PA

Stacked Regression and Poststratification

Joseph T. Ornstein

Brown School, Washington University in St. Louis, St. Louis, MO 63130, USA. Email: jornstein@wustl.edu

Abstract

I develop a procedure for estimating local-area public opinion called stacked regression and poststratification (SRP), a generalization of classical multilevel regression and poststratification (MRP). This procedure employs a diverse ensemble of predictive models—including multilevel regression, LASSO, k-nearest neighbors, random forest, and gradient boosting—to improve the cross-validated fit of the first-stage predictions. In a Monte Carlo simulation, SRP significantly outperforms MRP when there are deep interactions in the data generating process, without requiring the researcher to specify a complex parametric model in advance. In an empirical application, I show that SRP produces superior local public opinion estimates on a broad range of issue areas, particularly when trained on large datasets.

Keywords: public opinion estimation, machine learning, ensemble methods

1 Introduction

Subnational public opinion data are often difficult or costly to obtain. For political scientists who focus on lower-level units of government (e.g. legislative districts, counties, cities), this lack of local-area public opinion data can be a significant impediment to empirical research. And so, over the past two decades, political methodologists have refined techniques for estimating subnational public opinion data from national-level surveys. A now standard approach is multilevel regression and poststratification (MRP), first developed by Gelman and Little (1997) and refined by Park, Gelman, and Bafumi (2004).

MRP proceeds in two stages. First, the researcher estimates a multilevel regression model from individual-level survey data, using demographic and geographic variables to predict public opinion. The predictions from this first-stage model can then be used to estimate average opinion in each local area of interest. To do so, the researcher takes each demographic group's predicted opinion and computes a weighted average using the observed distribution of demographic characteristics. This second stage is called *poststratification*.

MRP has enabled a flowering of new research on political representation in states (Lax and Phillips 2012), Congressional districts (Warshaw and Rodden 2012), and cities (Tausanovitch and Warshaw 2014). But the method is not without its critics. Buttice and Highton (2013), for instance, find that MRP performs poorly in a number of empirical applications, particularly when the first-stage model is a poor fit for the public opinion of interest. They find that MRP works best for predicting opinion on cultural issues (like support for same-sex marriage), where there is greater geographic heterogeneity in opinion. In these cases, public opinion is more strongly predicted by geographic-level variables, yielding better poststratified estimates. But for opinions on economic issues, MRP yields a poorer fit. The authors conclude by emphasizing the importance of model selection, noting that "predictors that work well for cultural issues probably will not work well for other issue domains and vice versa". This finding echoes calls from other MRP scholars, who urge researchers to construct a first-stage model that is well-suited to the topic of study (Lax and Phillips 2009; Ghitza and Gelman 2013).

Fundamentally, MRP is an exercise in out-of-sample prediction, using observed opinions from survey respondents to make inferences about the opinions of similar individuals who were not surveyed. As such, the first-stage model should be selected on the basis of its out-of-sample

Political Analysis (2020)
vol. 28:293–301
DOI: 10.1017/pan.2019.43

Published
23 December 2019

Corresponding author
Joseph T. Ornstein

Edited by
Jeff Gill

predictive performance. Though classical MRP relies on multilevel regression models, there is no reason *ex ante* to believe that such models will perform best at this task; any method that produces regularized predictions can serve as a first-stage model (Gelman 2018).

In this paper, I introduce a refinement of classical MRP, called Stacked Regression and Poststratification (SRP). Rather than estimating public opinion from a single multilevel regression model, this technique generates predictions from a "stacked" ensemble of models, including regularized regression (LASSO), k-nearest neighbors, random forest, and gradient boosting. The stacking procedure selects an ensemble model average that minimizes cross-validation error, improving the ensemble's ability to predict out-of-sample cases, and thereby yielding better poststratified estimates.[1] In both a Monte Carlo simulation and empirical application, I show that this technique produces superior estimates of subnational public opinion, particularly when estimated using surveys with large samples. I conclude with guidelines for best practice and suggestions for future research.

2 The SRP Procedure

The SRP procedure differs from MRP in the first stage. Rather than predicting individual-level opinion using a single multilevel regression model, SRP employs an ensemble model average (EMA) from a diverse set of predictive models. This ensemble prediction is a weighted average defined as follows, where $f_k(X_i)$ denotes the predicted value from model k given covariates X_i:

$$f(X_i) = \sum_{k=1}^{K} w_k f_k(X_i). \tag{1}$$

Within this framework, one can think of classical MRP as a special case of SRP where the weight vector is constrained to $w_k = 1$ for a prespecified multilevel regression model and zero for all other models. SRP relaxes this constraint, and instead estimates the w_k weights to minimize cross-validated prediction error.

To estimate the w_k weights, I use an approach called *stacking*, first proposed by Wolpert (1992) and refined by Leblanc and Tibshirani (1996) and Breiman (1996).[2] This approach proceeds in two steps. First, the researcher generates out-of-sample predictions from each base model through k-fold cross-validation (holding out k-folds with $\frac{n}{k}$ observations each, training the model on the remaining data, and predicting the observations in each fold). This cross-validation step ensures that the ensemble model average does not place too much weight on complex models that overfit the training data.

Second, the researcher uses these out-of-sample predictions as the f_k values in Equation (1), estimating the w_k weights that minimize prediction error. Because the base models' predictions tend to be highly collinear (after all, they are all predicting the same outcome), OLS or logistic regression tend to yield highly unstable coefficient estimates, so it is best practice to constrain the weights to be nonnegative ($w_k \geq 0$) and sum to one ($\sum w_k = 1$) (Breiman 1996). The weights can then be estimated through a hill-climbing algorithm (Caruana *et al.* 2004) or quadratic programming (Grimmer, Messing, and Westwood 2017). For continuous outcome variables, I select

1 Stacking is similar in nature to Bayesian Model Averaging (Montgomery, Hollenbach, and Ward 2012). The latter method produces posterior predictive distributions through Bayesian updating, whereas stacking yields only point estimates. Though there has been recent work extending stacking to average Bayesian predictive distributions (Yao *et al.* 2018), for our purposes here point estimates will suffice.

2 A separate strand of the machine learning literature refers to this method as the "Super Learner"; van der Laan, Polley, and Hubbard (2007) prove that, asymptotically, the predictions from such ensemble model averages outperform any of ensemble's component models. This result, of course, does not guarantee that stacking will outperform its component models in all cases, particularly when trained on finite samples, but it does suggest that its relative performance will improve with larger datasets.

Table 1. The SRP Procedure.

Step	Procedure
1	Collect individual-level survey data on outcome of interest and predictors.
2	Fit a diverse ensemble of models to predict the outcome of interest, using demographic and geographic variables as covariates.
3	Obtain out-of-sample predictions for each observation in the dataset through cross-validation.
4	Use stacking to find the ensemble model weights that minimize cross-validated prediction error.
5	Generate predictions for each respondent type (demographic × geographic variables).
6	Poststratify by weighting these predictions against the known frequency of each type at the subnational level.

the ensemble that minimizes root mean squared error (RMSE); for binary outcomes, I maximize the log-likelihood.

In each of the applications that follow, I include five different predictive models in my ensemble. These include multilevel regression, regularized regression (LASSO), K-Nearest Neighbors (KNN), Random Forests (RF), and Gradient Boosting (GBM). I encourage readers who are unfamiliar with these methods to consult Tibshirani (1996), Breiman (2001), and Montgomery and Olivella (2018) for excellent primers. See the Supplementary Materials for a discussion of these models' properties, their implementation, and why I chose to include them while omitting other, similar machine learning techniques.

Once the ensemble weights are estimated, Equation (1) produces the first-stage predictions. The local-area estimates can then be generated through poststratification as in classical MRP. Table 1 summarizes the SRP procedure.

3 Monte Carlo Simulation

How well does SRP perform relative to classical MRP? And under what conditions does it perform best? To address these questions, I conduct a Monte Carlo analysis, simulating a data generating process where the outcome variable ($\mathbf{y}$) is a function of two individual-level covariates ($\mathbf{z_1}, \mathbf{z_2}$), one unit-level covariate ($\boldsymbol{\lambda}$), and a stylized geographic location. The DGP is linear-additive except in two geographic regions, where the Z variables have a multiplicative effect.[3] This produces a pattern that one might expect to observe in real data, where the relationship between demography, geography, and public opinion is not well captured by a linear and additively separable model (e.g. income is strongly associated with voting Republican in Mississippi, but weakly associated in Connecticut (Gelman *et al.* 2007)).

More formally, the data are generated through the following process. First, I create NM individuals, where M is the number of subnational units, and N is the number of observations per unit. Each individual has four latent (unobserved) characteristics, z_1 through z_4, drawn from a multivariate normal distribution with mean zero and variance–covariance matrix equal to

$$\begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}.$$

3 Replication materials are available at Ornstein (2019).

The variable $\mathbf{z_4}$ is used to assign each observation to a subnational unit, which ensures that there is cross-unit variation on the latent characteristics. Each subnational unit, in turn, is assigned a random latitude and longitude, drawn from a bivariate uniform distribution between $(0, 0)$ and $(1, 1)$. Once I assign each observation a $\mathbf{z}$ vector and subnational unit, I generate the outcome variable, y, using the following equation:

$$y_i = z_{1i} + z_{2i} + z_{3i} + \theta(D_i^0 z_{1i} z_{2i} - D_i^1 z_{1i} z_{3i}) + \varepsilon_i.$$

D^0 is a function that is decreasing in distance from (0,0), and D^1 is decreasing in distance to (1,1), so that multiplicative effects are strongest near those points. ε_i is an iid normal error term with mean zero and variance σ^2. The parameter θ governs the strength of the three-way interaction effect. When $\theta = 0$, the DGP is simply a linear-additive combination of the demographic variables, but as θ increases, the conditional effect of geography becomes stronger. Finally, I create discretized versions of the demographic variables $\mathbf{z_1}$ through $\mathbf{z_3}$, called $\mathbf{x_1}$ through $\mathbf{x_3}$. Although the outcome variable y is a function of the latent variables, Z, the researcher can only observe the discrete variables X. In addition, the researcher cannot observe $\mathbf{x_3}$ at the individual level, but instead observes its unit-level means (λ_m), which are included as a predictor in the first-stage model.

I repeatedly simulate this data generating process, varying the parameters ρ and θ. (See the Supplementary Materials for a more detailed technical description of the Monte Carlo and the combinations of parameter values used.) For each simulated population, I then draw a random sample of size n and generate three sets of subnational estimates: disaggregation, classical MRP, and SRP. The first-stage equation for the MRP estimation is a multilevel regression model of the following form, where x_1 and x_2 are the individual-level covariates and α_m^{unit} is a unit-specific intercept, itself a function of the unit-level covariate, λ_m:

$$\begin{aligned} y_i &= \beta^0 + \alpha_j[i]^{x_1} + \alpha_k[i]^{x_2} + \alpha_m^{\text{unit}} + \varepsilon_i; \\ \alpha_j^{x_1} &\sim N(0, \sigma_j^2); \\ \alpha_k^{x_2} &\sim N(0, \sigma_k^2); \\ \alpha_m^{\text{unit}} &\sim N(\beta^\lambda \cdot \lambda_m, \sigma_{\text{unit}}^2). \end{aligned}$$

For the first stage of the SRP, I fit a LASSO using the same predictors as the multilevel model and I train KNN, random forest, and GBM using $\mathbf{x_1}$, $\mathbf{x_2}$, $\boldsymbol{\lambda_m}$, latitude, and longitude as predictors. I use fivefold cross-validation for parameter tuning, and then estimate ensemble model weights through stacking.

Figure 1 illustrates the results of a representative run from the Monte Carlo simulation. Under certain conditions, SRP dramatically outperforms both disaggregation and classical MRP. When θ is large—and therefore the multilevel regression model is misspecified—the ensemble model average is better able to predict individual-level opinion than multilevel regression alone, which in turn produces better poststratified estimates.

Note that the component machine learning algorithms do not always perform strictly better than multilevel regression. When θ is small—and thus the true DGP is linear-additive—complex models provide no prediction advantage over ordinary least squares. Indeed, the flexibility of methods like KNN become a detriment when the sample size of the survey is small, as KNN

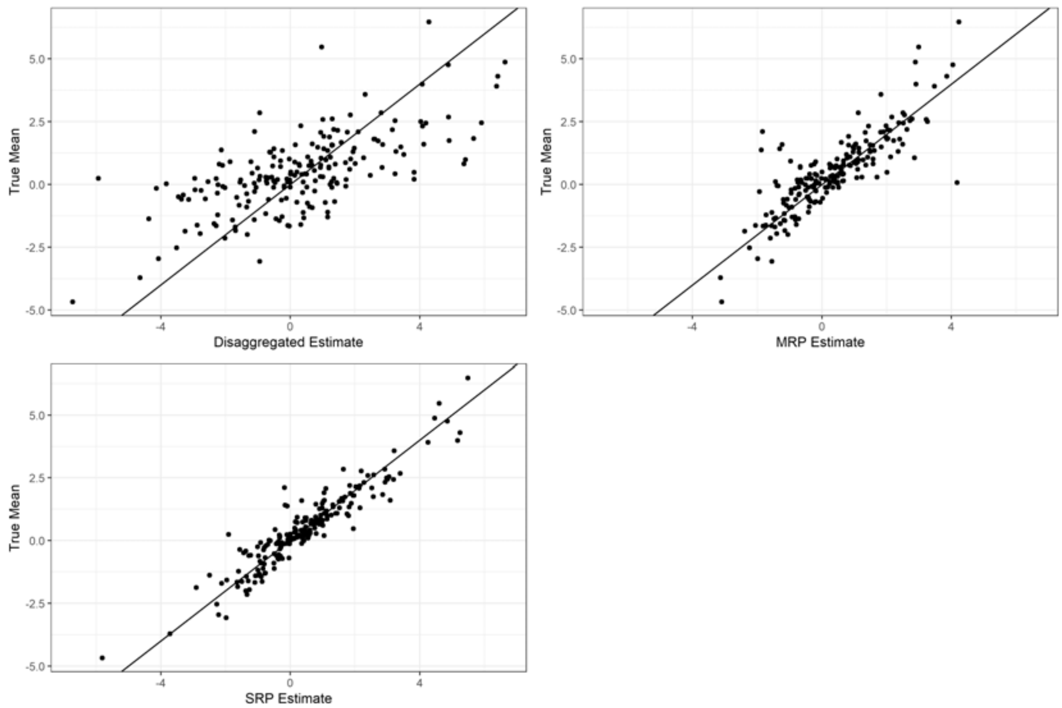

Figure 1. A representative simulation from the Monte Carlo analysis. Disaggregation, MRP, and SRP estimates are plotted against true subnational unit means. Parameter values: $\theta = 5$, $\rho = 0.4$, $N = 15\,000$, $M = 200$, $n = 5000$, $\sigma^2 = 5$.

performs poorly when the number of predictors is large relative to the size of the training set (Beyer *et al.* 1999).[4]

Nevertheless, the benefits of SRP can be dramatic under some conditions. In cases where θ and ρ are large, MRP performs only modestly better than disaggregation, while SRP produces estimates that are well-correlated with the true unit means. Figure 2 illustrates these relative performance gains for varying levels of θ. Even when SRP underperforms MRP, it never performs *poorly*: the worst correlation produced across all simulations was a 0.77 for SRP, compared to 0.78 for MRP and 0.25 for disaggregation.

4 Empirical Application

To demonstrate that SRP produces superior estimates in a wide variety of empirical applications, I replicate the results from Buttice and Highton (2013), comparing the performance of SRP and classical MRP. In that study, the authors use MRP to estimate state-level public opinion on 89 issues, drawn from two large-sample surveys of American public opinion, the National Annenberg Election Studies (NAES) and the Cooperative Congressional Election Studies (CCES). Because these surveys collect such a large sample within each state, the authors treat the state-level disaggregated means as the "true" values, and then test how well MRP performs at estimating these values after drawing smaller, random samples from the survey ($n = 1500$ or $n = 10\,000$). For each issue area, they model public opinion using a multilevel regression model with sex, age, race, and education as individual-level covariates, state-level covariates on presidential vote share and religious conservatism, and state- and region-specific intercepts.

4 More precisely, Beyer *et al.* (1999) show that KNN on high-dimensional data will perform poorly regardless of the size of n, owing to the "curse of dimensionality". Euclidean distance does not meaningfully measure "closeness" in spaces with more than 10–15 dimensions.

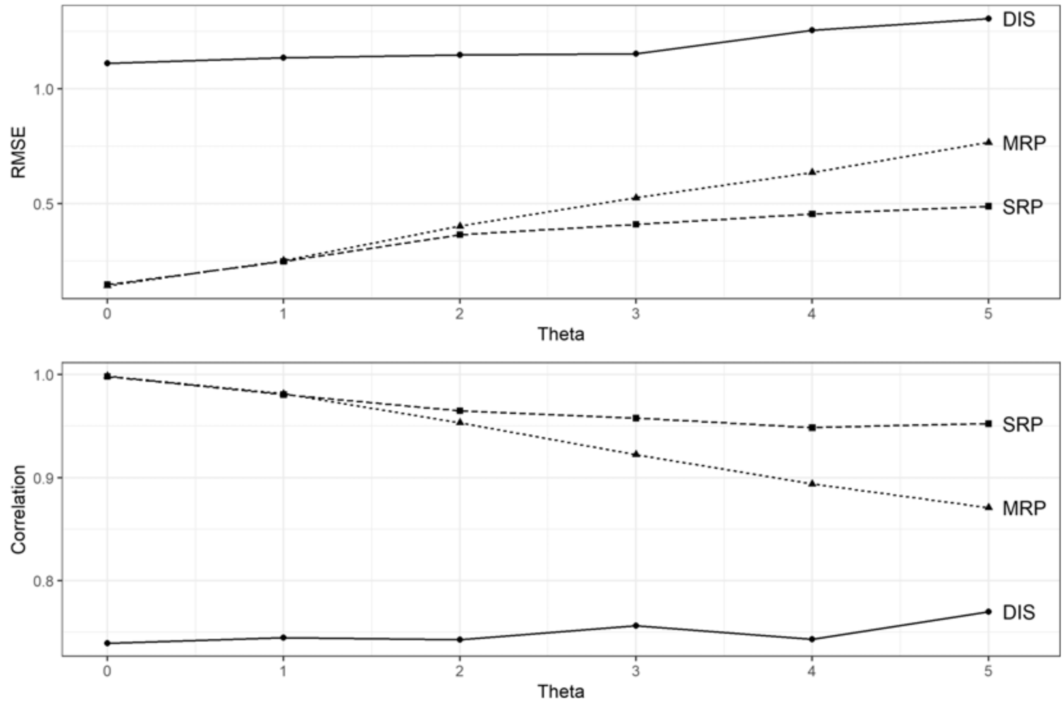

Figure 2. Relative performance of disaggregation, MRP, and SRP estimates, varying θ. Parameters Used: $\rho = 0.4$, $n = 5000$, $M = 200$, $N = 15\,000$, $\sigma^2 = 5$. Points denote 10-run averages.

I replicate this procedure for each of the 89 issue areas using both classical MRP and SRP, varying the size of the random sample drawn (n = 1500, 3000, 5000, and 10 000) and repeating the process five times for each combination. The MRP first-stage model is the same as in the original paper. The SRP ensemble includes a LASSO, KNN, random forest, and GBM model, using the same covariates included in Buttice and Highton (2013). For the KNN, random forest, and GBM, I substitute the latitude and longitude of each state's centroid for state- and region-level indicator variables. The stacking procedure typically selects weights that are mixtures of all five models. It is very rare in this empirical application that one model dominated the ensemble; fewer than 8% of ensembles yielded weights where $w_k > 0.8$ for any model k.

Figure 3 plots SRP's performance compared to classical MRP, varying the size of the random sample drawn. Across all simulations, SRP improves correlation in 79% of cases, and reduces mean average error (MAE) in 78% of cases. This performance improvement is the most consistent when working with larger sample sizes (rows 3 and 4). When $n = 10\,000$, SRP outperforms MRP in 88% of cases.

For issue areas where MRP already performs well, the gains from adopting SRP are modest. This is to be expected; when MRP produces a correlation of 0.9, there can be little improvement from fitting a more sophisticated first-stage ensemble. But in cases where MRP performs poorly, the gains from adopting SRP can be quite substantial. Examining Figure 3, one can observe numerous cases where SRP produces large improvements in correlation or MAE, and very few cases where it performs substantially worse. This is particularly true in Rows 3 and 4, where the estimates are constructed from larger samples.

To consider these performance gains in perspective, Figure 4 reports mean correlation and MAE across simulations, varying sample size. Adopting SRP yields a modest but statistically significant improvement across the board. In some places, this performance gain is comparable to a large increase in sample size. For example, when $n = 5000$, adopting SRP over MRP increases the

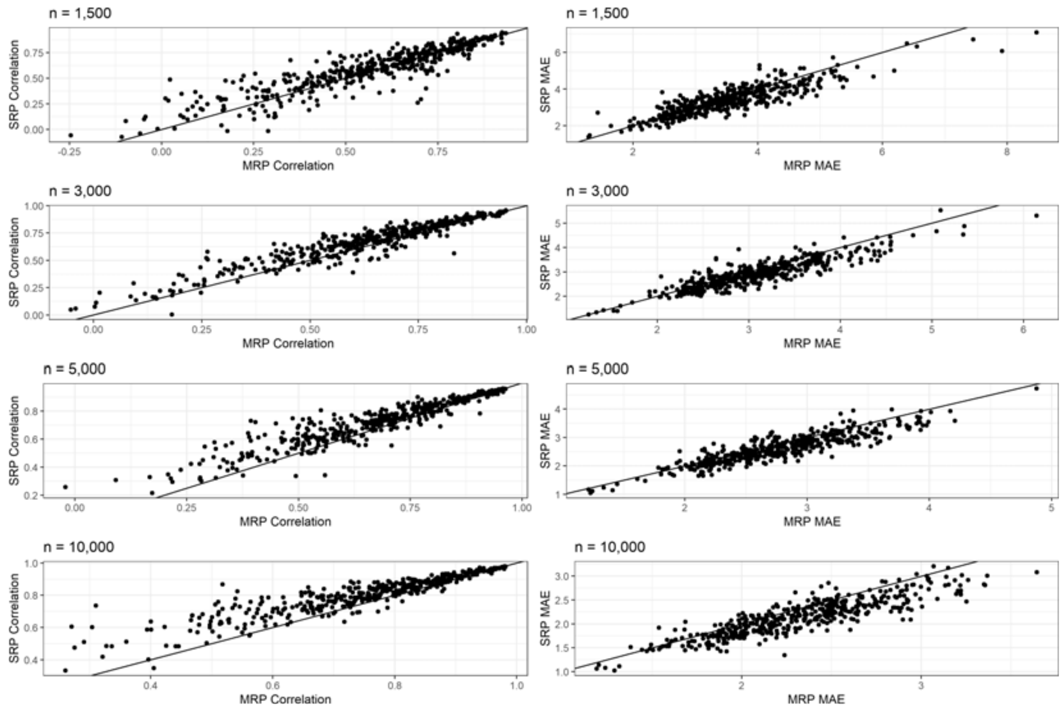

Figure 3. Plots in the left column compare SRP and MRP correlations, varying sample size. Plots in the right column compare Mean Absolute Error.

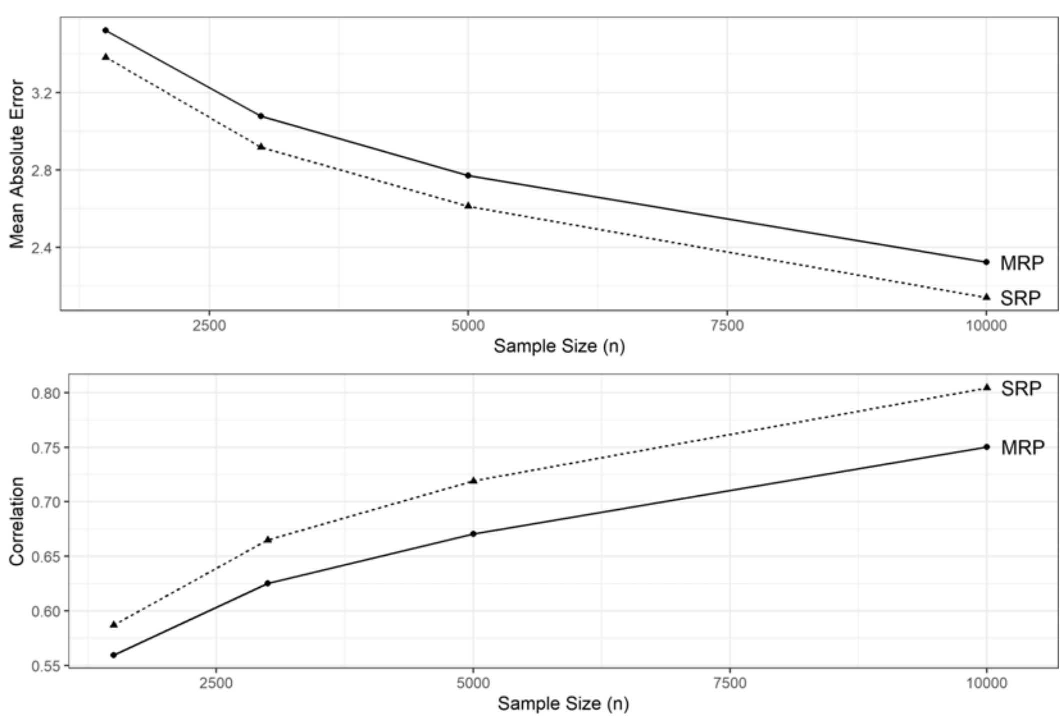

Figure 4. Mean performance of SRP and MRP across all simulations, varying sample size.

correlation from 0.68 to 0.72. This performance gain is comparable to doubling the sample size to $n = 10\,000$, a feat that is much more difficult in practice. (Doing both is, of course, best of all.)

The core finding from Buttice and Highton (2013) is that MRP's performance can be highly variable when using national-level surveys with a typical sample size of $n = 1500$. This remains true for SRP, albeit to a lesser extent. As the empirical application demonstrates, SRP performs best with large sample size. This is to be expected, given that more complex models require more data to fit. But it suggests that SRP alone is insufficient to expect good estimates from small, unrepresentative public opinion surveys. Rather, researchers should be most confident in estimates produced from large-sample surveys and cross-validated first-stage models.

5 Conclusion

In both Monte Carlo and empirical applications, SRP produces modest but consistent performance gains over classical MRP. These gains are largest in cases where the data generating process is complex, with nonlinear interactions between demographic and geographic variables that are unlikely to be specified in advance by the researcher's model. In the empirical analysis, SRP performs markedly better on issue areas where MRP performed poorly in Buttice and Highton (2013).

Despite this improvement in prediction performance, there are two reasons why a researcher might choose to forego SRP and employ classical MRP instead. The first involves "synthetic poststratification", a method proposed by Leemann and Wasserfallen (2017). This refinement of MRP proceeds as if the joint distribution of individual-level covariates is simply the product of their marginal distributions, allowing the researcher to include additional individual-level predictor variables in the first-stage model. In the Supplementary Materials (Appendix A), I prove that this procedure and classical MRP produce identical estimates, but only if the first-stage model is linear-additive. This suggests that, for applications where the researcher must conduct the poststratification stage synthetically —e.g. in countries that do not publish detailed census microdata— it is prudent to use a linear-additive model for the first-stage predictions rather than the ensemble methods proposed here.

The second cost of SRP is its added computational complexity; it takes significantly longer to estimate an ensemble of models than it does to estimate a single regression model. This is particularly true when gradient boosting is included in the ensemble, as it requires the most computationally intensive parameter tuning of all the methods I survey here (see the Supplementary Materials for details). To produce the results in this paper, I conducted several thousand simulations, which required a significant amount of computation time. But for applied researchers who need only to generate a single set of estimates, the added computation time is negligible. The maximum time spent on a single run was 15 minutes, the majority of which was spent tuning the GBM parameters. The average run time was closer to 3 minutes.

Given these results, I strongly recommend that researchers consider using SRP in place of MRP, particularly when working with large ($n > 5000$) public opinion datasets like CCES or NAES. To facilitate this, readers are welcome to adapt my replication code for their use, and I have developed an `R` package (`SRP`) currently available on my website. Ultimately, I hope that SRP will prove to be a useful addition to the empirical social scientist's toolkit, spurring further research into subnational politics.

Acknowledgements

The author thanks Rob Franzese, Scott Page, Liz Gerber, Paul Kellstedt, the anonymous reviewers, and workshop participants at the University of Michigan and the Midwest Political Science Association for their helpful comments and suggestions on earlier drafts.

Supplementary Material

For supplementary material accompanying this paper, please visit
https://doi.org/10.1017/pan.2019.43.

References

Beyer, K., J. Goldstein, R. Ramakrishnan, and U. Shaft. 1999. “When Is “Nearest Neighbor” Meaningful?” In *International Conference on Database Theory*, edited by C. Beeri and P. Buneman, 217–235. Berlin: Springer.

Breiman, L. 1996. “Stacked Regressions.” *Machine Learning* 24:49–64.

Breiman, L. 2001. “Statistical Modeling: The Two Cultures.” *Statistical Science* 16(3):199–231.

Buttice, M. K., and B. Highton. 2013. “How Does Multilevel Regression and Poststratification Perform with Conventional National Surveys?” *Political Analysis* 21(4):449–467.

Caruana, R., A. Niculescu-mizil, G. Crew, and A. Ksikes. 2004. “Ensemble Selection from Libraries of Models.” In *Proceedings of the Twenty-First International Conference on Machine Learning*, 18. New York: ACM.

Gelman, A. 2018. “Regularized Prediction and Poststratification (the generalization of Mister P).” *Statistical Modeling, Causal Inference, and Social Science* (blog), May 19. https://statmodeling.stat.columbia.edu/2018/05/19/.

Gelman, A., B. Shor, J. Bafumi, and D. Park. 2007. “Rich State, Poor State, Red State, Blue State: What’s the Matter with Connecticut?” *Quarterly Journal of Political Science* 2(June 2006):345–367.

Gelman, A., and T. C. Little. 1997. “Poststratification into Many Categories using Hierachical Logistic Regression.” *Survey Methodology* 23(2):127–135.

Ghitza, Y., and A. Gelman. 2013. “Deep Interactions with MRP: Election Turnout and Voting Patterns among Small Electoral Subgroups.” *American Journal of Political Science* 57(3):762–776.

Grimmer, J., S. Messing, and S. J. Westwood. 2017. “Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods.” *Political Analysis* 25:413–434.

van der Laan, M. J., E. C. Polley, and A. E. Hubbard. 2007. “Super Learner.” *Statistical Applications in Genetics and Molecular Biology* 6(1):1–25.

Lax, J. R., and J. H. Phillips. 2009. “How Should We Estimate Public Opinion in The States?” *American Journal of Political Science* 53(1):107–121.

Lax, J. R., and J. H. Phillips. 2012. “The Democratic Deficit in the States.” *American Journal of Political Science* 56(1):148–166.

Leblanc, M., and R. Tibshirani. 1996. “Combining Estimates in Regression and Classification.” *Journal of the American Statistical Association* 91(436):1641–1650.

Leemann, L., and F. Wasserfallen. 2017. “Extending the Use and Prediction Precision of Subnational Public Opinion Estimation.” *American Journal of Political Science* 61(4):1003–1022.

Montgomery, J. M., F. Hollenbach, and M. D. Ward. 2012. “Improving Predictions Using Ensemble Bayesian Model Averaging.” *Political Analysis* 20(3):271–291.

Montgomery, J. M., and S. Olivella. 2018. “Tree-Based Models for Political Science Data.” *American Journal of Political Science* 62(3):729–744.

Ornstein, J. T. 2019. “Replication Data for: Stacked Regression and Poststratification.” https://doi.org/10.7910/DVN/PRDURU, Harvard Dataverse, V1.

Park, D. K., A. Gelman, and J. Bafumi. 2004. “Bayesian Multilevel Estimation with Poststratification: State-level Estimates from National Polls.” *Political Analysis* 12(4):375–385.

Tausanovitch, C., and C. Warshaw. 2014. “Representation in Municipal Government.” *The American Political Science Review* 108(3):605–641.

Tibshirani, R. 1996. “Regression Shrinkage and Selection via the Lasso.” *Journal of the Royal Statistical Society, Series B (Methodological)* 58(1):267–288.

Warshaw, C., and J. Rodden. 2012. “How Should We Measure District-Level Public Opinion on Individual Issues?” *The Journal of Politics* 74(1):203–219.

Wolpert, D. H. 1992. “Stacked Generalization.” *Neural Networks* 5(2):241–259.

Yao, Y., A. Vehtari, D. Simpson, and A. Gelman. 2018. “Using Stacking to Average Bayesian Predictive Distributions.” *Bayesian Analysis* 13(3):917–1007.