Journal of Peace Research

Lessons from near real-time forecasting of irregular leadership changes

Journal of Peace Research
2017, Vol. 54(2) 141–156

DOI: 10.1177/0022343316680858
journals.sagepub.com/home/jpr
SAGE

Michael D Ward & Andreas Beger

Department of Political Science, Duke University

Abstract

Since 2014, we have been producing regular six-month forecasts of the probability of irregular leadership changes – coups, rebellions, protests that result in state leader changes – for most countries in the world for the Political Instability Task Force (PITF). During 2015, we issued new forecasts each month, with a delay as short as five days and no longer than two weeks into each six-month forecasting window. This article describes the approach we use to generate our forecasts and presents several examples of how we present forecasts. The forecasts are derived from a statistical ensemble of seven thematic models, each based on a split-population duration model that aims to capture a specific argument or related set of covariates. This approach is modular in that thematic models can be swapped out or new models integrated, and it lessens the need for generalist 'kitchen sink' models. Together, the models achieve high out-of-sample accuracy. Based on our experience, we draw conclusions about the practical, policy, and scientific aspects of this and similar undertakings. These include thoughts on how to evaluate and present forecasts, the potential role of ensembles in model comparison, the role of ensembles and prediction in causal research, and considerations for future efforts in forecasting and predictive modeling.

Keywords

coups, EBMA, ensemble, forecasting, ILC, PITF, prediction, protest, rebellion, split-population duration regression

Introduction

What if the President of Russia is overthrown this year? How will Turkey's policies towards Syria and mass migration to Europe change if the military were to reassert power? In 2014, there were five instances of similar unexpected and irregular transitions between sitting leaders of states, including one in Thailand and two in Burkina Faso, which has had another in 2015. Ukraine is still embroiled in a conflict that started when the then president was overthrown in a revolution, as is Yemen after an irregular change in 2015. Although the means by which the leaders of these countries were overthrown vary, ranging from military coups and revolutions to armed rebellion, we treat them as a common outcome, which we call *irregular leadership changes*, or ILC for short (Beger, Dorff & Ward, 2016).

Over the past several years, we have been providing regular global six-month forecasts of ILCs with a delay of as few as five days from the start of the forecast period to delivery of a forecast. The underlying focus on ILCs as an outcome of interest, rather than the mechanisms through which they can occur, is an important empirical addition to the literature. Processes of change are often theoretically not separable (Bueno de Mesquita & Smith, 2015). The coup in Mali in 2012, for example, was driven by dissatisfaction with how the civilian government handled the Tuareg rebellion in the north, while the most recent Thai coup involved mass protests. Trying to predict coups without considering the impact of mass protests or armed rebellion and vice versa may thus not be as fruitful as considering the joint outcome.

From a more general perspective, our effort flows from research into the duration of political authorities, regimes, and polities (Eckstein & Gurr, 1975). The basic

Corresponding author:
michael.don.ward@gmail.com

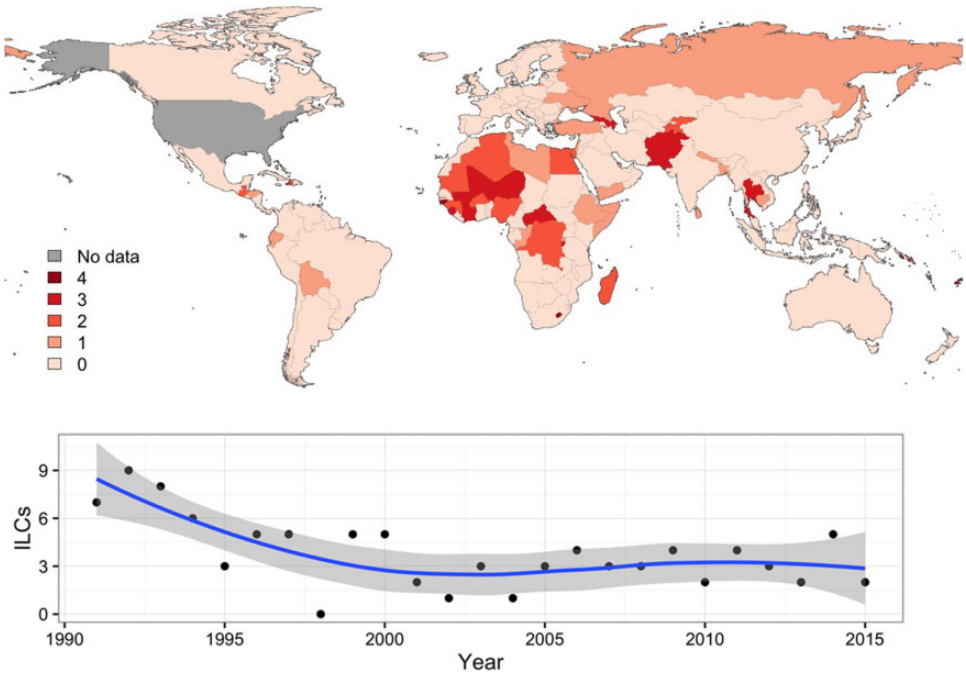

Figure 1. Spatial and temporal distribution of ILCs, 1991–2015

thrust of Eckstein and Gurr was to characterize situations by authority relations, defining underlying dimensions as a way of getting at the duration of different political constellations. 'Persistence and change in political systems, 1800–1971' (Gurr, 1974) first developed the idea of calculating the half-lives of different sorts of political polities, regimes, and authorities. We apply this basic notion to produce forecasts of their 'half-life'.[1]

Forecasting ILCs is also important in practice. To the extent that leaders of foreign states matter for foreign policy, we should care whether one is facing a risk of irregular and thus unexpected removal. Not only are ILCs often a result of violent and damaging processes, but the instability they cause can spark significant levels of violence and civil war (Powell & Thyne, 2011: 256) and also has negative economic effects (Alesina et al., 1996).

Forecasting rare political events

ILCs are rare events. Over the past 15 years, on average we see between two and four per year (Figure 1). One might question the utility of trying to develop models that will help to understand or to forecast rare events; but it is often said that reality is a low probability event. ILCs are low probability, high impact events that often change the world in consequential ways. Imagine the impact of a coup tomorrow in Russia or Turkey. Such events would certainly change the configuration of regional politics in Europe and the Middle East, let alone what might be the consequences for citizens living in these states. Despite – or because of – the rarity of such events, it is useful to develop better understandings of such occurrences. Toward that end, we develop a model of the duration of incumbent leaders.

Our forecasts are based on an ensemble of several split-population duration regression models using monthly data from 1991 to the present (actually to July 2015). The basic idea is to develop a model which is a mixture that can apply to situations where the risk of an irregular turnover is rare (for example, New Zealand) and have a separate but integrated model that applies to cases in which the risk of irregular turnover is high (e.g. Fiji or Ukraine). Such split-population models are widespread in the medical literature, but have only recently been adapted and employed in the social sciences. We present only a relatively brief summary of the methodology, but the Online appendix contains a more detailed description and estimation results.[2]

[1] Long ago in a distant galaxy, Ward was a student of Gurr, but it took him a long time to return to studying the persistence of political regimes.

[2] See also Beger et al. (2016) and the spduration R package.

It turns out that such split models can be combined in ensemble approaches, so that the goal becomes to get a diverse set of models, rather than search for the best model. Ensembles are better than their components on average, and this is true for split population models of regime change as well. At the same time, we do not employ ensemble techniques that prevent one from determining what changed from one period to the next. This is a constraint that we face that is important for being able to explain results to decisionmakers and non-academic consumers of such efforts.

What are ILCs?

Irregular leadership changes (ILCs) are transitions between political leaders that occur in contravention to the laws or conventions of a state at the time the transition took place. Hereditary succession can be regular in a monarchy, but if that state later became a democracy, another hereditary succession would be irregular. The key is that a leadership transition does not follow the expected pattern, whether established legally, by convention, or reasonable expectation.

We use the Archigos (Goemans, Gleditsch & Chiozza, 2009) Political Leaders dataset to determine whether and when an ILC occurred. Archigos identifies the political leaders of states, and for each, codes when they entered and left office. The entry and exit are classified as regular or irregular, *based on the established rules or conventions of a state at the time the change took place*. Further details are also coded, including exit due to death or illness, and for irregular exit the means (e.g. coups, rebellion, revolution) as well as whether there was foreign involvement. Since we need up-to-date information on this, we actually code each month anew, using the Archigos coding framework. Thus, our version of Archigos is current, and while we share our information with Hein Goemans and Kristian Gleditsch – curators of the database – we may reach slightly different conclusions about transitions. The monthly data are coded using our interpretation of the Archigos codebook.[3] Unlike the standard Archigos data, we do not maintain an elaborate textual history of each change.[4] Our data coding is done in an increasingly automated fashion using online sources such as rulers.org and modern text extraction techniques, then checked manually by the authors and shared with the Archigos team.

ILCs are coded based on whether there was either an irregular exit or entry from office. Often, but not always, the two co-occur. In Archigos, irregular exits are further classified into ten subcategories that give a sense of the variety of processes that can lead to an ILC. From the coding manual, but ordered by occurrence:

(1) Leader removed by domestic military actors without foreign support.
(2) Leader lost power as a result of domestic popular protest without foreign support.
(3) Leader removed by domestic rebel forces without foreign support.
(4) Leader removed by other domestic government actors without foreign support.
(5) Leader removed in a power struggle within military, short of coup, i.e. without changing institutional features such as a military council or junta.
(6) Leader removed through assassination by unsupported individual.
(7) Leader removed in an irregular manner through other means or processes.
(8) Leader removed by domestic rebel forces with foreign support.
(9) Leader removed by domestic military actors with foreign support.
(10) Leader removed by other domestic government actors with foreign support.

Figure 1 shows the spatial distribution of ILCs from 1991 through the end of 2014, along with a timeline of the number ILCs by year. ILCs seemingly cluster in West Africa and Central Asia. They also occurred more frequently in the early 1990s, before leveling off at the current level of 3–4 per year. Cases with ILC overlap with coups, armed rebellions, and popular revolutions. We view the relationship as one of outcome versus mechanism – for example, coups are one way in which the outcome of ILC can occur. Empirically, all *successful* coups, rebellions, and revolutions constitute an ILC, but the reverse is not completely the case

Data, real and imputed

The basic unit of observation for our data is the country-month. The dependent variable indicates whether an ILC occurred in a given country in a given month. Importantly for duration modeling this implies the length of time since the last change. The list of states

[3] http://www.rochester.edu/college/faculty/hgoemans/Archigos.2.9-August.pdf

[4] Occasionally, we disagree with the eventual Archigos coding and keep our own. An example is the case of the election of Medvedev as Russian president in 2008, which is not recognized as a leadership change by Archigos.

is based on the roughly 175 countries identified by the Gleditsch & Ward (1999) list of states from 1991 through the present. The covariates used in the models draw on four main sources: the World Bank World Development Indicators (World Bank Group, 2013), Polity (Marshall & Jaggers, 2015), Ethnic Power Relations (Wimmer, Cederman & Min, 2009), and various indicators constructed from the ICEWS event data (Ward, Ahlquist & Rozenas, 2012; Boschee et al., 2015; Lautenschlager, Shellman & Ward, 2015). The structural data constitute structural indicators that are measured annually or at similarly infrequent time intervals, and as a result vary mostly between rather than within countries. The event data aggregations vary widely across both time and countries, and lastly we compute spatial lags of several event aggregations, which have little spatial variation but vary significantly over time.[5]

All covariates are lagged by one time period, and missing values are imputed using a mix of copula imputation (Hoff, 2007) or carry forward and univariate time-series models for extrapolation of variables missing for recent time periods.

Ensemble and thematic models

We are modeling a rare event that is multicausal and complex. There have been fewer than 100 ILCs over the past 25 years. At the monthly level, the base rate of ILCs is 17 events per 10,000 country-months. In comparison, the data in Fearon & Laitin (2003) have a base rate of 167 civil war *onsets* per 10,000 country-*years*.

There is no single general model in the extensive literature in the social sciences for this newly conceptualized variable. Moreover, finding the best model may not be the best predictive strategy. We employ an ensemble that combines seven thematic models, each of which is based on a split-duration regression model.[6] The thematic models attempt to capture the basic, but plausible, explanations for irregular leadership changes.

The ensemble approach is robust, mimicking a wisdom of the crowds which has shown itself to produce – on average – better estimates. We use an ensemble not only because in principle it makes sense, but also because it generally predicts better and more robustly, meaning with less sensitivity to the specific data it was trained on, than any individual component model. It allows us to avoid searching for the best single model with available data, only to have that model fail in the prediction enterprise because it is too deeply attuned to extant information. Because it allows us to sidestep the issue of choosing a single 'best' model for forecasting ILCs, an ensemble approach provides a way to accommodate the multiple mechanisms that lead to ILCs. It is also modular, allowing us to expose the detailed workings of each theme in the ensemble. This is important in conveying the results of a forecasting effort to clients, who often want to know why a forecast has changed.

The combined forecast is created using ensemble Bayesian model averaging (EBMA). It constructs a weighted average of multiple predictions in such a way as to maximize fit against observed outcomes during a calibration period (Montgomery, Hollenbach & Ward, 2012). EBMA estimates three sets of parameters, W, a_0 and a_1, using an expectation maximization (EM) algorithm. To construct the ensemble prediction p, the predictions from each of K models are first transformed to the logit scale and bias-adjusted to reduce the effect of outliers, producing f_k, and then further transformed and weighed using the estimated parameters:

$$\begin{aligned} g_k(y|f_k) &= \text{logit}^{-1}(a_{k0} + a_{k1} \times f_k) \\ p &= \sum_{k=1}^{K} w_k g_k(y|f_k) \end{aligned} \quad (1)$$

EBMA is indifferent to the way in which a stream of predictions f'_k is generated, as long as they are probabilities. Model or method do not directly matter, and we could incorporate streams of expert forecasts or even dart-throwing chimps.

Unlike machine-learning methods like neural nets, SVM, and decision trees/random forests, EBMA preserves our ability to dig into and describe the factors driving high or low forecasts, as well as their monthly changes. Alternative approaches often are 'black box', meaning that the mapping from inputs to outputs is more complex and less easily interpreted. This conflicts with the premium in forecasting for being able to explain which factors are driving high forecasts. In the Online appendix, we detail a comparison exercise between our approach, random forest, and Lasso regression, which shows that the EBMA approach achieves similar performance to these alternatives, albeit with less sensitivity to small changes in data and without giving up either interpretability or control over specification.

In an ideal state, the ensemble would draw on a suite of models strongly grounded in the literatures on coups, rebellions, and revolutions. Instead we use seven 'thematic' models: leader characteristics, public discontent,

[5] See Table A12 in the Online appendix.

[6] We call the models 'thematic' or based on 'themes' because while they are not recreations of existing models from the literature, they are either inspired by one or were built around related variables, like financial indicators.

global instability, protest, contagion, internal conflict, and financial pressure. We developed these models where we could from existing models like the political instability model in Goldstone et al. (2010), or inspired by general concepts in relevant literatures like conflict contagion (e.g. Buhaug & Gleditsch, 2008), or, lastly, based on groupings of indicators like finance.

The reason we did not draw more strongly on existing literature is a major question we will return to. The basic point to note here is that our contribution is not in the underlying models used to forecast ILCs. On the contrary, the models are modular and exchangeable, and we welcome efforts to replace them with better alternatives.

The thematic models are estimated using split-population duration regression. This allows us to capitalize on the fact that many countries, like Canada, will for all practical purposes not experience an ILC, which helps ameliorate the sparsity of our outcome (for an application in political science, see Svolik, 2008). Split-population duration regression is a mixture of two equations, one that reflects a traditional duration model with time-varying covariates and accelerated failure time form to produce an estimate of hazard, and a second logistic equation that estimates whether a case belongs to a population that is immune and will never experience the event, or at risk and will experience it at some point. The likelihood is given as a product of the immunity π estimate, failure function $f(t_i)$, and survivor function $S(t_i)$, where δ_i indicates whether a spell ended in failure:[7]

$$L\{\theta|(t_1, \ldots, t_n)\} = \prod_{i=1}^{N}\{(1-\pi)f(t_i)\}^{\delta_i} \times \{\pi + (1-\pi)S(t_i)\}^{1-\delta_i} \quad (2)$$

Forecast procedure

At this point in the process of producing predictions, we have data as well as both model and ensemble parameters that have been estimated. This allows us to create theme model predictions and to aggregate these into a single ensemble prediction for each country-month. To generate actual forecasts requires two additional steps. The full algorithm is outlined in Figure 2.

First, to produce a single six-month forecast for each country rather than six separate forecasts for each month ahead – our unit of observation is the country-month – we combine the monthly predictions using probability disjunction, that is, by calculating the probability that at least one ILC will occur in a given country over the next six months.

With c indexing country, t month, m theme model, and using the last month of observed data, for $i = 1$ to 6:

1. Extrapolate data using updating and carry-forward
   $X_{c\times t+i} = f(X_{c,t})$
2. Calculate theme model predictions.
   $P_{c\times m\times t+i} = \text{Theme}_m(X_{c\times t+i})$
3. Calculate ensemble from theme predictions
   $p_{c\times t+i} = \text{EBMA}(P_{c\times m\times t+i})$
4. Collapse to single 6-month forecast.
   $p_c = 1 - \prod_{i=1}^{6}(1-p_{c,t+i})$

Figure 2. Algorithm to generate six-month forecast

Second, as our forecasts cover the actual future we observe neither outcomes nor covariates. To have a basis for forecasting we extrapolate covariate values available in the present month six months into the future, using a combination of carry-forward extrapolation, updating of predictable indicators like a leader's age, and exponential state space models for annually measured structural variables like GDP (Hyndman et al., 2008).

To have relevant fit statistics, we replicate these conditions in our test forecasts. For example, for the six-month test forecast from January 2014, we retain data through that month, but drop all data, both outcomes and covariates, past that date. Contrast this with conventional out-of-sample testing, which would drop outcomes, but retain covariate values.

Data partitions

As we are aiming for live forecasts, our data are regularly updated and we use the last available month to forecast. Otherwise, estimation and validation requires that we partition our data into several sets, which are summarized in Table I. To ameliorate left-censoring, we use information from Archigos back to 1955 when coding spells; otherwise the full data range from 1991 to the present. Of the latter data, we reserve the majority for estimating the thematic split-duration models, and another roughly 2.5 years for calibrating the ensemble. This leaves us with about three years for testing and the live forecasts.

Estimates and discussion

As we are less interested in the specific thematic models, their estimates are shown in the supplemental information in the Online appendix. Table II summarizes the

[7] A subject – country in our case – can experience multiple events. A spell is the period from after an event up to the next event or the end of our data.

Table I. Data partitions

Start	*End*	*t*	*Use*
1955-01	–	728	Archigos only, used to code duration and risk variables to reduce left-censoring.
1991-01	2009-12	226	Training data; estimate thematic models.
2010-01	2012-04	28	Calibration; fit EBMA ensemble using predictions from thematic models.
2012-05	2015-07	39	Test; rolling six-month test forecasts.
2015-08	2016-01	6	Forecast period.

ensemble results, and Table III shows a few of the top forecasts for the period from September 2015 to February 2016 derived from this ensemble.

In the EBMA results, of primary interest are the model weights W, which gives an indication of the strength each thematic model plays in the joint prediction. They are, along with a_0 and a_1, estimated to maximize calibration period fit and used according to Equation (1) to transform the theme model predictions – densities which are shown in the right part of plot (a) in Figure 3 – into the EBMA predictions summarized by the left-most density.

Higher weights are placed on the predictions from the leader characteristics and contagion models, but overall the weights are fairly even across the seven models. Compare these to the Brier scores, AUC-ROC, and the area under the precision-recall curve (AUC-PR) fit statistics for thematic models in the calibration period, when the EBMA parameters are estimated.[8] Intuitively, better fitting models should have a higher weight in the forecasts, but – hint – this is not exactly the case. This leads to two questions: *What is the relationship between ensemble weights and theme model fit?* and what is AUC-PR and why do we include it, more, more broadly: *How should we assess the accuracy of our forecasts?* We will also touch on a third question which by now surely has crossed the reader's mind, *Why are the theme models not more strongly grounded in the existing literature?*

Accuracy

In addition to the common Brier score and AUC-ROC fit statistics for our binary outcome, we also present the area under the precision-recall curve (AUC-PR). The PR curve, similar to a ROC curve, plots recall against precision over the threshold values used to dichotomize probabilistic predictions. While AUC-ROC is easier to interpret in the context of prediction than Brier scores, AUC-ROC values are misleading when used with rare events and should always be supplemented with AUC-PR.

The classification matrices in Figure 4 illustrate why. Each matrix plots a set of predictions against outcomes, for a frequent and rare outcome respectively. The AUC-ROC is based on the tradeoff between recall (sensitivity) and the false positive rate, that is, *false positive relative to true negative predictions*. Both matrices have identical values for these. The AUC-PR, since it depends on precision rather than the false positive rate, would correctly pick up on the large number of false positive *relative to true positive* predictions in the right matrix. Our models are in the world of the right-hand plot, which roughly matches the base rate in our data at 0.0019, or 1 in 530 country-months. So is most conflict research.

Lastly, conventional wisdom would suggest that we use out-of-sample testing to assess the accuracy of our models and guard against overfitting (see Ward, Greenhill & Bake, 2010). Since an indication of the accuracy of our forecasts is desirable, and when forecasting we do not observe future covariate values, unlike in conventional out-of-sample testing, we instead use our test data to perform multiple historical six-month forecasts. We do so by iterating over the available months, generating forecasts without any data beyond the current month, and then checking forecasts against outcomes. The results are reported in the last three columns of Table II, and give a better picture of the accuracy of our forecasts. Obvious once stated, but the best way to assess forecast accuracy is to practice forecasting with historical test data.

While there is no directly comparable work, our accuracy at the country-month level and in six-month forecasts is similar or even slightly better than other well-known work that reports fit values. Ward, Greenhill & Bakke (2010) report in-sample AUC-ROC values of 0.761 and 0.860 for two well-known country-year models of civil war. Jay Ulfelder's coup forecasts, also country-years, achieved cross-validation AUC-ROC values between 0.78 and 0.90 (Ulfelder, 2015), while Blair, Blattman & Hartman (2017) with village-level data report model values from to 0.52 to 0.67.

Weights, theme fit, and uniqueness

With a more complete picture of how we assess model accuracy, Figure 5 again plots the model weights against

[8] The Brier score corresponds to the mean squared error, that is, the mean of the squared differences between predicted and observed values.

Table II. Ensemble and individual models of ILCs

	EBMA parameters			*Calibration*			*Test forecasts*		
Model	*W*	a_0	a_1	*Brier*	*ROC*	*PR*	*Brier*	*ROC*	*PR*
Ensemble				0.00209	0.857	0.015	0.0101	0.823	0.059
Leaders	0.22	1.55	7.95	0.00209	0.766	0.011	0.0094	0.741	0.054
Public disc.	0.08	-5.69	0.25	0.00209	0.585	0.030	0.0097	0.591	0.013
Global instab.	0.11	-0.64	5.39	0.00209	0.753	0.019	0.0095	0.802	0.031
Protest	0.14	-7.85	-1.59	0.00210	0.724	0.006	0.0096	0.629	0.025
Contagion	0.2	1.66	7.43	0.00209	0.788	0.014	0.0095	0.805	0.034
Int. conflict	0.08	-6.43	-0.24	0.00211	0.645	0.006	0.0098	0.703	0.018
Financial	0.17	0.80	6.53	0.00209	0.781	0.014	0.0095	0.800	0.037

The ensemble prediction is calculated as $p = W \times \text{logit}^{-1}(a_0 + a_1 f_k)$.

Table III. Top forecasts for ILC between August 2015 and January 2016

	Country	*EBMA*
1	India	0.598
2	Syria	0.592
3	Somalia	0.045
4	Ukraine	0.037
5	Afghanistan	0.035
6	Egypt	0.034
7	Pakistan	0.031
8	CAR	0.030
9	Lesotho	0.030
10	Burundi	0.030
11	Guinea	0.029
12	Burkina Faso	0.027
13	Nigeria	0.026
14	Thailand	0.026
15	Guinea-Bissau	0.024
16	Honduras	0.023
17	Lebanon	0.023
18	Bangladesh	0.022
19	Nepal	0.021
20	Liberia	0.021
21	Rwanda	0.020
22	Mali	0.020
23	DR Congo	0.020
24	Niger	0.020
25	Madagascar	0.019
26	Cambodia	0.018
27	Turkey	0.017
28	Philippines	0.017

Expected recall 0.5, precision 1 in 38

the three fit statistics. It is easier to see here that while lower Brier scores and higher AUC-ROC are positively related to the weights, although higher AUC-PR is not, the overall variation is larger than any relationship one might infer.

Why do better fitting models not get much stronger weights? The reason is that models which predict well sets of cases that are missed by most other models can increase the accuracy of an ensemble more than their general accuracy might indicate. This holds even if these sets of cases are relative small. Table IV provides an example, using only positive cases for simplicity, and three models which have predictions for them. The general models are more accurate in terms of Brier score than the unique model, but the unique model correctly predicts a case missed by the other two. The weights that minimize the Brier score of an ensemble prediction are $(0.375, 0.375, 0.25)$. The optimal weight for the unique model is much higher than one might otherwise expect from its individual fit. More broadly, models do not have to predict everything well to be useful, as long as they predict something well and robustly.

This uniqueness is difficult to ascertain through bivariate correlations, like those shown in Figure 3, but Figure 6 shows a plot that is an extension of the idea behind Table IV. Rows are predictions from the EBMA and theme models, columns are country-months, and the plots are separated for positive and negative cases. Colors are based on the percentile of a prediction – necessary as the predictive densities are so different. The plot illustrates for example that the Leaders theme predicts one case missed by all other models very well (the left-most column in the ILC=1 facet in plot (a)).

Visual comparisons like these are less helpful as the number of cases increases, like in the ILC-0 facets in the plots, and a more abstract way to assess a set of predictions against a pool of existing predictions might be a better solution. Colaresi & Mahmood (2017) present an approach and techniques for iterative model development that are helpful in this regard.

The same logic for usefulness should intuitively apply not just to models but to the underlying explanations as

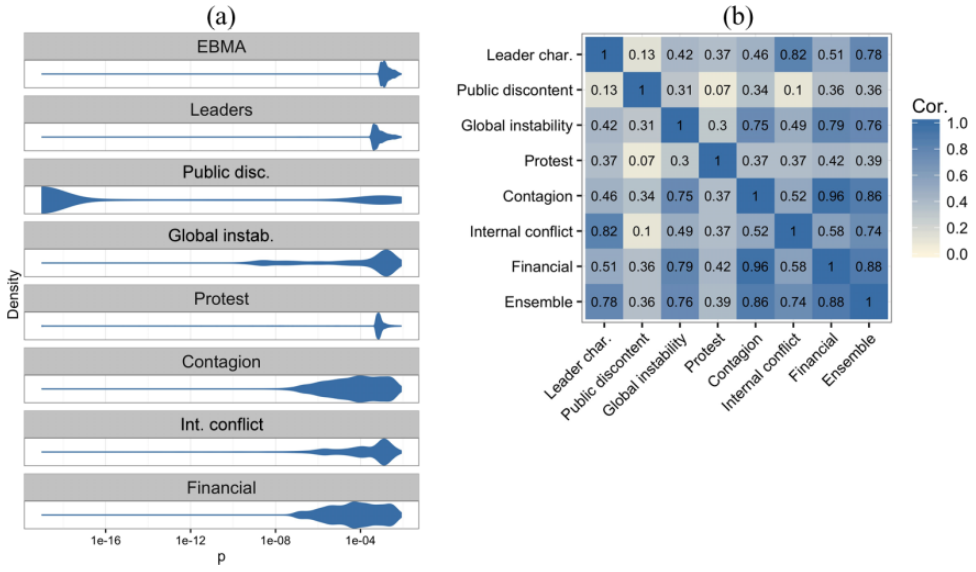

Figure 3. Densities of the theme and ensemble predictions (a) and their bivariate correlations (b)

(a) Balanced data, $\bar{y} = 0.5$

		Prediction 0	Prediction 1
Y	0	45	5
	1	25	25

(b) Sparse data, $\bar{y} \simeq 0.002$

		Prediction 0	Prediction 1
Y	0	22,500	2,500
	1	25	25

Figure 4. Classification matrices for two sets of data with different base rates but identical recall of 0.5 and false positive rate of 0.1

well. Take civil war as an example. Political grievances (e.g. Gurr, 1970), weak state capacity and other factors enabling insurgency (Fearon & Laitin, 2003), economic opportunities in rebellion through diamonds and other lootable commodities (Collier & Hoeffler, 2004), and other explanations have been proposed. One can think of illustrative cases for any of them – Tibet, Iraq, Liberia. Although the conjectures are sometimes posed as competing, it seems more plausible that each of them plays a role of varying extent in a conflict. This kind of multifaceted structure is also obvious with ILCs, which are a conscious construct of multiple potential mechanisms. Short of theoretical integration, ensembles are a way to combine such competing explanations into a single general framework that also evaluates each explanation's relative usefulness given the predictions from all others.

An ensemble might be especially preferable to a general model subsuming other models when a factor is conjectured to have opposing influences on two process that lead to the same outcome. For example, ILCs have historically occurred in part as the result of coups and successful rebellions. Counterbalancing, that is, the creation of multiple competing security organs as a way to forestall the possibility of coups has been shown to both reduce the chances of a coup, if done correctly Böhmelt & Pilster, 2015), and decrease the effectiveness of security forces in interstate war (Pilster & Böhmelt, 2011; also Belkin & Schofer, 2005 for a complementary view). It does not seem like a long stretch to hypothesize that counterbalancing will thus also impede the effectiveness of fragmented security organs to suppress domestic rebellion. If that is the case, what is the overall effect of counterbalancing on ILC? Would estimating a single model in which the competing effects average out to some value be the correct approach? Alternatively, one could estimate separate models of coups and rebellions respectively, and then use an ensemble to arbitrate.

In sum, ensembles offer the potential to evaluate competing arguments, maybe even with opposing hypotheses for the effect of a factor, on both their overall ability to predict an outcome and their ability to predict cases missed by other arguments.

Causality, basic research, and use in forecasting

It is common to draw a distinction between explanatory modeling aimed at identifying causal effects and predictive modeling like ours (e.g. Shmueli, 2010). There certainly can be trade-offs in practice, for example when we consider the range of models from random forests, lasso, and similar algorithmic approaches maximizing predictive accuracy to randomized controlled trials, regression discontinuities, and other modeling strategies like

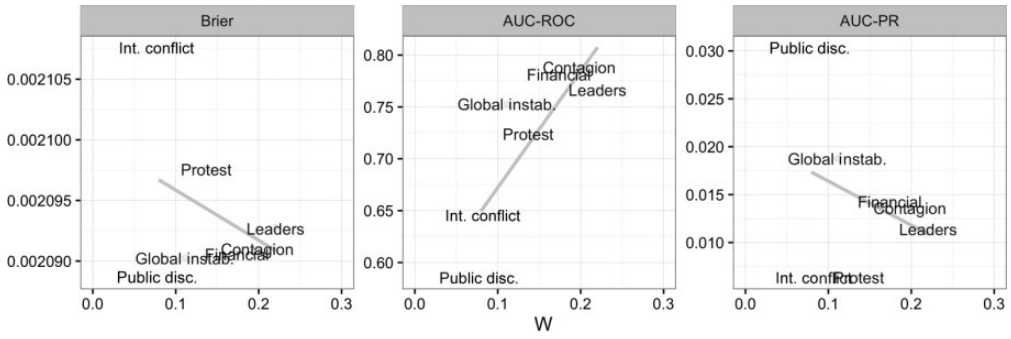

Figure 5. Model weights *W* in the ensemble forecast compared to calibration period Brier score, AUC-ROC, and AUC-PR

Table IV. Three sets of predictions for an outcome

Outcome	*1,*	*1,*	*1,*	*1,*	*1*	*Brier score*
General model 1	1,	1,	1,	0,	1	0.2
General model 2	1,	1,	1,	0,	1	0.2
Unique model	0,	0,	0,	1,	1	0.6

difference in difference and instrumental variable estimation that attempt to improve the ability to identify causal effects. Ward (2016) examines some of these issues. Basically, we believe that prediction is necessary to demonstrate the validity of causal claims, but it is not sufficient. In this exercise, we are focusing simply on a predictive model that is based on extant substantive ideas about irregular leadership changes.

Our effort is focused on prediction and we have tried to be careful about avoiding causal claims, although this does not equal evidence of their absence. Strictly speaking we can do this because prediction does not require assumptions of causality. Thus we do things like using protest counts to predict future outcomes that can be caused by mass protests. Still, predictions that come with credible causal claims are stronger. For consumers of predictions a natural second step is to question what can be done to change them, which requires knowledge of causes.

Why, then, do we not ground our predictions more strongly in extant arguments and models from the relevant literatures? This question has been a common reaction to our work from academic reviewers. We do not find a robust literature on this topic. But, more broadly, the answer is threefold: (1) data limitations, (2) the level of temporal resolution in most extant work, and (3) lack of information about the predictive performance of existing claims.

The first two are practical issues. To forecast in near-real time, we need data that have wide, and ideally global, coverage; are available for recent time periods, probably no more than a couple of months or years depending on the their measurement level; and are updated at least once a year or more. Few if any published studies use models derived from data that meet these requirements. While structural indicators like those from the World Bank have broad coverage and are fairly complete, much work relies on key indicators that are more limited in scope and possibly artisanal – coded by the authors themselves, and not necessarily maintained.

Another obstacle is that we are attempting to forecast at a scale of months, rather than the annual resolution more common. From a purely technical aspect, annual models will be limited for our prediction problem: annual data *in* means annual predictions *out*. Almost all extant models in the literature deal with at least annual aggregations. The problem is conceptually also deeper and concerns the level of temporal resolution at which current theories operate. Arguments about factors like regime types, counterbalancing and other forms of coup-proofing, youth bulges, climate change, and conflict all operate at a temporally slow-moving pace, and typically are evaluated using annual cross-national data. There is a paucity of extant work that considers what determines or at least might help anticipate the timing of an event.

The increasing availability of human- and machine-coded event data allows some finer-grained examinations to be empirically evaluated. But while they have been used extensively in spatial disaggregation to lower levels of resolution, scant work has attempted to disaggregate temporally to uncover the mechanisms behind the timing of events. Indeed, it seems that many political scientists remain uncomfortable with utilizing event data, beyond annual aggregations of conflict to augment existing structural indicators, and much of the published work has focused on biases in event data (Weidmann, 2015), especially those machine-coded from large volumes of news aggregations (for an overview of

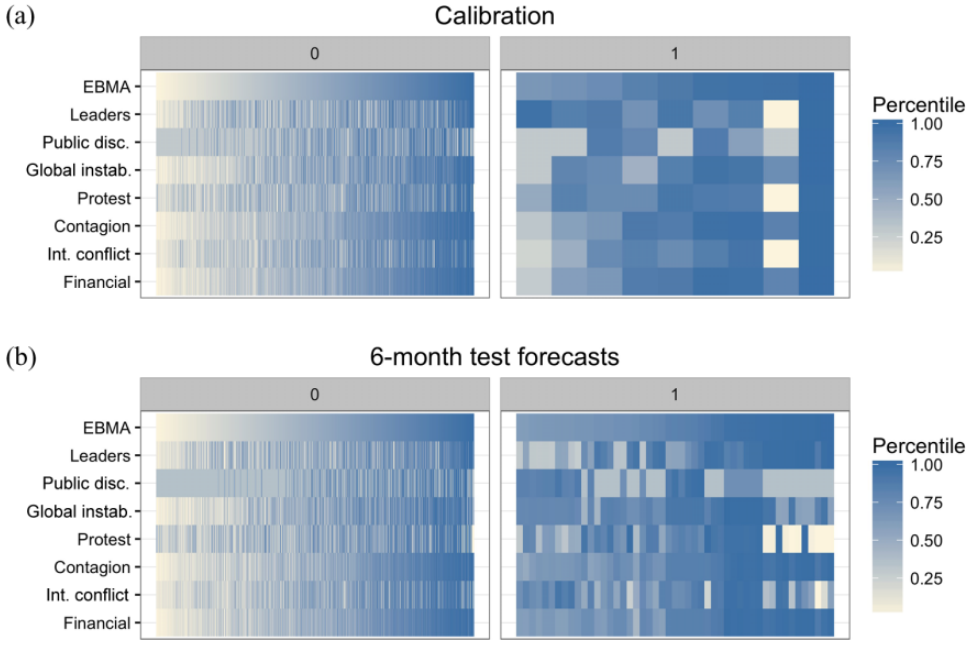

Figure 6. Comparison of model predictions for negative and positive outcomes

machine-coding specific issues, see Schrodt, 2015). These may be important issues for the estimation of causal effects (e.g. see Price & Ball, 2014), but less so for prediction. There is an increasing amount of evidence that event indicators add predictive power to purely structural models, especially at subannual temporal resolutions (e.g. see Chadefaux, 2014; Chiba & Gleditsch, 2017).

The last issue is paucity of information on predictive performance. There is a fundamental disconnect between current research practice and the evaluation needed for forecasting. The crux is that almost all relevant published research either does not report fit statistics at all, only in-sample, or only statistics like likelihood values that cannot be compared across non-nested models estimated on different data. As a result forecasters lack information on which arguments and models are worth utilizing. For a detailed view on this topic, see Tetlock (2005) and Mellers et al. (2015).

Is the lack of attention to model fit a problem for explanatory modeling? Good model fit is not needed to estimate marginal effects, unless the effects are expected to be substantial. Marginal effects can be calculated even if the overall effect is virtually zero. There is no inherent reason to expect that models and strategies like randomized trials, regression discontinuities, instrumental variables, or differences in differences will produce good absolute fit. Nobel prizes have been awarded for work with very weak fit. Unlike predictive models they do not have to be broad and comprehensive, and actually work better when they are focused on a single relationship. But they should still be assessed for their ability to predict, in their own right, if we care about placing any findings in broader context and interpreting their substantive importance. The medical literature is awash with marginal effects reports, even though many of these are very tiny overall effects. Drug X is twice as effective as drug Y in preventing some condition. However, neither of them help more than a couple individuals in 10,000. Fit and prediction are part of the same picture. Many statistical models are indifferent to fit, because they focus on marginal effects.

Out-of-sample predictions can show whether results generalize to other settings. One of the ways in which this can not occur is through overfitting, when model parameters pick up on idiosyncrasies and noise present in any particular sample of data. It is a well-known problem in prediction because it manifests itself in in-sample fit that is much better than on subsequent new data. But the common strategy for dealing with it, by partitioning data in some fashion to create some out-of-sample data, can also help identify whether generalizability is a problem for an explanatory model.

Consider a different angle: one of the common criticisms leveled by quantitative scholars against qualitative work is that it tends to be overly specific to the small

number of cases it is meant to explain and does not generalize well (e.g. see the discussion in Mahoney & Goertz, 2006). But theories and conjectures evaluated through statistical models that are not evaluated for fit at all or only in-sample, are also prone to overfitting on the data at hand, just on a larger scale.

Prediction is also important for assessing the substantive importance of an effect. Effect sizes and *p*-values themselves are not straightforward indicators of how well a factor accounts for variation in outcomes (Ward, Greenhill & Bakke, 2010). As a simple example, consider a difference of means test for two groups that is statistically significant. Effect size and confidence interval apart, it matters whether remaining individual variation overwhelms the difference in group means or not. The size of the confidence interval might give us some indication, but it also depends on sample size and other factors. Instead we could look at how much variation is accounted for by the difference in group membership.

Similarly, if the presence of oil increases the chances of ILC by 20%, but the underlying model only explains 10% of the variation in ILC outcomes between cases, what substantive conclusion are we to draw? Probably that we should look for other factors that have a stronger role – while keeping in mind that other things are still needed in order to argue for causality. The remaining variation between cases is much greater than the fraction that can be attributed to oil. One is reminded of medical studies that found that factor *x* increases the probability of cancer by 100%! – from 0.0001 to 0.0002.

Where do models and methods tailored for causal inference, like randomized trials or regression discontinuity designs, fit into this logic? Of course one cannot expect them to predict at the same scale, level, and way in which we predict ILCs. Indeed some of these methods work by transforming the dependent variable. But the models are fitted to something, and how well they and the key indicators they include can predict that something still contains information relevant for substantive interpretations of their results.

These somewhat narrow and technical considerations aside, the fundamental question for researchers interested in discovering the causes of ILC and other outcomes is this: What does a theory or model which cannot predict well actually explain or help us understand? In this sense the distinction between explanation and prediction, which Simon (2009) uses as the basis for distinguishing basic from applied research in political science, is false. As Schrodt (2014) notes, 'explanation in the absence of prediction is not scientifically superior to predictive analysis, it isn't scientific at all'.

A more common view seems to be that the obverse is true, that prediction without explanation is not scientific or valuable in understanding politics. This is wrong except in the narrow sense that it is not needed to establish causal effects. Prediction can establish boundaries for how systematic an outcome of interest is, as opposed to unobservable or intrinsically random factors, for example along the lines of Gartzke's (1999) 'war in the error term' argument. This also provides a further standard against which substantive interpretations of the importance of different explanations for a phenomenon can be evaluated. It is good to know that wealth is more important for predicting civil wars than regime types or commodity dependence (Ward, Greenhill & Bakke, 2010), but having an upper limit of civil war predictability can also tell us how worthwhile it is to look for additional explanations.

Lastly, predictive analysis can also identify fruitful avenues for further research that may have been unrecognized. If we find that something predicts very well, would it not be worthwhile to investigate why that is, that is, to explain an otherwise inductive pattern? Kepler was able to accurately describe and predict the motion of planets by 1619, even though it was not until Newton's theory of gravity and laws of motion in 1679 that a theory could account for this pattern.

Explanatory and predictive modeling require different considerations. Prediction is not a panacea, as much as explanatory modeling by itself is not either. If we harp on it, it is because common practice largely eschews prediction or model fit and does not recognize a role for it in theory-building and testing, a point explored in more depth by Ward (2016). Authors increasingly try to make statements about the substantive significance and policy relevance of their findings, and this is a positive trend, but also dangerous without considering how well the underlying models can predict.

Issues specific to forecasting

Forecasting introduces some considerations that are usually not faced in explanatory modeling. Foremost among these is what to present as 'the forecast'. Most forecasting models develop predictions of probabilities, but invariably users want these recoded into a scale they believe is intuitive, for example 'red', 'yellow', and 'green'. These, of course, destroy the underlying probability metric. Recently, some government agencies have developed overarching ways of describing probabilities. But in the end, many consumers of predictions simply want a rack and stack of countries, from high to low.

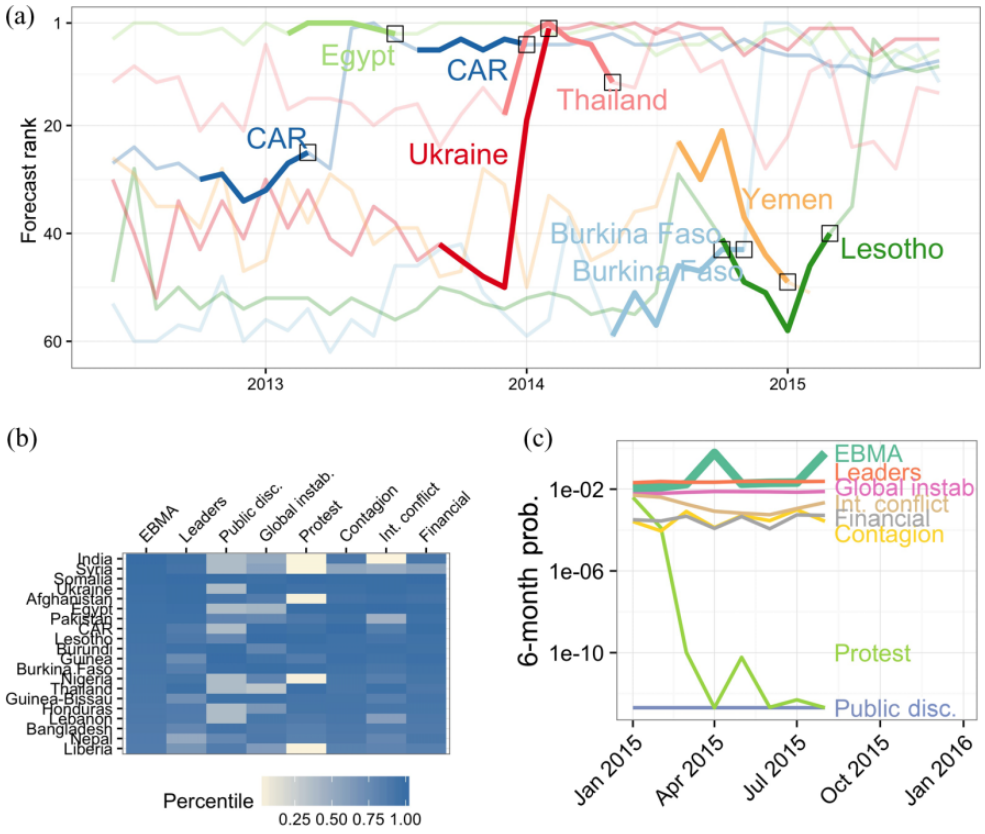

Figure 7. Example plots for presenting forecasts

(a) ranking in past forecasts for select countries over time – black squares mark ILC events in the country marked, and solid lines indicate forecasts that included that ILC; (b) table of forecasts with theme predictions; and (c) ensemble and theme predictions for Syria in the current and several past forecasts.

Kennedy (2015) discusses the same problem and recommends defining cost functions for weighing the trade-off between false positives and true positives in a forecast, and to create tables based on minimizing such a cost function. Extracting statements about relative costs is difficult in practice with consumers who are heterogenous and possibly non-technical. Indeed, you may be invited for a long walk off a short pier.

We have instead moved from arbitrary tables of the top 10 or 20 predictions to presenting tables targeted at specific values of accuracy. Based on test forecasts, we know what values of recall, precision, and accuracy to expect for specific probability thresholds, and thus we can use these to create targeted tables with known expected accuracy.

Furthermore, we have tried as much as possible to present graphics to guide interpretation. This leads 'naturally' to examine ranking in addition to raw probabilities. Figure 7 gives several examples of these plots. In the first, we present historical rankings for select countries, as a way to highlight whether high predictions are due to new or ongoing events. This can help identify likely false positives, as these generate fairly stable high rankings. The second plot is a visual equivalent of our forecast table, but with added information on the thematic predictions, all shaded by the model-specific percentile of a prediction. Plot (c) presents the information encoded in a single row of the previous table visually, by plotting the ensemble and theme predictions over multiple time periods. These are helpful, along with discussion of the specific variables that are driving (changes in) high or low predictions, for interpreting the forecasts for a specific country.

In addition, there are several more open-ended questions that could motivate future work:

(1) Foremost, forecasting would benefit from a systematic evaluation of the predictive performance of existing work. Even if research slowly is starting to evaluate model fit more carefully, there is a large number of existing arguments on which there is little information.

Ward, Greenhill & Bakke (2010) provide a template for how this can be done. Aside from assessing the absolute fit of a complete model, they also examine changes in fit associated with key variables in a model. Such examinations of the predictive power attributable to single variables are also a way to deal with focused models that consciously restrict the number of covariates without penalizing them for not resorting to the 'kitchen sink'.

Useful also would be to know how well an argument or model or variable fits gaps in existing knowledge. Our preceding discussion of uniqueness showed how they can be useful for prediction alongside other models. This may be harder to assess without a larger corpus of predictions, and thus would also be an interesting and worthwhile effort to pursue.

To build on such assessments, Colaresi & Mahmood (2017) present techniques for how to iteratively improve models based on their fit in specific cases, and Chiba & Gleditsch (2017), in a similar but applied spirit, extend an earlier civil war model by evaluating the utility of event-based variables.

(2) How should covariates be extrapolated? In out-of-sample testing, usually only the outcomes are held back but covariate values are used. In forecasting, however, future covariate values are also unknown. Practical solutions have included lagging by the number of forecast periods (Goldstone et al., 2010), carry-forward extrapolation of covariates (O'Brien, 2010), and scenario-based forecasting where researchers use multiple forecasts for long-term trends in key covariates or set them at interesting values. This is more commonly used in long-range forecasting based on simulation (e.g. Hegre et al., 2013; Hegre, Nygård & Ræder, 2017; Witmer et al., 2017).

Other approaches include univariate or multivariate time-series methods. Vector autoregression (VAR) can capture interdependencies between covariates and Brandt & Sandler (2012) extend them for count series, which is especially promising for the difficult problem of forecasting highly variable event count-based indicators (Brandt, Freeman & Schrodt, 2011). However, these approaches rely mostly on a weighted average of prior values, and may not be appropriate for rare event counts.

Scenario-based forecasting might be a useful alternative for these highly fluctuating factors. Instead of a single 'best' forecast we could evaluate multiple alternative forecasts under stated assumptions about the evolution of covariates. A benefit would be the ability to answer questions such as in which countries an uptick in protest activity would be of most concern. However, despite the broad range of possibilities for extrapolating covariates, there is little discussion and evaluation of these options for the purpose of forecasting conflict in political science.

(3) Robust out-of-sample validation with rare events and adaptive modeling. Having rare events means that for any given partition of our data, model estimates and fit statistics may vary widely depending on which positive cases are included in the data. This is problematic when developing or attempting to compare models as it decreases our confidence that any given set of comparisons is not the result of random sampling variation or substantive changes in the processes leading to an outcome of interest.

Another side effect of sparse data is that multiple out-of-sample tests have to be conducted on the same data partitions for the purpose of evaluating competing models or improving an existing model, for example using the Box method suggested by Colaresi & Mahmood (2017). Even out-of-sample test data can be overfitted through such adaptive modeling, which has shown itself in practice in Kaggle-like machine learning competitions where winning models often underperform in subsequent applications despite having been developed with out-of-sample validation and testing (Dwork et al., 2015). Although this is maybe not currently of concern to most political science and conflict research, as forecasting becomes more common it will become important to also find ways to address second-order overfitting. Despite these challenges, in our opinion the benefits of forecasting far outweigh the costs.

Conclusion

Forecasting and prediction have a wary profile in the social sciences. On the one hand, many feel that the social realm is too idiosyncratic, fluid, and complicated to be accurately predicted. On the other hand, those same individuals check the weather forecast daily, even though at one time, the same was said about the weather, another changing and non-deterministic system of many different variables. Progress has certainly been made on the weather, and there are long-term climate change predictions as well as derivative markets in the weather in Bozeman, Montana for your summer holidays in 2021. Progress in predicting social phenomena has been made too, perhaps not as dramatically as in the realm of weather. A simple point is that without some effort no progress is really feasible, and without forecasting efforts it is a self-fulfilling prophesy that prediction is not possible.

Others are worried that forecasting will become so accurate that it will be possible to change the future,

presumed to be a bad thing, or that only inherently unpredictable events will remain (see, in this issue, Chadefaux, 2017). We have a long way to go before that happens, but if we get there we might find a brighter future, one with more justice and peace, rather than a darker one. In any case, it is true that human actions now will affect the future. But that is a truism, unrelated to the ability of humans to forecast.

Another concern maybe is that our models are not accurate enough to make confident predictions. Are our forecasts accurate enough? We have generally achieved AUC-ROC values between 0.8 and 0.9 in out-of-sample tests of our forecasts, with base rates that are lower by about a factor of 10 compared to country-year civil war onset data. Due to the high resolution at which we are forecasting, the precision of our forecasts could be higher. Generally our accuracy matches reported information from other examples of conflict prediction, ranging from AUC-ROC values of between 0.82 and 0.94 for civil war onset, termination, and occurrence in models specifically tailored for good fit by Hegre et al. (2013), to 0.52–0.67, in the Blair, Blattman & Hartman (2017) models of subnational violence in Liberia.

In the end we know that the accuracy of our predictions is high enough for them to have been disseminated and used in practice. But what does that mean? Even low-level decisionmakers don't base their decisions on a single piece of information, and we can expect that consumers of our disciplinary forecasts will treat them as a piece of information, rather than the entirety of the relevant information. Isn't it better to be in the mix, than out of it?

Another concern is that forecasting will further empower the strong against the weak. Forecasting in political science certainly fills a demand for predictions within central governments. But it also is relevant for organizations that may be opposed to governments, and especially helpful to nongovernmental organizations. Consider that currently no forecasting model exists to help NGOs decide when their personnel are in danger and should be evacuated from dangerous situations. Such a model would be very useful.

What does applied work of this nature have to offer back to explanatory research? The big promise of forecasting is twofold. First, and despite its many difficulties, it holds the promise of assessing and adjudicating between the large number of arguments and conjectures that presently exist concerning many facets of political violence, and, through the ensemble approach we have presented, of integrating promising arguments and their models into a more coherent and more complete model of a phenomenon. Second, among the various threats to generalizable causal inference, forecasting provides a way in which to empirically evaluate causal arguments credibly without the possibility of dishonest research, honest mistakes, or contamination of arguments and models by knowledge of future states of the world.

Replication data

The Online appendix and replication R code and data are available at http://www.prio.org/jpr/datasets and https://github.com/andybega/jpr-forecasting-lessons.

Acknowledgements

Participants at the International Studies Association meeting in February 2015 in New Orleans, at the University of Essex in March 2015, the University of Washington CSSS Workshop on 21 April 2015, and the PRIO Forecasting Workshop 22–24 April 2015, Oslo, Norway, provided helpful comments.

Funding

The research described in this article was sponsored by the Political Instability Task Force (PITF). The PITF is funded by the Central Intelligence Agency. The views expressed in this article are the authors' alone and do not represent the views of the US Government.

References

Alesina, Alberto; Şule Özler, Nouriel Roubini & Phillip Swagel (1996) Political instability and economic growth. *Journal of Economic Growth* 1(2): 189–211.

Beger, Andreas; Cassy L Dorff & Michael D Ward (2016) Irregular leadership changes in 2014: Forecasts using ensemble, split-population duration models. *International Journal of Forecasting* 32(1): 98–111.

Beger, Andreas; Daniel W Hill Jr, Nils W Metternich, Shahryar Minhas & Michael D Ward (2016) Splitting it up: The spduration split-population duration regression package. Manuscript (https://github.com/dhill138/spduration-paper).

Belkin, Aaron & Evan Schofer (2005) Coup risk, counterbalancing, and international conflict. *Security Studies* 14(1): 140–177.

Blair, Robert A; Christopher Blattman & Alexandra Hartman (2017) Predicting local violence: Evidence from a panel survey in Liberia. *Journal of Peace Research* 54(2): 298–312.

Böhmelt, Tobias & Ulrich Pilster (2015) The impact of institutional coup proofing on coup attempts and outcomes. *International Interactions* 41(1): 158–182.

Boschee, Elizabeth; Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz & Michael D Ward (2015) ICEWS coded event data. Harvard Dataverse Network (http://dx.doi.org/10.7910/DVN/28075).

Brandt, Patrick T & Todd Sandler (2012) A Bayesian Poisson vector autoregression model. *Political Analysis* 20(3): 292–315.

Brandt, Patrick T; John R Freeman & Philip A Schrodt (2011) Real time, time series forecasting of inter- and intra-state political conflict. *Conflict Management and Peace Science* 28(1): 41–64.

Bueno de Mesquita, Bruce & Alastair Smith (forthcoming) Political succession: A model of coups, revolution, purges, and everyday politics. *Journal of Conflict Resolution*. DOI: 10.1177/0022002715603100.

Buhaug, Halvard & Kristian Skrede Gleditsch (2008) Contagion or confusion? Why conflicts cluster in space. *International Studies Quarterly* 52(2): 215–233.

Chadefaux, Thomas (2014) Earling warning signals for war in the news. *Journal of Peace Research* 51(1): 5–18.

Chadefaux, Thomas (2017) Market anticipations of conflict onsets. *Journal of Peace Research* 54(2): 313–327.

Chiba, Daina & Kristan Skrede Gleditsch (2017) The shape of things to come? Expanding the inequality and grievance model for civil war forecasts with event data. *Journal of Peace Research* 54(2): 275–297.

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

Collier, Paul & Anke Hoeffler (2004) Greed and grievance in civil war. *Oxford Economic Papers* 56(4): 563–595.

Dwork, Cynthia; Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold & Aaron Roth (2015) The reusable holdout: Preserving validity in adaptive data analysis. *Science* 349(6248): 636–638.

Eckstein, Harry & Ted Robert Gurr (1975) *Patterns of Authority: A Structural Basis for Political Inquiry*. New York: John Wiley & Sons.

Fearon, James D & David D Laitin (2003) Ethnicity, insurgency, and civil war. *American Political Science Review* 97(1): 75–90.

Gartzke, Erik (1999) War is in the error term. *International Organization* 53(3): 567–587.

Gleditsch, Kristian Skrede & Michael D Ward (1999) Interstate system membership: A revised list of the independent states since 1816. *International Interactions* 25(4): 393–413.

Goemans, Hein E; Kristian Skrede Gleditsch & Giacomo Chiozza (2009) Introducting Archigos: A dataset on political leaders. *Journal of Peace Research* 46(2): 269–283.

Goldstone, Jack A; Robert H Bates, David L Epstein, Ted Robert Gurr, Michael B Lustik, Monty G Marshall, Jay Ulfelder & Mark Woodward (2010) A global model for forecasting political instability. *American Journal of Political Science* 54(1): 190–208.

Gurr, Ted Robert (1970) *Why Men Rebel*. Princeton, NJ: Princeton University Press.

Gurr, Ted Robert (1974) Persistence and change in political systems, 1800–1971. *American Political Science Review* 74(4): 1482–1504.

Hegre, Håvard; Joakim Karlsen, Håvard Mokleiv Nygård, Håvard Strand & Henrik Urdal (2013) Predicting armed conflict, 2011–2050. *International Studies Quarterly* 57(2): 250–270.

Hegre, Håvard; Håvard Mokleiv Nygård & Ranveig Flaten Ræder (2017) Evaluating the scope and intensity of the conflict trap: A dynamic simulation approach. *Journal of Peace Research* 54(2): 243–261.

Hoff, Peter D (2007) Extending the rank likelihood for semiparametric copula estimation. *Annals of Applied Statistics* 1(1): 265–283.

Hyndman, Rob; Anne B Koehler, J Keith Ord & Ralph D Snyder (2008) *Forecasting with Exponential Smoothing: The State Space Approach*. Berlin: Springer.

Kennedy, Ryan (2015) Making useful conflict predictions: Methods for addressing skewed classes and implementing cost-sensitive learning in the study of state failure. *Journal of Peace Research* 52(5): 649–664.

Lautenschlager, Jennifer; Steve Shellman & Michael D Ward (2015) ICEWS coded event aggregations. Harvard Dataverse Network (http://dx.doi.org/10.7910/DVN/28117).

Mahoney, James & Gary Goertz (2006) A tale of two cultures: Contrasting quantitative and qualitative research. *Political Analysis* 14(3): 227–249.

Marshall, Monty G & Keith Jaggers (2015) Polity IV project: Political regime characteristics and transitions, 1800–2015 Dataset Users' Manual. (http://www.systemicpeace.org/inscr/p4manualv2015.pdf).

Mellers, Barbara; Eric Stone, Terry Murray, Angela Minster, Nick Rohrbaugh, Michael Bishop, Eva Chen, Joshua Baker, Yuan Hou, Michael Horowitz, Lyle Ungar & Philip Tetlock (2015) Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science* 10(3): 267–281.

Montgomery, Jacob M; Florian M Hollenbach & Michael D Ward (2012) Improving predictions using ensemble Bayesian model averaging. *Political Analysis* 20(3): 271–291.

O'Brien, Sean P (2010) Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International Studies Review* 12(1): 87–104.

Pilster, Ulrich & Tobias Böhmelt (2011) Coup-proofing and military effectiveness in interstate wars, 1967–99. *Conflict Management and Peace Science* 28(4): 331–350.

Powell, Jonathan M & Clayton L Thyne (2011) Global instances of coups from 1950 to 2010: A new dataset. *Journal of Peace Research* 48(2): 249–259.

Price, Megan & Patrick Ball (2014) Big data, selection bias, and the statistical patterns of mortality in conflict. *SAIS Review of International Affairs* 34(1): 9–20.

Schrodt, Philip A (2014) Seven deadly sins of contemporary quantitative political analysis. *Journal of Peace Research* 51(2): 287–300.

Schrodt, Philip A (2015) Event data in forecasting models: Where does it come from, what can it do? Manuscript (http://eventdata.parusanalytics.com/papers.dir/Schrodt.PRIO15.EventData.v1.1.pdf).

Shmueli, Galit (2010) To explain or to predict? *Statistical Science* 25(3): 289–310.

Simon, Herbert A (2009) Science seeks parsimony, not simplicity: Searching for pattern in phenomena. In: Arnold Zellner, Hugo A Keuzenkamp & Michael McAleer (eds) *Simplicity, Inference and Modelling: Keeping It Sophisticatedly Simple*. Cambridge: Cambridge University Press, 32–72.

Svolik, Milan (2008) Authoritarian reversals and democratic consolidation. *American Political Science Review* 102(2): 153–168.

Tetlock, Philip (2005) *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton, NJ: Princeton University Press.

Ulfelder, Jay (2015) Statistical assessments of coup risk for 2015. Blog post (https://dartthrowingchimp.wordpress.com/2015/01/17/statistical-assessments-of-coup-risk-for-2015/).

Ward, Michael D (2016) Can we predict politics? Toward what end? *Journal of Global Security Studies* 1(1): 80–91.

Ward, Michael D; John S Ahlquist & Arturas Rozenas (2012) Gravity's rainbow: A dynamic latent space model for the world trade network. *Network Science* 1(1): 95–118.

Ward, Michael D; Brian D Greenhill & Kristin M Bakke (2010) The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research* 47(4): 363–375.

Weidmann, Nils B (2015) A closer look at reporting bias in conflict event data. *American Journal of Political Science* 60(1): 206–218.

Wimmer, Andreas; Lars-Erik Cederman & Brian Min (2009) Ethnic politics and armed conflict: A configurational analysis of a new global data set. *American Sociological Review* 74(2): 316–337.

Witmer, Frank DW; Andrew M Linke, John O'Loughlin, Andrew Gettelman & Arlene Laing (2017) Subnational violent conflict forecasts for sub-Saharan Africa, 2015–65, using climate-sensitive models. *Journal of Peace Research* 54(2): 175–192.

World Bank Group (2013) World Development Indicators 2013 (http://data.worldbank.org/products/wdi).

MICHAEL D WARD, b. 1948, PhD in Political Science (Northwestern University, 1976); Professor, Duke University (2009–).

ANDREAS BEGER, b. 1983, PhD in Political Science (Florida State University, 2012); Data scientist, Ward Associates (2016–).