Work in progress, please do not cite.

Early Warning for Forced Displacement

Geraldine Henningsen[1]*, Giulia Del Panta[1]

Abstract
Anticipatory action holds the potential to increase efficiency and timeliness of the humanitarian response in crises. Nevertheless, successfully exploiting the potential of anticipatory action requires the early–and especially reliable–detection of future crisis events. To accommodate the need for early event detection, multiple early warning initiatives are currently sprouting in the humanitarian sector. For example, early warning initiatives in natural hazards (e.g., OCHA) and conflict (e.g., ViEWS-Uppsala/PRIO) are increasingly able to accurately predict future events with the help of quantitative modeling.
However, neither the prediction of natural hazard events nor of conflict is sufficient to predict significant displacement movements within countries of origin or across their borders. Such events usually contribute to forced displacement, but the environment in which they occur impacts their overall effect on the magnitude of forced displacement. As such, they are too unreliable as stand-alone predictors of forced displacement flows.
This report presents an early warning model for the forced displacement of both IDPs and refugees/asylum seekers, combining predicted conflict events with various other environmental variables. Although still at an early stage, our analysis shows that a country's risk for significant displacement flows can be predicted with high accuracy for a 12 months prediction horizon.

[1] *Statistics and Demographics Section, Global Data Service, UNHCR*
***Corresponding author**: hennings@unhcr.org

Contents

Introduction

Anticipatory action, or the proactive planning and preparation for future events, has the potential to significantly improve the efficiency and timeliness of humanitarian responses in times of crisis. However, the success of anticipatory action relies heavily on the ability to accurately detect and predict future crisis events, especially by reducing the risk of a false alarm. Various early warning initiatives in the humanitarian sector have emerged in recent years, such as those focused on natural hazards and conflict, utilising quantitative modelling to improve their predictions.

Whilst these initiatives have made progress in predicting natural hazard events and conflicts, they have not been sufficient in predicting significant displacement movements within or across borders. The impact of conflict and natural hazard events on forced displacement, is heavily influenced by the environment in which they occur, making them unreliable as standalone predictors of forced displacement flows.

This report presents an early warning model for predicting the forced displacement of internally displaced persons (IDPs) and refugees/asylum seekers to address this issue. The model combines predicted conflict events with various environmental variables to provide a more comprehensive prediction of displacement risk. Using a gradient boosting machine algorithm for a classification model which predicts the probability of a country producing significant displacement flows within the next 12 months, we derive a country-specific risk index as an indicator of how likely forced displacement within and across borders is likely to occur. Initial analysis suggests that this model can predict a country's risk for significant refugee/asylum seeker flows for a 12-month prediction horizon with high accuracy. The model performs less well for IDP flows but still performs significantly better than a naive model. Although still in its early stages, this model holds promise for improving the support for anticipatory action in the humanitarian sector through quantitative modelling.

Users should note, though, that as with every estimation, also, this risk index comes with a certain degree of uncertainty. This uncertainty means that even very high risk indexes do not guarantee forced displacement, likewise, low index values do

not guarantee that no forced displacement will happen within the next 12 months. The index is merely an indicator of the probability of such an event. As such, the presented risk index will always remain a supporting tool for decision-makers who plan for anticipatory action.

Data

Size and structure of data set The current status of this work is a data set based on country-year observations, where a country is depicting the country of origin of forced displacement. The data set comprises $n = 178$ and $t = 13$ (from 2009:2022). The panel data set is slightly unbalanced, with 158 countries entering the entire 14 years, which results in a total of 2,537 observations.

Dependent variables The models' dependent variables are aggregated yearly refugee and asylum seeker flows from the country of origin and aggregated IDP flows from conflict, respectively. Whilst figures on refugee and asylum seeker flows have been produced by UNHCR, we use IDP figures published by IDMC.

Our risk model is interested in detecting two movements in the flow data:

1. Significant flows above a certain threshold, currently set to 2000 persons per year, as this is considered a significant outflow of refugees by UNHCR.

2. Sudden increases in flows above the above the threshold. These sudden increases can happen in both ongoing situations and new situations. It is important to note, that we only include sudden increases that go beyond the threshold, i.e., even if a country experiences a sudden spike in flows starting from a low value, we do not include this change unless it surpasses a certain magnitude, in the current set-up 2000 persons per year.

In order to identify sudden and significant changes in the time series of refugee and IDP flows, we follow the approach outlined in [1] and use change point detection based on significant changes in the mean and variance of the time series. We apply the 'changepoint' package [2] on all country individual time series for both refugees and IDPs to identify years with a sudden change in refugee/IDP flows (see examples for changepoints in the time series of Afghanistan, Iraq, and Syria in figure 1).

Combining threshold and change points for each time series leads to six different scenarios for a given point in time t.

Table 1. Possible scenarios for flow time series at point t

	Change point		
Threshold	Up	None	Down
Above	1	2	3
Below	4	5	6

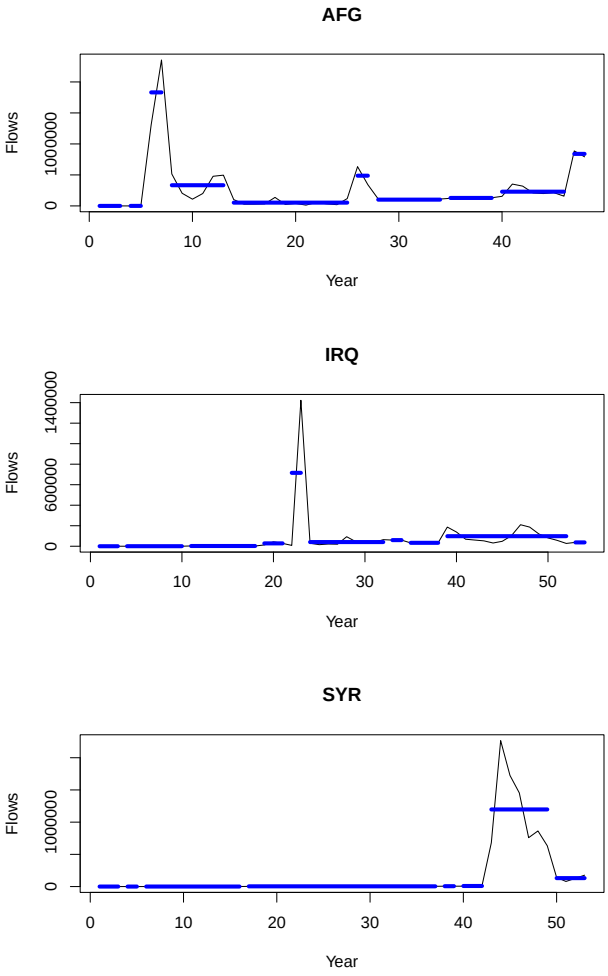

Figure 1. Changepoints in the time series of refugee flows for Afghanistan, Iraq, and Syria

However, from the perspective of an early warning model only scenarios 1: 'a sudden increase in flow above the threshold' and 2: 'a constant flow above the threshold' are of interest, which means that the remaining four scenarios can be grouped together with other scenarios. This leads to three final classes for the dependent variable at a given point in time t:

Class 1: Upwards changepoint above the threshold (scenario 1)

Class 2: No upward changepoint but above the threshold (scenarios 2 and 3)

Class 3: Everything else below the threshold (scenarios 4, 5, and 6)

Categorising refugee/asylum seeker and IDP flows into these three classes leads, unfortunately, to very unbalanced classes (see figure 2). As sudden and significant increases in the flow time series are a rather seldom event, it is unsurprising that the proportion of observations that fall into class 1 is notably lower than the other two classes. We further dis-

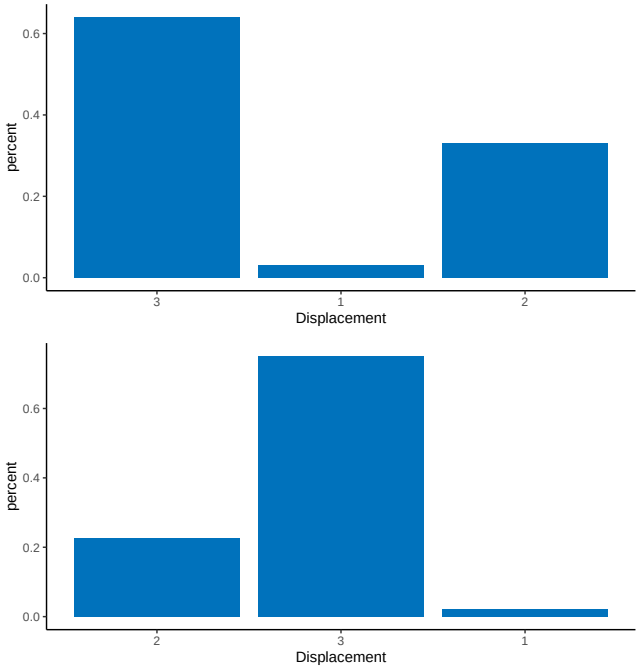

Figure 2. Proportion of observations in each class - Refugees/Asylum seekers and IDPs

cuss the impact of these significant class imbalances in the methodology section.

Feature variables We use 95 feature variables in our analysis. These feature variables cover a wide array of environmental factors and predictions for conflict events [3]. Variables covering economic, demographic, and geographic indicators have been provided by The Global Economy. Variables on natural hazard events, the financial damage they have caused, and the number of people affected by them has been derived from EM-DAT.

Missing values Missing values in the data set are imputed through multiple imputation using the R package Amelia II [4], which contrary to other imputation packages does take the longitudinal structure of the data into account. Multiple imputation uses a Bayesian approach to derive the most likely value for a missing cell, given the distribution of the other observations of the same variable and the distribution of the variable conditional on other variables in the data set. Based on conditional posterior distributions that are then derived for each missing cell, multiple draws from this conditional posterior are made for each missing cell, resulting in multiple versions of the data set. In our work, we generated five different data sets, i.e., each cell was imputed by five different draws from its respective conditional posterior.

The subsequent predictive analysis was conducted with each of the five data sets, whereafter the row mean was calculated across all five predictions. This approach has shown to be far superior to simpler imputation approaches, e.g., mean imputation, which artificially reduces the variance in the data and neglects and falsifies the correlation between variables, or row-wise deletion due to missingness, which can create significant bias in the data [5].

Method

We estimate a multiclass classification model with 1 = a sudden uptick in flow that lies above the threshold, 2 = a constant flow above the threshold, and 3 = no significant displacement for both refugees/ asylum seekers and IDPs using a gradient-boosting machine algorithm. We train the models with all feature variables–except for conflict prediction variables–lagged by one year. By training the models with lagged feature variables, we can run the predictions for the next year based on the known values of these variables.

Although this approach works for slow-moving environmental variables, it would be a poor fit for conflict variables whose impact on forced displacement is more immediate. We, therefore, use three unlagged variables for conflict predictions: armed conflict, other forms of violence, and fatalities from conflict and violence [3], i.e., the year of conflict prediction aligns with the year of displacement prediction. We would like to use the same approach for natural hazards. However, have yet to find a good global prediction model for natural hazards.

We use the package h2o [6], which is an R interface for the 'H2O' scalable machine learning platform, to run the gradient boosting machine. We perform a grid search over a 5×162 hyper-parameter space from which we randomly choose 80 hyper-parameter combinations to approach the most optimal model.

Because of the severe imbalance of the three classes (see figure 2) we use package 'UBL' [7] to re-sample our data using the SMOTE oversampling technique which creates synthetic observations for the minority classes–in our case classes 1 and 2. We, furthermore, under-sample the majority class, so that we end up with three classes of approximately the same size. To account for the multiple classes of our dependent variable and to further ensure that any imbalance in the classes does not impact the classification quality by favouring the majority class, we use mean-per-class-error as the optimisation metric.

We use 5k cross-validation as a validation set and a moving time window of length $9:1$ to partition the data set into a training and a testing data set that takes the temporal dependency of the data into account, which results in four training-testing combinations.

We apply this approach for each imputed data set, finding the best model for each imputation instance. We subsequently use those models to create five predictive values for each country. To calculate the final predicted risk index for each country we take the mean across all five predictive values which generates an unbiased estimator of the true prediction mean but also lets us assess the impact of the imputation on the results though the mean's variance [8].

Results and Discussion

We test the model described in the previous section for four moving time windows of length 9:1 years. The model results vary only slightly over all four time windows, i.e., they remain relatively stable irrespective of which data set we apply. In the following sections, we will, therefore, only present the results of the time window 2011-2020:2021 as a representative example.

Metrics

As the classes are in both cases highly unbalanced and we are operating in a multiclass setting, we use 'mean per-class error' as a metric to evaluate model quality.

Table 2. Confusion matrix test data - refugee estimation

	Predicted			
	1	2	3	Error
1	0	7	0	1.00
2	2	63	9	0.15
3	0	4	102	0.04
Totals	2	74	111	0.12

Table 3. Metrics training data set - refugee estimation

Metric	Value
Accuracy	0.93
Mean per-class error	0.07
LogLoss	0.20
R^2	0.92

Table 4. Metrics test data set - refugee estimation

Metric	Value
Accuracy	0.88
Mean per-class error	0.40
LogLoss	0.35
R^2	0.71

Tables 2–7 show that both models, although producing high accuracy and a low 'mean per-class error' on the re-balanced test data, struggle to correctly predict the minority class 1 in the test data set. A challenge for the evaluation process, are the very low counts for class 1 in the test data in both the refugee/asylum seeker and IDP case. However, the metrics show clear signs of model overfitting, most probably as a result from the considerably oversampling of the minority class to alleviate severe class imbalances, and might explain some of the lower performance regarding class 1. To overcome this problem, the class thresholds could be reconfigured to improve classification accuracy. But as our primary interest lies in the predicted risk probabilities and not necessarily in the correct classification of each observation, we have refrained from adding this correction.

Despite of these shortcomings and given a class distribution of 3:48:49 per cent between the three classes in the case of the refugee/asylum seeker model and 3:24:73 per cent in the IDP model, the accuracy of respectively 0.88 and 0.82 shows that both models still performs better than a simple benchmark model that only predicts the majority class. What is important to note, though, is that the errors across all three classes remain relatively balanced, indicating that the models are as prone to generate a type I as to generate a type II error.

Table 5. Confusion matrix test data - IDP estimation

	Predicted			
	1	2	3	Error
1	0	5	0	1.00
2	0	22	7	0.24
3	2	8	80	0.11
Totals	2	35	87	0.18

Table 6. Metrics training data set - IDP estimation

Metric	Value
Accuracy	0.93
Mean per-class error	0.06
LogLoss	0.19
R^2	0.92

Table 7. Metrics test data set - IDP estimation

Metric	Value
Accuracy	0.82
Mean per-class error	0.45
LogLoss	0.56
R^2	0.49

Variable Importance

This analysis is no inference study, and the impact of feature variables on the displacement classes should, therefore, only be seen as indicative. Also, the subset and ranking of most the impactful variables varies from estimation to estimation.

Figures 3 and 4 show the importance of the top 10 most important variables in the refugee/asylum seeker model and IDP model, respectively. Variables in the top 10 group of the refugee/asylum seeker and IDP model can be grouped into three main categories:

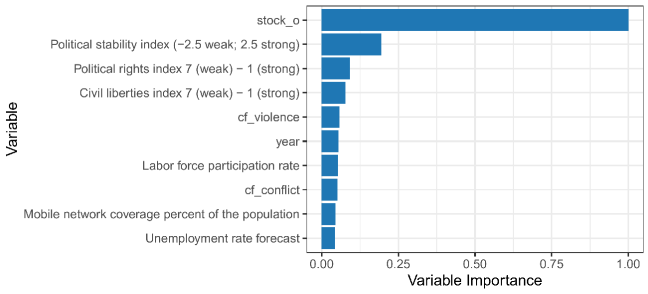

Figure 3. Variable importance - refugee estimation

- variables on forced displacement and migration,
- variables on conflict, violence, and political instability
- variables for economic development.

The variable with by far the most significant impact in the refugee/asylum seeker model is the existing refugee stock from the country of origin. A variable importance of 100% shows a clear impact of the size of the refugee stock on a country's risk index. This finding indicates that the best predictor for future displacement still is past displacement. In the case of the IDP model, both existing IDP and and refugee/asylum seeker stocks have large influences, but to a lesser degree than in the refugee/asylum seeker case. Unsurprisingly, the most dominant variables in the IDP model are all conflict related. Conflict and violence also score high in the refugee/asylum seeker model, ranking fifth and seventh place, respectively. Both clearly indicate that a high conflict/violence risk is linked with a high risk for significant future displacement.

Population size, despite being included as a feature variable, has a surprisingly low impact on the classification probabilities. This might be the result of the relative low threshold figure (2000) that we have set.

Low economic development and political instability, as well as low political freedom, are two other variable categories that show a significant impact on the categorisation of a country, in both models.

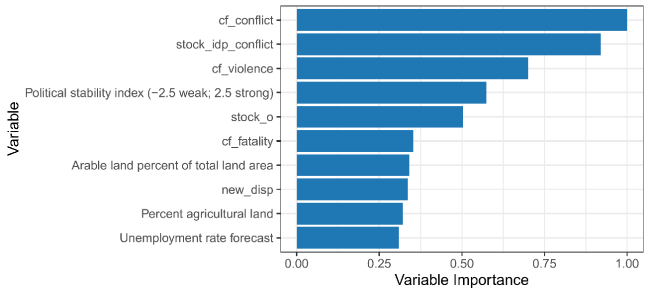

Figure 4. Variable importance - IDP estimation

Generally, one can also observe from figures 3 and 4 that whilst IDP flows are predominately driven by conflict, there is an indication that refugee/asylum seeker flows are more multi-causal being driven by political factors, as well as conflict.

Predicted probabilities

Figures 5, 6, 7, and 8 show the marginal effects of the respective stock variables and political stability indicators on the predicted probabilities of the displacement risk at a specific value of the feature variable. These plots help to gain a deeper understanding of the relationship between a specific feature variable and the predictive value of the outcome variable.

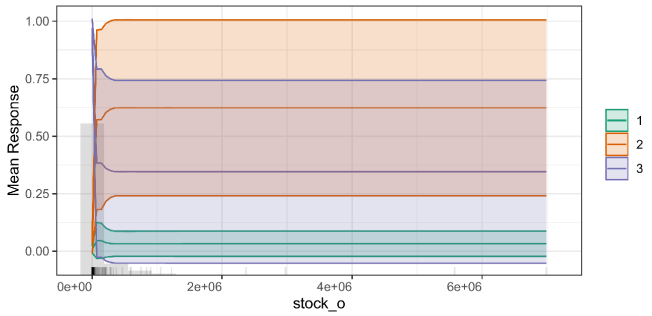

Figure 5. Predicted probabilities refugee stock - refugee estimation

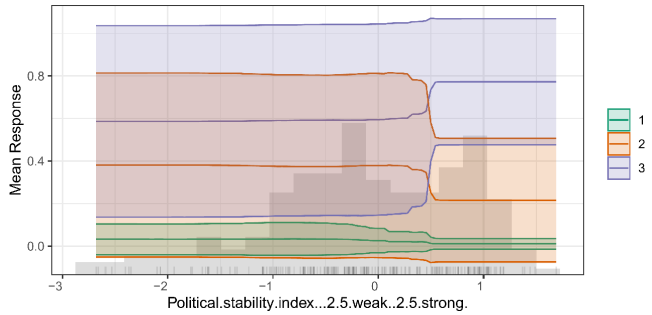

Figure 6. Predicted probabilities civil political satbility (-2.5 low–2.5 high) - refugee estimation

Figures 5 and 7 show that the stock figures' effect on the displacement risk is relatively immediate and impact full both in the case of refugees/asylum seeker flows and IDP flows. Both plots show that the mere presence of a refugee stock–even at low stock values–starkly elevates a country's risk for further future displacement. However, the effect quickly levels off and remains high irrespective of the stock size. Both feature variables function basically as dummy variables, where the absence or presence of the stock is the determining factor rather than the magnitude of the variable.[1]

The impact of existing refugee/asylum seeker and IDP stocks on a sudden increase in their respective flows is reversed. A country of origin's likelihood to experience a sudden spike in flow numbers, though low with zero pre-existing stock figures, experiences a sudden increase once stocks are present but levels of quickly with increasing stock figures. More notably, though, is the decrease in the confidence band

[1]The binary effect of both stock variables on the displacement risk could be a consequence of the classification model that we are estimating. This finding does not necessarily translate to analyses that look at the magnitude of the refugee flow.

around the mean effect that happens as stock figures increase (more pronounced in the case of refugees/asylum seekers). This indicates that although the existence of low stock figures do increase the average likelihood of experiencing a sudden spike in flows, the effect is rather uncertain. As pre-existing stock figures increase further, the likelihood of a country to experience sudden spikes in flow numbers decreases, indicating that sudden changes in flow numbers become less likely in ongoing situations.

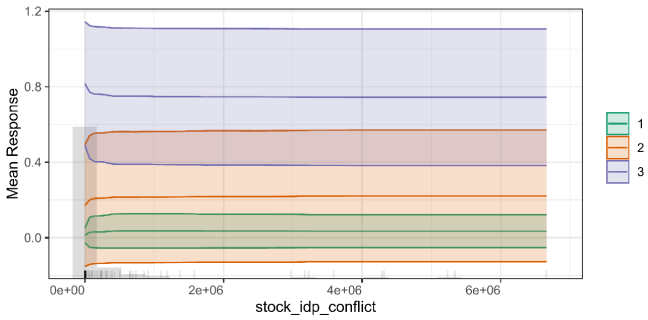

Figure 7. Predicted probabilities IDP stock - IDP estimation

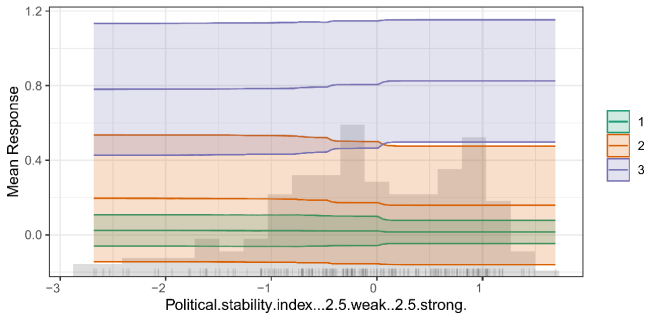

Figure 8. Predicted probabilities political stability - IDP estimation

We can observe a similar but less profound effect for the feature variable 'political stability'.[2] Both figures 6 and 8 show that both indices also can be interpreted as a binary feature variable, with an elevated risk for displacement for low values and a reduced risk for higher values after the breaking point. It seems the effect of civil liberty/political stability is about the same in magnitude for refugees/asylum seekers and IDPs, however, the breaking point for refugees/asylum seekers is much higher than for IDPs. One possible explanation could be that as IDPs are mainly conflict-driven, political instability affects IDPs only indirectly to the point where it causes conflict, whilst the effect of civil liberty on refugees and asylum seekers can be both direct–people fleeing the country from prosecution–and indirect through conflict. This discrepancy is an interesting finding which needs further investigation.

[2]Political stability is an index variable encoded with continuous values between -2.5 and $+2.5$. However, there seems to be a coding error in the data, which has shifted the data by -0.5 to the range of -3 to $+2$. We chose not to correct this error, as it does not impact the results being only a simple linear transformation of the variable.

It is also interesting to note, that both indexes impact the likelihood of a sudden increase in flow numbers in different ways. Whilst increasing political stability decreases the risk of a sudden increases in IDP flows, the feature variable 'Civil liberty' shows a non-linear effect on the likelihood of sudden increases in the flows of refugee/asylum seekers that peaks at mid level and levels of in both directions. Furthermore, like in the case of the stock variables, an increase in the average likelihood is accompanied by a wider confidence band of the marginal effect of the feature variable, i.e., the effect becomes more certain for higher values of the feature variable, most likely because occurrences of sudden spikes in flows at these levels of the feature variable are very rare.

Conclusion

We have presented a simple gradient-boosting classification model to estimate a country's risk of producing significant forced displacement. We developed prediction models for refugees/asylum seekers and IDPs based on a comprehensive data set. Our results show that significant refugee/asylum seeker and IDP flows can be predicted with high accuracy, however, the prediction of sudden increases in flow is more challenging and will require further work. Our model proves that quantitative methods can create valuable contributions to support the planning of anticipatory action within the humanitarian sector.

For future work, we would like to increase our predictions' spatial and timely granularity through the use of sub-yearly and sub-national data. This, though, will require the collection of new variables and the search for proxy variables as an alternative to national indicators that are not readily available at a highly granular sub-national level, e.g., GDP per capita.

That said, the increase in the spatial and temporal granularity will add valuable information about the characteristics of displaced populations and when specific displaced populations will move. This additional information will contribute to improved targeting of anticipatory action in the humanitarian sector.

References

[1] Marcello Carammia, Stefano Iacus, and Teddy Wilkin. Forecasting asylum-related migration flows with machine learning and data at scale.

[2] Rebecca Killick, Kaylea Haynes, and Idris A. Eckley. *changepoint: An R package for changepoint analysis*, 2022. R package version 2.2.4.

[3] Hannes Mueller and Christopher Rauh. The hard problem of prediction for conflict prevention. *Journal of the European Economic Association*, Forthcoming.

[4] James Honaker, Gary King, and Matthew Blackwell. Amelia II: A program for missing data. *Journal of Statistical Software*, 45(7):1–47, 2011.

[5] Roderick JA Little and Donald B Rubin. The analysis of social science data with missing values. *Sociological Methods & Research*, 18(2-3):292–326, 1989.

[6] Erin LeDell, Navdeep Gill, Spencer Aiello, Anqi Fu, Arno Candel, Cliff Click, Tom Kraljevic, Tomas Nykodym, Patrick Aboyoun, Michal Kurka, and Michal Malohlava. *h2o: R Interface for the 'H2O' Scalable Machine Learning Platform*, 2022. R package version 3.36.1.4.

[7] Paula Branco, Rita P. Ribeiro, and Luis Torgo. UBL: an r package for utility-based learning. *CoRR*, abs/1604.08079, 2016.

[8] Angela M. Wood, Patrick Royston, and Ian R. White. The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data. *Biometrical Journal*, 57(4):614–632, 2015.