SHORT ARTICLE

Improved Multilevel Regression with Poststratification through Machine Learning (autoMrP)

Philipp Broniecki, University of Oslo
Lucas Leemann, University of Zurich
Reto Wüest, University of Bergen

Multilevel regression with poststratification (MrP) has quickly become the gold standard for small area estimation. While the first MrP models did not include context-level information, current applications almost always make use of such data. When using MrP, researchers are faced with three problems: how to select features, how to specify the functional form, and how to regularize the model parameters. These problems are especially important with regard to features included at the context level. We propose a systematic approach to estimating MrP models that addresses these issues by employing a number of machine learning techniques. We illustrate our approach using 89 items from public opinion surveys in the United States and demonstrate that our approach outperforms a standard MrP model in which the choice of context-level variables has been informed by a rich tradition of public opinion research.

Multilevel regression with poststratification (MrP) has become the standard approach to estimating subnational public opinion using survey data that are only nationally representative. Compared to the older "disaggregation approach," which disaggregates the survey data by calculating subnational averages of public opinion, MrP relies on more structure to create more efficient opinion estimates. Most MrP models consist of two parts: a set of random effects for individual-level socioeconomic variables and a set of fixed effects for context-level variables. These models require researchers to choose variables, specify a functional form, and estimate parameters at the individual and context levels. Doing so at the context level is particularly challenging because the fixed effects parameters in a multilevel model are not shrunk toward the grand mean. Unlike individual-level variables that are included via random effects, context-level variables thus run the risk of overfitting the survey data. This risk is exacerbated by the fact that the number of observations at the second level is small in most applications.

In this article, we propose a systematic approach to measuring subnational public opinion. We borrow from the machine learning literature and modify the basic MrP model by introducing *systematic feature selection*, more *flexible functional forms*, and more *flexible regularization* of model parameters. Our approach is capable of providing an improved model that outperforms the standard MrP model as well as recent alternatives in terms of the mean squared prediction error (MSE). To showcase the benefits of the proposed approach, we use

Philipp Broniecki (philipp.broniecki@stv.uio.no; https://philippbroniecki.com/), Department of Political Science, University of Oslo, Norway. Lucas Leemann (leemann@ipz.uzh.ch; http://www.lucasleemann.ch), Department of Political Science, University of Zurich, Switzerland. Reto Wüest (reto.wueest@uib.no; http://www.retowuest.net/), Department of Comparative Politics, University of Bergen, Norway. The names of the authors are listed alphabetically.

Philipp Broniecki would like to acknowledge the support of the Business and Local Government Data Research Centre (ES/S007156/1) funded by the Economic and Social Research Council for undertaking this work. Lucas Leemann received funding from the Swiss National Science Foundation (grant 183120). Reto Wüest has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (ERC StG 2018 CONSULTATIONEFFECTS, grant 804288, and ERC AdG 2016 UneqDems, grant 741538). The R package autoMrP allows readers to readily apply the models presented in this article (see https://cran.r-project.org/package=autoMrP). Data and supporting materials necessary to reproduce the numerical results in the article are available in the *JOP* Dataverse (https://dataverse.harvard.edu/dataverse/jop). An online appendix with supplementary material is available at https://doi.org/10.1086/714777.

Published online October 19, 2021.

  Published by The University of Chicago Press for the Southern Political Science Association. https://doi.org/10.1086/714777

a large data set compiled by Buttice and Highton (2013) covering 89 survey items in the United States. We show that by using standard classifiers from the machine learning literature and employing a superlearner we can provide accurate estimates of subnational public opinion without relying on domain knowledge in public opinion research.

IMPROVING MRP

MrP has been successfully applied in a variety of contexts (Caughey and Warshaw 2018; Lax and Phillips 2009; Leemann and Wasserfallen 2017; Selb and Munzert 2011).[1] Meanwhile, its increased use has led to greater scrutiny, and some authors have offered a more cautionary view. Warshaw and Rodden (2012) show that MrP's performance depends on whether context-level information is exploited, and Buttice and Highton argue forcefully that strong context-level variables "emerge as a necessary but not sufficient condition for MrP to perform well" (2013, 464). However, to date, there is no clear guidance on how to systematically select and specify models that include context-level variables. Scholars agree that context-level variables are key to improving predictions, but they are generally selected in an ad hoc fashion, driven by personal intuition and domain knowledge.[2] We propose a systematic approach that allows scholars to make better use of context-level information. By relying on classifiers in addition to the multilevel model, we allow for more flexible functional forms and regularization.

MrP is a prediction model. The individual level of an MrP model includes only random effects, which are by definition shrunk toward the grand mean (Gelman and Hill 2007, 253) and provide (some) protection against overfitting. The contextual level commonly consists of two parts: the systematic part $\mathbf{X}_c'\beta$ and the random effect $\alpha_c^{\text{subnational unit}}$, where c indexes subnational units. The risk of overfitting comes from the elements of β, which are estimated as fixed parameters without shrinkage. Disregarding context-level information $\mathbf{X}_c$ may lead to underfitting since geographical variation can now only result from the random effect $\alpha_c^{\text{subnational unit}}$. Depending on the shrinkage of $\alpha_c^{\text{subnational unit}}$, subnational variation might well be underestimated. The extent to which it is underestimated is partly driven by subnational sample sizes, with smaller samples leading to more shrinkage. Both overfitting and underfitting diminish the prediction accuracy of the model. Hence, the question is how to best specify a model that increases prediction accuracy. We focus on context-level features for three reasons. First, as mentioned above, shrinkage at the individual level already provides protection against overfitting. Second, context-level variables have been shown to provide larger improvements (see, e.g., fig. 6 in Warshaw and Rodden 2012). Third, the risk of overfitting at the context level is typically larger due to the low number of subnational units.

A SYSTEMATIC APPROACH TO PREDICTION

Our approach relies on five classification methods to model individual response behavior and combines them via ensemble Bayesian model averaging (EBMA; Montgomery, Hollenbach, and Ward 2012). Note that our approach is fully flexible, allowing scholars to easily extend the set of classifiers by adding models. The classifiers we use are (i) multilevel regression with best subset selection of context-level predictors (Best Subset), (ii) multilevel regression with best subset selection of principal components of context-level predictors (PCA), (iii) multilevel regression with L1 regularization (Lasso), (iv) gradient boosting (GB), and (v) support vector machine (SVM).[3]

We combine the predictions of the individual classifiers by relying on a *superlearner* as is common in computer science (e.g., Van der Laan, Polley, and Hubbard 2007). Recent contributions in political science that use superlearners include Grimmer, Messing, and Westwood (2017) and Samii, Paler, and Daly (2016). Our approach relies on EBMA as proposed by Montgomery et al. (2012, 2015). The weights that determine each classifiers' contribution to the overall prediction depend on the classifiers' performance on new (i.e., previously unseen) data. The hyperparameter in EBMA is the tolerance. Following Montgomery et al. (2015), we optimize over seven candidate values for the tolerance that range from 1×10^{-2} to 1×10^{-5} (see app. sec. 5 for details).

Several recent contributions have exploited machine learning techniques to measure public opinion. Caughey and Hartman (2017) use L1 regularization to select variables for weighting to overcome nonresponse bias. Closer to our contribution, Goplerud et al. (2018) include L1 regularization in MrP. Our approach differs from theirs in that we rely on not only Lasso but also a number of other classifiers. Ornstein (2019) proposes an approach that is similar to ours but uses a slightly different set of classifiers. Two major differences are that we emphasize the context level and rely on EBMA rather than stacking to combine the individual classifiers. Finally, Bisbee

1. We provide an overview of the standard MrP model in app. sec. 1 (app. secs. 1–12 are available online).

2. Leemann and Wasserfallen (2016) provide an exception by selecting context-level variables based on the Akaike and Bayesian information criterion. They hence rely on penalized in-sample fit of the survey data rather than out-of-sample data fit.

3. See app. sec. 2 for a discussion of these classifiers. Montgomery and Olivella (2018) show how tree-based methods can be used in political science; one of their illustrations involves MrP.

(2018) modifies MrP by replacing the multilevel model with Bayesian additive regression trees (BARP), leading to significant improvements in prediction performance. While he restricts the set of covariates in the model to those that have been used by Buttice and Highton (2013), we leverage additional context-level information and combine the predictions of various classifiers. In what follows, we compare the performance of our approach to that of the standard MrP model and, in the appendix, also to the performance of the BARP model (where we show a test in which autoMrP outperforms BARP).

PERFORMANCE OF OUR APPROACH

To illustrate the performance of our approach, we use public opinion data from the United States compiled by Buttice and Highton (2013). The data consist of 89 items that were asked of at least 25,000 respondents in either of two surveys, the National Annenberg Election Studies (2000, 2004, and 2008) and the Cooperative Congressional Election Studies (2006 and 2008) (Buttice and Highton 2013, 454–55). We follow Buttice and Highton and treat all respondents who answered an item as the item-specific population. For each such population, the "true" public opinion in a state is calculated as the share of respondents in that state answering yes to the respective item. We then draw a sample of 1,500 respondents from the population and, using this sample, predict state public opinion. To evaluate the performance of our approach, we compare our predictions to the true state opinions.[4]

Our prediction of state public opinion involves three steps. First, we remove from each sample 1/3 of the respondents (i.e., 500 out of 1,500 respondents) and set them aside for the second step, the EBMA step. We then use the remaining 1,000 respondents to train and evaluate each individual classifier using K-fold cross-validation. In so doing, we randomly partition these respondents into $K = 5$ roughly equal-sized folds but include all respondents from the same state in the same fold. For each fold $k \in \{1, \ldots, K\}$, we train our five classifiers on all folds but the kth, on the basis of which we evaluate them by calculating the MSE. Averaging the MSEs over all held-out folds provides an estimate of the expected extrasample MSE (Hastie, Tibshirani, and Friedman 2009, 241–45). Note that we use the average individual error in our calculation of the MSE (see app. sec. 2 for details). Using five folds turned out to be a reasonable choice for our data. We also performed cross-validation with other values of K (e.g., $K = 10$). These led to similar results but increased computing time. Second, in the EBMA step, we combine the models of the individual classifiers with the lowest average MSE to generate an ensemble prediction for each respondent profile defined by the sociodemographic and geographic variables. The weights of the individual models are determined on the basis of the 500 respondents we have set aside, thus avoiding "double dipping." Third, we poststratify the demographic-geographic profiles to obtain state-level predictions that we can then compare to the true state public opinions.

The MrP model used by Buttice and Highton (2013) includes, at the individual level, random effects for respondents' age group (four categories), education level (four categories), and gender-race combination (six categories). At the context level, it contains variables for states' share of votes for the Republican candidate in the previous presidential election and percentage of Evangelical Protestant or Mormon respondents. We treat this as the baseline model against which we compare our approach. Since our approach aims to provide researchers with a disciplined, automated way of building a prediction model that does not require extensive domain knowledge, we augment the set of context-level variables. In addition to the two variables included in the baseline model, we consider states' percentage of the population living in urban areas, unemployment rate, share of Hispanics, and share of whites as candidate variables.

In addition to our combined approach (EBMA), we poststratify the predictions of each of our individual classifiers (Best Subset, PCA, Lasso, GB, SVM), the baseline model (Baseline), an "empty" model that does not contain any context-level information (No Vars), and a "full" model that includes all available context-level variables (All Vars). This allows us to compare our approach not only to the Buttice and Highton (2013) baseline model that is informed by years of public opinion research in the United States but also to a model that maximizes parsimony and one that maximizes in-sample data fit.

Figure 1 shows the MSEs of our combined approach, our five individual classifiers, the baseline model, the model without any context-level variables, and the model including all context-level variables. EBMA outperforms all other approaches. Most importantly, it improves on the baseline model by 12%.[5] We consider this a significant improvement since the Buttice and Highton (2013) model likely provides a hard test for the relative performance of our approach. The 89 survey items in our data are all about political issues on which we expect the baseline model specified by Buttice and Highton to perform well in predicting state public opinion. Contemporary US politics is characterized by a single dimension of conflict (Poole

4. We also employ an alternative strategy in which we first rake the megasample and create state-level truths. This leads to virtually identical results (see app. sec. 4).

5. We also note that the improvement offered by our approach is somewhat larger than that resulting from BARP (see app. sec. 3 for a comparison on the same 89 items; Bisbee 2018).

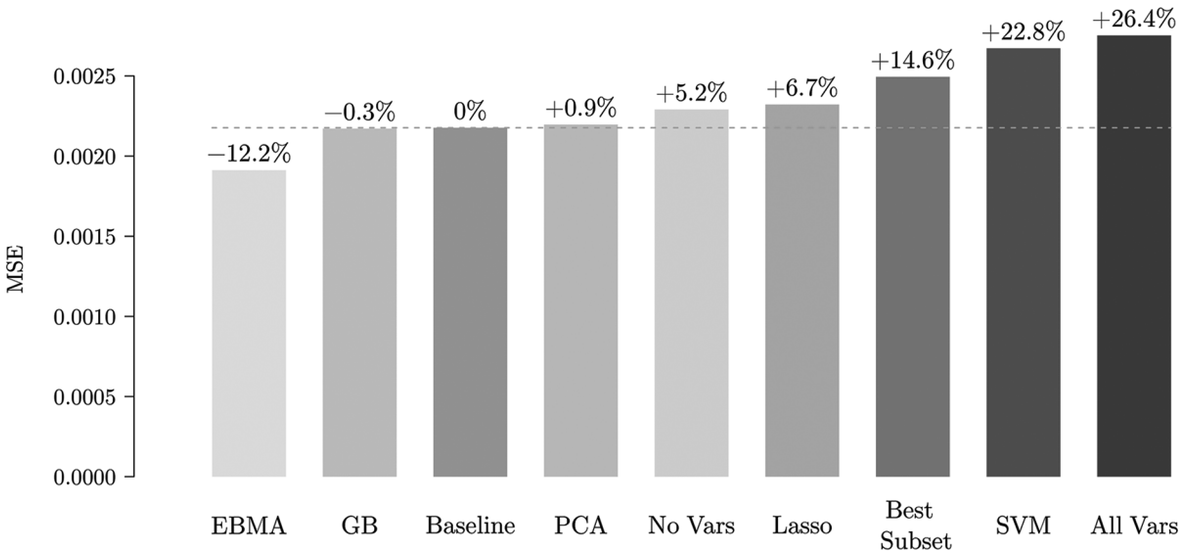

Figure 1. Comparison of prediction performance. Average mean squared prediction error (MSE) of state-level predictions over 89 survey items. Baseline model is from Buttice and Highton (2013), No Vars is empty at the context level, and All Vars includes all six context-level variables. *Dashed line*: MSE of Buttice and Highton model. *Percentages*: Comparison to Buttice and Highton model. Color version available as an online enhancement.

and Rosenthal 2011), and the two context-level predictors in the Buttice and Highton model likely explain much of the state-level variation on this dimension. This might explain why the (relatively sparse) baseline model performs well and why our approach shows only moderate improvement, in this application. In contexts in which strong predictors, such as presidential vote share, are not available, we may expect the improvement of our approach to be larger. Specifying a model based on substantive knowledge is likely to be more difficult in countries with a larger number of political conflict dimensions or less extensive research on public opinion, in which case our approach might also lead to a more significant improvement over a standard MrP model.

With regard to the other approaches commonly used in the literature, the approach we propose reduces the MSE of the model including all available context-level variables by 31% and the MSE of the model without any context-level variables by 17%. Our results show that our combined approach outperforms every single classifier taken individually.

CONCLUSION

We leverage insights from the machine learning literature and bring feature selection, flexible functional forms, and regularization to bear on the problem of how to best specify an MrP model for small area estimation. We focus on the context level since in a multilevel model parameters at the individual level are already moderated by shrinkage (partial pooling), whereas parameters at the context level are not regularized. Disregarding context-level information altogether may appear to be an easy solution, yet Warshaw and Rodden (2012) have shown that including context-level variables can greatly improve the performance of MrP.

We propose a data-driven approach to specifying MrP models. Our approach tunes five classifiers and combines them via EBMA into an overall prediction. We also provide an R package (autoMrP) that allows researchers to apply our approach easily. Appendix section 10 provides an example of how the package can be used.[6] We evaluated the performance of our approach using public opinion data from the United States. The results show that it outperforms alternative approaches commonly used in the literature. Most importantly, it reduces by 12% the MSE of a model informed by substantive knowledge. We consider this application to be a "hard test" since US public opinion is well studied and US political conflict tends to be structured by a single dimension. In contexts that are less well studied and characterized by multiple dimensions of conflict, our approach might lead to even larger improvements over models informed by substantive knowledge. The results also showed that the combined approach dominates all of its constituent classifiers. The combination of classifiers thus is important, and our approach can easily be extended by the inclusion of additional methods.

ACKNOWLEDGMENTS

We thank Sandra Boyd, Andy Guess, Guy Grossman, Nils Metternich, Slava Jankin, Santiago Olivella, and Marco Steenbergen for helpful discussions and comments. Earlier versions

6. In app. sec. 9 we show how users can derive the uncertainty of the estimates.

of the article were presented at the 2016 annual meetings of the European Political Science Association, American Political Science Association, and Southern Political Science Association as well as at the 2017 PSA meeting.

REFERENCES

Bisbee, James. 2018. "BARP: Improving Mister P Using Bayesian Additive Regression Trees." *American Political Science Review* 113 (4): 1060–65.

Buttice, Matthew K., and Benjamin Highton. 2013. "How Does Multilevel Regression and Poststratification Perform with Conventional National Surveys?" *Political Analysis* 21 (4): 449–67.

Caughey, Devin, and Erin Hartman. 2017. "Target Selection as Variable Selection: Using the Lasso to Select Auxiliary Vectors for the Construction of Survey Weights." Unpublished manuscript.

Caughey, Devin, and Christopher Warshaw. 2018. "Public Opinion in Subnational Politics." *Journal of Politics* 81 (1): 352–63.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Goplerud, Max, Shiro Kuriwaki, Marc Ratkovic, and Dustin Tingley. 2018. "Sparse Multilevel Regression (and Poststratification (sMRP))." Unpublished manuscript, Harvard University.

Grimmer, Justin, Solomon Messing, and Sean J. Westwood. 2017. "Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods." *Political Analysis* 25 (4): 413–34.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. 2nd ed. New York: Springer.

Lax, Jeffrey R., and Justin H. Phillips. 2009. "How Should We Estimate Public Opinion in the States?" *American Journal of Political Science* 53 (1): 107–21.

Leemann, Lucas, and Fabio Wasserfallen. 2016. "The Democratic Effect of Direct Democracy." *American Political Science Review* 110 (4): 750–62.

Leemann, Lucas, and Fabio Wasserfallen. 2017. "Extending the Use and Prediction Precision of Subnational Public Opinion Estimation." *American Journal of Political Science* 61 (4): 1003–22.

Montgomery, Jacob M., Florian M. Hollenbach, and Michael D. Ward. 2012. "Improving Predictions Using Ensemble Bayesian Model Averaging." *Political Analysis* 20 (3): 271–91.

Montgomery, Jacob M., Florian M. Hollenbach, and Michael D. Ward. 2015. "Calibrating Ensemble Forecasting Models with Sparse Data in the Social Sciences." *International Journal of Forecasting* 31 (3): 930–42.

Montgomery, Jacob M., and Santiago Olivella. 2018. "Tree-Based Models for Political Science Data." *American Journal of Political Science* 62 (3): 729–44.

Ornstein, Joseph T. 2019. "Stacked Regression and Poststratification." *Political Analysis* 28 (2): 239–301.

Poole, Keith T., and Howard L. Rosenthal. 2011. *Ideology and Congress*, vol. 1. Piscataway, NJ: Transaction.

Samii, Cyrus, Laura Paler, and Sarah Zukerman Daly. 2016. "Retrospective Causal Inference with Machine Learning Ensembles: An Application to Anti-recidivism Policies in Colombia." *Political Analysis* 24 (4): 434–56.

Selb, Peter, and Simon Munzert. 2011. "Estimating Constituency Preferences from Sparse Survey Data Using Auxiliary Geographic Information." *Political Analysis* 19 (4): 455–70.

Van der Laan, Mark J., Eric C. Polley, and Alan E. Hubbard. 2007. "Super Learner." *Statistical Applications in Genetics and Molecular Biology* 6 (1): 25.

Warshaw, Christopher, and Jontahan Rodden. 2012. "How Should We Measure District-Level Public Opinion on Individual Issues?" *Journal of Politics* 74 (1): 203–19.