Forecasting Bilateral Refugee Flows with High-dimensional Data and Machine Learning Techniques

BSE Working Paper 1387| March 2023

Konstantin Boss, Andre Groeger, Tobias Heidland, Finja Krueger, Conghan Zheng

Forecasting Bilateral Refugee Flows with High-dimensional Data and Machine Learning Techniques[*]

Konstantin Boss Andre Groeger Tobias Heidland
Finja Krueger Conghan Zheng

March 2, 2023

Abstract

We develop monthly refugee flow forecasting models for 150 origin countries to the EU27, using machine learning and high-dimensional data, including digital trace data from Google Trends. Comparing different models and forecasting horizons and validating them out-of-sample, we find that an ensemble forecast combining Random Forest and Extreme Gradient Boosting algorithms consistently outperforms for forecast horizons between 3 to 12 months. For large refugee flow corridors, this holds in a parsimonious model exclusively based on Google Trends variables, which has the advantage of close-to-real-time availability. We provide practical recommendations about how our approach can enable ahead-of-period refugee forecasting applications.

Keywords: Forecasting, refugee flows, asylum seekers, European Union, machine learning, Google Trends
JEL codes: C53, C55, F22

[*]Boss: Universitat Autonoma de Barcelona (UAB), Barcelona School of Economics (BSE), konstantin.boss@bse.eu; Groeger: Universitat Autonoma de Barcelona (UAB), Barcelona School of Economics (BSE) and Markets, Organizations and Votes in Economics (MOVE), andre.groger@uab.es; Heidland: Kiel Institute for the World Economy (IfW), Kiel University, tobias.heidland@ifw-kiel.de; Krueger: Kiel Institute for the World Economy (IfW), Kiel University, and Institute for the Study of Labor (IZA), finja.krueger@ifw-kiel.de; Zheng: Universitat Autonoma de Barcelona (UAB), Barcelona School of Economics (BSE), conghan.zheng@uab.cat.

Acknowledgements: We thank Joop Adema, Joshua Blumenstock, Stefano M. Iacus, Hannes Mueller, Jasper Tjaden, Ingmar Weber, Emilio Zagheni and conference and seminar participants at the 12th CEPII-LISER-OECD Annual Conference on "Immigration in OECD Countries" 2022 (Paris), Workshop on New Data and New Methods for Migration Studies 2022 (Paris), Forecasting for the Social Good Workshop 2022 of the International Symposium of Forecasters (Oxford), Bertelsmann Foundation Migration Prediction Workshop 2022 (Berlin), and the Workshop on Computational Conflict Research 2022 at Ludwig-Maximilian University (Munich) for very useful comments. We acknowledge financial support from the European Commission's Horizon 2020 program through the project "IT Tools and Methods for Managing Migration Flows" (ITFLOWS; grant number: 882986). Boss, Groeger, and Zheng acknowledge financial support from the Spanish Agencia Estatal de Investigación (AEI) through the Severo Ochoa Programme for Centres of Excellence in R&D (Barcelona School of Economics CEX2019-000915-S). Boss also acknowledges support through grant FPI PRE2020-093941. Groeger also acknowledges financial support through grant PID2021-122605NB-I00 funded by MCIN/AEI/10.13039/501100011033 and by ERDF A way of making Europe, through MCIN/RYC2021-033208-I/AEI/10.13039/501100011033 and the European Union "NextGenerationEU"/PRTR, as well as from the Generalitat de Catalunya (2021-SGR-00571). The usual disclaimer applies.

1 Introduction

In 2013-2015, the European Union experienced a substantial increase in refugee flows, with more than 1.2 million asylum applications registered at its peak in 2015 alone.[1] The experience of 2015 has triggered widespread political debate and stressed the importance of preparedness for governments and humanitarian organizations to ensure safe conditions for asylum seekers and refugees en route and upon arrival. That increased political attention, combined with improved data availability and methodological and technological advances over the past decade, has moved the objective of migration forecasting to the center of scientific attention across different disciplines. Forecasting refugee flows is especially challenging due to the particular circumstances of forced migration, such as the lack of media reporting or data collection in conflict areas, making it particularly difficult to obtain early signals about international population movements. Additionally, violent conflict, natural disasters, or economic crises tend to occur unexpectedly and infrequently, which further complicates prediction (Disney et al., 2015).

In this context, Böhme et al. (2020) have demonstrated that migration-related Google Trends Indices (GTI) recorded in migrant origin countries hold additional *in-sample* predictive power over classical predictors when explaining bilateral migration flows to the OECD in a yearly gravity-type regression framework.[2] Building on that proof-of-concept paper, we create a realistic policy prediction framework (Kleinberg et al., 2015) – simulating the information set available to the analyst including Google Trends – and assess the *out-of-sample* performance of different forecasting model specifications using bilateral high-frequency flows of asylum seekers to the European Union. Google Trends allows studying keyword-specific search behavior at an aggregate, anonymous level to follow trends by country over time. The platform enables us to track the search interest in certain keywords such as "passport" or specific (destination) country names in migrant countries of origin over time, thus learning about interest in migration in general, and certain bilateral corridors in particular.

[1] That peak constituted an increase of nearly 130% compared to the previous year alone. Recent figures show that asylum registrations have once again reached similar levels in 2022 (Eurostat, 2022).

[2] Most people who perform online searches for information, knowingly or unknowingly, use the Google search engine and their queries are thus captured by Google Trends. Google has a market share of well over 95 percent of the search engine market in the vast majority of countries of the world. China is the major exception with most searches being done on Baidu.

This paper provides three main contributions. First, we construct a data set of asylum seeker registrations by applicant nationality of unprecedented frequency and scope featuring a real-life practitioner's information set to forecast asylum seeker arrivals at destination. To this end, we combine a range of Google Trends time series capturing internet searches for migration-related keywords at origin with administrative records of asylum registries at destination as well as hundreds of predictor candidate variables capturing the occurrence of natural disasters and violent conflict at origin and a wide range of socio-economic, labor and monetary statistics in both origin and destination countries. The richness of this dataset goes clearly beyond those that are typically used in the migration literature and allows us to exploit monthly frequency in the panel of bilateral refugee flow corridors.[3] Second, we compare the forecasting performance of different models and approaches that are common in data science forecasting applications. We test model performance in a forecasting-at-scale exercise where the models are assessed based on their average forecasting power over many migration corridors and across a range of different specifications that involve different compositions of information sets, including high-frequency Google Trends. In other words, our objective is to identify the best aggregate performing forecasting model specification, composed of a particular combination of available predictor candidates, including Google Trends. This is inherently different from existing work, which has focused on assessing corridor-specific customized forecasting approaches using a single algorithm (Carammia et al., 2022). Our approach allows us to analyze model performance differences at the corridor- as well as at the aggregate level. Third, we provide novel insights into the forecasting power of Google Trends data by systematically comparing forecast performance across specifications with and without these indices, as well as assessing their performance in specifications using only Google Trends predictors. Given that Google Trends are available at up to daily frequency and close to real-time, the latter specification, if well performing, would hold particular attractiveness for the analyst as it is independent of data publishing lags, which typically restrict the viability of short-term forecasts decisively.

Our results show that the ensemble forecast composed of the Random Forest (RF) and Extreme Gradient Boosting (XGBoost) algorithms based on a host of classical push- and pull-factor in combination with around 200 unique GTI predictors consistently outperforms alternative machine learning methods such as Elastic Net and Factor Approaches.

[3] By corridor we mean a single origin-destination relationship, such as migration from Afghanistan to Germany.

The ensemble forecasting model also performs better than our evaluation benchmark - the Random Walk[4] - for forecasting horizons of 3, 6, and 12 months forecasts out-of-sample. We also show that these positive results carry over to a more parsimonious ensemble forecasting model exclusively based on GTI predictors for the group of large refugee flow corridors of high policy-relevance. This simplified model is available on a close to real-time basis, independently of the data availability and publishing lags of classical predictor variables. The results thus suggest that our approach may be successfully exploited in refugee forecasting tools to create ahead-of-period predictions for forecasting horizons of between 3 and 12 months.

2 Literature on Migration Forecasting

The literature on international migration is rich and multi-disciplinary. Here, we briefly review a selection of the studies most closely related to our research. For an overview of the models used and common data sources, see Appendix A.1 and A.2. Since refugee migration flows depend on many social, economic, and political factors, forecasting them is a highly complex and complicated undertaking. Shock events such as violent conflicts, economic crises, or policy changes make accurate prediction a difficult task. Furthermore, forecasts may be subject to large prediction errors because the data quality of migration and refugee flows is often poor (Disney et al., 2015). A related problem is the absence of a universal theory to explain the push- and pull factors of migration and refugee flows, which could provide a guideline regarding the choice of variables to include in such forecasting exercise (Bijak et al., 2019). For these reasons, the migration forecasting literature has focused largely on developing early warning systems aimed at i) predicting short-term fluctuations or ii) on in-sample exercises to gauge the importance of different push- and pull factors empirically. For example, Napierała et al. (2021) develop an early warning model using high-frequency (i.e., weekly and monthly) data of asylum application registrations in several European countries. Their model is inspired by statistical control theory and generates alerts whenever they detect that a threshold in the number of asylum applications is passed. Shellman & Stewart (2007) build an early warning model predicting civil violence, poor economic conditions, and foreign interventions, all factors leading to an increase in forced migration. The model is then applied to the case of Haitian citizens

[4] The Random Walk prediction is a "naive" forecasting technique in which the forecast for the next period is equivalent to the value of the dependent variable of the previous period.

fleeing to the US on a weekly basis. While the model performs well in the specific case under consideration, the paper does not investigate out-of-sample performance in other corridors.

In a paper closely related to ours, Carammia et al. (2022) develop an early warning and forecasting pipeline using data from Google Trends applied to monthly asylum data for EU destination countries to forecast arrivals. While their research objective and important data sources coincide with those used in our paper, their approach differs significantly from ours. Most importantly, the authors assess a *corridor*-specific forecasting approach using a single algorithm (here: Elastic Net). They apply a three-step procedure that involves an early warning step that selects input variables for the forecasting steps. The latter consists in first estimating an Adaptive Elastic Net on the selected variables and, second, using standard time-series models to forecast the explanatory variables retained by the Elastic Net as new data inputs. This is different from our approach which leaves the predictor selection task fully to the respective forecasting model, thus keeping all potential information available for prediction *ex-ante* and reducing uncertainty arising from introducing additional modeling steps. Consequently, our ambition is to assess the forecasting performance at scale across many corridors as opposed to creating a corridor-specific model. In addition, they use Google Trends bundled in "topics", which are clusters of keywords. Such clusters leave no control over the choice of keywords and cannot be combined with destination country names as they are pre-specified by Google. Using customized keywords allows us to construct migration-related "bilateral" search terms tailored to the specific origin-destination relation such as "visa Germany". This approach yields corridor-time-specific variation in Google Trends, which likely increases the predictive power of our GTI information set. Lastly, the forecasting performance in their paper is evaluated against an ARIMA(1,0,1) model in terms of percentage errors for a narrowly selected time period between 30 April 2017 to 1 September 2019 – a period which was characterized by stationary asylum lodgings that fluctuated around their mean values for many migration corridors. However, such stationary models are not appropriate when modeling trending episodes in the data, which is precisely one of our objectives. The important spikes in migration flows during the refugee migration episode of the years 2015/2016 are, therefore, left out of their assessment. We take a different approach and include all data available from our sources and cover the maximum time period available between January 2008 and April 2021. This implies that we are testing our models on the challenging task of predicting all large forced migration events during that time period across all corridors, including the major refugee spells of the years 2015/2016.

3 Data

To test the out-of-sample performance of the previous forecasting models, we rely on a panel data set of bilateral asylum seeker flows between 150 origin countries to the EU27 destination countries with monthly frequency. We combine this panel with a wide range of potential predictor variables, composed of almost 200 migration-related Google Trends variables and "classical" push- and pull-factor predictor variables. Our primary outcome variable is the monthly number of asylum applicants registered in an EU member state by country of origin of the applicant as provided by Eurostat. For the collection of Google Trends data, we follow Böhme et al. (2020) and use a list of migration-related search terms that we collect in several languages. In what follows, we describe the data extraction and construction procedure and provide brief descriptive statistics.

3.1 Asylum Seeker Flows

The composition of the panels we use in the forecasting exercise is dictated by the availability of the target variable, i.e., asylum seeker flows. The Eurostat database contains asylum applications from individuals of international origins to the EU27 plus the United Kingdom and Norway. The first observations date back to January 2008.[5] We drop countries with populations of less than 100,000 from the sample. Moreover, we discard origin countries without monthly migration flow data.

In particular, the variable we use to capture refugee flows is "asylum applicants" as registered in a member state by country of origin of the applicant. This variable is recorded as a continuous count variable, rounded to five. In the Eurostat database an asylum applicant is defined as a person who has submitted an application for international protection or was included in such an application as a family member during the

[5] Since the refugee data provided by The European Border and Coast Guard Agency (Frontex) aggregates border-crossing flows by routes rather than by destination country, we choose the Eurostat data source, which enables us to introduce a bilateral dimension of refugee flows from a specific origin to a specific destination. See, `https://frontex.europa.eu/we-know/migratory-map/`, accessed February 26, 2023.

reference period. By definition, asylum-seeking individuals differ from other types of people on the move (who may be engaged in voluntary movement). While they may cross a border *irregularly* before submitting their application for international protection, asylum seeker movement is considered a *forced displacement*, or involuntary.[6]. Given the original aim of the ITFLOWS project to assist policymakers and civil society in managing "mixed migration" flows to the EU, including both irregular migration and refugee flows, we consider the number of asylum applications the most suitable outcome variable for this prediction exercise. Also, given the continuous nature of this variable, it is more challenging to predict, compared to alternative options such as a binary indicator for an increase or decrease in the number of applicants.

3.2 Languages

To determine the languages for which we want to extract Google Trends data, we proceed in the following way. First, we collect all country names that were defined as either "origin countries," "transit countries," "potential additional origin countries," and "watchlist origins" within the ITFLOWS project. Second, we cross-reference the languages spoken in these countries with the languages associated with each country in the database of Melitz & Toubal (2014). While there are many more languages spoken in these countries, we focus on languages that are defined as "official languages". Their definition of "official languages" requires a language to be the official language as per the CIA World Factbook in at least two countries. This approach results in the following languages being considered in this paper: English, French, Arabic, Spanish, Portuguese, Turkish, Farsi, Pashto, Hausa, and Fulfulde.[7]

The origin countries are then sorted into the different categories by the languages

[6] The terms "migrant", "refugee" and "unrecognized refugee" are understood as follows: In its widest scope and for some authors, the term "migrant" also includes asylum seekers and refugees. However, we emphasizes that the two latter categories are regulated by additional instruments and are given specific guarantees in international and European law. Individuals who have been formally granted refugee status are understood as "recognized refugees." Any person (asylum applicant, irregular migrant or not) who meets the eligibility criteria, but have not applied or applied and have not yet been granted asylum by a state is understood as a "non-recognized refugee" (to be distinguished from unsuccessful asylum applicants); and both recognized and unrecognized are refugees.

[7] We widen the set for Arabic by including also the Algerian, Moroccan, Egyptian, Saudi Arabian, and Hassaniya versions, as well as the Afghan type of Farsi called Dari. Comparing some results, there were no substantial gains from tailoring the languages to the specific dialects, which is why we focus simply using standard Arabic and Farsi in the particular cases mentioned.

spoken in each country. To associate each origin country to one or more languages, we use the database provided in Melitz & Toubal (2014). We do not discriminate between spoken, common, and official languages. If a language is referenced as either of the three, we associate the origin country with that language. Moreover, those countries that are not in the Melitz & Toubal (2014) database are added to the English origins list unless an official language from our set can be identified from other sources, such as the CIA World Factbook. As shown in Appendix Table 2, the categorisation by language leads to 86 English, 30 French, 27 Arabic, 24 Spanish, 9 Portuguese, 7 Turkish, 6 Farsi, 6 Fulfulde, 2 Hausa, and 2 Pashto origin countries.

3.3 Keywords

The selection of keywords is crucial for our approach, as we need to cover a set of search terms that refugees will likely look up before (or during) their trip. To cover a broad and relevant set, the keyword selection strategy rests on two pillars. First, we follow Böhme et al. (2020) by extracting several terms related to each of the two base keywords "immigration" and "economic" through the website *Semantic Link*. This service checks the English version of the *Wikipedia* encyclopedia for the most common co-occurrences between the given keywords and words that appear at least 1,000 times in the database. By choosing the initial terms "immigration" and "economic", we aim to cover two main fields of search terms related to migration intent. Of the 200 keywords generated in this way, we discard 27 already included in the set of 67 keywords in Böhme et al. (2020). Then, we manually exclude 20 terms from the list which are seemingly ambiguous or irrelevant. These are mostly numbers, names, and acronyms. In total, from this first exercise, we obtain a list of 153 keywords.

Second, we feed a list of keywords consisting of the 153 terms from the *Semantic Link* exercise and the 40 unique keywords remaining from Böhme et al. (2020) for a total of 193 keywords into the Google Trends API to fetch the so-called *related queries* in each of the 86 English-speaking countries of origin. The *related queries* category outputs the Google searches similar to the keyword the user provided to the API, which occurs most often in a given geographic area. For example, queries related to the search term "visa" are "application", "authorization", or "lottery". By including such related queries, we aim to complement the *Semantic Link* list with frequently searched terms related to our two categories of "immigration" and "economic" conditions. The output from this exercise is then manually scraped for the most relevant new terms, which yields 37 additional terms. Manual selection is necessary since many terms are not strictly related to human migration, such as the word "server migration" which refers to an information technology process.

Combining the lists obtained from the two exercises and the unique terms in Böhme et al. (2020) result in a list of 230 keywords. Since, in many cases, we have singular and plural forms (e.g., "migrant" and "migrants") or semantically similar terms (e.g., "application" and "applicant"), we combine these terms using the search operator "+". This operator ensures that the query delivers all searches for all terms connected by it. For example, a Google Trends search for "migrant+migrants" will provide data for a combination of the interest over time in both terms. By combining terms in this fashion, we obtain a final list of 192 keyword expressions. The full list is reported in Appendix Table 3.

This English list is the starting point for translating all terms into the selected languages. In the case of the Roman languages in our set, we adapt the translation always to include male/female adjustments and accented/not-accented versions when applicable. Again, we add these additional spellings using the "+" operator to keep the different lists per language of equal lengths and comparable.

3.4 Google Trends Data

An advantage of our Google Trends customized search term approach is that we can exploit bilateral variation in search intensities in origin countries concerning a specific destination country. To obtain these, we interact the 192 keywords with each of the EU27 country names (see Appendix Table 4) to obtain bilateral search terms. That yields a combination of topical keywords related to migration in relation to a specific destination country, e.g., "visa Germany" or "embassy France". For each origin country, we add these 5,184 (192×27) bilateral terms to the list of keywords. This list also contains the original list of 192 topical keywords and the individual EU27 destination country names separately. Examples are search terms like "visa" or "unemployment" for topical keywords and "France" for EU27 destination country names. The final list has 5,403 entries and is the same length for each language considered.

We extract the Google Trends data for each of the terms specified on the final list of keywords through an API provided by Google. We proceed by language groups to account for the fact that people likely use Google searches in their native language. We collect this data at a monthly frequency between January 2004 and April 2021, the maximum time period available at the time of data construction. A noteworthy measurement issue with Google Trends data is that the API returns zero values when the search volume falls below an undeclared threshold. The resulting positive values of the index reflect what Google terms "interest over time" and are calculated as a fraction of searches for the given term relative to the total number of searches over a month in the same country. In the second step, the data for the entire period is automatically scaled to values between 0 and 100 such that 100 corresponds to the highest relative search intensity in a month measured over the entire period.[8] Table 3 lists all 192 search terms and provides descriptive statistics for each resulting time series.

3.5 Classical Predictors

We also collect a host of monthly economic indicators for each origin country and each destination country in our sample. These variables correspond to more "classical" predictors as used in the economics literature on migration and are supposed to capture various migration push- and pull factors, as listed in Appendix Table 5 and 6. We have an abundance of variables to choose from in the macroeconomic and agricultural domains, with over 300 variables collected for each. In addition, we have over 100 variables that sufficiently cover conflict and disaster data, political variables, labor force data, and short-term business data.

The first set of "classical" predictors of asylum seeker flows are indicators capturing socio-economic dimensions on both sides of the migration corridor. For example, macroeconomic and financial indicators at the origin country level, such as GDP growth or Consumer Price Index, provide proxies of the current economic situation during a specific month of observation and may provide signals on push factors. In contrast, the same variables at the destination level constitute proxies for the attractiveness of the EU27 country and may thus capture pull factors. For monthly macroeconomic statistics, we

[8] For example, suppose there are only three months in the sample, and the fractions of searches for the keyword "visa" relative to the entire search activity in a given country are 0.1, 0.2, and 0.5. Then 100 would be assigned to the third month, 40 to the second, and 20 to the first.

combine sources including the Statistical Office of the European Union (Eurostat), the International Labour Organization (ILO), and the International Monetary Fund (IMF).

An essential push factor is human-caused or natural shocks in the origin country. Adverse environmental shocks and political conflicts may decrease income and cause social instability. Political events and indicators from the Rulers, Elections, and Irregular Governance Dataset (REIGN), political violence indicators from Armed Conflict Location and Event Data Project (ACLED), and disaster indicators from International Disaster Database (EM-DAT) are merged into the monthly panel. Event entries are aggregated into monthly counts.

4 Forecasting Methodology and Performance Comparison

Migration data is often marked by changing dynamics and fluctuations in the set of relevant predictors. Rossi (2021) discusses the effect of such instabilities on forecasting methods and performances and suggests that Big Data approaches may help to improve forecasts. In particular, she suggests considering three different approaches: First, aggregating, then forecasting as in principal component models. Second, forecasting while aggregating as in regularization models. Third, forecasting and then aggregating by computing ensemble predictions across different models. Based on these suggestions, we select a range of empirical models and combinations thereof to forecast migration flows that have been used in the migration or violent conflict forecasting literature and then rank them by their forecasting performance. The models we compare are the Elastic Net (Carammia et al., 2022), the Random Forest (Mueller & Rauh, 2022), as well as the Extreme Gradient Boosted Regression Tree and factor models. To assess the forecasting performance of these approaches, we compare the forecasts to those obtained from our benchmark model, which is obtained by using the last available data point as the forecast at all horizons. We refer to this benchmark as the Random Walk (RW) forecast. For a discussion of the weaknesses of our approach, see Appendix C.

The model comparison is carried out individually over all possible bilateral relationships of countries of origin and EU27 destinations in our monthly data set. The variable to be forecast is the number of *asylum applications* by individuals of nationality from a country of origin o in a country of destination d, which is the best proxy of refugee flows available to us for the member states of the European Union. The information set used to make the forecast is denoted by X_{od} and contains different combinations of migration push- and pull factors such as, e.g., socio-economic characteristics in origin- and destination countries, migration-related GTI variables, and the lagged dependent variables. To compare the performance of different model compositions in terms of the information set, for each algorithm, we perform three sets of forecasts: first, motivated by the migration literature in economics using bilateral flow data in a gravity-type regression specification, we specify the predictors to be only the "classical" push- and pull factors. Second, based on the first specification, we add our GTI data to the vector of predictors. Third, we perform the forecast using only GTI variables without any "classical" predictors. The latter specification is particularly appealing for the analyst as it involves only one data source (Google Trends) available at up to daily frequency starting in 2004 and with a maximum publishing lag of 24 hours. That provides obvious advantages for the viability of conducting forecasts close to real-time.

We conduct a moving-window pseudo-out-of-sample exercise to assess the accuracy of the out-of-sample forecasts. That is, we use a window of constant size $w = 50$ to train each model and predict the number of *asylum seekers* at different forecast horizons, namely $h = 1, 3, 6, 12$ months ahead. The choice of the window size is arbitrary, but experimenting with different sizes has shown that longer training sets do not necessarily improve performance or change the ranking of models. Importantly, given that the sample starts in January 2008, this choice implies that all models are asked to forecast the large migration change in 2015, which is a novelty in the migration prediction literature. We obtain all forecasts using a direct approach. This requires offsetting the dependent and independent data by the forecast horizon for training and then feeding the latest available independent data to the trained model for prediction. Although we do not use real-time data for training or evaluation, this is a proper out-of-sample forecast. Importantly, we never include lags of the variable *asylum seekers* as an explanatory variable, as this may render especially short-term predictions unfeasible due to the publication lag in migration-related data. Moreover, note that the explanatory data contains only the lag corresponding to h and no deeper lags that would add potentially useful information. We check lag-augmentation as a robustness exercise and find limited use for it.

Motivated by our objective, we select three forecasting models that can deal with variable selection, which is a crucial criterion given our rich data set. These are the Elastic Net, the Random Forest, and the Extreme Gradient Boosted Regression Tree (XGBoost). While the Elastic Net is a linear model, the latter two can accommodate non-linearities. For performance comparison purposes, we rely on the Theil ratio (T), which we compute as the ratio of the root mean squared error ($RMSE$) statistics of the candidate model over the Random Walk, as follows:

$$T = \frac{RMSE_X}{RMSE_{RW}}$$

with X representing the respective model under investigation.[9] A value larger than 1 thus indicates that, on average over all migration corridors of the selected sample, the Random Walk outperformed the specific model. A value smaller than 1 implies that the candidate model outperforms in terms of forecasting performance over the forecasts using the Random Walk. In what follows, we briefly discuss the forecasting models under investigation.

4.1 Elastic Net

The Elastic Net (EN) is a linear regression model which allows for parameter shrinkage according to the following penalized regression:

$$\hat{\beta}_{EN} = arg \min_{\beta} |y - X\beta|^2 + \alpha\lambda|\beta| + (1-\alpha)\lambda|\beta|^2$$

Parameter regularization is performed against the L^1 (LASSO) and the L^2 (Ridge) penalties which ensure model selection through the LASSO part and allow for some control for multicollinearity through the Ridge part. We set the weight on each of the two to $\alpha = 0.5$ as in Carammia et al. (2022) and select the regularization parameter λ as the one yielding the smallest MSE in training.[10]

4.2 Random Forest

The Random Forest (RF) is an ensemble prediction algorithm that relies on decision trees (Breiman, 2001). In each decision tree, the algorithm finds the optimal variable to split

[9] The RMSE is calculated as: $\sqrt{(\frac{1}{n})\sum_{t=1}^{n}(\hat{y}_{t|t-h} - y_t)^2}$, with h being the respective forecasting horizon.

[10] The errors are assumed to follow a Poisson distribution.

the input data, thus creating a node. The metric for computing the optimal split for our regression problem is the MSE. Nodes in the tree are grown using bootstrapped data sets ("bagging") obtained from the original inputs. While individual decision trees can be prone to bias, averaging across many individual trees can produce predictions that are less driven by idiosyncratic error. Three hyperparameters have to be set to run the analysis. Firstly, the number of trees, which we set to $10,000$ to ensure that the many variables in our data set are used with high enough frequency in the regression trees. Secondly, the number of randomly chosen features by each tree m. Reducing the number of features chosen in each tree reduces both correlation and strength. Strength is defined as the error rate of each tree. Breiman (2001) shows that the forest error rate depends negatively on the correlation between any two trees and positively on the strength. We follow the rule-of-thumb recommendation to set $m = p/3$ where p is the number of variables. Thirdly, the depth of each tree represents the number of splits of each tree in the forest. We only require the standard minimum of five terminal nodes but allow the trees to be more complex at the cost of computational speed.

4.3 Extreme Gradient Boosting Regression Tree

While the RF is an extension of decision trees by bagging, the Extreme Gradient Boosting Regression Tree (XG) advances the model by boosting it (Friedman, 2001). Building upon weak learners (in our case, the decision trees), the model calls the original method repeatedly, using a different subset of the data each time. By sequentially fitting over the residuals (MSE) of the previous weak learner, the model finally aggregates the results of the steps into one strong learner. While the sequential setup increases computational time, the model's prediction often has higher accuracy. For extreme gradient boosting, the first hyperparameter that needs to be set is the number of rounds B the model is run, which we keep at a high value of $10,000$ to ensure convergence of the loss function at the risk of potentially overfitting the sample. The second hyperparameter is the learning rate, $\eta \in (0, 1)$, which controls the contribution of each past tree to the current approximation. By decreasing the learning rate, the model becomes more robust to overfitting but requires more computation time. We choose a low value of $\eta = 0.3$ to counter the overfitting problem we may obtain from the high number of rounds.

4.4 Factor Approach

Finally, we use a factor approach to forecast the following model for asylum applicants.

$$y_{t+h} = c + \beta' f_t + \epsilon_{t+h}$$

Kim & Swanson (2018) show that in the context of economic forecasting, such models can outperform simple benchmarks. Stock & Watson (2002) show that the factors f_t can be consistently estimated using the Principal Component (PC) estimator. After extracting PCs from the set of predictors, the model is estimated assuming a Poisson error distribution.

4.5 Ensemble Forecasts

Ensemble forecasts are constructed by equally weighting the forecast made by each model in the given combination and summing up. Suppose, for example, the Random Forest predicts 100 asylum seekers, and the XG Boost method suggests 120 asylum seekers. Given that there are two models in the combination, the equally weighted ensemble forecast would be $1/2(RF + XG) = 110$. Such ensemble forecasts can sometimes improve upon single model forecasts (Kim & Swanson, 2018).

5 Results

We start the discussion of our forecasting exercise by focusing on the best-performing model, the ensemble forecasting model composed of the Random Forest and XGBoost algorithms. The main results from the ensemble model are summarized in Figure 1.[11] The figures depict the Theil ratio for six subsamples of bilateral corridors of different importance for refugee flows, ranging from the top 20 most important corridors, in terms of total asylum seeker numbers over the 2008 to 2020 period, to the top 1000 corridors.[12] Each

[11] We report results based on the pure RF and XGBoost models in Appendix Figures 3 and 4, respectively. The general results for all models are reported in Appendix Table 7.

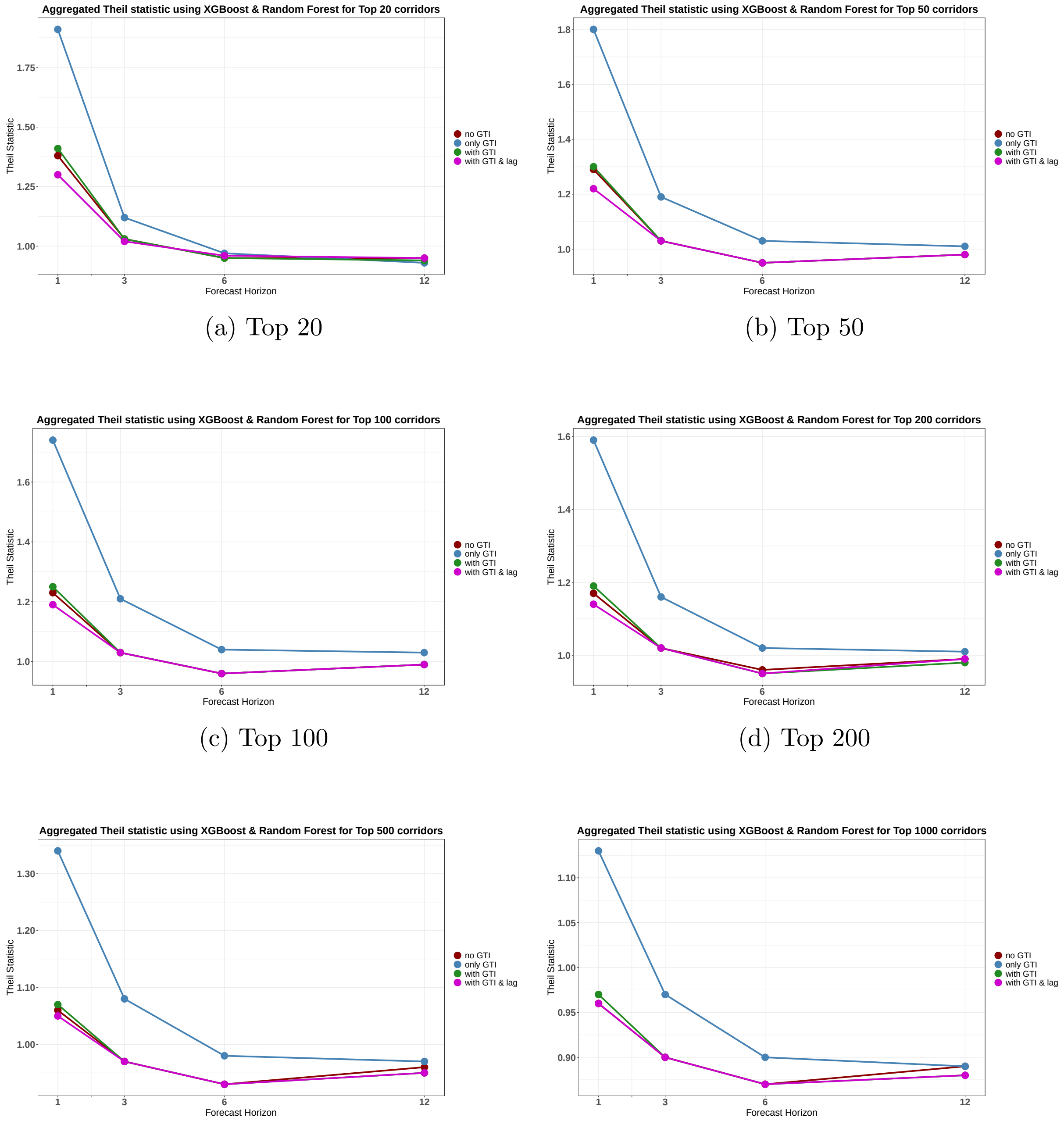
[12] In the case of the top 20 corridors of asylum seeker flows to the group of EU27 destination countries this includes (in order of magnitude of total registered asylum cases by nationality of origin over the period 2008-2020): Syria-Germany [719,650], Afghanistan-Germany [265,215], Iraq-Germany [249,040], Syria-Sweden [130,265], Serbia-Germany [124,090], Nigeria-Italy [121,255], Venezuela-Spain [110,795], Albania-Germany [95,040], Iran-Germany [91,430], Syria-Greece [84,330], Pakistan-Italy [78,440], Syria-Hungary [78,355], Colombia-Spain [75,860], Eritrea-Germany [73,995], Afghanistan-Hungary [73,885], Russia-Poland [72,990], Afghanistan-Greece [71,680], Afghanistan-Austria [71,420], Afghanistan-Sweden

panel (a)–(f) reports results from a different subsample. In each panel, the horizontal axis represents the four forecasting horizons of 1, 3, 6, and 12 months and there are four lines corresponding to different specifications of the information set used in each forecast (i.e., no GTI, only GTI, with GTI, and with GTI and lagged dependent variable). The vertical axis, on the other hand, reflects the Theil ratio of the RF-XG ensemble model against the benchmark forecast based on the RW.

We start by describing the general findings that emerge from comparing the out-of-sample forecasting performances across the six subsamples. A recurrent pattern in all panels is that the shape of the lines is convex with an L-shape, indicating that the performance of the RF-XG ensemble model is generally better for longer forecasting horizons. In particular, performance for the one-month forecasting horizon is relatively poor, as reflected by the Theil statistics larger than one, especially in panels (a) through (e), indicating that the RW clearly outperforms, on average, in the case of short-term forecasts in the top 500 group. We believe that this is due to two factors: First, the RW "forecast" may be relatively more powerful in the very short term in which the temporal proximity to the last month is small, and hence the number of asylum seekers of the last period may be a good approximation for that of the current period. Second, the predictive power of our candidate models hinges on the fact that classical predictors and/or digital trace data measured at the origin during the previous time period (i.e., last month) are predictive of registrations of asylum seekers from that origin at the destination. Considering the top 20 corridors, it may easily take several months for irregular migrants and refugees to arrive at their desired destination and register for asylum. If this is the case, we expect our approach to have low predictive power in the very short term.

A second general pattern relates to the relative performance of the four different compositions of the information set, depicted by each line. The GTI-only model (blue lines) generally performs relatively worse than the other specifications. This model shows the worst performance over all horizons for panels (b) through (e), indicating that, despite its attractiveness due to the ease of implementation, the GTI-only model is not a silver bullet for refugee flow forecasting. In other words, classical predictors seem to carry important signals that cannot be substituted entirely by digital trace data. Furthermore, for the remaining three models composed of only classical predictors (brown line), classical plus

[69,265], Syria-Austria [68,500].



(a) Top 20

(b) Top 50

(c) Top 100

(d) Top 200

(e) Top 500

(f) Top 1000

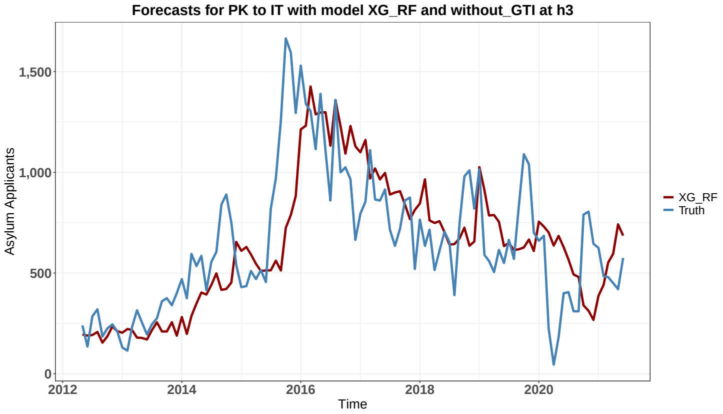
Figure 1: Main forecasting results from the ensemble model composed of the Random Forest and XGBoost algorithms for six subsamples of bilateral corridors of different importance for refugee flows (top 20 – top 1000 corridors in terms of total asylum seeker numbers over the 2008 to 2020 period) over four forecasting horizons (1, 3, 6, and 12 months) and four specifications of the information set (no GTI, only GTI, with GTI, and with GTI and lagged dependent variable. The curves in each panel depict the Theil ratio of the ensemble model against the benchmark forecast based on the RW. Source: Author calculations.

GTI (green line), and classical plus GTI and lagged dependent variable (purple line), the performance differences are often minor, indicating that the GTI variables provide limited additional predictive power over the (large) set of classical predictors used. Next, we analyze the relative performance of the RF model in comparison to the RW.

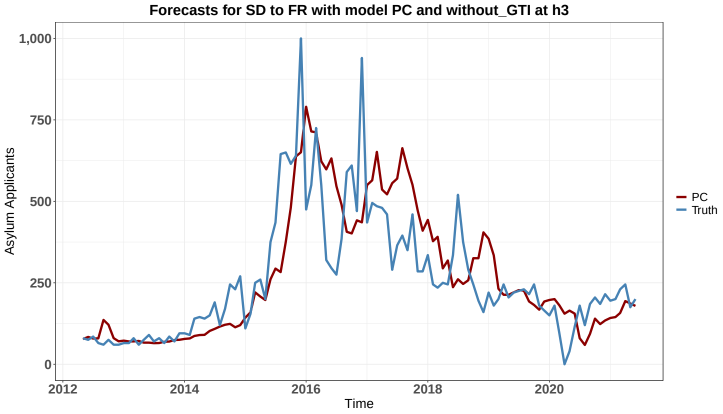
In all subsamples, we observe that the Theil statistic drops below one for any model including classical predictors at longer horizons, indicating that the ensemble model outperforms the RW. The precise horizon from which the ensemble model is superior to the RW in terms of forecasting performance varies between 6 and 1 months horizons for the top 20 to the top 1000 corridor subsamples, respectively. Interestingly, the magnitude of performance gains increases from high-importance to low-importance corridors. One of the reasons behind this pattern is that there tend to be fewer large shocks in refugee flows in low-importance corridors such that those time series are stationary, which, in turn, can be forecasted more accurately by the models. It is important to emphasize that the lines in Figure 1 represent the simple average of Theil statistics for each sample, respectively. This implies that the results for specific corridors may be significantly better or worse within the respective subsample.

We now turn to the role of our GTI predictors for the forecasting exercise. Focusing on the historically largest corridors for asylum seeker flows in absolute numbers in panel (a), we find that the GTI-only model outperforms the RW at horizons of six months and above. Further, the GTI-only model also performs equally well at $h = 6$ and even slightly better than any other specifications of the information set at $h = 12$. Since the curves represent average results across all corridors included in the sample, it is clear that the GTI-only model performs very well for some of these corridors. Due to the obvious implementation advantages of a forecasting model based on a single data source, this underlines the prospects of the GTI approach for selected policy prediction applications.

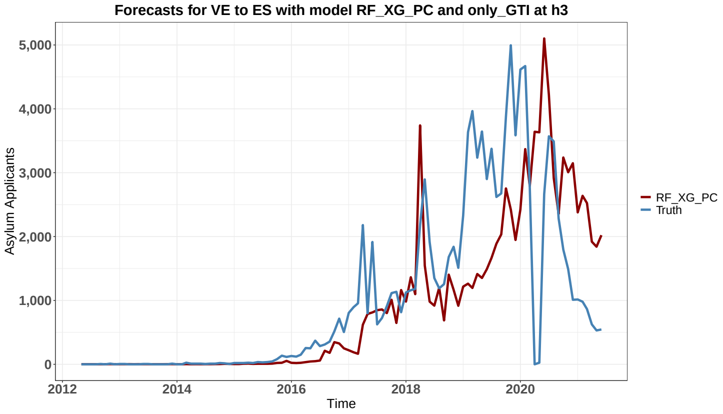
We now turn to the discussion of the performance of the Elastic Net and Factor approach using principal components, as reported in Appendix Table 7, which belong to the group of generalized linear regression models. Such models can extrapolate trends into the future, making them suitable for migration forecasting as shown, for example, in Carammia et al. (2022) for the case of EN. Given the large number of forecasts across corridors, horizons, and pseudo-samples we have to carry out, such forecasts must be



(a) Forecast using RF-XG ensemble model in no-GTI specification for the corridor Pakistan to Italy at 3 month horizon



(b) Forecast using PC model in no-GTI specification for the corridor Sudan to France at 3 month horizon



(c) Forecast using RF-XG-PC ensemble model in only-GTI specification for the corridor Venezuela to Spain at 3 month horizon

Figure 2: Time series plots comparing recorded asylum seeker flows (truth) to selected out-of-sample forecasts

appropriately automated. We follow their approach by setting the weight on the LASSO and Ridge component of the Elastic Net equal to $\alpha = 0.5$. The regularization penalty λ is chosen on a grid of 100 potential values on a log scale where the last value regularizes all parameters to zero. We select the λ associated with the smallest in-sample MSE. Regarding the principal components, a crucial choice is the number of principal components to use for the model. Typically, the variation in large data sets of socio-economic variables can be described by a small set of principal components that are useful for forecasting (Stock & Watson, 2002). While statistical criteria are available to determine the (in-sample) number of principal components to be used (Bai & Ng, 2013), this would render the exercise extremely time-consuming. Instead, we use a simple threshold and include principal components if they explain more than 5% of the dataset's variation.

These choices for the principal components and elastic net forecasts work well in many forecasts. However, in some cases, both models can grossly overfit the data or simply fail due to a lack of variation in the dependent variable. In such cases, the forecast errors can be nearly infinitely large, which then affects the aggregate error statistics accordingly. We observe this phenomenon in our results for these models. While some of the overestimates are obvious and could potentially be corrected by a human forecaster in the loop, this problem can be critical in the case of full automation of the forecasting process. Since the error sizes depend on the migration corridor at hand, it is not possible (and not prudent) to introduce arbitrary cutoffs for excluding certain values from the evaluation of the model. Despite the very poor relative performance across our sample, we emphasize that on certain migration corridors with good variation in the dependent variable and with largely stationary behavior, both the Elastic Net and the principal components estimator yield valuable forecasts.

In addition to the relative performance measures compared to the RW as reflected by the Theil statistics, we also report absolute forecasts for selected corridors in Figure 2. The three panels report time series plots comparing the recorded asylum seeker flows, depicted by the blue lines (i.e., ground truth data), to the selected out-of-sample forecasts for specific forecasting models and horizons, represented by the red lines. The examples have been selected from the group of top 50 corridors conditional on outperforming the RW (i.e., from those with a Theil statistic below one).[13]

[13] Among the Top 50 corridors there are a total of 1436 corridor/model/information set/horizon com-

Panel (a) depicts forecasts based on our best-performing RF-XG ensemble model in the no-GTI specification for the corridor from Pakistan to Italy with a 3-month horizon. A visual inspection shows that the model performs relatively well in this corridor, as reflected by the high correlation between the blue and the red line. Reassuringly, the model predicts both pronounced increases and decreases relatively well, for example, the sudden increase in flows during the 2015/2016 refugee wave. However, the inspection of the ground truth data around the beginning of the COVID-19 pandemic in early 2020 also shows that the model performs worse in the face of sudden disruptive shocks. The related shutdowns during the pandemic led to a general halt of human migration and refugee flows, as can be seen by the sudden drop in the blue lines across all panels in the first and second quarters of 2020. This is precisely the period when larger forecasting errors can be observed and the red line reflects that the model would have clearly overpredicted refugee arrivals during that period (this applies to all three models depicted). Panel (b) depicts forecasts using the PC model in the no-GTI specification for the corridor from Sudan to France with a 3-month horizon. Note that this is an example of the PC model outperforming the RW in the specific corridor despite an extremely large average Theil statistic in the top 50 sample. Panel (c) depicts forecasts based on the RF-XG-PC ensemble model in the only-GTI specification for the corridor from Venezuela to Spain with a three month horizon. Here, the forecasting errors are generally negative, implying that the model overpredicts, especially starting in the year 2019. An extreme overprediction occurs during the pandemic when flows drop to zero and the model continues to overpredict thereafter. Despite these prediction errors, this example shows that the only-GTI approach can work relatively well for the top 20 corridors of high policy relevance. Overall the figures provide evidence of the good absolute forecasting performance of selected models and compositions of the information set, in the context of specific refugee flow corridors.

6 Towards a Feasible Forecasting Approach for Refugee Flows

This paper evaluates the practical feasibility of building a refugee flow forecasting model combining high-dimensional data with machine learning techniques. We implement an out-of-sample forecasting model as suggested by Böhme et al. (2020). We extract migration-

binations where the model outperforms the Random Walk. Among these, 103 come from the Elastic Net, Factor Models or their ensemble. The examples presented here are chosen purely for illustration.

related Google Trends time series that can be used as predictors for bilateral refugee flows and combine them with asylum seeker flows and an extensive range of classical predictor variables with monthly frequency to construct a high-dimensional data set for performance testing. We have built a bilateral refugee panel database capturing flows between more than 150 origin countries and the group of EU27 destination countries with monthly frequency. We then evaluate the out-of-sample performance of different statistical models against a naive forecasting procedure constituted by the Random Walk, which tends to be “hard to beat” in forecasting exercises.

The results are encouraging in the following sense. Among the machine learning models we analyze, the ensemble forecast composed of the Random Forest and XGBoost models shows the best average performance in our context. In particular, we have presented evidence that this model consistently outperforms the Random Walk for forecasting horizons of 3, 6, and 12 months forecasts out-of-sample, depending on the sample of corridors analyzed. Our results also show that when comparing the predictive power of the Google Trends predictors in the same specification to those of hundreds of “classical” predictors capturing different types of push- and pull factors, performance gains from the GTI predictors are marginal on aggregate. This underlines that digital trace data is not a silver bullet for refugee flow forecasting if the forecaster includes a large vector of predictor candidates proxying for a diverse set of migration push- and pull factors.

Yet, focusing on the subsample of the top-20 corridors in terms of aggregate refugee flows between 2008 and 2020, we have shown that the positive results carry over to a specification of the ensemble model exclusively based on GTI predictors. The latter has the practical benefit of being available on a close to real-time basis, independently of the data availability and publishing lags of “classical” predictor candidates such as GDP growth or consumer price indices. In this selected sample of high policy relevance, the GTI-only specification outperforms all other models on average, including the Random Walk, for forecasting horizons of 6 and 12 months. This average performance improves further when focusing on specific corridors within the group of top 20 corridors. Our results imply that the Google Trends indices we extracted offer positive predictive power for such corridors, which could potentially be exploited for a refugee forecasting tool customized to selected corridors of particularly large flows.

For policy applications, we thus recommend customizing the selection of forecasting models according to the maximum performance within a specific refugee corridor, following our approach. We have provided average results and selected examples that represent high-performance combinations of the forecasting model, the information set, and the forecasting horizon. For origin countries with poor push-factor data availability, we suggest including Google Trends data, particularly in cases where the forecast horizon is longer than three months. In spite of the wide information set used in this exercise, our findings also reflect the fundamental limitations of the forecasting exercise: anticipating large and sudden shocks, such as the COVID-19 pandemic, which lead to an unexpected halt in migration flows, is literally impossible. Our approach is not exempted from that limitation and the selected examples show this clearly for the forecasts during the year 2020 and beyond.

It is important to note that the performance statistics we present are averaged over the respective bilateral refugee flow corridors. As discussed, there is heterogeneity in the models' predictive performance across corridors. In other words, the models tested work better for some corridors and worse for others. In other words, for specific corridors, the single best-performing model may be different than the one suggested by the average results. This heterogeneity should be made transparent whenever forecasting results are published or used to inform policymakers and migration management agencies. Another disclaimer is that this exercise's only available benchmark for predictive performance is data from previous years. There is no guarantee that the proposed approach will successfully predict future flows. Anyone applying the method should investigate the approach critically in their specific use case before scaling it up. Yet, overall, our results show that the approach does help enabling refugee forecasting applications with ahead-of-period predictions.

References

Adema, J. A. H., & Guha, M. (2022). Following the online trail of ukrainian refugees through google trends. *CESifo Forum*, *23*(04), 62-66.

Alexander, M., Polimis, K., & Zagheni, E. (2019). The impact of hurricane maria on out-migration from puerto rico: Evidence from facebook data. *Population and Development Review*, 617–630.

Alexander, M., Polimis, K., & Zagheni, E. (2020). Combining social media and survey data to nowcast migrant stocks in the united states. *Population Research and Policy Review*, 1–28.

Alvarez-Plata, P., Brücker, H., & Siliverstovs, B. (2003). *Potential migration from central and eastern europe into the eu-15: An update.* European Commission, Directorate-General for Employment and Social Affairs ....

Bai, J., & Ng, S. (2013). Principal components estimation and identification of static factors. *Journal of econometrics*, *176*(1), 18–29.

Bijak, J., Disney, G., Findlay, A. M., Forster, J. J., Smith, P. W., & Wiśniowski, A. (2019). Assessing time series models for forecasting international migration: Lessons from the united kingdom. *Journal of Forecasting*, *38*(5), 470–487.

Böhme, M. H., Gröger, A., & Stöhr, T. (2020). Searching for a better life: Predicting international migration with online search keywords. *Journal of Development Economics*, *142*, 102347.

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

Burzynski, M., Deuster, C., & Docquier, F. (2020). Geography of skills and global inequality. *Journal of Development Economics*, *142*, 102333.

Carammia, M., Iacus, S. M., & Wilkin, T. (2022). Forecasting asylum-related migration flows with machine learning and data at scale. *Scientific Reports*, *12*(1), 1457.

Dao, T., Docquier, F., Maurel, M., & Schaus, P. (2018). Global migration in the 20th and 21st centuries: the unstoppable force of demography.

Disney, G., Wiśniowski, A., Forster, J. J., Smith, P. W., & Bijak, J. (2015). Evaluation of existing migration forecasting methods and models. *ESRC Centre for Population Change, University of Southampton*.

Dustmann, C., Casanova, M., Fertig, M., Preston, I., & Schmidt, C. M. (2003). *The impact of eu enlargement on migration flows* (No. 25/03). Research Development and Statistics Directorate, Home Office.

Eurostat. (2022). *Annual asylum statistics.* Retrieved 2022-03-18, from `https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Annual_asylum_statistics`

Fertig, M., & Schmidt, C. M. (2005). *Aggregate-level migration studies as a tool for forecasting future migration streams.* Routledge.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Golenvaux, N., Alvarez, P. G., Kiossou, H. S., & Schaus, P. (2020). An lstm approach to forecast migration using google trends. *arXiv preprint arXiv:2005.09902*.

Hanson, G., & McIntosh, C. (2016). Is the mediterranean the new rio grande? us and eu immigration pressures in the long run. *Journal of Economic Perspectives*, *30*(4), 57–82.

Hausmann, R., Hinz, J., & Yildirim, M. A. (2018). Measuring venezuelan emigration with twitter. *Kiel Working Paper*.

Kim, H. H., & Swanson, N. R. (2018). Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods. *International Journal of Forecasting*, *34*(2), 339–354.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015, May). Prediction policy problems. *American Economic Review*, *105*(5), 491-95.

Melitz, J., & Toubal, F. (2014). Native language, spoken language, translation and trade. *Journal of International Economics*, *93*(2), 351–363.

Mueller, H., & Rauh, C. (2022). The Hard Problem of Prediction for Conflict Prevention. *Journal of the European Economic Association*, *20*(6), 2440–2467.

Napierała, J., Hilton, J., Forster, J. J., Carammia, M., & Bijak, J. (2021). Toward an early warning system for monitoring asylum-related migration flows in europe. *International Migration Review*, 01979183211035736.

Palotti, J., Adler, N., Morales-Guzman, A., Villaveces, J., Sekara, V., Garcia Herranz, M., … Weber, I. (2020). Monitoring of the venezuelan exodus through facebook's advertising platform. *Plos one*, *15*(2), e0229175.

Rossi, B. (2021). Forecasting in the presence of instabilities: How we know whether models predict well and how to improve them. *Journal of Economic Literature*, *59*(4), 1135–90.

Sardoschau, S. (2020). *The future of migration to germany: Assessing methods in migration forecasting.* DeZIM-Institut.

Shellman, S. M., & Stewart, B. M. (2007). Predicting risk factors associated with forced migration: An early warning model of haitian flight. *Civil Wars*, *9*(2), 174–199.

Spyratos, S., Vespe, M., Natale, F., Weber, I., Zagheni, E., & Rango, M. (2018). Migration data using social media: a european perspective. *Publications Office of the European Union*.

Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, *97*(460), 1167–1179.

Suleimenova, D., Bell, D., & Groen, D. (2017). A generalized simulation development approach for predicting refugee destinations. *Scientific reports*, *7*(1), 1–13.

Wanner, P. (2021). How well can we estimate immigration trends using google data? *Quality & Quantity*, *55*(4), 1181–1202.

Zagheni, E., Garimella, V. R. K., Weber, I., & State, B. (2014). Inferring international and internal migration patterns from twitter data. In *Proceedings of the 23rd international conference on world wide web* (pp. 439–444).

Zagheni, E., & Weber, I. (2012). You are where you e-mail: using e-mail data to estimate international migration rates. In *Proceedings of the 4th annual acm web science conference* (pp. 348–351).

Zagheni, E., Weber, I., & Gummadi, K. (2017). Leveraging facebook's advertising platform to monitor stocks of migrants. *Population and Development Review*, 721–734.

Online Appendix

A Literature

A.1 Trends in Approaches

Traditionally, the most common approach to predicting migration flows has been to apply **time series models** (Disney et al., 2015). In the most basic specifications, time series models predict future migration based only on past migration, using autoregressive models. For example, Dustmann et al. (2003) apply this method to predict immigration to the United Kingdom and Germany following the EU enlargement in 2004. However, such simple models disregard factors other than past migration that determine refugee flows. Thus, they can only provide reliable predictions as long as these factors do not vary, such that migration flows remain stable over time. Indeed, the migration flows estimated by Dustmann et al. (2003) and other authors modeling migration after the EU enlargements in the 2000s (e.g., Alvarez-Plata et al. (2003); Fertig & Schmidt (2005)) turned out to be erroneous and lacking power. To remedy this issue at least partly, it is possible to include additional control variables, such as unemployment rates, into autoregressive distributed lag models (Bijak et al., 2019).

Gravity-models add an important layer of complexity. They allow for the incirporation of factors from both the sending and the receiving countries as well as the bilateral relationship, thus going far beyond time series models, which tend to rely entirely on variations within one country. Gravity models use differences between countries and variation over time in these differences to predict future migration. Due to past data limitations, the popularity of gravity models has been on the rise only recently. For example, Hanson & McIntosh (2016) estimate the impact of geographical distance, the presence of shared borders and languages, former colonial ties, and GDP differentials on future immigration to several European countries. However, their results are only partly plausible because the model relies on average relationships and is not tailored to a specific origin or destination country. For Germany, Hanson's and McIntosh's model predicts that the stock of people born in foreign countries will be close to zero in 2020 and turn negative by 2030. Despite such examples of gravity models giving unrealistic results due to the inability to sufficiently tailor them to each country, the advantages of gravity models compared to time series models should not be underestimated. In addition to the characteristics mentioned above, they also allow modeling future migration flows to one country based on the experiences of other countries. Hence, scenario modeling becomes possible even for cases with a lack of data availability.

A third approach to predict medium-term migration flows is **theory-based structural modeling**. These models are primarily used to predict longer-term migration, as the factors they consider gain relevance only in the long run. Dao et al. (2018) develop a structural migration model in which migration depends on income differentials between countries. In turn, these income differentials are spurred by differences in education. Burzynski et al. (2020) refine this model such that it considers differences in educational and labor costs, as well as countries' consumption levels. However, while these theory-based studies have the potential to model relevant drivers of migration, caution is also warranted when interpreting these findings. For example, Burzynski et al. (2020) forecast that about a million immigrants from Mexico would be present in Germany by 2020, which, as Sardoschau (2020) points out, is a considerable overestimation. Again, this points out the conflict between a model's aims (here: understanding long-run forces) and short- or medium-run predictions.

A.2 Trends in Data Sources

Migration predictions rely predominantly on official data (e.g. immigrant registries) and if explanatory variables are used, they typically come from national statistics or cross-national databases such as those of the World Bank.

In recent years, a new source of data has become available and is increasingly used in migration forecasting: Digital trace data, i.e., the traces individuals leave behind from online behavior. Examples include internet searches, locations tied to services such as email accounts, and social media data (see Table A.2).

The main drawback of such data is that they are not representative of the overall population as there exists self-selection of users into these specialized services. The generalization of the forecasts depends on who uses a service, e.g., in most countries, Instagram covers a far lower share of the population above the age of 50 than Facebook, which in turn is less representative of the population than the users of Google search. Google search, with a global market share of about 92 percent of all internet user who search online globally (the outlier being China with only 3 percent, which strongly decreases the global average), has a particularly broad user base. Yet, not everyone has access to the internet in the first place, with intersectional difference or disadvantage along lines of race, gender, ethnicity and class, among others. Furthermore, as Böhme et al. (2020) summarize, searching for something online implies some interest in a topic but not necessarily that it will lead to behavior. Our literature survey shows that the majority of studies using “big data”-based approaches to predict migration flows use Google Trends and Facebook Ads Manager (see Table A.2). These studies typically do not forecast future migration, except in the very short term (“nowcasting”). Mostly, they develop models and test if they can predict actual migration flows (i.e. the present).

As summarized by Böhme et al. (2020), a few applications use internet metadata to approximate migration dynamics and patterns. For example, Zagheni et al. (2014) predict migration flows with the help of geo-referenced Twitter data while Zagheni & Weber (2012) perform the task based on IP addresses.

Data	Study	Sample	Methods and findings
Bing Maps: to obtain locations of major settlements and routing information between the camps, conflict zones and other settlements; combined with UNHCR and conflict data	Suleimenova et al. (2017)	Refugees from Burundi, Central African Republic and Mali to neighboring several countries	Agent-based modelling, using a generalized simulation development approach. Simulations predict more than 75% of the refugee destinations correctly after the first 112 days. The simulations are validated using actual data
Facebook: Ads Manager, through which advertisers can select subgroups (e.g., Puerto Ricans living in California) and get an estimate of the "potential reach" (monthly active users)	Zagheni et al. (2017)	Expats from several countries in the US	Proof of concept that people classifying themselves as "expats" on Facebook are quite representative of actual immigrant stocks in the US. Also evaluate biases in the data and show how accounting for these biases leads to better estimates and predictions; discuss the challenges and opportunities ahead in this area of research.
Facebook: Ads Manager (see above)	Spyratos et al. (2018)	Expats in 17 EU countries	Based on Zagheni et al. (2017) (see above), but accounting for bias in Facebook data. Estimated the number of expatriates in 17 EU countries based on the number of Facebook Network users who are classified by Facebook as "expats". Methodology allowed for the timely capture of the increase of Venezuelan migrants in Spain (though some other estimates were implausible).
Facebook: Ads Manager (see above)	Alexander et al. (2019)	Refugees from Puerto Rico to the US	Diff-in-Diff to obtain estimates of the percent change in migrants. Estimated that there was a 17 percent increase in the number of Puerto Rican migrants present in continental US over the period from October 2017 to January 2018 (in line with previous studies)

Data	Study	Sample	Methods and findings
Facebook: Ads Manager (see above); combined with survey data	Alexander et al. (2020)	Migrants from Mexico, India and Germany to the US	Combination of Bayesian Hierarchical Model, time series, spatial and data models. Produce timely ‘nowcasts’ of migrant stocks. Compare projections with data from American Community Survey, show that the model outperforms alternatives that rely solely on either social media or survey data.
Facebook: Ads Manager (see above)	Palotti et al. (2020)	Refugees and migrants from Venezuela to other Latin American countries	Estimated monthly active users from Ads Manager vs. estimates from Response for Venezuela, a high correlation between the two (r = 0.99)
Google Trends Index (GTI): time series of search intensities of the user’s choice of keywords, by (origin) country. The GTI can be restricted by geographical area, date, a set of predefined general search categories.	Böhme et al. (2020)	Search behavior of an estimated 842 million speakers from 107 countries of origin in which at least one of the three selected languages (English, French, Spanish) is officially spoken; 34 OECD destination countries	Filter Google data for keywords related to “immigration” and economics. Estimate a range of fixed effects panel models (with a range of controls) using two different specific actions: a unilateral and a bilateral (gravity) model. Approach provides strong additional predictive power for international migration flows when compared to reference models. Evidence, based on survey data, that measures partly reflect genuine migration intentions.
Google Trends Index (GTI); see above	Golenvaux et al. (2020)	Data from Böhme et al. (2020), see above	A long short-term memory (LSTM) approach (=machine learning), combined with Google Trends data. It outperforms two existing approaches in predicting the one-year ahead incoming international migration: the linear gravity model (from Böhme et al. (2020) and an artificial neural network model.

Data	**Study**	**Sample**	**Methods and findings**
Google Trends Index (GTI), see above	Wanner (2021)	Labor immigrants to Switzerland from Spain, Italy, France and Germany	Linear regression was used in order to measure the relationship between the number of searches (x) and the number of moves (y) that finally occurred. The results show that Google Trends can predict, to some extent, current and future (short-term) migration flows of adults arriving from Spain or Italy. However, the predictions appear not to be satisfactory for other flows (from France and Germany).
Google Trends Index (GTI), see above	Adema & Guha (2022)	Refugee flows from Ukraine to the European Union and their spatial distribution in EU host countries	Linear regression was used in order to measure the correlation between the predicted number of refugees using Google Trends and residence permits (x) and the actual number of registered refugees (y) recorded in the destination countries. The results show a strong positive association between predicted and actual values and a high coefficients of determination.
Twitter: Streaming API, which provides a 1% random sample of all geo-localized tweets at any given moment	Hausmann et al. (2018)	Emigrant from Venezuela	Define as migrants those people who tweeted in Venezuela in spring 2017 and from another country in spring 2018. Estimate that up to 2,9 million Venezuelans have left the country in the past year.
Twitter: Streaming API (see above)	Zagheni et al. (2014)	Emigrants from OECD countries	Diff-in-diff approach to reduce selection bias when inferring trends in out-migration rates for single countries. Methods can be used to predict turning points in migration trends, which are particularly relevant for migration forecasting. Geolocated Twitter data can substantially improve understanding of the relationships between internal and international migration.

Data	**Study**	**Sample**	**Methods and findings**
Yahoo: Self-reported age and gender of anonymized e-mail users were linked to the geographic locations (mapped from IP addresses) from where users sent e-mail messages over time (2009-2011).	Zagheni & Weber (2012)	Emigrants from 11 European countries	Estimates of age profiles of migration are qualitatively consistent with existing administrative data sources (Eurostat migration rates). Document the recent increase in human mobility and observe that female mobility has been increasing at a faster pace. Findings suggest that e-mail data may complement existing migration data.

B Tables

Table 2: List of Origin Countries by Language

English			French	Spanish	Portuguese	Arabic	Turkish	Fulah	Hausa	Persian	Papasena
Albania	Japan	Serbia	Algeria	Argentina	Angola	Algeria	Armenia	Cameroon	Niger	Afghanistan	Afghanistan
Armenia	Kazakhstan	Sierra Leone	Andorra	Aruba	Brazil	Bahrain	Azerbaijan	Gambia	Nigeria	Bahrain	Pakistan
Australia	Kenya	Singapore	Benin	Belize	Cape Verde	Chad	Georgia	Mali		Iran	
Bahamas	Kiribati	Solomon Islands	Burkina Faso	Bolivia	Equatorial Guinea	Comoros	Iran	Mauritania		Qatar	
Bangladesh	Kyrgyzstan	Somalia	Burundi	Brazil	Guinea-Bissau	Djibouti	North Macedonia	Niger		Tajikistan	
Barbados	Laos	South Africa	Cameroon	Chile	Mozambique	Egypt	Turkey	Senegal		Uzbekistan	
Belarus	Lebanon	South Korea	Canada	Colombia	Paraguay	Eritrea	Turkmenistan				
Belize	Lesotho	South Sudan	Central African Republic	Costa Rica	São Tomé & Príncipe	Iraq					
Bhutan	Liberia	Sri Lanka	Chad	Cuba	Timor-Leste	Israel					
Bosnia & Herzegovina	Malawi	St. Lucia	Comoros	Dominican Republic		Jordan					
Botswana	Malaysia	St. Vincent & Grenadines	Congo - Brazzaville	Ecuador		Kuwait					
Brunei	Maldives	Suriname	Congo - Kinshasa	El Salvador		Lebanon					
Cambodia	Mauritius	Switzerland	Côte d'Ivoire	Equatorial Guinea		Libya					
Cameroon	Mexico	Taiwan	Djibouti	Guatemala		Mauritania					
Canada	Micronesia	Tanzania	Equatorial Guinea	Honduras		Morocco					
China	Moldova	Thailand	Gabon	Mexico		Niger					
Equatorial Guinea	Mongolia	Tonga	Haiti	Morocco		Oman					
Eswatini	Montenegro	Trinidad & Tobago	Lebanon	Nicaragua		Palestinian Territories					
Ethiopia	Myanmar	Turkey	Madagascar	Panama		Qatar					
Fiji	Nepal	Uganda	Mali	Paraguay		Saudi Arabia					
Gambia	New Zealand	Ukraine	Mauritius	Peru		South Sudan					
Ghana	Nigeria	United States	Morocco	United States		Sudan					
Grenada	Norway	Vanuatu	Niger	Uruguay		Syria					
Guyana	Pakistan	Vietnam	Rwanda	Venezuela		Tanzania					
Iceland	Papua New Guinea	Zambia	Senegal			Tunisia					
India	Philippines	Zimbabwe	Seychelles			United Arab Emirates					
Indonesia	Russia		Switzerland			Yemen					
Israel	Rwanda		Togo								
Jamaica	Samoa		Tunisia								

[a] For each country, the most spoken language is chosen.

Table 3: Descriptive Statistics of Google Trends Data

Search Term	Mean(U)	Std.Dev.(U)	Mean(B)	Std.Dev.(B)	Share of Non-zeros (U)	Share of Non-zeros (B)
advisers+advisors	4.62	13.07	0.02	0.91	0.23	0.00
agent	13.88	19.95	0.09	1.78	0.58	0.01
aliens	6.38	13.63	0.02	0.86	0.38	0.00
applicant+applicants+application+apply	19.08	22.72	0.37	3.67	0.70	0.02
appointment	10.99	18.34	0.11	2.07	0.49	0.01
arrival+arrivals	12.62	18.61	0.09	1.88	0.53	0.00
assimilate+assimilation	4.42	12.05	0.01	0.59	0.23	0.00
asylum	5.40	12.85	0.05	1.37	0.31	0.00
asylum seeker	0.70	5.16	0.00	0.48	0.04	0.00
austerity	2.83	9.95	0.00	0.32	0.16	0.00
bailout	2.78	9.42	0.01	0.53	0.22	0.00
benefit+benefits	22.60	23.77	0.17	2.47	0.71	0.01
bilateral	7.21	15.37	0.01	0.40	0.37	0.00
biometric	3.94	11.77	0.01	0.77	0.23	0.00
births	5.98	14.70	0.02	0.77	0.29	0.00
border controls+border control	0.71	4.88	0.01	0.67	0.04	0.00
bureau of immigration	0.47	3.88	0.00	0.33	0.04	0.00
business+businesses	19.08	22.30	0.34	3.48	0.68	0.02
card	20.79	22.23	0.37	3.75	0.79	0.02
certificate	15.50	21.59	0.13	2.16	0.63	0.01
check	20.15	22.61	0.23	2.99	0.70	0.01
checkpoint+checkpoints	3.68	10.86	0.01	0.49	0.23	0.00
citizen	11.48	18.36	0.18	2.66	0.54	0.01
citizenship+citizenships	7.21	14.83	0.19	2.67	0.40	0.01
compensation+compensations	7.97	16.23	0.05	1.42	0.40	0.00
competitiveness	3.43	10.79	0.00	0.35	0.25	0.00
consulate+consulates	7.09	14.77	0.23	2.86	0.39	0.02
contract+contracts	16.82	20.79	0.11	2.05	0.67	0.01
cooperation	6.81	13.88	0.02	0.93	0.47	0.00
crises+crisis	9.89	15.79	0.14	2.20	0.57	0.01
curtail	5.49	15.02	0.00	0.44	0.21	0.00
customs	9.91	16.12	0.07	1.60	0.54	0.00
cyclical	5.80	14.12	0.01	0.67	0.28	0.00
decentralization+decentralisation	2.12	8.06	0.00	0.33	0.19	0.00
decreased	5.32	13.62	0.01	0.79	0.24	0.00
deficits	1.59	7.82	0.00	0.31	0.09	0.00
democratization+democratisation	1.04	5.92	0.00	0.17	0.09	0.00
demographic+demography	4.78	11.92	0.03	1.01	0.34	0.00
department	13.54	19.23	0.15	2.30	0.67	0.01
deportation+deportations+deported	3.01	9.93	0.02	0.91	0.17	0.00
deregulation	1.09	6.26	0.00	0.24	0.09	0.00
detain+detained+detention	5.85	13.80	0.02	0.83	0.29	0.00
determinants	5.20	13.05	0.01	0.48	0.31	0.00
devaluation	2.18	8.69	0.00	0.27	0.16	0.00
diaspora	3.91	11.11	0.01	0.49	0.27	0.00
discriminate+discriminatory	2.41	9.40	0.00	0.40	0.12	0.00
disparities	2.15	9.00	0.00	0.40	0.11	0.00

Table 3: Descriptive Statistics of Google Trends Data *(continued)*

Search Term	Mean(U)	Std.Dev.(U)	Mean(B)	Std.Dev.(B)	Share of Non-zeros (U)	Share of Non-zeros (B)
diversification	3.55	11.30	0.00	0.44	0.22	0.00
diversity	7.70	15.76	0.04	1.25	0.44	0.00
documents	12.40	19.60	0.10	1.81	0.58	0.01
downturn	3.86	11.91	0.00	0.40	0.19	0.00
dual citizenship	1.82	7.96	0.06	1.40	0.11	0.00
dual nationality	0.72	4.61	0.00	0.15	0.06	0.00
earning+earnings	8.33	16.01	0.04	1.27	0.40	0.00
economically	4.10	11.48	0.02	0.80	0.28	0.00
economist+economists	5.42	12.98	0.03	0.99	0.33	0.00
economy+economies	12.62	17.91	0.26	3.05	0.67	0.02
elites	3.71	11.35	0.01	0.63	0.21	0.00
embassy+embassies	14.44	18.95	0.59	4.55	0.72	0.05
emigrant+emigrants	2.88	9.55	0.01	0.68	0.18	0.00
emigrate+emigrated	3.88	10.71	0.03	1.03	0.23	0.00
emigration	4.19	11.60	0.05	1.44	0.30	0.00
employer+employers	8.38	16.62	0.05	1.46	0.40	0.00
employment	13.26	18.92	0.12	2.06	0.65	0.01
empowerment	5.76	14.17	0.00	0.30	0.32	0.00
enforcement+enforces	6.31	15.17	0.03	1.05	0.30	0.00
exclusion	4.42	12.32	0.01	0.69	0.26	0.00
exports	5.08	12.76	0.07	1.51	0.36	0.00
extension	12.46	19.01	0.07	1.54	0.57	0.00
foreigner+foreigners	9.87	16.43	0.14	2.22	0.51	0.01
form	22.82	23.94	0.25	3.16	0.74	0.02
GDP	8.40	15.88	0.21	2.72	0.46	0.01
geopolitical	2.36	9.03	0.00	0.41	0.16	0.00
globalisation+globalization	5.52	12.89	0.03	0.99	0.41	0.00
growth	14.13	19.19	0.11	2.01	0.65	0.01
H.R.+HR	13.84	21.51	0.06	1.45	0.55	0.00
hardship+hardships	4.38	12.36	0.02	0.93	0.25	0.00
hiring	9.06	17.54	0.07	1.80	0.41	0.00
homeland	9.04	16.75	0.03	1.06	0.44	0.00
ignoring	5.53	15.00	0.00	0.45	0.22	0.00
illegal+illegally	6.86	14.83	0.09	1.86	0.38	0.00
immigrant+immigrants	6.17	13.35	0.11	1.98	0.37	0.01
immigrate+immigrated	3.33	9.93	0.06	1.47	0.23	0.00
immigration	9.96	16.25	0.32	3.36	0.62	0.03
incentives	4.10	11.89	0.02	0.76	0.25	0.00
income+incomes	14.11	19.82	0.20	2.74	0.60	0.01
indentured	3.05	10.40	0.01	0.61	0.13	0.00
indicators	6.93	14.16	0.01	0.70	0.46	0.00
individualism	2.82	10.26	0.00	0.42	0.19	0.00
industrialisation+industrialization	2.50	9.14	0.03	1.01	0.19	0.00
industrialised+industrialized	0.99	5.74	0.01	0.64	0.07	0.00
inefficiency	1.27	6.78	0.00	0.20	0.09	0.00
inequalities+inequality	5.56	13.07	0.03	1.04	0.32	0.00
inflation	8.39	15.83	0.06	1.44	0.49	0.00

Table 3: Descriptive Statistics of Google Trends Data *(continued)*

Search Term	Mean(U)	Std.Dev.(U)	Mean(B)	Std.Dev.(B)	Share of Non-zeros (U)	Share of Non-zeros (B)
influx	4.21	12.64	0.00	0.46	0.19	0.00
instability	4.27	11.93	0.01	0.52	0.25	0.00
insurance	16.10	22.20	0.24	2.94	0.64	0.01
intermarriage	0.40	3.71	0.00	0.17	0.03	0.00
internship+internships	9.91	16.87	0.07	1.61	0.47	0.00
interview	15.53	19.94	0.10	1.88	0.61	0.01
job+jobs	27.88	25.32	0.76	5.11	0.81	0.05
labor+labour+laborers+labourers	14.84	19.13	0.12	2.11	0.66	0.01
layoff+layoffs	4.50	12.70	0.01	0.44	0.20	0.00
legalization+legalisation+legalisations+legalizations	2.57	8.57	0.01	0.61	0.18	0.00
liberalization+liberalisation	2.05	7.72	0.00	0.13	0.17	0.00
lottery	12.15	19.87	0.12	2.17	0.54	0.01
macro+macroeconomic	11.61	19.37	0.03	0.98	0.50	0.00
marriage	18.31	20.50	0.13	2.08	0.71	0.01
migrant+migrants	4.33	11.25	0.04	1.26	0.30	0.00
migrate	5.48	13.81	0.07	1.63	0.30	0.00
migration	9.20	16.28	0.12	2.12	0.55	0.01
minimum	11.28	18.74	0.21	2.84	0.49	0.01
mismanagement	0.72	5.11	0.00	0.00	0.05	0.00
monetary	7.01	14.24	0.03	0.97	0.46	0.00
monopolies	1.73	8.33	0.00	0.35	0.10	0.00
multicultural+multiculturalism	2.37	8.67	0.01	0.58	0.16	0.00
nationality+nationalities	10.95	17.70	0.13	2.12	0.52	0.01
nationalization+nationalisation	1.63	7.25	0.00	0.19	0.12	0.00
naturalization+naturalisation+naturalisations+naturalizations	1.93	8.09	0.01	0.59	0.14	0.00
news	27.53	22.11	0.78	5.17	0.85	0.06
passport+passports	15.34	21.65	0.31	3.30	0.56	0.02
payroll+payrolls	9.72	17.60	0.04	1.27	0.43	0.00
pension+pensions	9.14	17.93	0.11	2.01	0.40	0.01
permit	10.78	18.50	0.14	2.36	0.50	0.01
pogroms	1.33	7.59	0.00	0.15	0.06	0.00
policies	8.82	15.77	0.07	1.67	0.52	0.00
policymakers	1.65	7.51	0.01	0.62	0.10	0.00
political asylum	1.08	6.17	0.01	0.68	0.07	0.00
political refugee	0.26	3.36	0.00	0.20	0.01	0.00
populate	3.83	11.75	0.00	0.11	0.21	0.00
privatization+privatisation	2.60	9.25	0.00	0.37	0.22	0.00
productivity	5.99	13.74	0.01	0.65	0.36	0.00
prosperity	6.05	14.39	0.02	0.86	0.29	0.00
quarantine	2.54	10.60	0.08	2.25	0.22	0.00
quota+quotas	8.18	16.29	0.02	0.94	0.39	0.00
recession+recessions	2.80	9.13	0.02	0.88	0.22	0.00
recruitment+recruitments	11.56	17.63	0.06	1.48	0.55	0.00
reforms	4.45	12.26	0.04	1.30	0.27	0.00
refugee+refugees	4.12	10.66	0.05	1.25	0.32	0.00
remuneration+remunerations	6.17	14.45	0.03	1.09	0.32	0.00
renewal	10.94	19.11	0.08	1.77	0.45	0.00

Table 3: Descriptive Statistics of Google Trends Data *(continued)*

Search Term	Mean(U)	Std.Dev.(U)	Mean(B)	Std.Dev.(B)	Share of Non-zeros (U)	Share of Non-zeros (B)
repatriation	2.36	9.67	0.00	0.43	0.13	0.00
required documents+required document	3.01	11.03	0.03	1.02	0.13	0.00
requirements	16.00	23.03	0.29	3.20	0.54	0.02
resettlement	1.68	7.53	0.00	0.34	0.12	0.00
restrict+restricting	4.21	11.55	0.01	0.71	0.26	0.00
restriction	6.08	14.23	0.03	1.05	0.34	0.00
restrictive	2.34	9.56	0.00	0.29	0.13	0.00
reunification	1.76	7.83	0.01	0.71	0.11	0.00
revitalization+revitalisation	1.44	7.08	0.00	0.31	0.09	0.00
salary+salaries	18.77	21.93	0.45	4.19	0.65	0.02
sanctions	5.60	13.38	0.03	1.01	0.30	0.00
Schengen	6.23	14.27	0.09	1.79	0.35	0.01
sectors	6.04	13.99	0.02	0.98	0.34	0.00
seekers	3.89	11.87	0.01	0.50	0.26	0.00
slump	8.06	15.65	0.02	0.77	0.34	0.00
smuggler+smugglers+smuggling	5.16	12.71	0.01	0.68	0.27	0.00
social security	6.64	14.25	0.05	1.29	0.41	0.00
sponsor	6.16	14.59	0.04	1.30	0.32	0.00
spouses	4.85	12.25	0.01	0.40	0.26	0.00
stabilisation+stabilization	4.88	12.12	0.00	0.36	0.28	0.00
stagnation	2.41	9.24	0.00	0.31	0.13	0.00
stateless	1.54	7.63	0.00	0.33	0.12	0.00
status	17.73	21.55	0.25	2.97	0.66	0.02
stimulus	5.24	13.68	0.01	0.87	0.28	0.00
student visa	4.01	11.95	0.11	2.02	0.22	0.01
sufficiency	3.25	10.78	0.00	0.00	0.18	0.00
tariffs	5.83	13.68	0.01	0.64	0.35	0.00
tax+taxes	14.35	19.58	0.38	3.69	0.62	0.02
test	24.98	23.03	0.29	3.26	0.81	0.02
tightened+tightening	5.42	13.62	0.00	0.41	0.26	0.00
tourist+tourists	8.50	15.37	0.21	2.63	0.48	0.02
trafficked+trafficking	7.26	15.82	0.04	1.31	0.34	0.00
unauthorised+unauthorized	2.76	9.15	0.00	0.35	0.17	0.00
underdeveloped	1.82	8.09	0.00	0.39	0.12	0.00
undocumented	1.06	6.73	0.01	0.56	0.06	0.00
unemployment	7.10	14.01	0.10	1.90	0.48	0.01
union+unions	14.41	20.45	0.23	2.90	0.64	0.01
unskilled	1.46	7.87	0.00	0.32	0.07	0.00
unsustainable	1.45	7.94	0.00	0.00	0.06	0.00
vacancy+vacancies	11.91	18.84	0.08	1.70	0.53	0.01
viability	4.64	12.40	0.00	0.31	0.27	0.00
visa free	6.30	14.25	0.04	1.24	0.33	0.00
visa+visas	17.95	22.50	0.99	5.93	0.65	0.06
wage+wages	10.95	17.41	0.24	3.02	0.50	0.01
waiver+waivers	5.09	13.06	0.02	0.90	0.25	0.00
welfare	6.93	15.35	0.04	1.19	0.35	0.00
wellbeing	6.70	14.97	0.01	0.71	0.32	0.00

Table 3: Descriptive Statistics of Google Trends Data *(continued)*

Search Term	Mean(U)	Std.Dev.(U)	Mean(B)	Std.Dev.(B)	Share of Non-zeros (U)	Share of Non-zeros (B)
woes	1.62	7.85	0.00	0.18	0.10	0.00
work visa	4.67	12.70	0.16	2.45	0.25	0.01
worker	9.64	16.05	0.07	1.57	0.50	0.00
worsening	1.48	8.11	0.00	0.11	0.05	0.00
DESTINATION			7.23	14.27		0.46

Note: Search terms 1 to 192 are migration-related single keywords and keyword combinations. “U” and “B” in brackets indicate “unilateral” and “bilateral”., respectively. Unilateral indices are the relative search intensities for each migration-related search term, the descriptive statistics are calculated over all origins and all months for each term. The bilateral index of each term is the search intensity for the query combining the term and each destination, the descriptive statistics are taken over all origin-destination binaries and all months. Search term 193 is the destination country name. The search intensities of it reflect migration intentions at the bilateral level, therefore its unilateral descriptive statistic is not available. [a] U: unilateral, B: bilateral, S.D.: standard deviation. [b] Search term ’DESTINATION’ represents the 27 EU destinations country names used as search terms. [c] Google Trends data ranges from 0 to 100.

Table 4: List of EU Destination Countries

Country	
Austria	Ireland
Belgium	Italy
Bulgaria	Lithuania
Cyprus	Luxembourg
Czechia	Latvia
Germany	Malta
Denmark	Netherlands
Estonia	Poland
Spain	Portugal
Finland	Romania
France	Sweden
Greece	Slovenia
Croatia	Slovakia
Hungary	

Table 5: Descriptive statistics for selected classical predictor variables, monthly

Variable	Count	Mean	Std.Dev.	Min	Max
Unemployment rate, %	2941	6.19	2.31	2.10	19.46
Working-age population, million	2941	53.48	69.67	0.88	261.34
No. of Natural Disaster Events (destination)	1477	0.10	0.34	0	3
No. of Natural Disaster Events	13094	0.20	0.53	0	9
No. of Technological and Complex Disaster Events (destination)	1477	0.03	0.18	0	3
No. of Technological and Complex Disaster Events	13094	0.06	0.28	0	5

Table 6: Data Collected and Databases

Type of data	Frequency	Number of Variables Collected	Data source	Data source description
Political violence events and fatalities	Fully-recorded	12	ACLED	Armed Conflict Location and Event Data Project
Disaster indicators	Fully-recorded	12	EMDAT	International Disaster Database
Leadership characteristics and election outcomes	Fully-recorded	19	REIGN	Rulers, Elections, and Irregular Governance Dataset
Election violence outcome indicators	Monthly	17	ELVI	Election Violence Events Dataset
Asylum and managed migration	Monthly	2	eurostat	Statistical Office of the European Union
Short-term business statistics	Monthly	22	eurostat	Statistical Office of the European Union
Agri-environmental indicators	Monthly	348	FAO	The Food and Agriculture Organization
Labor statistics	Monthly	49	ILO	International Labour Organization
Consumer prices	Monthly	4	ILO	International Labour Organization
Macroeconomic and financial indicators	Monthly	456	IMF	International Monetary Fund

C Appendix on Methodology

We now discuss some difficulties with our approach. First, the panel data set we use is unbalanced. That means substantial numbers of missing observations in the explanatory variables for some but not all bilateral relationships. Hence, the information set available for forecasts is not always the same for each bilateral corridor. We do not perform any interpolation to replace missing values as this may add unnecessary error to the data, especially in the case of dummy variables or count variables such as disaster and conflict data. Second, due to the moving-window forecasting approach, some explanatory variables in the training window may contain missing observations. The R routines we use to train the models typically do not allow for missing observations. In such cases, the explanatory variable must be discarded, even if it contains only a single missing observation. Third, the testing data we use to make forecasts may contain missing explanatory variables. These are neither forecast outside the model nor interpolated from the preceding data. The affected variables are also discarded from the training set in such cases. Fourth, the dependent variable for some corridors exhibits very low variation as some bilateral relationships are not used for international migration at all or to a minimal extent. That may lead to very low-quality regression results as there is no variation to be explained. In the baseline case, this issue is ignored. Fifth, the models we use are all taken "off-the-shelf", meaning that we do not fine-tune hyperparameters, which can be detrimental to forecasting performance. On the one hand, we do this for time reasons, as there are thousands of bilateral relationships. For each of these, we need to calculate over a hundred forecasts at different forecast horizons and with varying specifications. On the other hand, experimenting with different tuning strategies has yielded minimal performance gains at increasing time costs, which do not appear to warrant tuning at each forecasting point. Finally, the dependent variable in the forecast period may be missing. This lack of ground truth makes it impossible to compute errors. So, while the forecast can be recorded, the error cannot be used for forecast performance evaluation. That is a relatively minor issue due to the overall good coverage of the dependent variable.

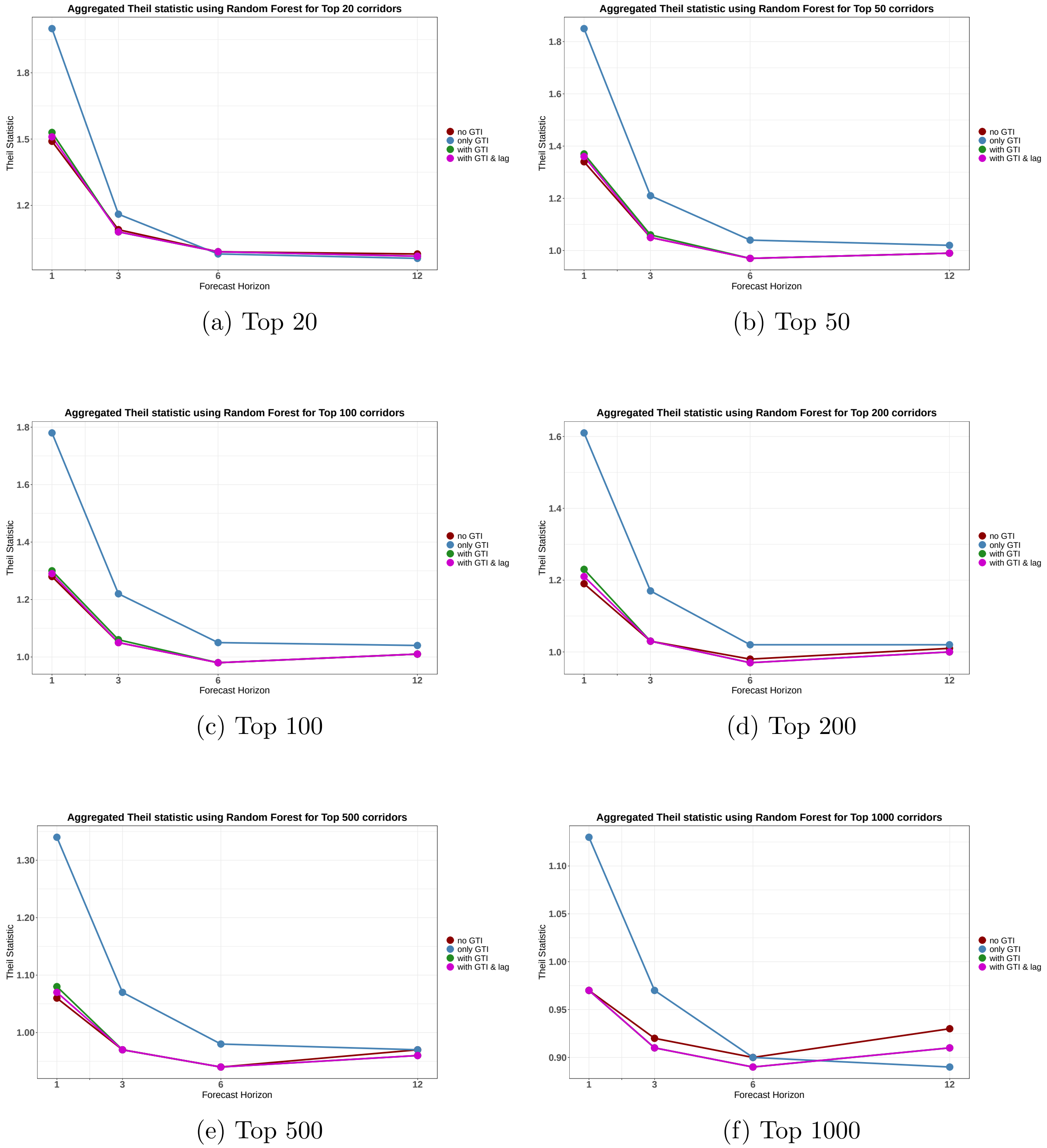
For the computation of Theil's U, it has to be noted that the Random Walk can make more forecasts than many of our models because, for certain training windows, the dependent variable may contain only a constant value, usually zero. The Random Walk prediction will then mechanically be zero again, whereas a proper regression has no variation to exploit and will not record any forecast. Hence, we restrict the error computation to those instances where all of the models and the Random Walk have made a proper forecast to not distort the evaluation sample in favor of models that forecast "simpler to forecast" periods.

D Additional Results

Table 7: Theil statistics for specification with socioeconomic regressors, GTI and lagged asylum seekers

Horizon	EN	RF	XG	PC	EN_RF	EN_XG	XG_RF	RF_PC	EN_PC	XG_PC	EN_RF_XG	EN_RF_PC	RF_XG_PC	EN_XG_PC	EN_RF_XG_PC
A. Top 20 Corridors															
1	1.62939E+20	1.51	1.27	2.26	8.14694E+19	8.14694E+19	1.3	1.72	8.14694E+19	1.55	5.43129E+19	5.43129E+19	1.46	5.43129E+19	4.07347E+19
3	2.23858E+13	1.08	1.03	4.54	1.11929E+13	1.11929E+13	1.02	2.67	1.11929E+13	2.62	7.46195E+12	7.46195E+12	2.04	7.46195E+12	5.59646E+12
6	3.4422E+68	0.99	0.98	10059.3	1.7211E+68	1.7211E+68	0.96	5029.98	1.7211E+68	5029.96	1.1474E+68	1.1474E+68	3353.56	1.1474E+68	8.60551E+67
12	1.97939E+13	0.97	0.97	67.44	9.89697E+12	9.89697E+12	0.95	33.94	9.89697E+12	33.91	6.59798E+12	6.59798E+12	22.79	6.59798E+12	4.94848E+12
B. Top 50 Corridors															
1	1.45861E+35	1.36	1.25	10.97	7.29303E+34	7.29303E+34	1.22	6.04	7.29303E+34	5.92	4.86202E+34	4.86202E+34	4.34	4.86202E+34	3.64652E+34
3	5.44149E+42	1.05	1.07	24331.38	2.72074E+42	2.72074E+42	1.03	12166.11	2.72074E+42	12166.09	1.81383E+42	1.81383E+42	8111.03	1.81383E+42	1.36037E+42
6	1.43425E+68	0.97	1	4051.01	7.17126E+67	7.17126E+67	0.95	2025.83	7.17126E+67	2025.81	4.78084E+67	4.78084E+67	1350.79	4.78084E+67	3.58563E+67
12	1.34981E+61	0.99	1.02	8113.14	6.74906E+60	6.74906E+60	0.98	4056.81	6.74906E+60	4056.8	4.49937E+60	4.49937E+60	2704.73	4.49937E+60	3.37453E+60
C. Top 100 Corridors															
1	7.36822E+34	1.29	1.23	6.57	3.68411E+34	3.68411E+34	1.19	3.81	3.68411E+34	3.73	2.45607E+34	2.45607E+34	2.85	2.45607E+34	1.84206E+34
3	2.80996E+118	1.05	1.08	12273.24	1.40498E+118	1.40498E+118	1.03	6137.03	1.40498E+118	6137.01	9.36655E+117	9.36655E+117	4091.64	9.36655E+117	7.02491E+117
6	7.32384E+67	0.98	1	2248.08	3.66192E+67	3.66192E+67	0.96	1124.35	3.66192E+67	1124.32	2.44128E+67	2.44128E+67	749.78	2.44128E+67	1.83096E+67
12	1.41949E+110	1.01	1.02	4309.79	7.09743E+109	7.09743E+109	0.99	2155.12	7.09743E+109	2155.1	4.73162E+109	4.73162E+109	1436.92	4.73162E+109	3.54872E+109
D. Top 200 Corridors															
1	3.07435E+128	1.21	1.21	4522330347	1.53718E+128	1.53718E+128	1.14	2261165174	1.53718E+128	2261165174	1.02478E+128	1.02478E+128	1507443450	1.02478E+128	7.68588E+127
3	8.40792E+142	1.03	1.09	9.9928E+128	4.20396E+142	4.20396E+142	1.02	4.9964E+128	4.2253E+142	4.9964E+128	2.80264E+142	2.81686E+142	3.33093E+128	2.81686E+142	2.11265E+142
6	6.39105E+122	0.97	1.01	3.14692E+33	3.19552E+122	3.19552E+122	0.95	1.57346E+33	3.19552E+122	1.57346E+33	2.13035E+122	2.13035E+122	1.04897E+33	2.13035E+122	1.59776E+122
12	5.58381E+126	1	1.03	1.09711E+66	7.72258E+147	7.72258E+147	0.99	5.48556E+65	7.72258E+147	5.48556E+65	5.14839E+147	5.14839E+147	3.65704E+65	5.14839E+147	3.86129E+147
E. Top 500 Corridors															
1	5.95739E+138	1.07	1.17	3.9276E+110	2.9787E+138	2.9787E+138	1.05	1.9638E+110	2.98471E+138	1.9638E+110	1.9858E+138	1.98981E+138	1.3092E+110	1.98981E+138	1.49236E+138
3	9.68767E+145	0.97	1.08	4.00921E+128	4.84384E+145	4.84384E+145	0.97	2.0046E+128	4.8842E+145	2.0046E+128	3.22922E+145	3.25613E+145	1.3364E+128	3.25613E+145	2.4421E+145
6	8.41865E+146	0.94	1.02	1.44132E+103	4.20932E+146	4.20932E+146	0.93	7.20662E+102	4.25401E+146	7.20662E+102	2.80622E+146	2.83601E+146	4.80441E+102	2.83601E+146	1.79454E+149
12	8.18465E+147	0.96	1.03	6.22826E+135	7.2556E+147	7.2556E+147	0.95	3.11413E+135	7.33446E+147	3.11413E+135	4.83706E+147	4.88964E+147	2.07609E+135	4.88964E+147	3.66723E+147
F. Top 1000 Corridors															
1	2.98773E+138	0.97	1.08	7.19194E+142	1.49387E+138	1.49387E+138	0.96	3.59597E+142	3.62918E+142	3.59597E+142	9.9591E+137	2.41945E+142	2.39731E+142	2.41945E+142	1.81459E+142
3	4.80906E+145	0.91	1	1.8834E+149	2.40453E+145	2.40453E+145	0.9	9.41701E+148	9.63709E+148	9.41701E+148	1.60302E+145	3.3624E+149	3.28637E+149	3.3624E+149	2.5218E+149
6	4.15934E+146	0.89	0.96	1.6166E+146	2.07967E+146	2.07967E+146	0.87	8.08302E+145	2.99344E+146	8.08302E+145	1.38645E+146	1.99563E+146	5.38868E+145	1.99563E+146	9.09242E+148
12	6.2881E+147	0.91	0.96	4.74611E+146	4.70987E+147	4.70987E+147	0.88	2.37305E+146	5.35906E+147	2.37305E+146	3.13991E+147	4.02052E+149	3.74491E+149	4.02052E+149	3.01539E+149

Note: Theil statistics are aggregated over all forecasts and all corridors in a given Top-X specification. For example, Top 20 refers to the twenty corridors in the sample with the largest number of asylum seekers during the sample period. A value below one indicates better performance than a random walk forecast for the same corridors over the same forecasting period. EN=Elastic Net; RF=Random Forest; XG=Extreme Gradient Boosting; PC=Principal Component. EN_RF=Ensemble model composed of Elastic Net and Random Fores.



(a) Top 20

(b) Top 50

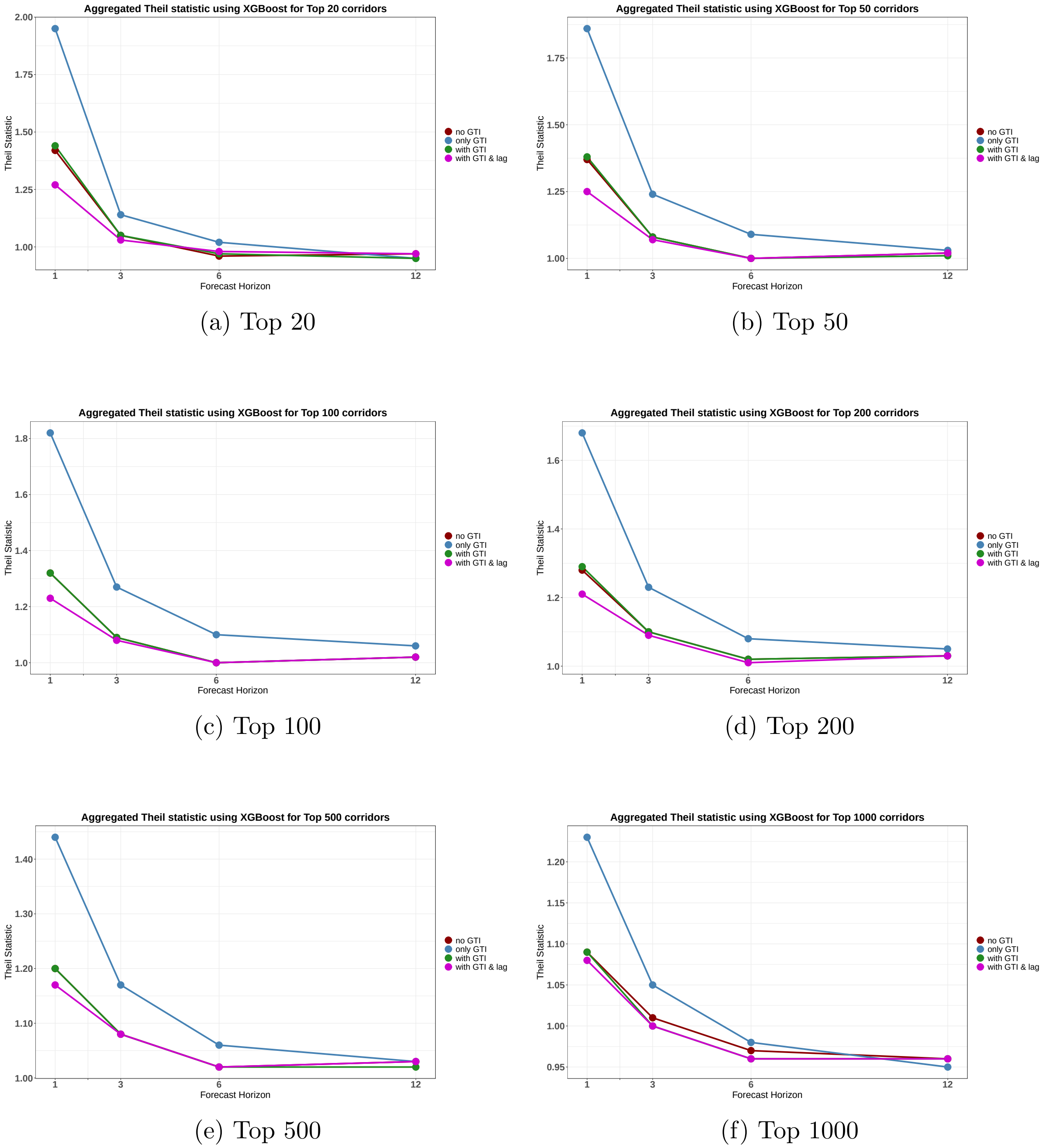
(c) Top 100

(d) Top 200

(e) Top 500

(f) Top 1000

Figure 3: Main forecasting results from the Random Forest model for six subsamples of bilateral corridors of different importance for refugee flows (top 20 – top 1000 corridors in terms of total asylum seeker numbers over the 2008 to 2020 period) over four horizons (1, 3, 6, and 12 months) and four specifications of the information set (no GTI, only GTI, with GTI, and with GTI and lagged dependent variable. The curves in each panel depict the Theil ratio of the RF model against the benchmark forecast based on the RW. Source: Author calculations.



(a) Top 20

(b) Top 50

(c) Top 100

(d) Top 200

(e) Top 500

(f) Top 1000

Figure 4: Main forecasting results from the XGBoost model for six subsamples of bilateral corridors of different importance for refugee flows (top 20 – top 1000 corridors in terms of total asylum seeker numbers over the 2008 to 2020 period) over four horizons (1, 3, 6, and 12 months) and four specifications of the information set (no GTI, only GTI, with GTI, and with GTI and lagged dependent variable. The curves in each panel depict the Theil ratio of the XGBoost model against the benchmark forecast based on the RW. Source: Author calculations.