Latent Civil War: Improving Inference and Forecasting with a Civil War Measurement Model

Andrew Halterman*
Benjamin J. Radford†

July 26, 2023

Abstract

This paper contributes to the substantive study of civil war onset and the methodological literature on handling latent dependent variables. We first introduce a new Bayesian measurement model of civil war status, using eight datasets on civil war status and a dynamic Bayesian IRT to produce measures of latent civil war status. Using the latent data, we then re-analyze canonical work on civil war onset, showing that some previous findings are sensitive to the uncertainty in civil war status. We then turn to our second methodological contribution, introducing a technique ("posterior bagging") for building improved forecasting models for civil war onset.

1 Introduction

Political scientists have developed Bayesian measurement models for concepts such as democracy (e.g. Treier and Jackman 2008; Pemstein, Meserve and Melton 2010; Ulfelder and Taylor 2015), respect for human rights (e.g. Fariss 2014, 2019), wartime sexual violence (Krüger and Nordås 2020), and other expert coded data (Marquardt and Pemstein 2018), along with many techniques for estimating latent political ideology.[1] Measurement models allow researchers to address uncertainty and missing data in a principled way and to account for the differences across raters or datasets.

*Department of Political Science, Michigan State University. Email: ahalterman0@gmail.com.
†Department of Political Science and Public Administration, Public Policy Ph.D. Program, School of Data Science.
[1]For helpful comments, we thank Edo Airoldi and Max Goplerud. AH acknowledges support by an NSF Graduate Research Fellowship.

Civil war status is a natural application for measurement models, but no latent variable model of civil war status has yet been published. Using a measurement model to estimate latent civil war status has several appealing features. First, it allows us to reduce our dependence on any specific dataset's coding of civil war, which may have idiosyncratic decisions or errors. Second, it allows us to extend the range of coverage for the dataset beyond the years that any particular dataset covers, which is important for forecasting. Finally, the estimates from a latent variable model of civil war status more accurately reflect the inherent uncertainty in civil war status, reporting the probability that a country-year is experiencing civil war, rather than a binary rating.

Recent work has explored the special requirements that using latent variables in regression models imposes. For instance, Fong and Grimmer (2019) discuss the extra assumptions that are required to recover unbiased causal estimates when treatment is a latent variable. Specifically, using latent variables as outcomes requires special techniques. If the variance in the latent estimate is independent of the explanatory variables in the model, the extra variance should increase the variance of our parameter estimates but will not induce bias. To incorporate the uncertainty of the latent variable into our parameter estimates, we follow a sampling-based technique used in Mislevy (1991) and Schnakenberg and Fariss (2014). They suggest drawing repeatedly from the posterior distribution of the latent variable and use Rubin's method, developed for accounting for the variance caused by multiple imputation methods (Rubin 1976), to aggregating uncertainty from draws into the final parameter estimates (Mislevy 1991). We replicate the Fearon and Laitin (2003) model of civil war, using our new latent measure of civil war status and accounting for the variance of the latent estimate using the technique described above.

However, the techniques described above assume that the variance in the latent variable is independent of the explanatory variables in the model. This is not the case for our model of civil war status. We show that the variance in the latent civil war model is correlated with the variables we observe, raising concerns that error in measures of civil war is dependent on omitted variables as well, which would bias parameter estimates in studies of civil war onset. Recent work has brought renewed attention to the problem of non-classical error in dependent variables (Millimet and Parmeter 2022), which this current work emphasizes as well.

Finally, we show how our latent civil war measure can be used to build improved forecasting models for civil war onset. We begin with two insights. First, fitting models on a binary outcome measure of civil war status is equivalent to dichotomizing the continuous latent civil war status measure. In general, dichotomizing continuous dependent variables greatly reduces the power of our models. In civil war onset, this is especially problematic

as the onset of civil war is a rare event. We show that using the new latent civil war status measure instead of the binary measure increases the power of our models.

Second, we introduce an extension of the bootstrap aggregating ("bagging") method that uses draws from the posterior estimate of civil war status, alongside the traditional bootstrapping method. Traditional bagging, which underlies algorithms such as the random forest classifier, take repeated bootstrap draws from a dataset, fit a classifier on each sample, and aggregate the set of classifiers' predictions into a final prediction. We extend this method to use draws from the posterior distribution of the latent civil war status measure. "Posterior bagging" allows us to train predictive models on a wider range of possible civil war outcomes, which allows us to increase the stability and predictive power of our forecasting model. Note that the purpose of this technique is not to improve the uncertainty of our parameters (as Mislevy (1991) and Schnakenberg and Fariss (2014) do), but to improve the accuracy of our predictions by providing the model with a wider range of possible outcomes.

2 Error in the Dependent Variable

Measurement error in the dependent variable in a regression context has received less attention than errors in independent variables because under common assumptions about classical measurement in the dependent variable, measurement error in the dependent variable is incorporated in the regression error term.

If the true model is given by $y^{\text{true}} = X\beta + \epsilon$, but the observed data is measured with error γ, we can write the observed dependent variable as $y^{\text{obs}} = y^{\text{true}} + \gamma$. Rearranging terms yields $y_k^{\text{true}} = X\beta + \epsilon + \gamma$. If the additional measurement error term γ is mean zero and independent of X, the regular proof of the unbiasedness of OLS holds and shows unbiased estimates β. Thus, a regression with the observed values y^{obs} produces unbiased coefficient estimates for the true model with y^{true}.

However, running the regression with y^{obs}, rather than the error-free y^{true}, introduces three complications:

1. Greater, unaccounted variance in the coefficient estimates, leading to incorrect inference.

2. Greater variance leading to degraded predictive performance.

3. If γ is not independent of X, the regression on y^{obs} will produce biased coefficient estimates.

We illustrate how a measurement model that produces an estimate $\widehat{y^{\text{true}}}$ can be used to address the first and second issues. Following previous work (Schnakenberg and Fariss 2014), we show that using the estimates from a measurement model of the dependent variable allows us to incorporate additional uncertainty in our parameters estimates. Next, we introduce a technique, "posterior bagging", that improves the predictive performance of a model, compared to training on the original observed data.

The final issue is empirically testable and we leave for future versions of this work. If measurement error is *not* independent of the covariates in common explanatory models of civil war, this implies that the coefficient estimates in many models of civil war are biased.

3 Civil War as a Latent Variable

We identify eight separate datasets on civil war status. Table 1 shows the eight civil war datasets we use in our latent variable model, along with their definitions. Note that we split the UCDP-PRIO armed conflict dataset into two separate datasets, following the two definitions of civil war that they provide. All of the datasets have as a core component of their definition that a civil war involves armed conflict that produces casualties, and involves a non-state armed group fighting with the government. The primary differences across the datasets concern the battle death threshold and distribution of causalities between sides, along with different rules about when an interrupted conflict generates separate incidents or a single, prolonged incident.

Dataset	Definition

ICEWS Ground Truth Dataset (Lustick et al. 2015)	• An insurgency is "organized, active, violent opposition with substantial arms, whose objective is to overthrow the central government" (3) • Rebellion is "organized, active, violent opposition with substantial arms, whose objective is to seek autonomy or independence from the central government." (3-4) • No battle death threshold, but "there must be multiple instances of violence either on a specific month or in the surrounding months" (3) • ICEWS reports civil war at the monthly level: we code a country-year as having civil war if more than 6 months have either insurgency or rebellion
UCDP (major) (Melander, Pettersson and Themnér 2016; Gleditsch et al. 2002)	• armed conflict between a government and organized opposition over government or territory • >1,000 battle deaths per year • We include both "internal armed conflict" and "internationalized internal armed conflict", where a second state intervenes.
UCDP (major or minor)	Same as above, but with a lower death threshold: • >25 battle deaths per year

PITF Major Episodes of Political Violence (Marshall 2019)	• We code both "civil war" and "ethnic war" as civil war. • "information regarding the degree of militant organization, tactical and strategic characteristics, and expressed level of commitment to the use of violence are taken into consideration; the designation of war carries with it a stronger institutional, or institutionalized, component and more definite objectives." (2) • "The 'begin' and 'end' years [...] are those considered by the author to be those most likely to capture the transformative 'moments' (beginning and ending) of the episodes" (7)
Fearon and Laitin (2003)	• Violent conflict state and organized non-state actor • >1,000 total deaths • Yearly average deaths >100 • >100 deaths on both sides • Anti-colonial wars included
Doyle and Sambanis (2000)	• >1,000 total battle deaths • >1,000 battle deaths in at least one year • organized armed opposition challenges the sovereignty of a state within its borders
Collier and Hoeffler (2002)	• Internal conflict • Government and identifiable rebel group each suffer >5% casualties • >1,000 battle deaths per year • Anti-colonial wars excluded

Correlates of War (Sarkees and Schafer 2000)	• Internal (non-colonial) war involving the government • "Effective resistance" by each side, measuring by the ratio of casualties • >1,000 battle deaths per year • Ceasefires lasting longer than 6 months produce a new episode

Table 1: *Civil war datasets with civil war coding criteria*

4 A Dynamic Bayesian Measurement Model of Latent Civil War

While each of the eight datasets has a different definition of civil war, we assume that the similarities in their definitions are similar enough to treat them as measuring the same underlying latent concept, albeit with different stringency, usually based on the number of battle deaths per year. Thus, we wish to combine the information contained in each dataset to estimate an underlying latent probability of civil war.

4.1 Simple two parameter IRT

We begin with a discussion simple two parameter item response theory (2PL IRT) model to introduce notation and provide a foundation for the dynamic models we propose below. In a standard IRT, the estimated probability that rater k assigns observation i a value of 1 is given by Eq 1. The parameter θ_i is the unobserved latent value for observation i. The "difficulty" parameter for rater k, δ_k, corresponds to a rater-specific intercept: we can think of it as a bias or offset term that shifts rater k's threshold for a positive class label relative to the latent variable θ. The "discrimination" parameter for rater k, $\alpha_k > 0$, corresponds to a slope. This parameter controls how precisely rater k distinguishes positive from negative cases along the range of θ or, alternatively, how much uncertainty rater k demonstrates when evaluating cases that call in the middle range of θ. Here we suggest standard normal and exponential priors for these parameters, though other priors may be preferable in practice (Eqs 2 and 3).

$$P(y_{ik} = 1|\theta_i, \alpha_k, \delta_k) = \text{logit}^{-1}(\alpha_k(\theta_i - \delta_k)) \quad (1)$$
$$= \frac{\exp(\alpha_k(\theta_i - \delta_k)}{1 + \exp(\alpha_k(\theta_i - \delta_k)}$$
$$\delta_k, \theta_i \sim \text{Normal}(0, 1) \quad (2)$$
$$\alpha_k \sim \text{Exponential}(1) \quad (3)$$

4.2 "Robust" dynamic IRT

Static IRT models do not account for the autocorrelation in latent status between years within a country. By ignoring the potential information between years, the posterior estimates that do not include autocorrelation are very wide and have large variances from year to year. However, simply including a normally distributed random walk does not allow sharp jumps in latent status (Ulfelder and Taylor 2015; Reuning, Kenwick and Fariss 2019), which does not match the sudden onset and termination dynamics that are often present in civil war. Ulfelder and Taylor (2015) suggest modeling the temporal autocorrelation as a Cauchy distribution, centering the current year's status on the previous year's status, while still allowing large shifts in status. However, the Cauchy distribution's variance induces large variance in the posterior estimate and can make it difficult for the model to converge.[2] Reuning, Kenwick and Fariss (2019) point out that a Student's t distribution with few degrees of freedom offers the ability to allow large shifts in latent status while still usefully constraining the dynamic model.

The robust dynamic IRT follows the suggestion in Reuning, Kenwick and Fariss (2019) and modifies the simple 2PL IRT. We index our latent variable, $\theta_{c,t}$, by c and t, representing countries and time units (years), respectively. For the first observation per country we draw $\theta_{c,t=1}$ from a Student's t distribution with four degrees of freedom and mean zero. For $t > 1$ we draw $\theta_{c,t}$ from a Student's t distribution with four degrees of freedom and mean $\theta_{c,t-1}$ (Eq 4). We place an Exponential(1) prior on the scale of the t distribution (Eq 5). We also modify our prior for α_k to maintain consistency between this model and our subsequent models (Eq 6). Here, we draw α_k from a standard Normal distribution and exponentiate this value when computing the inverse logit from Eq 1. Therefore, Eq 1 becomes $\text{logit}^{-1}(\exp(\alpha_k) \times (\theta_{c,t} - \delta_k))$.

[2]We experimented with the model suggested by Ulfelder and Taylor (2015) but found convergence issues, even on their original democracy application.

$$\theta_{c,t} \sim \begin{cases} \mathrm{T}_4(\theta_{c,t-1}, \sigma) & \text{if } t > 1 \\ \mathrm{T}_4(0, \sigma) & \text{if } t = 1 \end{cases} \tag{4}$$

$$\sigma \sim \text{Exponential}(1) \tag{5}$$

$$\alpha_k \sim \text{Normal}(0, 1) \tag{6}$$

Reuning, Kenwick and Fariss (2019) find that this "robust" dynamic model outperforms static IRTs or a simple "dynamic" model that uses a Gaussian random walk. We use this model as a baseline and suggest improvements below.

4.3 Random effects IRT

Our first proposed model introduces random coder effects across countries. We can imagine scenarios in which rater difficulty and discrimination parameters are heterogeneous across cases. Some countries may pose greater challenges to raters than others if, for example, information from those areas is particularly poor. For a given country c then, we may want to leverage information about not only in the mean difficulty or discrimination parameter for a rater, but also country-rater idiosyncrasies. In this model, we draw random variables $\alpha_{k,c}$, $\delta_{k,c}$, and $\sigma_{c,t}$ as given in Eqs 7 through 10. Hyperpriors are given in Eqs 11 through 15.

$$\theta_{c,t} \sim \begin{cases} \mathrm{T}_4(\theta_{c,t-1}, \exp(\sigma_{c,t})) & \text{if } t > 1 \\ \mathrm{T}_4(0, \exp(\sigma_{c,t})) & \text{if } t = 1 \end{cases} \tag{7}$$

$$\alpha_{k,c} \sim \text{Normal}(\mu_{\alpha_k}, \Sigma_\alpha) \tag{8}$$

$$\delta_{k,c} \sim \text{Normal}(\mu_{\delta_k}, \Sigma_\delta) \tag{9}$$

$$\sigma_{c,t} \sim \text{Normal}(\mu_{\sigma_c}, 1) \tag{10}$$

$$\mu_{\alpha_k}, \mu_{\delta_k}, \mu_{\sigma_c} \sim \text{Normal}(0, 1) \tag{11}$$

$$\Sigma_\alpha = \text{Diag}(\tau_\alpha) \times \Omega_\alpha \times \text{Diag}(\tau_\alpha) \tag{12}$$

$$\Sigma_\delta = \text{Diag}(\tau_\delta) \times \Omega_\delta \times \text{Diag}(\tau_\delta) \tag{13}$$

$$\tau_\alpha, \tau_\delta \sim \text{Cauchy}(0, 2.5) \tag{14}$$

$$\Omega_\alpha, \Omega_\delta \sim \text{LKJCorr}(2) \tag{15}$$

4.4 Bayesian Data Reweighting IRT

While there may be many motivating reasons for estimating a latent variable from observed realizations, one is the desire to leverage multiple observations to minimize the impacts of anomalous (and possibly miscoded) observations. To this end, we incorporate Bayesian data reweighting as introduced by Wang, Kucukelbir and Blei (2017). We convert our models into robust probabilistic models (RPM) by raising the likelihood of the model to the power of an observation-specific weight, $w_i \in \mathbf{w} \times N$, a simplex vector of weights scaled by the total number of observations (Eq 16).[3] $\mathbf{\Theta}$ is the vector of all model parameters except the weights and $\mathbf{w}$ is the length-N vector of observation weights. These parameters down-weight the influence of observations that poorly match the assumptions of the model (i.e., outliers). These weights are themselves latent parameters that are modeled with a Dirichlet distribution as shown in Eq 17.

$$p(\mathbf{\Theta}, \mathbf{w}|y) \propto p_{\mathbf{\Theta}}(\mathbf{\Theta}) p_{\mathbf{w}}(\mathbf{w}) \prod_{i=1}^{N} \ell(y_i|\mathbf{\Theta})^{N w_i} \tag{16}$$

$$\mathbf{w} \sim \text{Dirichlet}(\mathbf{1}) \tag{17}$$

4.5 "Switching" IRT

Finally, we introduce a "switching" dynamic IRT that explicitly accounts for shifts in latent status in a theoretically motivated way. We estimate three latent states for each country year (Eq 18) along with a switching parameter, π (Eq 20). The means for the latent states are drawn from a Normal distribution such that they preserve the order given in Eq 19. Thus, the war status is constrained to have a higher value than the transition and peace states. Each country-year is associated with a simplex vector $\pi_{c,y}$ over the three latent states. The small prior on the Dirichlet distribution encodes a preference for country-year membership probability mass to be centered on one of the three latent states rather than mixed between them.

[3]Note the possible confusion in terminology. Reuning, Kenwick and Fariss (2019) use "robust" to refer to their use of a Student's t rather than a normal walk. Wang, Kucukelbir and Blei (2017) use the term "robust" to refer to their data reweighting scheme.

$$\theta_{c,t}^{\text{peace}} \sim \begin{cases} \text{T}_4(\theta_{c,t-1}^{\text{peace}}, \sigma^{\text{peace}}) & \text{if } t > 1 \\ \text{T}_4(\mu^{\text{peace}}, \sigma^{\text{peace}})) & \text{if } t = 1 \end{cases}$$

$$\theta_{c,t}^{\text{transition}} \sim \text{T}_4(\mu^{\text{transition}}, \sigma^{\text{transition}})) \tag{18}$$

$$\theta_{c,t}^{\text{war}} \sim \begin{cases} \text{T}_4(\theta_{c,t-1}^{\text{war}}, \sigma^{\text{war}}) & \text{if } t > 1 \\ \text{T}_4(\mu^{\text{war}}, \sigma^{\text{war}})) & \text{if } t = 1 \end{cases}$$

$$\mu^{\text{peace}}, \mu^{\text{transition}}, \mu^{\text{war}} \sim \text{Normal}(0, 5) \text{ s.t. } \mu^{\text{peace}} < \mu^{\text{transition}} < \mu^{\text{war}} \tag{19}$$

$$\sigma^{\text{peace}}, \sigma^{\text{transition}}, \sigma^{\text{war}} \sim \text{Exponential}(2)$$

$$\pi \sim \text{Dirichlet}(\mathbf{0.01}) \tag{20}$$

4.6 Combined models

We consider two further models that combine these techniques. We fit a model that employs both random effects (RE) and the Bayesian data weighting scheme (W). We also consider a switching model that also incorporates data weighting (W). In total, we fit six models.

5 Evaluating the Models

Our validation of the six models consists of several steps parts. First, we evaluate their performance on semi-simulated data with a known true latent probability of civil war and compute accuracy statistics. We then generate a plot for visual comparison between the six models on simulated data. Finally, then apply the model to the actual dataset of civil war status, which allows further validation in a real world setting.

5.1 Validation on Simulated Data

To validate our models, we begin by producing semi-simulated data with a known "true" probability of civil war. Because we never have access to the true underlying probability of civil war in real datasets, we cannot directly compare the probability returned by our model to a ground truth dataset. Moreover, purely simulated data may differ in important ways from the true process of civil war onset and termination, with its complicated temporal dependence. To produce our evaluation data, we produce "semi-simulated" data, beginning with Fearon and Laitin's (2003) dataset of civil war. When Fearon and Laitin code a civil war onset or termination, we shift our latent war variable upward and toward by a random

	Model	RE	W	Switch	RMSE	Acc.	F1	α correct	δ correct
1	RE	✓			0.07	0.99	0.97	1.00	0.50
2	RE-W	✓	✓		0.07	0.99	0.97	1.00	0.25
3	Simple "Robust"				0.08	0.99	0.98	1.00	1.00
4	W		✓		0.08	0.99	0.98	1.00	1.00
5	Switching			✓	0.11	0.99	0.96	1.00	0.50
6	Switching-W		✓	✓	0.11	0.99	0.96	1.00	0.50

Table 2: *Evaluation results of six models on semi-simulated data. "Simple robust" is the model proposed by Reuning, Kenwick and Fariss (2019) described in section 4.2. "RE" refers to the random effect model described in section 4.3. "W" is the Bayesian data weighting approach described in section 4.4, and "Switch" refers to the model described in section 4.5.*

amount. By adding noise and autocorrelation, we can produce a dataset of latent civil war status that qualitatively matches the dynamics of civil war status. Further details on the semi-simulated approach are in Appendix A. Table 2 shows the results of the six models on our semi-simulated data.

For each of the six models, we compute three standard accuracy measures. First, because we are interested in the probability of civil war, we compute the root mean squared error (RMSE) of the model's predicted probability of civil war against the "true" probability of civil war in our simulated data. Next, we dichotomize both the inferred and true probabilities of civil war by coding civil was as 1 if Pr(civil war > 0.5). While this sacrifices information about the underlying probability, it allows us to calculate familiar accuracy measures such as simple accuracy $\left(\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(\widehat{\text{civil war}} = \text{civil war})\right)$ and the F1 score (the harmonic mean of precision and recall). Finally, because we know the true rater bias and discrimination in our simulated data, we can compare the inferred rater parameters against their true values. The final two columns in Table 2 show the whether the model recovers the correct ordering of the parameters. A value of 1 indicates that the model provides the correct rank for all parameters, 0.5 indicates that half of the parameters are in the correct order, and so on.

We find that the models are similar in their accuracy performance, but that the two switching models perform worse than the other four models. Only two models, the reweighting model (W) and the simple robust model, correctly recover the order of the rater parameters.

A second form of validation on simulated data is to visually inspect the latent variable estimates on simulated data. Figure 1 shows the outputs of the six models on four simulated countries. Because our simulated data includes the true probability of civil war, we can plot both the true probability (red) and the inferred probability (blue) with 90% credible intervals. The underlying observed civil war status from our simulated raters are shown in a

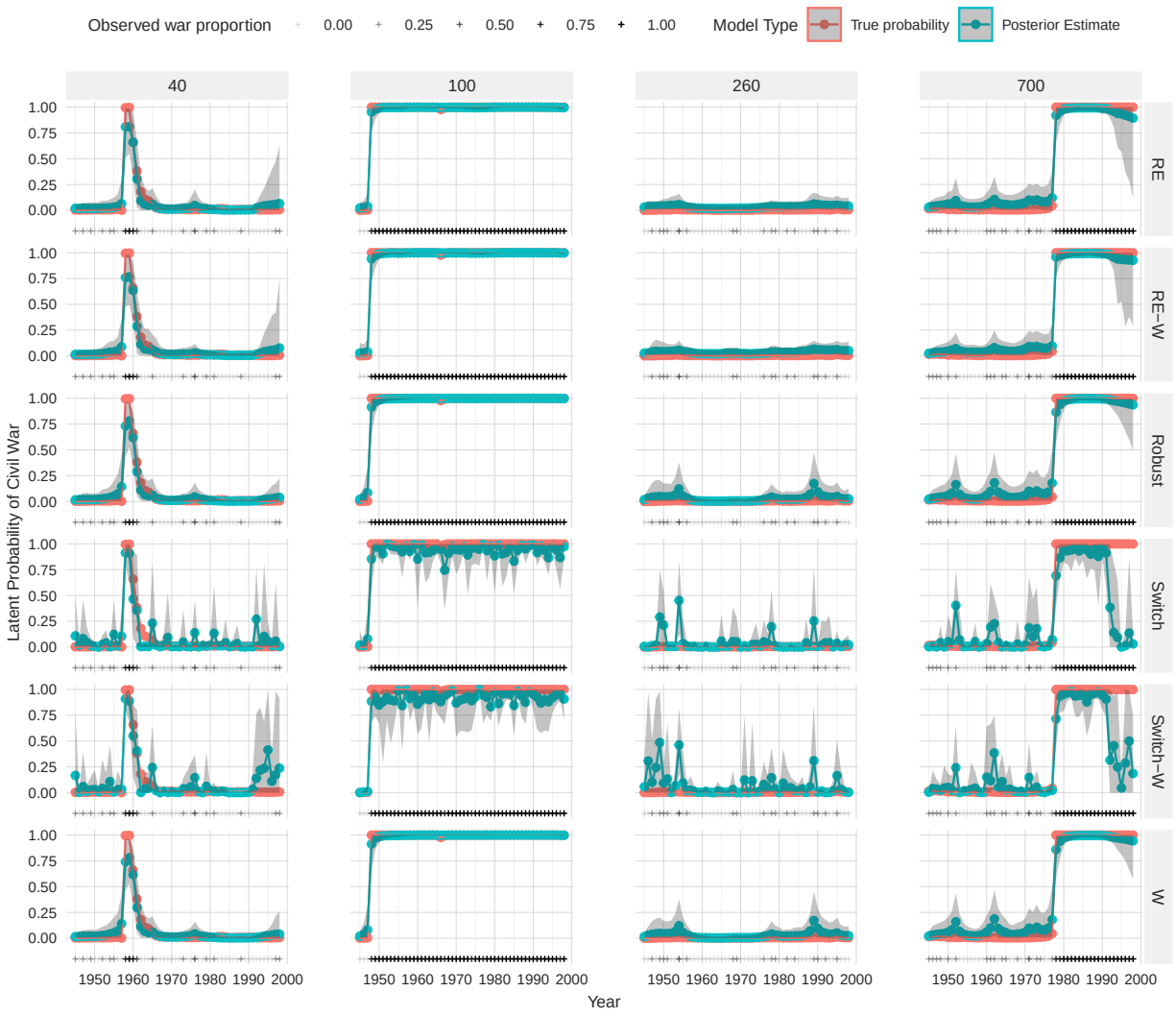


Figure 1: *Simulation results: Posterior estimates of latent civil war probability for four countries (columns) and six models (rows) on simulated data. Red lines show the true latent probability and the blue lines show the posterior estimate, with 90% credible intervals. The transparency of the + symbols indicates the proportion of the four raters who code civil war in that country-year. Because the data is semi-simulated, country names are replaced by numeric codes to prevent over-interpretation.*

modified rug plot under panel. The figure reveals several features of the models that are not apparent in the performance statistics. The two switching models perform quite poorly, with phantom onsets and terminations that deviate greatly from the true underlying probability. The robust model and the model with random effects + weighting (RE-W) are qualitatively the best models, closely matching the underlying probability and with few extraneous jumps. The RE-W model has a slight advantage in allowing sharper jumps in probability when civil war status changes, when compared to the robust model. Dynamic latent variable models often face a tradeoff between stability and allowing for sharp changes in the underlying latent variable (Reuning, Kenwick and Fariss 2019), and the RE-W model appears to provide both desirable traits.

5.2 Real Model Results

We then fit the models on the real dataset of civil war status described in section 3. Figure 2 shows the results of the two models that performed best on the simulated data: the baseline "robust" model and the random effects + reweighting model (rows) on four countries (columns). The black line and points show the mean posterior prediction with shaded 90% credible intervals. A modified rug plot below shows the proportion of datasets code civil war in each year. The four countries are selected to show a range of civil war dynamics. Afghanistan begins at peace and remains at war from 1979 onward, with high agreement between raters. Colombia shows a much more uncertain process of civil war, with at least a quarter of raters coding civil war in each year after 1950, but large disagreements. Cuba has a short war with high agreement, and the Netherlands show unanimous agreement on peace throughout the period.

The RE-W model provides greater differences from the simple moving average than the robust model, with sharper jumps between years and more instances of mean predictions at either 0 or 1. This is desirable in the case of Cuba, but is ambiguous in the cases of Afghanistan and Colombia. Qualitatively, the results for Afghanistan match our expectations for civil war status in 2002–2010. While the conflict was certainly internationalized through the involvement of NATO, the period also saw major Taliban efforts to regain control of the country. The results for Colombia are more ambiguous. While violence was at a much lower level in the 1970s than the rest of the conflict, fighting continued and assigning zero probability to civil war status seems overconfident.

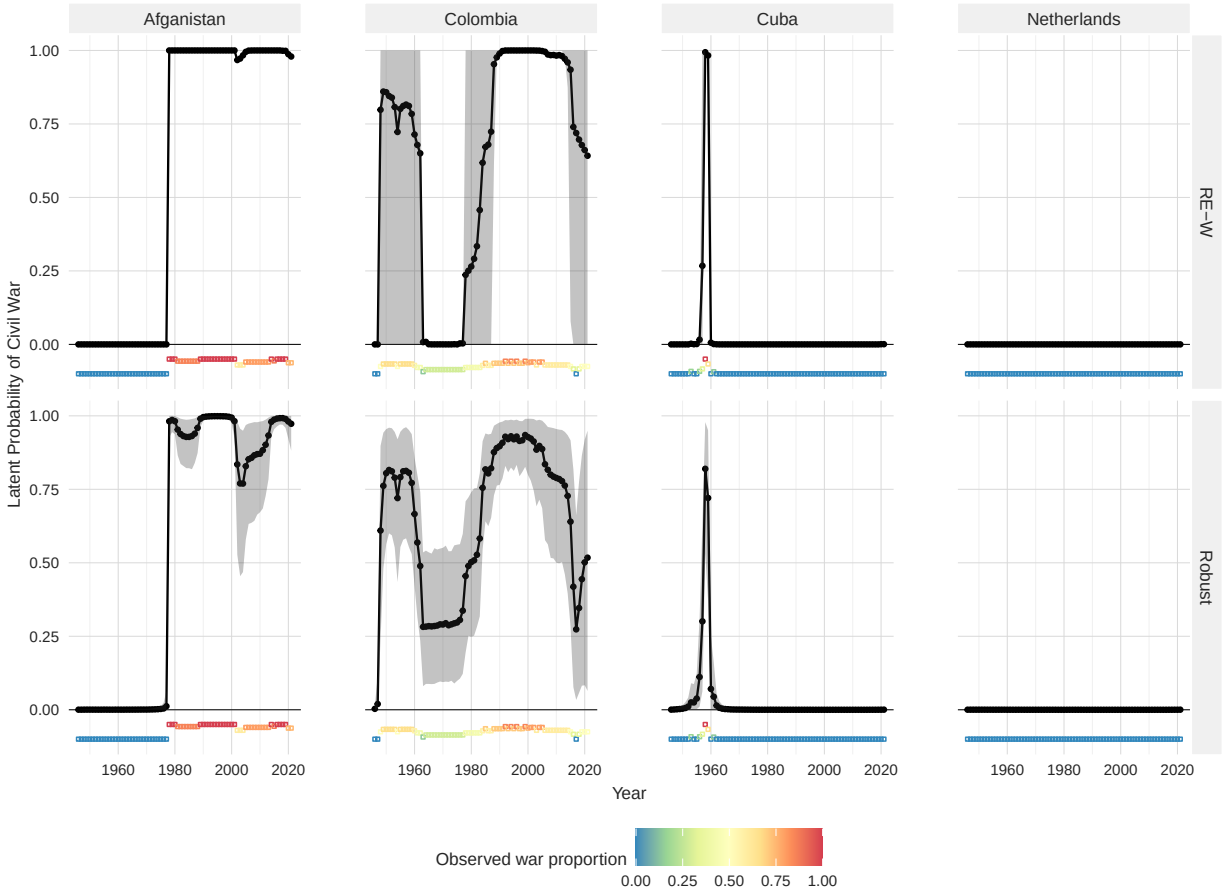


Figure 2: *Posterior estimates of latent civil war on real data, showing the probability for four countries (columns) and two models (rows) on real data. The top row shows the results of a random effect + weighting model, and the second row shows a "robust" dynamic model (Reuning, Kenwick and Fariss 2019). The colored points below the latent variable lines show the proportion of datasets that code each country-year as a civil war.*

5.3 Evaluating Bias Terms

As a final validation, we examine the estimated bias terms for each dataset in the RE-W model. In Figure 3 we depict the posterior distribution of the mean estimated rater difficulty for the RE-W model. Lower values correspond to lower thresholds for civil wars along the latent variable scale. ICEWS has the lowest estimated difficulty (bias) parameter, which matches its coding definition: it has no lower limit to the number of annual (or total) casualties required to code a violent domestic conflict. Similarly, UCDP minor civil wars have a threshold of 25 battle deaths per year, much lower than the remaining datasets. That ICEWS and UCDP major or minor obtain the two lowest difficulty parameters makes sense.

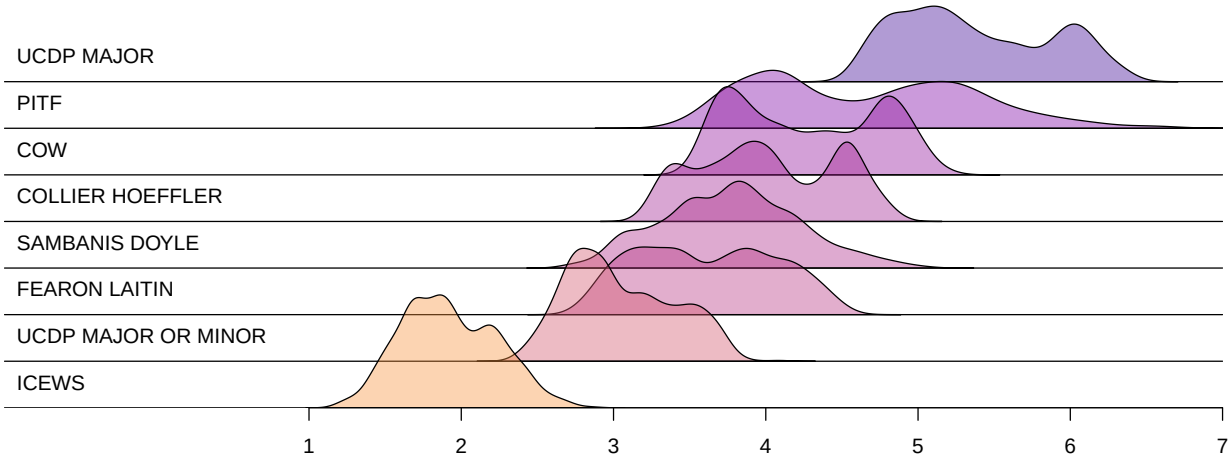


Figure 3: *Mean difficulty (bias) parameters by rater for the RE-W model, fit on real data.*

Fearon and Laitin requires only 100 battle deaths per year on average. Sambanis and Doyle, Collier and Hoeffler, COW, and UCDP Major all require some variation on 1,000 annual battle deaths. Therefore, that Fearon and Laitin obtains the third smallest bias parameter properly reflects the relative coding rules.

6 Incorporating Measurement Uncertainty

Because previous work on civil war onset uses single datasets of civil war status, they cannot incorporate measurement error into their analysis, only the sampling error estimated by regression models. We follow the recommendations by Schnakenberg and Fariss (2014) and Crabtree and Fariss (2015) to incorporate additional measurement uncertainty into the main results of the model. Building on the multiple imputation literature (King et al. 2001), they propose fitting M separate models on M copies of the original data, where each dataset is built with separate draws from the posterior estimate. To combine the sampling uncertainty within each model with the measurement uncertainty across different draws from the posterior. These are combined using a rule proposed by Rubin (1976, 1987):

$$\text{SE}(\theta) = \sqrt{\frac{1}{M}\sum_{i=1}^{m}\text{SE}(\theta_i)^2 + S_\theta^2(1+1/m)}$$

where the sample variance of θ across m estimates is: $S_\theta^2 = \sum_{i=1}^{m}\frac{(\theta_j-\bar{\theta})^2}{(m-1)}$.

In Figure 4 we show the results of re-analyzing the main result from Fearon and Laitin

(Table 1, Model 1), incorporating additional measurement uncertainty.[4] We take 250 draws from the posterior distribution of civil war status for each country-year. Each of these represent a complete set of possible war status for all country years. We then take a Bernoulli draw from the latent probability for each country-year and fit the original Fearon and Laitin model on that draw. Using Rubin's rule, we combine the sampling variance from within the models with the variance across each model.[5] The new variances that account for measurement error are very large, with no variable reaching statistical significance. This result diverges from our prior understanding of civil war and may indicate problems with the model or its convergence. We have very strong theoretical and empirical results to support the belief that GDP and population are predictive of civil war onset, in a way that should not depend on measurement uncertainty. While this approach to including measurement uncertainty in the dependent variable has promise, it depends on having an well functioning measurement model.

7 Improving Forecasting with Posterior Bagging

Civil war forecasting faces many challenges. One of these problems is a well-known (but normatively good) rare data problem. Civil war occurs relatively infrequently as a proportion of country-years since 1945. This problem is further exacerbated when researchers wish to rely on covariates that are only available for more recent times, such as text-derived event data. A second challenge is the problem of borderline cases, where fighting is occurring, but may not rise to the level of civil war.

We use draws from the posterior distribution of latent statuses to partially mitigate both of these problems. Incorporating the uncertainty in civil war status should improve our forecasts by increasing the amount of information we can include in our model. Rather than including or excluding borderline cases altogether as a hard 0/1 label would do, we can include borderline cases in proportion to their predicted probability. For instance, in Colombia in 2015, the ceasefire between the government and FARC broke down and several dozen combatants were killed in each side. While the casualties were below the 100 battle death threshold that some datasets use, it met the criteria for other datasets. If the measurement model gives a probability of 0.4 for civil war status in this year, it would allow some models in the ensemble to learn to predict civil war in this year, while other models would learn to predict no civil war.

[4]We thank Max Goplerud for detailed comments that corrected an earlier error in generating the figure.

[5]We conduct an additional test, where instead of repeatedly sampling the latent probability of war from the posterior, we average the posterior draws to generate a single posterior probability of civil war status for each country year. Result are similar, with slightly smaller variances.

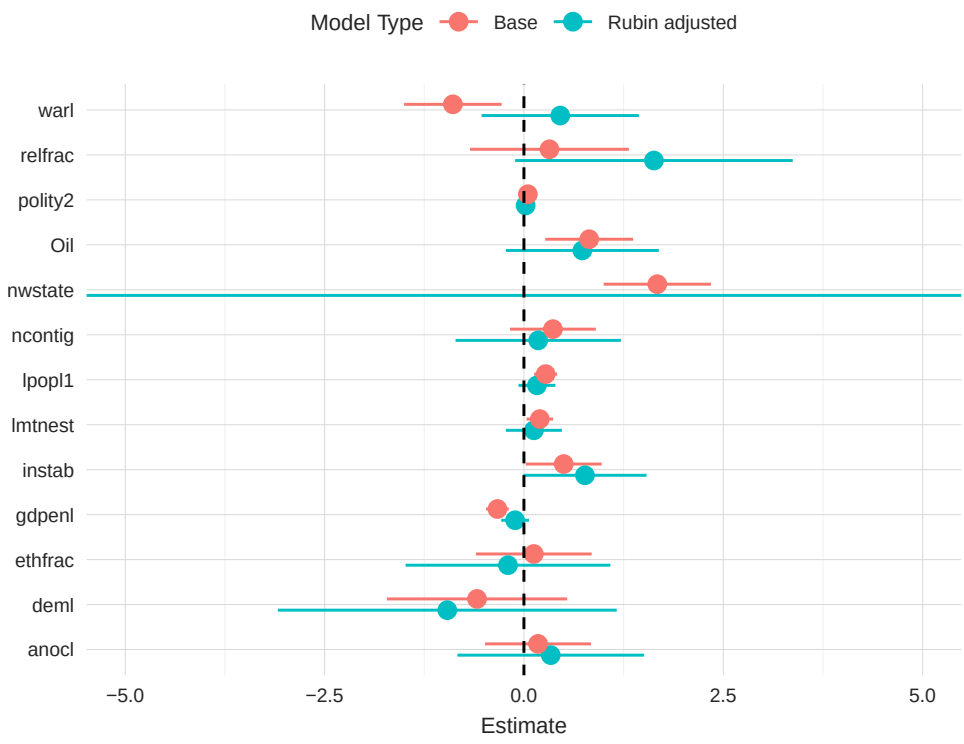


Figure 4: *Re-analyzing Fearon and Laitin's Model 1 (Table 1) with 95% confidence intervals that account for measurement uncertainty using Rubin's rule. We fit 250 separate models, each fit on a separate draw from the posterior probability of civil war for each country year. Binary war status is generated with a Bernoulli draw from the posterior probabilities.*

Breiman (1996) introduces *bagging*, a technique for aggregating multiple models trained on bootstrap samples of a dataset to improve their predictive accuracy.

$\mathcal{L}$: set of bootstrap draws $\{(x, y), (x, y)...\}$ from original data X.

$\phi(X, \mathcal{L})$: predictor for y given $\mathcal{L}$

Define the aggregate predictor $\phi_A(X) = \mathbb{E}_{\mathcal{L}}[\phi(X, \mathcal{L})]$

We then decompose the expectation over $\mathcal{L}$ of the mean squared error:

$$\begin{aligned}
\mathbb{E}_{\mathcal{L}}[(y - \phi(X, \mathcal{L}))]^2 &= y^2 - 2y\mathbb{E}_{\mathcal{L}}[\phi(X, \mathcal{L})] + \mathbb{E}_{\mathcal{L}}[\phi(X, \mathcal{L})]^2 \\
&\text{Recall the definition of the aggregate predictor:} \\
&= y^2 - 2y\phi_A(X) + \phi_A(X)^2 \\
&= (y - \phi_A(X))^2 \\
&\text{by Jensen's inequality } (\mathbb{E}[Z^2] \geq \mathbb{E}[Z]^2) : \\
\mathbb{E}_{\mathcal{L}}[y - \phi(X, \mathcal{L})^2] &\geq (y - \phi_A(X))^2
\end{aligned}$$

This implies that the mean squared error from the aggregate (bagged) predictor will be lower (or the same) as the expected MSE from a non-aggregated model. The derivation (omitted) is similar for a classifier predicting the maximum probability class.

The result that the aggregate predictor outperforms the original predictor depends on the variance of the predictive models across the datasets in $\mathcal{L}$: if the variance is low, taking the expectation of the predictor over the different draws will be very close to the predictor trained on the original dataset. Thus, bagging is most useful in cases where models have higher variance.

While bagging takes bootstrap draws from the entire dataset, our "posterior bagging" approach uses the posterior distribution of civil war status as a source of variability.[6] We generate $m = 1...M$ copies of the original dataset by sampling y values from the posterior distribution $y_m \sim Y$ and leaving X unchanged for each draw. We fit a model on each posterior draw and original X and aggregate the models into an ensemble forecast.

7.1 Posterior Bagging Simulation Results

We begin by providing forecasting results using our simulated data to show the effectiveness of the posterior bagging approach. We begin with a standard civil war forecasting setup, building on Muchlinski et al. (2015) and Colaresi and Mahmood (2017). While the specific models and techniques in Muchlinski et al. (2015) contain errors (Wang 2019; Heuberger

[6]Note that posterior bagging can easily be combined with traditional bagging, for example, by fitting multiple random forests on posterior draws.

2019), the dataset and base models have been used as foundations for other forecasting work (Colaresi and Mahmood 2017).

Specifically, we use the same country-year dataset as Muchlinski et al. (2015) and Colaresi and Mahmood (2017), and use the same set of 88 variables that each of those works employ. Rather than comparing logistic regression and random forest models, we opt for a consistent, simple logistic regression model to predict civil war onset.

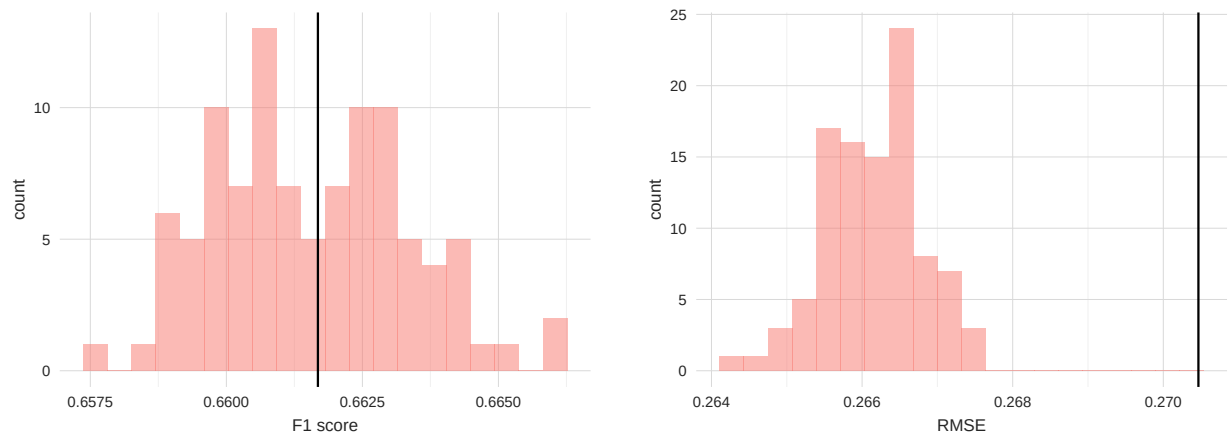
Most civil war forecasting projects evaluate the performance of their models using a set of metrics defined for binary outcomes, including Brier scores, the area under the receiver-operator curve (AUC), raw accuracy, or precision, recall, and F1 scores. The premise of our project is that civil war is a latent variable, best thought of as a probability of being in a civil war, rather than a simple binary coding. As a result, even in our simulated data, we have no binary ground truth for evaluating forecasting models. We evaluate our models using two metrics:

- root mean squared error (RMSE) between the predicted probability of civil war, and the "true" latent civil war status from our simulation. This choice is theoretically motivated, given that Brier scores, a common proper scoring rule for forecasts, can be seen as the binary case of the mean squared error loss.

- F1 score, using a dichotomization of the latent variable, where an observation with latent probability $\Pr(\text{civil war}) > 0.5$ is coded as having a civil war. Dichotomizing the latent variable sacrifices some information, but allows us to more easily compare our results to the forecasting literature.

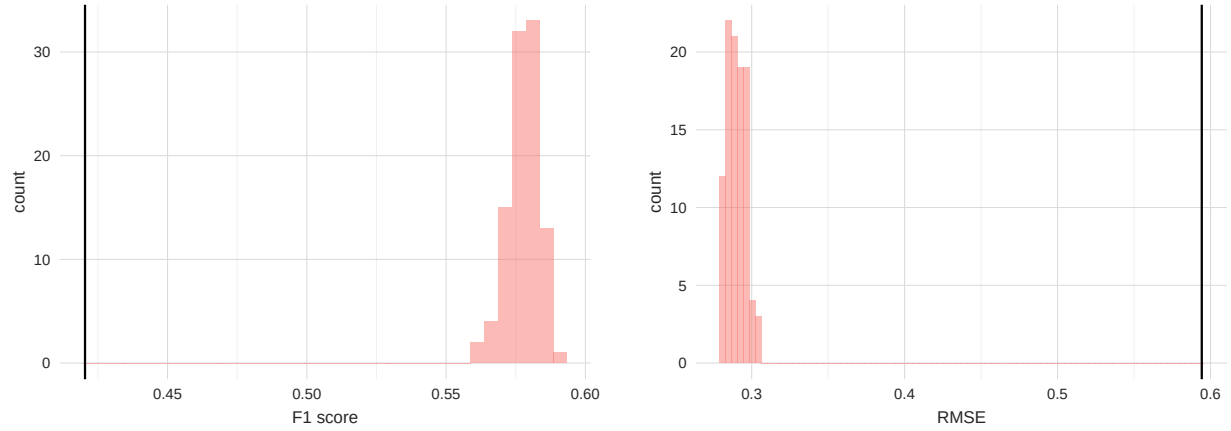
Figure 5 shows the out of sample F1 and RMSE scores for a forecasting model trained on the period 1946–1980 and evaluated on 1980–2000. (Note that the outcome is our semi-simulated latent variable.) The posterior bagging technique shows no improvement in F1 score, but shows marginal but significant gains in RMSE. The limited improvement offered by posterior bagging is not surprising, given the theoretical results above. Even with a relatively large number of covariates, the logistic regression model is fairly stable, yielding an aggregate predictor that is similar to the simple predictor.

However, posterior bagging does show major improvements in a limited data setting. Figure 6 shows the results of a forecasting model trained only on the 1980–1990 time period. While the performance of the posterior bagged model is lower than the model trained on the full time period, it offers very large improvements over the ordinary logistic regression model. The F1 for the posterior bagged model is around 0.58, compared to an F1 score around 0.42 for the ordinary logit model. The RMSE shows similarly dramatic improvements.



(a) *F1 Score* **(b)** *RMSE*

Figure 5: *Out of sample results for full time period forecasting on semi-simulated data. F1 and root mean squared error for ordinary logit (vertical line) and posterior-bagged logit (histogram). Trained on 1946–1980 and evaluated on 1981–2000.*



(a) *F1 Score* **(b)** *Root Mean Squared Error*

Figure 6: *Out of sample results for forecasting using a restricted date range on semi-simulated data. F1 and root mean squared error for ordinary logit (vertical line) and posterior-bagged logit (histogram). Trained on 1980–1990 and evaluated on 1991–2000.*

Note that fitting a single measurement model on the entire time period is unlikely to cause train/test set contamination because we do not use the latent probability of civil war as a feature in our model to predict civil war in future time periods. However, a promising direction for future forecasting research would be to fit separate measurement models for all years through $t-1$ and incorporate the latent probability of civil war as a feature to predict war in time t.

7.2 Forecasting Civil War Onset

The results in the previous section used semi-simulated data to show that posterior bagging offers improvements over simple logistic regression, especially in a limited data setting. Next, we turn to applying the posterior bagging approach to real data, using draws from the measurement model we discuss in Section 4.

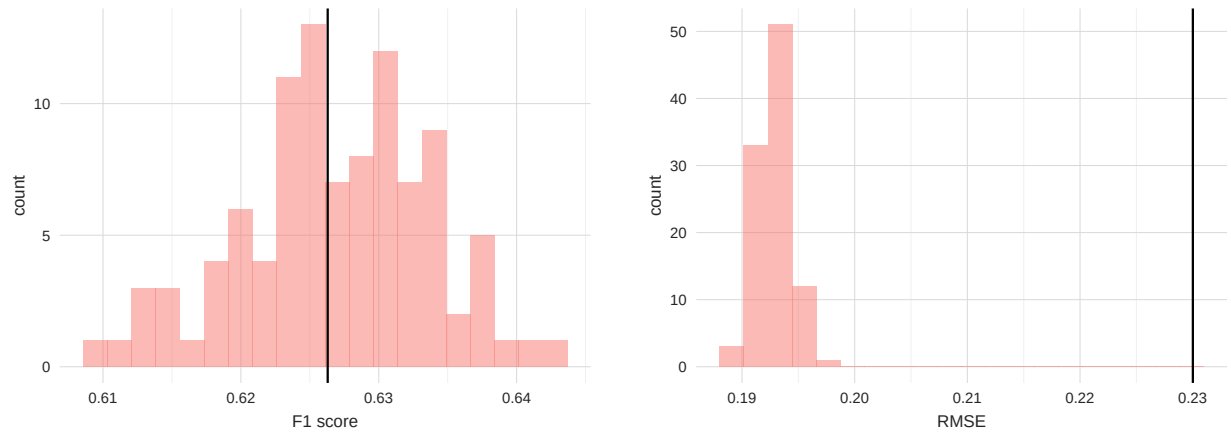
Figures 7 and 8 show the F1 and RMSE of the forecasting models. Here, the model differs from the previous forecasting model on simulated data, because we do not have access to a true underlying probability of civil war. Instead, we generate binary civil war status by dichtotomizing the latent probability of civil war from our measurement model. The results are similar to the forecasting model fit on simulated data. When trained on the longer time period, the posterior bagged model shows significant improvement in RMSE, but not in F1. On a restricted date range, however, the posterior-bagged model greatly outperforms the simple logistic regression model. When researchers fit models on shorter time periods, perhaps as a result of data limitations, posterior bagging their predictive models can greatly improve their models' accuracy.

8 Future Work

This project offers several promising directions for future work.

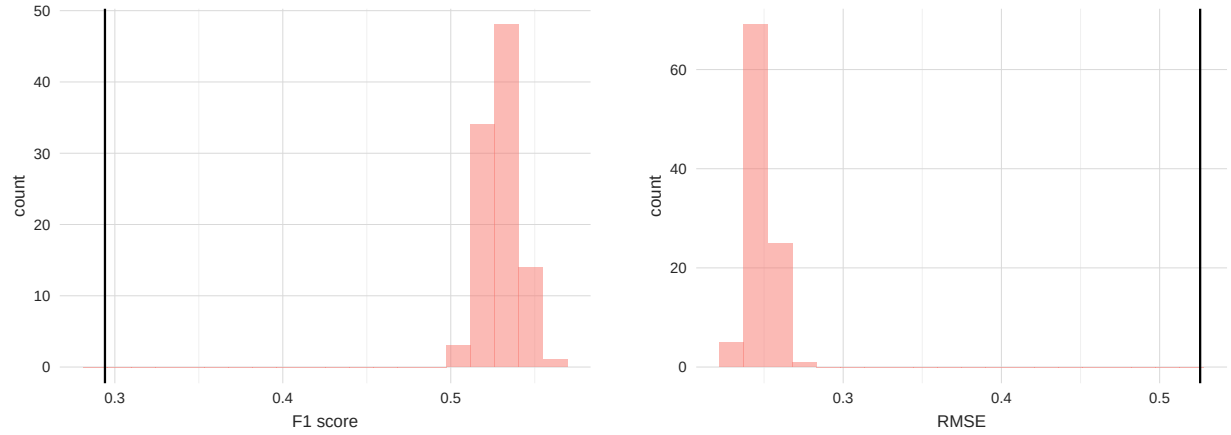
First, the measurement model provides the possibility of extending a single dataset's coding forward and backward in time. By using the estimated latent civil war status and the estimated rater bias and discrimination terms, a researcher can generate predicted codings for a single dataset. For instance, given the prominence of the original Fearon and Laitin dataset in the forecasting community, a researcher could extend the Fearon and Laitin dataset forward in time, inferring how the dataset would have coded other civil wars. This allows for greater consistency in evaluating models over time.

Second, the measurement model can be used to identify non-classical errors in civil war as a dependent variable, as discussed in Section 2. If civil war measurement error is correlated



(a) *F1 Score* **(b)** *RMSE*

Figure 7: *Out of sample results for full time period forecasting on semi-simulated data. F1 and root mean squared error for ordinary logit (vertical line) and posterior-bagged logit (histogram). Trained on 1946–1980 and evaluated on 1981–2000.*



(a) *F1 Score* **(b)** *Root Mean Squared Error*

Figure 8: *Out of sample results for forecasting using a restricted date range on actual data. F1 and root mean squared error for ordinary logit (vertical line) and posterior-bagged logit (histogram). Trained on 1980–1990 and evaluated on 1991–2000.*

with the covariates in a model explaining or predicting civil war, this induces a form of omitted variable bias that will yield biased coefficient results. We can obtain estimates of measurement error by comparing our posterior estimate of civil war status with the coding provided by each dataset and regressing the difference on the covariates used in a model. Our reweighting models offer a second approach to this: the weights provide more direct information on the countries where coders perform poorly, which may reveal systematic regional or income-based errors, similar to the patterns in missing political economy data (Lall 2016).

Finally, future improvements to the models, by reparameterizing or changing priors and running for more iterations, may provide better convergence and more precise estimates. This is especially likely in the case of the switching models, which display poor convergence.

9 Conclusion

Whether a country is experiencing a civil war is clearly latent–experts can code civil war status, but their binary decisions depend on their coding rules, varying interpretations, data limitations, and noise. We introduce a new measurement model to generate estimates of the probability that a country is experiencing a civil war in a given year. Our results show that, similarly to work on measurement models for other political phenomena, quantifying the uncertainty around civil war has implications for our conclusions about the causes of civil war onset.

We provide several contributions. First, we introduce the first treatment of civil war status as a latent variable, drawing on eight datasets of civil war status. Similar to other latent variables in political science, providing an estimate of latent civil war is useful in understanding the differing coding rules of each dataset, the uncertainty in civil war status, and its implications for both predictive and explanatory variables, as we demonstrate. We identify coder bias terms that match our expectations from the definitions of civil war in their codebooks.

Second, we provide new methodological contributions on dynamic latent variable models. We propose several extensions to previous work on "robust" dynamic models, including the use of random coder–unit effects, Bayesian data reweighting, and a switching model. We find that a combination of random effects and data reweighting provide good estimates of dynamic civil war status. We find that the switching model, while theoretically motivated, does not perform well on our simulated data.

Third, we show that measurement uncertainty around civil war status has implications for explanatory models of civil war onset. Incorporating measurement uncertainty into Fearon

and Laitin's classic model greatly expands the confidence intervals, but we remain cautious in interpreting the new confidence intervals given their implausibly wide range.

Finally, we introduce posterior bagging, a technique for improving forecasting models. Extending Breiman's original work on bagging, we show that taking draws from the posterior of civil war status, fitting a model on each draw, and aggregating their predictions improves the accuracy of forecasting models, especially in a limited data setting.

References

Breiman, Leo. 1996. "Bagging predictors." *Machine learning* 24:123–140.

Colaresi, Michael and Zuhaib Mahmood. 2017. "Do the robot: Lessons from machine learning to improve conflict forecasting." *Journal of Peace Research* 54(2):193–214.

Collier, Paul and Anke Hoeffler. 2002. "On the incidence of civil war in Africa." *Journal of conflict resolution* 46(1):13–28.

Crabtree, Charles D and Christopher J Fariss. 2015. "Uncovering patterns among latent variables: Human rights and de facto judicial independence." *Research & Politics* 2(3):2053168015605343.

Doyle, Michael W and Nicholas Sambanis. 2000. "International peacebuilding: A theoretical and quantitative analysis." *American Political Science Review* 94(4):779–801.

Fariss, Christopher J. 2014. "Respect for human rights has improved over time: Modeling the changing standard of accountability." *American Political Science Review* 108(2):297–318.

Fariss, Christopher J. 2019. "Yes, human rights practices are improving over time." *American Political Science Review* 113(3):868–881.

Fearon, James D and David D Laitin. 2003. "Ethnicity, insurgency, and civil war." *American Political Science Review* 97(01):75–90.

Fong, Christian and Justin Grimmer. 2019. "Causal inference with latent treatments." *American Journal of Political Science* .

Gleditsch, Nils Petter, Peter Wallensteen, Mikael Eriksson, Margareta Sollenberg and Håvard Strand. 2002. "Armed conflict 1946-2001: A new dataset." *Journal of peace research* 39(5):615–637.

Heuberger, Simon. 2019. "Insufficiencies in Data Material: A Replication Analysis of Muchlinski, Siroky, He, and Kocher (2016)." *Political Analysis* 27(1):114–118.

King, Gary, James Honaker, Anne Joseph and Kenneth Scheve. 2001. "Analyzing incomplete political science data: An alternative algorithm for multiple imputation." *American political science review* 95(1):49–69.

Krüger, Jule and Ragnhild Nordås. 2020. "A latent variable approach to measuring wartime sexual violence." *Journal of Peace Research* 57(6):728–739.

Lall, Ranjit. 2016. "How multiple imputation makes a difference." *Political Analysis* 24(4):414–433.

Lustick, Ian, Sean O'Brien, Steve Shellman, Timothy Siedlecki and Michael Ward. 2015. "ICEWS Events of Interest Ground Truth Data Set.".
URL: *https://doi.org/10.7910/DVN/28119*

Marquardt, Kyle L and Daniel Pemstein. 2018. "IRT models for expert-coded panel data." *Political Analysis* 26(4):431–456.

Marshall, Monty G. 2019. "Major episodes of political violence (MEPV) and conflict regions, 1946-2018." *Center for Systemic Peace* .

Melander, Erik, Therése Pettersson and Lotta Themnér. 2016. "Organized violence, 1989–2015." *Journal of Peace Research* 53(5):727–742.

Millimet, Daniel L and Christopher F Parmeter. 2022. "Accounting for skewed or one-sided measurement error in the dependent variable." *Political Analysis* 30(1):66–88.

Mislevy, Robert J. 1991. "Randomization-based inference about latent variables from complex samples." *Psychometrika* 56(2):177–196.

Muchlinski, David, David Siroky, Jingrui He and Matthew Kocher. 2015. "Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data." *Political Analysis* 24(1):87–103.

Pemstein, Daniel, Stephen A. Meserve and James Melton. 2010. "Democratic Compromise: A Latent Variable Analysis of Ten Measures of Regime Type." *Political Analysis* 18(4):426–449.

Reuning, Kevin, Michael R Kenwick and Christopher J Fariss. 2019. "Exploring the dynamics of latent variable models." *Political Analysis* 27(4):503–517.

Rubin, Donald B. 1976. "Inference and missing data." *Biometrika* 63(3):581–592.

Rubin, Donald B. 1987. *Multiple imputation for survey nonresponse*. New York: Wiley.

Sarkees, Meredith Reid and Phil Schafer. 2000. "The correlates of war data on war: An update to 1997." *Conflict Management and Peace Science* 18(1):123–144.

Schnakenberg, Keith E and Christopher J Fariss. 2014. "Dynamic patterns of human rights practices." *Political Science Research and Methods* 2(1):1–31.

Treier, Shawn and Simon Jackman. 2008. "Democracy as a latent variable." *American Journal of Political Science* 52(1):201–217.

Ulfelder, Jay and Sean J Taylor. 2015. "A Measurement Error Model of Dichotomous Democracy Status." *Available at SSRN: http://dx.doi.org/10.2139/ssrn.2726962* .

Wang, Yixin, Alp Kucukelbir and David M. Blei. 2017. Robust Probabilistic Modeling with Bayesian Data Reweighting. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17 JMLR.org p. 3646–3655.

Wang, Yu. 2019. "Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data: A comment." *Political Analysis* 27(1):107–110.

A Semi-Simulated Civil War Details

- $Y_{it} \in \{0, 1\}$ is the original binary coding of civil war, where 0 indicates peace and 1 is war.
- $Y^*_{it} \in \mathcal{R}$ is a latent status.
- $S_{it} \in \{0, 1\}$ is a change in status, where 1 indicates an onset, and 0 indicates a termination

We initialize our latent variable, Y^*_{it} at -6 if the country begins at peace, at 4 if it begins at war. Then, for following years, we generate the latent variable as follows:

$$
\begin{aligned}
Y^*_{it} &= 0.8 \cdot Y^*_{i,t-1} && \text{autocorrelation in status} \\
&+ (1 - Y_{it})\mathcal{N}(-1, 0.3) && \text{downward trend during peace} \\
&+ Y_{it}\mathcal{N}(0.6, 1) && \text{upward trend during war} \\
&+ S_{it}\mathcal{N}(5, 1) && \text{positive shock for onset} \\
&+ (1 - S_{it})\mathcal{N}(-4, 1) && \text{negative shock for termination} \\
&+ S_{i,t+1}\mathcal{N}(1, 1) && \text{increase in the year before onset} \\
&+ S_{i,t+1}\mathcal{N}(1, 1) && \text{increase in the year following termination}
\end{aligned}
$$

To convert the latent variable into a latent probability of civil war, we apply a logistic transformation: $\pi_{it} = \text{logit}^{-1}(Y_{it})$.

To generate observed codings for each rater $Y_{itk} \in \{0, 1\}$, we apply the two parameter IRT model using rater-specific parameters α_k and δ_k for $k = 1...4$ raters:

$$
\begin{aligned}
Y^*_{itk} &= \text{logit}^{-1}\left(\alpha_k(Y^*_{it} - \delta_k)\right) \\
Y_{itk} &= \text{Bernoulli}(Y^*_{itk})
\end{aligned}
$$