Forecasting Global Refugee Flows

A Machine Learning approach using non-conventional data

Daniela de los Santos, Eric Frey and Renato Vassallo

A thesis presented as part of the fulfillment of the Barcelona School of Economics Master's in Data Science for Decision Making

July 4, 2023

Forecasting Global Refugee Flows: A Machine Learning Approach using Non-conventional Data*

Daniela De los Santos†, Eric Frey‡ and Renato Vassallo§

July 4, 2023

Abstract

This study presents a novel forecasting framework for global refugee flows, incorporating non-conventional data sources such as Google Trends, the GDELT project event dataset, conflict forecasts, among others. Our main objective is to generate accurate one-step ahead predictions for the number of new refugee arrivals per country pair. These predictions play a crucial role in facilitating effective humanitarian response and informed infrastructure planning. While existing literature focuses on developed countries, this research develops a comprehensive global model, considering the majority of refugees seeking refuge in neighboring countries. In addition, to overcome challenges with imbalanced and low frequency data, two strategies are proposed using a rolling window framework: modeling refugee outflows from origin countries (outlfow level), and modeling refugee flows between country pairs (dyad level). Our results reveal a significant improvement in prediction accuracy by augmenting traditional variables from UNHCR with high-frequency non-conventional data, with Random Forest and Gradient Boosting as effective regressors.

Keywords: *Forecasting, refugee flows, machine learning, Google Trends.*
JEL codes: *C53, C55, F22.*

*We appreciate the comments by Konstantin Boss, Christian Brownlees, Elliot Gaston, Andre Groeger, Geraldine Henningsen, Jack Jewson and Hannes Mueller. We also thank Giovanna Chaves, Margherita Phillip, and Luis Quiñones for providing us with additional data resources. The views expressed in this document are those of the authors and do not necessarily reflect the official position of the Barcelona School of Economics.

†Barcelona School of Economics, daniela.de@bse.eu
‡Barcelona School of Economics, eric.frey@bse.eu
§Barcelona School of Economics, renato.vassallo@bse.eu

Contents

1 Introduction

Forecasting refugee flows is a challenging task due to its complex nature, influenced by geopolitical events, conflicts, and humanitarian crises. However, accurate and timely forecasts provide invaluable insights to governments, humanitarian organizations, and policymakers, allowing them to prepare and respond effectively to the needs of refugees and the host communities.

As shown in Figure 1, the total number of new refugees remained more or less stable between 2000 and 2012. However, in recent years this flow has grown significantly, meaning non-trivial implications for host countries mainly associated with resource allocation (such as food, shelter, healthcare, and education), humanitarian response (to ensure the provision of essential services and protection) and infrastructure planning (refugee camps or temporary housing facilities).

When we examine the breakdown of the total number of refugees by continents of origin, it becomes evident that Africa and Asia collectively account for over 80% on average. Notably, Asia has emerged as one of the primary contributor since 2013, during which a significant surge is observed. Furthermore, starting from 2018, South America has experienced a substantial increase in the number of refugees, which is in stark contrast to its historical levels. Lastly, the advent of the pandemic has led to a decline in the overall count of refugees, owing to lockdowns and restrictions that hindered international movement.

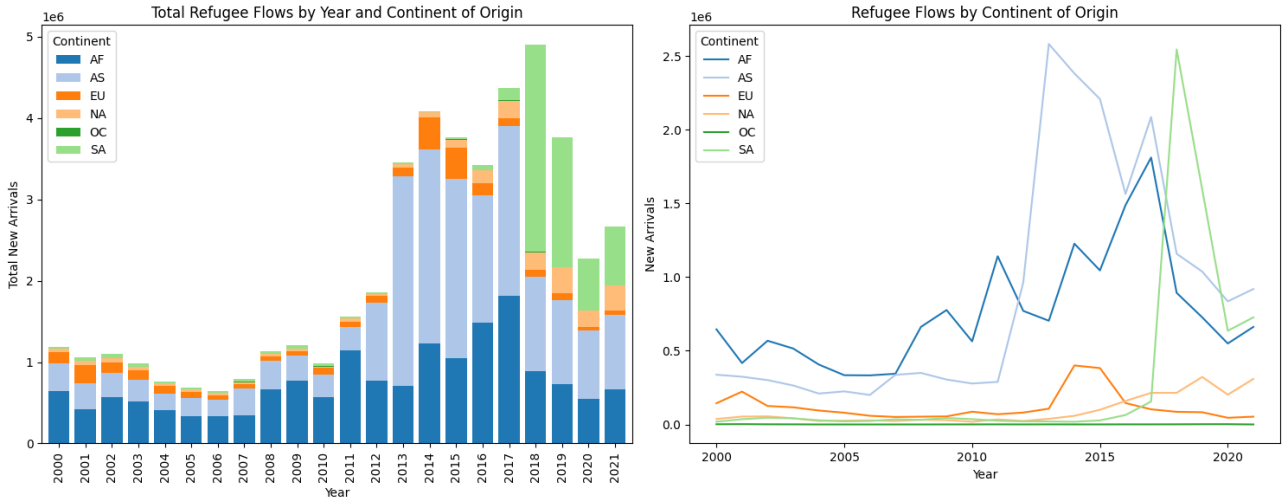

Figure 1: Trends in the aggregate number of refugee flows

The United Nations High Commissioner for Refugees (UNHCR) has accumulated a comprehensive database over the years, documenting the counts of individuals classified as refugees or refugee-like, categorized by their respective countries of origin and asylum/residence. In addition, in this study we developed a novel database with a wide variety of unconventional indicators, with the objective of informing in real time of possible events that trigger refugee or asylum-seeking flows worldwide.

The approach proposed in this document aligns with recent studies on immigration and refugee forecasting. However, our work stands out from existing research through **three main contributions**:

1. This research focuses on modeling global refugee flows, representing a large-scale machine learning effort to address this prediction problem. This perspective enables us to gain a deep understanding of the complex dynamics underlying refugee movements worldwide.

2. We integrate non-conventional data sources into our forecasting framework. Specifically, we leverage Google Trends data, which offers insights into the search patterns of individuals seeking information about key semantic terms, as well as neighboring countries and cities that often serve as potential destinations for refugees.

3. To tackle the challenges posed by imbalanced and low-frequency data, we propose two complementary strategies: modeling refugee outflows from origin countries and modeling refugee flows between pairs of countries. This segregation of modeling approaches allows for a more comprehensive understanding of the intricate dynamics driving refugee flows.

After applying a one-step rolling window approach on different tree ensemble models (specifically Random Forest and Gradient Boosting), the findings demonstrate significant improvements in terms of RMSE by incorporating high-frequency data sources like GTrends, GDELT, and conflict data, compared to models that solely rely on UNHCR covariates. Notably, our preferred model outperforms a benchmark model by approximately 20% when examining at dyad level.

The remainder of the document is divided as follows. Section 2 briefly summarizes the literature on empirical evidence on refugee forecasting. Section 3 describes the wide variety of data sources we use for the prediction problem. Section 4 describes the proposed strategies considering the complexities of global refugee movements. Section 5 examines the results for the refugee outflow model and for the dyad-specific model. Section 6 summarizes the conclusions. It should be noted that this document shows only the best-fitting models. An appendix showing important definitions, as well as the results for other models, are available in Section A.

2 Brief Literature Review

According to the 1951 Refugee Convention and the 1967 Protocol (UNHCR, 2011), a refugee is someone who 'owing to a well-founded fear of being persecuted for reasons of race, religion, nationality, membership of a particular social group or political opinion, is outside the country of his nationality, and is unable to, or owing to such fear, is unwilling to avail himself of the protection of that country'.

The existing empirical evidence consistently demonstrates that armed conflict in the country of origin is the primary factor influencing forced out-migration (Ball et al., 2002). Equally significant is the lack of civil liberties or political rights within the same country (e.g., Neumayer 2004, 2005; Moore and Shellman 2007). Furthermore, Barthel and Neumayer (2015) have identified geographic proximity, shared language, and historical colonization as additional determinants of forced migration.

Two marked approaches stand out in the recent literature on immigration and refugee forecasting. On the one hand, there are gravity models (inherited from international trade analysis), which exploit the structural relationships between the main determinants of these flows and which take into account common characteristics between groups of countries (fixed effects, for example). Whilst gravity models describe spatial patterns of international migration very well, they fail to capture basic temporal dynamics, indeed, often worse than even the time-invariant average of the historical flows (Echevarria and Gardeazabal, 2016; Beyer et al., 2022).

On the other hand, there are reduced-form models, which sacrifice interpretability to exploit (often non-

linear) patterns among the forces that generate these international flows. Within the latter branch, non-conventional indicators have become important inputs to improve the predictive capacity of models.

Within this branch, the study of Carammia et al. (2020) presents an adaptive machine learning algorithm that integrates administrative statistics and non-traditional data sources at scale to effectively forecast asylum-related migration flows. The authors focus on asylum applications lodged in countries of the European Union (EU) by nationals of all countries of origin worldwide. In this line, Boss et al. (2023) develop monthly refugee flow forecasting models for 150 origin countries to the EU27, using machine learning and high dimensional data. They find that an ensemble forecast combining Random Forest and Extreme Gradient Boosting algorithms consistently outperforms for forecast horizons between 3 to 12 months.

Our research is aligned with the latter two studies, as we also utilize high-frequency data and machine learning techniques to generate forecasts. However, our distinctive contribution lies in the global perspective we adopt when examining refugee flows. In this regard, it is crucial to acknowledge that the primary destination for the majority of refugee migration is neighboring countries. This observation is supported by Figure 2, which presents the cumulative number of new refugee arrivals categorized by degrees of separation from 2000 to 2021.

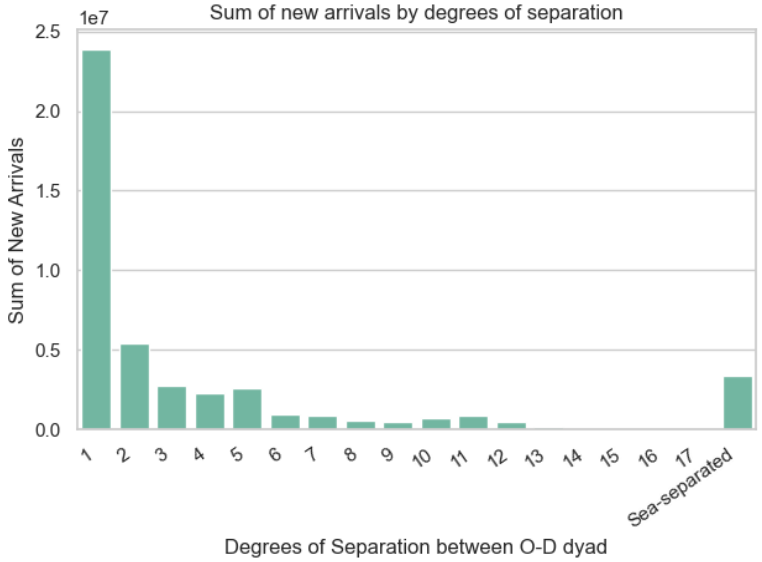

Figure 2: Sum of new arrivals of refugees by degrees of separation (2000-2021)

By focusing solely on forecasting refugee flows into Europe, the larger picture of global displacement trends may be overlooked. Therefore, it becomes imperative to develop forecasting models that consider the dynamics and complexities of displacement on a global scale.

3 Data

In this section we describe the wide variety of data sources we collect and how they are processed for this prediction problem. Table 1 summarizes the data we use.

Data	Freq	Level	Years	Countries	Source
UNHCR Refugee flows	Yearly	Country, Dyad	2000-2022	196	UNHCR
Migration flows	5 years	Dyad	2000-2020	194	Our World In Data
Internal Displacement Figures	Yearly	Country	2008-2022	85	IDMC
Fragile States Index	Yearly	Country	2007-2022	179	Fund for Peace
Topic-modeled newspaper text	Monthly	Country	2000-2023	190	Conflict Forecast
Conflict Forecast	Monthly	Country	2000-2023	190	Conflict Forecast
GDELT	Monthly	Country	2000-2023	195	The GDELT Project
Google Trends	Monthly	Country, Dyad	2005-2023	195	Google Trends

Table 1: Summary of different datasets

3.1 Traditional predictors

3.1.1 UNHCR refugee data

The UNHCR dataset, provided by the UNHCR as part of our master's thesis, contains annual observations on the number of new refugees between country origin-destination pairs from 2000 to 2021. It also contains 45 covariates, which include information on the origin country, destination country, and the specific pair. The country-specific data includes the number of fatalities due to conflict, the number of years of conflict, GDP, and inflation, whereas the dyad-specific data includes data on whether or not the two countries share a common language, common colonizer, were in a colonial relationship, whether or not they are bordering, and the distance between the two countries. See A.1 for a full description of the covariates.

We feature engineer this data by binary-encoding the origin country, destination country, origin-destination pair, origin continent and destination continent. We compute rolling sums of newarrivals between countries for the past 3, 5, and 10 years. We include lagged variables of variables we believe to be most important up to three years. We also construct a feature that describes the degrees of separation between countries.

3.1.2 Migration Data

Obtained from Our World In Data, this migration data describes the stock of migrants from an origin country living in a destination country. We believe this data is useful for two reasons: (1) people seeking asylum are more likely to seek it where emigrants are already living and (2) an increase in migration could be an early indicator of an increase in asylum seekers. We first process this data by forward filling in missing years, as the stock of migrants is available every five years. We then feature engineer this data by dividing the stock of migrants by the origin country population, the destination country population, and leaving the numerical stock as is.

3.1.3 Internal Displacement Data

The European Civil Protection and Humanitarian Aid Operations estimates that in 2022 there were 53.2 million internally displaced peoples (IDPs) and 27.1 million refugees (European Civil Protection and Humanitarian Aid Operations). As the number of IDPs is almost twice that of refugees and can be an indicator of the extent of potential refugees, we include data on the number of IDPs obtained from the IDMC. Unfortunately, the data obtained from the IDMC only includes 85 countries from 2008 to 2022, and some years are missing. To make it usable, we first transform the data, $log(\frac{\text{number of IDPs}}{\text{country population (in 1000s)}} + .1)$, which has the effect of transforming the IDP data to a distribution with a range of 20. We then round the data to the nearest integer, convert the data to character data, then target encode it. Thus, observations that have data on IDP are included in terms of the average number of newarrivals per IDP bin, and the missing IDP data is essentially imputed as the mean of newarrivals.

3.1.4 Fragile States Index

The Fragile States Index, produced by the Fund For Peace includes an overall fragility measure as well as twelve specific cohesion, economic, political and social indicators of vulnerability to collapse or conflict. It is produced annually for 179 countries from 2006 to today. To estimate the remaining countries we take an average of the bordering countries' fragility indices. In the case of island countries we take an average of the nearest neighbors by distance.

3.2 Non-conventional predictors

3.2.1 Topic-modeled newspaper text and ConflictForecast data

We obtain topic-modeled newspaper text and conflict forecasts provided by Hannes Mueller as part of his work with the ConflictForecast.org. The topic modeled newspaper text data describes the proportion of country-specific articles written about 15 topics that have been modeled using a Latent Dirichlet Allocation model. The Conflict-risk dataset contains data on political violence and escalations into internal armed conflict. Both of these datasets are at a monthly frequency for approximately 190 countries. To convert to yearly data, we compute summary statistics: minimum, 25th percentile, mean, 75th percentile, maximum, and average log difference for a given column and year. Similar to the Fragile States Index, we impute missing countries with bordering and neighboring countries' data.

3.2.2 GDELT

The Global Database of Events, Language, and Tone (GDELT) is a large-scale initiative that monitors global news media, providing a comprehensive database of events and emotions from around the world. It was created with the goal of capturing, analyzing, and understanding the global societal context in near real-time. GDELT collects data from news articles, television broadcasts, online media, and other sources in multiple languages. It processes this vast amount of information to identify and extract various attributes such as event types, actors involved, locations, and sentiments expressed.

A thesis team consisting of Giovanna Chaves, Luis Quiñones, and Margherita Phillip were kind enough to provide us with a dataset that they created from The GDELT Project as part of their master's thesis. This dataset contains the counts of 20 event categories related to conflict on a monthly basis on a subnational region level. We first sum by month and country, and then normalize by country. This is done because the count of events recorded varies by country and year, and this normalization is performed

to make the GDELT data comparable across countries and years. We aggregate the monthly data by taking the mean and max of the categories over the year.

3.2.3 Google Trends

Google Trends is a free tool that analyzes the popularity of keywords over time. It offers insights into the relative frequency of regional-specific searches. Boehme et al. (2020), Boss et al. (2023), and Carammia et al. (2020) have demonstrated that Google Trends recorded in migrant origin countries holds additional predictive power over classical predictors when explaining asylum and non-asylum migration flows.

In this line, we collect data from Google Trends of users searching for terms that may indicate interest in leaving an origin country. These include searches related to travel, conflict, economics, and destinations. We collect data on both terms and topics (clusters of related keywords irrespective of a specific language). The terms fall into two categories:

- **Origin country searches:** These keywords may indicate intention to leave, irrespective of the destination. We use a combination of terms and topics, based on Boss et al's search terms used in their work. For the terms, we search in English as well as up to two primary languages for each country. See A.2.1 for a complete list of terms and topics.

- **Dyad-specific searches:** These keywords may indicate intention to leave to a specific destination. These include destination countries, up to two main cities per destination country, and up to two border cities per origin-destination country pair that shares a border. Because it was infeasible to gather a full data set of every single origin-destination search for every single term, we prioritize neighboring countries up to two degrees of separation as well as the remaining top 500 country pairs in terms of the number of new arrivals. We also were able to obtain an interaction term, `visa + destination country` for every single origin-destination pair in two primary languages of each country.

As a motivation for the usage of Google Trends, Figure 3 shows the evolution of refugee flows between Venezuela and two countries: Colombia (top panel) and Brazil (bottom panel). Also, the blue line in both graphs represents the Google Trends index for searches for specific terms made in Venezuela, in particular searches for border cities such as Cúcuta and Boa Vista. A significant jump is evident for Google trends indicators in periods around 2018, a period in which the number of refugees increased considerably, associated with the social and political crises in Venezuela.

We treat the data in two ways that improved the predictive power: smoothing 'spikes' and weighting the Google Trends data by the percent of the population with internet access. We aggregate using monthly rolling sums, and taking the mean and max of a year.

4 Model

Modeling global refugee flows poses several unique challenges, namely that the data on these flows are imbalanced, and low frequency.

- **Imbalanced :** The data is highly imbalanced, with 90% of the observations listed as zero, 99% as less than 100 new arrivals, and the remaining 1% of observations somewhere between 100 and 1

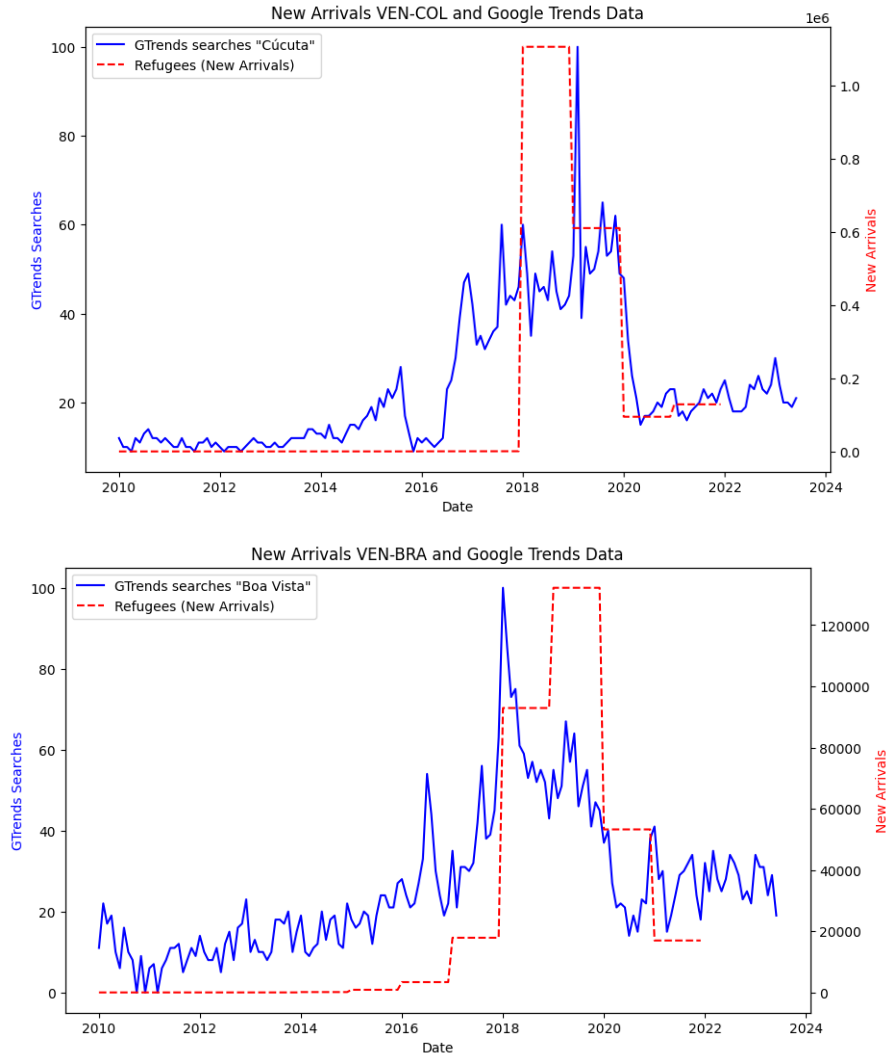

Figure 3: Google Trends data for Venezuela

million new arrivals. As the prediction task involves an origin country and a destination country, working with 196 countries means that one year of forecasting yields 196*195 = 38,220 dyad forecasts. This makes the prediction task quite challenging, as there are relatively few observations to learn from when predicting large flows.

- **Low Frequency:** The global refugee data available from the UNHCR, while extensive in the sense that it covers 196 countries and territories, is limited in the sense that it is only on a yearly basis. This means that approaches of a different model for each corridor are infeasible, because there are only about 20 observations per origin-destination dyad. This task naturally requires an approach that incorporates learning across time series.

  Working with yearly data poses additional challenges if one wishes to incorporate higher-frequency data for prediction. Possible solutions include aggregating high-frequency data at a yearly level or using a model such as MIDAS to explicitly work with mixed-frequency data. As discussed in Section 3, we aggregate the high-frequency data by taking the mean, max, and rolling sums of variables based on our priors but an interesting extension would be to work a model that allows for mixed-frequency data.

We approach the modeling process in **two separate stages**: modeling **refugee outflows** from origin countries, and modeling **refugee flows between dyads** of countries. Both predictions can be approached differently due to their distinct underlying drivers.

Refugee flows, which encompass the movement of individuals into various countries, are influenced not only by the conditions in the countries of origin but also by the political decisions of destination countries regarding accepting refugees. These decisions can significantly impact the scale and pattern of refugee inflows to different regions. On the other hand, refugee outflows primarily stem from the conditions in the countries of origin, such as conflicts, environmental disasters, or socio-economic challenges. Unlike refugee flows, outflows are less contingent on external political decisions and may occur independently, driven by the urgent need for safety and protection. Thus, separating the modeling approaches allows for a more comprehensive understanding of the complex dynamics driving forced displacement and enables tailored strategies for predicting and responding to both types of movements.

As practical notes, we can mention two additional aspects. First, optimizing the outflow predictions as a first stage led to a faster and more robust optimization process for the dyad-level predictions. Second, the distribution of flows and outflows in the datasets is quite different, becoming less imbalanced when aggregated to an outflow level, which may make the prediction task easier to model as displayed in Figure 4.

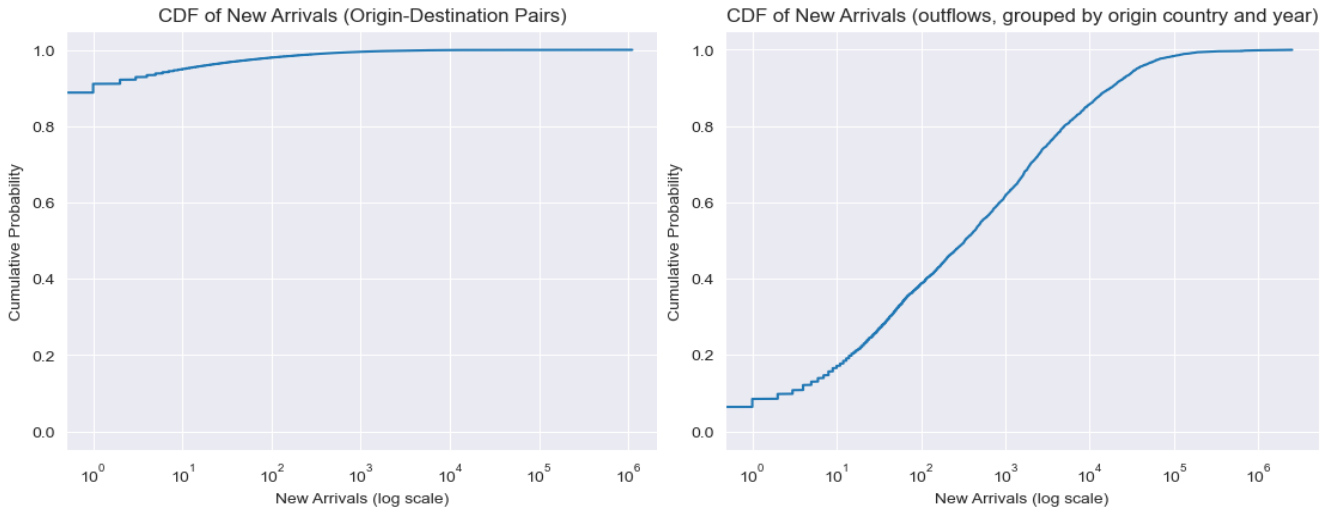

Figure 4: Comparison between distribution of target variable (new arrivals of refugees) on a dyad level and an outflow level. Outflow target distribution is significantly less imbalanced.

5 Results

5.1 Modeling refugee outflows

The exercise of modeling outflows included feature engineering and testing all the different covariates reported in the data section that are relevant for outflow prediction, as well as combining them in different ways. Different tree ensemble models were tested: the best results were obtained with Random Forest and Gradient Boosting regressors.

The models were trained on a rolling window basis, and set to predict on a one-year horizon. Some of the models are set to predict levels directly, while others predict a log transformation of the target variable,

which is then re-converted to levels. The training data was set to start always in 2005, as this is when most covariates become available. Predictions were tested out-of-sample for the 2018-2021 period. In practice, this means that, for example, when the model aimed at predicting 2021, it was trained with data from 2005 to 2019, and then used 2020 covariates to make the predictions.

We examine the performance of our model using the relative Root Mean Squared Error, which can be calculated using the following formula:

$$\text{Relative RMSE} = \frac{\text{RMSE}_{\text{Model}}}{\text{RMSE}_{\text{Naive}}} \tag{1}$$

where the naive prediction is simply last year's outflow value for the country.

Table 2 summarizes some of the combinations tested. Data sources which we found overlapped with others providing similar information and therefore not improving accuracy, were left out (i.e. FDI, IDP, and topic-modeled newspaper text data).

Model	**Root Mean Squared Error**				
	Testing Period				**Overall**
	2018	2019	2020	2021	(Relative RMSE)
Naive	186459,00	79330,73	69261,51	15408,33	1
History model: UNHCR covs.. including lags of outflows	200535,74	59296,63	51742,74	24249,31	1.10
GTrends	182600,74	109160,21	31006,37	43626,45	**1.08**
Augmented 1: History model + GTrends	192868,42	83155,21	37726,97	31612,19	1.00
Augmented 2: Aug1 + Migration features	209066,05	77796,26	74016,37	38933,53	1.10
Aug2 (log Y)	208175,28	61634,98	75045,65	33826,17	0.91
Augmented 3: Aug2 + GDELT	174695,55	78948,88	15411,88	36499,59	1.08
Aug3 (log Y)	177111,55	32300,74	22388,75	44581,17	**0.87**
Aug4: Aug3 + Conflict	203941,98	53386,28	67003,01	34913,55	1.04
Aug4 (log Y)	176951,53	77064,87	28886,73	44478,41	0.93

Table 2: RMSE of outflow models

Note: with the exception of the GTrends model, which uses Random Forest, the results were obtained with Gradient Boosting Regressor models.

There are several key points to highlight from the obtained results. Firstly, it should be noted that the accuracy of the naive predictions varies significantly from one year to another. For instance, the naive RMSE for 2021 represents only 8% of the RMSE for 2018. The naive prediction approach assumes a constant or incremental change in outflows over time, which has not been the case during the last decade. As shown in Figure 1, there was a historical peak in 2018, followed by rapidly changing behavior partly due to COVID-19 movement restrictions. Predicting these peaks and troughs is be especially demanding for any model.

Starting with the more basic models, we observe that a model that learns from its own history will have relatively similar results to the naive predictions, even when accounting for some covariates like historic conflict features and economic factors. On the other hand, it is worth noting that the model that does

not include any historical features but only uses Google Trends to predict outflows produces similar estimates as the naive model. This confirms the validity of Google Trends as a good predictor for refugee dynamics.

Different augmentations of these initial models improve the predictions to varying degrees, though the variability by year is noticeable. For example, several models achieve very good results in predicting 2020 outflows (augmentations 2 and 3 with logarithmic transformations achieve a relative RMSE of .22 and .32, respectively). Different augmentations of these initial models improve the predictions to varying degrees, though the variability by year is noticeable. For example, several models achieve very good results in predicting 2020 outflows (augmentations 2 and 3 with logarithmic transformations achieve a relative RMSE of .22 and .32, respectively). On average, the third augmentation of the model (which includes UNHCR covariates, Google Trends features, migration and GDELT features) is the most balanced in obtaining relatively low RMSEs for most years (.94, .40, and .32 for the 2018-2020 period), with the exception of 2021.

In general, 2021 proves to be the most challenging year to predict, and no model performs better than the naive prediction. It is likely that the complex dynamics of the COVID-19 pandemic affected refugee outflows differently in the 2020-2021 period. While both years faced quarantines and international movement restrictions, in 2021 the outflows started growing again, perhaps feeding the model contradictory information. As a note, including COVID-19 related covariates [1] did not improve the results.

Figures 5 and 6 intend to show the advantages of predicting a target variable that has been temporarily converted to logarithms. On the one hand, we see that RMSE decreases with the logged prediction especially for larger outflows, which are typically harder to predict as they are outliers in the dataset. On the other hand, it is easy to see that logged predictions are less biased, while predicting directly in levels leads to overestimations (the more points we find in the shaded area of the plots, the more our model tends to overshoot).

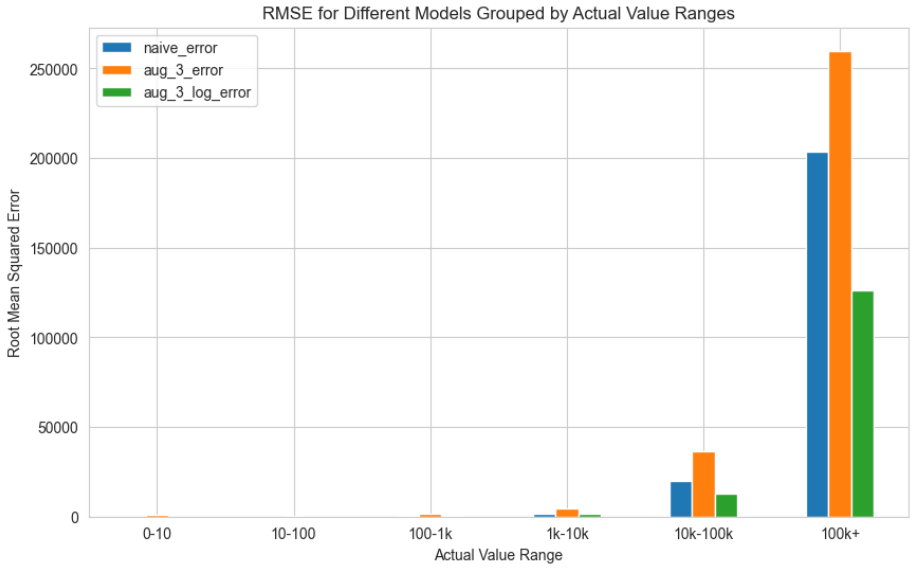

Figure 5: RMSE for Augmented model (4), conditional on outflow size

[1] Two sets of COVID-19 related covariates were tested: the Stringency Index dataset computed daily by Oxford University, available here: https://ourworldindata.org/covid-stringency-index, which also includes health-related outputs; and a dataset that provides daily information about international flights restrictions: https://ourworldindata.org/covid-international-domestic-travel

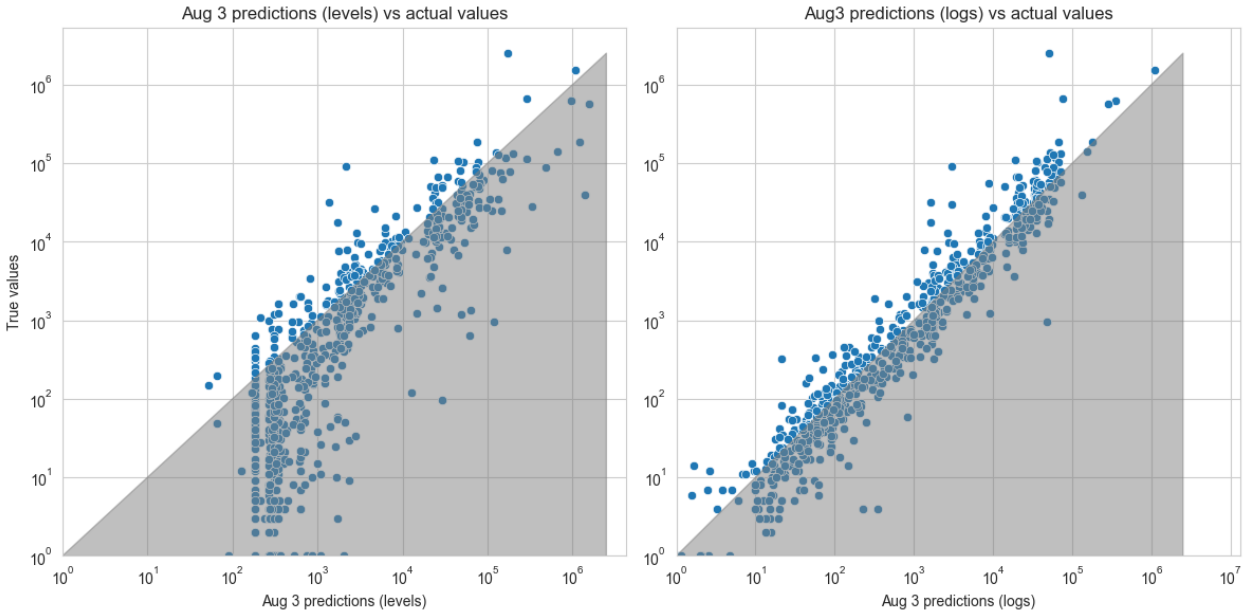

Figure 6: Comparison of prediction bias for the same model (aug 4), with and without logarithmic transformation of the target variable

Finally, Figure 7 shows the outflow results for the best-performing model grouped by continent, compared with real values and naive results. Accounting for the levels reported on the Y axis, we see that our model still underestimates the results for some of the larger outflow-producing continents such as Africa, Asia, and South America.

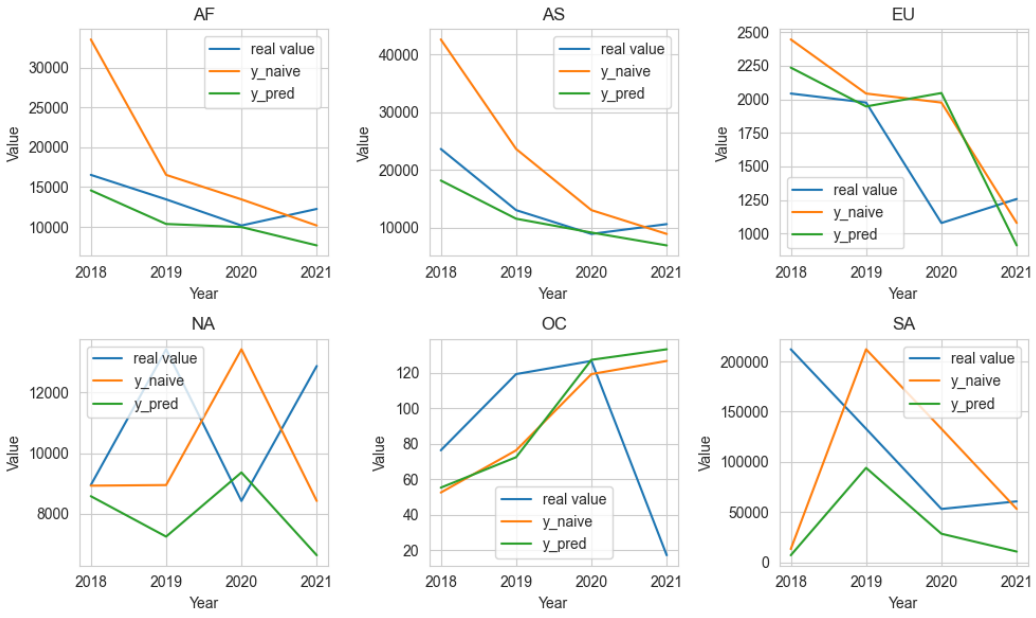

Figure 7: Comparison of prediction at the continent level (aug 3 model vs naive vs real values)

5.2 Dyad-specific model

In this section, we present our results for predicting flows between countries. Similar to the outflow prediction problem, we implement a rolling window to predict on a one-year horizon. All models are trained on logs of the target variable, and then converted back to levels before computing the error.

Table 3 and Figure 8 compares performance across testing years and models. Except for the naive which is simply the previous year's target values, each model is a GradientBoostingRegressor trained on different combinations of the datasets. We find that a combination of high-frequency data sources yields the best results, though adding any high-frequency data source improves the predictions slightly relative to simply using the UNHCR covariates. Similar to the outflow predictions in the previous section, 2021 seems to be a difficult year to predict and all models perform worse than the Naive prediction.

We do not include the comparison of the Migration Data, FragileStatesIndex, Internal Displacement nor the Topic-modeled newspaper text as these generally did not have a significant improvement of the results. We also experimented with undersampling the observations where the target values were zero, sampling only 20% of the zero-values for example, but did not see improvement. Additionally, we tried including an ensemble model, including covariates plus the predictions of a model only trained on observations whose target value is greater than 100 and the predictions a model trained on observations whose target value is less than or equal to 100, but did not see great improvement with this specialization approach. Finally, we attempted an approach where dimensionality was reduced by including the outflow predictions generated in the first modeling stage as a covariate. However, this did not give better results than including outflow-predicting covariates directly in the dyad-level models.

Model	**RMSE**					
	Testing period				**Overall**	
	2018	2019	2020	2021	RMSE	(Relative RMSE)
Naive	7957.3	3182.8	3016.8	1226.9	4584.1	1.00
UNHCR	6811.0	3295.2	1187.7	1768.3	3930.2	0.857
UNHCR + GoogleTrends	6730.9	2810.1	1161.1	1728.3	3792.7	0.827
UNHCR + GDELT	6747.4	2939.4	1177.2	1781.8	3831.7	0.836
UNHCR + Conflict	6773.9	2959.4	1168.2	1769.4	3845.1	0.839
UNHCR, GTrends, GDELT, Conflict	6821.5	1723.2	1138.1	1705.6	3664.3	0.799

Table 3: RMSE of different models by year

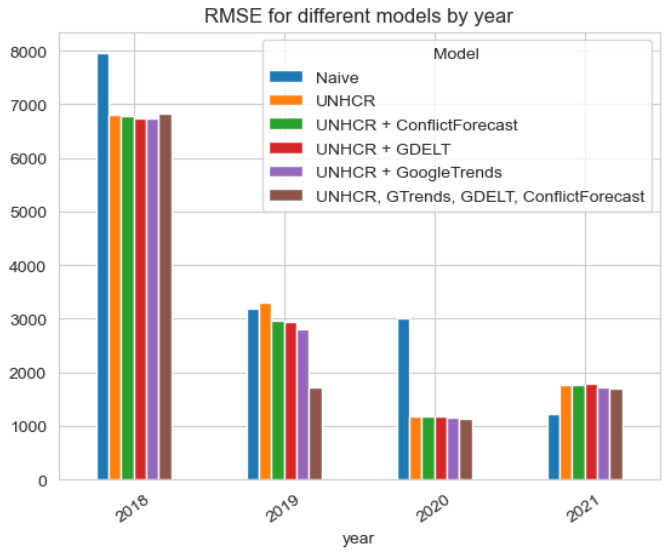

Figure 8: Plot of RMSE of different models by year

We provide a few examples of forecasts of the largest flows from the best performing model in Figure 9. The model is generally conservative in its estimates of the number of new arrivals, which could be due to the fact that the majority of the target values are quite small.

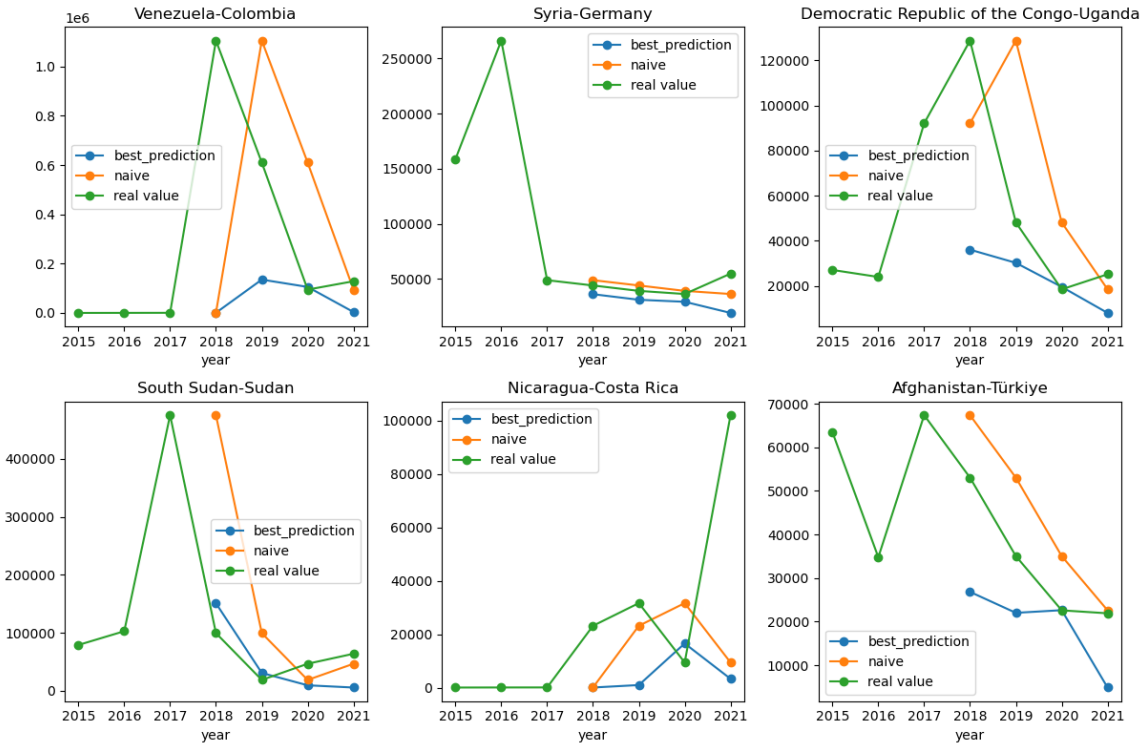

Figure 9: A sample of predictions from the best performing model trained on UNHCR covariates, GoogleTrends GDELT, and the ConflictForecast. The actual values and naive predictions are included for comparison.

Figure 10 displays the RMSE of the naive and best model's predictions, grouped by the actual new arrival values. It appears that it outperforms the naive prediction up to values less than 100,000. At that point the sample size becomes extremely small, which makes it quite difficult for the model to learn from these outliers in the training process.

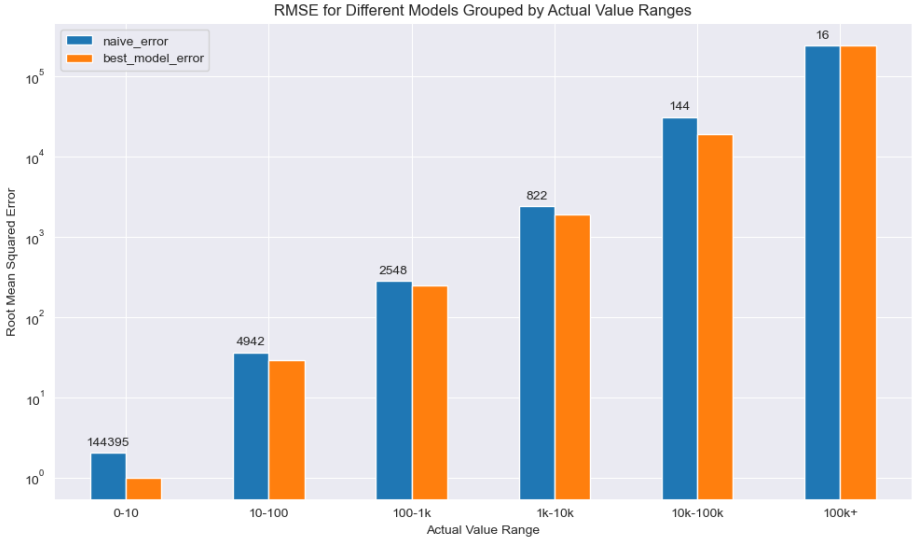

Figure 10: Log scale of the error grouped by the actual value.

6 Conclusions and Recommendations

Forecasting refugee flows is a complex task due to various factors such as geopolitical events, conflicts, and humanitarian crises. However, accurate and timely forecasts are crucial for governments, humanitarian organizations, and policymakers to effectively respond to the needs of refugees and host communities.

In this research, we aim to address the challenge of modeling global refugee flows, representing the first large-scale machine learning attempt to tackle this prediction problem to the best of our knowledge. Our study introduces two complementary strategies for modeling refugee flows, focusing on both outflows from origin countries and flows between pairs of countries. The results indicate significant improvements in forecasting accuracy by incorporating high-frequency non-conventional data sources compared to models relying solely on traditional UNHCR covariates. The proposed model outperforms a benchmark by approximately 20% at the dyad level, demonstrating the usefulness of diverse data sources in this task.

In addition to this, our results highlight the relevance of alternative and freely-available high-dimensional data sources as standalone predictors for refugee outflows and flows. For instance, Google Trends models that do not have any information about the refugee flows history at the country level, can still make predictions that are similar to a history-based model. This can be an especially useful tool in scenarios of data scarcity, where refugee flows are suspected to be under-reported or have stopped being reported for political reasons. Augmenting Google Trends features with other available sources that perform well in predicting conflict, such as GDELT and ConflictForecast, can allow for more robust predictions in cases where the data is unknown.

In general terms, it is worth flagging that our best-predicting models for both outflows and flows are mostly conservative in their estimates, prioritizing caution. However, policymakers may prefer less conservative projections to proactively prepare for a broader range of scenarios. To address this, the models can be tweaked by adjusting parameters and assumptions, allowing for a more pessimistic outlook when necessary.

Based on the findings of our study, several **recommendations** can be made for future research. One potential avenue is building models of hierarchical forecasting of outflow and dyad together. By incorporating the hierarchical structure of migration data, such as considering the interdependencies between countries of origin and destination, we can improve the predictive capabilities of our models. Moreover, it is essential to explore modeling mixed-frequency data instead of aggregating it. This approach would account for the inherent variability in migration data, allowing for more precise predictions by considering the fluctuations at different time intervals. Future research may also explore dealing with the imbalanced nature of this regression problem as well as predicting the onset of these flows, as the largest flows are quite anomalous and can be quite sudden (see Venezuela-Colombia in Fig. 9). To further enhance the predictive power of migration models, it is crucial to incorporate additional data sources that capture a destination country's willingness to receive refugees. For example, considering factors such as EU countries' agreements to accept refugees could provide valuable insights into migration patterns.

In conclusion, the integration of advanced machine learning techniques with diverse and high-dimensional data sources has the potential to improve the accuracy and effectiveness of forecasting refugee flows, enabling better preparation and response strategies for governments, humanitarian organizations, and policymakers in addressing the complex challenges associated with refugee crises.

7 References

[1] Ball, P., Betts, W., Scheuren, F., Dudukovich, J., and Asher, J. (2002). Political killings in kosova/kosovo, March–June 1999. Technical report. Washington, DC: American Association for the Advancement of Science.

[2] Barthel, F., and Neumayer, E. (2015). Spatial dependence in asylum migration. *Journal of Ethnic and Migration Studies* **41(7)**, 1131–1151.

[3] Beyer, R. M., Schewe, J., and Lotze-Campen, H. (2022). Gravity models do not explain, and cannot predict, international migration dynamics. *Humanities and Social Sciences Communications* **56(9)**.

[4] Bohme, M. H., Groger, A., and Stohr, T. (2020). Searching for a better life: Predicting international migration with online search keywords. *Journal of Development Economics* **142**, 102347.

[5] Boss, K., Groeger, A., Heidland, T., Krueger, F., and Zheng, C. (2023). Forecasting Bilateral Refugee Flows with High-dimensional Data and Machine Learning Techniques. *BSE Working Paper* **1387**, Barcelona School of Economics.

[6] Carammia, M., Iacus, S. M., and Wilkin, T. (2020). Forecasting asylum-related migration flows with machine learning and data at scale. *Sci Rep* **12**, 1457.

[7] Echevarria, J. and Gardeazabal, J. (2016). Refugee gravitation. *Public Choice* **169**, 269-292.

[8] Moore, W. H., and Shellman, S. M. (2007). Whither will they go? A global study of refugees' destinations, 1965–1995. *International Studies Quarterly* **51(4)**, 811–834.

[9] Mueller, H. and Rauh, C (2022). The Hard Problem of Prediction for Conflict Prevention. *Journal of the European Economic Association*, forthcoming.

[10] Neumayer, E. (2004). Asylum destination choice: What makes some west european countries more attractive than others? *European Union Politics* **5(2)**, 155–180.

[11] Neumayer, E. (2005). Bogus refugees? The determinants of asylum migration to Western Europe. *International Studies Quarterly* **49(3)**, 389–410.

[12] European Civil Protection and Humanitarian Aid Operations. (2022). "Forced displacement Factsheet". Retrieved June 10, 2023, from https://civil-protection-humanitarian-aid.ec.europa.eu/what/humanitarian-aid forced-displacement-refugees-asylum-seekers-and-internally-displaced-persons-idps_en

[13] UNHCR (2011). United Nations high commissioner for the refugees: Convention and protocol relating to the status of refugees.

A Appendix

A.1 UNHCR Dataset Variables

iso_o	Iso code country of origin
iso_d	Iso code host country
year	Year
newarrival	New refugees
CL_d	Civil Liberty of host country
CL_o	Civil Liberty of country of origin
col45	Dummy variable of country pairs in colonial relationship post 1945
colony	Dummy variable of country pairs ever in colonial relationship
comcol	Dummy variable for countries with common colonizer post 1945
comlangethno	Dummy variable if a language is spoken by at least 9% of the population in both countries
comlangoff	Dummy variable common official of primary language
contig	Dummy variable for contiguity
CPI_d	Inflation in host country
CPI_o	Inflation in country of origin
$Dead_d$	Dummy variable number of fatalities in host country > 100
$Dead_o$	Dummy variable number of fatalities in country of origin > 100
$Deadlog_d$	Logarithm of number of fatalities in host country
$Deadlog_o$	Logarithm of number of fatalities in country of origin
dist	Simple distance (most populated cities, km)
$GDPPP_d$	Gross domestic product per capita, constant prices of host
$GDPPP_o$	Gross domestic product per capita, constant prices of origin
$GDPPPP_d$	Gross domestic product based on purchasing-power-parity (PPP) share of world tot of host country
$GDPPPP_o$	Gross domestic product (PPP) country of origin
index0asylum	Dummy variable if host country gives temporal asylum to country of origin
$Nyearconf_d$	Dummy variable years of conflict in host country > 0
$Nyearconf_o$	Dummy variable years of conflict in country of origin > 0
$Nyearlog_d$	Logarithm of number of years of conflict in host country
$Nyearlog_o$	Logarithm of number of years of conflict in country of origin
Pop_d	Population of host country
Pop_o	Population of country of origin
PR_d	Politics Rights of host country
PR_o	Politics Rights of country of origin
smctry	Dummy variable countries were or are the same country
$typeofviolence_d$	Type of UCDP conflict in host country, 1 == state-based conflict, 2 == non-state conflict, 3 == one-sided violence

A.2 Google Trends

A.2.1 List of keywords used

List of terms used:

asylum
asylum seeker
border controls+border control
bureau of immigration
citizen
citizenship+citizenships
consulate+consulates
country
crises+crisis
deportation+deportations+deported
devaluation
diaspora
dual citizenship
dual nationality
emigrant+emigrants
emigrate+emigrated
embassy+embassies
emigration
foreigner+foreigners
immigrant+immigrants
immigrate+immigrated
legalization+legalisation+legalisations+legalizations
migrate
migration
nationalization+nationalisation
nationality+nationalities
naturalization+naturalisation+naturalisations+naturalizations
political asylum
political refugee
recruitment+recruitments
refugee+refugees
recession+recessions
repatriation
Schengen
smuggler+smugglers+smuggling
student visa
undocumented
visa free
visa+visas
work visa

List of topics used:

Armed Forces
Bureau de change
Civilian
Conflict
Coup d'état
Crisis
Currency
Economy
Genocide
Government
Immigration
Lottery
Militia
Passport
Protest
Refugee
Travel Visa
Violence
Wage
War

A.2.2 Preprocessing steps

We introduce an algorithm that, if there is a spike which goes from 0 to 100 or 100 to 0, we rescale the trends data.

Algorithm 1: RescaleTrend Algorithm

procedure RescaleTrend(s, t, max_iter)
  Input: Google Trends series s (array), threshold t (float), maximum iterations max_iter (int)
  Output: Rescaled trend if there is a spike
  $max\_value\_index \leftarrow \text{argmax}(s)$
  $i \leftarrow 0$
  while $(s[\text{max\_value\_index} - 1] = 0) \vee (s[\text{max\_value\_index} + 1] = 0) \vee (i < max\_iter)$ **do**
    $next\_largest\_value \leftarrow$ the second largest value
    if $next\_largest\_value \neq 0$ **then**
      $scale\_factor \leftarrow 100/\text{next\_largest\_value}$
    else
      $scale\_factor \leftarrow 0$
    end if
    if $\text{max\_value\_index} < 2$ **then**
      $s[\text{max\_value\_index}] \leftarrow$ average of neighboring values ▷ Replace the spike value
    else
      $s[\text{max\_value\_index}] \leftarrow$ average of neighboring values ▷ Replace the spike value
    end if
    $s \leftarrow s \times scale\_factor$
    $i \leftarrow i + 1$
    $max\_value\_index \leftarrow \text{argmax}(s)$
  end while
  return s
end procedure

A.3 Other Models

A.3.1 Dyad-specific model using conventional data

As our initial approach, we employ a recursive rolling forecast algorithm to estimate a time series model for each country pair within our sample, utilizing the data provided by the UNHCR. It is worth mentioning that our target variable *newarrival*, is subject to a certain reporting lag caused by the delay in the official statistical data collection process. However, we leverage this characteristic to our advantage by incorporating exogenous variables that are available without lag, enabling us to integrate real-time information and enhance the forecasting process. This allows us to establish a nowcasting method for our target variable.

To evaluate the effectiveness of our proposed model, we conducted a comparative analysis against a benchmark model. Initially, we employed a naive prediction approach based solely on the previous value of the target variable. Additionally, we introduced a second model that solely relies on lagged values of the same target (autoregressive model, AR).

Subsequently, we expanded our autoregressive (AR) model by incorporating relevant UNHCR covariates, which provided valuable additional information for our predictions. Moreover, to further enhance the accuracy of our forecasts, we incorporated contemporaneously available exogenous variables.

Figure 11 illustrates the process adopted for this approach. We utilized a recursive rolling window forecast methodology. This involved training distinct models for each step within the forecast horizon, enabling us to capture the evolving dynamics and adapt the model's predictions accordingly.

To perform the prediction, we use the *Skforecast* library, in particular the *ForecasterAutoreg* class. The model is created and trained from a *RandomForestRegressor* with a time window of 4 lags. We split this dataset between train and test, where observations between 2007 and 2017 are included in the train set and observations between 2018 and 2021 are included in the test.

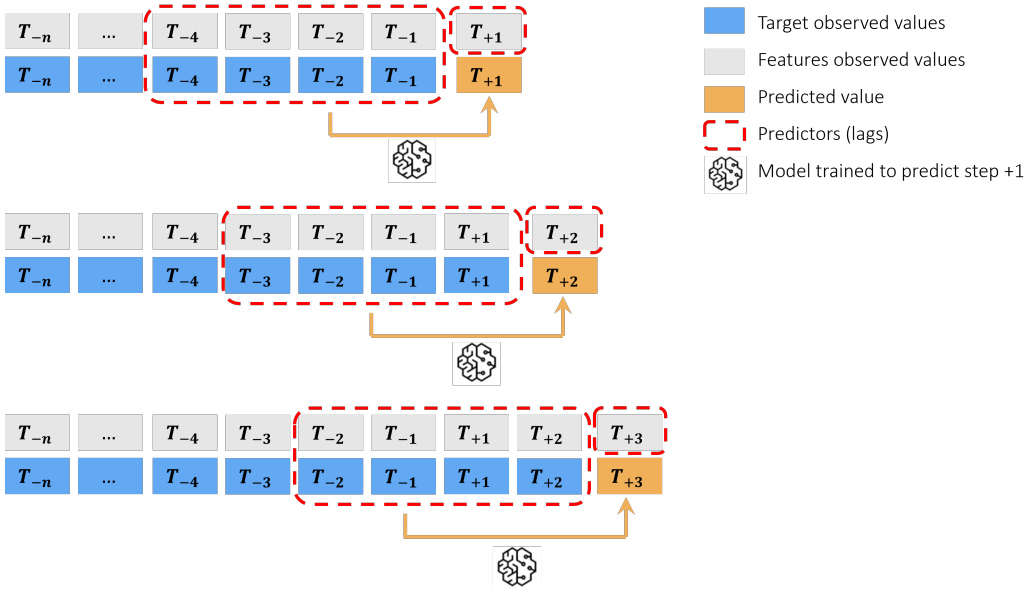

Figure 11: Recursive rolling forecast process diagram

The metrics that allow us to compare the performance of the models are presented in Table 4. In particular, we show the Mean Squared Error (MSE) for the 4 years of the test, as well as the overall MSE. We find that the AR model improves by 8% the prediction of the naive model, while the model that includes UNHCR covariates and exogenous variables improves by 15% the benchmark prediction. This gain is transversal to the period evaluated (2018-2021).

Model	**Root Mean Squared Error**					
	Testing period				**Overall**	
	2018	2019	2020	2021	RMSE	(Relative RMSE)
Naive	7956.9	3776.9	4608.1	3693.1	5302.2	1.00
AR	7305.7	3867.6	4425.2	2684.6	4876.5	0.92
AR + UNHCR	7287.4	3166.3	3297.6	2624.9	4497.1	0.85

Table 4: Root Mean Squared Error by periods

Since we have noticed that our data is highly unbalanced and contains many outliers, we evaluate the Root Mean Squared Error (RMSE) for different intervals of our target *newarrival*. The results are shown in Figure 12. The main takeaway from this analysis is that our rolling forecast is undoubtedly superior to the naive model only for large values of our target. However, there is much room for improvement in prediction at medium and small values.

Range	RMSE		
	Naive	AR	AR + UNHCR
0-10	2.0	4.5	3.9
10-100	36.2	74.2	68.9
100-1k	280.2	460.6	512.1
1k-10k	2386.5	3596.4	3104.1
10k-100-k	31011.2	35081.9	31471.7
100k+	240486.6	208526.7	191992.6

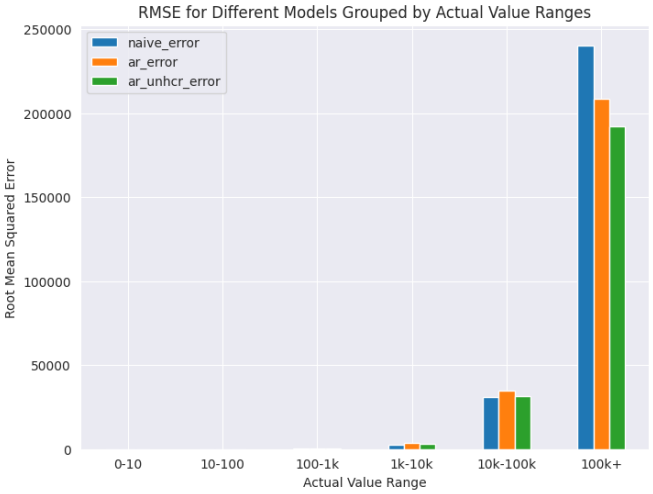

Figure 12: Root Mean Squared Error for specific intervals of refugee flows

Finally, Figure 13 shows the aggregate forecast predictions for the 2018-2021 period by continent. The results are heterogeneous, mainly because the test periods are characterized by abrupt jumps and drops, not observed in the past.

Figure 13: Refugee flow forecast by Continent

Some advantages of this model are associated with the fact that it works with conventional data, and therefore facilitates its interpretation. In contrast, we must also note disadvantages: (i) the model is independent of other country pairs, which does not allow for interaction and cross experiences with common events in a region or cluster; (ii) the model does not include high frequency and non-conventional variables to incorporate for example exit intentions; and (iii) the performance is systematically poor for medium or low scale ranges of our target, associated to the large imbalance of our data.