British Journal of Political Science (2020), **50**, 1405–1417
doi:10.1017/S0007123418000443

British Journal of Political Science

ARTICLE

The Future is a Moving Target: Predicting Political Instability

Drew Bowlsby[1], Erica Chenoweth[2], Cullen Hendrix[1] and Jonathan D. Moyer[1*]

[1]Josef Korbel School of International Studies, University of Denver and [2]John F. Kennedy School of Government, Harvard University
*Corresponding author. Email: jmoyer@du.edu

(Received 5 December 2017; revised 24 August 2018; accepted 17 September 2018 ; First published online 20 February 2019)

Abstract

Previous research by Goldstone et al. (2010) generated a highly accurate predictive model of state-level political instability. Notably, this model identifies political institutions – and partial democracy with factionalism, specifically – as the most compelling factors explaining when and where instability events are likely to occur. This article reassesses the model's explanatory power and makes three related points: (1) the model's predictive power varies substantially over time; (2) its predictive power peaked in the period used for out-of-sample validation (1995–2004) in the original study and (3) the model performs relatively poorly in the more recent period. The authors find that this decline is not simply due to the Arab Uprisings, instability events that occurred in autocracies. Similar issues are found with attempts to predict nonviolent uprisings (Chenoweth and Ulfelder 2017) and armed conflict onset and continuation (Hegre et al. 2013). These results inform two conclusions: (1) the drivers of instability are not constant over time and (2) care must be exercised in interpreting prediction exercises as evidence in favor or dispositive of theoretical mechanisms.

Keywords political instability; prediction; quantitative modeling; state failure

Episodes of acute state-level political instability – a catch-all category encompassing civil wars, democratic reversals, genocides and politicides, and state collapse – are comparatively rare but exceptionally important. Aside from their manifest and massive human costs, civil wars and other episodes of abrupt socio-political change often entail massive economic losses, to both unstable countries and their immediate neighbors, which are rarely offset by post-instability growth (Cerra and Saxena 2008; Collier and Hoeffler 2004; Gates et al. 2012; Ghobarah, Huth and Russett 2004). Moreover, episodes of political instability frequently necessitate multilateral security responses, such as interventions in Somalia in 1991 and Libya in 2011, and nearly always require large-scale humanitarian responses.

For these reasons, there is significant interest in predicting the onset of political instability. Goldstone et al. (2010, henceforth GEA) use an exceptionally parsimonious model to accurately predict 85 per cent of instability onsets in out-of-sample testing in the 1995–2004 period. The GEA model focuses on the importance of political institutions, rather than economic conditions or social factors, as the primary determinants of political instability.[1] Whereas prior research had identified many structural and slowly changing covariates of instability onset – income per

[1]GEA model political institutions by categorically representing governance regime type (autocratic, partially autocratic, partially democratic, democratic) and by modifying two of these categories (partially autocratic and partially democratic) to measure elite factionalism. Vreeland (2008) criticized the inclusion of factionalism as a driver of political instability because violence is partly definitional to factionalism, making the understanding of causality problematic.

capita, mountainous terrain, population size, age structure and resource endowments – the GEA model instead focused on a five-category measure of regime type along with a set of three additional variables: infant mortality (normalized), armed conflict in bordering states and state-led discrimination.[2] This regime-centric model emphasized the institutional nature of the regime, which GEA assert is more dynamic and therefore amenable to policy reform. Thus the GEA model has been interpreted as evidence that understanding domestic political institutions, low development, neighborhood effects and state-led discrimination are critical to understanding political instability.

The model is also notable for its exceptional accessibility and reproducibility. With its reliance on just five widely available indicators, the model allows researchers to update predictions on an annual basis, providing the model with a wide range of practical applications for academics and practitioners seeking to anticipate major political instability events and geopolitical crises. Thus their results have been taken as evidence that domestic political institutions are the most important determinant of political stability. Moreover, because time is not treated explicitly within this predictive framework, there is an implicit assumption that the drivers of political instability are time invariant within the study period.

We revisit and extend the GEA model in light of a decade of accumulated post-2004 evidence, and find the predictive power of the model has waned substantially since the original period of 1995–2004. We demonstrate that (1) the model's predictive power varies substantially over time; (2) its predictive power peaked in the period used for out-of-sample validation (1995–2004) in the original study and (3) the model performs relatively poorly in the more recent period. Moreover, this decline in performance is not just a function of the Arab Uprisings, which occurred in comparatively developed autocracies.

These issues – substantial variation in predictive power over time and failure to accurately predict recent episodes of instability – are not unique to the GEA model. We further explore these results by replicating[3] two other forecasting studies that predicted instability using quantitative methods: Chenoweth and Ulfelder (2017, henceforth CU) and Hegre et al. (2013, henceforth HEA). CU assesses the structural drivers of nonviolent campaigns. HEA uses a multinomial logit model to predict the onset of different levels of violent conflict to 2050. In both studies, we confirm significant temporal variation in the models' predictive power.

The extension of GEA, and additional replications, inform two conclusions. First, the drivers of instability vary over time. This has implications for the growing field of scholarship that relies on out-of-sample prediction to test theories (Cederman and Weidmann, 2017; Chenoweth and Ulfelder 2017; Ward, Greenhill and Bakke 2010). Our results suggest that we must be careful in interpreting the results of prediction exercises as evidence for or against time-invariant, unconditional, theoretical relationships.

Secondly, this finding has implications for policy makers, especially those who engage in data-driven attempts to predict instability. To the extent that their mental models – derived from empirical models like GEA – are fixed around a particular set of drivers, their predictions may be wrong. There is no holy grail in predicting political instability; the tools and predictors will necessarily evolve over time.

The remainder of this article proceeds as follows. The next section presents the GEA model in detail and demonstrates the variation in its predictive power over time, as well as the (in) sensitivity of its recent decline in predictive power to the inclusion/exclusion of Arab Uprising cases. We then describe the replication of CU and HEA. The following section concludes by discussing the implications of our analysis for scholars using predictive accuracy as evidence of

[2]For a review, see Hegre and Sambanis (2006).

[3]We replicate but do not extend these findings, since data for updating their analyses are not publicly available.

https://doi.org/10.1017/S0007123418000443 Published online by Cambridge University Press

given theoretical mechanisms, and for policy makers and applied modelers using prediction to anticipate major global events.

GEA Revisited and Extended

GEA used case-controlled conditional logistic regression to fit a model to the Political Instability Task Force problem set. This aggregate measure of political instability defines an event as adverse regime change, revolutionary war, ethnic war and/or genocide/politicides that resulted in at least 1,000 conflict deaths (Goldstone et al. 2010, 191–2).[4] The original study used data from 1955–2004, though these data have been updated for the period 2005–2014.[5]

To select cases, GEA randomly matched one country-year onset case with three country-year cases that (1) were stable at the period of time in which the onset case failed and (2) shared a geographic region. This approach is similar to a region-year and time fixed-effects model in that it explores within-group effects of one onset country and three non-onset countries in the same year and region (Goldstone et al. 2010, 194). They applied this technique to three random samplings of non-onset cases for sensitivity analysis and advanced independent variables by two years to make the model predictive.[6]

Time is treated only indirectly in GEA. Temporal effects are included in two ways with limited overall impact. First, the case control method considers time in its groupings. Secondly, infant mortality is normalized to annual global levels that have reduced over time.

GEA fit data from 1955–94 to their onset cases (in-sample model fitting). They then tested out-of-sample model performance from 1995 to 2004 using the following steps. First, they predicted the probability of conflict for all countries in a given year. Next, they folded that year into the full test sample and recalibrated the model. They then proceeded to the subsequent year and repeated the process. For example, they made out-of-sample predictive values for 1995 using a model calibrated in 1955–1994, and out-of-sample predictions for 2004 using a model calibrated in 1955–2003. The GEA model performed well in iterative out-of-sample testing, correctly classifying 85.7 per cent of political instability onset countries within the top quintile of predicted country-year values for the period 1995–2004.[7]

[4]"Revolutionary wars are episodes of violent conflict between governments and politically organized groups (political challengers) that seek to overthrow the central government, to replace its leaders, or to seize power in one region' (Marshall, Gurr and Harff 2016, 5). 'Ethnic wars are episodes of violent conflict between governments and national, ethnic, religious, or other communal minorities (ethnic challengers) in which the challengers seek major changes in their status' (Marshall, Gurr and Harff 2016, 6). 'Adverse Regime Changes are defined by the Political Instability Task Force as major, adverse shifts in patterns of governance, including major and abrupt shifts away from more open, electoral systems to more closed, authoritarian systems; revolutionary changes in political elites and the mode of governance; contested dissolution of federated states or secession of a substantial area of a state by extrajudicial means; and or near-total collapse of central state authority and the ability to govern' (Marshall, Gurr and Harff 2016, 10). 'Genocide and politicide events involve the promotion, execution, and/or implied consent of sustained policies by governing elites or their agents – or in the case of civil war, either of the contending authorities – that result in the deaths of a substantial portion of a communal group or politicized non-communal group' (Marshall, Gurr and Harff. 2016, 14).

[5]The GEA model used was trained on data through 2003 but extended the data to 2004 for out-of-sample testing. For this article we report 'new data' findings from 2005–2014.

[6]For example, when fitting the model to the onset of political instability in Libya in 2010, Goldstone et al. (2010) used independent variables measuring development in 2008.

[7]The authors predicted failure rates for all country-years in out-of-sample testing. These were then ranked from most unstable to least. The number of correct onset cases captured in the top quintile was then divided by the total number of onset cases in the out-of-sample period. Within the top quintile of predicted country-year values (in the original out-of-sample period from 1995–2004), the original model captured 85 per cent of onset cases. In this period there were twenty-one cases of political instability. The model correctly classified 18 of these in the top quintile of predicted values (18/21 = 0.857). If 190 countries were tested over a ten-year period, and if the rate of correct classification were the same, this process would generate 362 false positive predictions.

https://doi.org/10.1017/S0007123418000443 Published online by Cambridge University Press

Variables

The model includes (1) four categorical measures of regime type: democracy, partial democracy without factionalism, partial democracy with factionalism and partial autocracy, (2) the level of infant mortality relative to a global annual average, (3) the presence of state-led discrimination and (4) an indicator measuring whether four or more neighboring states were experiencing civil war.[8] Their key finding was that partially democratic states with factionalized party systems – where political demands are highly parochial in nature and bargaining positions between factions or parties are intransigent – are particularly vulnerable to political instability. This led the authors to conclude '[…] political institutions, properly specified, and not economic conditions, demography, or geography, are the most important predictors of the onset of political instability' (Goldstone et al. 2010, 190).

Results and Extension

To assess predictive power over time, we use an alternative metric:[9] the area under the receiver operator curve (AUROC). The AUROC is equal to the probability that the model will rank a randomly chosen positive instance (instability onset) higher than a randomly chosen negative one (no onset) according to the predicted probability of a positive outcome (Fawcett 2006, 863). AUROC values can range from 1 (perfect predictive power) to 0.5 (random predictive power – no better than a coin flip). Generally speaking, AUROC values lower than 0.8 are considered inaccurate.[10]

Figure 1 shows the GEA model's AUROC scores on a rolling decade basis for the period 1956–2014. The figure includes the full sample (the same as estimated by GEA), a sample excluding the Arab Uprising onsets (Egypt, Libya and Syria 2011), and a sample excluding the Middle East and North Africa from the analysis for the entire period.

Three findings emerge. First, the model's performance peaked during the period used for out-of-sample validation (1995–2004). This finding is unsurprising, because the published model was always going to be the one that performed best at the specific task of out-of-sample prediction for a given period. However, this finding is nevertheless important because it illustrates how the specific out-of-sample period selected can generate confidence about a model's performance.

Secondly, the GEA model's performance in earlier periods varies substantially. From 1973–82, the model performed nearly as well as during the out-of-sample validation period. For the period 1962–1971, however, the model performs nearly as poorly as in the most recent period.

Thirdly, in the more recent period (2005–2014), the model's performance has declined precipitously. Instead of correctly classifying 85.7 per cent of onset cases, the model predicts only 35.3 per cent of instability onsets.[11] There are three potential reasons for this decline: (1) measurement error in either the dependent or independent variable, (2) time trends in slope coefficients (that is, time-varying effects in the independent variables) or (3) simple model uncertainty due to misspecification. Of these three, the least likely would be measurement error: the GEA model's variables have been coded and curated by the same research team since the project's inception and are validated and rechecked often by Political Instability Task Force

[8]Factionalism is defined as political competition that is 'intense, hostile, and frequently violent' (Gurr 1970, 12). For a critique of the inclusion of factionalism in a theory-testing model of political instability, see Vreeland (2008). They did not include full autocracy in the published model. Regime type odds ratios for regime type independent variables should be interpreted as relative to autocratic states.

[9]Appendix A compares AUROC and out-of-sample accuracy over time.

[10]Tape, Thomas G., 'Interpreting Diagnostic Tests: the area under an ROC curves', http://gim.unmc.edu/dxtests/ROC3.htm.

[11]A result we replicated. See footnote eleven for a description of this evaluative method. We conducted this test using onsets from 2005–2014. The GEA main model uses data from 1955–2003 but also out-of-sample testing in 1995–2004. If we included 2004 in the more recent time window, the GEA model would correctly classify 24 per cent of onset countries in the top quintile of results.

https://doi.org/10.1017/S0007123418000443 Published online by Cambridge University Press

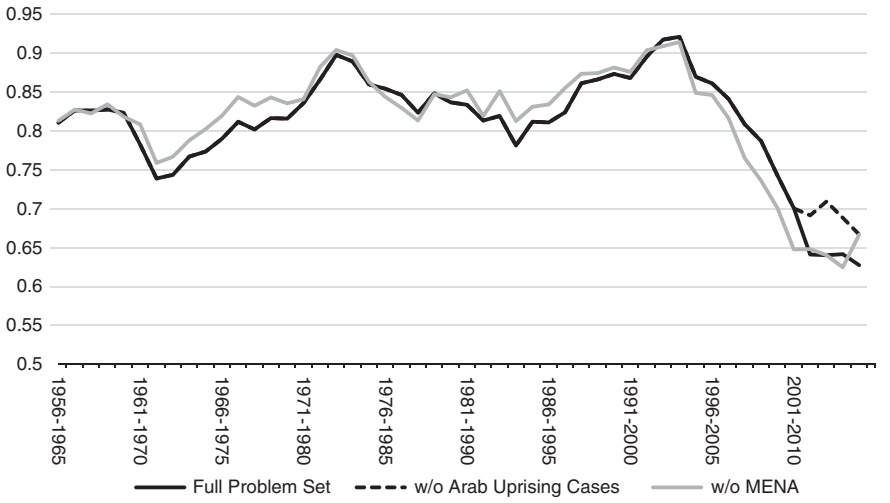

Figure 1. Moving-decade AUROC scores for GEA
Note: scores reported for full GEA problem set, excluding the recent Arab Uprising cases, and excluding the Middle East and North Africa Region.

Table 1. PITF full problem set cases 2005–14, correctly and incorrectly classified[12]

True positives (correctly classified)	False negatives (incorrectly classified)
Bangladesh 2007 (Abrupt Regime Change)	China 2009 (Ethnic War)
Chad 2005 (Revolutionary War)	Egypt 2011 (Revolutionary War)
Haiti 2010 (Abrupt Regime Change)	Fiji 2006 (Abrupt Regime Change)
Ukraine 2014 (Complex)	Guinea-Bissau 2012 (Abrupt Regime Change)
Nigeria 2006 (Complex)	Madagascar 2009 (Abrupt Regime Change)
Ethiopia 2007 (Ethnic War)	Mali 2012 (Complex)
	Mauritania 2008 (Abrupt Regime Change)
	Mexico 2007 (Revolutionary War)
	Niger 2009 (Abrupt Regime Change)
	Syria 2011 (Ethnic War)
	Libya 2011 (Complex)

(PITF)-affiliated researchers. We return to the second possible explanation for the decline – that drivers of instability have time-bounded effects – in the discussion.

Table 1 classifies onset cases (between 2005 and 2014) by their placement in the top quintile of GEA predicted values. Many of the false negatives were very significant events: Egypt (2011), Syria (2011) and Libya (2011). The latter two cases have led to ongoing, large-scale civil wars involving major power intervention. This decline remains evident, though slightly less precipitous, when we exclude the Arab Uprising cases. Overall, excluding the Middle East and North Africa does not change the predictive accuracy of the model a great deal, though the model excluding those cases outperforms the full GEA model in earlier periods and underperforms in the more recent period. Including the Middle East and North Africa actually improves model fit in the most recent period.[13]

[12]'Complex' is a mixed category used by PITF to identify onset cases across multiple categories. South Sudan (2011) is included in the problem set but not this analysis because of data limitations.

[13]Middle East and North Africa cases: Afghanistan, Algeria, Azerbaijan, Bahrain, Egypt, Georgia, Iran, Iraq, Israel, Jordan, Kazakhstan, Kuwait, Kyrgyzstan, Lebanon, Libya, Morocco, Oman, Qatar, Saudi Arabia, Syria, Tajikistan, Tunisia, Turkey, Turkmenistan, UAE, Uzbekistan and Yemen.

https://doi.org/10.1017/S0007123418000443 Published online by Cambridge University Press

While the prominence of the GEA model makes it a natural candidate for replication and investigation, these issues are not unique to its modeling strategy. We now turn to a replication of the CU and HEA.

Replicating Additional Models that Predict Rare Events

To determine the generalizability of the results identified above, we replicated two studies that use quantitative models to predict rare events: (1) the onset of maximalist nonviolent campaigns, which seek to remove the incumbent government, liberate the nation from colonial control or secede and (2) a 'tiered' armed conflict outcome, ranging from no conflict to minor and then major conflict. We address each model in turn (see Appendix A for additional detail).

Chenoweth and Ulfelder

CU derive their dependent variable from the Major Episodes of Contention (MEC) dataset (Chenoweth 2015) in a study designed to explore competing structural explanations for the onsets of maximalist nonviolent campaigns. Instead of relying on traditional measures of statistical significance (the approached used in GEA), these authors use two forms of model validation to compare the explanatory and predictive accuracy of distinct conceptual approaches for the onset of mass uprisings: k-fold cross-validation and out-of-sample prediction. The research identifies a final 'culled' model that best predicts the onset of maximalist nonviolent protest. The dependent variable used in this study is conceptually related to GEA: it measures large-scale political instability and is a rare event, with onsets occurring in only 2 per cent of eligible country-years from 1955–2013. Although CU limit their study to explaining the onset of mass uprisings, the theoretical approaches they explore – resource mobilization, political opportunity, grievance and modernization – and the culled approach that combines the highest-performing covariates from each of these approaches are relevant to political instability events beyond just mass nonviolent uprisings.

Thus we replicated their culled model using a pooled approach, similar to our approach above, to calculate a moving-decade AUROC score and assess out-of-sample predictive accuracy.

Hegre et al.

HEA used a dynamic multinomial logit model to estimate a transition probability matrix across levels of conflict using PRIO Peace Research Institute Oslo (PRIO) armed conflict data (Themnér and Wallensteen 2012) measuring three levels of conflict: no conflict, minor conflict (twenty-five or more battle deaths per year) and major conflict (1,000+ battle deaths per year). HEA use model simulations to predict country-level conflict to 2050 with exogenously specified drivers. The dependent variable in this study is similar to the PITF data used in GEA, though it does not capture abrupt regime changes that do not involve significant violence.

The authors sample a wide selection of independent variables using a split-sample design to estimate multiple model specifications for the period 1970–2001. They use these alternative models to predict out-of-sample onset from 2001–09 to determine model accuracy.[14] These models are then projected to 2050 using exogenous forecasts of variables like population and oil wealth.

Compared to GEA, HEA is more technically involved, uses a tiered dependent variable, and forecasts both onset and conflict continuation. These differences make an 'apples-to-apples' comparison challenging. To approximate such a comparison, we recoded HEA's onset variable dichotomously, collapsing major and minor conflict together and restricting the outcome of interest to onsets instead of onsets and ongoing conflict. We examined only the core predictors

[14]Notably, this period corresponds to that during which the GEA model's predictive power declined significantly.

https://doi.org/10.1017/S0007123418000443 Published online by Cambridge University Press

(oil, ethnic dominance, infant mortality, population, education and regions). We again used a pooled approach to determine moving-decade AUROC scores. This greatly simplifies the HEA approach and will impact their model's predictive accuracy but make a direct comparison across models feasible.

Results

Figure 2 compares the moving AUROC scores of the replication models for all three studies analyzed here. CU demonstrates similar patterns of predictive accuracy compared with GEA, declining to the decade 1983–92, increasing in predictive accuracy to 1989–98, and then declining again to the present. HEA, however, shows a very different pattern of temporal variation, experiencing a nadir in predictive accuracy in 1981–1990 and then increasing through the period in which both GEA and CU begin to decline in predictive accuracy.

Each model in Figure 2 uses a distinct dependent variable. To understand whether the dependent variable was driving these patterns, we replicated GEA, CU and HEA using PITF, MEC and PRIO onset data.[15]

We find that the dependent variable does determine the pattern of temporal variation in the predictability of conflict, but in distinct ways. For example, each model run using PRIO data follows a similar pattern of temporal variation in predictive accuracy, with the least predictive period occurring in the 1980s and then improvement in predictive accuracy to the present in all models except CU, which declines starting in 1991–2000. Using MEC data, model behaviors diverge significantly in the period 1990–99, with CU predicting the most accurately and GEA the least accurately. The CU model then begins to lose accuracy, converging to HEA and GEA in more recent periods. Using PITF onset data, GEA and CU predictive accuracy moves in opposite directions: CU is more accurate when GEA is low (in 1984–93 and 2004–2013) and GEA is more accurate when CU is low (1973–82 and 1994–2003).

Why Does This Matter?

The waning predictive power of these forecasting efforts over time – and the temporal variation shown across model specifications – is important for three reasons: (1) its implications for how we think about replication, (2) the use of prediction to test theories and (3) how it affects the policy community, an under-discussed but nevertheless important consumer and funder of prediction efforts.

Prediction and Replication

This exercise complicates discussions of what constitutes replicable social science. Standards for replication in quantitative studies range from describing the analysis in sufficient detail to facilitate replication – tracing the steps and processes of the original researchers – to posting code in order to reproduce statistical analysis to complex files that build working datasets from their constituent publicly available datasets. According to this definition, replication is an emergent norm in the social sciences (Dafoe 2014; Nosek et al. 2015).

However, replication may also refer to the ability to recover a finding out of sample using similar techniques, a definition that has become common among scholars using randomized controlled trials to study development (for example, determining whether an intervention that worked in Malawi will work in India) (Banerjee and Duflo 2009). Although some scholars have used out-of-sample predictive modeling to evaluate theoretical mechanisms, it is not yet common in political science or international relations to use out-of-sample predictive

[15]See Appendix B for moving-decade AUROC scores for each of these tests.

https://doi.org/10.1017/S0007123418000443 Published online by Cambridge University Press

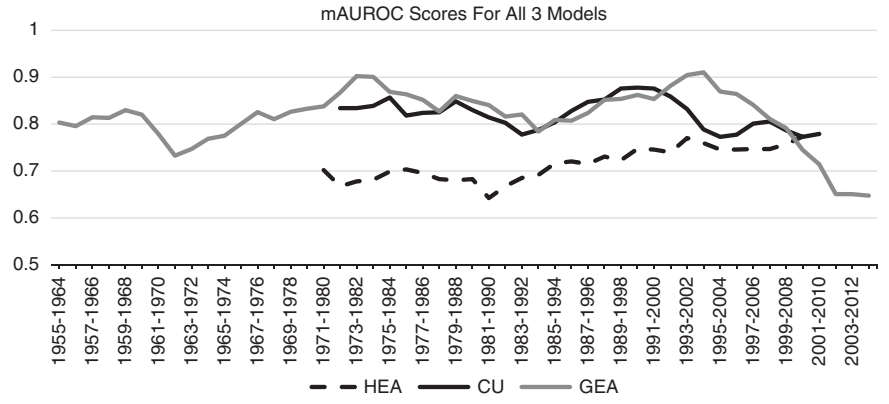

Figure 2. Moving-decade AUROC scores for HEA, CU and GEA studies predicting political instability and conflict

exercises to validate and replicate core substantive findings (Evanschitzky and Armstrong 2010).[16]

While we were able to replicate the GEA model by following their described methods using data from the same (or similar) sources, we could not replicate the core findings when applied to a different time period (2005–14). Instead, we find the model provides inaccurate out-of-sample predictions of political instability, and that the predictions do not provide evidence of the centrality of political institutions for explaining outcomes. The same was true to varying extents of our replications of CU and HEA.

Adopting the latter approach to replication – that studies replicate if their findings hold in out-of-sample analysis – would impose more stringent requirements on researchers, which would require them to develop automated data pipelines for updating variables and code. While pursuing more standardized replication practices would improve our ability to track research over time (and would be one way to better understand how temporal patterns change), it could also be prohibitively costly. Gathering, standardizing and updating data can be extremely expensive, and could privilege larger teams of researchers over individuals.

Prediction and Theory Testing

These results also highlight the complexity of using predictions as evidence for or against theoretical mechanisms. Over the past decade there has been more discussion of predictive power, rather than statistical significance, as the criterion for assessing theoretical models of political instability (Cederman and Weidmann 2017; Chenoweth and Ulfelder 2017; Ward, Greenhill and Bakke 2010). While theirs is not a theory-testing exercise, GEA note that '[…] confidence in an explanation is clearly influenced by the degree to which future events conform to the expectations that it engenders', and conclude that '[…] political institutions, properly specified, and not economic conditions, demography, or geography, are the most important predictors of the onset of political instability' (Goldstone et al. 2010, 190). If predictions, rather than statistical significance, are to inform our understanding of conflict dynamics, there may be tension between the goal of accurately predicting outcomes within a given temporal window and understanding more enduring relationships.

Table 2 presents logistic regression coefficients for the GEA model run on (1) the period for which out-of-sample validation was used in GEA 2010 (1995–2004), (2) the more recent decade (2005–2014) and (3) the full period for which data are available (1956–2014). Based on these

[16]K-fold cross-validation is conceptually similar but based on dividing within-sample observations into subsamples and analyzing them. See Cedermann and Weidmann 2017; Chenoweth and Ulfelder 2017; Ward, Greenhill and Bakke 2010.

https://doi.org/10.1017/S0007123418000443 Published online by Cambridge University Press

Table 2. Logistic regression estimates of GEA

DV (full problem set)	1995–2004	2005–2014	1956–2014
Regime Types			
Full Democracy			−0.174
			(0.778)
Partial Autocracy	1.462	0.006	0.835**
	(0.237)	(0.994)	(0.008)
Partial Democracy w/Factionalism	3.503***	0.335	1.866***
	(0.001)	(0.694)	(0.000)
Partial Democracy w/o Factionalism	1.677	0.419	0.698*
	(0.160)	(0.572)	(0.025)
Transitional			1.328**
			(0.008)
Infant Mortality (logged, normalized)	1.207	0.621	0.762***
	(0.059)	(0.284)	(0.000)
Discrimination	0.340	0.654	0.599**
	(0.553)	(0.231)	(0.003)
4+ Border Conflict	0.681		1.641***
	(0.561)		(0.000)
Regional Dummies			
E. Asia/Pacific	0.481	0.251	0.250
	(0.556)	(0.754)	(0.419)
Europe/N. Americas	1.396	−0.043	0.464
	(0.233)	(0.977)	(0.341)
Latin America	−1.097	−0.060	−0.019
	(0.371)	(0.952)	(0.953)
Middle East/N. Africa	0.549	0.334	0.505
	(0.601)	(0.729)	(0.131)
Constant	−7.064***	−4.906***	−5.272***
	(0.000)	(0.000)	(0.000)
N	1004	1026	6169
Chi-squared	41.36	4.72	130.98
Prob > chi-squared	0.000	0.858	0.000

Note: p-values in parentheses. Missing coefficient estimates indicate variable was dropped due to perfect collinearity with outcome variable.
*p < 0.05, **p < 0.01, ***p < 0.001

results, a model fit to the period 2005–2014 (Model 2) would *not* support the conclusion that political institutions are central to understanding political instability, even if such a relationship emerges in both the earlier period and the full sample (Models 1 and 3).

Indeed, for the period 2005–14, none of the core variables in the GEA model are statistically significant. If GEA had undertaken their analysis a decade later and thus developed a model to maximize predictive power for the period 2005–14, it is possible that political institutions would not have been included in the model at all. Many subsequent studies that cite the GEA model but that do not develop predictive models – which is over half of the 662 studies that cite it[17] – would likely have placed considerably less emphasis on the role of political institutions, even though they emerge as incredibly important when analyzing the full sample (1956–2014). Counterfactual reasoning of this type is of course difficult, but highly influential articles – by definition – shape subsequent discourse to a remarkable degree.[18] While predictive models can be used to test theories and can indicate which variables are most important for understanding outcomes (Hegre et al. 2017), doing so makes the specification of the out-of-sample period – that is, the period to be predicted – extremely consequential.

[17]Based on a Google Scholar search conducted on 16 August 2018. The specific search was of articles citing GEA 2010 but excluding those that discuss forecasting and/or prediction; 382 such studies were identified, accounting for 57 per cent of the total.

[18]GEA 2010 is the second-most cited article to appear in the *American Journal of Political Science* since 2010, following only Honaker and King (2010).

https://doi.org/10.1017/S0007123418000443 Published online by Cambridge University Press

Table 2 shows that model fit varies significantly over time in GEA (and impacts their substantive theoretical findings), and prediction validates that this finding is important. But model fit and prediction perform different roles vis-à-vis evaluating modeling exercises. The former can evaluate the relative strength of competing claims within a temporally bounded research exercise and allow us to assess the precision with which we have estimated comparative statics: how much the probability of a particular outcome changes/would change given a change in an explanatory variable. The latter can determine whether a model is supported 'in the real world', through out-of-sample testing. We explore this concern next.

Prediction and Policy Audiences

Finally, the decline in the predictive power of the GEA model, temporal variation shown with other model specifications, and our limited ability to accurately predict out of sample in meaningful ways has implications for policy audiences, both narrowly and broadly defined. Narrowly speaking, the GEA model was first developed with funding from the PITF, which is funded by the US intelligence community. Analysts and policy makers have relied on this and other similar models, as both prediction tools and heuristic models for thinking about political instability. Since the model's performance has declined significantly in the past decade, those using the model should update their empirical and mental models accordingly. Of course, this is already being done, including by researchers affiliated with the PITF (Chiba and Gleditsch 2017; Nyseth Brehm 2017; Ward and Beger 2017). But many of these advances have been motivated by a desire to probe the validity of the theoretical model of instability that underpins the GEA model or interrogate specific types of instability in further detail, rather than the simple fact that the model appears to have stopped working – at least for the most recent decade. More broadly, since the model's power has risen and fallen in the past, we should expect it to continue to do so in the future, necessitating continual revision and validation over time. Prediction of rare events in social systems should be understood through the perspective of the agent-structure problem (Wendt 1987). Structural explanations can provide important contextual clues, but agency remains a key focal point for predicting outcomes (Chenoweth and Ulfelder 2017).

Treating Temporal Variation in the Drivers of Political Instability

The problem of temporal variability in forecasting models is not unique to the study of political instability. Econometricians, for example, have developed advanced techniques for identifying 'structural breaks' – shifts in time series data that can lead to large forecasting errors – and improving out-of-sample forecasts in their presence (Giacomini and Rossi 2009; Pesaran, Pettezuno and Timmermann 2006; Stock and Watson 1996). While these techniques have helped improve out-of-sample macroeconomic forecasting, they cannot be easily applied to predicting the onset of political instability.

These advances have come in the context of trying to forecast comparatively better-understood outcomes, like treasury bond yields and inflation rates, which are affected by factors like real GDP growth, unemployment and industrial output – factors that have been well known for decades (Friedman 1968; Litterman and Scheinkman 1991; Lucas 1973). Moreover, these forecasting efforts are designed to predict a single phenomenon. Lastly, these efforts are based on comparatively high-frequency (monthly) time-series data that are highly autoregressive: recent past values are highly predictive of current values (Stock and Watson 1996).

By contrast, forecasting the onset of political instability poses a different set of challenges. First, it requires researchers to confront a high degree of fundamental uncertainty about the drivers of the outcome. For example, though the output–inflation relationship (commonly known as the Phillips Curve) continues to be contested and refined, conflict scholars have no similarly robust relationship they can point to that flows from straightforward quasi-deterministic modeling. Secondly, political instability, at least as it is treated in the field and

https://doi.org/10.1017/S0007123418000443 Published online by Cambridge University Press

in this modeling exercise, is not a single outcome but rather an aggregated outcome that lumps together disparate types of instability.[19] Inflation in Afghanistan is the same as inflation in Zimbabwe: an increase in prices and decline in the purchasing power of money. But the Syrian civil war (2011) is qualitatively different from the abrupt regime change in Madagascar (2009). Finally, political instability onsets represent, by definition, sharp breaks from a comparatively stable recent past; otherwise, they would not be onsets.[20] Forecasting political instability is thus more akin to forecasting complex phenomena like financial crises than bond yields or inflation rates – and financial crises are notoriously difficult to anticipate (Reinhart and Rogoff 2009).

It is possible to create models that predict continuity in political instability, that is, instability continuation, using autoregressive models. These models would accurately predict out-of-sample values, as current instability is an excellent predictor of instability in the next time period ($r = 0.86$). These types of models of instability continuation would be more conducive to methodologies used in other fields that explicitly identify and account for structural breaks.[21]

But these models would have little or no policy relevance. Typical goals of forecasting political instability are to shine light on unfolding crises before they explode and to inform decision making to respond to (or prevent) potential humanitarian catastrophes. Thus the emphasis is on the *onset* of instability, rather than the continuation of periods of (in)stability.

A new toolkit of methods is needed to better understand and measure temporal variability in the drivers of instability and mitigate this variability to improve model performance. While it may be impossible to predict fundamental changes in the international system that alter which causal drivers of instability events are relevant from period to period – or even how to define those periods, except in retrospect – it is possible to use forecasting techniques as inductive tools to diagnose when such shifts have occurred and when priors about drivers should be re-evaluated. In this article, we rely on out-of-sample prediction and moving AUROC scores as two approaches to more effectively diagnose this problem. Future research should extend our ability to measure these phenomena and build techniques that can address the problem of temporal variability.

Conclusions

This article revisited the influential GEA model of political instability and found that: (1) the model's predictive power varies substantially over time; (2) its predictive power peaked in the period used for out-of-sample validation (1995–2004) in the original study and (3) the model performs relatively poorly in the most recent period (2005–2014). This model was incredibly influential in pushing the field to focus both theoretically on political institutions and practically on using predictive power, rather than statistical significance, as a practical metric for model quality. However, academics and practitioners should be cognizant that the model's power has declined significantly since it was published. Moreover, this issue is not unique to the GEA model; we found similar dynamics in reanalysis of efforts to predict mass nonviolent uprisings and armed conflict onset and continuation. The CU and HEA models both demonstrate temporal variability, though these trends differ significantly across models with alternative assumptions. In some tests, models performed with relatively similar temporal patterns of predictive accuracy. Other tests, however, showed much greater variation in levels of predictive accuracy.

[19]Some may wonder why political instability is treated as an aggregated outcome, as opposed to decomposing it into different types of instability, e.g., civil war, abrupt regime change, etc. The reason is a practical one: several generations of political instability forecasting effort found no benefit, in terms of predictive accuracy, of doing so (Goldstone et al. 2010).

[20]The correlation between instability onset and its one-year lag is −0.01.

[21]For example, a two-period ($t - 1$ and $t - 2$) autoregressive model of instability incidence – rather than onset, which we seek to forecast here – returns a ROC of 0.94 over the full sample (1961–2014).

https://doi.org/10.1017/S0007123418000443 Published online by Cambridge University Press

These findings inform two paths for future research. First, they suggest the drivers of political instability are likely time variant. The GEA model was designed to predict political instability in the decade after the end of the Cold War, and emphasized factors like infant mortality and neighboring conflict that are typically higher in developing countries with comparatively weak political institutions – precisely the types of states that benefitted significantly from overt and covert major power support during the Cold War (Boix 2011). These states experienced both partial democratization and political instability at significantly higher rates in the post-Cold War era due to a retrenchment of major powers from intervening in their domestic politics and providing material support for weak regimes. GEA may have been right to argue that political institutions were the most important drivers of instability during most of the period under inquiry. An approach that emphasizes system-level factors that might condition the effects of country-level factors may be a path forward – with events like major regional transitions following the US invasion of Iraq in 2003, or the global financial crises of 2007–2008 – fundamentally reordering the most relevant factors associated with instability.

Secondly, this article illustrates how historical analysis of predictive accuracy over time (using a moving-decade AUROC score, for example) can help us identify the time periods in which major shifts and reordering may have occurred (or are occurring) in the international system. By evaluating sudden shifts in model performance, we can see where major political, economic, demographic or normative changes are altering the ways in which political instability occurs. Identifying these patterns can help researchers to confirm or disconfirm their intuitions about when key processes in global politics are changing – and when they are staying the same. Indeed, identifying periods of systemic instability and reordering – like the period contemporaneous with the Arab Uprisings – may be a more important task than identifying any individual instance of political instability.

Supplementary material. Replication data sets are available in Harvard Dataverse at: https://doi.org/10.7910/DVN/XMGVO2, and online appendices at https://doi.org/10.1017/S0007123418000443.

Acknowledgements. The authors would like to thank the Minerva Research Initiative (award W911-NF-14-1 0538) as well as colleagues Barry Hughes and Tim Sisk for their support in conducting this analysis.

References

Banerjee MV and Duflo E (2009) The experimental approach to development economics. *Annual Review of Economics* **1**, 151–178.

Boix C (2011) Democracy, development, and the international system. *American Political Science Review* **105** (4):809–828.

Bowlsby D, **Chenoweth E**, **Hendrix C and Moyer J** (2018) Replication Data for: The Future is a Moving Target: Predicting Political Instability, https://doi.org/10.7910/DVN/XMGVO2, Harvard Dataverse, V1, UNF:6:s2+Q3W+hPkdB/e24Vhx/IQ== [fileUNF].

Cederman LE and Weidmann NB (2017) Predicting armed conflict: time to adjust our expectations? *Science* **355** (6324):474–476.

Cerra V and Saxena SC (2008) Growth dynamics: the myth of economic recovery. *American Economic Review* **98** (1):439–457.

Chenoweth E (2015) *Major Episodes of Contention Dataset*, Vol. 1. Denver, CO: University of Denver.

Chenoweth E and Ulfelder J (2017) Can structural conditions explain the onset of nonviolent uprisings? *Journal of Conflict Resolution* **61** (2):298–324.

Chiba D and Gleditsch KS (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.

Collier P and Hoeffler A (2004) Greed and grievance in civil war. *Oxford Economic Papers* **56** (4):563–595.

Dafoe A (2014) Science deserves better: the imperative to share complete replication files. *PS: Political Science & Politics* **47** (1):60–66.

Evanschitzky H and Armstrong JA (2010) Replications of forecasting research. *International Journal of Forecasting* **26** (1):4–8.

Fawcett R (2006) An introduction to ROC analysis. *Pattern Recognition Letters* **27** (8):861–874.

Friedman M (1968) The role of monetary policy. *American Economic Review* **58** (1):1–17.

Gates S *et al.* (2012) Development consequences of armed conflict. *World Development* **40** (9):1713–1722.
Ghobarah HA, **Huth P and Russett B** (2004) The post-war public health effects of civil conflict. *Social Science & Medicine* **59** (4):869–884.
Giacomini R and Rossi B (2009) Detecting and predicting forecast breakdowns. *The Review of Economic Studies* **76** (2):669–705.
Goldstone JA *et al.* (2010) A global model for forecasting political instability. *American Journal of Political Science* **54** (1):190–208.
Gurr TR (1970) *Why Men Rebel*. Princeton, NJ: Princeton University Press.
Hegre H and Sambanis N (2006) Sensitivity analysis of empirical results on civil war onset. *Journal of Conflict Resolution* **50** (4):508–535.
Hegre H *et al.* (2013) Predicting armed conflict, 2010–2050. *International Studies Quarterly* **57** (2):250–270.
Hegre H *et al.* (2017) Introduction: forecasting in peace research. *Journal of Peace Research* **54** (2):113–124.
Honaker J and King G (2010) What to do about missing values in time-series cross-section data. *American Journal of Political Science* **54** (2):561–581.
Litterman L and Scheinkman J (1991) Common factors affecting bond returns. *Journal of Fixed Income* **1** (1):54–61.
Lucas RE (1973) Some international evidence on output–inflation tradeoffs. *American Economic Review* **63** (3):326–334.
Marshall MG, **Gurr TR and Harff B** (2016) *PITF State Failure Problem Set*. Vienna, VA: Center for Systemic Peace.
Nosek BA *et al.* (2015) Promoting an open research culture. *Science* **348** (6242):1422–1425.
Nyseth Brehm H (2017) Re-examining risk factors of genocide. *Journal of Genocide Research* **19** (1):61–87.
Pesaran M, **Pettenuzzo D and Timmermann A** (2006) Forecasting time series subject to multiple structural breaks. *Review of Economic Studies* **73** (4):1057–1084.
Reinhart C and Rogoff KS (2009) *This Time is Different: Eight Centuries of Financial Folly*. Princeton, NJ: Princeton University Press.
Stock JH and Watson MW (1996) Evidence on structural instability in macroeconomic time series relations. *Journal of Business & Economic Statistics* **14** (1):11–30.
Themnér L and Wallensteen P (2012) Armed conflicts, 1946–2011. *Journal of Peace Research* **49** (4):565–575.
Vreeland JR (2008) The effect of political regime on civil war: unpacking anocracy. *The Journal of Conflict Resolution* **52** (3):401–425.
Ward MD and Beger A (2017) Lessons from near real-time forecasting of irregular leadership changes. *Journal of Peace Research* **54** (2):141–156.
Ward MD, **Greenhill BD and Bakke KM** (2010) The perils of policy by p-value: predicting civil conflicts. *Journal of Peace Research* **47** (4):363–375.
Wendt AE (1987) The agent-structure problem in international relations theory. *International Organization* **41** (3):335–370.

Cite this article: Bowlsby D et al. (2020). The Future is a Moving Target: Predicting Political Instability. *British Journal of Political Science* **50**: 1405–1417, doi: 10.1017/S0007123418000443