Modeling Refugee Presence Across Europe: Insights from the Ukrainian Crisis and Random Forest Analysis

Daniel C. MacLeod

This project proposes a model to evaluate refugee migration flows from Ukraine to European Union countries, addressing the critical question of why refugees go where they go once displacement occurs. Building on established migration theory and advances in machine learning, the model integrates five key causal themes consistently observed in refugee research: accessibility, safety, familiarity, opportunity, and gravity. Drawing from Eurostat, Freedom House, and custom geographic datasets, the model captures monthly refugee flows, economic conditions, existing diaspora networks, political freedom scores, and proximity to Ukraine. To account for complex interactions, temporal variability, and nonlinear relationships across these factors, ensemble methods will be employed to pursue these effects. This approach combines the structural logic of gravity models with the flexibility of machine learning to produce a forecasting tool capable of supporting host nations in anticipatory policy decisions.

Table of contents

1 Introduction

Diaspora is a phenomenon of growing global importance, as rising tensions, international conflicts, and domestic crises drive refugees from their homes in increasing frequency. While the causes of forced migration have been widely studied, far less attention has been given to a question of arguably equal importance to host nations: why do refugees go where they go? Understanding the factors that shape refugee destination choices is critical for governments seeking to prepare for inflows, allocate resources effectively, and manage the resulting political and social strain.

The purpose of this research is not to reexamine the conditions that force people to flee their countries of origin, but rather to model where they go once they do. Understanding patterns in refugee movement empowers stakeholders to respond preemptively to incoming waves of refugees. Refugee flows have affected regions as disparate as the Middle East to Europe. This research will consider the various stems of crisis: domestic conflict, economic collapse, and international war. Better forecasting tools can help host nations prepare to both serve their own citizens while also moderating its support to displaced populations.

The research focuses on five key factors that shape where refugees choose to go: accessibility, safety, familiarity, opportunity, and gravity. Accessibility refers to how easily refugees can reach a country, including geographic distance, visa rules, and border controls. Safety involves the level of protection from violence and the strength of legal safeguards in the host country. Familiarity includes shared language, religion, or culture. Opportunity captures economic factors like job availability, quality of education, and state sponsored services. Gravity refers to the pull of existing diaspora communities or previous refugee flows, which have been found to greatly influence future movements. Each of these factors operates differently depending on the type and phase of the crisis, and this research is designed to account and weigh these differences.

The ultimate goal is to answer the research question: *"How do aspects of political freedom in Ukraine affect refugee migration to EU countries, given the aspects of those countries and the five main refugee drivers?"* By developing a generalizable, empirically grounded model that integrates these dimensions, this research seeks to provide governments with a pragmatic forecasting tool, one that can help anticipate refugee inflows and make informed policy decisions before the next crisis arrives.

2 Literature Review

Classic migration theory establishes the push-pull models of migration, where the detracting push factors compel departures and the attractive pull factors determine the destination choice. Recent research has also derived primary, secondary, and nascent migration as key stages in the process. Nascent migration precedes primary migration, characterized by middle to upper class citizens leaving for better opportunities abroad. Primary migration is the immediate flow of refugees with safety and accessibility foremost in consideration(Frith et al. 2019), corresponding with immediate state collapse or conflict. Secondary migration is marked by lower crisis intensity, or when refugees who initially fled to a nearby country during primary migration are now able to prioritize opportunity over safety and move onward to a second destination. The push-pull model and the three categories of migration refine the research to understand how means, timing, and urgency impact the balanced pursuit of safety and opportunity.

Recent refugee research converges on five key causal themes: accessibility, safety, familiarity, opportunity, and gravity. All pull factors will fall under one of these five categories. By analyzing these pillars, foundational assumptions begin to manifest. Distance decay is described as a refugee's preference for nearer destinations when facing initial displacement (Hierro and Maza

2024). The safety pull is marked by political stability, rule of law, low crime and terrorism, which are immediately enticing for refugees (Frith et al. 2019). A Syrian study fortified the well-established notion that security threats as measured by terrorism, crime and conflict are inversely related to refugee flow. Cultural proximity, marked by shared language, religion, and history fall under the theme of familiarity, and can be observed in the Syrian refugee crisis, where 2.7 million of 4.4 million registered refugees have landed in Turkey (UNHCR 2025).

Gravity models are increasingly popular and have been used to analyze refugee flow from Venezuela, Syria, and Ukraine. Gravity models (inspired by Newtonian gravity and are structured like a regression) successfully implement variables that include: existing diaspora, proportionality of population, physical distance, and social networks. Gravity models also control for cultural proximity (shared history, religion, etc.), opportunity, anti-immigration sentiment, and others (Hierro and Maza 2024). With sufficient data gravity models' lay a baseline which can then be refined by more control variables or by being integrated into ensemble methods (Lanati and Thiele 2024). Social network theory is closely interwoven with the theory behind gravity models, being that diaspora communities created linkages which facilitate yet more movement (Greene et al. 2023). Gravity is often boosted when augmented with social ties in a social network considerate gravity model.

This research paper will group social networks, social gravity, diaspora, and cumulative flows under the singular term "gravity". Within this context, diaspora and cumulative flows refer to existing migrant and refugee communities' influence (or pull) on new refugees. In several recent studies, gravity models appear especially strong. In the case of Ukrainian refugees, both prewar diaspora communities and newly accumulated refugee populations exert a measurable pull on subsequent flows. A 1% increase in prewar social networks leads to a 0.25% rise in monthly refugee inflows, while a 1% increase in accumulated refugee migration corresponds to a 0.36% rise.

In this case, while pre-existing communities shape early movement, the gravitational influence of new refugee networks grows over time and can surpass that of older diasporas. This statistically significant finding lends credence to gravity models' robust nature and why it is a practical starting point. Policy responsiveness can redirect asylum flows in measurable and often immediate ways (Guichard and Machado 2024).

The research has also developed to analyze how the world has reacted to certain crises. For example, when the Russo-Ukraine conflict began, the EU issued temporary protection orders to support Ukrainian refugees, and as the Syrian crisis begins to evolve, studies have also met the shift in policy with the analysis required to derive a line of effects. For example, studies following Germany's efforts to take in Syrian refugees have found that reductions in processing time significantly increased asylum applications. One simulation showed that Germany's drop in average processing time, from 15.7 to 9.4 months, accounted for 13.5% of the rise in applications lodged there, with a corresponding 7.9% drop in applications to other European countries (Bertoli, Brücker, and Fernández-Huertas Moraga 2022).

The methodology used to analyze refugee crises has gone through its own evolution over the last three decades. Originally constrained by comparatively small datasets, researchers primarily employed cross sectional, time series analysis (Schmeidl 1997), panel regression analysis and OLS and logistic regression. Studies like (Neumayer 2005) exemplified this early approach, linking asylum applications to political oppression and economic conditions through country-year regressions. These methods were pioneering in the 1990s and 2000s, but they had limitations. Early studies were limited by high dimensionality and under reported data. As detailed UNHCR data became available and computing power grew researchers were enabled by more sophisticated models to gather higher dimensions of these relationships.

The aforementioned gravity models marked a turning point in refugee research, particularly when coupled with models like the Poisson Pseudo Maximum Likelihood (PPML). These models were previously used to model trade flow, but their ability to model bilateral flows proved useful for this sector of research. PPML offers the ability to account for origin and destination fixed effects and has been used to model determinants of asylum applications to the EU as recently as 2023 (Di Iasio and Wahba 2024). Gravity modeling stands in contrast to prior research, in that it abides by theory of mass and distance, while also being able to deal well with overdispersion, heteroskedasticity, and keeps zeros in the data by assigning count values to countries which receive no refugees. These are all key elements that build a robust starting point capable of reducing bias, preserving the integrity of the dataset, and yielding more accurate and theoretically grounded estimates of refugee flows.

Other studies have acknowledged the complexity of multistage movement and the diaspora characterizations of primary versus secondary migration. For example, researchers split analysis of Ukrainian migrants into sub-periods (initial versus sustained) and found that determinants shifted over time. Many studies neglect temporal shifts and rather focus on the immediate time of the research.

An interesting approach is marked by a Swedish study on Syrian refugees that assessed the direct impact of policy change on asylum flows. Using a quasi-experimental interrupted time series design with multiple control groups, the researchers examined how Sweden's 2013 decision to grant permanent residence to Syrians affected application volumes (Andersson and Jutvik 2022). By combining high-frequency national data with UNHCR figures and comparing flows from other origin countries and Germany, the study isolated the policy's effect from broader conflict trends. It stands as a rare causal inference effort in refugee research and shows that even in a field somewhat saturated with gravity models, novel approaches still push the boundaries of

how we understand displacement dynamics. Another creative methodological approach includes Bayesian Hierarchical Clustering (Cottier 2024) and Agent Based Modeling (ABM) which can project probability distributions in the face of unseen data to make predictions and enable scenario analysis respectively. The ABM (when using the FLEE simulation framework) is a particularly novel approach allowing simulations of open and closed camps and borders which was able to match 75% of destinations in Africa over a twelve day simulation period (Suleimenova, Bell, and Groen 2017). These are notable evolutionary offshoots of research which branch away from strict equation based modeling and pivot to computational simulations.

The last five years have been marked by rapid developments in machine learning, which have increasingly been applied to refugee study. Complex nonlinear multidimensional data can now be handled by models such as Random Forests, gradient boosting, and neural networks, allowing researchers to detect interactions and relationships that simpler models might miss (Micevska 2021). When paired with traditional approaches, these techniques offer a powerful balance between prediction and explanation. This integration is exemplified in ensemble methods and in tools like the 2025 World Bank AI-powered refugee forecasting model (World Bank 2025). Ensemble methods refer to modeling approaches that combine multiple algorithms, each optimized for a specific aspect or stage of the problem. Rather than relying on a single model to handle all tasks, ensemble systems assign different models to different components of the pipeline, leveraging their respective strengths and synergizing their outputs to produce more robust, accurate, and generalizable results (Frith et al. 2019).

The greatest advance across these methodologies is the fact that the data and computational advances are only improving. ACLED has vastly expanded and refined conflict data while EUROSTAT now provides monthly asylum statistics allowing research to geo-reference and gain near real time data encompassing millions of datapoints which did not exist in the early years of

study. However, while the future of research in this field is promising, there are coverage gaps in the existing literature.

Despite researchers having significantly advanced our understanding of refugee destination patterns, through the development of gravity models, social network theory, and the classification of migration phases, some key context is often left out. Gravity models, while widely used, often perform inconsistently in low-data or high-volatility environments, limiting use in less-documented crises. Traditional push-pull theory struggles to account for nascent decline, where early, often elite, migrants respond to subtle precursors of collapse. And despite advances in causal inference and machine learning, most forecasting models remain tailored to individual case studies, lacking potential uses in forecasting.

Case studies further reinforce that no single factor dominates across all contexts. Gravity mechanisms are pronounced in the Ukrainian crisis, where cumulative networks amplified flows over time. In contrast, economic opportunity played a more central role in Venezuelan secondary migration, and accessibility shaped early movements during the Syrian conflict. The five variables are largely present across all cases, but their intensity varies by region. These variations suggest that predictive modeling must be context-aware, modular, and capable of adapting to shifting crisis dynamics.

This paper is a case study on Ukrainian migration into EU countries which have signed the temprorary protection order. This study uses attempts to responds directly to previous gaps in researchg by integrating thee five most consistently observed causal themes: accessibility, safety, opportunity, and gravity, and marrying them with political freedom scores. This case study will use a target variable of total presence of migrants to avoid the volatility of monthly predictions. For this reason, it is expected that the model will be best effective at post intial modeling (i.e.,

mid to late primary and secondary migration). This model is intended to be transferable across crises and usable by host nations for anticipatory planning. The model uses iterative Random Forest Modeling from the Ranger Package to identify the most valuable pull predictors while also preserving political push facors. The goal is to bridge the gap between explanatory insight and implimentable forecasting, contributing to both established theory and practical policy application.

3 Data and Methods

Data

This project draws upon multiple data sources to examine refugee migration patterns from Ukraine to European Union (EU) countries, with the goal of developing a predictive model for refugee flows. The most substantial and consistent dataset originates from Eurostat, providing migration flow data across EU member states. Additional data sources include the Freedom House political freedom indicators and custom-constructed geographic distance data, both of which contribute to capturing the key factors influencing refugee destinations.

The scope of this project includes both the five primary pull factors for refugees and the broader push factors originating from Ukraine. The five pull factors considered for this study include accessibility, safety, familiarity, opportunity, and gravity. However, familiarity was unable to be fully modeled due to a lack of consistent data. These factors collectively shape refugee decision-making and are reflected across the selected data sources.

In March 2022, the European Union passed a temporary protection order which has permitted up to 4 million refugees asylum across member states(Council of the European Union

2025). This order has been extended into the present and has been monitored through various datasets employed by this project, sourced from Eurostat. The policy decision provides a valuable temporal marker and structural explanation for observed migration patterns and will be a key consideration when modeling flows over time.

The target variable for the model is the total presence of refugees in a given European country. This stratified granularity is necessary to align with the temporal structure of Eurostat datasets and to allow for responsiveness to external events such as policy shifts or conflict escalations. The following countries have been excluded from analysis due to a lack of consistent or complete data: Albania, Andorra, Armenia, Azerbaijan, Belarus, Bosnia and Herzegovina, Georgia, Kosovo, Liechtenstein, Moldova, Monaco, Montenegro, North Macedonia, Russia, San Marino, Serbia, Ukraine, United Kingdom, and Vatican City. These exclusions, while limiting, are consistent with data-driven modeling practices aimed at ensuring integrity and comparability across observations.

The World Bank (World Bank 2025) provides critical socioeconomic data on countries and is commonly used for historical assessments in case studies. However, this source lags by at least a year, making monthly analysis into August of 2025 impossible. Therefore, World Bank data has been excluded from this study. This exclusion reinforces the need to rely heavily on Eurostat datasets, which provide both higher-frequency and more regionally tailored data.

To address this gap, multiple datasets provided by Eurostat have been aggregated to make this project feasible. Economic health of included countries is modeled based on the GDP and Main Components (Eurostat 2025f) dataset, which captures a country's economic viability through metrics including: GDP, gross value added, consumption expenditure, exports and imports of goods and services, employee compensation, and wages and salaries. All of these

indicators will be included in the model for individuals between the ages of 20 and 64, ensuring that economic pull factors are consistently measured across the relevant working-age population.

Population by Educational Attainment Level (Eurostat 2025e) captures the educational attainment of those between the ages of 15 and 64 and is recorded annually. This provides an important proxy for opportunity, as countries with higher levels of educational attainment may offer greater professional and economic opportunities to arriving refugees.

This is supplemented by the Unemployment by Sex and Age dataset (Eurostat 2025b), which provides monthly unemployment statistics across the included countries and the relevant date range. This dataset provides much-needed temporal resolution for labor market conditions, which represent a critical component of the opportunity pull factor. Annual employment data is also provided by the Employment and Activity by Sex and Age (Eurostat 2025b) dataset; however, its contribution to modeling will be limited as the data only extends to 2023. Nevertheless, the inclusion of both monthly and annual labor market indicators provides additional context to the opportunity structures present in each destination country.

The pull factor of safety is not explicitly captured within the provided data, as all destination countries included in the model are within Europe and broadly meet a baseline standard of safety and political stability. As a result, safety offers limited explanatory power in this context compared to analyses across more diverse global regions where safety differentials between destinations are more pronounced. For the purposes of this study, safety is treated as a near-constant, allowing other pull factors to play a more significant explanatory role.

The push factors are accounted for in the Freedom House (Freedom House 2025) dataset, which is recorded annually. The Freedom House dataset is used to measure the freedom of Ukrainian citizens relative to every other country. Scores are assigned based on the assessment of

various criteria including due process, freedom of assembly, election freedom, and governmental transparency. Less relevant components of the dataset have been filtered out, leaving the total freedom score and its fluctuation between 2020 and 2025 as the measure of interest. This variable provides a critical time-series perspective on how political conditions within Ukraine have evolved and how those changes may influence refugee flows.

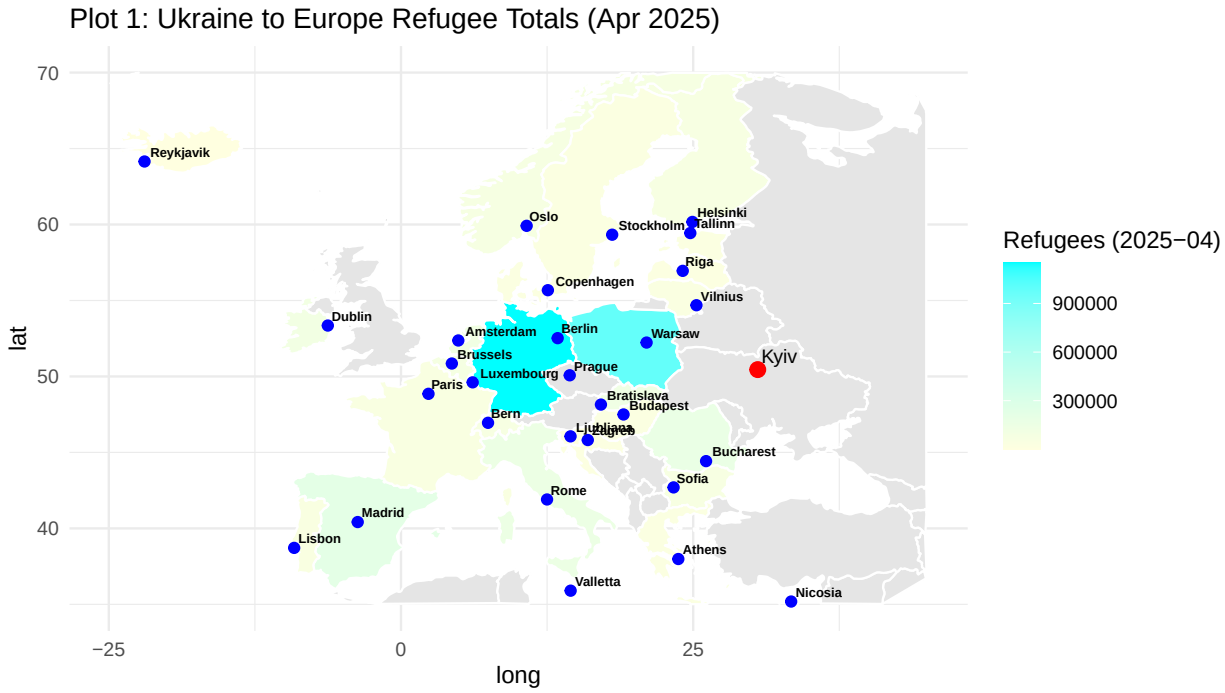
Established diaspora networks are fundamentally the most important aspect of existing gravity models. In response to the EU resolution to permit temporary protection, three Eurostat datasets were employed to capture monthly asylum trends: Asylum Applicants by Type (Eurostat 2025c), Beneficiary Country Refugee Totals (Eurostat 2025d), and decisions to grant temporary protection to applicants (Eurostat 2025a). The Beneficiary Country Refugee Totals dataset was used to establish both existing diaspora networks and the target variable, which is the fluctuation of migrants between months. During the modeling process, particular attention will be paid to the relationship between the size of existing diaspora networks and monthly migration fluctuations to avoid introducing endogeneity or spurious correlations into the model. This consideration is especially important given the use of Random Forest Models, where complex interaction effects and nonlinear relationships, if left unchecked, could obscure meaningful causal relationships.

An inherent aspect of refugee flow is distance from the home country to the destination country. The CEPII provides detailed geographic data on countries, which prior research has used to establish trends in displacement distances. However, due to processing limitations, this data was not accessible. The solution to this limitation was a dataset constructed by calculating the distance from each included country's capital city to Ukraine. Whether the country shares a border with Ukraine has also been included as a binary variable within this dataset. Together, these measures provide a pragmatic, if simplified, operationalization of geographic accessibility as a pull factor.

The objective of the model is to use temporally variable data on the political freedom of a country at war to assess the flow of refugees into host nations, while considering the pull factors of the host nations. The Random Forest Modeling approach is preferred for this project, as it is well-suited to handle the mix of continuous and categorical predictors, pervasive non-linear relationships, and complex interaction effects present in the integrated dataset.

Methods

To construct the modeling dataset, all relevant Eurostat, Freedom House, and geographic sources were cleaned, filtered, and harmonized to a common monthly structure spanning 2020–2025. Annual indicators such as political freedom scores and employment rates were expanded to monthly resolution using a custom stratification function, while datasets already reporting at the monthly level, such as refugee presence, unemployment, and inflation, were preserved in their native format. Country names were normalized across files, and extraneous or malformed entries were removed. Gravity-related features were constructed using cumulative refugee counts per country and proximity data, which was calculated from Kyiv to each EU capital using the haversine formula. The final dataset integrates push factors (Ukraine's political deterioration) with four primary pull mechanisms (accessibility, opportunity, safety, and gravity) across all EU/EEA nations with complete data coverage. While some processing steps required manual corrections due to inconsistent formatting, the result is a unified modeling frame exported as modeling_df_with_ukraine_freedom.csv. A full breakdown of the cleaning pipeline and source files is available on the referenced GitHub for replication and audit. This structured dataset now allows for the exploration of spatial migration patterns, such as those depicted in the choropleth visualization below.



This choropleth is a useful visual aid for readers to reference for the countries included in the study as well as the number of migrants present in each country in April of 2025.

Iterative Sizing and Variable Importance Strategy

Each model was trained on the same input data, using identical bootstrapping logic and default hyperparameters, with num.trees (number of trees) as the only changing input. The 500-tree model performed best in terms of out-of-bag (OOB) R^2, suggesting that the added complexity offered marginal but measurable gains in explanatory power. Specifically, R^2 improved from 0.962 (50 trees) to 0.965 (500 trees). This essentially has a negligible impact on performance, but it was

consistent across multiple iterations and reflected an actual gain in predictive stability, which is important given the moderate dimensionality and nonlinearity of the modeled data.

Once the model was finalized, attention turned to interpreting its internal logic. The Random Forest Model calculates variable importance using the total decrease in impurity (Gini or variance, depending on the outcome type) attributable to each predictor, averaged across all trees. This yields an "impurity importance" score for each variable, a proxy for how useful the feature was in splitting the data to reduce prediction error.

To make this output digestible, we converted the raw importance vector into a three-column, multi-row table, with variables grouped side-by-side for readability. This table does not just enumerate top predictors, it communicates scale, redundancy, and diminishing returns. Unsurprisingly, high-ranking variables include well-known economic indicators such as Imports of Goods and Services, GDP at Market Prices, and Employment Rate, as well as structural factors like Distance to Kyiv and Border Status. These reflect classical pull mechanisms found in migration literature.

What this process confirmed is that variable importance is not a ranking of theoretical significance but of statistical leverage. Some variables may be conceptually important yet provide little marginal benefit once better proxies or composite features are introduced. This is a necessary check for both modelers and policymakers: even theoretically justified variables don't always move the needle.

Finally, this entire process, benchmarking tree count, tuning ensemble depth, and surfacing variable importance, serves a dual purpose. First, it ensures the model is technically robust. But second, and more importantly, it ensures that models are appropriately iterated and that insights drawn from the model are grounded in the underlying mechanics, and that those mechanics are

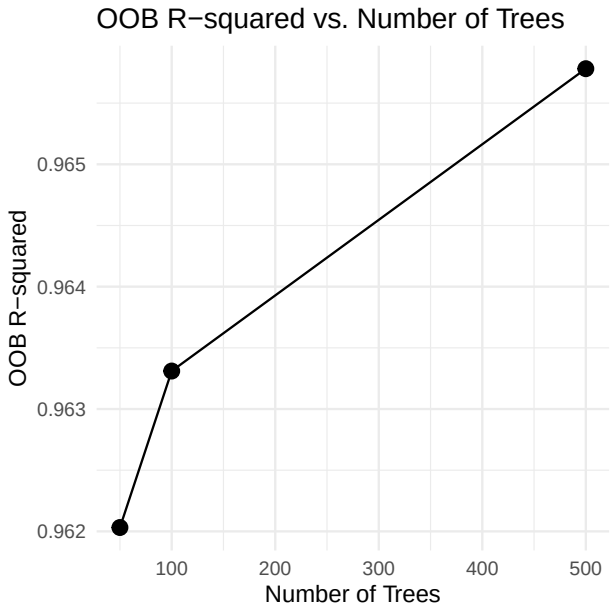
transparent.

After assessing Random Forest performance using an initial set of economic, demographic, and structural indicators, we proceeded to formalize our variable inclusion process and build the final ensemble model. The goal at this stage was twofold: (1) maximize out-of-bag performance, and (2) ensure a principled representation of all five core drivers of refugee migration, accessibility, opportunity, familiarity, gravity, and safety, with a particular focus on capturing political deterioration within Ukraine over time.

To justify the final ensemble depth, we benchmarked model performance using three commonly used tree counts: 50, 100, and 500 trees. All models were seeded identically to maintain consistency in bootstrapping and feature sampling.

Table 1: OOB R Squared Random Forest Models by Tree Count

Number of Trees	OOB R^2
50	0.96203
100	0.96331
500	0.96578



The 500-tree model slightly outperformed the 100-tree alternative (OOB $R^2 = 0.96578$ vs. 0.96331), confirming its selection as the final ensemble configuration. These improvements, though marginal in scale, consistently appeared across reruns and reflect stronger generalization rather than overfitting. More importantly, this step set the foundation for evaluating which

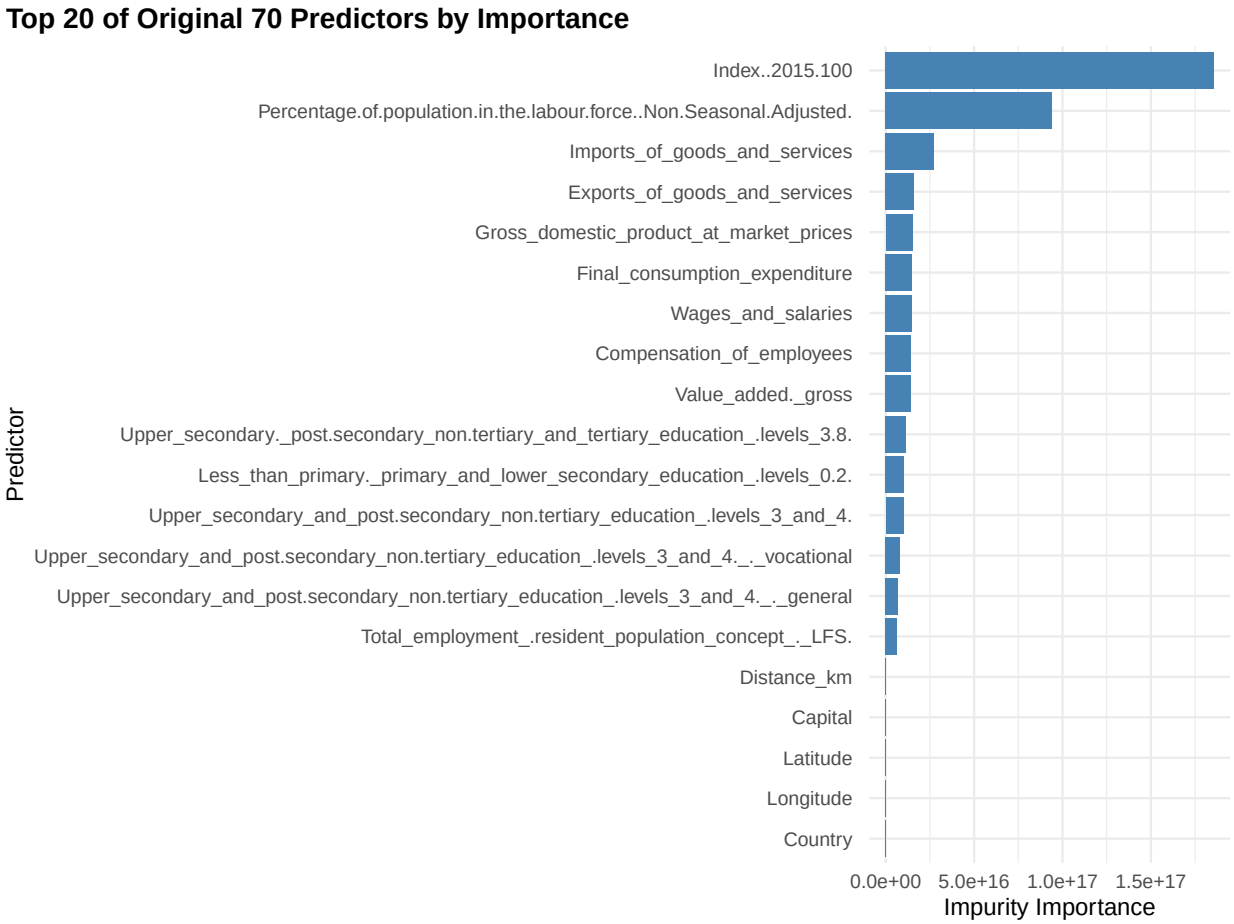
predictors were consistently valuable as trees were added.

Figure 1: Variable Importance Scores from Random Forest

Variable Importance Scores from Random Forest

Variable	Importance	Variable	Importance	Variable	Importance
Country	2.788630e+13	Class	0.000000e+00	YearMonth	1.386995e+12
Dataset.x	8.330207e+10	Index..2015.100	1.855769e+17	Percentage.of.population.in.the.labour.force..Non.Seasonal.Adjusted.	9.410451e+16
Unit_of_measure	5.901349e+12	Dataset.y	5.127376e+12	Compensation_of_employees	1.442823e+16
Exports_of_goods_and_services	1.612133e+16	Final_consumption_expenditure	1.488286e+16	Gross_domestic_product_at_market_prices	1.518620e+16
Imports_of_goods_and_services	2.719460e+16	Value_added_gross	1.436690e+16	Wages_and_salaries	1.486440e+16
Less_than_primary_primary_and_lower_secondary_education_levels_0.2.	1.033783e+16	Upper_secondary_and_post.secondary_non.tertiary_education_levels_3_and_4.	1.010032e+16	Upper_secondary_and_post.secondary_non.tertiary_education_levels_3_and_4__general	7.042945e+15
Upper_secondary_and_post.secondary_non.tertiary_education_levels_3_and_4__vocational	8.109948e+15	Upper_secondary_post.secondary_non.tertiary_and_tertiary_education_levels_3.6.	1.140441e+16	Total_employment_resident_population_concept_LFS.	6.268190e+15
Capital	5.235002e+13	Latitude	3.236966e+13	Longitude	2.919014e+13
Distance_km	5.316108e+13	Borders_Ukraine	1.642546e+13	Year	3.379681e+11
Region	0.000000e+00	CT	0.000000e+00	Status	0.000000e+00
PR.rating	1.200038e+11	CL.rating	1.580914e+11	A1	4.965972e+10
A2	0.000000e+00	A3	0.000000e+00	A	6.300034e+10
B1	1.001300e+11	B2	1.189423e+11	B3	0.000000e+00
B4	0.000000e+00	B	2.684662e+11	C1	1.673462e+11
C2	0.000000e+00	C3	0.000000e+00	C	1.419659e+11
Add.Q	7.275188e+10	Add.A	0.000000e+00	PR	3.408077e+11
D1	0.000000e+00	D2	0.000000e+00	D3	6.687714e+10
D4	1.247678e+11	D	9.002441e+10	E1	0.000000e+00
E2	0.000000e+00	E3	1.151291e+11	E	9.925257e+10
F1	0.000000e+00	F2	0.000000e+00	F3	1.043727e+11
F4	0.000000e+00	F	1.114618e+11	G1	1.350670e+11
G2	0.000000e+00	G3	0.000000e+00	G4	0.000000e+00
G	1.356892e+11	CL	1.553075e+11	Total	3.012856e+11

With the initial model trained, we visualized variable importance by impurity reduction. As shown below, the top-ranked variables were overwhelmingly economic and structural:



Here, Index 2015, Labor Force Participation, Imports, and Exports dominated, with geospatial indicators like Latitude, Longitude, and Country falling to the bottom. The education variables ranked in the midrange, and nearly all governance-related indicators, those from the Freedom House dataset, contributed no measurable importance.

This prompted two corrective actions: 1) Truncate the predictor set. Variables that were functionally inert (e.g., spatial coordinates, duplicated country labels) were removed. 2) Reevaluate political variables. Since Freedom House was the main proxy for Ukrainian push factors, it cannot be ignored, even if some indicators showed low raw importance. Instead, we isolated only the theoretically strongest measures.

To retain explanatory fidelity while eliminating noise, we manually filtered Freedom House

indicators using both empirical performance and codebook based rationale. The retained variables reflect democratic process, corruption, legal rights, association freedoms, and judicial integrity, all factors likely to affect refugee push pressure in a modern autocracy. The following Freedom House indicators were retained: (A1) Electoral Process (B1), (B2) Political Pluralism and Functioning of Government (C1) Freedom of Expression and Belief (D), (D3), (D4) Associational and Organizational Rights (E), (E3) Rule of Law and Due Process (F), (F3) Personal Autonomy and Individual Rights (G1) Control of Corruption. These measures were chosen not because they scored highly in preliminary impurity rankings, but because they offer a signal and represent fundamental dimensions of political collapse and freedom. They form the backbone of the model's representation of "push factors" from Ukraine.

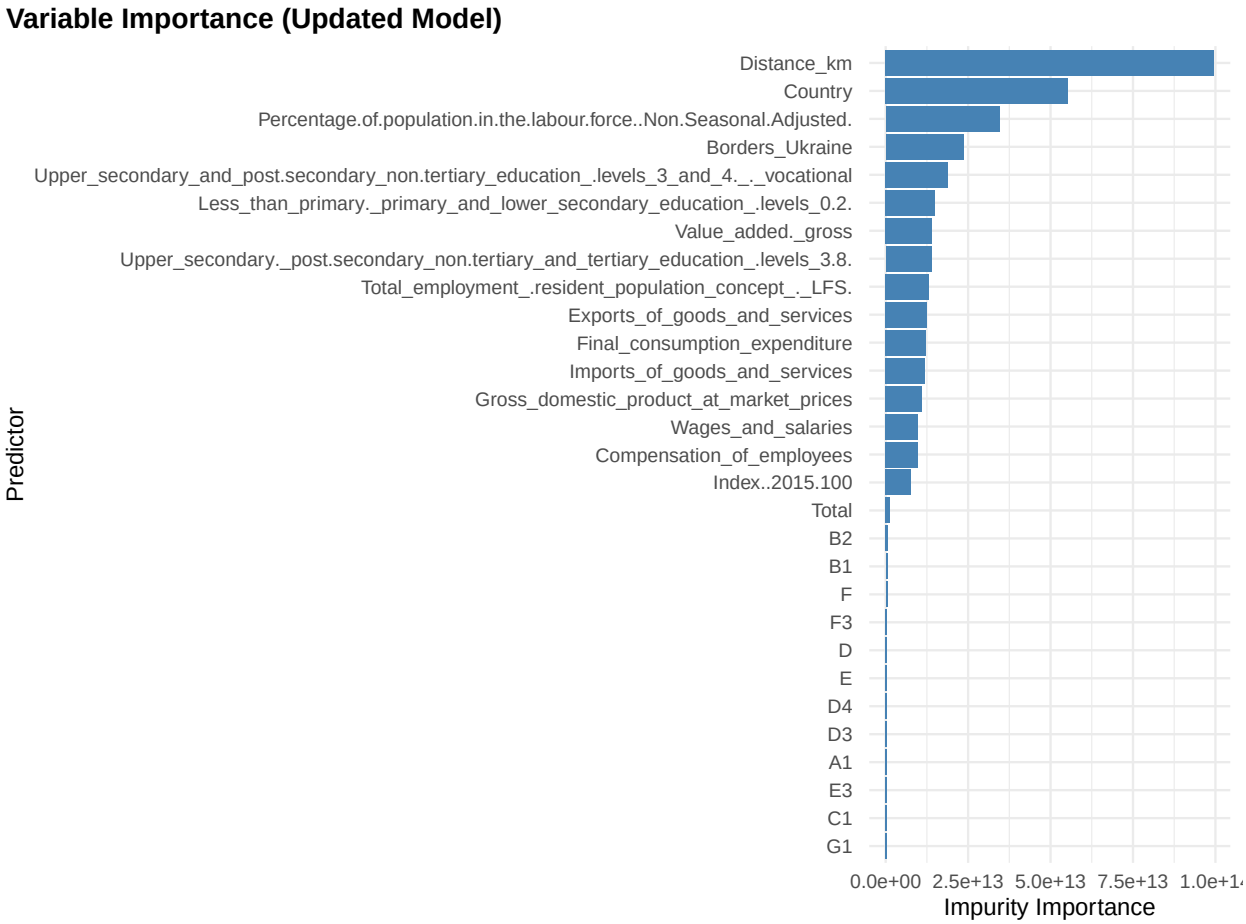
After finalizing the Freedom House subset, we constructed a new training dataset containing only the top-performing economic, geographic, and political variables.

Predictor	Predictor	Predictor
Index.2015.100	Total	Percent.Labor.Force
Imports_Goods	Exports_Goods	GDP_Market_Prices
Consumption_Final	Wages_Salaries	Comp_Employees
Value_Added_Gross	UpperSec&PostLvl3.8	LessThanPrim&LowerSec
UpperSecVoc_Lvl3&4	Employment_LFS	Distance_km
Borders_Ukraine	Country	A1
B1	B2	C1
D	D3	D4
E	E3	F
F3	G1	-

The model was retrained using this cleaner, more focused dataset. As shown in the updated variable importance graph below, the rankings shifted dramatically.

4 Results

Distance to Kyiv (Distance_km), border adjacency (Borders_Ukraine), and labor participation again dominate, but now several Freedom House indicators break into the visible range. While they don't outperform economic factors, their inclusion now enhances multidimensional fidelity across all five refugee migration drivers.



In fact, the model now better reflects: Accessibility (via Distance_km & Borders_Ukraine) , Opportunity (via Employment Rate, Wages, Education Levels) , Gravity (reflected in past flows and aggregated protection status), Safety / Governance (via assumed European standard of freedom). We must again consider that familiarity is not included in the model in any measurable way.

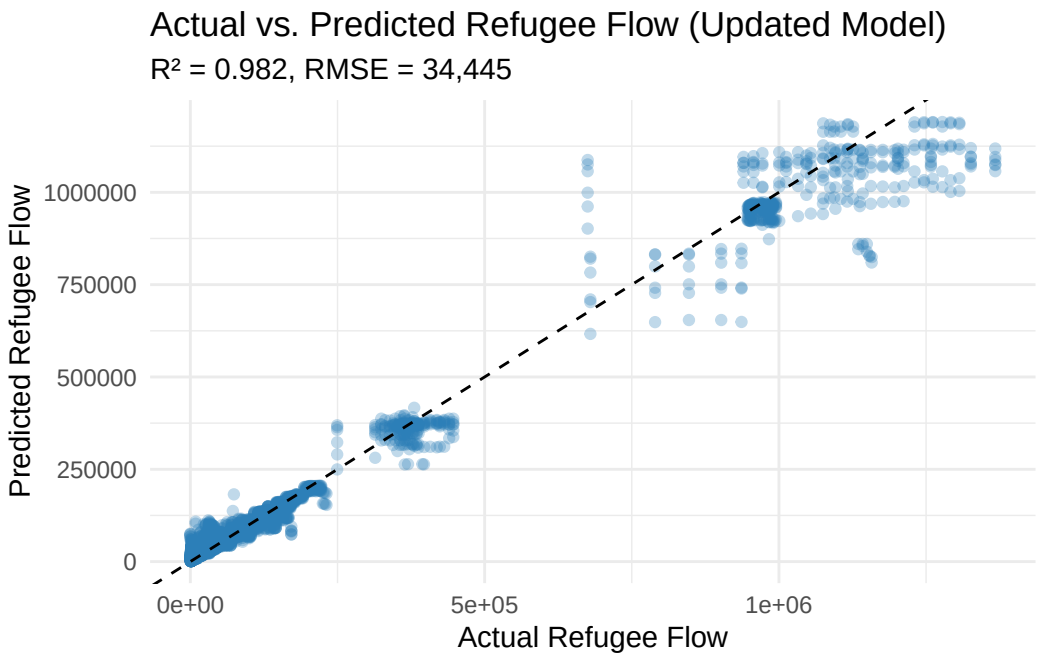
The final Random Forest Model contains only the strongest performing variables from each domain, empirically, theoretically, and temporally. We moved from a bloated, macroeconomic-heavy predictor space to a streamlined feature set that still covers four of the five included pillars of refugee theory as well as a few freedom indicators. The Freedom House indicators used were handpicked not for their ability to represent state failure with longitudinal

integrity. The final model produces cleaner importance signals and a more reliable platform for predictor importance, actual predictions, and policy insight.

This design supports not just Ukraine-specific modeling, but extensibility to future crises where both pull and push variables must be integrated under time constraints and data scarcity.

At this point, recall the model's target variable was not monthly flow, but rather the total number of Ukrainian refugees present in each country during a given month. This distinction is crucial. Rather than forecasting near-term spikes or border inflows, the model estimates cumulative refugee presence, the lasting footprint of displacement across Europe.

Model Fit and Predictive Accuracy



The updated Random Forest Model achieved an R^2 of 0.982 and a root mean square error (RMSE) of 34,445. This RMSE, while seemingly large, is quite strong given that the dependent variable ranges into the millions for countries like Poland and Germany. In relative terms, the model's average error is less than 3% of the highest observed refugee counts, suggesting a tight fit

and effective pattern recognition.

The scatterplot of predicted vs. actual values shows near-perfect alignment in the low to mid-range, with mild overdispersion at the upper end of the scale. These deviations are acceptable given the heterogeneity of humanitarian policy and the lack of explicit policy ceilings encoded in the dataset. Overall, the model accurately captured the structural determinants of where refugees reside, not just where they arrive. Germany and Poland top the list, each with over 1.1 million projected Ukrainian refugees. This aligns with observed policy trends, diaspora concentration, and logistical proximity to Ukraine. Czechia, Romania, and Slovakia form the next tier, representing high-gravity secondary destinations. Italy, Hungary, and Spain round out the top ten, with the rest acting as capacity absorbers rather than frontline recipients.

Table 3: Predicted Total Ukrainian Refugee Presence by Country

Country	Refugees	Country	Refugees	Country	Refugees
Germany	1190396	Belgium	142831	Portugal	102937
Poland	1119924	Slovenia	136900	Switzerland	101557
Czechia	438120	Lithuania	136049	Sweden	98237
Romania	265469	Croatia	135893	Malta	97275
Slovakia	238688	Estonia	135285	France	96280
Spain	220820	Cyprus	124996	Iceland	93343
Bulgaria	206230	Ireland	123035	Ukraine	91392
Italy	176692	Norway	120661	Denmark	82439
Hungary	171277	Finland	119363	Greece	78712
Netherlands	143115	Latvia	115642	Luxembourg	70372

The decision to model total refugee presence rather than inflows offers notable conceptual and

methodological strengths. Monthly inflow forecasting is unstable due to data volatility, policy reversals, and lagging indicators. Presence, by contrast, captures the cumulative footprint of displacement, offering a more policy-relevant and stable target. This shift elevates the model's utility: housing, schooling, and medical planning hinge more on who stays than who arrives. Moreover, it reduces the noise introduced by short-term fluctuations and instead brings into focus the geographies where displacement has hardened into a structural reality. In this regard, the model reflects a more realistic and strategic lens through which long-term burden and infrastructure needs can be assessed.

Empirically, the model demonstrates impressive performance. The final Random Forest Model achieves an R^2 of 0.982 and a root mean square error of just 34,445, less than 3% of the highest observed refugee counts. This tight fit is particularly notable given the wide range of refugee totals across countries and the policy heterogeneity embedded in the target region. The model accurately ranks countries by refugee burden, with Germany, Poland, and Czechia emerging as consistent high-gravity destinations. These rankings align well with both proximity and historical patterns of diasporic linkage, reinforcing confidence in the model's output. Importantly, the model was built after selecting variables that represent four pillars of refugee migration theory, accessibility, opportunity, gravity, and safety, thereby ensuring multidimensional validity.

However, key limitations remain. First, the model does not include explicit policy ceiling variables, and thus assumes open-ended absorptive capacity in each country. This may overstate real-world capacity, particularly in nations with restrictive immigration policies or saturation-level strain on services. Second, while the model predicts presence, it does not account for onward migration or returns, nor does it capture dynamic feedback loops, such as how prior arrivals influence future flows. This makes it a cross-sectional snapshot rather than a reactive forecasting tool. The data was also unable to encapsulate existing welfare support provided by EU states.

This is a notable limitation since an established correlation exists between developed welfare states and displaced populations. Additionally, the model's reliance on annualized indicators (like GDP, Freedom House scores) stratified into monthly rows introduces a lag that limits responsiveness to sudden within-year political or economic shocks. While this approach was necessary for harmonization across data sources, it sacrifices temporal granularity and may mute the impact of fast-moving events such as military escalations or asylum law reforms. The reality is that the model lags, and the biggest mitigating factor to that lag is the total population being the target variable.

Compounding this is the fact that many datasets were originally structured for different analytical purposes and required significant preprocessing. The absence of harmonized real-time data forced repeated imputation and backward propagation of previous-year values to fill gaps. Some variables, such as governance indicators, showed low importance not due to conceptual irrelevance but due to collinearity or proxy saturation. Others, like geographic coordinates, under performed because more synthetic spatial features (e.g., distance to Kyiv) captured the relevant information more effectively. This reflects a broader challenge in migration modeling: theoretically important variables often provide limited marginal gain once higher-leverage proxies are introduced.

5 Conclusion

The employed Random Forest Model demonstrates utility in identifying the most influential pull factors shaping Ukrainian refugee settlement patterns within the EU. While forecasting initial inflows is often unstable due to policy shifts, volatility, or data lag, modeling refugee presence offers a more stable and policy-relevant output. It reflects not just border crossings, but the

enduring footprint of diaspora networks, where displaced populations ultimately remain, and where long-term support structures must follow.

By focusing on total presence, this model excels not at capturing early-stage flight, but rather at characterizing late primary and sustained secondary migration, where refugees seek opportunity, not just safety. This distinction is vital. As a diaspora network solidifies, refugees become more responsive to structural conditions within host countries. The model leverages this reality and identifies which features shape that response most strongly.

Across over 70 tested predictors, and after granulating for the top 29, the most important variables for predicting refugee presence were overwhelmingly economic and structural: labor force participation, imports and exports, GDP, and wages. These findings suggest that opportunity metrics, not border proximity alone, are what sustain and expand refugee communities after initial arrival. Distance to Kyiv and whether a country shares a border with Ukraine remained salient but fell behind opportunity variables in overall predictive leverage.

Conversely, political freedom indicators from Freedom House consistently ranked among the least important predictors. While Ukraine's internal conditions undeniably drive initial displacement, they appear to play a minor role in shaping patterns in secondary migration. This supports the conclusion that once safety is broadly guaranteed, material and structural opportunity, not ideological freedom, determines movement. The culminating point is that political freedom metrics are significant indicators for primary migration, but socioeconomic factors in host countries are significantly more important when determining destinations of secondary migration.

For policymakers, the implications are relatively straightforward. Countries aiming to retain or deter refugee populations should focus less on border management and more on economic

signals. Those offering stable employment, higher education, and strong wage markets are more likely to retain refugees long term, regardless of geographic proximity. Finally, this project offers a replicable framework for rapid modeling during future crises. By integrating gravity-based logic with Random Forests, it balances interpretability and predictive strength, enabling both scholarly insight and real-world application.

References

Andersson, H., and K. Jutvik. 2022. "Do Asylum-Seekers Respond to Policy Changes? Evidence from the Swedish–Syrian Case." *The Scandinavian Journal of Economics*. doi:10.1111/sjoe.12510.

Bertoli, S., H. Brücker, and J. Fernández-Huertas Moraga. 2022. "Do Applications Respond to Changes in Asylum Policies in European Countries?" *Regional Science and Urban Economics* 93: 103771. doi:10.1016/j.regsciurbeco.2022.103771.

Cottier, F. 2024. "Projecting Future Migration with Bayesian Hierarchical Gravity Models of Migration: An Application to Africa." *Frontiers in Climate* 6. doi:10.3389/fclim.2024.1384295.

Council of the European Union. 2025. "EU Member States Agree to Extend Temporary Protection for Refugees from Ukraine." *Press Release 477/25*.

Di Iasio, V., and J. Wahba. 2024. "The Determinants of Refugees' Destinations: Where Do Refugees Locate Within the EU?" *World Development* 177: 106533. doi:10.1016/j.worlddev.2024.106533.

Eurostat. 2025a. "Asylum Applicants and Pending Applicants by Citizenship Age and Sex – Monthly Data (Migr_asyappctzm)."

Eurostat. 2025b. "Employment Rate by Sex Age and Nationality – Quarterly Data

(Lfsi_emp_q_h)."

Eurostat. 2025c. "First Instance Decisions on Asylum Applications by Type of Applicant – Monthly Data (Migr_asytpsm___custom_17211992)."

Eurostat. 2025d. "First Instance Decisions on Asylum Applications for Family Members – Monthly Data (Migr_asytpfm)."

Eurostat. 2025e. "Labour Force Survey Activity Rates by Sex Age and Educational Attainment Level – Annual Data (Edat_lfse_03)."

Eurostat. 2025f. "National Accounts: Gross Domestic Product and Main Components (Output Expenditure and Income) – Quarterly Data (Nama_10_gdp)."

Freedom House. 2025. "Freedom in the World 2025."

Frith, M. J., M. Simon, T. Davies, A. Braithwaite, and S. D. Johnson. 2019. "Spatial Interaction and Security: A Review and Case Study of the Syrian Refugee Crisis." *Interdisciplinary Science Reviews* 44(3-4): 299–317. doi:10.1080/03080188.2019.1670439.

Greene, R. N., J. M. Hess, B. Soller, S. Amer, D. T. Lardier, and J. R. Goodkind. 2023. "Expanding Social Network Conceptualization, Measurement, and Theory: Lessons from Transnational Refugee Populations." *Journal of Applied Social Science* 17(3). doi:10.1177/19367244231172426.

Guichard, L., and J. Machado. 2024. *The Externalities of Immigration Policies on Migration Flows: The Case of an Asylum Policy.* Institute of Labor Economics (IZA). IZA Discussion Papers. https://hdl.handle.net/10419/295958.

Hierro, M., and A. Maza. 2024. "How Social Networks Shape Refugee Movements in Wartime: Evidence from the Russian Attack on Ukraine." *International Migration Review.* doi:10.1177/01979183241240712.

Lanati, M., and R. Thiele. 2024. "South-South Refugee Movements: Do Pull Factors Play a Role?" *Economics and Politics* 36(2): 928–58. doi:10.1111/ecpo.12275.

Micevska, M. 2021. "Revisiting Forced Migration: A Machine Learning Perspective." *European Journal of Political Economy* 70: 102044. doi:10.1016/j.ejpoleco.2021.102044.

Neumayer, E. 2005. "Bogus Refugees? The Determinants of Asylum Migration to Western Europe." *International Studies Quarterly* 49(3): 389–409. doi:10.1111/j.1468-2478.2005.00370.x.

Schmeidl, S. 1997. "Exploring the Causes of Forced Migration: A Pooled Time-Series Analysis, 1971–1990." *Social Science Quarterly* 78(2): 284–308. http://www.jstor.org/stable/42864338.

Suleimenova, D., D. Bell, and D. Groen. 2017. "A Generalized Simulation Development Approach for Predicting Refugee Destinations." *Scientific Reports* 7(13377). doi:10.1038/s41598-017-13828-9.

UNHCR. 2025. “Syria Regional Refugee Response.”

World Bank. 2025. “AI-Powered Refugee Forecasting: Preparing for Refugee Movements Before They Happen.”