An Early Warning System for Democratic Resilience: Predicting Shocks to Civic Space*

Xiaoyin Chen *Duke University*
Jeremy Springman *University of Pennsylvania*
Erik Wibbels *University of Pennsylvania*

DEVLAB@Penn

September 23, 2022

Civil society is a powerful force for political change and democratic accountability (Carothers 2020). Civil society movements have been credited with sparking instances of popular mobilization ranging from local land disputes all the way to regional 'colour revolutions' (Gilbert 2020; Gilbert and Mohseni 2018). Understanding this, a growing number of governments have cultivated a diverse repertoire of repressive tactics (Bagozzi, Berliner, and Welch 2021), ranging from legal sanctions to outright physical coercion. Advances in big data analytics are endowing governments with new tools, including the ability to anticipate citizen action and engage in preemptive repression (Feldstein 2019), which likely contributes to the waning effectiveness of traditional non-violent resistance (Chenoweth 2020).

To reverse this alarming trend, civil society needs new tools to navigate increasingly sophisticated repression. The Machine Learning for Peace (MLP) project leverages new technologies to bolster civil societies. Of utmost importance in this regard is the capacity of local civic actors to strategize around major changes to civic space. In this research note, we report on our ability to forecast *major shocks* to civic space using the MLP dataset. Analyzing 9 different civic space event types and 39 countries, we find:

- Predicting shocks to civic space is difficult. For most country-event pairs, we cannot reliably predict shocks. However, we are able to predict certain shocks in certain places with considerable precision. Specifically, we accurately forecast censorship in Armenia and Colombia, legal changes in Colombia, troop/police mobilization and purges in Nigeria, purges in Tanzania, legal actions in Turkey, and non-lethal violence in Uganda.

- We accomplish this using 'interpretable' models that reveal the model's decision-making process. Interpretable models provide a way for practitioners with contextual knowledge to judge how reliable models are in the real world. Specifically, we show the precise variables that lead a model to predict a future shock. Using the case of legal actions in Turkey, we show that our model follows a simple decision-making process driven largely by sensible changes in substantively interesting variables.

The major events that we forecast constitute notable shocks to civic space. Thus, we provide the basis for an 'early warning system' that could help civil society strategize around

*This study is part of the Illuminating New Solutions and Programmatic Innovations for Resilient Spaces (INSPIRES) project funded by the United States Agency for International Development (USAID) Center for Democracy, Human Rights, and Governance.

repressive government action. This work contributes to a growing body of academic research aimed at producing actionable tools and insights for citizens engaged in nonviolent resistance in restrictive political environments (Manekin and Mitts 2021; Chenoweth 2021; Pinckney, Butcher, and Braithwaite 2022).

Measuring & Forecasting Civic Space Events

The "third wave of autocratization" has brought renewed attention to the study of political transitions broadly, and "democratic backsliding" in particular (Lührmann and Lindberg 2019; Waldner and Lust 2018). This attention has been accompanied by a proliferation of annual measures of regime type, including the Varieties of Democracy project (V-Dem) and the Civil Society Organization Sustainability Index. However, civic space closures often happen abruptly, as governments seize on crises to restrict fundamental rights or expand executive authority. Thus, standard data are not able to identify rapidly changing events that might predict changes that will occur over weeks or months rather than years or decades.

Our approach conceptualizes changes in civic space as a product of specific important events. To measure changes in civic space, the MLP team built a massive data production pipeline that tracks reporting on 20 types of events bearing on civic space. To select event-types of interest, we consulted existing scholarship and civil society and governance experts. To track these events, we scrape the text of online news and use recent advances in natural language processing (NLP) to identify articles that are reporting on each of our event types. For each month, we calculate the share of all articles published about each country that are reporting on each event type. These "event data" provide a structured record of politically relevant occurrences, such as protests or changes in a country's laws.[1]

To date, MLP has scraped more than 70 million articles from more than 100 international, regional, and domestic online news sources in 22 different languages for a sample of 43 countries. We update these data for each country on a quarterly basis to accelerate the provision of data to practitioners and policymakers who need to take timely, evidence-based action to counter attacks on civil society. In addition to tracking these events, MLP also produces monthly forecasts predicting future levels of activity (proxied by future levels of reporting) for a subset of 11 of our 20 events. When these forecasts are accurate, they can provide civic actors with a sense of future trends in arrests, protests, legal changes, etc.

Here we are concerned with a slightly different task. Rather than forecasting modest changes in different civic space events, we seek to forecast large, discontinuous changes– these are the major civic events that governments often use to fundamentally alter the rules of the game. We identify these 'shocks' by calculating the change in the value of the target event type between month X and month $X + k$ for every month in the sample, with k being the number of months into the future that the model is forecasting. In this report, we look at how well models can predict shocks in 9 different event types 3-months in the future. This 3-month forecasting window allows us to focus on the relatively near-future, posing a less challenging forecasting task (forecasting the near-future is easier than the far-future). However, it still produces potentially useful information by providing at least 2-months of advanced warning for major events (new data is produced during month $X + 1$).

[1]See our Technical Report on the Production of Civic Space and RAI Event Count Data for details.

We select the 20% of months with the largest increases between month X and $X + k$ and classify those months as 'shocks' (with all other months not being shocks). This transforms the task into a classification problem and simplifies interpretation by focusing only on providing a warning about very large increases in activity. We select the largest 20% of increases because testing suggests this threshold strikes a balance between the performance of our models and the magnitude of events. Our predictive models achieve higher levels of accuracy when a larger share of the sample is classified as shocks. Specifically, our models perform significantly more accurately predicting the largest 20 and 25% of shocks than the largest 10 or 15% of shocks. However, the higher the threshold, the more common the events that are classified as shocks. 20% results in 18-24 observed shocks depending on the number of months available for each country. This intermediate threshold ensures that shocks capture major events rather than modest fluctuations in the month-to-month share of articles reporting on each target event.

Model Selection, Training, and Evaluation

To train a model that can predict when shocks to civic activity are likely to occur, we use a simple 'ensemble' of 'boosting' algorithms using a method called *AdaBoost*.[2] While there are many alternative methods, we select AdaBoost because it is relatively simple compared to many similar approaches. This simplicity maintains an ability to model non-linear relationships and simplifies interpretation while guarding against over-fitting (a concern given the relatively small sample of data we have for each country). In other words, AdaBoost can model highly complex relationships while remaining transparent in it's decision-making.

As with most machine learning models, AdaBoost requires a process of tuning various parameters of the algorithm (parameterization) to improve performance. This process allows researchers to control features like model complexity. While more complex models can learn about more complex relationships between predictor and target variables, they risk 'learning' complex patterns that are chance features of the data (i.e. 'over-fitting') and reduce performance when trying to make predictions in the real world.[3]

For each country-event pair, we train an AdaBoost classifier and tune the parameters using a grid search with 10-fold cross-validation. Cross-validation is designed to test the performance of a range of different parameter settings across multiple samples of the data to reduce the risk of over-fitting.[4] For each observed month of the target variable X, AdaBoost returns the probability that there will be a shock in activity in month $X + k$. To generate a prediction, we must convert this probability into a binary classification. For this task, we

[2] AdaBoost uses sequential classification trees on samples of the data and applies higher weights to misclassified observations, gradually learning how to classify more observations in the sample by correcting past mistakes. The final prediction is made by a weighted average across all trees.

[3] The parameterization of AdaBoost has two tunable hyperparameters: the number of stumps (iterations) and the learning rate. The number of stumps controls the complexity of the model, with more iterations allowing for more complex relationships between predictor and target variables. The learning rate controls how much each stump in the ensemble contributes to the final prediction. A lower learning rate requires more stumps to learn from, while a higher learning rate can learn more quickly from fewer stumps.

[4] The range for the number of iterations under considerations is 10-100 (10 per step) and the range for the learning rate is 0.1-2 (0.1 per step).

classify as a shock any observation with a greater than 0.5 probability of being a shock. We use the F1-score as the metric for identifying the values for each hyperparameter that yield the best performance.[5]

Model Performance

In this section, we assess the ability of our models to predict civic space shocks across country-event pairs. For each country-event model, we 'train' the model by testing its performance using different values of the hyperparameters and selecting the values that yield the best F1 score. Importantly, we use only the first 80% of months in the sample during this process, referred to as the training set. We then measure the model's performance on the final 20% of months, referred to as the test set, to estimate how well the lessons learned from the training set allow the model to make accurate predictions about new data that it did not learn from.

To identify models that may provide reliable information to practitioners, we want models that can predict as many shocks as possible while avoiding false-positives (predicting a shock when there is not one). Because we want these models to provide actionable information, we prefer to tolerate more false-negatives than false-positives. In other words, we care more about *avoiding* incorrectly warning practitioners about shocks that will not actually happen than we do failing to warn them about events that will happen. For this reason, we identify models that have at least 0.7 precision and at least 0.5 recall. We consider models that meet these criteria to be high-performing. This criteria is arbitrary, but provides a reasonable minimum performance to consider our models useful.

Figure 1 shows the precision scores across each country-event pair for models estimating shocks in civic space activity 3-months into the future. Of the 429 country-event pair models (11 events across 39 countries), the majority exhibit values below our performance criteria, indicating that we are unable to accurately predict shocks in civic space activity for many country-event pairs. This may be the result of overfitting, underfitting, or measurement error, but they almost certainly reflect the unpredictability of many civic space shocks (Kuran 1991).

However, we are able to predict certain events in certain countries with a high degree of precision, even with these relatively simple models. Specifically, our models accurately forecast censorship in Armenia and Colombia, legal changes in Colombia, troop mobilization and purges in Nigeria, purges in Tanzania, legal actions in Turkey, and non-lethal violence in Uganda. The points representing each of these pairs in Figure 1 have a black circle indicating their high-performance.

Figure 2 plots the performance of the model predicting legal actions in Turkey, our best performing model. This figure visualizes performance on the last 20% of the months (test set) that are withheld from the model during the training process. Of the 24 months in the testing data, we observe 6 shocks in legal action. Our model accurately predicts 4 of these shocks (67% recall) and predicts 1 shock that doesn't happen (80% precision).

[5]The F1-score combines the two most common measures of predictive performance of classification models, precision and recall, into a single metric by taking their harmonic mean. Both scores can be thought of as measuring a model's ability to make 'true positive' predictions while avoiding 'false positive' predictions. By averaging across these two measures, F1 provides a balanced assessment of how model performance is impacted by the values of the hyperparameters.

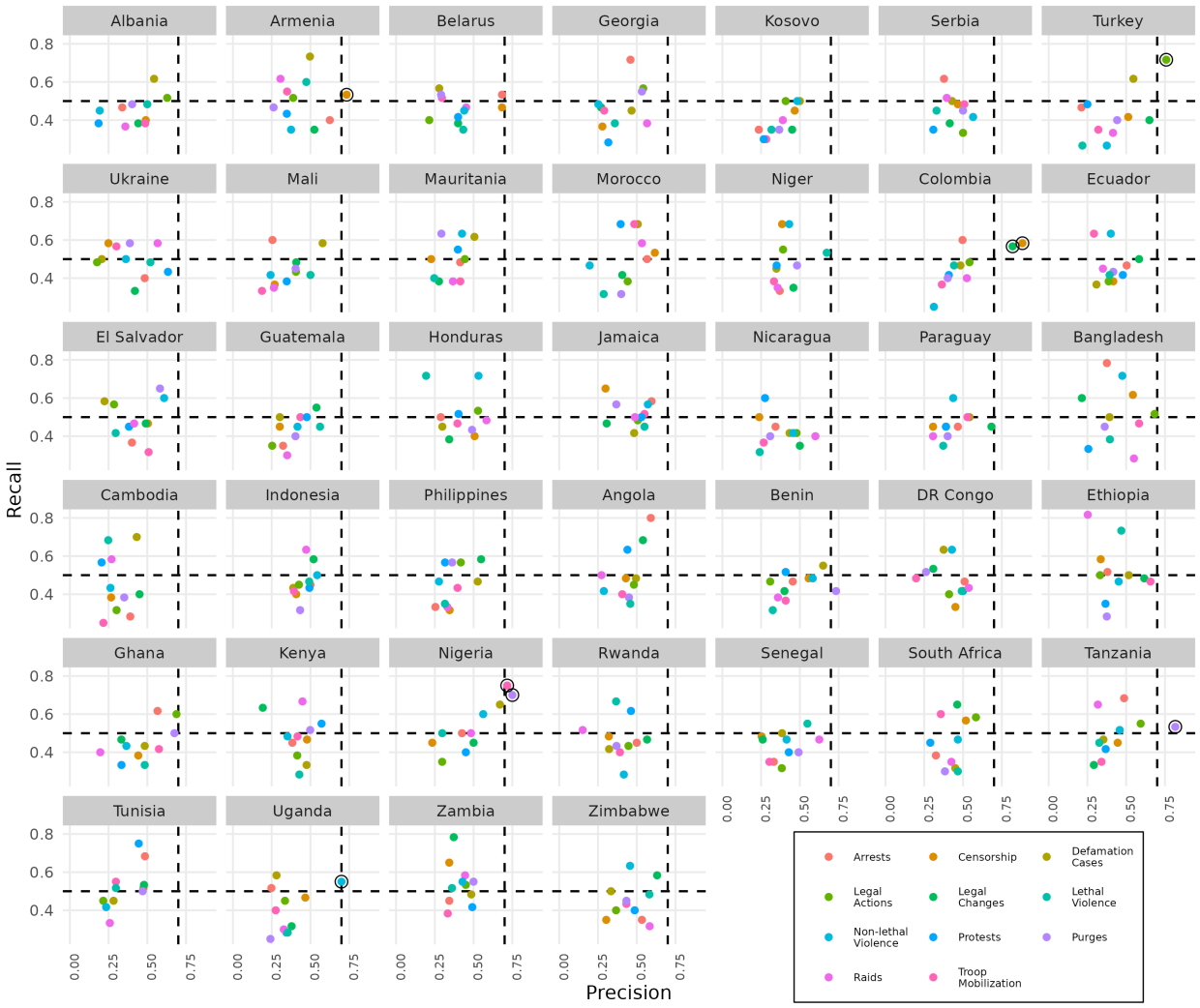

Figure 1: AdaBoost model precision and the number of stumps using 10-fold cross-validation. Black circles around dots indicate high performing models with at least 0.7 precision and at least 0.5 recall.

We believe this provides a 'proof-of-concept' that the MLP data can produce actionable insights for citizens engaged in civic activism. By providing advanced warning of when practitioners should expect increases in different civic event types, these actors may have an opportunity to prepare. However, for this tool to truly inform strategic decisions, it is critical that we understand why a model is predicting shocks in some months but not others. This ability to 'interpret' is important so that information about conditions on the ground (only a fraction of which is captured by our data and known by the model) can be used to assess the credibility of predictions on a case-by-case basis.

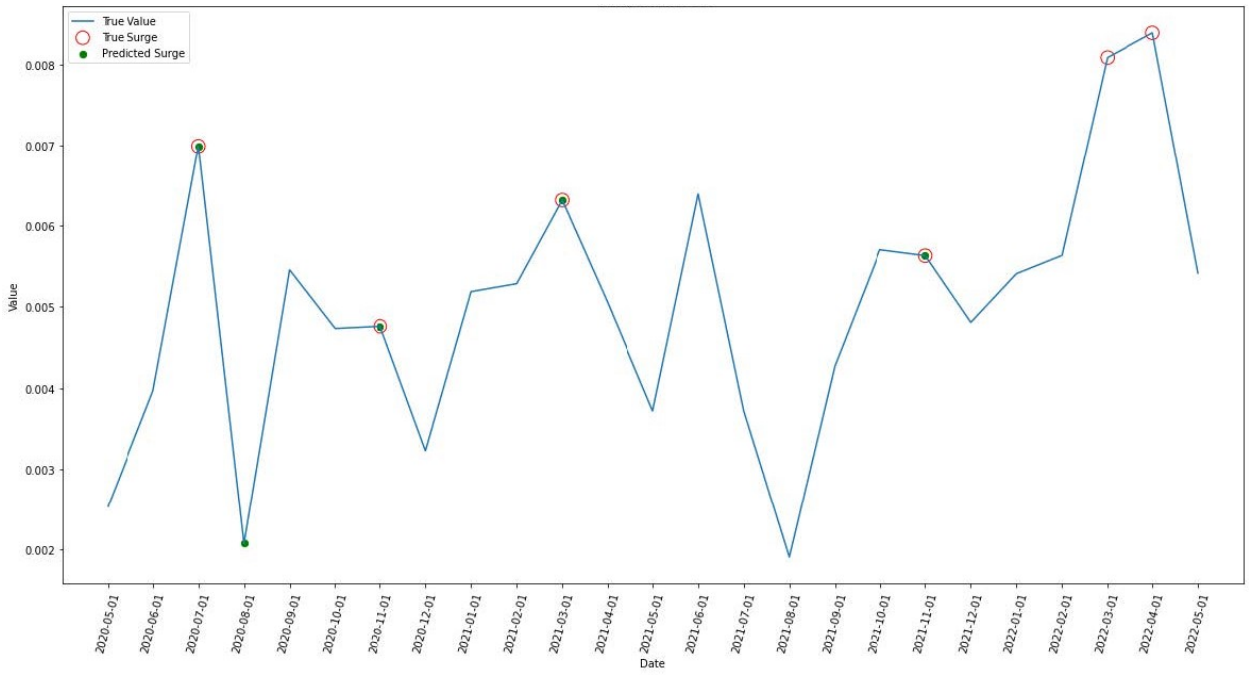

Figure 2: AdaBoost model performance on 20% reserved testing data.

Model Interpretation

In this section, we illustrate how interpretable machine learning models such as Adaboost allow informed practitioners to judge the credibility of forecasts on a case-by-case basis. Models use information about historical patterns between variables to make predictions based on current conditions. However, this approach relies on the persistence of past patterns into the future. Interpretable models allow substantive experts to judge the credibility of forecasts using their knowledge about how the world is changing and how well historical patterns are likely to predict future events (Rudin 2018). In this exercise, we focus on three pieces of information that we think may be useful to practitioners: the list of variables that the model identifies as being predictive of shocks in the target variable, the direction(s) of the relationship between each predictor variable and the target, and the specific changes in these predictor variables that lead the model to expect shocks in a specific future month.

To provide an example, we focus on the largest 3-month shock in the testing data for our highest-performing model, the July 2020 shock in legal actions in Turkey. That month saw an unusually large increase in reporting on legal actions in the country, with the number of articles jumping from 78 in June to 147 in July and the total share of articles reporting on legal actions jumping from 0.4% to 0.7%. This reporting captured a number of significant events, including the arrest of two opposition mayors on charges of terrorism, criminal charges or sentences against at least five journalists, the bringing of charges against two actors for "insulting the president," and charges against at least four high-profile members of civil society on terrorism charges.

What information lead the model to expect a shock in activity in July 2020? We first inspect the list of variables that the model identifies as being predictive of shocks in legal

actions. The variables provided to the model includes the share of articles reporting on each of our 20 civic space and 22 RAI events, the raw count of articles reporting on each event, binary indicators capturing the the quarter each month belongs to (to account for seasonality), and 38 economic variables from TradingEconomics. Of these variables, the model identifies past values of five civic space events as predictive of future changes in legal action, including current levels of legal actions, protests, defamation cases (a specific type of legal action), political threats, arrests, and three RAI events, including trade agreements, diaspora activation, and economic aid. Changes in these variables in the current month lead the model to expect changes in legal action three months into the future.

To better understand the model's decision-making process, we now turn to looking at the direction of the relationship between each predictor and future changes in the target. Of the nine predictor variables, the three strongest predictors are legal action, defamation cases, and arrests. For legal action and defamation cases, high values are associated with a decreased probability of future legal action shocks, while low values signal an increased probability of a shock. For arrests, the opposite is true, with high levels of arrest in the current month predicting elevated levels of legal action three months into the future.

These directions generate important insights into the historical patterns that the model is relying on. Specifically, the initiation of major legal actions and defamation cases are spread-out across time, suggesting that Turkey's use of legal repression follows cycles of intense activity followed by relative inactivity. Interestingly, arrests follow the opposite pattern, with low levels of arrest predicting low levels of legal action, whereas historically high levels of arrest predict high levels of legal action three months in the future. This makes sense, given that legal actions against individuals often follow their arrest.

Finally, we look at the precise values in our data that lead our model to predict a shock in activity in July 2020. Consistent with the directional relationships already discussed, we see that in April 2020, both legal action and defamation cases were at very low levels relative to their historical average. Alternatively, the arrests variable was slightly above it's historical mean. By identifying the specific historical patterns and current conditions that cause our model to predict a shock, practitioners may assess how robust relationships are likely to be and consider whether the trajectory of the near future is likely to adhere to historical patterns.

Conclusion

In this report, we test an application of the MLP data as an early warning system designed to forecast major shocks to civic space. Using simple, interpretable machine learning forecasting models, we demonstrate an ability to forecast a subset of country-event pairs with a high degree of accuracy. We will use these results to build a public-facing early-warning system into the MLP website.

We also demonstrate that interpretable models can generate useful information about the model's decision-making. Using the case of legal actions in Turkey, we discuss the specific types of information that result in the prediction of shocks. We discuss this information in detail and illustrate how practitioners might combine their understanding of the model's decision-making with their substantive knowledge to judge the credibility of predictions.

References

Bagozzi, Benjamin E, Daniel Berliner, and Ryan M Welch (2021). "The diversity of repression: Measuring state repressive repertoires with events data". In: *Journal of Peace Research* 58.5, pp. 1126–1136.

Carothers, Thomas (2020). "Rejuvenating democracy promotion". In: *Journal of Democracy* 31.1, pp. 114–123.

Chenoweth, Erica (2020). "The future of nonviolent resistance". In: *Journal of Democracy* 31.3, pp. 69–84.

— (2021). *Civil Resistance: What Everyone Needs to Know*. Oxford University Press.

Feldstein, Steven (2019). *The global expansion of AI surveillance*. Vol. 17. Carnegie Endowment for International Peace Washington, DC.

Gilbert, Leah (2020). "Regulating Society after the Color Revolutions: A Comparative Analysis of NGO Laws in Belarus, Russia, and Armenia". In: *Demokratizatsiya: The Journal of Post-Soviet Democratization* 28.2, pp. 305–332.

Gilbert, Leah and Payam Mohseni (2018). "Disabling dissent: the colour revolutions, autocratic linkages, and civil society regulations in hybrid regimes". In: *Contemporary politics* 24.4, pp. 454–480.

Kuran, Timur (1991). "Now out of never: The element of surprise in the East European revolution of 1989". In: *World politics* 44.1, pp. 7–48.

Lührmann, Anna and Staffan I Lindberg (2019). "A third wave of autocratization is here: what is new about it?" In: *Democratization*, pp. 1–19.

Manekin, Devorah and Tamar Mitts (2021). "Effective for Whom? Ethnic Identity and Nonviolent Resistance". In: *American Political Science Review*, pp. 1–20.

Pinckney, Jonathan, Charles Butcher, and Jessica Maves Braithwaite (2022). "Organizations, Resistance, and Democracy: How Civil Society Organizations Impact Democratization". In: *International Studies Quarterly*.

Rudin, Cynthia (2018). "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead". In: DOI: 10.48550/ARXIV.1811.10154. URL: https://arxiv.org/abs/1811.10154.

Waldner, David and Ellen Lust (2018). "Unwelcome change: Coming to terms with democratic backsliding". In: *Annual Review of Political Science* 21, pp. 93–113.