American Political Science Review (2024) 118, 1, 529–536
doi:10.1017/S0003055423000035

Letter

Re-Evaluating Machine Learning for MRP Given the Comparable Performance of (Deep) Hierarchical Models

MAX GOPLERUD *University of Pittsburgh, United States*

Multilevel regression and post-stratification (MRP) is a popular use of hierarchical models in political science. Multiple papers have suggested that relying on machine learning methods can provide substantially better performance than traditional approaches that use hierarchical models. However, these comparisons are often unfair to traditional techniques as they omit possibly important interactions or nonlinear effects. I show that complex ("deep") hierarchical models that include interactions can nearly match or outperform state-of-the-art machine learning methods. Combining multiple models into an ensemble can improve performance, although deep hierarchical models are themselves given considerable weight in these ensembles. The main limitation to using deep hierarchical models is speed. This paper derives new techniques to further accelerate estimation using variational approximations. I provide software that uses weakly informative priors and can estimate nonlinear effects using splines. This allows flexible and complex hierarchical models to be fit as quickly as many comparable machine learning techniques.

The growing popularity of machine learning continues to revolutionize parts of political science by allowing easy estimation of flexible and powerful models. One increasingly popular application is using machine learning when performing "multilevel regression and post-stratification" (MRP) to extrapolate nationally representative surveys to smaller geographic units such as states (Lax and Phillips 2009; Park, Gelman, and Bafumi 2004). MRP is a two-step process that begins by fitting a predictive model to the survey using demographic and state-level information. Next, opinion estimates for the states are obtained by a weighted average of the predicted values for various demographic groups inside of that state using their known distribution. While the performance of MRP depends on *both* steps, multiple papers have found that using machine learning for the predictive model outperforms traditional methods ("multilevel regression") by considerable margins (e.g., Bisbee 2019; Broniecki, Leemann, and Wüest 2022; Goplerud et al. 2018; Ornstein 2020). A plausible justification is that the linear, additive, nature of traditional models is insufficient to capture the complex relationship between the covariates and the outcome.

The reliance on *simple* hierarchical models, however, unnecessarily limits their usefulness. Unlike some machine learning methods that can automatically estimate interactions (or nonlinear effects of continuous predictors), hierarchical models can only estimate interactions that the researcher has explicitly included. This is both a strength and a weakness. While hierarchical models are highly modular and allow the researcher to explicitly incorporate domain-specific knowledge as to important predictors or interactions, there is a risk of mis-specification—and thus worse performance—if important interactions are omitted. Thus, a "fair" test of MRP's performance must examine a model that explicitly includes many possibly relevant interactions. Ghitza and Gelman (2013) illustrate this by adding a broad set of interactions and uncover considerably more subtle results than traditional methods could identify. Following their usage, I refer to complex hierarchical models that explicitly include interactions or nonlinear effects as "deep MRP."[1]

Thus, despite the understandable enthusiasm for applying machine learning to MRP, it is simply unknown in a systematic way whether machine learning outperforms deep MRP. The main reason for this gap in the literature is a practical one. Existing uses of deep MRP sometimes include nearly 20 random effects to capture the underlying heterogeneity and thus are usually very slow to estimate. Given that one might wish to fit these models repeatedly (e.g., comparing different specifications), this has quite reasonably caused researchers to "rule out" deep MRP.

Fortunately, recent work has shown that deep MRP can be estimated very quickly using variational inference while producing very similar point estimates to traditional methods (Goplerud 2022). However, that paper only tested those algorithms on the single dataset

Max Goplerud, Assistant Professor, Department of Political Science, University of Pittsburgh, United States, mgoplerud@pitt.edu.

Received: August 16, 2021; revised: May 31, 2022; accepted: January 18, 2023. First published online: March 03, 2023.

[1] This method is distinct from "deep learning" (e.g., involving neural networks).

from Ghitza and Gelman (2013) and did not compare them against machine learning techniques. My initial systematic tests found that those algorithms performed unfavorably against machine learning. This paper provides two improvements to existing variational methods that result in competitive performance: First, Goplerud (2022) relied on an improperly calibrated prior that often resulted in too little regularization. Second, those algorithms cannot capture nonlinear effects of continuous covariates (e.g., presidential vote share).

Those concerns are addressed by, first, extending the variational algorithms to include a weakly informative prior (Huang and Wand 2013) that can more appropriately regularize random effects and, second, allowing the use of penalized splines for continuous predictors. After implementing a number of novel computational techniques to accelerate estimation, the accompanying open-source software can fit highly flexible deep MRP in minutes—rather than the hours possibly needed for traditional approaches.

This paper illustrates the importance of deep MRP by reanalyzing two papers that suggest that machine learning methods clearly outperform MRP (Bisbee 2019; Ornstein 2020). It demonstrates two important stylized facts: deep multilevel models (i) are given considerable weight in an ensemble of machine learning methods and (ii) are competitive with Bayesian additive regression trees (BART) in terms of performance. While recent work reports that BART performs noticeably better than traditional MRP (Bisbee 2019), I demonstrate that this is not the case. I show that, especially at moderate sample sizes, BART usually only slightly outperforms even traditional MRP. Thus, while machine learning methods that combine many methods together in an ensemble can improve performance, (deep) MRP should continue to be used as a highly competitive single method or in any ensemble approach.

FITTING DEEP MRP FAST

The key limitation in fitting MRP with interactions is the speed of estimation. Earlier research has shown that fitting a single deep MRP model can take multiple hours (e.g., Goplerud 2022). This is because of the presence of high-dimensional integrals that traditional methods either numerically approximate or address using Bayesian methods.

Variational inference provides a different approach for fast estimation; the goal is to find the best approximating distribution to the posterior given some simplifying assumption—usually that blocks of parameters are independent (Grimmer 2011). However, the accuracy of this approximation can depend heavily on the specific problem, and thus needs extensive testing to ensure its reliability. Goplerud (2022) derived a new general algorithm for binomial hierarchical models and conducted extensive explorations of its performance on the single dataset considered in Ghitza and Gelman (2013). Those algorithms fit an extremely complex hierarchical model in around 1 minute—versus hour(s) for existing approaches. It demonstrated excellent performance by recovering posterior means on coefficients and predictions that closely aligned with the gold standard approach of Bayesian estimation.[2] Appendix A of the Supplementary Material provides a full exposition of the variational algorithm.

To illustrate the extensions in this paper, I focus on a simplified MRP model: Equation 1 shows a hierarchical model without fixed effects and with a random intercept for state and a random intercept for race, where y_i is the (binary) response for observation i. The notation follows Gelman and Hill (2006) where $\alpha^{\text{state}}_{g[i]}$ selects the random effect for the state g of which observation i is a member. Appendix A of the Supplementary Material shows the generalization to random slopes, arbitrary numbers of random effects, and fixed effects.

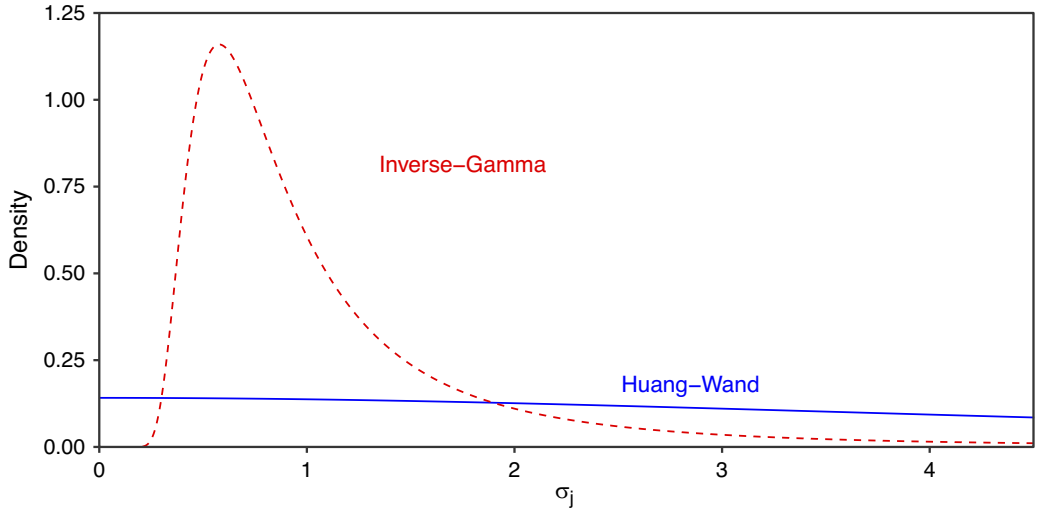
$$\begin{aligned} &y_i \sim \text{Bern}(p_i); \quad p_i = \frac{\exp(\psi_i)}{1+\exp(\psi_i)}; \quad \psi_i = \beta_0 + \alpha^{\text{state}}_{g[i]} + \alpha^{\text{race}}_{g'[i]}; \\ &\alpha^{\text{state}}_{g} \sim N(0, \sigma^2_{\text{state}}); \quad \alpha^{\text{race}}_{g'} \sim N(0, \sigma^2_{\text{race}}); \quad p(\beta_0) \propto 1; \\ &\sigma^2_j \sim p_0(\sigma^2_j) \quad \text{for} \quad j \in \{\text{state}, \text{race}\}. \end{aligned} \tag{1}$$

The choice of prior on the variance of the random effect $p_0(\sigma^2_j)$ is a difficult task. Some inferential techniques assume a flat prior. A risk of this strategy is that point estimates of σ^2_j could be degenerate and equal zero; this sets all random effects estimates equal to zero. This problem is rather common for the Laplace approximation in `glmer` (Chung et al. 2015). A proper prior prevents this problem and thus is preferable. An Inverse-Gamma prior is a popular choice, but it is difficult to calibrate the strength correctly (Gelman 2006). Figure 1 illustrates this by showing the prior used in Goplerud (2022) ("Inverse-Gamma") against the Huang and Wand (2013) prior employed in this paper. The latter prior implies the popular half-t prior on the standard deviation σ_j (Gelman 2006), and it can be generalized to multidimensional random effects. In that case, it imposes half-t priors on each marginal standard deviation while maintaining (if desired) a uniform prior on the correlations. Appendix A.1 of the Supplementary Material provides more information on this prior such as its density.

Figure 1 shows that Goplerud's (2022) prior puts effectively no mass on small values of σ_j (e.g., $P(\sigma_j \leq 0.25) \approx 0.0003$). Thus, in the event where the true value is small (i.e., the random effect is mostly irrelevant), the prior results in too large estimates of σ_j, thereby underregularizing the coefficients, which likely results in poorer performance. By contrast, the Huang–Wand prior puts nontrivial weight on very small σ_j and thus allows for strong regularization when appropriate. Appendix A.1 of the Supplementary Material provides a stylized example of this phenomenon.

[2] As with most variational methods, it underestimates posterior uncertainty; Goplerud (2022) provides a post-estimation adjustment to mitigate some of this problem.

FIGURE 1. Comparing Prior Density for Random Effect Standard Deviation σ_j



Note: The dashed line shows the prior density on σ_j given an Inverse-Gamma prior on σ_j^2 with $\alpha_0 = 1$ and $\beta_0 = 1/2$. The solid line shows a Huang–Wand prior on σ_j^2 with $\nu = 2$ and $A = 5$ (i.e., half-*t* on σ_j).

Unfortunately, Appendix A.5 of the Supplementary Material illustrates that naively incorporating the Huang–Wand prior dramatically increases estimation time. While it does increase the time per iteration, the major problem is that estimation requires 5–10 times more iterations to converge. Thus, a key contribution of this paper is to accelerate variational algorithms when this more appropriate prior is employed. Appendices A.2 and A.3 of the Supplementary Material provide, respectively, full explanations of the techniques employed: (i) a squared iterative method and (ii) a novel application of parameter expansion.

The model in Equation 1 can be extended by adding many interactions between geographic and demographic factors ("deep MRP"; Ghitza and Gelman 2013). However, Broniecki, Leemann, and Wüest (2022) note that additional state-level predictors (e.g., unemployment rate) may also provide considerable benefits. Unlike hierarchical models, many machine learning methods can automatically estimate nonlinear effects or interactions between these continuous predictors, whereas they must be specified explicitly for MRP.

I address that scenario by allowing estimation of nonlinear effects using splines as in a generalized additive model. Appendix A.4 of the Supplementary Material demonstrates how splines can be represented as additional hierarchical terms and thus estimated using the same variational algorithms.

Appendix B of the Supplementary Material provides simulations to illustrate the importance of using hierarchical models that include interactions or nonlinear effects. It shows that ignoring important interactions or nonlinearities hurts the performance of hierarchical models vis-à-vis alternative models, especially as the sample size increases. However, after those terms are included, hierarchical models perform well even against machine learning methods.

COMPARING METHODS FOR FITTING MRP

To compare deep hierarchical models against machine learning systematically, I use Buttice and Highton's (2013) popular dataset for validating new methods for MRP (e.g., Bisbee 2019; Broniecki, Leemann, and Wüest 2022; Ornstein 2020). It consists of 89 policy questions that are collected from multiple years of the National Annenberg Election Studies (2000, 2004, and 2008) and the Cooperative Congressional Election Studies (2006 and 2008). The benefit of these large samples is that it is possible to use the entire dataset to get a "ground truth" by taking the observed average in each state while drawing a smaller subsample (e.g., 1,500 respondents) to mimic the conditions under which a researcher would need to apply MRP to obtain reliable state-level estimates.

Existing comparisons, however, only rely on a simple hierarchical model outlined below (Equation 2), following the original specification in Buttice and Highton (2013). The model includes random effects for age, education (`educ`), gender–race combination (`gXr`), state, and region. The state-level continuous predictors `pvote` (state-level Republican presidential two-party vote share) and `relig` (share of population identifying as Evangelical Protestant or Mormon) are indexed with $g[i]$ as they are constant within a state.

$$\Pr(y_i = 1) = \text{logit}^{-1}\left(\begin{array}{l}\beta_0 + \beta_{\text{pvote}} \cdot \text{pvote}_{g[i]} + \beta_{\text{relig}} \cdot \text{relig}_{g[i]} + \\ \alpha^{\text{age}}_{g[i]} + \alpha^{\text{educ}}_{g[i]} + \alpha^{\text{gXr}}_{g[i]} + \alpha^{\text{state}}_{g[i]} + \alpha^{\text{region}}_{g[i]}\end{array}\right),$$
$$\alpha^j_g \sim N\left(0, \sigma^2_j\right) \text{ for all } j \text{ and } g. \tag{2}$$

This model includes no interactions between variables or nonlinear effects on continuous predictors,

and thus is likely insufficiently rich to capture the true underlying relationship. It is reasonable to suspect that a "properly specified" MRP model should include at least some interactions to be competitive with methods that can automatically learn interactions or nonlinearities.

I consider three expansions of this model's hierarchical component. First, I consider a deep MRP where all two-way interactions between demographics and geography are included (e.g., age–education and age–state), as well as a triple interaction between the three demographic variables. Second, I add splines to capture possible nonlinear effects in the state-level continuous variables. A third model includes both extensions. Table 1 summarizes the specifications; Appendix F of the Supplementary Material provides a demonstration of how to fit these models in the accompanying software (Goplerud 2023).[3]

It is important to stress that this paper tracks the existing analyses comparing machine learning and MRP as closely as possible. There are thus other specifications that likely improve upon Table 1, although I show that adding this set of interactions enables MRP to perform competitively against state-of-the-art machine learning techniques.

DEEP MRP IN AN ENSEMBLE

The first comparison explores whether deep MRP adds much benefit when used alongside a suite of machine learning methods. I begin by using a technique known as "stacking" that takes the predictions of many different methods and combines them into a single prediction known as an ensemble. Ornstein (2020) applied this method to MRP and found considerable gains over traditional methods. The method performs K-fold cross-validation to get an out-of-sample prediction for each observation in the survey using each constituent model. The out-of-sample predictions are combined to see which weighted average (convex combination) best predicts the outcome, and then these weights are used to combine predictions before post-stratification. It is often the case that the ensemble outperforms any single method (Broniecki, Leemann, and Wüest 2022; Ornstein 2020), although this is not guaranteed and can be empirically assessed by, for example, using a held-out dataset.

A useful property of ensembles is the ability to compare the weights given to the constituent models. The weights reflect both the performance of the method and its "distinctiveness" from the other methods in the ensemble. Using each survey in Buttice and Highton (2013), I drew 10 different samples of varying sizes and estimated an ensemble using fivefold cross-validation with the models in Ornstein (2020) where I swapped the traditional ("Simple") MRP model with the deep MRP model from Table 1. Figure 2 summarizes the weights given to each model, averaging across the surveys and simulations.

The results provide clear support for the importance of deep MRP in an ensemble: it is the highest weighted method when the sample size is over 1,500 and is given over 40% of the total weight when the sample size is 5,000 or higher. The performance of deep MRP is corroborated by the fact that, of the methods in the ensemble, it has the lowest cross-validated error on the survey data when $N > 1,500$. In terms of computational time, fitting this deep model on the full survey takes around 30 seconds for the largest sample size of 10,000 observations. Thus, deep MRP can be added to an ensemble with limited cost.

Appendix D of the Supplementary Material explores the trade-off between model complexity and sample size. It examines a larger ensemble that includes all four models from Table 1.[4] While corroborating Figure 2—MRP models collectively receive around 40%–50% of the weight—it shows an expected trade-off between traditional and deep MRP where traditional (noninteractive) methods are given decreasing weight as the sample size increases. This suggests that the ensemble upweights more complex methods as the amount of

TABLE 1. Deep MRP Specifications

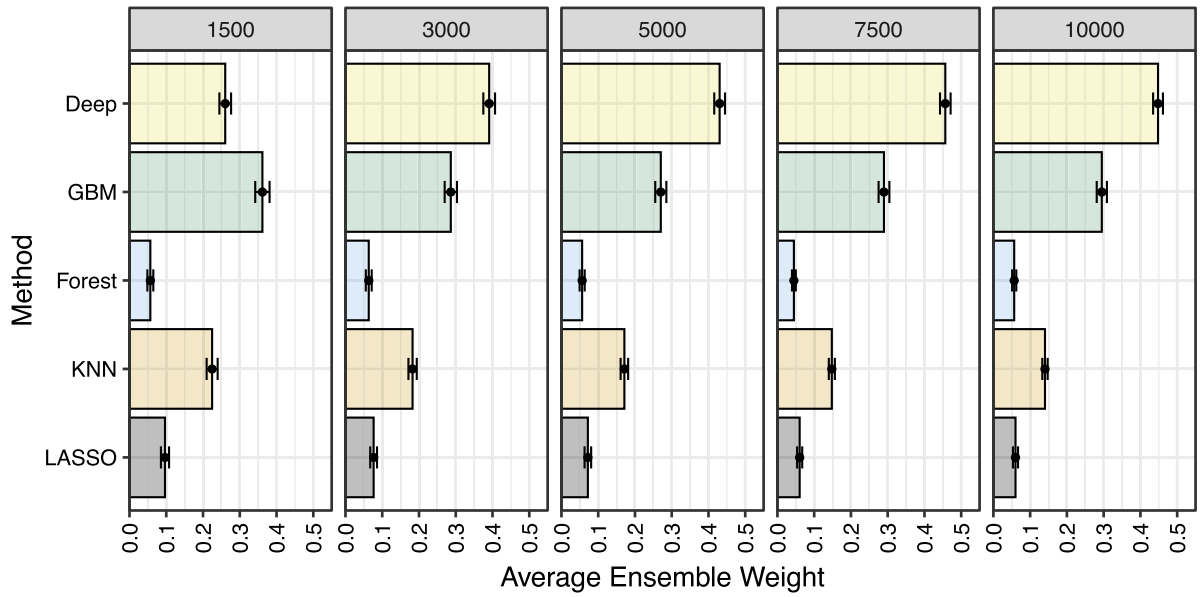
Model	`pvote` and `relig`	Demographics and state
Simple	Linear	Additive
Deep	Linear	Interacted
Splines	Splines	Additive
Combined	Splines	Interacted

Note: The second column indicates how these two variables are included. It is either "Linear" (Equation 2) or "Splines" where a spline is used to allow for nonlinear effects for each variable. The third column indicates how the random effects on age, education, gender × race, state, and region are included. "Additive" refers to five random effects added together (Equation 2). "Interacted" refers to including the interactions noted in the main text alongside the additive terms.

[3] All methods use a Huang–Wand prior for each random effect; hyperparameters are identical to Figure 1.

[4] It also examines the "deep" hierarchical model with an Inverse-Gamma/Wishart prior.

FIGURE 2. Weights Given to Models in Ensemble



Note: This figure shows the ensemble weights averaged across all surveys and 10 simulations per survey. Each panel reports the sample size of the survey. Ninety-five percent confidence intervals are shown. The first four methods are from Ornstein (2020): LASSO, *k*-Nearest Neighbors (KNN), Random Forest (Forest), Gradient Boosting Machine (GBM). The final method is "Deep" MRP from Table 1.

data increases. The spline-based methods receive relatively low weight, but this may be due to the limited variation in the continuous variables that are measured at the state level.[5] Appendix D of the Supplementary Material also shows that, in terms of raw performance, a well-designed ensemble usually beats any single constituent method. The five-model ensemble beats all of its constituent methods by more than 5% on at least one sample size considered.

MRP AND BART

One limitation of ensembles is appropriately quantifying uncertainty of the post-stratified estimates. This is challenging because it may be difficult to quantify the uncertainty of the estimates from the individual machine learning methods used in the ensemble.[6] It also requires careful work to interpret the effects of the included variables. Thus, researchers often seek to rely on a single model that can incorporate uncertainty and remains highly flexible. To that end, BART (Chipman, George, and McCulloch 2010) is an attractive choice. The method is related to popular tree-based methods such as "random forests," but it is implemented in a Bayesian framework that allows for quantification of uncertainty. Bisbee (2019) applies BART to MRP and reports that it substantially outperforms (traditional) MRP. The magnitude of the improvement is large (e.g., around 20%–30% decrease in mean absolute error [MAE]). This motivates an initial question: does BART improve upon deep MRP?

After some preliminary exploration, I discovered an error in Bisbee's (2019) replication archive. Appendix E of the Supplementary Material describes it in detail; see also a corrigendum to Bisbee (2019) (Goplerud and Bisbee 2022). In brief, the error arbitrarily injected random noise into the MRP estimates at the prediction stage. When this is corrected, traditional MRP's performance increases markedly and is only slightly beaten by BART.

Following the main analysis in Bisbee (2019), Figure 3 shows the predictive accuracy on the surveys in Buttice and Highton (2013) for a sample with 1,500 observations.[7] For simplicity, I show only three methods: the traditional ("Simple") MRP following Bisbee's (2019) provided code, a corrected traditional MRP, and deep MRP estimated using variational inference (see Table 1).

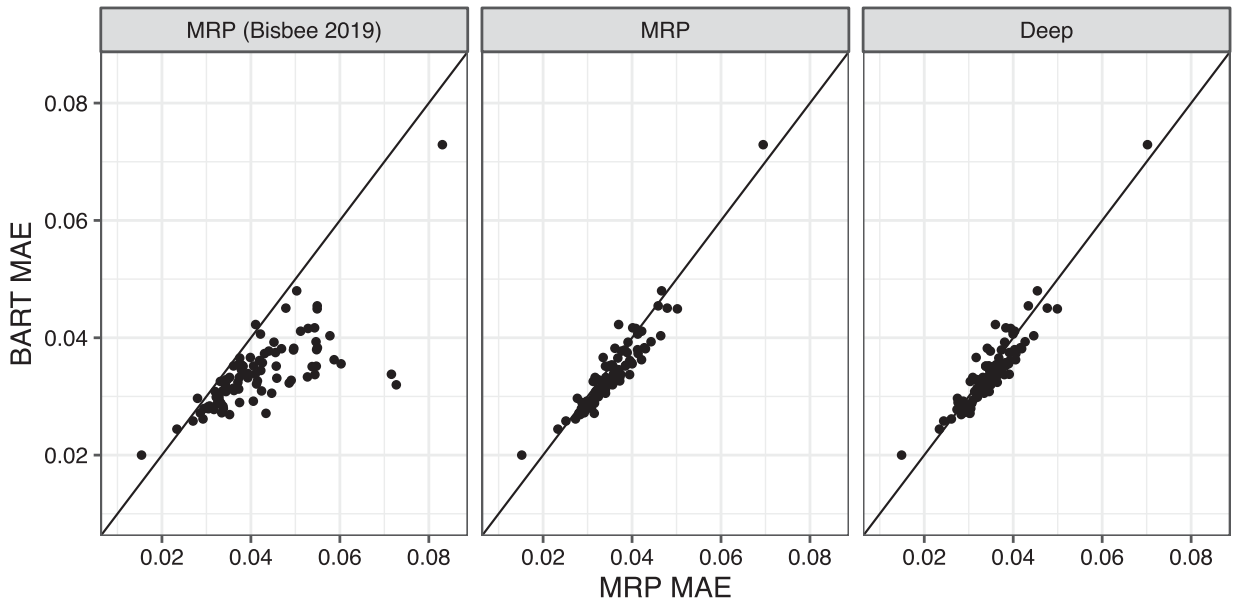
Fixing the error shows a noticeably different story; rather than being clearly beaten by BART, traditional MRP looks visually similar to BART in terms of its error across surveys. Table 2 provides a more concise quantitative summary. It shows the percentage gap in MAE versus BART averaged across the 89 surveys: $(\text{MAE}_k - \text{MAE}_{\text{BART}})/\text{MAE}_{\text{BART}} \times 100$, where MAE_k indicates the MAE of the model k averaged across two-

[5] Appendix B of the Supplementary Material provides simulations where splines are important for strong performance.

[6] Broniecki, Leemann, and Wüest (2022) suggest bootstrapping. They show promising results, but this can be computationally expensive and thus sometimes a single method is desirable.

[7] Appendix E of the Supplementary Material replicates other analyses in Bisbee (2019) and finds similar results.

FIGURE 3. Visualizing Performance: MRP versus BART



Note: Each plot compares the mean absolute error (MAE) of BART and MRP, averaged across two-hundred simulations per survey. The points below the 45-degree line indicate that MRP performs worse. The three methods shown are the traditional MRP and prediction method in Bisbee (2019) ("MRP (Bisbee 2019)"), an MRP with the same specifications but a corrected prediction ("MRP"), and a deep MRP fit using variational inference ("Deep").

TABLE 2. Relative Mean Absolute Error versus BART

	Sample size					
Method	1,500	3,000	4,500	6,000	7,500	10,000
MRP (Bisbee 2019)	22.55	26.29	30.95	35.18	38.78	44.25
MRP	4.52	1.66	1.22	1.07	1.07	1.08
Deep	2.61	−1.04	−1.05	−0.59	0.14	1.11

Note: This table reports percentage gap in mean absolute error between BART and the alternative methods; positive numbers indicate that BART outperforms its competitor. Figure 3 defines the abbreviations.

hundred simulations. A positive number indicates that BART outperforms the other method. This measure is *relative* as BART and MRP both decrease the observed MAE as sample size increases.

Table 2 shows that BART outperforms traditional MRP, but it does so by quite small margins (1%–4%) and its relative advantage declines as sample size increases. Deep MRP performs slightly better versus BART; for modest sample sizes (3,000–6,000), it actually slightly outperforms BART, although it does slightly worse at small and very large sample sizes. In terms of performance, across all sample sizes, deep MRP outperforms BART between 45% and 55% of the time and, thus, they can be considered to reach an effective "draw" in terms of performance. The table also suggests a small-but-systematic improvement of deep MRP over traditional MRP. Comparing the traditional ("Simple") MRP against deep MRP shows that, except for the largest sample sizes, traditional MRP performs around 1%–3% worse than deep MRP in terms of MAE and is beaten around 60%–70% of the time.

CONCLUSION

This paper has shown that with recent advances in variational inference and novel technical extensions, it is possible to rapidly estimate deep MRP; Appendix C of the Supplementary Material shows that deep MRP can be often estimated much more quickly than machine learning methods that require tuning of external hyperparameters and as quickly as BART without tuning. It found that deep MRP is highly competitive in performance—effectively tying the state-of-the-art BART method. Compared with traditional MRP, adding

interactions provided a small, but systematic and nontrivial, gain in performance at most observed sample sizes.

The key implications of this paper for applied MRP research are twofold. First, the fast variational methods developed in this paper allow researchers to easily perform the well-established process of model comparison (e.g., by using cross-validation) when deciding which (complex) hierarchical model to use. Rather than selecting a single (traditional) specification for MRP, the results in this paper suggest that considering and comparing a variety of possible models—traditional MRP, deep MRP, or perhaps machine learning—can result in better performance for the post-stratified estimates. It is not possible to know *a priori* whether deep or traditional hierarchical models will perform better on a specific survey, but the methods in this paper allow for this to be easily tested rather than assumed. If quantification of uncertainty in the estimates is desired, one might employ a hybrid strategy where the variational methods are used for initial model comparison to winnow down the possible models before using a fully Bayesian approach for the final estimates.

Second, in terms of re-evaluating the role of machine learning for MRP, this paper suggests that the major benefit of machine learning comes from *combining* models in an ensemble. However, it also shows a clear important role for (deep) hierarchical models in those ensembles; these deep hierarchical models often have the strongest out-of-sample predictive accuracy on the survey itself and thus are given high weight in an ensemble. There is usually little downside to including additional methods in an ensemble—especially when the cost of estimation is rather low—and thus a well-specified ensemble should probably include multiple versions of MRP (traditional and deep). As in the case of model comparison, this allows for the data to provide guidance in terms of which type of hierarchical model is most appropriate.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/S0003055423000035.

DATA AVAILABILITY STATEMENT

Research documentation and data that support the findings of this study are openly available in the American Political Science Review Dataverse at https://doi.org/10.7910/DVN/XRD6YG.

Open-source software implemented the methods in this paper is available at https://cran.r-project.org/package=vglmer or https://github.com/mgoplerud/vglmer.

ACKNOWLEDGMENTS

I would like to thank Michael Auslen, James Bisbee, Danny Choi, Jeff Gill, Kosuke Imai, Gary King, Shiro Kuriwaki, Dustin Tingley, and participants at PolMeth 2021 for comments on this draft. A demonstration of how to use the accompanying software is available in Appendix F of the Supplementary Material. All remaining errors are my own.

FUNDING STATEMENT

This research was supported in part by the University of Pittsburgh Center for Research Computing, RRID: SCR_022735, through the resources provided. Specifically, this work used the H2P cluster, which is supported by NSF award number OAC-2117681.

CONFLICT OF INTEREST

The author declares no ethical issues or conflicts of interest in this research.

ETHICAL STANDARDS

The author affirms this research did not involve human subjects.

REFERENCES

Bisbee, James. 2019. "BARP: Improving Mister P using Bayesian Additive Regression Trees." *American Political Science Review* 113 (4): 1060–65.

Broniecki, Philipp, Lucas Leemann, and Reto Wüest. 2022. "Improved Multilevel Regression with Post-Stratification through Machine Learning (autoMrP)." *Journal of Politics* 84 (1): 597–601.

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

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. 2010. "BART: Bayesian Additive Regression Trees." *Annals of Applied Statistics* 4 (1): 266–98.

Chung, Yeojin, Andrew Gelman, Sophia Rabe-Hesketh, Jingchen Liu, and Vincent Dorie. 2015. "Weakly Informative Prior for Point Estimation of Covariance Matrices in Hierarchical Models." *Journal of Educational and Behavioral Statistics* 40 (2): 136–57.

Gelman, Andrew. 2006. "Prior Distributions for Variance Parameters in Hierarchical Models." *Bayesian Analysis* 1 (3): 515–33.

Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Ghitza, Yair, and Andrew Gelman. 2013. "Deep Interactions with MRP: Election Turnout and Voting Patterns among Small Electoral Subgroups." *American Journal of Political Science* 57 (3): 762–76.

Goplerud, Max. 2022. "Fast and Accurate Estimation of Non-Nested Binomial Hierarchical Models using Variational Inference." *Bayesian Analysis* 17 (2): 623–50.

Goplerud, Max. 2023. "Replication Data for: Re-Evaluating Machine Learning for MRP Given the Comparable Performance of (Deep) Hierarchical Models." Harvard Dataverse. Dataset. https://doi.org/10.7910/DVN/XRD6YG.

Goplerud, Max, and James Bisbee. 2022. "BARP: Improving Mister P using Bayesian Additive Regression Trees—Corrigendum." *American Political Science Review*, 1–3. https://doi.org/10.1017/S0003055422001435.

Goplerud, Max, Shiro Kuriwaki, Marc Ratkovic, and Dustin Tingley. 2018. "Sparse Multilevel Regression (and Poststratification [sMRP])." Unpublished Manuscript.

Grimmer, Justin. 2011. "An Introduction to Bayesian Inference via Variational Approximations." *Political Analysis* 19 (1): 32–47.

Huang, Alan, and Matt P. Wand. 2013. "Simple Marginally Noninformative Prior Distributions for Covariance Matrices." *Bayesian Analysis* 8 (2): 439–52.
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.
Ornstein, Joseph T. 2020. "Stacked Regression and Poststratification." *Political Analysis* 28 (2): 293–301.
Park, David K., Andrew Gelman, and Joseph Bafumi. 2004. "Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls." *Political Analysis* 12 (4): 375–85.