Estimating Constituency Preferences from Sparse Survey Data Using Auxiliary Geographic Information

Peter Selb*
Simon Munzert*

*University Konstanz, Department of Politics and Management

http://www.uni-konstanz.de/cms
Estimating Constituency Preferences from Sparse Survey Data Using Auxiliary Geographic Information

Peter Selb
Simon Munzert

Department of Politics and Management and Center for Quantitative Methods and Survey Research University of Konstanz

ABSTRACT

Measures of constituency preferences are of vital importance for the study of political representation and other research areas. Yet, such measures are often difficult to obtain. Previous survey-based estimates frequently lack precision and coverage due to small samples, rely on questionable assumptions, or require detailed auxiliary information about the constituencies’ population characteristics. We propose an alternative Bayesian hierarchical approach that exploits minimal geographic information readily available from digitalized constituency maps. If at hand, social background data is easily integrated. To validate the method, we use national polls and district-level results from the 2009 German Bundestag election, an empirical case for which detailed structural information is missing.
1. INTRODUCTION

Constituency preferences, commonly conceptualized in terms of the distribution of the voters’ issue attitudes or value orientations within electorally relevant geographical units, are central for studying classic questions about democratic representation. For example, does the roll-call behavior of legislators reflect their constituents’ policy views? Under what political and institutional conditions? What will be the electoral consequences if legislators fail to do so? Providing an empirical answer to such questions inevitably necessitates measures of constituency preferences. Yet, such measures are often difficult to obtain. Previous measures of constituency preferences based on aggregate proxies of public opinion have been criticized both in validity and specificity terms (see Jackson and King, 1989). Direct survey measures either lack precision due to small samples and do not provide estimates for off-sample constituencies (e.g., Miller and Stokes, 1963), or pool surveys over long time periods and therefore have to assume temporally stable preferences (e.g., Erikson, Wright and McIver, 1993). Early model-based approaches that use census in addition to survey data involve cross-level inferences that rest on strong assumptions (see Seidman, 1975). Current hierarchical estimators of constituency preferences have been demonstrated to outperform previous approaches (e.g., Park, Gelman and Bafumi, 2004), but require detailed structural information that will often be missing, for example, if the boundaries of the geographic units of interest (e.g., electoral districts) do not coincide with the boundaries of the administrative units for which this information is collected or provided.

In this paper, we propose an alternative Bayesian hierarchical estimation strategy that exploits auxiliary geographic information about the constituencies’ neighborhood structure and surface area, which is readily available from digitalized constituency maps alone. If at hand, social background information is easily integrated in a manner compatible with other state-of-the-art strategies to estimating constituency preferences (in particular, Park, Gelman and Bafumi, 2004). In doing so, we draw on (and call attention to) recent developments in small area estimation (SAE) – a field that covers a variety of methods used to produce survey-based estimates for geographical areas in which the sample sizes are too small to provide reliable direct estimates (see

\begin{footnote}
\textsuperscript{1}Alternative definitions of constituency preferences emphasize functional, not geographical, criteria, which of these definitions is more appropriate seems to not least dependent on the electoral structure: geographical conceptions of constituency preferences are commonly used in the single-member district plurality (SMP) context, while functional conceptions are more prevalent in connection with proportional representation (PR) electoral systems (see Powell, 2009).
\end{footnote}
Rao, 2005), but has been largely ignored by scholars of representation.

To validate our method, we estimate district-level party vote shares from two post-election surveys conducted after the German Bundestag election in 2009 in order to compare them to official election results. Germany is a particularly relevant case, since detailed census data as required by alternative approaches is not available – partly for structural reasons, i.e., non-nested administrative and electoral geographies, partly due to political sensitivities arguably dating back to the massive abuses of census data during the Nazi regime (see Luebke and Milton, 1994).

2. PREVIOUS APPROACHES

This section reviews previous methods of estimating constituency preferences from survey data. As we will see, the definition of constituencies (e.g., legislative districts, federal states) differs between these approaches, and some of them employ additionally aggregate data available from other sources such as statistical offices. What they have in common, though, is their focus on survey measures of public opinion, which from our point of view, gives them an edge over alternative approaches using aggregate proxies of public opinion both in terms of validity and specificity. To be sure, many if not most previous attempts to obtain measures of constituency preferences exclusively rely on aggregate, not individual-level, data. For instance, constituency-level sociodemographics (e.g., Kalt and Zupan, 1984; Krehbiel, 1993; Levitt, 1996), presidential electoral returns (e.g., Ansolabehere, Snyder and Stewart, 2001; Canes-Wrone, Brady and Cogan, 2002; Erikson and Wright, 1980), and referenda outcomes (e.g., Kuklinski, 1978; McCrone and Kuklinski, 1979; McDonagh, 1993) have been frequently used as surrogates of constituency preferences. Yet, while sociodemographics may antecede preferences, and preferences may determine voting behavior, these concepts are clearly not the same, so that such measurement by proxy is always prone to validity issues (for a very pronounced critique, see Jackson and King, 1989). Recent advances on estimating constituency preferences from aggregate data take these issues seriously, and attempt to retrieve the voter preference distributions underlying election results (e.g., Kernell, 2009; Levendusky, Pope and Jackman, 2008) or ballot proposition outcomes (e.g., Gay, 2007; Gerber and Lewis, 2004; Selb and Pituctin, 2010; Snyder, 1996) using scaling and decomposition techniques, often grounded solidly in decision theory. As a

---

2The only explicit references to SAE, that we came across, in the vast political science literature on measuring constituency preferences appear in Jackson (1989) and Park, Gelman and Bafumi (2004).
potential drawback, however, the former usually recover a single latent preference dimension which is broadly interpreted in terms of a constituency’s ‘partisanship’ or ‘ideology’ – concepts probably too fuzzy when one is interested in more specific aspects of public opinion. The latter, while inherently policy-specific, require data on referenda and are thus essentially limited to states and countries that make heavy use of direct democracy.

2.1. Direct estimators

In their seminal study of policy congruency between constituents and legislators, Miller and Stokes (1963) used survey data from the Survey Research Center (SRC) of the University of Michigan’s 1958 congressional election study to directly estimate constituency preferences on three policy issues (social welfare, civil rights, foreign policy) through averaging the respondents’ opinions on an issue by district (also see, i.a., Hurley and Hill, 2003; McCrone and Stone, 1986). The overall sample of 1’517 respondents only covered 116 of the 435 congressional districts, with an average of 13 respondents ranging from 1 to over 40 per district – numbers presumably too small (and too variable) to obtain reliable direct estimates of constituency preferences even for the districts covered. Moreover, attempts to at least gauge the reliability of the direct estimators and, ultimately, to adjust policy congruency measures and models for varying estimator reliability depended on design variance formulae that assume random sampling within districts (Achen, 1978; Cnudde and McCrone, 1966; Jackson, 1989) – a questionable assumption given that the primary sampling units (PSUs) of the 1958 survey were counties not districts (see Erikson, 1978). In 1978 and 1980, the Center of Political Studies’ (CPS) election study temporarily switched to congressional districts as the PSUs, which made the calculation of the direct estimators’ reliabilities more straight-forward (see Erikson, 1981), and slightly increased (and equilibrated) the average number of respondents per district to 21, though at a modest expense of cross-district coverage (108 out of 435). Indeed, with a given budget constraint, there is an obvious tradeoff between within-district precision and cross-district coverage. Furthermore, even the recent reduction of survey costs through the proliferation of telephone and online instead of face-to-face interviewing does not really promise a way out of this dilemma in the near future.

One strategy to increase the size of survey samples in the face of limi-
The major drawbacks of the direct estimator in normal circumstances, that is, its imprecision and potential lack of coverage with sparse survey data and large numbers of districts, arise from its exclusive reliance on survey information from within constituencies. Alternative strategies use cross-area information in different ways. One approach involves a two-step strategy based on standard regression methods. For example, in his reanalysis of the Miller-Stokes data, Erikson (1978) used dummy variables for urban resident, foreign stock, blue-collar occupation, nonwhite, and Democratic presidential vote in 1956 to model the respondents’ opinions on issues. He then predicted mean district preferences based on the district-level distributions of the covariates as provided by official statistics, weighted by the estimated regression parameters from the individual-level equation. The great advantage of this strategy over the direct estimator is that it allows for the prediction of constituency preferences even for the districts not covered by the survey.
as long as the population characteristics are known from census data – an opportunity Erikson could not capitalize on, since complementary data on Congressmen had only been collected for the 116 districts covered by the CPS survey. Leaving the problem of integrating first-stage estimation uncertainty into the second-stage model aside for a moment, the major drawback of this approach is that it involves a bottom-up, cross-level inference and is based on what is known as the ‘constancy assumption’ in the (top-down oriented) ecological inference literature (see Goodman, 1953). In particular, respondents are assumed to form their opinions similarly, that is, as a fixed function of their personal background, regardless of their district of residence – an assumption all too often demonstrated not to hold in many connections (see, i.a., Cho and Manski, 2008).

In an earlier attempt to estimate state-level constituency preferences, de Sola Pool, Abelson and Popkin (1965) already set the conceptual stage for fixing this problem. They stratified the electorate into five regional, four occupational, three size-of-place, two racial, two partisan, and sex categories, yielding a total of $5 \times 4 \times 3 \times 2 \times 2 \times 2 = 480$ synthetic voter types, which they thought to be relatively homogeneous regarding their preferences. Subsequent studies in this vein expanded the number of synthetic voter types to 960 (see Weber et al., 1972). Provided the homogeneity assumption held and that estimates of the voter types’ preferences were available, state-level constituency preferences could then be calculated as the census-weighted sums of voter types’ preferences. The advantage of this strategy is that it allowed

---

4In small area statistics, such estimators are frequently dubbed ‘synthetic’ (see Rao, 2005), which we find more appropriate than ‘simulation-based’, the term usually used to describe this approach in the policy congruency literature. This obviously dates back to the Simulmatics project (de Sola Pool and Abelson, 1961), whose originators introduced the method.

5Ardoin and Garand (2003) took a top-down approach in estimating mean preferences in congressional districts by first regressing (large-sample) state-level direct estimators of opinion on state-level characteristics. The first-stage regression weights were then used to predict mean district-level preferences from known population characteristics. Ardoin and Garand argued that this strategy is superior to the bottom-up approach, as it is normally easier to predict aggregate opinion from the aggregate distribution of sociodemographic characteristics than it is to predict individual opinion from individual social background. Again, relationships between variables just need not be the same across spatial scales, a phenomenon that is similar to the ecological inference problem and known as the ‘modifiable areal unit problem’ in the spatial statistics literature (see Openshaw and Taylor, 1979).

6Erikson (1978) himself obviously suspected that the relationship between individual opinion and background characteristics was not the same across regions, which led him to exclude the 20 southern districts from the estimation procedure. However, the southern vs. non-southern divide is but one contextual factor that potentially causes varying preferences across constituencies, social structure being equal.
different individuals from different constituencies (or, at least, regions) to make up their minds differently, and that it easily disaggregated to the level of legislative districts, provided the distributions of the cross-classified population characteristics were known (e.g., Sullivan and Minns, 1976). Moreover, by using census data for poststratification, this approach would implicitly correct for survey nonresponse, given that the nonresponse was related to the observed sociodemographic characteristics. However, directly estimating the mean preferences of 480 or more voter types from sparse survey data would raise the same difficulties as those already discussed in Section 2.1. Instead, de Sola Pool and his colleagues directly estimated the mean preferences for each category of the stratification variables separately, and calculated a given voter types’ preference as an additive function of the separate effects of each particular defining characteristic. Though practical, this estimation strategy rests on some very strong assumptions, thereby giving away much of the general approach’s conceptual leverage. First, it assumes that variables not included in the stratification can be ignored. Any differences in constituency preferences are attributed to the differential social composition of the states. Second, it is assumed that the stratification variables affect preferences in an additive manner, i.e., they do not interact. Third and relatedly, it assumes that the effect of an individual’s social background on their preferences is insensitive to the social composition of the constituency’s population. For instance, ethnicity is assumed to impinge on an individual’s preferences regardless of the ethnic composition of the constituency where the individual lives. These assumptions seem difficult to maintain without empirical evidence the approach itself is ill-suited to deliver (for a slating critique, see Seidman, 1975).

2.3. Hierarchical estimators

More recently, Gelman and Little (1997) picked up on de Sola Pool, Abelson and Popkin’s idea to stratify survey samples into voter types and to use census information for poststratification in order to obtain state-level constituency preferences, dramatically improving the estimation strategy (also see Gelman, 2008; Park, Gelman and Bafumi, 2004). Gelman and Little developed a multilevel (logistic) model, with (binary) preferences being modeled as a function of 64 sociodemographic voter types nested in 51 states (including Washington D.C.), making for a total of $64 \times 51 = 3'264$ synthetic subconstituencies. This model ‘borrows strength’ by partially pooling voter types across states, that is, all the respondents in a survey, no matter where they live, contribute information about differences in preferences between sociodemographic voter types to an extent warranted by the data. At the same
time, the state of residency is used to estimate non-demographic state-level effects, which themselves are modeled using additional state-level predictors (previous Republican vote share, region). In a second step, Gelman and Little obtained measures of state-level constituency preferences by weighting estimated voter type preferences with the portion of each type in the actual state populations. Park, Gelman and Bafumi (2004) used the model to estimate state-level vote shares for George Bush sen. (therefore the logistic specification) at the 1988 and 1992 Presidential elections using CBS News-New York Times national polls with sample sizes of 2'193 and 4'650 plus census data for poststratification. Validation by comparing to the actual election returns yielded an average absolute error of about 4% in both cases – far better than predictions based on separate state analyses (a strategy that is akin to the direct estimator based on small samples) and on complete pooling of all respondents across states (an approach similar to the synthetic estimator discussed above). Lax and Phillips (2009) further compared this method to the direct estimator based on a large-scale sample from several merged polls, and concluded that the hierarchical estimator using small samples was as accurate as utilizing ten times as much survey data for direct estimation. Thus, the hierarchical estimator with poststratification appears to clearly outperform previous survey-based approaches to estimating constituency preferences in terms of accuracy and efficiency. Moreover, as is the case with other synthetic estimators, it offers coverage and nonresponse corrections even for constituencies not included in the surveys used, provided constituency-level information is available. As opposed to the two-step procedures applied in previous approaches, the hierarchical estimator incorporates the uncertainty inherent in direct estimation from survey data through shrinkage of directly observed differences between subconstituencies according to their reliability (i.e., the variability and number of individual responses). On the other hand though, the hierarchical estimator with poststratification requires detailed census information for the population of each synthetic voter type (e.g., the number or population share of white females aged over 65 who are college graduates) – information that might not be available in other contexts, or at lower federal levels (for a modified application of this approach at the level of U.S. school districts, see Berkman and Plutzer, 2005). In addition, the available census information must cover demographic variables that are associated with the preferences of interest. Finally, whenever the aim is to estimate constituency preferences at lower federal levels, the number of synthetic voter types will rapidly increase. For example, a model including the above 64 sociodemographic categories and 435 congressional districts will yield 27'840 synthetic types. Although multilevel models borrow strength across all observations, and thus do not require that each and every strat-
ification cell be populated with survey respondents, such complexities will inevitably necessitate larger numbers of survey responses.

3. AN ALTERNATIVE APPROACH

So what to do if the current gold standard in estimating constituency preferences is impractical due to data limitations? Surely, the major advantages of a hierarchical estimation strategy – borrowing strength from across constituencies and shrinkage of direct estimators according to their reliability – should be retained. Therefore, our point of departure is a simple ‘empty’ hierarchical logistic model of the probability that respondent \(i\) in constituency \(j\) holds a certain preference, \(p_{ij} = \Pr(y_{ij} = 1)\), being a function of a global propensity to hold that preference, \(\alpha\), and a district-specific deviation from that global propensity, \(u_j\):

\[
\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \alpha + u_j. \tag{1}
\]

In assuming that the district deviations vary, for instance, independent identically according to a normal distribution,

\(u_j \sim N(0, \sigma_u^2)\), \hspace{1cm} \tag{2}

we can use information from all the districts to estimate the proportion holding a preference, \(\pi_j\), in any of the districts,

\[
\hat{\pi}_j^{\text{Uns}} = \frac{\exp (\hat{\alpha} + \hat{u}_j)}{1 + \exp (\hat{\alpha} + \hat{u}_j)}, \tag{3}
\]

where the acronym ‘Uns’ stands for the unstructured random effects (REs), \(u_j\). The benefits of such a simple RE model in connection with SAE problems are already considerably high. For small samples, the direct estimates of the district proportions (with \(N_j\) indicating the number of survey respondents per district)\(^8\)

\[
\hat{\pi}_j^{\text{Dir}} = \frac{\sum_{i(j)=1}^{N_j} y_{ij}}{N_j}, \tag{4}
\]

\(^7\)We use a logistic specification since, in the following empirical section, we will estimate party vote shares from post-election surveys in order to compare them to official election results (also see Park, Gelman and Bafumi, 2004). If indicated, one could equally start from a linear model specification.

\(^8\) The direct estimators may be weighted inversely proportional to the respondents’ selection probability as determined by the sampling design of the survey. However, integrating such weights into a hierarchical model is not a trivial exercise (see Gelman, 2007).
often display much more variability than the true values, $\pi_j$. In particular, we would expect many district proportions directly observed from sparse survey data to be exactly zero (or one). Giving $u_j$ a distribution will, in effect, shrink the observed district proportions toward the overall sample mean, with the amount of shrinkage increasing with decreasing $N_j$ and $\sigma^2_u$. This ‘workhorse’ model for SAE problems is easily extended to include district- as well as individual-specific covariates, $z_j$ and $x_{ij}$, and to accommodate more flexible covariance structures (see Rao, 2005). In the following paragraphs, we will demonstrate how auxiliary geographic information may be incorporated in order to further improve the estimates from this simple measurement model.

3.1. Using auxiliary geographic information

Our approach starts from the observation that political predispositions, preferences and behaviors often come along in geographic clusters (see, i.a., Agnew, 2002; Gelman, 2008; Johnston and Pattie, 2006). Explanations of these geographical clusterings emphasize regional differences in historical settlement and immigration patterns (Elazar, 1994), the geography of industrialization and urbanization (Rodden, 2010), and the resulting differences in current cultural and ethnic composition (Hero, 1998), and local economic conditions (Heppen, 2003). Such broad historical processes are unlikely to abruptly halt at the geographical boundaries of constituencies, particularly if these boundaries are arbitrarily rather than historically drawn to satisfy numerical considerations. For example, SMP electoral districts are required to unite approximately equal numbers of voters to warrant evenly weighted votes, and are thus often subject to more or less arbitrary redistricting to reflect changing shares of electoral population. If the geographical distribution of political preferences cross-cuts constituency boundaries, this will lead to similar preference distributions among neighboring constituencies, or, in technical terms, to spatial autocorrelation. Potential spatial autocorrelation is often conceived of as a nuisance for statistical estimation, as it may violate the assumption of independent errors that is typical of many regression models (see the distributional assumptions for $u_j$ in the above model). Far from being treated as a nuisance, spatial autocorrelation is the major source of information in our model. Provided constituency preferences are

\footnote{Contrary to linear models, though, the prediction of constituency preferences in Equation 3 is not so straightforward with individual-level regressors because the response and the covariates are not linked linearly and additively any more, so that an appropriate aggregate form of the individual-level model needs to be specified (see Denk and Finkel, 1992). See Park, Gelman and Bafumi (2004) and Section 4.4 for a simple solution to this problem using dummy predictors and poststratification.}
spatially correlated, the preferences of neighboring constituencies are informative about preferences in any given constituency. Moreover, since most constituencies have more than one neighbor, the mean preferences in the neighborhood of a given constituency can usually be estimated much more precisely than in the target area itself (with the number of survey respondents per constituency being relatively constant). Therefore, we will extend the basic model’s covariance structure to borrow strength from across neighboring constituencies.

In particular, the (logit of the) probability of respondent $i$ in constituency $j$ to hold a certain preference of interest is again a function of a global propensity, $\alpha$, and this time a vector of constituency-specific covariates, $z_j$, plus their respective regression weights,

$$\log \left( \frac{p_{ij}}{1 - p_{ij}} \right) = \alpha + \beta z_j + u_j + v_j,$$

where $u_j$ is an independent identically distributed RE as above, and $v_j$ represents a spatially correlated RE for which we assume an intrinsic conditional autoregressive (CAR) distribution (see Besag, York and Mollié, 1991). Under this specification, the conditional distribution of $v_j$ given the values $v_k$ in all the other constituencies $k \neq j$ only depends on the $v$'s in the $j$-specific subset, $c_j$, of contiguous constituencies, $l$,

$$v_j \mid v_k \sim N \left( \sum_{l \in c_j} \frac{v_l}{L_j}, \sigma^2_v \right),$$

where $L_j$ represents the number of constituencies in the neighborhood of $j$. Hence the expected conditional mean of $v$ in $j$ corresponds to the average value of $v$ in $c_j$, with its variance, $\sigma^2_v$, also being normalized by the number of neighbors. In this manner, the model borrows strength from neighbors to estimate the (mean) constituency preference in $j$, which is then

$$\hat{\pi}^{\text{CAR}}_j = \frac{\exp \left( \hat{\alpha} + \hat{\beta} z_j + \hat{u}_j + \hat{v}_j \right)}{1 + \exp \left( \hat{\alpha} + \hat{\beta} z_j + \hat{u}_j + \hat{v}_j \right)}.$$
The advantage of this model over direct and synthetic estimators in terms of coverage is that by exploiting the conditional distribution of $v_j$, it will equally inform estimates of constituency preferences for areas not covered by the survey even if the $z_j$’s were unknown, provided a constituency is not an island (i.e., it has neighbors to draw information from). In fact, in the following empirical application, we will use a single constituency-level covariate, namely the log inverse surface area as a proxy of urbanity, which has been repeatedly demonstrated to be a crucial contextual determinant of political preferences, and therefore has the potential to disrupt otherwise smooth spatial preference distributions (see Rodden, 2010). To better be able to separate between eventual model improvements due to the inclusion of the covariate and due to the spatially structured REs, we will also report validation results for an unstructured estimator (see Equation 3) with log inverse area as a covariate, $\pi_{j,uns}$. Quite intuitively, we would expect the estimator with spatially structured REs to a fortiori outperform its simpler alternatives, the stronger the spatial autocorrelation of the (true) preferences of interest.

3.2. Bayesian estimation

Given the paucity of survey and auxiliary data one usually confronts, the REs $u_j$ and $v_j$ are essential components of our model. Most frequentist hierarchical modeling approaches treat such effects as fixed nuisance parameters and will, at best, generate point estimates (so-called ‘empirical Bayes residuals’) of these quantities (e.g., Snijders and Bosker, 1999). Without assessments of uncertainty, however, it is problematic to compare these point estimates, to validate the models, and to use the estimates in subsequent analyses. We choose a Bayesian estimation strategy instead (for some forceful arguments in favor of a Bayesian estimation strategy in this connection, see Levendusky, Pope and Jackman, 2008). In a Bayesian framework, $u_j$ and $v_j$ (including missing values for districts not being covered by the survey) are treated as

---

Note that, by exploiting second- and higher-order neighborhood relations, this is also true for constituencies that altogether lack in-sample neighbors, i.e., all its neighbors have missing values as well. However, the estimation will be imprecise and unstable in these circumstances. We will get back to this point later.

If (and only if) the number of residents is relatively constant across (SMP) electoral districts, surface area is logically related to population density. In cases where some structural information is available one could, of course, directly use log population density as an indicator of urbanity. In fact, information about population density is available in our empirical case, and log inverse area and log population density are correlated at 0.99. For the time being, we would like to see how far we get using minimal geographic information alone.
any other model parameters, for which a joint posterior density is computed using Markov chain Monte Carlo (MCMC) methods. Point estimates and uncertainty assessments are then obtained by summarizing arbitrarily many samples from the joint posterior density. We use WinBUGS to run the required computations (Lunn et al., 2000). Complex covariance structures as those in Equation 6 can be implemented using GeoBUGS, an add-on module to WinBUGS for manipulating spatial data (Thomas et al., 2004). Commented replication code is given in the Appendix.

We use uninformative priors to let the data determine the parameters. In particular, we use flat priors for the fixed parameters, \( \alpha \) and \( \beta \), and vague inverse Gamma priors of \((0.001, 0.001)\) for the variances of the REs, \( \sigma_u^2 \) and \( \sigma_v^2 \). To monitor convergence, we set up three chains with randomly chosen starting values for the parameters, each with 20,000 iterations, of which we discard the first 10,000 before summarizing the model’s parameters’ posterior probabilities.

4. VALIDATION

To validate our models, we estimate district-level party vote shares from two post-election surveys conducted after the 2009 German Bundestag election within the framework of the German Longitudinal Election Study (GLES), and compare them to their true distributions which are known from official election results.\(^{14}\) Germany’s (so-called ‘mixed-member PR’) electoral system divides the country into \( J = 299 \) primary electoral districts (Wahlkreise) which constitute our target areas. The districts are nested within a higher electoral tier, the 16 Länder (federal states). Each voter has two votes: a candidate vote (Erststimme) which governs the allocation of the district seats (Direktmandate) using SMP, and a second vote (Zweitstimme) for a länder-specific party list that translates into seats (Listenmandate) according to the LR-Hare method, restricted by a national five per cent clause (for details, see Saalfeld, 2008). Since the Erststimme is particularly prone to strategic voting which may disrupt otherwise smooth spatial distributions of (partisan) preferences contingent on district-specific incentives, we will focus our analysis on the Zweitstimme. All five parliamentary parties will be considered: the Christian Democrats (CDU/CSU), the Social Democrats (SPD),

\(^{13}\)Note that, due to its independent distribution, \( u_j \) cannot be identified for constituencies which are not covered by the survey and thus has to be set to zero in these instances.\(^{14}\) The survey data as well as technical reports are available from the GLES website [www.gesis.org/gles](http://www.gesis.org/gles). Official election statistics, digitalized maps and additional structural information can be downloaded from the website of the Bundeswahlleiter (federal elections officer) at [http://www.bundeswahlleiter.de/de/bundestagswahlen/ BTW_BUND_09/](http://www.bundeswahlleiter.de/de/bundestagswahlen/BTW_BUND_09/).
the Liberals (FDP), the Greens (Bündnis 90/Die Grünen), and the Left (Die Linke). Altogether, these parties received 94% of the Zweitstimmen in 2009, which almost makes for the complete choice set with which the voters were confronted. Instead of adding a lumping category for marginal parties and setting up a (more efficient) multinomial logistic regression, we model the party vote shares separately. First, in order to better explore the conditions of how well or poorly the estimates perform; and second, to be consistent with the more general modeling strategy proposed in Section 3.

4.1. Data

Clearly, estimating constituency preferences for the 299 electoral districts from national polls puts some demands on the survey data even if model-based (let alone direct) estimation strategies are employed. We can benefit from the fortunate circumstance that the 2009 GLES conducted two relatively generous post-election surveys after the election on September 27: a face-to-face (F2F) survey of a three-stage random sample of 2'117 eligible individuals, and the post-election wave of a rolling cross-section survey of 4'027 individuals conducted by telephone (CATI).

For the time being, we will pool these data, which makes for a total sample of 6'144 respondents, 5'067 of which are self-reported voters that indicated a party choice (recall that our aim is to estimate party vote shares, which are fractions of actual voters). Subsequently, we will also use the two surveys separately in order to see how our estimates behave with more common sample sizes. Table 1 gives an overview over the sample characteristics. The pooled sample covers 297 of the 299 districts, with an average of $\bar{N}_j = 17$, ranging from 1 to 51 respondents per district – barely sufficient to directly estimate the vote shares of the five parties in a reliable manner. Of course, this holds all the more true for the separate surveys. On the other hand, at the national level the sample vote shares of the parties quite reasonably reflect their true values, with B90/Die Grünen being the most notable exception (13.7% according to

\footnote{The respondents from the F2F survey were sampled using the standard design of the consortium of German market research institutes (Arbeitsgemeinschaft deutscher Marktforchungsinstitute, ADM), with electoral wards as the PSUs being randomly sampled within regions (East and West), households being selected through random-route methods within electoral wards (Wahlbezirke), and persons being randomly chosen within households. The sample of the CATI survey was drawn using the so-called Gabler/Häder design, where area codes within regions serve as the PSUs. None of these designs will normally warrant equal probability samples for the electoral districts. Absent the information necessary to construct weights that compensate for unequal selection probabilities within districts (and the quibbles with including such weights in hierarchical models; see Footnote 8), we will treat the district samples as if they were randomly drawn.}
the pooled sample vs. an official return of 10.7%) – an issue we will return to in Section 4.4.

[Table 1 about here.]

4.2. Conditions and validation criteria

In the modeling section, we have conjectured that the performance of the spatially structured estimator will depend on the magnitude of spatial autocorrelation in the (true) distribution of preferences across districts. A widely used global measure of spatial autocorrelation is Moran’s $I$, that is, the correlation between a variable, $y_j$, with its spatial lag of $Wy_j$, where $W$ is a $J \times J$ adjacency matrix whose elements, $w_{jk}$, assume $1/L_j$ if units $j$ and $k$ have a common border (or, at least, a common vertex), otherwise 0. $Wy_j$ is thus the average value of $y$ in the neighborhood set, $c_j$. We initially calculate Moran’s $I$ for the official election returns at the district level to see how our estimates’ performance depends on the true spatial correlation of partisan preferences. Figure 1 plots the official vote shares of the five parties (henceforth ‘true values’) against their spatial lags. The slopes of the solid lines is given by the Moran values. The dashed lines represent averages of the true values and their spatial lags. Altogether, the magnitude of spatial autocorrelation is considerable, indicating that neighborhoods should be highly informative for estimating constituency preferences in any district. In particular, the SPD vote shares exhibit the strongest spatial correlation ($I = 0.79$), followed by the FDP ($I = 0.76$), and the CDU/CSU ($I = 0.67$). While, at first glance, there seems to be tremendous autocorrelation in Die Linke vote shares ($I = 0.85$), this turns out to be an artifact of its pronounced regional stronghold in eastern Germany (see the upper cloud in the graph). Separate calculations of Moran’s $I$ yield relatively low values of 0.57 for eastern, and 0.49 for western districts. Finally, B90/Die Grünen vote shares also exhibit comparatively modest levels of spatial autocorrelation ($I = 0.56$).

[Figure 1 about here.]

Since true spatial patterns will be unknown in substantively more interesting applications, we will, in the subsequent section, also compute $I$ for the direct estimator (see Equation 4) to see whether spatial patterns readily observable from a sparse survey sample properly indicate the underlying true spatial patterns. Finally, we will use Moran’s $I$ as a diagnostic tool to test our models’ assumption that $u_j$ is, indeed, independently distributed.

In validation terms, one important criterion will be the distance between true values and their estimates expressed as the empirical root mean squared error.
error (RMSE) of (the median of) the estimated \( \pi \)'s posterior probabilities. Clearly, smaller RMSEs indicate better point estimates. As to these estimates’ uncertainty, we calculate 90% Bayesian credible intervals from the highest posterior density regions that can be immediately interpreted in terms of the probability that the true value of the estimated parameter is inside a given interval.\(^{16}\) A second validation criterion is the coverage probability of the credible intervals, that is, the proportion of the time that the intervals actually contain the true value of interest. The actual coverage probability should approximate the nominal level of 90% as closely as possible. A final validation criterion will be the parsimony or efficiency of the credible intervals in terms of their width. With their coverage probability being equal, a narrower interval is, of course, preferable.

4.3. Results

Figure 2 plots the true party vote shares versus the direct estimators and the medians of the model-based estimators’ posterior densities. Above all, it is easy to see that the direct estimators graphed in the leftmost panels are essentially useless in all cases. They are way too variable, and wide off the mark in most instances (in fact, so wide off the mark that some of the estimates could not even be plotted within the displayed range of 0 to 60% of the votes). As the RMSEs reported in Table 2 show, they misestimate the true vote shares by an unsatisfactory 7.5 to 10%. The hierarchical estimators with unstructured REs in the second column generally go to the other extreme and shrink the direct estimators toward the overall sample means (represented by the dashed lines in Figure 2) to an extent that there is barely any cross-district variability left – a clear indication that even the pooled data carry too little information for reliable direct estimation. Adding log inverse area as a proxy of population density/urbanity to the unstructured RE model in the third column improves the estimates of the CDU/CSU, B90/Die Grünen, and very modestly, the SPD vote shares.\(^{17}\) A more considerable improvement

\(^{16}\)Assessing the uncertainty of the direct estimators in Equation 4 via (frequentist) confidence intervals is far from straight-forward with small samples (see Agresti and Coull, 1998). Moreover, frequentist and Bayesian intervals are hardly comparable, as the former treat the estimated parameters as fixed and the confidence intervals as random, so that the probability that the true parameter value is inside the given interval is either 0 or 1. In the subsequent empirical analysis, we will therefore only report RMSEs for the direct estimators, but no uncertainty assessments.

\(^{17}\)CDU/CSU vote shares are negatively, and SPD and B90/Die Grünen vote shares are positively related to urbanity (see the estimated coefficient plots for the parameters denoted above by Greek letters in the Appendix). FDP and Die Linke vote shares are also related to urbanity, although in a non-monotonic way that is difficult to interpret, let
of the estimates result from adding the CAR-structured REs, \( v_j \), in the final column. The models’ point estimates come out very close to their true values, with an average error ranging from just 2.5 for the FDP to 4.5% for the CDU/CSU. The CDU/CSU estimates are too conservative in the tails, which also leads to coverage probabilities relatively far below their nominal level. Nevertheless, the coverage probabilities of the credible intervals of \( \hat{\pi}_{j}^{\text{CAR},*} \) are in all cases generally much closer to their nominal 90% level than those of the previous estimators, meanwhile the intervals are narrower and thus more efficient. Finally, as indicated by the Moran’s \( I \) values in Table 2, the inclusion of \( v_j \) has helped to rid most of the spatial autocorrelation of \( u_j \) (which is assumed i.i.d.) in the previous models, which supports the view that the CAR-structured REs provide a good representation of the observed spatial patterns. Overall, our results suggest that the \( \hat{\pi}_{j}^{\text{CAR},*} \)-estimators perform remarkably well with the pooled data, particularly considering the models’ exclusive reliance on auxiliary geographic information. As conjectured, the estimators’ performance seemingly depends on the amount of spatial autocorrelation inherent in the underlying distributions of the true preferences. In particular, the model performed better in estimating SPD and FDP vote shares, and worse in estimating the vote shares of the other parties, most notably those of B90/Die Grünen, with an acceptably low mean coverage probability of 0.76. A closer inspection of Figure 2 reveals, though, that the modest performance of the estimates of the B90/Die Grünen’s vote shares is primarily due to the general upward bias in reported votes that has already been remarked in the data section above. We will further discuss this issue in the following section.

How well do the estimators perform with more regular sample sizes, that is, the F2F (\( N = 1'527 \)) and the CATI (\( N = 3'540 \)) surveys taken separately? Not surprisingly, the RMSEs are generally larger than those of the pooled estimators, ranging at best (i.e., with the spatially informed estimators) from 3 to 7% in the former, and from 3 to 5% in the latter case (see the two lower panels in Table 2 and also the true values versus estimates plots in the Appendix). It is difficult to tell at first sight whether these differences are due to the latter survey providing larger \( \bar{N} \) (and thus, larger \( \bar{N}_j \)), or fuller cross-district coverage in terms of \( J \) (see Table 1), or due to the possibility that the

\footnote{Another violation obvious from Figure 2, the assumption that \( u_j \) is normally distributed also seems difficult to maintain for the Die Linke model. Rather, there seem to be distinct modes for eastern and western districts. We therefore added a regional dummy to the vector of covariates, \( z_j \), which reduced the Die Linke model’s RMSE to 0.02, although at the expense of over-confident estimates, as indicated by a decline in mean coverage probabilities to 0.83. Detailed results are available upon request.}
CATI survey sample more evenly represents the district populations than the F2F sample based on electoral wards within districts as the PSUs (see Footnote 15). On closer inspection (see the RMSE versus $N_j$ and $L_j$ plots in the Appendix), the RMSEs of the spatially informed estimates proved relatively insensitive to the number of respondents per district, $N_j$. Furthermore, even the estimates for the off-sample districts (i.e., $N_j = 0$) on average did not come out wider off target than those for the in-sample districts. Rather, the accuracy of the CAR-estimates seem to moderately depend on the number of respondents in the neighborhood of a given district, $L_j$. Curiously, though, the estimators based on the CATI survey are over-confident, and thus exhibit lower coverage probabilities than those based on the F2F survey. In general, these results indicate that, for obtaining accurate model-based estimates for so many small areas, one still necessitates relatively large-scale national polls.

[Figure 2 about here.]

[Table 2 about here.]

4.4. Extension: poststratification to population controls

One of the drawbacks of our approach is that it does not naturally correct for survey non-response. This problem has, perhaps, become most evident in the upward shift of our district estimates of Green party vote shares. In this final part of our empirical analysis, we will therefore demonstrate how our approach may be combined with the hierarchical approach using poststratification discussed in Section 2.3 to correct for non-response bias. It is well known that both the individual propensity to support postmaterialist parties and to participate in election surveys is positively related to educational attainment (e.g. Brehm, 1993). If so, we would expect the higher educated to be overrepresented in our sample, and the upward bias in our estimates to decrease after poststratification to the ‘true’ district-level distributions of attainment levels. While census data on education is not available, the Bundeswahlleiter at least provides district-level school graduation statistics as of 2007 which we, in the absence of better alternatives, use as proxies for the whole district populations (see Footnote 14). Indeed, individuals with tertiary education seem to be massively overrepresented in our sample (42% compared to 27% according to the official statistic). In contrast, those with secondary (33% compared with 41%) and (at most) primary education (24% compared with 32%) are underrepresented.\footnote{Considering that younger cohorts as those having graduated in 2007 are, on average, higher educated than older cohorts, we may even suspect the true educational bias in our sample to be stronger than indicated.} Following Park, Gelman and
Bafumi (2004), we re-run our models of Green party vote shares using individual dummy variables for the two higher levels of educational attainment (with regression weights $\gamma_2$ and $\gamma_3$) and obtain predicted probabilities of having cast a Green vote for each of the $299 \times 3 = 897$ cross-classifications of electoral districts and attainment levels. The district estimate is then simply the population-weighted average of the type-specific predicted probabilities by district. The results for the unstructured and CAR-structured RE models are reported in Table 3. Obviously, the upward bias in the estimates observed in the previous section is considerably reduced by using poststratification. Both RMSEs are markedly lower, and the coverage probabilities are much closer to their nominal level than before poststratification – most notably, once again, for the spatially informed estimators.

[Table 3 about here.]

5. DISCUSSION

In this paper, we proposed a hierarchical Bayesian approach to estimating constituency preferences from sparse survey data that uses auxiliary geographic information to compensate for small samples and lack of coverage. In particular, we set up a logistic regression model with spatially structured random effects to obtain district-level estimates of party vote shares at the 2009 German *Bundestag* election from national post-election polls that could be validated using official election results. These estimates proved quite accurate and efficient as well as outperformed spatially ignorant alternatives both in terms of precision and uncertainty assessment. Expectedly, the validation results were particularly convincing when we used a relatively generous pooled national sample of approximately 5,000 respondents to estimate the constituency preferences of the 299 German electoral districts, but the estimators’ properties were still somewhat satisfactory with more regularly sized national samples of 3,500 and even 1,500 respondents.

While we sense the most fruitful area of application in situations in which the number of target areas is high and detailed structural information is missing, we also demonstrated how this approach may be integrated with other state-of-the-art strategies that exploit additional information about the constituencies’ population characteristics – specifically, Park, Gelman and Bafumi’s (2004) hierarchical estimator with poststratification. Other extensions are thinkable. For example, one could make use of the fact that

---

20 As opposed to Park, Gelman and Bafumi (2004), we model education with fixed, not random effects.
primary electoral districts are frequently nested within higher-level political units (e.g., states) which themselves emerged from broad historical processes that may have partly shaped their citizens’ preferences. Additional higher-level random effects could exploit the induced correlation between constituencies that belong to the same upper-tier units. Such an estimator will presumably be less sensitive to incomplete constituency coverage, and particularly to the prevalence of off-sample constituencies that lack in-sample neighbors (of course, as long as all the higher-level units are covered by the survey). \textsuperscript{21} These higher-level random effects could, in turn, also be given a spatial structure.

In essence, the strategy we proposed boils down to the argument that spatial correlation is much more than a nuisance for statistical estimation. Rather, spatial patterns may be conceived of as an invaluable source of information when other data are sparse – a common situation when estimating constituency preferences from sparse survey data. In this manner, our approach nicely fits into the ‘spatial turn’ in political science as recently demanded by Ethington and McDaniel (2007) and Franzese and Hays (2008).

\textsuperscript{21} We have also included higher-level random effects for the 16 German Länder in our models. However, the improvement of the district estimates over the specification with CAR-structured random effects were negligible. Detailed results are available on request.
REFERENCES


Table 1: Summary statistics of survey data

<table>
<thead>
<tr>
<th></th>
<th>$N$</th>
<th>$J$</th>
<th>$N_j$</th>
<th>s.d.($N_j$)</th>
<th>min($N_j$)</th>
<th>max($N_j$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F2F</td>
<td>1'527</td>
<td>201</td>
<td>7.60</td>
<td>5.21</td>
<td>1</td>
<td>29</td>
</tr>
<tr>
<td>CATI</td>
<td>3'540</td>
<td>291</td>
<td>12.16</td>
<td>5.60</td>
<td>2</td>
<td>45</td>
</tr>
<tr>
<td>Pooled</td>
<td>5’067</td>
<td>297</td>
<td>17.06</td>
<td>8.31</td>
<td>1</td>
<td>51</td>
</tr>
<tr>
<td>Pooling</td>
<td>N=5'067</td>
<td>Pooled</td>
<td>N=1'527</td>
<td>F2F, N=1'527</td>
<td>CATI, N=3'540</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>---------</td>
<td>--------</td>
<td>---------</td>
<td>--------------</td>
<td>---------------</td>
<td></td>
</tr>
<tr>
<td>SPD</td>
<td>0.096</td>
<td>0.096</td>
<td>0.096</td>
<td>0.096</td>
<td>0.107</td>
<td></td>
</tr>
<tr>
<td>FDP</td>
<td>0.079</td>
<td>0.079</td>
<td>0.079</td>
<td>0.079</td>
<td>0.089</td>
<td></td>
</tr>
<tr>
<td>CDU/CSU</td>
<td>0.101</td>
<td>0.101</td>
<td>0.101</td>
<td>0.101</td>
<td>0.114</td>
<td></td>
</tr>
<tr>
<td>Die Linke</td>
<td>0.074</td>
<td>0.074</td>
<td>0.074</td>
<td>0.074</td>
<td>0.078</td>
<td></td>
</tr>
<tr>
<td>B90/Die Grünen</td>
<td>0.075</td>
<td>0.075</td>
<td>0.075</td>
<td>0.075</td>
<td>0.085</td>
<td></td>
</tr>
</tbody>
</table>

### Table 2: Validation results

<table>
<thead>
<tr>
<th></th>
<th>RMSE</th>
<th>Mean width of 90%-CI</th>
<th>Coverage probabilities</th>
<th>Moran’s I in u</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>Mean width of 90%-CI</td>
<td>Coverage probabilities</td>
<td>Moran’s I in u</td>
</tr>
<tr>
<td>Pooled, N=5'067</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPD</td>
<td>0.096</td>
<td>0.047 0.047 0.034</td>
<td>0.178 0.161 0.143</td>
<td>0.876 0.833 0.923</td>
</tr>
<tr>
<td>FDP</td>
<td>0.079</td>
<td>0.029 0.030 0.025</td>
<td>0.155 0.156 0.137</td>
<td>0.943 0.940 0.963</td>
</tr>
<tr>
<td>CDU/CSU</td>
<td>0.101</td>
<td>0.052 0.047 0.045</td>
<td>0.175 0.146 0.149</td>
<td>0.796 0.786 0.839</td>
</tr>
<tr>
<td>Die Linke</td>
<td>0.074</td>
<td>0.044 0.044 0.032</td>
<td>0.168 0.167 0.123</td>
<td>0.880 0.883 0.906</td>
</tr>
<tr>
<td>B90/Die Grünen</td>
<td>0.075</td>
<td>0.037 0.033 0.034</td>
<td>0.144 0.128 0.115</td>
<td>0.776 0.776 0.759</td>
</tr>
<tr>
<td>F2F, N=1'527</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPD</td>
<td>0.101</td>
<td>0.051 0.051 0.042</td>
<td>0.130 0.153 0.188</td>
<td>0.629 0.712 0.930</td>
</tr>
<tr>
<td>FDP</td>
<td>0.100</td>
<td>0.031 0.031 0.035</td>
<td>0.132 0.140 0.198</td>
<td>0.803 0.823 0.973</td>
</tr>
<tr>
<td>CDU/CSU</td>
<td>0.119</td>
<td>0.061 0.062 0.065</td>
<td>0.211 0.223 0.273</td>
<td>0.726 0.753 0.906</td>
</tr>
<tr>
<td>Die Linke</td>
<td>0.093</td>
<td>0.064 0.065 0.055</td>
<td>0.171 0.169 0.203</td>
<td>0.592 0.612 0.863</td>
</tr>
<tr>
<td>B90/Die Grünen</td>
<td>0.076</td>
<td>0.032 0.029 0.030</td>
<td>0.083 0.095 0.126</td>
<td>0.676 0.759 0.923</td>
</tr>
<tr>
<td>CATI, N=3'540</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPD</td>
<td>0.107</td>
<td>0.052 0.051 0.044</td>
<td>0.114 0.098 0.114</td>
<td>0.625 0.518 0.702</td>
</tr>
<tr>
<td>FDP</td>
<td>0.089</td>
<td>0.031 0.033 0.030</td>
<td>0.088 0.075 0.109</td>
<td>0.709 0.615 0.826</td>
</tr>
<tr>
<td>CDU/CSU</td>
<td>0.114</td>
<td>0.058 0.053 0.050</td>
<td>0.124 0.084 0.117</td>
<td>0.619 0.475 0.692</td>
</tr>
<tr>
<td>Die Linke</td>
<td>0.078</td>
<td>0.045 0.046 0.033</td>
<td>0.162 0.163 0.120</td>
<td>0.873 0.883 0.920</td>
</tr>
<tr>
<td>B90/Die Grünen</td>
<td>0.085</td>
<td>0.048 0.045 0.044</td>
<td>0.125 0.077 0.114</td>
<td>0.615 0.395 0.619</td>
</tr>
</tbody>
</table>

*p < 0.05 (for values of Moran’s I)
Table 3: Validation results

<table>
<thead>
<tr>
<th></th>
<th>RMSE</th>
<th>Mean width of 90%-CI</th>
<th>Coverage probabilities</th>
<th>Moran’s I in u</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{x}_j^{\text{Uns,x}}$</td>
<td>0.028</td>
<td>0.026</td>
<td>0.138</td>
<td>0.876</td>
</tr>
</tbody>
</table>
Figure 1: Moran’s plots of party vote shares (true values)
Figure 2: True values versus estimates by party

- SPD
- FDP
- CSU/CSU
- Die Linke
- Bündnis 90/Die Grünen

True values
CMS Discussion Paper 2010-05

For a complete list of the Discussion Papers published by the CMS please visit http://www.uni-konstanz.de/cms.

01 “The Effect of Grade Retention on School Outcomes: An Application of Doubly Robust Estimation Method” by Selver Derya Uysal, January 2010

02 “Can we help finance professionals to predict the Euribor rate?” by Fabian Krüger, Frieder Mokinski and Winfried Pohlmeier, February 2010

03 “Bayesian Inference for Factor Structure Models via Gibbs Sampling” by Rémi Piatek, February 2010

04 “Forecasting Euro-Area Macroeconomic Variables Using a Factor Model Approach for Backdating” by Ralf Brüggemann and Jing Zeng, February 2010

05 “Estimating Constituency Preferences from Sparse Survey Data Using Auxiliary Geographic Information” by Peter Selb and Simon Munzert, October 2010