Exploring the performance of multilevel modeling and poststratification with Eurobarometer data

Dimiter Toshkov*

April 21, 2015

1 Introduction

Multilevel regression modeling and poststratification (MRP) is a promising and increasingly popular strategy[1] to derive disaggregated data for political and policy analysis from aggregate-level surveys. In short, MRP adjusts the simple disaggregated estimates of federal-level public opinion by building a multilevel regression model with individual-level and state-level predictors and state indicators, and then combining these estimates with census data for poststratification in order to arrive at the final state-level estimates.

While the promise of MRP is appealing, its ability to derive valid and useful state-level estimates in practice is still being explored (Lax and Phillips 2009, Pacheco 2011, Warshaw and Rodden 2012, Stollwerk 2013, Buttice and Highton 2013). In a recent article, Buttice and Highton (2013) compare MRP estimates to the 'true' state-level values derived directly from very large (but still federal-level only) US surveys of public opinion. Their results are generally positive, but also highlight the need for more work to understand when the adjusted estimates get close enough to the 'true' values. However, a major limitation of this study - pointed out in the ensuing discussion[2] - is that the

*Associate Professor of Public Administration at Leiden University, The Netherlands (d.d.toshkov@cdh.leidenuniv.nl). I would like to thank Thomas Konig and the participants of the Seventh Pan-European Conference on the European Union, 5-7 June 2014, The Hague, The Netherlands, for useful comments and suggestions.

[1] Although it is based on previous work, the technique has been introduced to political science by Park, Gelman, and Bafumi (2004). Applications include Kastellec et al. (2010), Lax and Phillips (2012), and Canes-Wrone et al. (2014). Comprehensive expositions to the statistical theory behind the method are available in Gelman and Little (1997) and Park et al. (2004). Gelman and Hill (2007, Chapter 14) and Kastellec et al. (2014) provide accessible introductions to the method, as well as practical guides to its implementation using the R statistical program.

[2] See `http://andrewgelman.com/2013/10/09/mister-p-whats-its-secret-sauce/`

benchmarks against which the performance of MRP has been compared cannot, in fact, be assumed to represent the 'true' state-level values, as they are not based on representative state-level surveys.

2 Approach

This letter reports the results of an analysis of the performance of MRP using real political data and a different approach[3]. I rely on the fact that each (recent) Standard Eurobarometer survey wave contains as many as 30 nationally-representative polls based on a multi-stage, random sampling design with more than 1,000 respondents per each state. As a consequence, the real national-level public opinion in the EU states can be estimated with high precision - (sampling) margins of error vary between 1.4% and 3.1% (95% confidence level). I use these national estimates as a benchmark.

To mimic MRP, I start with drawing a sample of size 1,500 from the entire Eurobarometer pool[4]. The sample is drawn in two different ways - *weighted* by state population share and with *equal* number of individuals per state.

Then I apply MRP to reconstruct the state-level estimates of public opinion from the sample data only (combined with census data for poststratification). I estimate a multilevel regression model of the public attitude of interest which has state indicators, a number of individual-level predictors (age, sex, occupation status, and education of the respondent), and two state-level predictors which differ per item being modeled[5]. There are 4032 distinct categories created by the intersection of these variables. For the poststratification stage, I use data from the 2001 EU-wide census available through *Eurostat* which provides the relative shares of citizens cross-classified by age, sex, occupation

[3] A longer version of this text featuring a more detailed presentation of the approach and additional results is available at the website of the author at: ...

[4] In practice, due to the limited availability of census data, I use only 24 of the states included in the survey which still provides more than 24,000 respondents.

[5] Linear regressions of the 'true' state means on the selected two state-level predictors return adjusted R^2s ranging from 0.32 to 0.75 with a mean of 0.58 for the ten items being modeled. See the Supplementary materials for details.

status, and education for 24 EU member states.

Finally, I compare these estimates to the 'true' values from the full survey. I employ four measures of the fit between the two sets of state means - Pearson's product-moment correlation (ρ), Kendall's rank-order correlation τ, mean absolute error (MAE), and coverage (the number of states for which the MRP-derived estimates fall within by the 95% sampling margins of error of the respective 'true' state means).

To capture the variability of the MRP estimates, I use simulation. I repeat the process of drawing a sample, fitting a model and poststratification 100 times, and report the means and standard deviations of the resulting sets of 100 comparisons for each of the four performance indicators. The process is replicated for ten survey items from different Eurobarometer waves and on different topics. The ten items differ in the range and variability of the 'true' state means, in the share of inter-state versus within-state variation, and in the extent to which individual-level variation is captured by the available demographic predictors[6].

3 Results

The results of the analysis are summarized in Table 1 below. In line with existing studies (Buttice and Highton 2013, Warshaw and Roden 2012, Lax and Philips 2009), MRP appears generally successful in producing estimates which are highly correlated with the 'true' values. In fact, the mean correlation ρ across the ten items (0.90) is much higher than the one reported in Buttice and Highton (2013) from their analysis of US data, and even the lowest one observed here (0.80) would be in the top 10% of their estimates.

However, the approach is less capable of reconstructing the relative rankings of the country means. Although the average Kendall's τ would easily pass a statistical significance test, the rank order of the country means is in fact not very well preserved by

[6] See Buttice and Highton (2013) for the potential importance of these parameters for explaining the varying performance of MRP.

	Mean (St.dev.)							
	Weighted				**Equal weight**			
	ρ	τ	MAE	*cov*	ρ	τ	MAE	*cov*
Trust in the EU	0.84	0.63	0.05	8.3	0.79	0.59	0.05	9.0
	(0.05)	(0.06)	(0.01)	(2.9)	(0.06)	(0.08)	(0.01)	(2.3)
Immigration importance	0.89	0.66	0.03	10.4	0.86	0.68	0.03	8.2
	(0.05)	(0.08)	(0.01)	(2.5)	(0.07)	(0.07)	(0.01)	(2.7)
Life satisfaction	0.95	0.80	0.05	7.3	0.93	0.78	0.06	6.0
	(0.01)	(0.03)	(0.01)	(1.8)	(0.02)	(0.05)	(0.01)	(2.3)
Interpersonal trust	0.94	0.72	0.04	9.8	0.93	0.70	0.05	7.6
	(0.02)	(0.04)	(0.01)	(2.2)	(0.02)	(0.02)	(0.01)	(2.2)
Organizational membership	0.89	0.71	0.07	6.8	0.93	0.76	0.07	6.5
	(0.02)	(0.04)	(0.01)	(2.0)	(0.02)	(0.04)	(0.01)	(2.3)
EU membership is good	0.80	0.58	0.06	7.5	0.76	0.53	0.07	6.8
	(0.07)	(0.08)	(0.01)	(2.2)	(0.09)	(0.12)	(0.01)	(2.2)
Belief in God	0.92	0.74	0.06	7.8	0.93	0.75	0.07	6.7
	(0.02)	(0.04)	(0.01)	(2.0)	(0.03)	(0.05)	(0.01)	(2.0)
Gender equality	0.89	0.68	0.05	10.4	0.87	0.66	0.05	8.0
	(0.03)	(0.06)	(0.01)	(2.7)	(0.04)	(0.06)	(0.01)	(2.5)
Unemployment importance	0.92	0.76	0.06	7.9	0.90	0.74	0.07	5.8
	(0.02)	(0.04)	(0.01)	(2.2)	(0.03)	(0.06)	(0.01)	(2.1)
Democracy satisfaction	0.92	0.78	0.06	7.7	0.93	0.79	0.06	6.6
	(0.02)	(0.04)	(0.01)	(2.2)	(0.03)	(0.04)	(0.01)	(2.0)

Table 1: Comparisons between the 'true' and the MRP-based state means (based on 100 simulations)

the MRP procedure. The average *mean absolute error* (MAE) is in the range between 0.03 and 0.07 with an average of 0.05 (which is lower than the average error reported in Buttice and Highton's study). Surprisingly however, the relatively high correlations and low errors are not translated into good coverage with on average 8.4 states (from the total of 24) falling within the 95% sampling error margins of the true state means. In other words, typically 65% of the state means would not be covered by the MRP estimates, despite the good overall performance of MRP.

These results confirm the general utility of MRP to produce estimates that are highly correlated with the true values, but also point to some of the limitations of MRP to provide useful data for political and policy research. When it is the *ranking* of the states that matters, MRP appears less useful. More importantly, even when the correlations between the MRP estimates and the 'true' values are very high, the coverage can be rather low with less than one third of the MRP state estimates falling within the plausible range of values of the state means. One implication of these results is that MRP might be more useful for creating variables for inclusion as covariates in models of other outcomes of interest (because the correlation ρ with the real values which they would substitute is high), but less so for deriving valid descriptive inferences about the absolute values of the state means (low coverage) and their relative rankings (moderate τ)[7].

Figure 1 presents a visual overview of how the MRP estimates compare to the 'true' values for one of the items being modeled ('Trust in the EU', *weighted* sampling). The graph shows the means (dots) and associated 95% sampling margins of error for each state based on the full Eurobarometer survey (in black) and the means (dots) and the 0.05 and 0.95 quantiles of the MRP estimates from 100 simulations (in red).

The performance of MRP across the ten items is not related in a straightforward way to factors like the strength of the state-level predictors, the range and variabilty of 'true' state means, or the share of inter-state versus intra-state variation (cf. Buttice and Highton 2013). The *weighted* and *equal shares* sampling procedures do not seem to lead to significant differences in performance, and in some cases the *equal shares* one does actually worse. This is surprising since some of the smaller states get no more than a couple of respondents in the *weighted* sample.

The indicators discussed so far assessed the performance of MRP in absolute terms but it is important to consider the relative improvement of MRP vis-a-vis alternative approaches and to trace where the power of MRP comes from. Simple disaggregation of the samples by state to derive state means offers one reference point for comparison with the MRP estimates. As expected, MRP always outperforms simple disaggregation with respect to all four indicators. The improvement is substantial when the sample is *weighted*

[7] But note that often the 'true' state ranking cannot be uncovered easily even using the full Eurobarometer survey, with many of the state means being within the sampling margins of error of their neighbours.

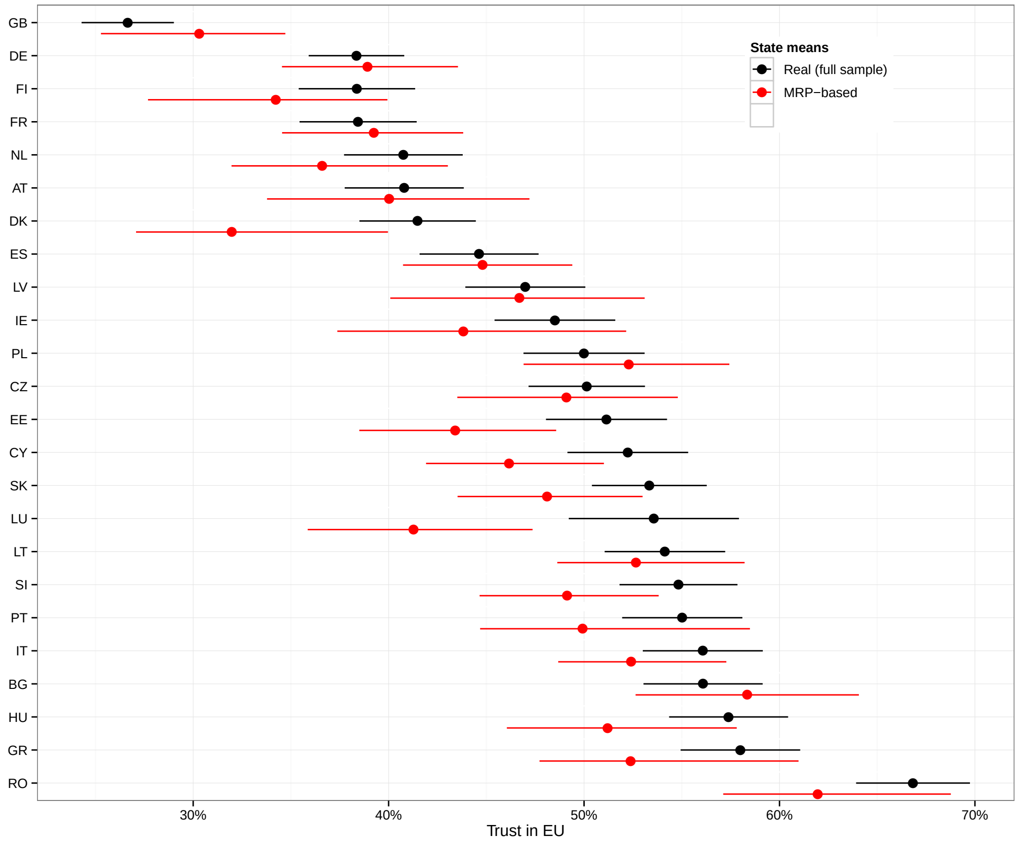

Figure 1: Comparison of state means of 'Trust in EU' based on the full Eurobarometer sample and on MRP estimates based on a total of 1,500 individuals (weighted sample by population shares)

(probability of inclusion proportional to state size), but often only marginal when the sample is drawn with an equal number of individuals per state (in the simulations, a sample of 1,500 units for 24 states gives 62 or 63 units per state).

The average improvement of the correlation ρ with the 'true' values when the sample is drawn proportionately to state size is around 0.17, but as low as 0.03 when each state has an equal number of individuals in the sample. The improvement in MAE is between 0.05 and 0.02 with an average of 0.03 for the *weighted* sample and much smaller (sometimes less than 0.01) for the *equal-shares* sample. The average improvement in coverage is between 1 and 2 countries[8].

How does MRP compare to two simpler alternatives - a model with only a state in-

[8] For details, see the Supplementary materials.

dicator entered as a random effect, and a model with a state random effect plus the two state-level predictors? It turns out that the big share of the improvements with respect to all performance indicators comes from the modeling of the state means as a random effect (cf. Lax and Phillips 2009). The additional improvements from including the state-level predictors are often very small, despite these predictors accounting for substantial parts of the variation in the state means. Surprisingly, adding the individual-level random effects and poststratification by relative population shares actually slightly *decreases* the performance of the models with respect to all four criteria. This result is significant because poststratification is the more data-demanding stage of MRP as it requires the availability of estimates of population shares for each combination of individual-level predictors for each state. In the context of the data analyzed and presented here, it turns out that the best approach is to model the state means and perhaps add state-level predictors. But the inclusion of individual-level predictors and poststratification brings no added value and can, in fact, decrease the quality of the resulting estimates[9]. This is in line with the finding of Buttice and Highton that there is little observed improvement in [MRP] performance associated with stronger individual-level models (2013, 9), but contradicts Lax and Phillips who conclude that including even one individual-level predictor leads to significant gains (2009, 115). More work is needed to ascertain the conditions under which the inclusion of individual-level predictors and poststratification significantly improves on simpler models. [10]

Altogether, the results summarized here suggest that just by imposing a distribution on the simple disaggregated state means (and possibly including a couple of state-level predictors as well), one can get state-level estimates which would often correlate as high as 0.9 with the true state means, although would generally not be as good in reproducing

[9] The demographic individual-level predictors included in this study account for a share of the individual-level variation in public opinion that is comparable to the corresponding quantity in the analysis of US data reported in Buttice and Highton (2013).

[10] Additional analyses conducted on the Eurobarometer data showed that having better measures of the individual-level effects (estimated from the total available pool of observations), interacting them with state-level variables, and estimating them as fixed rather than random effects did not improve significantly the performance of MRP.

relative state rankings and the actual 'true' values for the majority of the state means. Whether high correlation (ρ) coupled with poor coverage is *sufficiently good* for research purposes is a judgement that needs to be made in the context of specific research projects.

The results of the analyses also suggest that MRP brings relatively little added value when there are 62 individuals available per state. Future research should explore exactly how many individuals per group are needed so that simple disaggregation becomes significantly worse option than MRP.

This study has been based on European data and it is important to consider how its implications would travel across the Atlantic. In the US, there are twice as many states as there are European countries with comparable census data in Europe. This should help MRP as the models have more groups to work with. But the same total amount of respondents is distributed over twice as many states which implies that with simple disaggregation the estimates of the smaller states in particular would suffer (especially under sampling not stratified by state, which would be the equivalent of the weighted samples discussed above); hence, there will be more space for improvements by MRP. It also seems that the ratio of inter-state to intra-state variation is on average higher in the EU than in the US, which would favor the performance of MRP. Future research should explore further the similarities and dissimilarities between the structures of European and US public opinion and their implications for the use and evaluation of MRP.

References

.

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

Canes-Wrone, Brandice, Tom S. Clark, and J. P. Kelly. 2014. Judicial Selection and Death Penalty Decisions. *American Political Science Review* 108(1): 23-39.

Gelman, Andrew. 2007. Struggles with Survey Weighting and Regression Modeling. *Statistical Science* 22(2): 153-64.

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

Gelman, Andrew, and Thomas Little. 1997. Poststratification into Many Categories Using Hierarchical Logistic Regression. *Survey Methodologist* 23(1): 127-35.

Kastellec, Jonathan P., Jeffrey R. Lax, and Justin H. Phillips. 2014. Estimating State Public Opinion with Multi-Level Regression and Poststratication Using R. `http://www.princeton.edu/~jkastell/MRP_primer/mrp_primer.pdf`.

Kastellec, Jonathan P., Jeffrey R. Lax, and Justin H. Phillips. 2010. Public Opinion and Senate Confirmation of Supreme Court Nominees. *Journal of Politics* 72(3): 767-84.

Lax, Jeffrey R., and Justin H. Phillips. 2012. The Democratic Deficit in the States. *American Journal of Political Science* 56(1): 148-66.

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.

Pacheco, Julianna. 2011. Using National Surveys to Measure Dynamic U.S. State Public Opinion: A Guideline for Scholars and an Application. *State Politics & Policy Quarterly* 11(4): 415-39.

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.

Stollwerk, Alissa. 2013. The Application of Multilevel Regression with Post-Stratication to Cluster Sampled Polls: Challenges and Suggestions. `http://www.uiowa.edu/~stpols13/papers/Stollwerk_SPP_2013.pdf`.

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