Measuring the Ideology of State and Congressional Districts Using Universal Kriging

Jeff Gill

Departments of Government and Mathematics & Statistics

American University

jgill@american.edu

James E. Monogan III

Department of Political Science

University of Georgia

monogan@uga.edu

December 21, 2018

Abstract

In this paper, we develop and make available measures of public ideology in 2010 for the 50 American states, 435 congressional districts, and state legislative districts. We do this using the geospatial statistical technique of Bayesian kriging, which uses the locations of survey respondents, as well as population covariate values, to predict ideology for simulated citizens in districts across the country. In doing this, we improve on past research that uses the kriging technique for forecasting public opinion by incorporating Alaska and Hawaii, making the important distinction between ZIP codes and ZIP code tabulation areas, and introducing more precise data from the 2010 Census. We show that our estimates of ideology at the state, congressional district, and state legislative district levels appropriately predict the ideology of legislators elected from these districts, serving as an external validity check.

† For helpful assistance, we thank Stephen Jessee, Min Hee Seo, and Yunkyu Sohn. This study is based upon work supported by the National Science Foundation under grant nos. SES-1630265 and SES-1630263. This study also was supported in part by resources and technical expertise from the Georgia Advanced Computing Resource Center, a partnership between the University of Georgia's Office of the Vice President for Research and Office of the Vice President for Information Technology. Previous versions of this work have been presented at the Annual Summer Meeting of the Society for Political Methodology, July 2016, Houston, and the Sixth Asian Political Methodology Meeting, January 2019, Kyoto. Questions may be directed to Jamie Monogan as corresponding author. Complete replication information and our estimates of ideology in 2010 are available at our Dataverse page: http://dx.doi.org/10.7910/DVN/NSTBR7.

In the study of state politics, a constant struggle when studying representation in the American republic is finding reliable measures of public sentiment for the constituencies elected officials serve. In order to see the degree to which voters shape or constrain legislators' actions, a sense of where the voters stand is critical. However, it is hard to find public opinion surveys that are taken at regular intervals, include respondents from all districts of interest, and have a large enough sample for constituency-based subsets of respondents to be big enough to obtain meaningful district-based measures of opinion. For example, if we want to consider how state-level public ideology affects U.S. Senators' behavior in roll call votes, there are scarce options for surveys that cover all 50 states, include a large sample size in each state, and are observed at regular intervals.[1] This problem led Erikson, Wright and McIver (1993) to address this issue by pooling several CBS/*New York Times* polls over time to create a static measure of state ideology, thus sacrificing temporal change to obtain respectable state-level sample sizes. The problem is exacerbated in studies of U.S. House members, which require coverage in 435 smaller districts, and the problem becomes an order of magnitude harder in the study of state legislators (requiring coverage in 1,972 upper chamber districts and 5,411 lower chamber districts). Thus, there is a running challenge in measuring constituency-level public opinion, particularly in smaller districts.

There are several primary strategies for dealing with the difficulty of measuring public ideology. The first is to simply *subset* the survey data by the unit of geographic distinction, which has the advantages of being simple and relying on direct observations of individuals. The main problem with this approach, however, is that the sample size can become quite small even at the state level, much less when looking at districts for the state legislature, and even more challenging when studying demographic subgroups. Additionally, many surveys such as the American National Election Study stratify on region so subsamples are not going to be representative at the state level or lower. A second approach is to use *election returns* as a proxy for ideology in a district (Ansolabehere, Snyder and Stewart 2001; Erikson and Wright 1980; Berry et al. 1998). This approach either uses presidential returns (with the logic being that because the candidates' ideologies are constant nationwide, the vote share will change only in response to the median voter) or uses votes in congressional races (scaling vote shares with measures of the ideology of both incumbents and challengers). While vote-based measures use abundant data that are simple to gather, vote choice is conceptually distinct from ideology. Besides general ideology, votes might be based on regional appeals, personality traits, or economic well-being, thereby inducing added measurement error. Also, vote choice

[1] An important, recent exception to this is the annual Cooperative Congressional Election Study (CCES), which is an internet-based survey that covers every congressional district in the nation (Ansolabehere 2011).

alone may be a misleading measure in that it does not account for the relative dispersion of ideological positions in a district (Kernell 2009). A third possibility is to use *poststratification*, which fits a training model based on survey data and then uses that model to forecast public opinion based on known population data. Several scholars have used weighting and forecast-based measures of public opinion over the years (Pool, Abelson and Popkin 1965; Weber and Shaffer 1972; Weber et al. 1972; Jackson 1989, 2008). The most recent technology is to use multilevel regression with poststratification (MRP), which finds constituency-specific random effects in the survey data (Gelman and Little 1997; Park, Gelman and Bafumi 2004, 2006; Lax and Phillips 2009; Tausanovitch and Warshaw 2013). The idea of incorporating a constituency-specific random effect is reasonable because of all of the unobserved factors that can shape public sentiment in an area. However it can be improved upon in two ways: ideally we would be able to predict random effects even in constituencies where we do not observe survey data, and the geographic variation in random effects may be even more precise than defined borders dictate.

A fourth option, which we build on, is the *universal kriging* approach developed by Monogan and Gill (2016). Universal kriging follows a similar logic to MRP, but uses covariate values measured at the most precise geographical level possible and a smoothed residual structure over geographic space to improve forecasts. The smoothed structure does not abruptly break at border definitions, and it is possible to make forecasts from it even in constituencies without observed survey respondents. While the previous work shows that kriging produces externally valid measures of public sentiment, this study improves on that method in several ways: First, the previous work ignored Alaska and Hawaii as discontiguous states. Here, we propose a solution of relocating these states next to their ideological neighbors in the contiguous 48 states to obtain measures of ideology in all 50 states. Second, the previous work erroneously located survey respondents with ZIP Code Tabulation Areas (ZCTAs), when the survey recorded respondent ZIP code. These are not equivalent, so we address this problem here. Third, we improve upon prior work by using newer data from the 2010 Census, and we specifically krige with much more precise information. The 2010 Census reports data at the census block level, allowing us to draw simulated citizens closely in line with population density. Covariates also are now sampled from the most precise possible level—often the block level itself. Hence, our estimates should be more accurate in smaller constituencies. Fourth, we apply this method not only at the state level, but also in congressional and state legislative districts. Consequently, a product of our work is that we now release for public use measures of public ideology in 2010 for the 50 states, 435 congressional districts, districts for upper chambers of state legis-

latures, and districts for state legislative lower chambers. Fifth, we present a new program for estimating for kriging models, a full-information Metropolis-Hastings algorithm. Altogether, this work represents a marked advance in the universal kriging technique.

We proceed first by reviewing the model itself, why substantively it should work, and the data we use in this application. Second, we describe in detail the new advances that we make in how universal kriging can be applied to measuring public opinion. Third, we describe the results from our estimated model using 2008 Cooperative Congressional Election Study (CCES) data. Fourth, we present our forecasted measures of ideology and several validity checks. Fifth, we present our new Metropolis-Hastings algorithm with an application that examines mineral deposit levels near the Meuse River. Sixth, we describe remaining challenges of our work. Finally, we describe the implications of our study and room for future work.

1 Point-to-Block Realignment with Universal Kriging

Our method for translating a public opinion survey into measures of constituent-level ideology follows the logic of point-to-block realignment (Banerjee, Carlin and Gelfand 2015, Chapter 7; Monogan and Gill 2016). The intuition behind this technique is to estimate a model of observations that are located at points in space (such as latitude and longitude), make several predictions from this model at a wider range of points in space using known covariate values, and then use the predictions falling within a block (or border-referenced area in space) to produce a block-level forecast. In our case, we will locate survey respondents in geographic space using known information about their address (treating them as points in space), use population Census data at various geographic locales to make predictions throughout the United States, and then average all predictions falling within an electoral district to determine the average ideology of that constituency. Hence, the 50 states, 435 congressional districts, or state legislative districts form our block, or areal, units of interest in this point-to-block realignment.

Meanwhile, our middle step of using population Census data to make forecasts of several simulated citizens in districts across the nation follows a similar logic to weighting, regression, or MRP techniques, except our forecasts can include a spatial error term that borrows strength from nearby observed survey respondents. To do this, the model we estimate over our training data must be a kriging model that allows for covariance among geographically proximate respondents. Kriging has had some uses in Political Science, both in predicting potential

campaign contributions at residences (Tam Cho and Gimpel 2007) and the wind direction at major pollution sites (Monogan, Konisky and Woods 2017). The two general types of kriging are *ordinary kriging*, which relies purely on a spatial error process to make predictions, and *universal kriging* which also allows spatial trend terms and even location-specific exogenous predictors to shape the prediction. We follow the universal kriging approach advanced by Monogan and Gill (2016), which uses a linear prediction based on demographic predictors and a polynomial trend term, plus the spatial error prediction. The nice feature of this is that our spatial error term forms a density blanket such that we can make predictions for any constituency spanned by the locations of our respondents, even if there were no observed survey respondents within the district of interest.

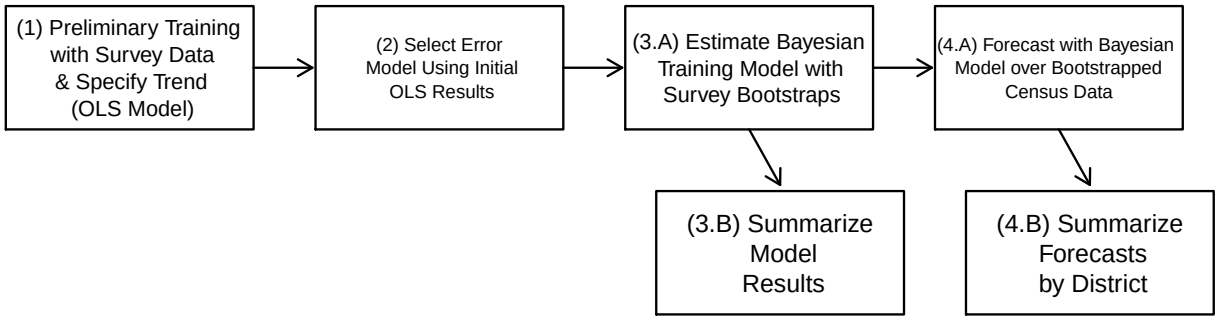


Figure 1: Flowchart showing the steps of point-to-block realignment.

Figure 1 illustrates the steps of point-to-block realignment. Step (1) is to fit a preliminary linear model with ordinary least squares (OLS). Here we try several specifications of the OLS model to gauge the proper functional form for a trend term in longitude and latitude (or eastings and northings) based on model fit. In step (2), we can examine the OLS residuals from the best-fitting model in the prior step and determine what the best-fitting functional form of the spatial error process is. That is, given the covariates and our chosen geographic trend term, how do our errors spatially correlate and what function best summarizes that correlation structure? Possible error process models for the residuals include (among others) the exponential, Gaussian, spherical, wave, or Matérn processes (Banerjee, Carlin and Gelfand 2015, 25-30). Step (3.a) is to estimate the Bayesian model with survey bootstraps. This model treats the conditional mean of ideology as a function of individual covariates and the geographic trend term, and it simultaneously estimates the parametric error structure decided on in the previous step. Due to computational limitations, the model is estimated for subsets of the survey data, each a random draw from the larger survey sample. At the end of this step, we pool all parameter posterior samples together across bootstrapped runs to form one

large posterior sample with which to summarize our parameters of interest. After estimating the model, we can proceed in two ways: In step (3.b) we can summarize our model's results by reporting descriptive statistics from the pooled posterior draws. Meanwhile, in step (4.a) we can begin forecasting with bootstrapped Census data. With the forecasting, we take a set of bootstrapped results from our training model, and we forecast over a random sample of Census data. We draw a fresh random sample of the Census data for each bootstrapped sample, thereby forecasting for a wide range of individuals in our kriged Census sample at locations spread throughout the nation in proportion to population density. In step (4.b), our final step, we summarize our forecasts by district: We simply pool all of our kriged Census individuals from bootstrapped forecasts and organize the larger pool into the districts of interest in which these simulated citizens reside. Once organized based on constituency, we compute descriptive statistics of these kriged forecasts by district. This provides us with our district-level forecasts of the mean and variance of ideology within each constituency—be the district a state, congressional district, state legislative district, or something else.

1.1 Specifying the Model

The method of *point-to-block realignment* assumes that the observed point-level (person) data and the extrapolated block-level averages (area) have a joint Gaussian distribution. We start by specifying how the *training* side of the model works to fit a model over observed survey data. Define now $\mathbf{s}$ as a set of n observed sites $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$, where each $\mathbf{s}_i$ represents the location of a survey respondent in space—either in latitude and longitude, or in northings and eastings (as we use in this application).[2] Here $\mathbf{Y}(\mathbf{s})$ is an associated collection of outcomes $\mathbf{Y}(\mathbf{s}) = \{Y(\mathbf{s}_1), Y(\mathbf{s}_2), \ldots, Y(\mathbf{s}_n)\}$, the survey response of interest for the survey-taker at each site. $\mathbf{X}^*(\mathbf{s}) = \{\mathbf{x}^*(\mathbf{s}_1), \mathbf{x}^*(\mathbf{s}_2), \ldots, \mathbf{x}^*(\mathbf{s}_n)\}$ is a collection of covariates for each survey respondent observed at his or her respective point in space. We specify a linear model as follows:

$$\mathbf{Y}(\mathbf{s}) = \boldsymbol{\mu}(\mathbf{s}) + \boldsymbol{\omega}(\mathbf{s}) + \boldsymbol{\epsilon}(\mathbf{s}), \tag{1}$$

where: $\boldsymbol{\mu}(\mathbf{s}) = \mathbf{X}(\mathbf{s})\boldsymbol{\beta}$ is the mean structure based on a linear additive component (like a standard regression model), $\boldsymbol{\omega}(\mathbf{s})$ are realizations from a mean-zero stationary (usually)

[2]Northings and eastings are an alternative to latitude and and longitude advocated by the U.S. National Imagery and Mapping Agency and used by most militaries. These are defined by the Universal Transverse Mercator (UTM) which establishes 60 curved vertical "strips" across the globe, each with 6 degrees of longitude starting at 180 degrees. Within this UTM, grid points are offsets in meters where northing is the distance from the equator and easting is the distance from the closest western line of the 60 vertical zone boundaries. The southern hemisphere is made positive in northings by adding a constant. There are a variety of possible projections and reference points, and we define ours later in the paper.

Gaussian spatial process that captures spatial association (closer points are more informative than distant points), and $\boldsymbol{\epsilon}(\mathbf{s})$ is a regular uncorrelated disturbance term.

An important feature of Equation 1 is that the variance is split into two disturbance terms: one that captures spatial association, and the other that is a traditional independent and identically distributed error term with homoscedastic variance. We thereby use the following distributional assumptions for these two terms: $\boldsymbol{\omega}(\mathbf{s}) \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{H}(\phi))$, and $\boldsymbol{\epsilon}(\mathbf{s}) \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I})$. Several of the parameters of these variance components have a substantive interpretation. From the idiosyncratic error term, $\boldsymbol{\epsilon}$, we call the variance term τ^2 the *nugget*. This is the amount of error variance in the outcome that is independent from spatial separation. We can think of this as the variance in the error when the geographic separation between observations is negligible. Turning to the spatial $\boldsymbol{\omega}$ error terms, σ^2 is called the *partial sill*. The partial sill reflects the variance that can be driven by geographic distance between two observations, with the assumption being that more distant observations have a higher variance. The partial sill equals the maximum amount of variance among observations due *strictly* to geographic separation. In fact, the nugget plus the partial sill equals the *sill*, which is the maximum total variance possible among distant observations. Finally, the other parameter feeding into the spatial $\boldsymbol{\omega}$ error terms is the *range* term, $R = 1/\phi$. (ϕ itself is called the decay term.) When the distance between observations is as great or greater than the range term R, then the variance among those observations equals the sill. In other words the range term tells us the threshold distance at which error variance is maximized.

The last piece of specifying $\boldsymbol{\omega}(\mathbf{s})$ is that we must specify the function $\mathbf{H}(\phi)$. This is a parametric spatial correlation function that typically only requires us to estimate the decay parameter ϕ. We typically assume an *isotropic* model, which means that the level of spatial correlation does not depend on direction but only on the distance between the observations $d_{ij} = \|\mathbf{s}_i - \mathbf{s}_j\|$. In this case, we must choose a parametric model—the exponential, spherical, wave, and Gaussian are a few common options—that captures the patterns of residual association in our data. Each of these parametric models specifies both a spatial correlation function (stating simply how much observations should correlate given their distance apart), as well as a semivariogram function (specifying how much observations should vary given their distance apart). The two fit naturally together with a high correlation implying a low variance and vice versa. Once we determine the best parametric correlation function, the product of the correlation function $\mathbf{H}$ with the partial sill σ^2 builds the spatial covariance structure into the joint distribution of the $\boldsymbol{\omega}(\mathbf{s})$ disturbance terms.

When determining the exact parametric specification of $\mathbf{H}$, we normally focus on the

related semivariogram to determine the right parametric structure. We choose the right model through an empirically-driven process wherein the *empirical semivariogram* is calculated from the residuals of an initial model. The formula for the empirical semivariogram is (Cressie 1993, 69):

$$\hat{\gamma}(d) = \frac{1}{2|N(d)|} \sum_{(i,j)\in N(d)} |z(\mathbf{s}_i) - z(\mathbf{s}_j)|^2, \tag{2}$$

where $z(\mathbf{s}_i)$ is the residual term for the respondent located at site $\mathbf{s}_i$ from an initial linear model, d is an approximate distance of interest (possible distance values are usually coarsened into bins), $N(d)$ is the set of all pairs of observations such that $|z(\mathbf{s}_i) - z(\mathbf{s}_j)| \approx d$, and $|N(d)|$ is the number of pairs in the set that are separated by distance d. The semivariogram equals two quantities: the variance of all observations separated by distance d when pooled together, as well as half the variance of the differences $(z(\mathbf{s}_i) - z(\mathbf{s}_j))$ between observations separated by distance d. Using the empirical semivariogram, we then determine which parametric model is most appropriate for our data, choosing from the exponential, spherical, wave, Gaussian, or some other parametric semivariogram. Once we have that, we know the related spatial correlation function (Banerjee, Carlin and Gelfand 2015, 28-29). In our case here, the best-fitting model is the Gaussian semivariogram, so implies that our spatial correlation function should be $H(\phi)_{ij} = \exp(-\phi^2 d_{ij}^2)$.

With all of these elements in place for modeling the responses of survey respondents, we now step back and think about where this *training* model fits relative to our *forecasting* process of state, congressional district, and state legislative district ideology. Since our model assumes the observed point-level data and the extrapolated block-level averages have a joint Gaussian distribution, we get:

$$f\left(\begin{pmatrix} \mathbf{Y_s} \\ \mathbf{Y_B} \end{pmatrix} \middle| \boldsymbol{\beta}, \sigma^2, \phi\right) = \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu_s}(\boldsymbol{\beta}) \\ \boldsymbol{\mu_B}(\boldsymbol{\beta}) \end{pmatrix}, \sigma^2 \begin{pmatrix} \mathbf{H_s}(\phi) & \mathbf{H_{s,B}}(\phi) \\ \mathbf{H}_{\mathbf{s,B}}^T(\phi) & \mathbf{H_B}(\phi) \end{pmatrix}\right),$$

where $\mathbf{Y_s}$ represents the vector of ideology among individual citizens, $\mathbf{Y_B}$ represents the vector of ideology in all block-referenced constituencies of interest, and $\mathbf{H}$ defines the correlation matrix of observations as before. Note that this presents the simplified case where there is no nugget effect (τ^2), but the result still holds if the variance-covariance terms do include a nugget. By standard normal theory (e.g. Ravishanker and Dey 2002), the conditional distribution of

our extrapolated block averages is:

$$\begin{aligned}\mathbf{Y_B}|\mathbf{Y_s}, \boldsymbol{\beta}, \sigma^2, \phi \;\; &\sim \;\; \mathcal{N}(\boldsymbol{\mu}_\mathbf{B}(\boldsymbol{\beta}) + \mathbf{H}^T_{\mathbf{s},\mathbf{B}}(\phi)\mathbf{H}^{-1}_\mathbf{s}(\phi)(\mathbf{Y_s} - \boldsymbol{\mu}_\mathbf{s}(\boldsymbol{\beta})), \\ &\qquad \sigma^2[\mathbf{H_B}(\phi) - \mathbf{H}^T_{\mathbf{s},\mathbf{B}}(\phi)\mathbf{H}^{-1}_\mathbf{s}(\phi)\mathbf{H}_{\mathbf{s},\mathbf{B}}(\phi)]). \qquad (3)\end{aligned}$$

These quantities can be estimated with Monte Carlo integration:

$$\begin{aligned}(\hat{\boldsymbol{\mu}}_\mathbf{B}(\boldsymbol{\beta}))_k &= L_k^{-1} \sum_\ell \mu(\mathbf{s}_{k\ell}; \boldsymbol{\beta}) \\ (\hat{\mathbf{H}}_\mathbf{B}(\phi))_{kk'} &= L_k^{-1} L_{k'}^{-1} \sum_\ell \sum_{\ell'} \rho(\mathbf{s}_{k\ell} - \mathbf{s}_{k'\ell'}; \phi) \\ (\hat{\mathbf{H}}_{\mathbf{s},\mathbf{B}}(\phi))_{ik} &= L_k^{-1} \sum_\ell \rho(\mathbf{s}_i - \mathbf{s}_{k\ell}; \phi)\end{aligned}$$

We conduct this Monte Carlo integration using the technique of Bootstrapped Random Spatial Sampling (BRSS) developed by Monogan and Gill (2016). Doing so allows us to forecast the average ideology with:

$$\hat{\boldsymbol{\mu}}_\mathbf{B}(\boldsymbol{\beta}) + \hat{\mathbf{H}}^T_{\mathbf{s},\mathbf{B}}(\phi)\hat{\mathbf{H}}^{-1}_\mathbf{s}(\phi)(\mathbf{Y_s} - \hat{\boldsymbol{\mu}}_\mathbf{s}(\boldsymbol{\beta})) \qquad (4)$$

We account for the spatial element by forecasting $\hat{Y}(\mathbf{s}_{k\ell}; \boldsymbol{\beta}, \sigma^2, \tau^2, \phi)$ and using this quantity in our Monte Carlo integration. With the nugget effect, from $\mathbf{Y}(\mathbf{s}) = \boldsymbol{\mu}(\mathbf{s}) + \boldsymbol{\omega}(\mathbf{s}) + \boldsymbol{\epsilon}(\mathbf{s})$, we also get $\mathbf{Y_s} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where we still require: $\boldsymbol{\Sigma} = \sigma^2\mathbf{H}(\phi) + \tau^2\mathbf{I}$, $H(\phi)_{ij} = \rho(\phi, d_{ij})$, and $d_{ij} = ||\mathbf{s}_i - \mathbf{s}_j||$. Again, for this application the Gaussian semivariogram function was the best fitting, meaning that our correlation function is $H(\phi)_{ij} = \exp(-\phi^2 d_{ij}^2)$.

1.2 Why Proximity Matters for Public Opinion

Tobler's *First Law of Geography* states: "Everything is related to everything else, but near things are more related than distant things" (Tobler 1970, 236). Here we assume that this law holds for individuals' opinions and ideology, also, with more physically proximate Americans holding a more similar political outlook. While there are physical barriers, such as highways and rivers, that separate populations and therefore can change ideology dramatically, our kriging approach connects these smoothly with no sudden shift.

Proximal influence in politics is supported Gimpel and Schuknecht who describes two different approaches to understanding regionalism in this way (2003, 2-4). First Gimpel gives a *compositional approach* asserting that political behavior is similar within a region due to economic interests, racial origin, ethnic ancestry, religion, social structure, and other related factors (Gastil 1975; Garreau 1981; Fischer 1989; Lieske 1993). Therefore if all of these factors

could be included in an empirical model, then the variability with geographic units in the model would be small. Clearly, though, it is often impossible to measure every relevant demographic and socioeconomic variable, or even to identify every critical variable for inclusion. Since some relevant inputs may be overlooked or unmeasured, we assume that neighboring individuals will have a relatively similar political outlook, even holding included covariates constant. Second, Gimpel gives a *contextual approach* that offers the idea that citizens' political attitudes and behaviors are influenced by political socialization and by interactions with other citizens in their social network, which is supported by a large literature in political science (Putnam 1966, 1993; Huckfeldt and Sprague 1995; DeLeon and Naff 2004; Djupe and Sokhey 2011). For instance, "the first place to look for political networks is within the immediate physical proximity of each individual" (Sinclair 2012, 26). This means that we expect under the contextual approach as well that geographically proximate individuals will have relatively similar opinions, even in a general setting.

Furthermore, Erikson, Wright and McIver propose that, "the unique political cultures of individual states exert an important influence on political attitudes" (1993, 48). This idea goes as far back as Elazar (1966), who proposes that U.S. states can be categorized based on an individualist, moralist, or traditionalist view of government's role. Erikson, Wright and McIver also demonstrate that a higher proportion of variance in ideology and partisanship is be predicted by state-level dummies than by demographic information, although state-level residuals will pick up some of the effects of unmeasured individual-level variables (1993, 56-68). We also find evidence that this holds urban areas where political culture shapes the impact of identity on public opinion and political participation, even in cities with heterogeneous neighborhoods (2004, 703). Our method of kriging increases the ability to accurately model the effects of political culture, omitted predictors, and social context by including weighted neighbors' residuals in forecasts of public opinion and ideology. For example, western Kentucky and southeast Illinois are similar places that are likely to be populated with similar people, both culturally and in demographic terms.

1.3 Data: 2008 CCES and 2010 Census

In this study, we use the 2008 Cooperative Congressional Election Study (CCES) as training data to estimate our model of individual ideology as a function of demographics. The 2008 CCES offers 21,849 observations spread across the American states and congressional districts. This survey asks respondents to place themselves ideologically on a scale from 0 to 100, with 0 representing the most liberal and 100 the most conservative. Our training model predicts

responses as a function of age, education, race, sex, income, religion, urban-rural status, homeownership, employment status, and a geographic trend term. CCES respondents were located geographically by ZIP code. Our procedure for locating these respondents is described in greater detail in the next section.

After estimating the training model over the CCES, we used 2010 Census data to forecast the ideology of 250,000 simulated citizens throughout the continental United States. We simulated by census block, the most precise geospatial unit the Census Bureau keeps track of, drawing proportionally by the population of the block. We used Census Bureau maps of census blocks to place simulated citizens in latitude and longitude (or more exactly in eastings and northings), and we drew predictor values based on the variables' local distribution for that block.

The 11 million census blocks perfectly tesselate all higher-level geospatial indicators of which the Census Bureau keeps track, so there are no gaps and no overlaps of areal units. Hierarchically, blocks are nested within block groups, block groups are nested within tracts, and tracts are nested within counties. When simulating covariate values for a kriged point, if a predictor was not reported at the block level we drew from the most precise level for a given location. More specifically, we simulate age, race, sex, and homeownership based on block-level data. We simulate education and income based on block group-level data. We simulate employment status based on tract-level data. We simulate religion and urban-rural status with county-level data. By using the 2010 census block data, all of these predictors are simulated with greater geographic precision and with more up-to-date data than in Monogan and Gill (2016).

2 Innovations in Kriging for Measuring Public Opinion

Besides using more recent and more precise data, we offer two more methodological advances for the technique of kriging to forecast public opinion. Specifically, past work did not include estimates of opinion in Alaska and Hawaii because they are not contiguous with the rest of the United States. In this article, we offer a solution to this problem and create new estimates for these states and their component districts. Additionally, past work erroneously used Census ZIP Code Tabulation Areas (ZCTAs) to locate survey respondents in state with their ZIP code. We discuss why that is a problem and introduce new data that resolve this issue, thereby creating better forecasts. We discuss each innovation in turn.

2.1 Moving Noncontiguous States

Any method of measuring the ideology of public constituencies ought to be comprehensive in covering all fifty states as well as all state legislative and congressional districts falling within each. A major challenge of using spatial data analysis to measure public opinion in the United States is that it is difficult to measure opinion in the non-contiguous states of Alaska and Hawaii. In fact, measuring public opinion in these two states is often difficult anyway on account of having few, if any, survey respondents in many national polls. Kriging, like many methods of spatial analysis, requires observations to have neighbors. If we attempted to train a kriging model using Alaska's and Hawaii's data as they are located on a map, then we would have extreme geographic outliers that could distort the estimation of our spatial error process model. This, in turn, would diminish our ability to make accurate predictions of opinion as we turned to forecasts because the partial sill would treat ocean-distance as regular geographic space and create a smoothed spatial surface over broad swarths of the Pacific Ocean and Canada. When the goal is to understand and predict the opinions of American citizens, this is not a sound substantive approach.

To address this, we proceed in two ways: First, when estimating the model itself, we omit Alaska and Hawaii from the training data. Their extreme outlier values could unduly affect the spatial variance components, so only the continental 48 states were included in the training data. Second, when forecasting ideology in these two states, we relocate Alaska and Hawaii to sit next to the west coast of the United States. Doing so greatly narrows the out-of-sample space that falls within the convex hull of our forecasting space. That is, the areas that are part of Canada, Mexico, or the Pacific Ocean that are included within the space of our smoothed kriging surface are shrunken dramatically relative to a model that uses these two states at their actual geographic location.

For the sake of forecasts, we locate Alaska and Hawaii near their *ideological neighbors*, or locales on the west coast as similar as possible ideologically. To find these states' ideological neighbors, we estimated a training model on the continental 48 states and then chose the locations off the west coast that minimized predictive error for the two discontiguous states when using that model to predict Alaska and Hawaii's observations in survey data. (Our full process is described in more depth in Appendix A.) Figure 2 shows the result of our procedure, illustrating how we relocated Alaska and Hawaii for the sake of the forecasting data. Specifically, that map shows a dot at the location of each census block's centroid (a census block serving as our primary unit for sampling forecasting observations in kriging). The census blocks for the continental 48 states are in black at their original locations in eastings

and northings. The census blocks for Hawaii (in blue) and for Alaska (in red) are at their new ideological neighbor locations in eastings and northings.

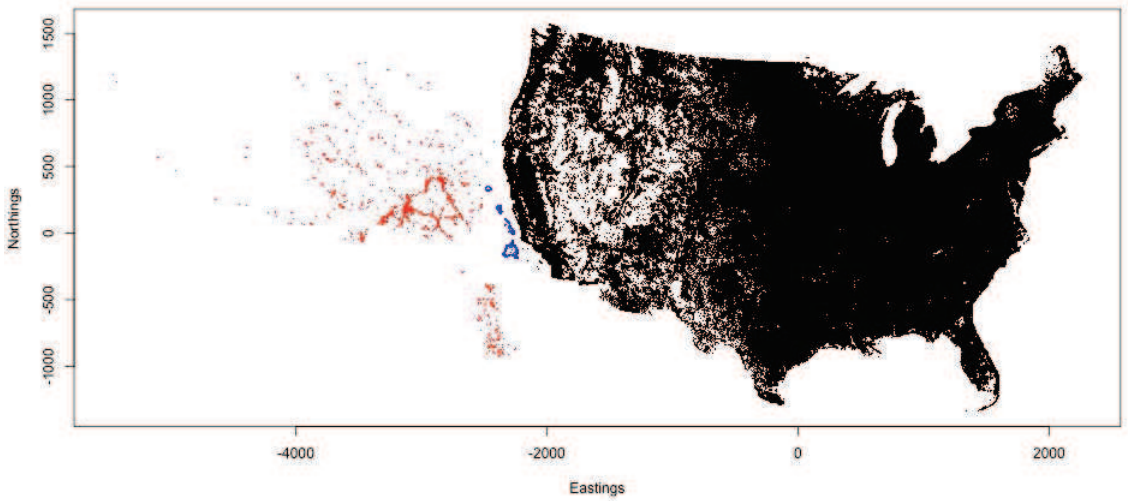


Figure 2: Map of census block centroids when Alaska and Hawaii are placed near their ideological neighbors from kriging forecasts.

Substantively in Figure 2, Hawaii has been relocated so that Honolulu is an ideological neighbor with San Francisco. Alaska has been repositioned so that Anchorage is located close to Santa Barbara. These positions, again, are the positions that minimize forecasting errors in the two discontiguous states, as detailed in the appendix. This allows each state to have a west coast neighbor that is ideologically similar without producing any overlap between either state and the continental states. In order to preserve area and point-to-point distances, the original locations of census blocks (with Alaska and Hawaii at their actual locations) were reprojected from longitude and latitude into eastings and northings first. After this reprojection, Alaska and Hawaii were relocated to the positions shown on the map. This two-step process was repeated for the training data (ZIP-code referenced CCES respondents) and for the forecast data (census blocks with centroid coordinates). This solution of finding ideological neighbors is far superior to the common solution to simply drop these two states from measurement models: According to 2010 Census numbers, dropping these states would mean ignoring over 2 million U.S. citizens (710,249 + 1,360,301 = 2,070,550). It is possible to model these two states separately, but that imposes the assumption that there is no influence back and forth between these two states and the contiguous 48 states. Our solution is a compromise between these two extremes that allows inclusion without deteriorating the quality of the total model.

2.2 ZIP Codes versus ZIP Code Tabulation Areas

Prior kriging work placed survey respondents on a map using ZIP Code Tabulation Areas (ZCTAs), as computed by the Census, when respondents' geographic identifier was ZIP code. Mechanically, if a respondent is known to reside in a geographic area, he or she has to be placed at a specific coordinate using eastings and northings. This has been done by starting at the centroid of the areal unit and jittering within the radius of the unit's area. The problem of doing this with ZCTAs when ZIP code is the true geographic identifier is that the area of ZCTAs does not exactly overlap with the areas covered by ZIP codes themselves (Beyer, Schultz and Rushton 2008; Grubesic and Matisziw 2006; Grubesic 2008). Hence, respondents could be placed at a position on the map that puts them in the wrong ZIP code, adding unnecessary measurement error to the model.

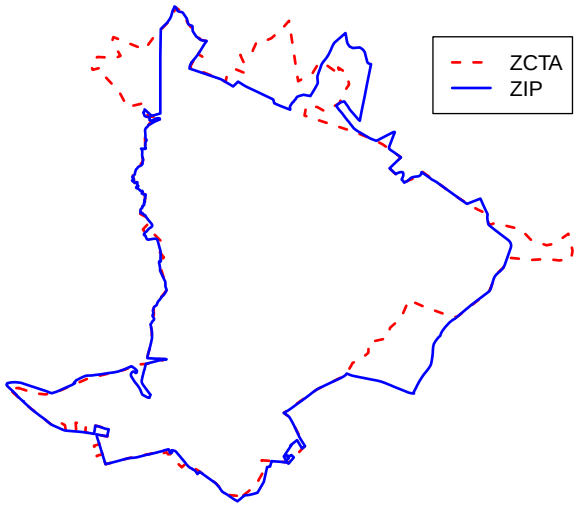


Figure 3: Map illustrating non-overlap of an example ZIP code compared with corresponding ZCTA.

Figure 3 illustrates the problem. This figure draws the real map of the 30601 ZIP code in Georgia using data obtained by TomTom as well as the 30601 ZCTA using data obtained from Census. The solid blue line shows the ZIP code boundary, and the dashed red line shows the ZCTA boundary. As can be seen, if we knew a resident lived in the 30601 ZIP code but placed them at a location in the ZCTA, we could make several mistakes. First, there are observable points in the ZCTA that are outside of the ZIP code. In the east (right) and the north (top) in particular, there are several places in the ZCTA that stray well outside of the ZIP code. If we placed a survey respondent who identified 30601 as his or her ZIP code in one of these portions of the ZCTA, we would have placed them in the wrong ZIP code. A second problem that emerges is that the ZCTA does not cover all of the ZIP code. In the southeast

(bottom-right) in particular, there is a large block of land where residents of the 30601 ZIP code could live. If we proceed to locate these individuals using the ZCTA, then we have no chance of putting them at the correct location on the map.

This problem emerges because of the nature of ZIP codes and how the Census has had to deal as best as possible with the issue of investigators' need of ZIP code-referenced demographic data. ZIP codes themselves are not areal units with defined borders. Rather, ZIP codes are routes defined by the U.S. Postal Service prescribing how to efficiently deliver mail. Hence, *there is no official map of where one ZIP code ends and the next begins.* To create demographic and geographic data by ZIP code (since it is a common locator recorded for Americans), the Census Bureau created ZCTAs for the 2000 Census—mindful to warn users that ZIP codes cross-cut even census blocks, the smallest geographic unit the Bureau records. As a best alternative, the Census records the ZIP code that a majority of addresses in a census block use. A ZCTA then is formed as a combination of all census blocks with the same majority ZIP code. This is certainly an important tool that the Census Bureau provides, and in cases in which demographics need to be measured by ZIP code it is the best alternative available. However, residents who have a ZIP code that is held by a minority of addresses in their census block will be placed in a ZCTA that differs from their ZIP code.

In our case, we only need ZIP code locations in order to locate respondents of the CCES training data. Hence, we turn to a new alternative that deals with the issue of locating the position of ZIP codes themselves in space. Specifically, we use a 2014 dataset that draws from TomTom navigation services. This map defines ZIP code boundaries based on actual addresses, drawing a border around the complete set of addresses with a particular ZIP code. We therefore were able to compute the centroid and radius of actual ZIP codes and then link this information to the CCES to place survey respondents in space. This allowed us to estimate a model over our CCES training data that allowed for spatial correlation among nearby respondents.

Importantly, we only use ZIP codes at the *training* stage of estimating the model. When *forecasting* or kriging ideology, we use extremely precise census block data from the 2010 Census. At the forecasting stage, we can sample from the population using any geographic unit we wish, as long as we know both the location of the unit and the distribution of demographic predictors within that unit. A census block is the smallest possible geographic unit we can sample from, with 11 million of them defined in 2010. By forecasting using census block data, we can make predictions in places that are often as small as a city block using records of the U.S. Census recorded from that small area to sample demographic predictors. This maximizes

predictive accuracy and avoids the ZIP code question altogether at the predictive stage of the model. Between the TomTom ZIP code data for the training stage and the U.S. Census Bureau's block data for the forecasting stage, we maximize the accuracy in estimation and prediction.

3 Training Model with CCES Data

With Alaska and Hawaii moved to sit next to the west coast of the continental United States and the survey respondents placed geographically by their actual ZIP code, we turn to the estimation of our training model. As described before, we are estimating a model over the 2008 CCES, which we will then use to make forecasts with 2010 Census population data throughout electoral constituencies in the United States. Our first step, then, is actually estimating the spatial model with the CCES data. The full specification of our Bayesian model for the training data is as follows:

$$\begin{aligned}
\mathbf{Y_s} &\sim \mathcal{N}(\mathbf{X_s}\boldsymbol{\beta}, \boldsymbol{\Sigma}) \\
\boldsymbol{\Sigma} &= \sigma^2\mathbf{H}(\phi) + \tau^2\mathbf{I} \\
H(\phi)_{ij} &= \exp(-\phi^2 d_{ij}^2) \text{ (Gaussian correlation function)} \\
\pi(\boldsymbol{\beta}) &\sim \text{flat} \\
\pi(\tau^2/\sigma^2) &\sim \text{Unif}(6, 8) \\
\pi(\sigma^2) &\sim 1/\sigma^2 \\
\pi(1/\phi) &\sim \text{Unif}(0, 12000)
\end{aligned} \tag{5}$$

Here, $\mathbf{Y}$ refers to the individual's self-reported ideology on a 0-100 scale, $\mathbf{s}$ refers to the individual's location in eastings and northings, $\mathbf{X}$ refers to a vector of individual-level demographic predictors of ideology, $\boldsymbol{\beta}$ is the vector of regression coefficients, $\boldsymbol{\Sigma}$ is the covariance matrix of $\mathbf{Y}$ given the predictors, σ^2 is the partial sill term, τ^2 is the nugget effect, $\mathbf{H}$ is the correlation matrix of observations, ϕ is the decay term, and d_{ij} is the geographic distance between observations i and j. Of note, the third line of the specification shows that each cell of the correlation matrix is defined by a Gaussian correlation function: This means that the correlation between observation i and observation j depends solely on the distance (d_{ij}) between them as prescribed by the correlation function. Each coefficient (β) has a flat prior, the ratio of the nugget to the partial sill has a uniform prior from 6 to 8 (based on our observation that the nugget variance is about 7 times the partial sill variance), the partial sill itself has a

(conservative) reciprocal prior, and the range term ($1/\phi$) has a uniform prior from 0 to 12,000 kilometers.

Table 1: Bayesian Spatial Model of Self-Reported Ideology Using BRSS

Parameter	Estimate	Std. Error	90% CI	
Age	0.1730	0.1618	[-0.0950:	0.4377]
Education (six categories)	-1.7418	2.8893	[-6.5364:	2.9300]
Age×education	-0.0189	0.0538	[-0.1054:	0.0701]
African-American	-8.9397	5.9852	[-18.8885:	0.7519]
Nonwhite, nonblack	-0.4998	4.8544	[-8.3121:	7.6381]
Female	-3.7610	2.4136	[-7.8902:	0.0579]
African-American female	6.9816	7.8155	[-5.6070:	20.0377]
Nonwhite, nonblack female	-1.4811	6.7514	[-13.1171:	9.0445]
Income ($10000-$14999)	-3.0830	9.3233	[-17.9325:	12.7183]
Income ($15000-$19999)	-3.9677	9.8787	[-20.4196:	12.1683]
Income ($20000-$24999)	-0.5696	8.4009	[-14.5332:	13.1658]
Income ($25000-$29999)	-0.8421	9.3513	[-16.6975:	14.0028]
Income ($30000-$39999)	-0.5540	8.3337	[-14.0389:	13.3959]
Income ($40000-$49999)	-0.0847	8.2990	[-13.5631:	13.8652]
Income ($50000-$59999)	0.2999	8.5763	[-14.0713:	14.2191]
Income ($60000-$69999)	-0.5245	9.0996	[-15.7085:	14.2139]
Income ($70000-$79999)	1.6559	8.4463	[-12.0908:	15.8484]
Income ($80000-$99999)	0.6219	9.0557	[-14.0566:	15.5985]
Income ($100000-$119999)	0.6195	8.6163	[-13.6012:	14.8625]
Income ($120000-$149999)	0.9644	9.1339	[-14.0129:	15.9063]
Income ($150000 or more)	-0.2145	8.2411	[-13.4899:	13.6303]
Catholic	7.1849	2.9065	[2.3648:	11.9254]
Mormon	21.2255	7.8989	[8.1493:	34.1530]
Orthodox	6.6772	15.3959	[-19.0482:	31.3295]
Jewish	-5.3877	7.1392	[-16.9935:	6.5644]
Islam	-2.9257	21.4662	[-38.0358:	32.5910]
Mainline	5.2233	3.5823	[-0.6456:	11.1997]
Evangelical	16.1484	3.0724	[11.0782:	21.2105]
Ruralism (nine categories)	0.7858	0.6876	[-0.4005:	1.8820]
Homeowner	4.9572	2.7090	[0.6070:	9.5330]
Unemployed	-1.3962	5.0840	[-9.9154:	6.7799]
Not in workforce	-0.1537	2.4769	[-4.1260:	4.0965]
Eastings	0.1292	2.5024	[-3.9961:	4.1866]
Northings	-2.3598	3.3950	[-7.5980:	3.4632]
Eastings^2	-0.4953	1.1702	[-2.4270:	1.2934]
Northings^2	-0.6886	3.5232	[-6.4463:	5.0737]
Eastings×northings	-0.3293	2.0707	[-3.6050:	3.0592]
Intercept	47.0194	14.2282	[23.7830:	70.4446]
σ^2	86.6253	8.8827	[73.2144:	102.0180]
$1/\phi$	5922.7947	2843.5643	[1493.2441:	10364.8758]
τ^2	602.7219	34.1616	[547.3352:	659.6399]

Notes: N =21,849. Data from 2008 CCES.
Results based on 50 subsamples of 5% original data.
1,000 iterations were run for each subsample, for 100,000 total
Computed with the `geoR` 1.7-5.1 library in `R` 3.2.3.
Eastings and northings rescaled to megameters (Mm) in this table.

We report the results of this model in Table 1. For each parameter in the table, the first numeric column reports the mean of the marginal posterior distribution for the parameter, which serves as a point estimate of our term. The second numeric column reports the standard deviation of the marginal posterior distribution for the parameter, which serves as a standard error. The last two columns report the 90% credible interval, meaning there is a 90% probability that the parameters falls within that range. The first 38 rows report summary statistics for the regression coefficients included in the model. Our goal with this model is to maximize predictive ability, so we include any predictor that is both known to predict ideology and for which population data are observed. As the table shows, these predictors include age, eduction, race, sex, income, religion, rural-versus-urban, homeownership, and employment status. We also model trends in geographic space by including the respondents' coordinates in eastings and northings in the model—in linear, interactive, and quadratic forms. The last three rows of the table summarize the marginal posteriors for the three terms that characterize the spatial error process: the partial sill (σ^2), range ($1/\phi$), and nugget (τ^2) terms.

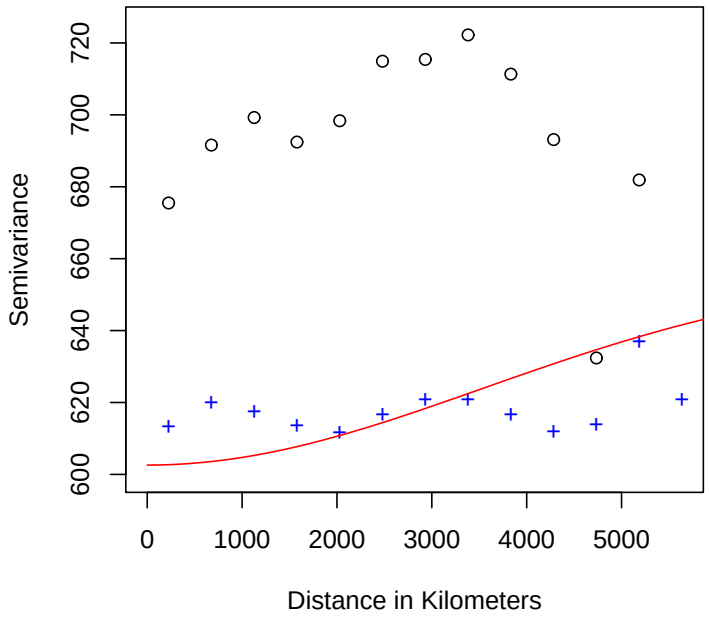


Figure 4: BRSS Estimated Semivariogram.

Figure 4 offers another illustration of how the spatial error process works given these parameters. The horizontal axis of this plot represents the approximate distance between two survey respondents' locations. The vertical axis represents the semivariance of observations

separated by this distance—again referring to either half the variance of the difference between observed responses separated by that distance or the whole variance of undifferenced responses separated by that distance when pooled together. The open black circles along the top show the empirical semivariance of raw survey responses from the 2008 CCES. The blue crosses show the empirical semivariance of residuals from an initial model estimated with OLS that used the same predictors reported in Table 1.

Finally, the red line in Figure 4 shows the functional form of the Gaussian semivariance function estimated in our full Bayesian model. This line is computed by assuming that the nugget, range, and sill are at their mean values from the posterior distribution. As is the typical case, the semivariance starts lower at more proximate values and rises as distance increases. A low semivariance means that the correlation between observations is high, and similarly a high semivariance implies a low correlation between observations. Our result therefore means that in our forecasting model, the responses of nearby survey respondents will get greater weight in predicting ideology at a particular location than the responses of farther survey respondents.

4 Forecasts of Public Ideology

With a training model of ideology in hand, we now turn to using this model to make forecasts of public opinion throughout electoral constituencies following the point-to-block realignment strategy described earlier. To implement this plan we proceed in four steps. First, we kriged 250,000 simulated citizens. These citizens were located in proportion to the population distribution in 11 million census blocks in 2010. This has the advantage of placing citizens in locations reflective of the true population density, which later on will make it easier to cover legislative districts that are compact in size. For each draw, we started at the centroid of the census block and jittered from the block's midpoint to the extent of the block's radius. This allowed us to place each simulated citizen in eastings and northings. As with the CCES training data, Alaska and Hawaii's census blocks were relocated to sit off of the continental west coast.

Second, once we kriged a simulated citizen, we assigned this citizen covariate values consistent with population data for the location. For each census block we know the block's distribution of age, sex, race, and homeownership, so we draw covariate values for the simulated citizen in proportion to the local distribution. For other covariates, we have to go to a higher level of aggregation, but we always use the most local possible distribution to simulate covariate values. For instance, we simulate education and income based on block group-level

data, and we simulate employment status based on tract-level data. We also simulate religion and urban-rural status with county-level data, using government data besides the 2010 Census (Grammich et al. 2012; United States Department of Agriculture 2013).

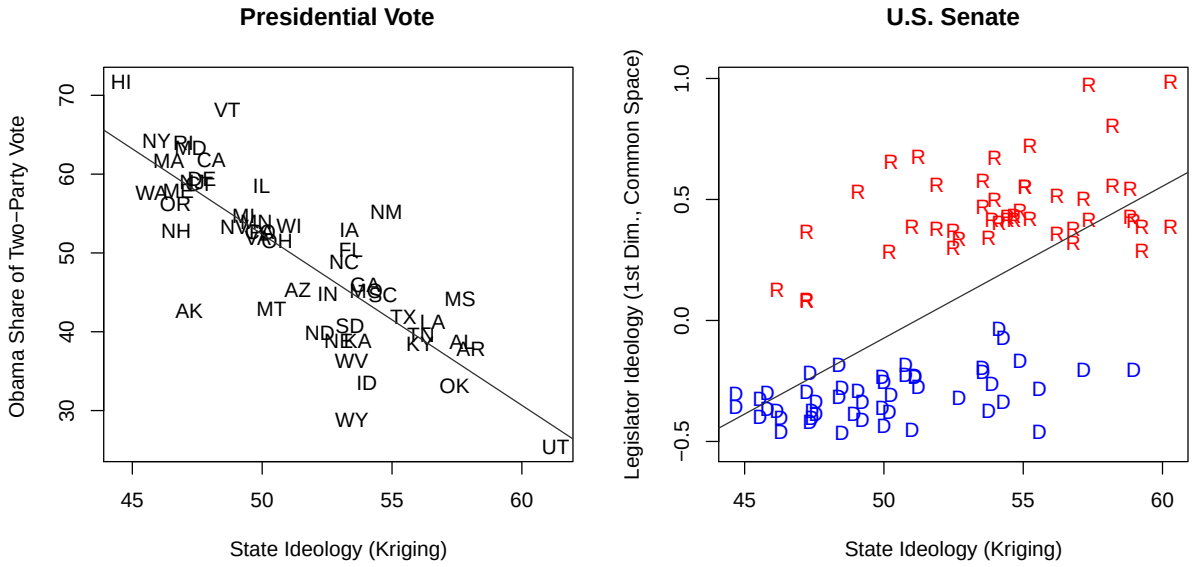
Third, we forecasted ideology for each simulated citizen using the model estimated over the training data. This meant placing all simulated covariate values into the mean model. Additionally, we use the spatial variance process model to predict a spatial error term for each simulated citizen as a weighted combination of the training model residuals, with more proximate training observations getting a higher weight. Fourth, we gathered all simulated citizens falling within a constituency and used the average of their forecasts to compute a district average ideology score. This allowed us to make forecasts for states, congressional districts, upper chambers of state legislatures, and lower chambers of state legislatures.

4.1 Measures of Ideology and Validity Checks

Figure 5 presents our estimates of ideology in all 50 states. In both panels of the figure, the horizontal axis represents our estimates for the average state ideology, with higher values meaning more conservative. In Subfigure 5(a) the vertical axis represents the percentage of the two-party vote that Obama won in 2012. Each state is represented by its two-letter postal code, and the line represents the best fit from a regression that models Obama's vote share as a function of our kriged ideology scores. As the scatterplot and best fit line both show, there is a close relationship between our measures of kriged ideology and presidential vote share, which serves as an external validation of our scores.

Additionally, Subfigure 5(b) illustrates how well our measures of public ideology predict the ideology of U.S. senators elected from these respective states. In this subfigure, the horizontal axis again is our kriged measure of public ideology. The vertical axis the the first dimension score of DW-NOMINATE, which frequently is used as a measure of member ideology (McCarty, Poole and Rosenthal 1997; Poole and Rosenthal 1997). Republicans are represented by a red "R" and Democrats by a blue "D." The line represents a linear regression predicting member's NOMINATE score with public opinion ideology. As can be seen, more conservative states are more likely to elect conservative members and more likely to elect Republicans. Even within party, the scatterplot shows that within-party variance conforms to expectations: moderate Republicans are elected from more liberal states, and moderate Democrats are elected from more conservative states.

Figure 6 turns to the 435 districts for the U.S. House of Representatives and displays our measure of public ideology by district. The horizontal axis displays our measure of public



(a) Obama Vote Share 2012

(b) Senate Ideology 2011

Figure 5: Scatterplots of 2012 presidential vote by state and 2011 U.S. Senators' ideology, each against kriged measure of 2010 state public ideology.

opinion ideology by district, and the vertical axis represents each House member's first dimension DW-NOMINATE score. Again, every Democrat is represented with a blue "D" and every Republican is represented with a red "R." The line shows the results of a regression of elected members' ideology as a function of district ideology. Even at this smaller level of geographic precision, we still see that we can use an electorate's ideology to predict whether those voters will choose a Republican or Democrat and how conservative or liberal the member will be. Again, moderate members of each party tend to be drawn from districts that normally would not elect a member of their respective party. Hence, for both chambers of Congress, we see a relationship between voters' ideology and the ideology of their members. The fact that this well-established electoral connection continues to be supported by our data further validates our kriging approach.

Finally, we applied our kriging technique to constituencies as precise as state legislative districts. Figure 7 illustrates our measures of constituency ideology in both lower and upper chambers of the state legislatures. In both panels, the horizontal axis captures public ideology with our kriging measure, while the vertical axis measures state legislators' ideology with the common space measure developed by Shor and McCarty (2011). In both panels, red dots represent Republican legislators, and blue dots represent Democratic legislators. Districts and legislators for lower chambers are presented in Subfigure 7(a), while upper chambers are

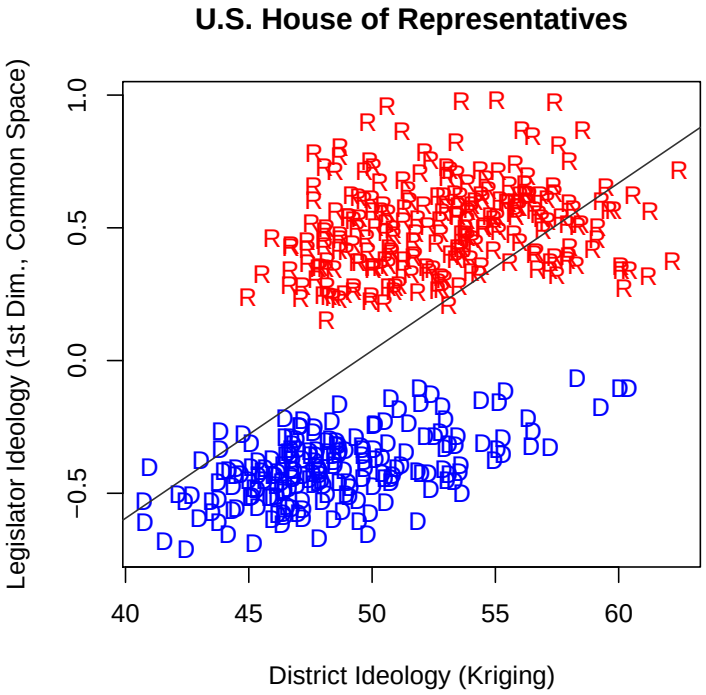
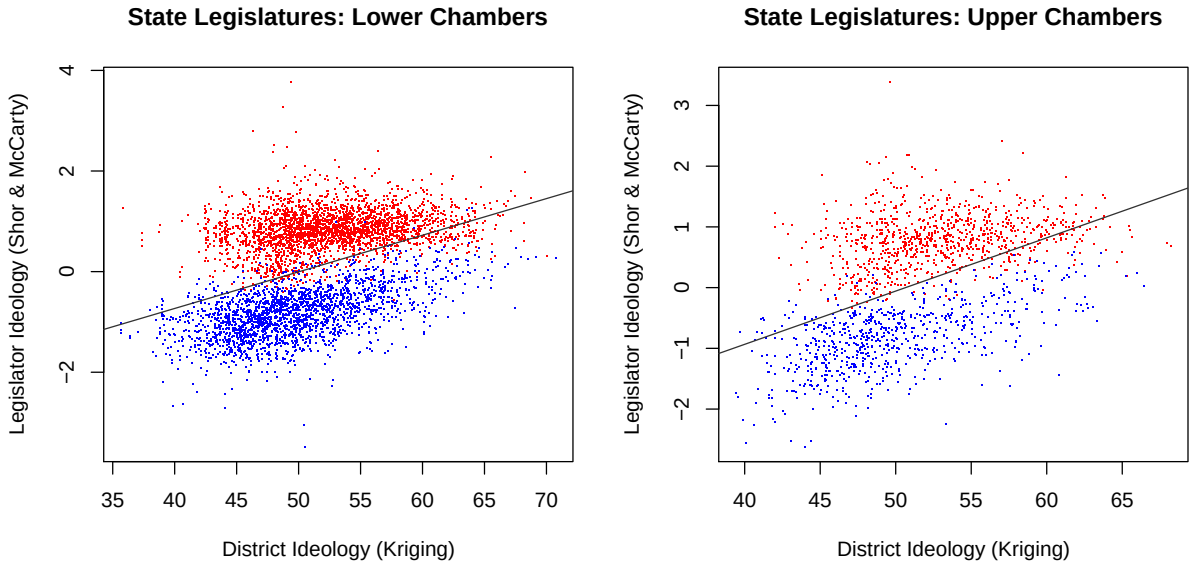


Figure 6: Scatterplot of 2011 U.S. House of Representatives members' ideology against kriged measure of 2010 state public ideology.

presented in Subfigure 7(b). In each panel, the regression line shows a positive association between electoral conservatism and legislator conservatism. So even at this most precise level where many geographic constituencies are no larger than a neighborhood, our measures of public opinion correspond to the electoral connection that we would expect for state legislators. Hence, for many sizes of electoral constituencies, our measures of public ideology pass the external validity checks we consider.

5 Full information Markov Chain Monte Carlo: Meuse River Example

In this section we present an emerging improvement to our estimation strategy using an example dataset of observed zinc levels in the Meuse River floodplain. One of the key limits of our procedure so far is the estimator that we use. Recall from Figure 1 that in step (3.a) we estimate our training model with bootstrap samples from the full data. When we estimate the model with each replicate sample, we apply a five-step algorithm developed by Diggle and



(a) Lower Chambers (b) Upper Chambers

Figure 7: Scatterplots of 2011 state legislators' ideology in lower and upper chambers, each against kriged measure of 2010 state public ideology.

Ribeiro (2002, 141).[3] First, we draw several values from a discrete version of the uniform priors for $\frac{\tau^2}{\sigma^2}$ and $\frac{1}{\phi}$. Second, we estimate the conditional posterior distribution, $p(\frac{\tau^2}{\sigma^2}, \frac{1}{\phi}|\mathbf{Y})$ by placing our draws from the discrete prior into the following formula:

$$p\left(\frac{\tau^2}{\sigma^2}, \frac{1}{\phi}\middle|\mathbf{Y}\right) \propto \pi\left(\frac{\tau^2}{\sigma^2}\right)\pi\left(\frac{1}{\phi}\right)|V_{\tilde{\beta}}|^{\frac{1}{2}}\left|\mathbf{H}(\phi)+\left(\frac{\tau^2}{\sigma^2}\right)\mathbf{I}\right|^{-\frac{1}{2}}(\hat{\sigma}^2)^{-\frac{n}{2}}, \tag{6}$$

where $V_{\tilde{\beta}}$ is the correlation matrix of the regression coefficients estimated with feasible generalized least squares (FGLS) using the current draw of $1/\phi$, n is the sample size, and $\hat{\sigma}^2$ is an estimate of the partial sill based on residuals drawn from the FGLS coefficient estimates.[4] All other terms are defined as before. Third, we draw a single set of sample posterior values for $\frac{\tau^2}{\sigma^2}$ and $\frac{1}{\phi}$ from (6). Fourth, we attach the set of sampled values to $p(\boldsymbol{\beta}, \sigma^2|\frac{\tau^2}{\sigma^2}, \frac{1}{\phi}, \mathbf{Y})$ and compute the corresponding conditional posterior distributions as:

$$\begin{aligned}\sigma^2|\mathbf{Y}, \frac{\tau^2}{\sigma^2}^*, \frac{1}{\phi}^* &\sim \chi^2_{ScI}(n, \hat{\sigma}^2) \\ \boldsymbol{\beta}|\mathbf{Y}, \sigma^2, \frac{\tau^2}{\sigma^2}^*, \frac{1}{\phi}^* &\sim \mathcal{N}(\tilde{\boldsymbol{\beta}}, \sigma^2 V_{\tilde{\beta}})\end{aligned} \tag{7}$$

[3] See also Diggle and Ribeiro (2007, Chapter 7).

[4] Specifically, $\tilde{\beta} = (\mathbf{X}'\mathbf{H}(\phi)^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{H}(\phi)^{-1}\mathbf{Y}$. Hence, $V_{\tilde{\beta}} = (\mathbf{X}'\mathbf{H}(\phi)^{-1}\mathbf{X})^{-1}$. In addition, $\hat{\sigma}^2 = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{H}(\phi)^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$.

The terms in these equations are again drawn from the FGLS estimates from the initial draw of ϕ. After taking a draw from the scaled inverse χ^2 distribution for the partial sill σ^2, this term is linked with the draws from the relative nugget and range terms when drawing the regression coefficients from a normal distribution. By repeating the third and fourth steps to generate a sufficiently large sample from each of the conditional posteriors, we build a sufficient Monte Carlo sample to reflect the joint posterior for the full parameter set $(\frac{\tau^2}{\sigma^2}, \frac{1}{\phi}, \sigma^2, \boldsymbol{\beta})$.

Rather than using this Monte Carlo-FLGS iterative method, a better approach would be a full-information Markov chain Monte Carlo estimator. To do this, we have programmed a new Metropolis-Hastings estimator. To motivate this, we use a real data example. Specifically, our applied example is an analysis of zinc levels in a floodplain of the Meuse River near Stein, The Netherlands (Rikken and Van Rijn 1993).[5] Suppose we specify our model as follows:

$$\begin{aligned}
\mathbf{Y_s} &\sim \mathcal{N}(\mathbf{X_s}\boldsymbol{\beta}, \boldsymbol{\Sigma}) \\
\boldsymbol{\Sigma} &= \sigma^2\mathbf{H}(\phi) + \tau^2\mathbf{I} \\
H(\phi)_{ij} &= \exp(-\phi d_{ij}) \text{ (exponential correlation function)} \\
\pi(\boldsymbol{\beta}) &\sim \mathcal{MVN}(\mathbf{0}, \boldsymbol{\Delta}) \\
\pi(\tau^2) &\sim \mathcal{G}^{-1}(\eta, \nu) \\
\pi(\sigma^2) &\sim \mathcal{G}^{-1}(\alpha, \psi) \\
\pi(\phi) &\sim \mathcal{G}^{-1}(\zeta, \theta)
\end{aligned} \tag{8}$$

In Equation 8, $\mathbf{Y}$ is the dependent variable, which is the logged level of zinc concentration in topsoil (milligrams per kilogram). These data are located at sites $\mathbf{s}$, which are recorded in eastings and northings (both scaled in meters); however, we have rescaled the domain of the data to form a unit square. $\mathbf{X}$ is matrix of predictors, which are also geolocated: In this case a constant and the square root of the number of meters from the location to the Meuse River are the two predictors. $\boldsymbol{\beta}$ is the vector of the two regression coefficients. $\boldsymbol{\Sigma}$ is the variance-covariance matrix of errors, which is structured with nugget τ^2, partial sill σ^2, and decay term ϕ. The correlation of errors follows an exponential correlation function. The regression coefficients, $\boldsymbol{\beta}$ have a multivariate normal prior. The nugget, partial sill, and decay terms each have an inverse gamma prior.

With this model specified for our Meuse River data, we can form the joint posterior distribution of the parameters $(\boldsymbol{\beta}, \tau^2, \sigma^2, \phi)$ by taking the product of the likelihood of $\mathbf{Y_s}$ with

[5]This example was presented in Bivand, Pebesma and Gómez-Rubio (2008, Chapter 8).

the priors of each of these four parameters. That is:

$$\pi(\boldsymbol{\beta}, \sigma^2, \tau^2, \phi|\mathbf{Y}) \propto f(\mathbf{Y}|\boldsymbol{\beta}, \sigma^2, \tau^2, \phi)\pi(\boldsymbol{\beta})\pi(\sigma^2)\pi(\tau^2)\pi(\phi). \tag{9}$$

If we compute this product for the Meuse model of Equation 8 and then take the logarithm of it, then we obtain the following log-posterior distribution:

$$\begin{aligned} \ln(2\pi) - &\frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) + \\ &\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Delta}| - \frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\beta} + \\ &\eta\ln(\nu) - \ln(\Gamma(\eta)) - \eta\ln(\tau^2) - \ln(\tau^2) - \frac{\nu}{\tau^2} + \\ &\alpha\ln(\psi) - \ln(\Gamma(\alpha)) - \alpha\ln(\sigma^2) - \ln(\sigma^2) - \frac{\psi}{\sigma^2} + \\ &\zeta\ln(\theta) - \ln(\Gamma(\zeta)) - \zeta\ln(\phi) - \ln(\phi) - \frac{\theta}{\phi} \end{aligned} \tag{10}$$

Sampling from the log-posterior distribution is computationally simpler than using the original posterior and yields equivalent results.

With the log-posterior distribution of Equation 10 we first proceed by running a hill climber on the posterior distribution. This gives us good initial values and a sense of the uncertainty on each parameter. We then proceed to run several iterations of a Metropolis-Hastings sampler. This proceeds as follows:

1. Treat the estimates of the parameters from the hill climber as starting values: $(\boldsymbol{\beta}_0, \tau_0^2, \sigma_0^2, \phi_0)$

2. At the first iteration, simulate candidate values of the three variance terms $(\tau^{2'}, \sigma^{2'}, \phi')$ each using a draw from a gamma distribution that is based on initial estimates of the respective term's mean and error variance from the hill climber as well as a fixed tuning parameter. Similarly, draw a vector of candidate regression coefficients $(\boldsymbol{\beta}')$ from a multivariate normal distribution based on the hill climber's initial mean and variance-covariance estimates of the regression coefficients.

3. Compare the value of the log-posterior of the new candidate values to the value of the log-posterior with the starting values:

   - If $\ln(\pi(\boldsymbol{\beta}', \sigma^{2'}, \tau^{2'}, \phi'|\mathbf{Y})) > \ln(\pi(\boldsymbol{\beta}_0, \sigma_0^2, \tau_0^2, \phi_0|\mathbf{Y}))$ then accept the candidate vector $(\boldsymbol{\beta}', \sigma^{2'}, \tau^{2'}, \phi')$ as the sample values for the first iteration $(\boldsymbol{\beta}_1, \sigma_1^2, \tau_1^2, \phi_1)$.
   - If $\ln(\pi(\boldsymbol{\beta}', \sigma^{2'}, \tau^{2'}, \phi'|\mathbf{Y})) < \ln(\pi(\boldsymbol{\beta}_0, \sigma_0^2, \tau_0^2, \phi_0|\mathbf{Y}))$ then with probability $\pi(\boldsymbol{\beta}', \sigma^{2'}, \tau^{2'}, \phi'|\mathbf{Y})/\pi(\boldsymbol{\beta}_0, \sigma_0^2, \tau_0^2, \phi_0|\mathbf{Y})$ choose the candidate vector $(\boldsymbol{\beta}', \sigma^{2'}, \tau^{2'}, \phi')$

as the sample values for the first iteration $(\boldsymbol{\beta}_1, \sigma_1^2, \tau_1^2, \phi_1)$. With the complementary probability $1 - \pi(\boldsymbol{\beta}', \sigma^{2\prime}, \tau^{2\prime}, \phi' | \mathbf{Y}) / \pi(\boldsymbol{\beta}_0, \sigma_0^2, \tau_0^2, \phi_0 | \mathbf{Y})$ retain $(\boldsymbol{\beta}_0, \sigma_0^2, \tau_0^2, \phi_0)$ as the sample values for iteration 1.

4. Repeat steps 2 & 3 for each iteration j of the sampler such that candidate parameters $(\boldsymbol{\beta}', \sigma^{2\prime}, \tau^{2\prime}, \phi')$ are compared to the prior iteration's parameters $(\boldsymbol{\beta}_{j-1}, \sigma_{j-1}^2, \tau_{j-1}^2, \phi_{j-1})$ in order to select iteration j's parameter vector $(\boldsymbol{\beta}_j, \sigma_j^2, \tau_j^2, \phi_j)$.

5. Stop at iteration M, a predetermined length of the MCMC chain. It would be advisable to discard several of the early iterations in the chain as burn-in.

In this way, our Metropolis-Hastings algorithm ought to incorporate full information about all parameters when sampling possible parameter values.

To illustrate how this procedure works, we present our results for the Meuse data in Table 2. Recall that the dependent variable is logged zinc levels in the soil. The rows of the table represent the parameters of the model: first the variance-covariance terms of the nugget, decay, and partial sill, and then the two regression coefficients for the constant and the root of distance to the river. Three models are presented: The first two numeric columns show the estimate and standard error from a simple maximum likelihood model that has no priors. The third and fourth numeric columns show the estimate and standard error if the log-posterior distribution is optimized with a hill climber—meaning just the mode of the log-posterior is found. The last two columns of the table show the mean and standard error found when summarizing MCMC results from applying our Metropolis-Hastings algorithm to the log-posterior distribution.

Table 2: Three Objective Function Estimates with Meuse Data

	Log-Likelihood		Posterior Mode		Posterior MCMC	
Parameter	Est.	S.E.	Est.	S.E.	Est.	S.E.
Nugget	0.0448	0.0308	0.0272	0.0366	0.0225	0.0191
Decay	18.2478	6.8334	24.3525	9.2716	25.1224	5.9875
Partial Sill	0.1445	0.0419	0.1547	0.0440	0.1654	0.0272
Intercept	6.9954	0.1206	6.9758	0.1106	6.9745	0.0779
Root of Distance	-2.5823	0.2287	-2.5516	0.2128	-2.5488	0.1503

Notes: N=155. Final column presents summary statistics from MCMC chain.
100,000 iterations (after a burn-in of 10,000 iterations)
with an acceptance rate of 22%.

As Table 2 shows, for the regression coefficients the estimates are all pretty similar. For both the intercept and the coefficient on the square root of distance, the second and third

model (both of which use the log-posterior instead of the log likelihood) report coefficients that are slightly closer to zero, but hardly different enough to be noticed. Namely, the negative coefficient for square root of distance consistently hovers around -2.55 with a small standard error, indicating that sites farther from the Meuse river on average have less zinc in the soil. With the three variance-covariance terms, however, the story is different. The maximum likelihood model presented first has a much bigger nugget, a much smaller decay term, and a somewhat smaller partial sill term than the two models using the log-posterior. This is notable because it implies that the log-likelihood model is attributing a larger share of the error variance to non-spatial causes in the nugget. Meanwhile the bigger decay terms in the log-posterior models specify a larger degree of spatial correlation among observations. This is an important shift in the unexplained variance, and in a case of forecasting with kriging the spatial variance terms are essential to making predictions. As a final and important point, the smallest standard errors by far emerge from the Metropolis-Hastings sampler, shown in the last column. While the standard errors are not grossly different among the three models, actually summarizing the iterations of a Markov chain shows the tightest level of uncertainty.

In this section, we have described how our newly programmed Metropolis-Hastings algorithm works for point-referenced geospatial models, and we have applied it to a model of zinc deposits near the Meuse River. Our results are similar to what can be found with maximum likelihood, albeit with smaller standard errors and an error variance structure that places more weight on geospatial correlation. With this test example of 155 observations, we believe we have demonstrated that our new sampler works well and can be expanded to other applications, such as our project of forecasting ideology.

6 Current Challenges

As the previous section may imply, we are still working through some challenges on this project. To start, our Metropolis-Hastings algorithm for full-information MCMC is fairly new, and we would like to use our own estimator when modeling and forecasting ideology. As a next step, we would like to apply our Metropolis estimator to a subset of our ideology dataset, such as all CCES observations from New York City (which has a manageable sample size of 568 respondents). Once we have successfully completed a test run with those data, we would like to use the Metropolis sampler in step (3.a) from Figure 1. That is, we would like to continue with the bootstrap procedure but use our estimator instead of the FGLS-Monte Carlo hybrid.

As another point on step (3.a), we would like to demonstrate with Monte Carlo analyses

how effective this bootstrap procedure is and when it performs better or worse. With big data, such as the 21,849 observations in the CCES, kriging models become computationally intensive. All pairwise distances must be computed, and then correlations between each pair of observations based on the distance and the value of the decay parameter (ϕ) must be computed. Once all of this goes into a variance-covariance matrix of errors, that matrix must be inverted. Importantly, in a Bayesian model this step must be completed *with each iteration.* This is a big computation ask, so bootstrapping the sample makes the process considerably more feasible. With the Monte Carlos, we hope to show applied analysts how to krige with big data.

As a final point we are facing, the last step of the forecasting process can be computationally intensive as well. When making forecasts for kriged points, for each point in geographic space where a forecast of ideology is being made, the distance from that point to all observed points in the training data must be computed. Then, using the values of the spatial variance-covariance terms (σ^2, ϕ), all of the training errors are given weights to predict a kriged error at the new point. So for 250,000 kriged points and 21,849 training observations this produces over 5.4 billion distance calculations. Plus, to take full advantage of the Bayesian approach, for these 5.4 billion distances, 5.4 billion weights must be computed *for each iteration* of the MCMC sampler. That allows the analyst to obtain uncertainty over the forecast.

Since that computational problem is wholly infeasible, we propose replacing the stage of recomputing distances and weights with each forecast by instead finding a *smoother function.* A smoother function can use a tangible number of parameters to reproduce with accuracy what the kriged errors would be without making reference to the original training data. In particular, thin plate splines are a promising avenue through which we can define the forecasts purely with a multidimensional trend function and distances to a feasible number of knot points (Wood 2003). The problem we have encountered so far is that existing programs for thin plate splines either: (A) do not allow us to extract parameters in order to make forecasts without referencing the original data, (B) do not allow the model to be fitted over a set of knot points that is a subset of the data, or (C) have an error in a distance scaling algorithm. We may have to program our own thin plate spline function to work around this and make the final forecasting stage more feasible.

7 Implications for the Applied Researcher

In this paper, we have described and implemented the method of Bayesian universal kriging as a way of using survey responses to forecast public opinion in electoral constituencies. Using the 2008 CCES and the 2010 Census, we have created measures of public opinion for the year 2010 at several levels. In doing so, we have improved on past work with this method by correcting a problem of misalignment between ZIP codes and ZCTAs, and we also have found a means of incorporating Alaska and Hawaii into this type of measure. Using presidential vote share and measures of legislative ideology, we verify that this measure behaves as it ought to relative to other established measures in American politics.

Our resulting measures are now freely available for any researcher to use in his or her own analysis. These new measures serve the practical researcher in several ways: First, by releasing a measure for 2010 based on the most modern Census data, our measures are more recent than many alternative measures, even measures taken for the state level. Second, our measures capture ideology at multiple levels, serving as a means of capturing public sentiment not only for the 50 states, but also congressional and state legislative districts. We know of no other measure besides presidential vote share (which itself can be problematic, per Kernell 2009) that is available for all state legislative districts, so we have filled an substantial gap in the measurement of political ideology at low levels of aggregation.

The approach of point-to-block realignment with universal kriging has the potential to fill public opinion measurement needs in many ways. To start, the realignment of kriged points into constituencies need not be to existing legislative districts. A natural extension of this would be to allow users to draw hypothetical districts and extract public opinion in the proposed new district—which would have applications for state legislative and congressional redistricting. Another extension would be to expand this technique to allow ordinal responses from the survey respondents, such as when a public opinion question is asked on a three, five, or seven-point scale. Doing this would open up the possibility of forecasting ideology at the four levels we consider in more years (when only limited versions of ideology questions are available), and it would also allow for the creation of issue-specific public opinion measures based on questions of this type. Finally, the modeled outcome does not necessarily need to be ideology, any surveyed attitude with geocoded response is possible. Our approach can even be applied to epidemioligical outcomes. For now, we have produced measures that researchers can use for state-level, congressional-level, and state legislative-level research. However, we believe there is an even more promising research agenda with Bayesian kriging that will enable even better measures over time, space, and issue area.

A Appendix: Finding Ideological Neighbors for Alaska and Hawaii

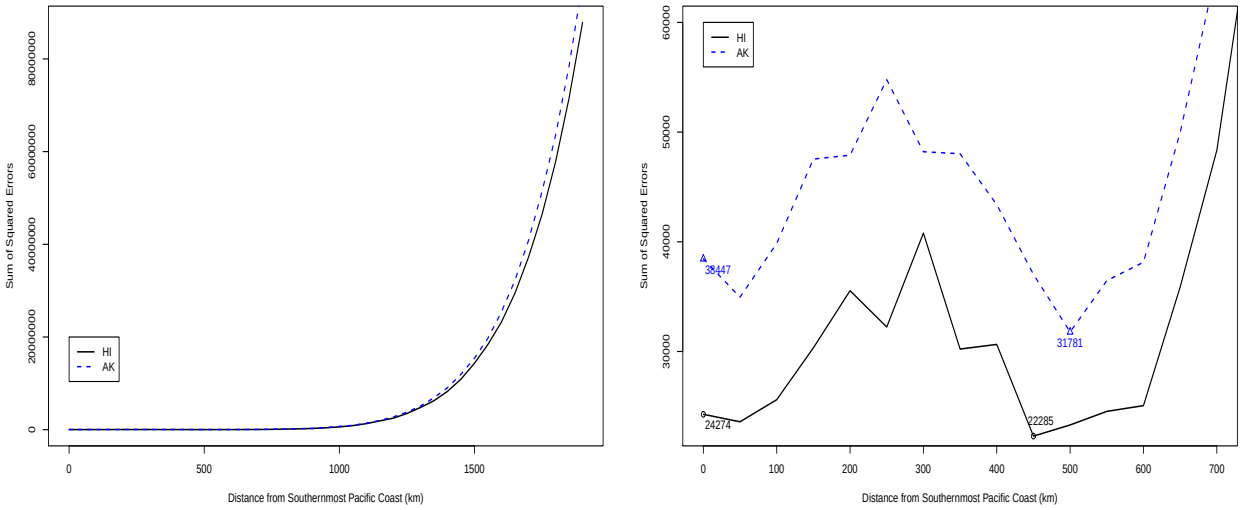
Since Alaska and Hawaii are geographic outliers relative to the continental 48 states, it could pose problems to estimate the model and forecast ideology with the states located as they are. Such geographic outliers could exert substantial leverage over both the geographic trend term and the smoothed error structure. Yet, we do want to forecast ideology for these two states and these constituencies. As a result, we did a search for the states' *ideological neighbors*. The idea behind finding ideological neighbors is that we can determine what geographical locales on the Pacific Coast work well as neighbors for the purposes of forecasting.

We searched for ideological neighbors for Alaska and Hawaii as follows: First, we estimated a preliminary point-referenced data model using OLS regression and an error structure on those residuals. In this initial model, we exclude Alaska and Hawaii from the training data. Second, we consider a variety of locations along the Pacific Coast that might serve as ideological neighbors for each state. For each possible location of the state, we compute the sum of squared errors if we forecasted ideology in the CCES data for each state using the model that was trained over the continental 48 states. We chose each state's location based on which locale minimized the sum of squared errors.

For each state, we considered a central block and started out by placing the state so that its center was as the same northing as the southernmost point on the U.S. Pacific Coast. For Alaska we set the eastings so that the easternmost point in Alaska could never overlap with the westernmost point in the continental states. For Hawaii, we started off by setting the eastings so that the easternmost point in Hawaii would not overlap with the westernmost point of the continental states, but then moved Hawaii 75 kilometers closer: given how the best fits for each state work this adjustment prevents Hawaii from sitting on top of either the continental coast or atop Alaska's new ideological neighbor position.

Figure 8 illustrates our comparative fit on this process. On each panel the horizontal axis represents the distance from the southernmost point on the continental Pacific Coast, with 0 representing the point at the far south and larger numbers indicating kilometers northward from there. The vertical axis on each panel represents the sum of squared errors (SSE) for the out-of-sample predictions to that state from the continental model. The black solid line represents the SSE for Hawaii at each position, and the dashed blue line represents the SSE for Alaska at each position. The left panel shows all positions along the Pacific Coast, which were considered in 50km increments. As that panel shows, 1,000km from the southernmost point

the SSE starts rapidly increasing for both states, indicating that the fit becomes remarkably bad if we place either state in the northwest. The right panel therefore focuses on the southern side of the Pacific Coast, ranging from the southernmost point at 0km to 700km north of that. For Alaska we see that the smallest SSE emerges 500km north of the U.S.-Mexico border, and for Hawaii the smallest SSE emerges 450km north of the border.



(a) Full Range of Distance

(b) Range of 0–700km

Figure 8: Sum of squared errors for forecasts of Alaska and Hawaii data for positions along the Pacific Coast.

Using these placements, Figure 2 in the main text plots the centroids of the census blocks we use for forecasting. The blocks for Alaska and Hawaii are now placed in their new ideological neighbor locations based on the best out-of-sample fit. Hawaii is placed so that its central census block is 450km north of the U.S.-Mexico border. Substantively, this places the island of Oahu off the coast of the San Francisco Bay, meaning that Honolulu, HI and San Francisco, CA are ideological neighbors. Alaska is placed so that its central census block is 500km north of the border. Substantively, this means that Juneau, AK is situated just south of San Diego, CA, which also puts the capital city as close as possible to Arizona and Texas for making forecasts of ideology. Anchorage, AK, meanwhile is a bit north of Santa Barbara, CA. Moving forward, when we forecast ideology for Alaska, Hawaii, and each state's respective legislative districts we use these new ideological neighbor locations.

B Appendix: Data Sources

- **Survey data of individual ideology in 2008:** The Cooperative Congressional Election Survey, Common Content, 2008. Accessed from `http://hdl.handle.net/1902.1/14003` on April 18, 2013 (Ansolabehere 2011).

- **Population demographic data in 2010:** U.S. Census 2010 block, block group, and tract-level data. Dataset `2010_SF1a` accessed from the National Historical Geographic Information System, `https://www.nhgis.org` on October 13, 2015 (Minnesota Population Center 2011).

- **Census block centroids and area:** U.S. Census TIGER shapefiles for 2010. Accessed from `http://www2.census.gov/geo/tiger/TIGER2010BLKPOPHU/` on December 12, 2015.

- **ZIP code centroids and area:** USA ZIP Code Areas, 2014. TomTom data held by ArcGIS. Accessed from `https://www.arcgis.com/home/item.html?id=8d2012a2016e484dafaac0451f9aea24` on September 2, 2015.

- **Religious affiliation by county:** The 2010 Religious Congregations and Membership Study, provided by the ARDA. Accessed from `http://www.thearda.com/Archive/Files/Descriptions/RCMSCY10.asp` on October 15, 2015 (Grammich et al. 2012).

- **Urban-rural classification continuum by county:** United States Department of Agriculture, 2013 measure. Accessed from `http://www.ers.usda.gov/data-products/rural-urban-continuum-codes.aspx#.UfCK11OE4xc` on October 15, 2015 (United States Department of Agriculture 2013).

- **Ideology of members of Congress:** Subset of Carroll, Lewis, Lo, McCarty, Poole, & Rosenthal's Common Space DW-NOMINATE scores. Restricted to first dimension for 112th Congress. Accessed from `http://voteview.com` on December 20, 2015 (McCarty, Poole and Rosenthal 1997; Poole and Rosenthal 1997).

- **Ideology of state legislators:** Subset of Shor & McCarty's June 2015 update of individual state legislator database, focusing strictly on 2011 scores. Accessed from `http://dx.doi.org/10.7910/DVN/THDBRA` on December 19, 2015 (Shor and McCarty 2011).

C Appendix: Crosswalk Between Census State Legislative District Names and Full Names, New Hampshire and Vermont

The 2010 Census files keep track of the state legislative and congressional districts that each census block falls in. Congressional districts follow a numbering scheme or are easily identified as a single at-large district. Most state legislative districts also follow a numbering scheme. All of this allows for merging with information such as legislator ideology scores.

As an important exception, the states of New Hampshire and Vermont name state legislative districts based on the county they are in. For the sake of file size, the names of state legislative districts are abbreviated in the Census files, though files such as the data by Shor and McCarty (2011), use the full name of the district. For the sake of facilitating future merging work, we list a crosswalk that ties the abbreviations to the full district names. Table 3 lists districts for the New Hampshire House of Representatives, Table 4 lists districts for the Vermont House of Representatives, and Table 5 lists districts for the Vermont Senate.

Table 3: Crosswalk between Names of State Legislative Districts and Census Bureau Abbreviations in the New Hampshire House of Representatives

Census Code	District Name	Census Code	District Name	Census Code	District Name
101	BELKNAP 1	617	HILLSBOROUGH 17	001	SULLIVAN 1
102	BELKNAP 2	618	HILLSBOROUGH 18	002	SULLIVAN 2
103	BELKNAP 3	619	HILLSBOROUGH 19	003	SULLIVAN 3
104	BELKNAP 4	620	HILLSBOROUGH 20	004	SULLIVAN 4
105	BELKNAP 5	621	HILLSBOROUGH 21	005	SULLIVAN 5
106	BELKNAP 6	622	HILLSBOROUGH 22		
201	CARROLL 1	623	HILLSBOROUGH 23		
202	CARROLL 2	624	HILLSBOROUGH 24		
203	CARROLL 3	625	HILLSBOROUGH 25		
204	CARROLL 4	626	HILLSBOROUGH 26		
205	CARROLL 5	627	HILLSBOROUGH 27		
301	CHESHIRE 1	701	MERRIMACK 1		
302	CHESHIRE 2	702	MERRIMACK 2		
303	CHESHIRE 3	703	MERRIMACK 3		
304	CHESHIRE 4	704	MERRIMACK 4		
305	CHESHIRE 5	705	MERRIMACK 5		
306	CHESHIRE 6	706	MERRIMACK 6		
307	CHESHIRE 7	707	MERRIMACK 7		
401	COOS 1	708	MERRIMACK 8		
402	COOS 2	709	MERRIMACK 9		
403	COOS 3	710	MERRIMACK 10		
404	COOS 4	711	MERRIMACK 11		
501	GRAFTON 1	712	MERRIMACK 12		
502	GRAFTON 2	713	MERRIMACK 13		
503	GRAFTON 3	801	ROCKINGHAM 1		
504	GRAFTON 4	802	ROCKINGHAM 2		
505	GRAFTON 5	803	ROCKINGHAM 3		
506	GRAFTON 6	804	ROCKINGHAM 4		
507	GRAFTON 7	805	ROCKINGHAM 5		
508	GRAFTON 8	806	ROCKINGHAM 6		
509	GRAFTON 9	807	ROCKINGHAM 7		
510	GRAFTON 10	808	ROCKINGHAM 8		
511	GRAFTON 11	809	ROCKINGHAM 9		
501	GRAFTON 1	810	ROCKINGHAM 10		
601	HILLSBOROUGH 1	811	ROCKINGHAM 11		
602	HILLSBOROUGH 2	812	ROCKINGHAM 12		
603	HILLSBOROUGH 3	813	ROCKINGHAM 13		
604	HILLSBOROUGH 4	814	ROCKINGHAM 14		
605	HILLSBOROUGH 5	815	ROCKINGHAM 15		
606	HILLSBOROUGH 6	816	ROCKINGHAM 16		
607	HILLSBOROUGH 7	817	ROCKINGHAM 17		
608	HILLSBOROUGH 8	818	ROCKINGHAM 18		
609	HILLSBOROUGH 9	901	STRAFFORD 1		
610	HILLSBOROUGH 10	902	STRAFFORD 2		
611	HILLSBOROUGH 11	903	STRAFFORD 3		
612	HILLSBOROUGH 12	904	STRAFFORD 4		
613	HILLSBOROUGH 13	905	STRAFFORD 5		
614	HILLSBOROUGH 14	906	STRAFFORD 6		
615	HILLSBOROUGH 15	907	STRAFFORD 7		
616	HILLSBOROUGH 16				

Table 4: Crosswalk between Names of State Legislative Districts and Census Bureau Abbreviations in the Vermont House of Representatives

Census Code	District Name	Census Code	District Name
A-1	ADDISON-1	L-2	LAMOILLE-2
A-2	ADDISON-2	L-3	LAMOILLE-3
A-3	ADDISON-3	L-4	LAMOILLE-4
A-4	ADDISON-4	LW1	LAMOILLE-WASHINGTON-1
A-5	ADDISON-5	OG1	ORANGE-1
AR1	ADDISON-RUTLAND-1	OG2	ORANGE-2
B-1	BENNINGTON-1	OA1	ORANGE-ADDISON-1
B21	BENNINGTON-2-1	OGC	ORANGE-CALEDONIA-1
B22	BENNINGTON-2-2	OL1	ORLEANS-1
B-3	BENNINGTON-3	OL2	ORLEANS-2
B-4	BENNINGTON-4	OLC	ORLEANS-CALEDONIA-1
B-5	BENNINGTON-5	OLF	ORLEANS-FRANKLIN-1
BR1	BENNINGTON-RUTLAND-1	R11	RUTLAND-1-1
CA1	CALEDONIA-1	R12	RUTLAND-1-2
CA2	CALEDONIA-2	R-2	RUTLAND-2
CA3	CALEDONIA-3	R-3	RUTLAND-3
CA4	CALEDONIA-4	R-4	RUTLAND-4
CAW	CALEDONIA-WASHINGTON-1	R51	RUTLAND-5-1
C11	CHITTENDEN-1-1	R52	RUTLAND-5-2
C12	CHITTENDEN-1-2	R53	RUTLAND-5-3
C-2	CHITTENDEN-2	R54	RUTLAND-5-4
C31	CHITTENDEN-3-1	R-6	RUTLAND-6
C32	CHITTENDEN-3-2	R-7	RUTLAND-7
C33	CHITTENDEN-3-3	R-8	RUTLAND-8
C34	CHITTENDEN-3-4	RY1	RUTLAND-WINDSOR-1
C35	CHITTENDEN-3-5	W-1	WASHINGTON-1
C36	CHITTENDEN-3-6	W-2	WASHINGTON-2
C37	CHITTENDEN-3-7	W31	WASHINGTON-3-1
C38	CHITTENDEN-3-8	W32	WASHINGTON-3-2
C39	CHITTENDEN-3-9	W33	WASHINGTON-3-3
C35	CHITTENDEN-3-10	W-4	WASHINGTON-4
C-4	CHITTENDEN-4	W-5	WASHINGTON-5
C51	CHITTENDEN-5-1	W-6	WASHINGTON-6
C52	CHITTENDEN-5-1	W-7	WASHINGTON-7
C61	CHITTENDEN-6-1	WC1	WASHINGTON-CHITTENDEN-1
C62	CHITTENDEN-6-2	X-1	WINDHAM-1
C63	CHITTENDEN-6-3	X-2	WINDHAM-2
C71	CHITTENDEN-7-1	X31	WINDHAM-3-1
C72	CHITTENDEN-7-2	X32	WINDHAM-3-2
C-8	CHITTENDEN-8	X33	WINDHAM-3-3
C-9	CHITTENDEN-9	X-4	WINDHAM-4
EC1	ESSEX-CALEDONIA	X-5	WINDHAM-5
EC2	ESSEX-CALEDONIA-ORLEANS	X-6	WINDHAM-6
F-1	FRANKLIN-1	XB1	WINDHAM-BENNINGTON-1
F-2	FRANKLIN-2	XBY	WINDHAM-BENNINGTON-WINDSOR-1
F-3	FRANKLIN-3	Y11	WINDSOR-1-1
F-4	FRANKLIN-4	Y12	WINDSOR-1-2
F-5	FRANKLIN-5	Y-2	WINDSOR-2
F-6	FRANKLIN-6	Y-3	WINDSOR-3
GC1	GRAND ISLE-CHITTENDEN-1-1	Y-4	WINDSOR-4
L-1	LAMOILLE-1	Y-5	WINDSOR-5
		Y61	WINDSOR-6-1
		Y62	WINDSOR-6-2
		YO1	WINDSOR-ORANGE-1
		YO2	WINDSOR-ORANGE-2
		YR1	WINDSOR-RUTLAND-1
		YR2	WINDSOR-RUTLAND-2

Table 5: Crosswalk between Names of State Legislative Districts and Census Bureau Abbreviations in the Vermont Senate

Census Code	District Name	Census Code	District Name
ADD	ADDISON	LAM	LAMOILLE
BEN	BENNINGTON	ORA	ORANGE
CAL	CALEDONIA	RUT	RUTLAND
CHI	CHITTENDEN	WAS	WASHINGTON
E-O	ESSEX-ORLEANS	WDM	WINDHAM
FRA	FRANKLIN	WSR	WINDSOR
CGI	GRAND ISLE		

References

Ansolabehere, Stephen. 2011. "CCES, Common Content, 2008." *http://hdl.handle.net/1902.1/14003* Ver. 4.

Ansolabehere, Stephen D., James M. Snyder and Charles Stewart. 2001. "Candidate Positioning in U.S. House Elections." *American Journal of Political Science* 45:136–159.

Banerjee, Sudipto, Bradley P. Carlin and Alan E. Gelfand. 2015. *Hierarchical Modeling and Analysis for Spatial Data*. 2nd ed. New York: Chapman & Hall/CRC.

Berry, William D., Evan J. Ringquist, Richard C. Fording and Russell L. Hanson. 1998. "Measuring Citizen and Government Ideology in the American States, 1960-93." *American Journal of Political Science* 42:327–348.

Beyer, Kirsten M.M., Alan F. Schultz and Gerard Rushton. 2008. Using ZIP Codes as Geocodes in Cancer Research. In *Geocoding Health Data: The Use of Geographic Codes in Cancer Prevention and Control, Research, and Practice*, ed. Gerard Rushton, Marc P. Armstrong, Josephine Gittler, Barry R. Greene, Claire E. Pavlik, Michele M. West and Dale L. Zimmerman. New York, NY: CRC Press pp. 37–68.

Bivand, Roger S., Edzer J. Pebesma and Virgilio Gómez-Rubio. 2008. *Applied Spatial Data Analysis with R*. New York: Springer.

Cressie, Noel A. C. 1993. *Statistics for Spatial Data*. Revised ed. New York: Wiley.

DeLeon, Richard E. and Katherine C. Naff. 2004. "Identity Politics and Local Political Culture: Some Comparative Results from the Social Capital Benchmark Survey." *Urban Affairs Review* 39(6):689–719.

Diggle, Peter J. and Paulo J. Ribeiro, Jr. 2002. "Bayesian Inference in Gaussian Model-based Geostatistics." *Geographical & Environmental Modeling* 6(2):129–146.

Diggle, Peter J. and Paulo J. Ribeiro, Jr. 2007. *Model-based Geostatistics*. New York: Springer.

Djupe, Paul A. and Anand E. Sokhey. 2011. "Interpersonal Networks and Democratic Politics." *PS: Political Science & Politics* 44(1):55–59.

Elazar, Daniel J. 1966. *American Federalism: A View from the States*. New York: Thomas Y. Crowell.

Erikson, Robert S. and Gerald C. Wright. 1980. "Policy Representaton of Constituency Interests." *Political Behavior* 2:91–106.

Erikson, Robert S., Gerald C. Wright and John P. McIver. 1993. *Statehouse Democracy: Public Opinion and Policy in the American States*. New York: Cambridge University Press.

Fischer, David Hackett. 1989. *Albion's Seed: Four British Folkways in America*. New York: Oxford University Press.

Garreau, Joel. 1981. *The Nine Nations of North America*. Boston: Houghton Mifflin.

Gastil, Raymond D. 1975. *Cultural Regions of the United States*. Seattle: University of Washington Press.

Gelman, Andrew and Thomas C. Little. 1997. "Poststratification into Many Categories Using Hierarchical Logistic Regression." *Survey Methodology* 23:127–135.

Gimpel, James G. and Jason E. Schuknecht. 2003. *Patchwork Nation: Sectionalism and Political Change in American Politics*. Ann Arbor, MI: University of Michigan Press.

Grammich, Clifford, Kirk Hadaway, Richard Houseal, Dale E. Jones, Alexei Krindatch, Richie Stanley and Richard H. Taylor. 2012. *2010 U.S. Religion Census: Religious Congregations & Membership Study*. Association of Statisticians of American Religious Bodies.

Grubesic, Tony H. 2008. "Zip Codes and Spatial Analysis: Problems and Prospects." *Socio-Economic Planning Sciences* 42(2):129–149.

Grubesic, Tony H. and Timothy C. Matisziw. 2006. "On the Use of ZIP Codes and ZIP Code Tabulation Areas (ZCTAs) for the Spatial Analysis of Epidemiological Data." *International Journal of Health Geographics* 5:58.

Huckfeldt, Robert and John T. Sprague. 1995. *Citizens, Politics, and Social Communication: Influence in an Election Campaign*. New York: Cambridge University Press.

Jackson, John E. 1989. "An Errors in Variables Approach to Estimating Models with Small Area Data." *Political Analysis* 1:157–180.

Jackson, John E. 2008. Endogeneity and Structural Equation Estimation in Political Science. In *The Oxford Handbook of Political Methodology*, ed. Janet M. Box-Steffensmeier, Henry E. Brady and David Collier. New York: Oxford University Press.

Kernell, Georgia. 2009. "Giving Order to Districts: Estimating Voter Distributions with National Election Returns." *Political Analysis* 17:215–235.

Lax, Jeffrey R. and Justin H. Phillips. 2009. "How Should We Estimate Opinion in the States?" *American Journal of Political Science* 53:107–121.

Lieske, Joel. 1993. "Regional Subcultures of the United States." *Journal of Politics* 55(4):888–913.

McCarty, Nolan M., Keith T. Poole and Howard Rosenthal. 1997. *Income Redistribution and the Realignment of American Politics*. American Enterprise Institute Studies on Understanding Economic Inequality. Washington: AEI Press.

Minnesota Population Center. 2011. *National Historic Geographic Information System: Version 2.0* [Machine-readable database]. Minneapolis, MN: University of Minnesota.

Monogan, III, James E., David M. Konisky and Neal D. Woods. 2017. "Gone with the Wind: Federalism and the Strategic Location of Air Polluters." *American Journal of Political Science* 61(2):257–270.

Monogan, III, James E. and Jeff Gill. 2016. "Measuring State and District Ideology with Spatial Realignment." *Political Science Research and Methods* 4(1):97–121.

Park, David K., Andrew Gelman and Joseph Bafumi. 2004. "Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls." *Political Analysis* 12:375–385.

Park, David K., Andrew Gelman and Joseph Bafumi. 2006. State-Level Opinions from National Surveys: Poststratification Using Multilevel Logistic Regression. In *Public Opinion in State Politics*, ed. Jeffrey E. Cohen. Stanford: Stanford University Press.

Pool, Ithiel de Sola, Robert P. Abelson and Samuel L. Popkin. 1965. *Candidates, Issues, and Strategies*. Cambridge, MA: MIT Press.

Poole, Keith T. and Howard Rosenthal. 1997. *Congress: A Political-Economic History of Roll Call Voting*. New York: Oxford University Press.

Putnam, Robert D. 1966. "Political Attitudes and the Local Community." *American Political Science Review* 60(3):640–654.

Putnam, Robert D. 1993. *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton, NJ: Princeton University Press.

Ravishanker, Nalini and Dipak K. Dey. 2002. *A First Course in Linear Model Theory*. Boca Raton, FL: Chapman & Hall/CRC.

Rikken, M.G.J. and R.P.G. Van Rijn. 1993. "Soil Pollution with Heavy Metals. An Inquiry into the Spatial Variation, Cost of Mapping and Risk-Evaluation of Copper, Cadmium, Lead and Zinc in the Floodplains of the Meuse West of Stein, the Netherlands." Internal M.Sc. thesis, Department of Physical Geography, Utrecht University.

Shor, Boris and Nolan McCarty. 2011. "The Ideological Mapping of American Legislatures." *American Political Science Review* 105(3):530–551.

Sinclair, Betsy. 2012. *The Social Citizen: Peer Networks and Political Behavior*. Chicago: University of Chicago Press.

Tam Cho, Wendy K. and James G. Gimpel. 2007. "Prospecting for (Campaign) Gold." *American Journal of Political Science* 51(2):255–268.

Tausanovitch, Chris and Christopher Warshaw. 2013. "Measuring Constituent Policy Preferences in Congress, State Legislatures, and Cities." *Journal of Politics* 75(2):330–342.

Tobler, Waldo R. 1970. "A Computer Movie Simulating Urban Growth in the Detroit Region." *Economic Geography* 46(2):234–240.

United States Department of Agriculture. 2013. "2013 Rural-Urban Continuum Codes." *http://www.ers.usda.gov/data-products/rural-urban-continuum-codes.aspx* .

Weber, Ronald E., Anne H. Hopkins, Michael L. Mezey and Frank Munger. 1972. "Computer Simulation of State Electorates." *Public Opinion Quarterly* 36:549–565.

Weber, Ronald E. and William R. Shaffer. 1972. "Public Opinion and American State Policy-Making." *Midwest Journal of Political Science* 16:683–699.

Wood, Simon N. 2003. "Thin Plate Regression Splines." *Journal of the Royal Statistical Society, Series B (Statistical Methodoogy)* 65(1):95–114.