From academia to policy makers: a methodology for real-time forecasting of infrequent events

Alfred Krzywicki1,4 · David Muchlinski2 · Benjamin E. Goldsmith3 · Arcot Sowmya1

Received: 28 November 2021 / Accepted: 27 July 2022 / Published online: 17 August 2022
© The Author(s) 2022

Abstract
The field of conflict forecasting has matured greatly over the last decade. Advances in machine learning have allowed researchers to forecast rare political and social events in near real time. Yet the maturity of the field has led to a proliferation of diverse platforms for forecasting, divergent results across forecasts, and an explosion of forecasting methodologies. While the field has done much to establish some baseline results, true, consensual benchmarks against which future forecasts may be evaluated remain elusive, and thus, agreed upon empirical results are still rare. The aim of this work is to address these concerns and provide the field of conflict forecasting with a standardized analysis pipeline to evaluate future forecasts of political violence. We aim to open the black box of the conflict forecasting pipeline and provide empirical evidence on how modeling decisions along all steps of the pipeline affect end results. In this way, we empirically demonstrate best practices that conflict forecasting researchers may utilize in future endeavors. We employ forecasts of targeted mass killings and genocides to support our methodological claims.

Keywords Machine learning · High-dimensional data · Data imbalance · Methodology

Alfred Krzywicki
alfred.krzywicki@adelaide.edu.au
David Muchlinski
david.muchlinski@inta.gatech.edu
Benjamin E. Goldsmith
ben.goldsmith@anu.edu.au
Arcot Sowmya
a.sowmya@unsw.edu.au

1 School of Computer Science and Engineering, UNSW Sydney, Sydney, Australia
2 Sam Nunn School of International Affairs, Georgia Institute of Technology, Atlanta, GA, USA
3 School of Politics and International Relations, Australian National University, Acton, ACT, Australia
4 School of Computer Science, University of Adelaide, Adelaide, Australia
Introduction

Conflict forecasting has matured as a field of research over the last decade. There now exists a multitude of platforms for predicting numerous kinds of political violence, including civil conflict across Africa [1], and civil war in general [2], targeted mass killings and genocides [3], mass atrocities [4], political instability [5], and coups [6]. Increasing ability to forecast these violent events can help bridge the gap between basic and applied research and warn policy makers of impending humanitarian crises [7]. Accurate prediction of infrequent but devastating events, such as mass killings or terrorist attacks, is important as it could be applied to initiate preventive actions [8] and is a gold-standard test of the validity of academic theories [9]. While substantial progress has been made in forecasting rare violent events, the methods generally chosen to benchmark may generalize poorly when applied to different types of political violence, and the challenges associated with using models in real time are rarely considered. We address this gap by developing a methodology for the forecasting of real-time events (FORTE) that can be used to generate predictions of political violence. We demonstrate with empirical evidence that forecasting performance may be optimized on a specific dataset by adopting the FORTE methodology.

We choose to demonstrate our forecasting methodology on a newly released dataset of genocide onsets and mass atrocities, namely, the targeted mass killing (TMK) dataset [10]. The TMK dataset employs a baseline measure of atrocity severity and perpetrator intent to identify 201 episodes (1946–2017), and includes indicators of higher-level intent and severity, precise start and end dates, perpetrator and target groups, and descriptions of triggering events and episode endings. We utilize the TMK because it contains numerous benefits over other established datasets in the field, including those commonly utilized to measure state repression, one-side violence against civilians, and other mass violence events. The TMK dataset (1) clarifies ambiguities and measurement problems associated with intent to destroy specific groups, while maintaining a distinction between mass killing, repression, and terrorism; (2) enables users to customize thresholds of intent and severity for particular research questions, for example, to identify genocide/politicide episodes as a subset of TMK events; (3) explicitly and systematically identifies state and non-state actors as perpetrators; (4) identifies attempted episodes of genocide; and (5) provides extensive documentation of coding decisions and source material [10].

Forecasting targeted mass killing is difficult due to the rarity of these events, which tends to bias predictions of genocide downward toward the majority class of no onset [11]. This work discusses two challenges: development of an accurate forecasting methodology and its application to the forecasting of genocide onsets for both academic researchers and practitioners alike. In the next section, we discuss details of the methodology, including building the dataset, metrics, parameter tuning, model training, and automation for those with little experience with machine learning methods. Methodology development is based on consideration of many choices, resulting in many paths that may be utilized for prediction [12]. Rarely is the black box of forecasting opened to reveal the effects of different methodological
choices on the results, and many of these choices, we demonstrate, result in sub-optimal predictions. Therefore, we make our pre-processing and other pipeline choices explicit and provide evidence suggesting that our recommendations should be preferred to the alternatives. In “Analysis of final results and selection of forecast model,” we demonstrate the utility of our forecasting methodology by presenting the final set of results obtained from the surviving paths in the methodology, together with some examples of successfully predicted genocide onsets in “Model training and forecasting event.”

For evaluation, we contend that it is important to identify true positives in addition to evaluating the overall performance of the model. Our methodology presents the researcher with a small number of cases to evaluate, allowing for practitioners, subject matter experts, and other qualitative researchers to evaluate the overall forecasting methodology. It is important that this number is chosen carefully, considering the relative cost associated with potentially missing events against raising unnecessary false alarms. Therefore, we automate this process to find an optimal number of forecasts. In “Conclusion and future work,” we summarize the main points of the paper and briefly discuss future extensions of the research to further improve prediction models and result presentation.

The overall contribution of this work is a unified and systematic methodology for the forecasting of rare events from a highly imbalanced dataset that contains commonly utilized observational time series cross-sectional data, complete with the known challenges associated with using such data, including missing values, and time-varying and time-invariant variables. We contend that this makes our methodology well suited to future benchmarks using alternative choices along our forecasting paths, including multiple models, ensembles, or alternative forecasting windows or methods of missing data imputation. Further, the forecasts are guided by metrics selected to maximize the benefits of the produced results for academic researchers as well as policy makers. While we detail the specifics of the methodology in this paper, this process can also be automated to any degree desired, such that even those with little background knowledge in machine learning and quantitative forecasting can utilize, interpret, and make recommendations based on the results of our methodology. Those who wish to have more direct control over the process can choose their own preferred degree of automation.

The FORTE methodology

In this section we introduce our FORTE methodology. Our goal is to design an accurate forecasting pipeline for the near real-time forecasting of political violence. While our process is general, we test our pipeline on a dataset of targeted mass killings [10] due to the challenge of successfully predicting such rare events, and because the forecasting of such events has important ramifications for practitioners. As previously mentioned, rather than presenting a single forecasting pipeline, we open the methodological black box to compare and discuss results from several forking paths along the pipeline to assess the efficacy of different modeling choices.
In Table 1 we list the stages of FORTE in the first column, while the second column summarizes their application to the genocide forecasting case. Each step is then described in more detail in subsequent subsections.

In Fig. 1, red boxes depict general methodology steps, while blue lines indicate paths specific to genocide forecasting. Broken blue lines mark options that have been rejected in the process of building our pipeline and are discussed in each subsection to provide empirical support for our modeling choices along these forking paths, while solid blue lines indicate all the options for which results are discussed in the Final results section. Based on these results, the best combination of options can be selected.

The order of stages shown in Fig. 1 corresponds to Table 1 and the following subsections are numbered accordingly.

**Prediction task definition**

The first and obvious step is the formulation of the prediction task, i.e., what we are trying to predict. In our case, given values of independent variables for past years and the current year, the main task is to predict if a genocide onset will occur in a

| Table 1 | Stages of FORTE methodology | Application to genocide forecasting |
|---------|-----------------------------|-------------------------------------|
| 1. Definition of prediction task | A True/False value indicating if a genocide will occur in future years |
| 2. Selection of metrics | Select MAP as a main metric |
| 3. Choosing forecasting algorithms | Presents results for GLM, GAM, and RF, the three forecasting methods commonly used for this type of prediction |
| 4. Dataset construction | Create country-year dataset for all countries available in data sources |
| 4.1 Pre-processing and merging sources | Merge sources by retaining all selected data from all sources. Convert logical to numerical and normalize |
| 4.2 Selecting missing values imputation method | Select imputation method based on a prior test |
| 4.3 Selecting prediction target | Choosing number of prediction years to predict if a genocide event will occur in any of these years |
| 4.4 Selecting class balancing methods | Select as part of parameter tuning |
| 5. Selection of predictors based on variable importance | Variable importance is obtained from trained models, variables below importance average are dropped and the rest used to re-train the model. |
| 6. Selection of windowing technique | Every year, data window moves one year at a time. Results of fixed window versus expanding window are compared |
| 7. Optimizing number of forecasts | Optimizing the number of predicted country-years in which genocide onsets are most likely to occur |
| 8. Model training and making forecasts | Based on results from previous steps, select and train the model and produce forecasts |
country in future years. This target takes on a binary value (i.e., 0, 1) for no onset or onset, respectively. Thus, our primary task is classification. Following similar research [8, 11, 13] we use country-years as a base unit of forecasting. Since we are predicting future genocide events with respect to the current year, we differentiate data year as a data point with independent variable values and ground truth, from prediction year, in which we predict the onset. For example, variable values from 1960 are used to predict onsets in any year from 1961 to 1963.

Selection of metrics

A rich literature [14, 15] demonstrates the importance of selecting metrics to evaluate overall forecast error. It is generally understood that common metrics, like the area under the Receiver Operating Characteristic (ROC curve, accuracy, classification error, and other metrics, can often provide misleading evaluations of forecast accuracy when data are class [16]. For this reason, area under the Precision/Recall curve and F-Score have often been utilized to evaluate forecasts on class imbalanced data, as these metrics are unaffected by rare events [15, 16]. We introduce an alternative metric, mean average precision (MAP) as the main evaluation metric to evaluate forecasts of rare events. Next, we justify its utility for evaluating the models.

For forecasts of political violence, the emphasis is on minimizing false negatives (missing genocide onsets) at the cost of some small increase in false positives, as the former is costlier than the latter. As part of our FORTE pipeline, after model training, a set of test cases is presented with a list of prediction scores, one for each test case,
corresponding to the predicted probability of true positive. These scores are then used to form a list of most probable predictions as illustrated in Fig. 2, with 10 numbered test cases that include four genocide events (2, 5, 9, and 10).

Higher scores correspond to higher predicted probabilities of genocide onset. This list of scores is then ordered from high to low and a subset of top cases (5 in the figure) are presented to the user. The number of cases in this prediction set can be arbitrarily selected or, as discussed further in this paper, optimized to maximize True Positives (TP) and minimize False Positives (FP) and False Negatives (FN). For a given size of the prediction set, correct predictions (denoted with red “1” in the figure) should ideally be at the top of the set. Consequently, in order to compare models, we need a metric that measures how close all correct predictions are to the top of the prediction set. The commonly used metrics, such as precision and recall, are not sufficient for this purpose. Consider, for example, the two models in Fig. 2 where Model 1 outputs “10011” and Model 2 “11100.” Both precision and recall are the same, while it is obvious that the second model is better, as it provides more correct predictions at the top of the prediction set. For this reason, we use Average Precision at k top predictions for a given year [7] as the overall metric, and Mean Average Precision at k (MAP@k) when AP@k is averaged over all test years. This metric is frequently used in recommender systems and has values between 0 and 1. For the two prediction sets shown, AP@5 is 0.42 and 1.0 respectively. For comparing general results, k is the number of country-years available for a given test year. The AP metric, defined in Eq. (1), can be interpreted as the precision at k for all country-year pairs for a given year.

\[
\text{AP@}k = \frac{1}{k} \sum_{i=1}^{k} (p_i \times \text{rel}_i),
\]  

(1)

![Fig. 2 Illustration of calculating mean average precision (MAP) evaluation metric](image-url)
where $p_i$ is the precision at $ith$ prediction, and $rel_i$ is the relevance at $i$, which is 1 for TP and 0 otherwise.

**Choosing forecasting algorithms**

The most widely used analysis methods for genocide forecasting and other similar domains are generalized linear models (GLM) [17] and generalized additive models [18], which are types of regression models where the outcome is binary, e.g., genocide or no genocide for a country-year pair.

The expected value of the dependent variable $Y$ in the GLM model is a function of a linear predictor $\mathbf{X}\beta$. $E(Y|X) = g^{-1}(\mathbf{X}\beta)$, where $g^{-1}$ is a link function that links the linear predictor with a probability distribution. For some distributions, the GLM model becomes logistic regression (Logit). Generalized Additive Model [18], a sub-family of GLM, is often used in Probit analysis [3, 8]. GAMs are often especially useful when the functional form of the relationship between $X$ and $Y$ is not theoretically well specified, as is the case with much of the political violence literature [11]. The main innovation in GAM is the use of smooth functions in the linear predictor $\mathbf{X}\beta$ that are differentiable functions of some order.

Another popular method belonging to the Machine Learning family is Random Forests [19] used, for example by Muchlinski et al. [11]. A combination of Decision Trees, usually a few hundred, is used, each supplied with a sub-sample of training data. The prediction is the class obtained by voting.

In this research, we compare these three forecasting algorithms since they are used in similar research as discussed above.

1. GLM with binomial link function, which is effectively logistic regression (Logit).
2. GAM, also with binomial link function and default smoothing parameters.
3. Random Forests. The most important parameters are the number of trees, node size (minimum size of each tree’s terminal nodes), and the maximum number of terminal nodes that trees in the forest can have. Based on initial experimentation with a data sample, the number of trees was set to 500, node size to 20 and max nodes to 10. Normally cross-validation would be used for parameter tuning. In our setting, however, a more appropriate tuning method was to use a forward moving window over whole years, in agreement with our methodology. This was done for two reasons: (a) to avoid training on future data and testing on past data and (b) to avoid splitting each year’s data between datasets, as the current year predictors can only be used to forecast future years.

In this work, these three methods are adopted for combination testing, with results discussed in the Final Results section. These results can then be used to select the best final combination of options.
Dataset construction

In this section, we show details of the dataset construction process that resulted in 13,295 observations (data years) used for model training and testing.

Data pre-processing and merging sources

For country-year pairs, we use the Correlates of War (Ghosn and Bennett 2003) [27] country codes listing of 219 countries. The data sources used for predicting genocide onsets (data years and prediction years) cover years between 1946 and 2017 for 209 countries, and 72 years; therefore, the first task was to create a set of country-year pairs for the 60 countries covered by available data, giving rise, after removing duplicates, to a total of 13,295 pairs. Based on related genocide studies [10, 20] and our domain knowledge, a total of 86 predictor variables were selected. These variables were collected from a variety of sources, plus some “dummy” variables used to include derived values for countries of interest. The main variables are listed in Table 2, and the rest are grouped by the source.

Dependent variables for this work come from the new TMK dataset [10], which flags genocide onsets as a subset of TMK. A value of 1 indicates a positive case (onset occurred), and 0 a negative case. For evaluation, we also added genocide onsets separately for each future year for the next 5 years, from current + 1 to current + 5 as dependent variables. These variables are potential prediction targets. The number and percentage of positive samples for these predictor variables in the dataset are shown below. Further in this section we discuss the selection of a single prediction target.

- onset_1 (genocide onset in current year + 1): 86 positives (0.65%).

Table 2 Summary of predictors

| Variable description                                      | Value type and range | Source                                      |
|----------------------------------------------------------|----------------------|---------------------------------------------|
| Year                                                     | 1946–2017            |                                             |
| Country code                                             | 41–950               |                                             |
| Whether a genocide onset occurred in year/country         | 0 or 1               | New TMK dataset [10]                        |
| Was this perpetrator a non-government actor that was      | 0 or 1               | New TMK dataset [10]                        |
| connected to the state (i.e., state militias)             |                      |                                             |
| Group of variables used for genocide forecasting in      |                      |                                             |
| Goldsmith 2013                                           |                      | [3]                                         |
| Group of variables used for genocide forecasting in      |                      | [8]                                         |
| Goldsmith 2018                                           |                      |                                             |
| Variables related to political violence by year/country   | Numbers              | MEPV [21]                                   |
| Total population of a country in given year               | Number               | World Bank [22]                             |
| Variables related to ethnic population                    | Numbers              |                                             |
| Number of years since last actotal > 0 (a variable from  | Number               |                                             |
| MEPV)                                                    |                      |                                             |
| Whether a genocide occurred before in this country        | 0 or 1               | Dummy variable                              |
• onset_2 (genocide onset in current year + 2): 82 positives (0.62%).
• onset_3 (genocide onset in current year + 3): 79 positives (0.59%).
• onset_4 (genocide onset in current year + 4): 78 positives (0.59%).
• onset_5 (genocide onset in current year + 5): 75 positives (0.56%).

As can be seen, the onset of genocide is (fortunately) an extremely rare event. This small proportion is not uncommon in forecasting of other types of political and societal events. For example, a dataset of irregular leadership changes [24] contains only 0.17% of positive cases. All logical variables were converted to 1/0 values.

Selecting and applying missing values imputation method

Missing data are common in Social Science domains. Gorard [25] identifies three types of missing data:

(i) Completely random loss of data points.
(ii) Biased loss, where cases from one or more categories are not present causing a bias in the dataset.
(iii) All categories are generally represented, but there are some missing attributes or values.

Type (iii) allows full or partial reproduction or prediction of missing values. Somewhat similar types of missing data have been analyzed [26], where three categories are identified: missing completely at random (MCAR), missing at random only in some groups (MAR), and not missing at random (NMAR). In our genocide case study, all three types (i) to (iii) occur; however, assuming that exhaustive data collection has been performed (i.e., no missing categories), imputation methods can only be applied to types (i) and (iii).

When dealing with missing data, incomplete cases could be ignored, or an attempt could be made to impute missing values. Ignoring cases may substantially reduce the data size and introduce extra bias [25]. In our case, imputation is the most reasonable solution to limit data loss. We identified two sources of missing values: original source datasets contained missing values, or values were missed as a result of data merging. Before considering the application of any imputation method, an effort was made to reduce the number of missing values. In cases where it was logical to do so, these values were filled in manually. For example, missing genocide onset values were set to 0; since all country-year onsets are defined in the ground truth dataset, unspecified values may be set to false.

Imputation is an important, but often overlooked, part of the prediction pipeline. Introducing bias by removing data with missing values, for instance, can skew resulting test set predictions in unknowable ways. Further, cross-validation generally works only when datasets are complete, making imputation of missing values a necessary part of pre-processing. While nearly every forecasting methodology imputes missing data, little is known about the ways in which imputation can affect the final predictions. Given that researchers want to be as accurate as possible in their final estimations, we believe it is important to make transparent this part of the pipeline.
In general, imputation works by predicting missing values using machine learning or statistical methods trained on available values. Since multiple predictors may have missing values, we consider multi-variate imputation. An imputation method can produce single or multiple approximations for each such value. It has been suggested that multiple imputation works better as it better reflects the uncertainty of the missing data [26]. Moreover, this process may create additional training cases, which is valuable when very infrequent positive cases are available. We therefore evaluate three imputation methods available in R: MICE with CART kernel [27], missForest, and Hmisc [29]. Each of them represents a different learning algorithm; therefore, a broader spectrum of available methods is covered.

MICE [28] generates imputations for each variable in turn, treating all other variables as inputs for the CART [27] model. In the next iteration, imputed values become inputs for a different variable. The number of iterations and the number of imputations can be specified as parameters, allowing more than one “version” of imputed data. We use the default recommended values of 5 imputations and 5 iterations, generating 5 “versions” of the base dataset.

missForest, an implementation of Random Forests [19], also proceeds in multiple iterations, calculating the difference between the previous and the new imputed data matrix until a stopping criterion is met. The missForest method in R does not generate multiple imputations per se; therefore, we used MICE with the Random Forests kernel to achieve this goal.

The Hmisc [29] imputation R package uses additive models for predicting missing values and does not inherently make multiple imputations [29]. To achieve multiple imputations, bootstrapping is used to approximate the process of drawing predicted values from a full Bayesian predictive distribution. Different bootstrap resamples are used for each of the multiple imputations.

To test the three imputation methods, namely, MICE, missForest, and Hmisc, we applied these methods to a sample from the merged dataset, using the three prediction algorithms, GLM, GAM, and Random Forests on datasets imputed by the three methods, and averaged the test results. Ideally this would be performed on an exclusive subset of data. Due to the small dataset, however, this was not practical. Therefore, a data sample was considered sufficient for selecting the imputation method. As a target variable, onset_123 (genocide onset occurring in any of the three future years) was used.

As shown in Table 3, the MICE with CART algorithm decisively outperforms the other two methods; therefore, it was selected for further processing.

| Table 3 | Results of testing imputation methods using MAP metric. Columns refer to imputation algorithms and rows refer to forecasting algorithms |
|---------|--------------------------------------------------|
|         | MICE CART     | MICE RF     | Hmisc     |
| GAM     | 0.17          | 0.14        | 0.11      |
| GLM     | 0.17          | 0.14        | 0.11      |
| RF      | 0.14          | 0.11        | 0.1       |
| Average | 0.16          | 0.13        | 0.11      |
Selecting prediction target

The next question to consider is the choice of the number of years to forecast. Here we compare the MAP values separately for each year from 1 to 5 years after the current data year. We used the GLM method and compared the two windowing processes (discussed in subsequent sections). Similar results, however, were observed when using the other two forecasting algorithms.

In Fig. 3, the first two years show steady performance, whereas in the next two years the metric values drop substantially. In Year + 5 an increase in MAP values is observed again. When looking at the vertical black lines on top of bars, depicting relative standard deviations over prediction years (from 1966 to 2017), these values tend to get larger with the number of future years, for example, when comparing Year + 1 with Year + 5. There are greater variations in the quality of predictions; hence, the predictions are less stable. Consequently, it makes sense to set the prediction target for at least the first and the second year, and therefore, we continue further analysis with the three-year prediction as a target variable (onset_123), which makes 1.86% of the total dataset.

Selecting class balancing methods

With a relatively small and highly imbalanced dataset, with less than 2% of genocide onsets, the use of up-sampling methods to increase the number of positive class instances was considered. Since any class balancing distorts the original data, the no class balancing option was also tested. Unlike the imputation methods, our study of forecasting genocide onsets includes class balancing in the automatic process; therefore, no prior selection is needed. The results are part of the final result set. Four methods were used:

- No balancing.
- Simple up-sampling (provided in R): randomly samples (with replacement) the minority class to be the same size as the majority class.

![Fig. 3 MAP values of GLM predictions for each year from 1 to 5](image)
• SMOTE (Synthetic Minority Over-sampling) [30]: generates synthetic instances that are close neighbors to the actual positive instance. For each $i$th neighbor of a given instance, where $i < k$, SMOTE generates a new instance that lies between the given instance and its $i$th neighbor in feature space. The default value $k = 5$ is used.

• ROSE [31]: this method also generates synthetic examples from the neighborhood of the instance. Theoretically, the sample generation is for both classes, with probability inversely proportional to the class probability. Effectively, with such class imbalance, ROSE oversamples the minority class creating a new dataset with balanced classes. Compared to SMOTE, another difference is that ROSE uses smoothed bootstrapping to draw artificial samples from feature space. Briefly, the distribution of class samples is smoothed by summing kernel distributions centered at each class sample. This allows for smooth generation of artificial samples, creating balanced class distribution.

Although Menardi and Torelli’s [31] tests show the superiority of ROSE over SMOTE, in this work we test all methods.

**Selection of predictors based on variable importance**

It is not unusual that the dataset for this kind of prediction problem may contain hundreds of predictors, many of them added under the assumption that they may supply valuable information for model training. There are 86 predictors for the genocide onset problem discussed in this paper. The performance of some prediction methods may, however, be negatively affected by unimportant variables.

Predictor selection is an important topic on its own and may be achieved in several ways, e.g., grid search (full search using all combination of variables) or greedy search that uses heuristics to optimize decisions at each stage, or statistical methods. In our case study, variable selection is performed inside the outer training/test loop (discussed in “Selection of windowing technique”). This allows for dynamic selection of predictors as their importance may gradually change over years. The model is trained on the full set of predictors and a set of scores, one for each predictor, is obtained from the model. These scores are then scaled, and predictors below the mean are excluded. The second round of the training is performed using the reduced set of predictors. The mean method was selected as best performing in our case study after testing a number of different thresholds, including elimination of the lower 25 and 50 percent of variables. Other methods of computing importance scores for predictor selection, however, may be more appropriate in other case studies.

For linear models, (GAM and GLM), the variable importance score is the absolute value of the t-statistic for each variable. In our study, the Random Forests model does not use variable selection, as feature sampling is already part of the algorithm.
Selection of windowing technique

When selecting the training/forecasting process, the usual methods to consider are cross-validation (CV), training/tuning dataset split, moving window, and presenting the data in random order. The genocide onset forecasting problem and dataset have their own characteristics that determine the suitability of any training and testing method. The most important one is that the data are temporal in nature, with a number of country-year cases in each year. Some of the predictors in a given year are country specific, but others pertain to more than one country, e.g., relative conditions in neighboring countries. This precludes the use of randomization and cross-validation methods. Training/test split in chronological order might be possible; however, we advise against it for two reasons. Firstly, political and societal conditions and context may change substantially over time; therefore, testing on too many years in the future may not be appropriate. The second reason is that this methodology is developed for real life applications, where training data are available up to the current year and the forecasting is made for the subsequent 1–3 years. Therefore, windowing techniques are considered the most suitable. The model is trained on the past \( n \) years and events are predicted in next (small) \( k \) years. Then the window is shifted forward by one year and the process is repeated.

Within this windowing framework, some choices are still available.

(i) The moving window is of fixed size (Fixed Window).
(ii) The moving window is expanded by one year at each iteration (Expanding Window).

The fixed window may better capture changes in the data, but it limits an already small dataset available for training. In our case study we decided to use both methods and compare the results.

Next, these methods are described in turn as they are applied to the genocide onset forecasting.

In Fig. 4, the training and testing method with a single moving Fixed Window (external dark blue rectangle) of size 20 years is shown, moving forward one year at a time. The window is divided into three parts from right to left: data year used to predict the onsets in prediction years (dotted light gray), 5 test and 14 training years for model tuning (dotted pink and continuous pink, respectively. Note that the test data year becomes one of the prediction years for the model tuning part). Three prediction years, where an onset may occur in any year, are shown in the dotted light gray rectangle.

The training/testing proceeds as follows:

1. For a given window of 20 data years, apply MICE imputation.
2. Training data are split into 5 test and 14 training years,
3. Training/testing for tuning is performed many times, each time with different model parameters and/or configuration,
4. The best set of parameters is selected as measured by the MAP metric.
The model with the best set of parameters is trained using 19 training data years and tested using the single test data year. The model then predicts onsets in the prediction years (light gray rectangle).

The window moves one year forward, and the process is repeated.

The final result is a list of MAP measurements, one for each test data year from 1965 to 2016.

The training/testing method for the Expanding Window is similar, except that the start year is fixed (1946), while the end year moves forward by one year in each iteration.

To avoid overfitting, an adjustment made to both the above scenarios is to remove ground truth from training and testing inside the loop, if the last data year is less than three years from the prediction years. For example, in Fig. 5, when the data year becomes 1966, the prediction year is only 1967. Ground truth for 1968 and 1969 is removed, as these are prediction years for data test year 1967.

**Post-processing**

The goal of post-processing is to prepare the test results for analysis or presentation to the user. The post-processing methods depend on the outputs generated by the models and the intended use of the predictions. We aim to optimize the number of predictions to reflect the relative cost of false positives and false negatives. The method is described in detail in “Model training and forecasting events.”
Summary of the FORTE methodology steps

The FORTE methodology described above is summarized in Fig. 5. Ideally all combinations would be tested separately. Due to a large number of combinations, several prior decisions were made with regard to selecting the options, resulting in accepting some choices (green ticks) and rejecting others (red crosses).

Several selections were not made explicitly, but left to final testing and comparison, symbolized by the multiple document shapes in the right part of the diagram.

Analysis of final results and selection of forecast model

Next, we discuss the final results specific to our case study. The results for the 24 remaining combinations, enclosed in the blue, dotted rectangle in Fig. 5, are summarized in Table 4, noting that variable selection, for reasons stated previously, was not applied to Random Forests. The bottom part of the table shows averages for each combination and the overall average for the Fixed and Expanding window options. It might be useful to know that the default MAP is 0.0354, obtained by completely random assignment of the same number of true positives. Generally, linear models produce better results than Random Forests. For all models, the fixed and expanding window options are comparable.

There are, however, a few more interesting differences at the methods level. Class balancing substantially increases the MAP values for Random Forests, which
could be due to the way that the classification trees are constructed. If the number of instances in the positive class is too small, the whole branch for that class may not be constructed due to the leaf threshold, resulting in greater misclassification. For linear prediction methods, class balancing produces mixed results, with the best overall MAP obtained by Fix Win GAM without class balancing (Model 2), whereas the second-best MAP was produced by Exp Win GLM with class balancing (Model 7). Variable selection generally works well with and without class balancing, as can be seen in the average rows of the table.

The summary in the second part of Table 4 also shows that class balancing, except for Random Forests, did not contribute to any increase in model performance and that variable selection generally has a positive effect with and without class balancing. While these observations cannot be generalized and are open to further research, we can point out a couple of reasons for this behavior. The first one is related to the FORTE methodology, where class balancing is performed locally in the incremental setting, as opposed to balancing the whole training dataset, which is commonly used in Machine Learning. The second reason could be the skewness and irregularity in the dataset itself, where each incremental window captures conflicting data trends.

Given the overall small MAP differences shown in Table 4, it could be informative to show an example of how a difference in the MAP metric is reflected in genocide case ranking. In 1974 there were 6 genocide cases, of which Model 2 (GAM with Fixed Window, No Class Balancing and Variable Selection) predicted three in the first 15 positions, ranked 1, 2, and 5. MAP for the first 15 positions, which

| Method | Class balancing | Var selection | Fixed win MAP | Exp Win MAP |
|--------|-----------------|---------------|---------------|-------------|
| 1. GAM | ✗               | ✗             | 0.19          | 0.18        |
| 2. GAM | ✗               | ✓             | 0.22          | 0.19        |
| 3. GAM | ✓               | ✗             | 0.18          | 0.17        |
| 4. GAM | ✓               | ✓             | 0.19          | 0.19        |
| 5. GLM | ✓               | ✓             | 0.19          | 0.18        |
| 6. GLM | ✓               | ✓             | 0.18          | 0.20        |
| 7. GLM | ✓               | ✗             | 0.18          | 0.21        |
| 8. GLM | ✓               | ✓             | 0.19          | 0.19        |
| 9. RF  | ✓               | N/A           | 0.12          | 0.12        |
| 10. RF | ✓               | N/A           | 0.16          | 0.19        |
| 11. RF | ✓               | N/A           | 0.16          | 0.19        |
| 12. RF | ✓               | N/A           | 0.16          | 0.19        |
| Average| ✗               | ✓             | 0.17          | 0.18        |
| Average| ✓               | ✓             | 0.20          | 0.19        |
| Average| ✓               | ✗             | 0.18          | 0.19        |
| average| ✓               | ✓             | 0.19          | 0.19        |
| Average overall | ✓  | ✓   | 0.18  | 0.18        |
is commonly used in predictions (e.g., forecasting report [15]) MAP$_{15} = 0.17$. If another correct prediction was made at position 15, the MAP would increase to 0.19. If that additional correct prediction were ranked 10, the MAP would be 0.2. The MAP metric is therefore sensitive to the rank of the correct prediction, as illustrated in “Selection of metrics.”

The final model for training and forecasting events can be selected from the above analysis. If these test results indicate a number of equally good models, these models can be used in an ensemble by averaging their MAP scores and ordering the resulting score to choose the top n predictions. In our case study of genocide onset predictions, we will use Model 2 from Table 4 for further analysis.

**Optimizing number of predictions**

Depending on the application of the forecasting model, the desirable number of predictions may vary. This section presents an analysis of predictive model flexibility in adjusting the number of optimal predictions.

**Analysis of precision and recall**

The precision and recall versus the number of predictions charts are shown in Fig. 6. These charts are made using Model 2 (Table 4). The right-hand chart is a magnified part of the curve for the first 20 predictions. The top 10 predictions correspond to 34% of true cases (recall = 0.34). For the top prediction, the model produces a maximal precision of 0.25 but with very low recall, which is usually too low for practical applications.

These charts also show that by varying the number of predictions, it is possible to obtain a desired recall with precision set to some minimum value, which could be very useful in practical prediction scenarios. Increasing the prediction set (containing the number of predicted genocide events), however, does not proportionally increase the number of true cases as the recall curve flattens. The question remains,

| Position on FORTE prediction list | Genocide event |
|----------------------------------|----------------|
| 2 in 1980                        | Lebanese Civil War: Lebanon-Right-wing Christian militias-1982 |
| 1 in 1981                        | OR War of the Mountain: Lebanon-People’s Liberation Army-1983 |
| 1 in 1982                        | Sri Lankan Civil War: Sri Lanka-LTTE-1989 |
| 4 in 1986                        | Yezidi genocide and other events: Iraq-IS-2014 |
| 1 in 1987                        | |
| 6 in 1988                        | |
| 2 in 2011                        | |
| 1 in 2012                        | |
| 7 in 2013                        | |
however, as to how many such predictions should be generated in order to optimize analysis by an end user. A solution to this problem is proposed in the next section.

Calculating optimal number of predictions

In the precision/recall curves it was observed that the forecasting system can be flexible: fewer predictions can decrease the number of false positives at the cost of losing some true predictions, and vice versa. A natural question would be about an optimal point on the curve. To answer this question, in this section, we perform further analysis.

We view the system as recommending a certain number of genocide onset cases for further analysis. The system is not perfect; only some predicted cases are true positives ($tp$); the rest are false positives ($fp$). The window of shown cases may also miss some true predictions and count them as false negatives ($fn$). The expectation is to find a certain number of positive cases ($tp$) without missing too many of them ($fn$). We can formulate this as a relative cost associated with $fp$ and $fn$.

In order to formalize this analysis, we introduce the concept of cost assigned to false positives and false negatives, both of which we are trying to minimize. Since, however, decreasing one of them usually increases the other, the cost value represents the relative importance between them. Let the cost of missing an event be $c_{fn}$, and the cost of false alarm be $c_{fp}$. In practice, it is very likely that $c_{fn}$ is much higher than $c_{tp}$. As an example, Kennedy [32] suggests that the acceptable cost of a false positive is no more than 7.7% of the cost of a prediction failure. This can be interpreted as $c_{fp} \leq 7.7\% \times c_{fn}$ or $13 \times c_{fp} \leq c_{fn}$; hence, the cost ratio $c_r = c_{fn}/c_{tp} \geq 13$; therefore, the minimal value of $c_r = 13$. The minimal relative total cost of $n$ predictions is then $C = fp + c_r \times fn$. In this equation, we do not consider the cost associated with desirable true predictions $tp$. For simplicity, we assume that the unit costs $c_{fn}$ and $c_{fp}$ are the same for all genocide onset events.

The next step is to consider the effect of the prediction model on cost $c$. The model built by training the prediction algorithm is expected to correctly deliver as many true positives as possible in the top $n$ predictions. Furthermore, it is also expected that there will be many more true cases in the top $k$ than in the remaining $n-k$ forecasts, where $k \ll n-k$. In other words, the $tp$ distribution over $n$
predictions should be skewed to the right (with a right long tail). A uniform distribution would indicate a random assignment of true positives.

The distribution of genocide onsets for GAM Fixed Window (Model 2) is shown in Fig. 7, with the moving average trendline. This distribution is directly related to the recall chart shown in Fig. 6; the recall at \( n \) predictions in these charts is just the cumulative number of true positives in the chart above divided by the real number of genocide onsets in the test. Similar plots were shown in other work [11] as separation plots when models for predicting civil war onsets were compared.

The number of true positives is an unknown function \( m \) of the shown predictions \( n \), \( tp = m(n) \) which can be calculated from the prediction model. Since \( n = tp + fp \), \( fp = n - m(n) \). The number of false negatives \( fn = g - tp = g - m(n) \), where \( g \) is the real number of genocide onsets in the test set over all test years. Replacing \( fn \) and \( fp \) with these expressions, we obtain

\[
C = c_r \times (g - m(n)) + (n - m(n))
\]  

Hence,

\[
C = g \times c_r - m(n) \times (c_r + 1) + n
\]  

Equation (3) is the total relative cost \( C \) for forecasts over all test years obtained by inserting the number of cumulative true positives versus \( n \) into Eq. (2). An average cost per year \( c \) can be obtained by dividing the total cost \( C \) by number of forecast years. This is a more meaningful value, since in practice forecasts for future years are produced in the current year in a single prediction step, yielding a yearly forecast. Charts of these cost values are shown in Fig. 8 for different values of \( c_r \).

These charts clearly show that the cost functions do have their lowest points (indicated by markers), although the curves are relatively flat around the minima. There are, however, differences between different cost ratios \( c_r = 5, 10, 20, \) and \( 30 \); for a higher cost ratio the minimum is at a higher \( n \). The flat line around the minimum means that for a wide range of \( n \) the cost is close to optimal. Visually, this is around 1–5 for cost 5, 2–10 for cost 10, and so on. Since this is an average cost over all forecast years, there may be scope to further optimize the number of shown predictions by additional model tuning. In simple terms, however, these results show

---

Fig. 7 Histogram of true positives for GAM with Fixed Window (Model 2)
that it is possible to find a range of optimal recommendation window sizes under some assumptions about the relative cost of misclassification.

**Model training and forecasting events**

In the final step of the FORTE methodology, the best model, or models, is trained to make forecasts. In our case of genocide forecasting we trained Model 2 (Table 4) on a number of subsets of data years, and predictions are listed for the corresponding three future test years. For example, if Model 2 was trained on years 1979–1998, predictions are made for 1999–2001.

In Table 5, three examples of successful genocide predictions from various parts of the world are listed. Following Kennedy [32], we used cost ratio $c_r = 13$; hence, according to Fig. 8, the number of forecasts is limited to top 8. In the first column of Table 5, the number indicates the rank of the event in the prediction set. This demonstrates that our forecasting methodology is correctly identifying likely onsets of genocide some years in advance. This early warning may be useful to practitioners who seek to draw international attention to such horrific conflicts.

**Conclusion and future work**

In this paper, we have taken a novel approach to genocide forecasting by focusing on the development and testing of a methodology (FORTE) that can be used to generate predictive models for forecasting rare political and societal events in real time. The methodology is analyzed by discussing and testing available options at each stage and finally displaying the results of predictions for several combinations of the building blocks. An important part of the FORTE methodology is its practical aspect. We built, trained, and analyzed several models that show sufficient flexibility to adjust to user preferences in terms of balancing the number of false positives and false negatives. We introduced an evaluation metric to measure the usefulness of predictions, when the number of predicted events is not known in advance. We show that an optimal range for the number of recommended predictions can be achieved.

![Relative cost of yearly forecasts vs number of predictions and cost ratio](image-url)
We suggest that this relatively flat range of values can be used for further model tuning.

In our analysis, we limited the choices of variable selection, model selection, and parameter tuning, instead focusing on building a general framework for future expansion and analysis. Both the range of variables and models can be expanded, considering the most successful settings used for similar problems in the literature. By comparing fixed versus expanding training windows, we could not find any general trend that supports one or the other; however, there are differences between individual models that may be explained by further analysis. A practical option to consider would be a solution that uses expanding window with limited size.

The latest machine learning models can be used for prediction. Recent rapid progress in deep learning research suggests that such models should be evaluated.

**Acknowledgements** The authors thank Charles Butcher and Sascha Nanlohy for valuable assistance in data collection and advice.

**Funding** Open Access funding enabled and organized by CAUL and its Member Institutions. The author(s) disclosed the following funding for the research, authorship, and/or publication of this article: Australian Research Council (Grant ID: DP160101122).

**Declarations**

**Conflict of interest** The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Data availability** All datasets used in this paper are publically available. The sources of data are listed or referenced in Table 2 of the paper.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

**References**

1. Hegre, H., Metternich, N. W., Nygård, H. M., & Wucherpfennig, J. (2017). Introduction: Forecasting in peace research. *Journal of Peace Research, 54*(2), 113–124.
2. D’Orazio, V., Honaker, J., Prasady, R., and Shoemate, M.: Modeling and forecasting armed conflict: AutoML with human-guided machine learning, In 2019 IEEE International Conference on Big Data (Big Data), pp. 4714–4723.
3. Goldsmith, B. E., Butcher, C. R., Semenovich, D., & Sowmya, A. (2013). Forecasting the onset of genocide and politicide: Annual out-of-sample forecasts on a global dataset, 1988–2003. *Journal of Peace Research, 50*(4), 437–452.
4. [https://earlywarningproject.ushmm.org/](https://earlywarningproject.ushmm.org/). Accessed 12 Aug 2022
5. Goldstone, J. A., Bates, R. H., Epstein, D. L., Gurr, T. R., Lustik, M. B., Marshall, M. G., Ulfelder, J., & Woodward, M. (2010). A global model for forecasting political instability. *American Journal of Political Science, 54*(1), 190–208.
6. Beger, A., Dorff, C. L., & Ward, M. D. (2014). Ensemble forecasting of irregular leadership change. *Research & Politics, 1*(3), 2053168014557511.

7. Halkia, M., Ferri, S., Schellens, M. K., Papazoglou, M., & Thomakos, D. (2020). The global conflict risk index: A quantitative tool for policy support on conflict prevention. *Progress in Disaster Science, 6*, 100069.

8. Goldsmith, B. E., & Butcher, C. (2018). Genocide forecasting: Past accuracy and new forecasts to 2020. *Journal of Genocide Research, 20*(1), 90–107.

9. Shmueli, G. (2010). To explain or to predict? *Statistical science, 25*(3), 289–310.

10. Butcher, C., Goldsmith, B. E., Nanlohy, S., Sowmya, A., & Muchlinski, D. (2020). Introducing the targeted mass killing data set for the study and forecasting of mass atrocities. *Journal of Conflict Resolution*. https://doi.org/10.1177/0022002719896405

11. Muchlinski, D., Siroky, D., He, J., & Kocher, M. (2016). Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis*. https://doi.org/10.1093/pan/mpv024

12. Gelman, A., & Loken, E. (2016). The statistical crisis in science. In M. Pitici (Ed.), *The best writing on mathematics* (pp. 305–318). Princeton University Press.

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

14. Koyejo, O., Natarajan, N., Ravikumar, P., and Dhillon, I. S. (2014). Consistent binary classification with generalized performance metrics. *Advances in neural information processing systems, 3*, 2744–2752.

15. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS One, 10*(3), e0118432.

16. Davis, J., and Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on machine learning, pp. 233–240.

17. Dobson, A. J., & Barnett, A. G. (2018). *An introduction to generalized linear models*. CRC Press.

18. Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical Science, 1*(3), 297–310.

19. Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5–32.

20. Butcher, C., & Goldsmith, B. E. (2017). Elections, ethnicity, and political instability. *Comparative Political Studies, 50*(10), 1390–1419.

21. https://www.systemicpeace.org/inscrdata.html. Accessed 12 Aug 2022

22. Ghosn, F., & Bennett, S. (2003). *Codebook for the dyadic militarized interstate dispute data, version 3.10*. The Pennsylvania State University.

23. Cederman, L.-E., Wimmer, A., & Min, B. (2010). Why do ethnic groups rebel? New data and analysis. *World Politics, 62*(1), 87–119.

24. Ward, M. D., & Beger, A. (2017). Lessons from near real-time forecasting of irregular leadership changes. *Journal of Peace Research, 54*(2), 141–156.

25. Gorard, S. (2020). Handling missing data in numeric analyses. *International Journal of Social Research Methodology, 23*(6), 651–660.

26. Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC Press.

27. Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC Press.

28. Van Buuren, S., & Groothuis-Oudshoorn, K. (2010). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v045.i03

29. Alzola, C., and Harrell, F. (2006). An introduction to S and the Hmisc and design libraries. http://biostat.mc.vanderbilt.edu/wiki/pub/Main/RSSintro.pdf.

30. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of artificial intelligence research, 16*, 321–357.

31. Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery, 28*(1), 92–122.

32. Kennedy, R. (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.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.