assignment. It is also important to note that Herfindahl index values in this range represent a large portion of all ZIP code-legislative district pairs – as made evident from the density plot at the bottom of the panel. Therefore, while substantive disparities in matching methods arise,

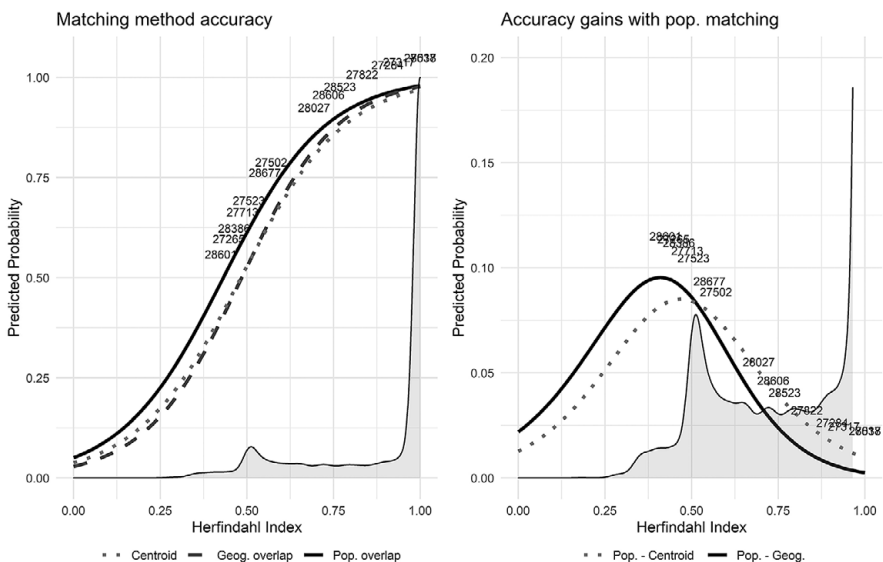


Figure 3. Predicted probability of correct matching.

researchers should hierarchically geocode their data if possible and utilize population overlap matching when hierarchical geocoding is not possible.

The light gray shaded area reflects the distribution of ZIP code nestedness within the North Carolina data. The right panel limits the analysis to ZIP codes with a Herfindahl index under 0.95 for the purpose of focusing on the changes in accuracy across methods. ZIP codes hovering above the solid lines represent a sampling of ZIP codes representative of ZIP code-district nestedness.

Application

By utilizing geographic matching methods to distinguish between constituent and non-constituent influence in lobbying, we can apply the various techniques using a real-world situation. Bishop and Dudley (2017) research the influence of lobbying relative to constituency interests among Pennsylvania state legislators voting for pro-fracking legislation. Their case study selection allows for a critical test of matching techniques that minimizes the impact of endogeneity that typically accompanies research on lobbying and policy outcomes. Bishop and Dudley (2017) tackle the challenges posed by Ansolabehere, de Figueiredo, and Snyder Jr. (2003) in identifying a causal relationship between lobbying and policy outcomes head on, choosing a case where a lobby formed within a few years, effectively precluding a mistake in the causal direction. Additionally, the authors employ geocoded gas-well data to measure constituency reliance upon the natural gas industry.

In a roll call model of voting, negative values of the dependent variable reflect a more conservative voting record in favor of the natural gas industry. The explanatory variables of interest are, first, natural gas production within a legislator's district, and, second, donations from the natural gas industry's political action committee and associated individuals. In their original analysis, the authors find lobbying, in general, exerts a significant, albeit modest, impact on legislator voting.

The only measurement shortcoming of Bishop and Dudley (2017) can be captured in not having distinguished between donations that arise from in- or out-of-district sources. As theorized by Kingdon (1977), representatives attempt to minimize the tension between their stakeholders, ideally never choosing between influential lobbyists and their constituents, hence their preference for committees relevant to their district. For example, Kalla and Broockman (2016) find that the combination of being both a constituent and donor leads one to be more likely to secure meetings with representatives in their randomized field experiments. If one could separately estimate lobbying by in-district versus out-of-district sources, it would be possible to ascertain how much power and influence lobbyists had relative to a legislator's own voters. If all lobbying arose from in-district, it would suggest that Pennsylvania representatives were acting within their constituency's interests. If out-of-district funds still retain an effect, that would suggest a degree of power more associated with fears of corruption and responsiveness to corporations raised by critics of the expansion of natural gas industry goals.

The benefit of population overlap to this analysis is the ability to better distinguish constituency versus non-constituency interests measured using the ZIP codes reported within the contribution data. Pennsylvania is not the ideal state for wholly assigning individuals dichotomously to a district using only a ZIP code. Fortunately, population overlap analysis allows us to weight donations as in-district based upon

the proportion of a ZIP code's population that is located within a legislative district to calculate the respective logged donations. Such coding allows us to estimate the impact of out-of-district lobbying. With these new data, we re-estimate the first model presented by Bishop and Dudley (2017) in their Table 3 (169), predicting legislator voting scores as measured by the Pennsylvania League of Conservation Voters (PLCV).

When analyzing the donation data, we find that approximately 97.2% of the donors to members of the state house lived in ZIP codes completely outside the state house member's district. The figure is 90.6% when looking at donors to members of the state senate. Approximately 0.4% of donors to members of the state house were completely nested within one district, and 7.5% of the state senate. This results in 97.6% of the donations made to members of the state house and 98.1% of the donations made to members of the state senate bypass the need for geocoding. In fact, only 4.7% of the donations made to members of the state house and 2.0% of the donations made to members of the state senate require the use of a geographic matching method to allocate them as in- or out-of-district. As a result, this suggests that model differences comparing geographic matching methods will be minimal.

Table 2 estimates separate models for the effect of in-district and out-of-district donations using each matching method for comparison. Furthermore, for population overlap and geographic overlap matching, we include separate models where donations are assigned wholesale to the legislative district with the greatest overlap with the donation's ZIP code and where assignment is weighted on the shared proportion of overlap between a ZIP code and its greatest overlapping legislative district.

Perhaps most telling about the usefulness of matching methods is the difference between coefficients for the effect of donations in the original model (column 1) to

Table 2. Comparisons in predicting PLCV scores

	Dependent variable					
	Original	Pop. weighted	Pop. plural	Geog. weighted	Geog. plural	Centroid
	(1)	(2)	(3)	(4)	(5)	(6)
Democrat	63.479*** (3.753)	63.367*** (3.747)	63.302*** (3.738)	63.416** (3.742)	63.270*** (3.726)	63.497*** (3.743)
NPAT score (ideology)	-4.801 (3.071)	-4.818 (3.066)	-4.947 (3.059)	-4.742 (3.062)	-4.895 (3.048)	-4.776 (3.063)
Senate	-0.406*** (0.153)	-0.341** (0.16)	-0.337** (0.158)	-0.326** (0.161)	-0.310* (0.158)	-0.338** (0.160)
Log dist. gas prod.	-9.062*** (2.187)	-8.826 (2.190)	-8.820*** (2.182)	-8.863*** (2.184)	-8.684*** (2.178)	-8.818*** (2.187)
Log industry donations	-0.613** (0.269)					
Logged out of district industry donations		-0.551** (0.273)	-0.540** (0.271)	-0.544** (0.272)	-0.533** (0.270)	-0.553** (0.272)
Logged in district industry donations		-1.188** (0.504)	-1.417*** (0.530)	-1.268** (0.502)	-1.599*** (0.530)	-1.273** (0.519)
Constant	31.057*** (2.389)	30.929*** (2.387)	30.921*** (2.380)	30.894*** (2.384)	30.918*** (2.372)	30.857*** (2.387)
Observations	248	248	248	248	248	248
R ²	0.877	0.878	0.878	0.878	0.879	0.878
Adjusted R ²	0.874	0.875	0.875	0.875	0.876	0.875

Note: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.1$.

Table 3. Cost/benefit analysis of geographic matching methods

Method	Accuracy	Allocation method	Replication impact	Costs	Program
Centroid	Least	Dichotomous	Moderate	Moderate	ArcGIS
Geographic overlap	Moderate	Dichotomous or weighted	Moderate	Low	GeoCorr
Population overlap	High	Dichotomous or weighted	Moderate	Low	GeoCorr

the coefficients of in-district and out-of-district contributions in each of the subsequent models. It is clear that the original impact of industry donations are somewhat muted relative to estimates obtained when the source of donations are distinguished from one another. For example, when examining the impact of in- and out-of-district donations using population overlap matching, it is clear that in-district donations are having an outsized impact relative to donations coming from outside the legislator's district. The coefficients for in-district donations tend to be around double that of out-of-district donations. This pattern is found regardless of matching method and without respect to how donations are aggregated. The lack of significant differences between estimations is not surprising, given the aforementioned lack of donor ZIP codes split between districts. Therefore, these results suggest that while the nature of the data makes it less meaningful *how* the user assigns data, any type of control for donation source can add more nuance as to the differential impact of lobbying by source. Regardless, their original conclusion not only holds up, but is strengthened by utilizing geographic matching to distinguish the source of contributions.

Discussion

As demonstrated, there are situations where geocoding might be necessary to ascertain who represents an individual. However, in the context of American state and local politics, the use of geographic assignment matching methods can reduce the geocoding burden by at least a quarter in every state. Even when an area is split between multiple districts it is possible to confidently assign individuals to legislative districts by gauging the degree of lower-level geography nestedness to make appropriate decisions about the use of geographic assignment methods. These tools can be of use to scholars – and reviewers – in making the call of whether more rigorous methodologies must be employed when assigning voters to legislative districts. Furthermore, it is possible to review previous research, such as Gimpel, Lee, and Pearson-Merkowitz (2008), and determine methodological soundness given matching method and geographic context. In the case of Gimpel, Lee, and Pearson-Merkowitz (2008), we estimate approximately 75% of the nation's population to reside within ZIP codes fully nested within congressional districts during the 2000s, with the greatest inaccuracies arising within Maryland, Nevada, Florida, New York, and North Carolina.²¹

Following our results, we present a cost/benefit analysis of each geographic matching method in Table 3. We organize the results by method, accuracy, allocation method, impact on replication, and the program used to conduct each.

²¹The data acquired to estimate these are from Curiel and Steelman (2018).

We ultimately find that in regard to accuracy, population overlap ranks highest, followed by geographic overlap and finally centroid. The impact of accuracy becomes most clear when a lower level geographic unit is split between two effective higher level units. Allocating individual data points using geographic matching proves to be most flexible when using population or geographic overlap, since each provides a continuous 0–1 score to allocate fractions of a value or assign an entire unit of geography based on the greatest degree of overlap. Costs were highest in terms of compromised accuracy and financial burden for the centroid method as it required access to the centroid geographic bounding tool through ArcGIS. This is compared to a free download of the program GeoCorr which can be used to facilitate geographic and population overlap.²²

From Table 3, it is apparent that population overlap weakly dominates the other two methods, and centroid assignment is weakly dominated by both population and geographic overlap. In order to aid in future research, we suggest the following rules by which to implement geographic matching assignment.

First, are both levels of geography available from the US Census? If so, then it is possible to employ GeoCorr. If either level of data is not from the US Census, employ a package that can read in raw shapefiles and use geographic operations to find the population or geographic overlap. These might consist of an R CRAN available package by Goplerud (2015) or the recently developed *arealOverlapr* package (Curiel 2022).

Second, to what extent does the higher level of geography split the lower level? This should be determined by finding the Herfindahl index/effective number of higher geographic units nested within the lower level. To avoid the need of discarding data with accuracy under 90% as practiced by Enos (2015), it is recommended to weight data by dyadic overlap should the Herfindahl index fall below 0.75.

Finally, should the researcher feel uncomfortable with partial weighted geographic assignment and they prefer to geocode individual observations, it is recommended that the researcher lessen the burden of geocoding. By identifying those lower level geographic units completely nested within the higher level, researchers can subset the data they must geocode to only those observations that are not fully nested within the higher level geography being employed. This procedure can save researchers hours of time and potentially thousands of dollars. As illustrated in Figure 1 and the replication of Bishop and Dudley (2017), it might be the case that only few observations even need to be partially weighted or geocoded after properly identifying the data to be geocoded.

Although there will always be uncertainty in geographic assignment where different geographies do not nest within each other, we have improved the confidence that one can have when researching such matters and utilizing such matching techniques. The improvements in population overlap analysis highlight the usefulness of more recent advances in GIS capability and ease of access to individuals. We assert that population overlap analysis offers a valuable tool to anyone pursuing research questions involving the geographic assignment of inconsistently nested geographies.

²²While free programs like R have the ability to find centroids, this approach might not be geographically bounded as they are in ArcGIS.

We conclude by noting that even though the field of state and local politics is highly variable in regard to the quality of data available, it is possible to overcome the challenges of identifying geographic constituencies scientifically. Moreover, more nuanced identification of these constituencies via geographic assignment and weighting can in turn improve and expand our understanding of state and local politics.

Data Availability Statement. Replication materials are available on SPPQ Dataverse at <https://doi.org/10.15139/S3/WIN7SM> (Curiel and Steelman 2022).

Funding Statement. The authors received no financial support for the research, authorship, and/or publication of this article.

Conflict of Interest. The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

- Amos, Brian. 2019. "Replication Data for: A Method to Audit the Assignment of Registered Voters to Districts and Precincts." Harvard Dataverse, V1. <https://doi.org/10.7910/DVN/Y18MK5>.
- Amos, Brian, and Michael P. McDonald. 2020. "A Method to Audit the Assignment of Registered Voters to Districts and Precincts." *Political Analysis* 28 (3): 356–71.
- Amos, Brian, Michael P. McDonald, and Russell Watkins. 2017. "When Boundaries Collide: Constructing a National Database of Demographic and Voting Statistics." *Public Opinion Quarterly* 81: 385–400.
- Ansolabehere, Stephen, John M. de Figueiredo, and James M. Snyder Jr. 2003. "Why Is There so Little Money in Congress?" *Journal of Economic Perspectives* 17 (1): 105–30.
- Bishop, Bradford H., and Mark R. Dudley. 2017. "The Role of Constituency, Party, and Industry in Pennsylvania's Act 13." *State Politics and Policy Quarterly* 17 (2): 154–79.
- Caughy, Devin, and Christopher Warshaw. 2018. "Policy Preferences and Policy Change: Dynamic Responsiveness in the American States, 1936–2014." *American Political Science Review* 112 (2): 249–66.
- Curiel, John A. 2022. "arealOverlapR." <https://github.com/jcuriel-unc/arealOverlapR>.
- Curiel, John A., and Tyler Steelman. 2018. "Redistricting Out Representation: Democratic Harms in Splitting Zip Codes." *Election Law Journal* 17 (4): 328–53.
- Curiel, John A., and Tyler Steelman. 2020. "A Response to "Tests for Unconstitutional Partisan Gerrymandering in a Post-Gill World" in a Post-Rucho World." *Election Law Journal* 19 (1): 101–9.
- Duque, Juan C., Henry Laniado, and Adriano Polo. 2018. "S-maup: Statistical Test to Measure the Sensitivity to the Modifiable Areal Unit Problem." *PLoS One* 13 (11): 1–25.
- Eicher, Cory L., and Cynthia A. Brewer. 2001. "Dasymeric Mapping and Areal Interpolation: Implementation and Evaluation." *Cartography and Geographic Information Science* 28: 125–38.
- Enos, Ryan. 2015. "What the Demolition of Public Housing Teaches Us about the Impact of Racial Threat on Political Behavior." *American Journal of Political Science* 60 (1): 123–42.
- Fenno, Richard. 1978. *Home Style: House Members in Their District*. Boston, MA: Little, Brown and Company.
- Ghitza, Yair, and Andrew Gelman. 2020. "Voter Registration Databases and MRP: Toward the Use of Large-Scale Databases in Public Opinion Research." *Political Analysis* 28 (4): 507–31.
- Gimpel, James G., Frances E. Lee, and Shanna Pearson-Merkowitz. 2008. "The Check is in the Mail: Interdistrict Funding Flows in Congressional Elections." *American Journal of Political Science* 52 (2): 373–94.
- Goplerud, Max. 2015. "Crossing the Boundaries: An Implementation of Two Methods for Projecting Data Across Boundary Changes." *Political Analysis* 24: 121–9.
- Imai, Kosuke, and Kabir Khanna. 2016. "Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Record." *Political Analysis* 24 (2): 263–72.
- Kalla, Joshua L., and David E. Broockman. 2016. "Campaign Contributions Facilitate Access to Congressional Officials: A Randomized Field Experiment." *American Journal of Political Science* 60 (3): 545–58.
- Kingdon, John. 1977. "Models of Legislative Voting." *Journal of Politics* 39: 563–95.

- Lewis, Daniel C. 2013. "Advocacy and Influence: Lobbying and Legislative Outcomes in Wisconsin." *Interest Groups and Advocacy* 2 (2): 206–26.
- Marigalt, Yotam. 2011. "Costly Jobs: Trade-Related Layoffs, Government Compensation, and Voting in U.S. Elections." *American Political Science Review* 105 (1): 166–88.
- Missouri Census Data Center. 2018. "Geocorr 2018: Geographic Correspondence Engine." <http://mcdc.missouri.edu/applications/geocorr2018.html>.
- Naman, Julia Marie, and Jacqueline MacDonald Gibson. 2015. "Disparities in Water and Sewer Services in North Carolina: An Analysis of the Decision-Making Process." *American Journal of Public Health* 105 (10): 20–6.
- Rao, J. N. K. 2003. *Small Area Estimation*. Hoboken, NJ: John Wiley and Sons.
- Rohde, David W. 1979. "Risk-Bearing and Progressive Ambition: The Case of Members of the United States House of Representatives." *American Journal of Political Science* 23: 1–26.
- Shepherd, Michael E., Adriane Fresh, Nick Eubank, and Joshua D. Clinton. 2021. "The Politics of Locating Polling Places: Race and Partisanship in North Carolina Election Administration, 2008–2016." *Election Law Journal* 20 (2): 155–177.
- Steelman, Tyler S., and John A. Curiel. 2022. "Replication Data for: The Accuracy of Identifying Constituencies with Geographic Assignment Within State Legislative Districts." UNF:6:hf3thMP7B4gHG7GM91WVwA== [fileUNF]. <https://doi.org/10.15139/S3/WIN7SM>.
- Swift, Jennifer N., Daniel W. Goldberg, and John P. Wilson. 2008. "Geocoding Best Practices: Review of Eighth Commonly Used Geocoding Systems." <https://spatial.usc.edu/wp-content/uploads/2014/03/gislabtr10.pdf>.
- Winburn, Jonathan, and Michael W. Wagner. 2010. "Carving Voters Out: Redistricting's Influence on Political Information, Turnout, and Voting Behavior." *Political Research Quarterly* 63 (2): 373–86.

Author biographies. Tyler Steelman is the survey research analyst for the Office of Institutional Research and Assessment at UNC Chapel Hill. He completed his PhD in political science and political psychology at the University of North Carolina at Chapel Hill and focuses his research on marginalized identities in contemporary American politics and surrogate representation.

John A. Curiel is an assistant professor of Political Science at Ohio Northern University and earned his PhD in political science from the University of North Carolina at Chapel Hill in August of 2019. He primarily researches how American institutions at the national and state levels mediate representation, in addition to Bayesian and spatial methods.

Cite this article: Steelman, Tyler, and John A. Curiel. 2023. The Accuracy of Identifying Constituencies with Geographic Assignment Within State Legislative Districts. *State Politics & Policy Quarterly* 23 (2): 218–232, doi:10.1017/spq.2022.27