Modeling Refugee Presence Across Europe: Ukrainian Crisis Case Study Using Random Forest Analysis

Daniel C. MacLeod

This project proposes a model to evaluate refugee migration flows from Ukraine to European Union countries, focusing on the critical question of why refugees settle where they do after displacement. Building on migration theory and advances in machine learning, the model integrates five key causal themes: accessibility, safety, familiarity, opportunity, and gravity. It also incorporates political freedom scores from Ukraine as dynamic push factors. Using Eurostat, Freedom House, and custom geographic datasets, the study models total refugee presence rather than inflow, capturing the lasting footprint of displacement. An iterative Random Forest approach combines the structural logic of gravity models with the flexibility of ensemble learning to handle nonlinear and temporal effects. Results show that once safety is assured, economic opportunity, border accessibility, and proximity outweigh political freedoms in determining long-term refugee presence. These findings offer policymakers a data-driven framework for anticipating sustained refugee settlement patterns in similar European crises where initial displacement has stabilized.

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 examines crisis drivers including domestic conflict, economic collapse, and international war. Better forecasting tools can help host nations prepare to both serve their own citizens while also moderating their 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.

This study asks: *"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 an identifying durable, high-importance predictors 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.

This study's novelty lies in integrating political freedom scores, representing dynamic push conditions in Ukraine, directly into a modular Random Forest framework built on the five most consistently observed pull factors in refugee research. While prior work has applied gravity models or economic variables in isolation, this approach blends the explanatory grounding of migration theory with the predictive flexibility of ensemble methodology.

2 Literature Review

Migration theory frames refugee decisions through push-pull dynamics, refined into primary, secondary, and nascent stages, which differ in urgency and drivers (Frith et al. 2019). Across contexts, five consistent pull factors emerge: accessibility, safety, familiarity, opportunity, and gravity (Hierro and Maza 2024). These shape initial displacement and long-term settlement, with proximity, political stability, and cultural ties often leading early movements, and economic opportunity and diaspora networks sustaining later flows.

Gravity models, inspired by Newtonian gravity and often structured like regressions, have been used to analyze flows from Venezuela, Syria, and Ukraine, successfully incorporating 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. 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 relevance 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 temporary protection order. This study attempts to respond directly to previous gaps in research by integrating the 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 initial modeling (i.e., mid to late primary and secondary migration). The model uses iterative Random Forest Modeling from the Ranger Package to identify the most valuable pull predictors while also preserving political push factors. The goal is to bridge the gap between explanatory insight and implementable 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 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. Due to inconsistent data across EU states, familiarity was excluded from the model, an omission noted as a limitation in interpreting results. 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. For this reason, 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 Eurostat Gross Domestic Product (GDP) and Main Components (Eurostat 2025f) datasets to capture 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.

Safety is not explicitly captured in the data, as all included countries meet baseline European standards for stability. In this context, safety is treated as a near-constant, though this limits generalization to global or less-stable regions.

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.

While past migration research has largely relied on linear regression, Poisson Pseudo Maximum Likelihood (PPML), and logistic regression models, these approaches often struggle with the complex, nonlinear, and high-dimensional structure of modern refugee data. For example, PPML models excel at handling bilateral flows and origin-destination fixed effects, and were instrumental in modeling early EU asylum flows (Di Iasio and Wahba 2024). However, their reliance on strong distributional assumptions and limited ability to capture interaction effects makes them ill-suited for evolving multi-factor refugee contexts, especially where monthly refugee counts fluctuate under rapidly changing conditions.

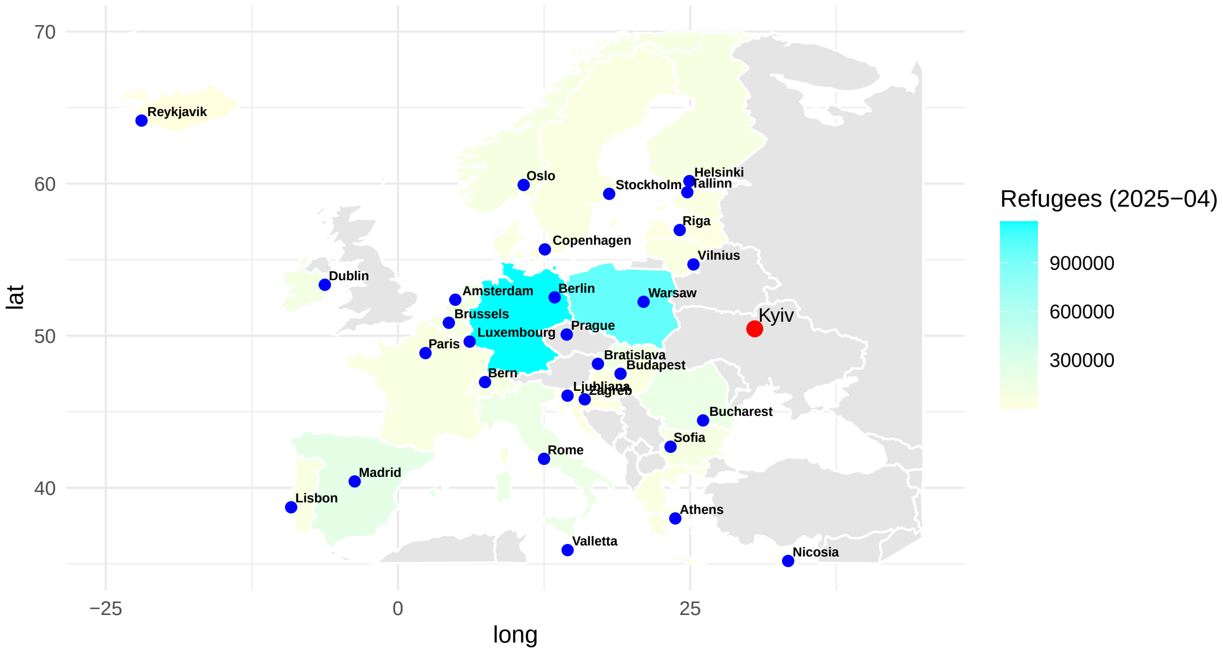
Random Forests present a pragmatic alternative. As an ensemble learning method, they mitigate overfitting through tree averaging and handle mixed-type variables without the need for transformation. Additionally, they consider nonlinearities and high-order interactions that would be obscured in additive models. While Random Forests are often critiqued for limited interpretability, this weakness is addressed through variable importance rankings and model simplification strategies deployed later in this paper.

Several candidate models were initially tested, including Ordinary Least Squares, Lasso Regression, and PPML; each failed to match the out-of-bag (OOB) performance of the Random Forests, particularly when modeling total refugee presence (as opposed to inflows). Given this modeling target, Random Forests provided the best fit while maintaining policy relevance. With that established, this choice has caveats. The method's flexibility increases the risk of overfitting to context-specific patterns. As such, generalizing the model to non-European or pre-protection-order contexts should be done with care and re-tuned variable sets. This approach, therefore, is best understood as a modular forecasting scaffold—effective when built upon theory, but not able to be adapted as a solution for all crises.

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 from 2020 to 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, calculated as the distance from Kyiv to each European capitol, 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 (MacLeod 2025) 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.

Plot 1: Ukraine to Europe Refugee Totals (Apr 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 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). The increased R^2 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, the next step was to interpret 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 only 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.

Some variables may be conceptually critical to understanding refugee behavior, yet contribute little to model performance when similar or composite features dominate the data structure. Their absence in importance rankings reflects statistical redundancy, not theoretical insignificance. 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. Second, and more importantly, it ensures that models are appropriately iterated, 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. First is to maximize OOB performance. The second is to ensure a principled representation of all five core drivers of refugee migration with a particular focus on capturing political deterioration within Ukraine over time.

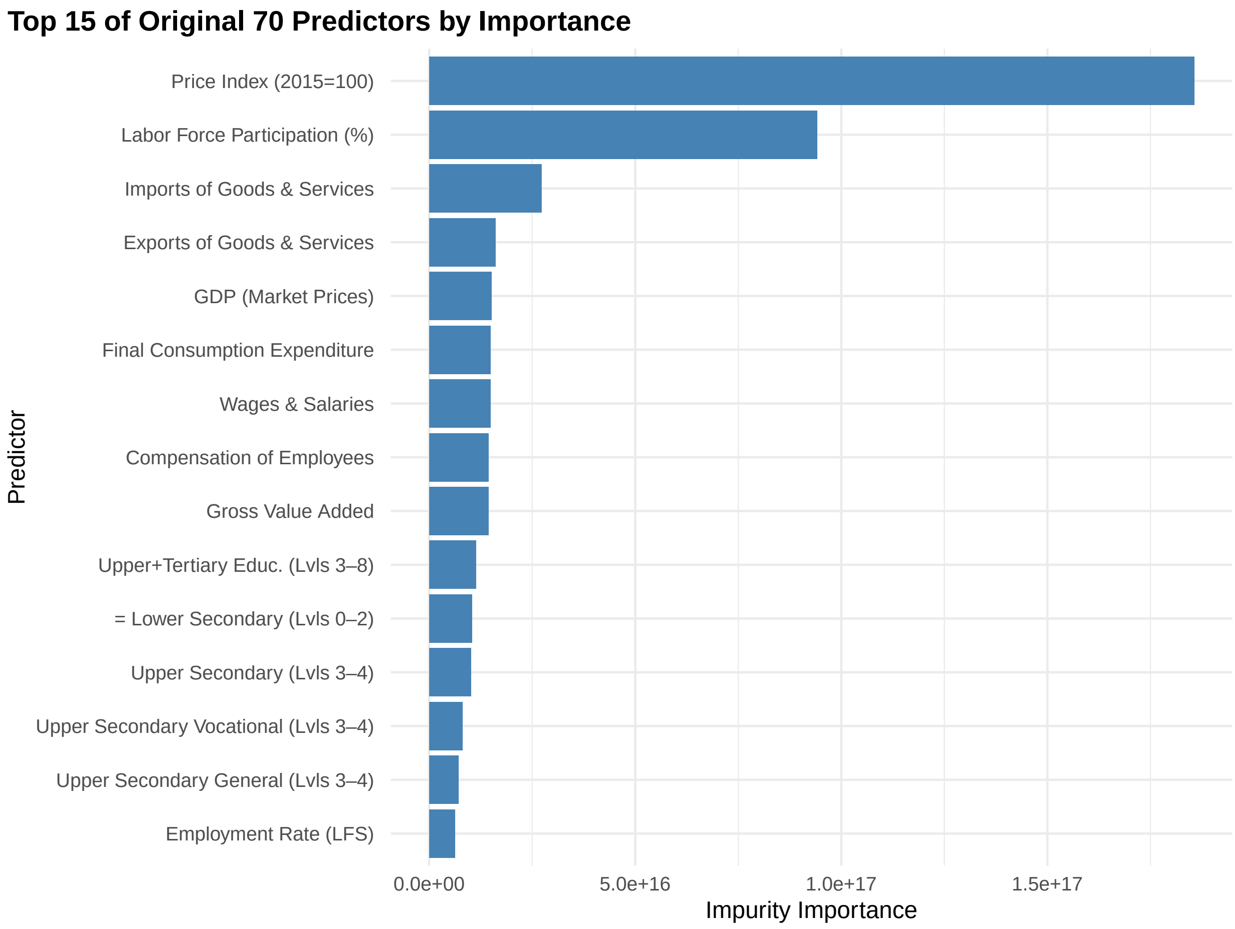
To justify the final ensemble depth, we benchmarked model performance using three commonly used tree counts: 50, 100, and 500. All models were seeded identically to maintain consistency in bootstrapping and feature sampling.

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 strongly on the models ability to generalize with available data. More importantly, this step set the foundation for evaluating which predictors were consistently valuable as trees were added. The full variable importance scores are provided in the project's GitHub repository (MacLeod 2025) for transparency and replication.

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



This chart shows the 15 most influential predictors from the original 70-variable Random Forest model, ranked by impurity importance. Although the original model contained a much larger set of predictors, truncating to the top 15 focuses attention on the variables with the greatest explanatory power. This snapshot highlights the model's initial tendency to prioritize macroeconomic and structural indicators over governance or proximity variables. The comparison with the updated model underscores how targeted feature refinement can shift the importance profile, improving interpretability while retaining predictive strength.

Here, Price Index (2015=100), 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.

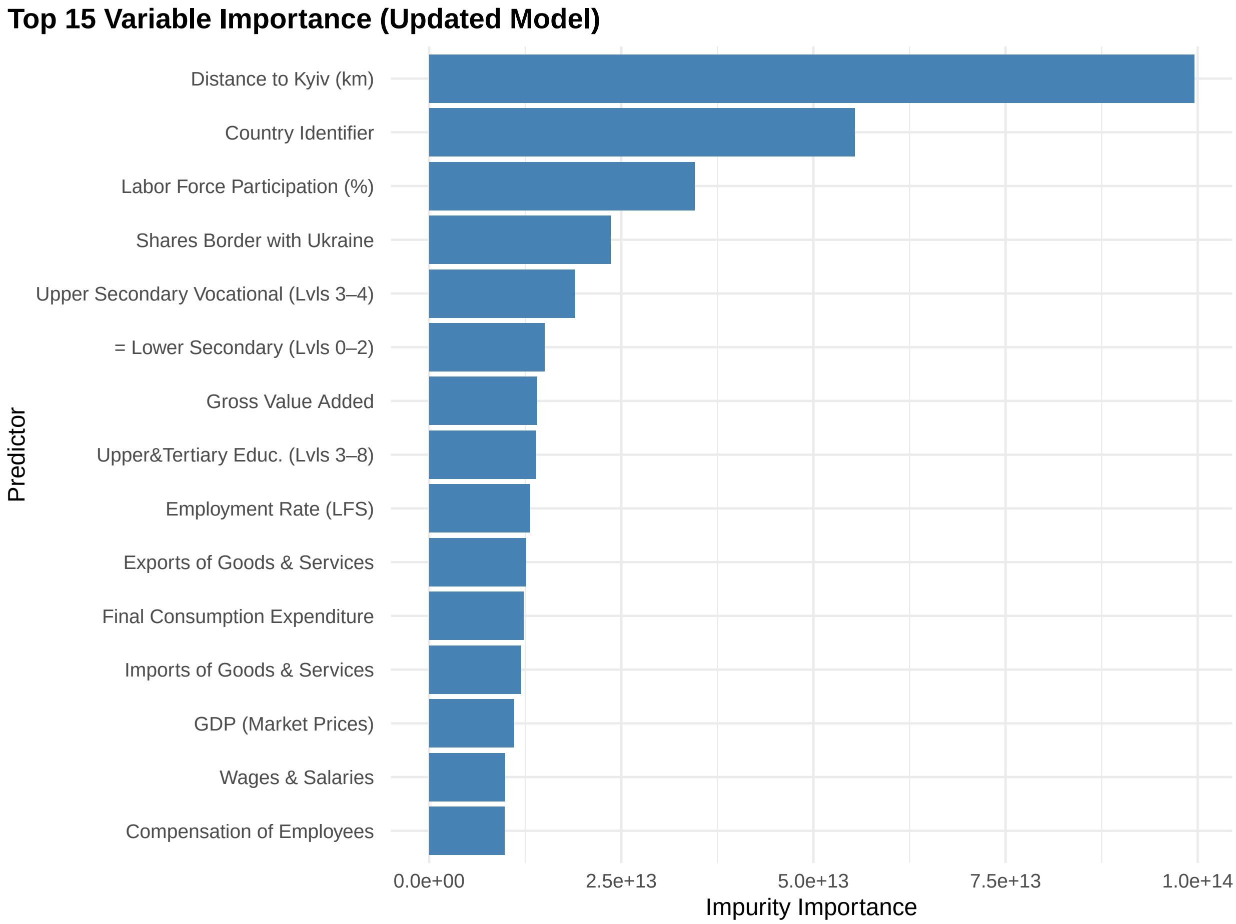
Predictor	Predictor	Predictor
Price Index (2015=100)	Total Population	Labor Force Participation (%)
Imports Goods & Services	Exports Goods & Services	GDP (Mrkt Prices)

Predictor	Predictor	Predictor
Final Consumption Expenditure	Wages & Salaries	Compensation of Employees
Gross Value Added	Upper Educ. (Lvls 3–8)	Lower Secondary (Lvls 0–2)
Secondary Vocational (Lvls 3–4)	Employment Rate	Distance to Kyiv (km)
Shares Border with Ukraine	Country Identifier	A1: Electoral Process
B1: Political Pluralism	B2: Govt Functioning	C1: Expression & Belief
D: Assoc. & Org. Rights	D3: Assoc. Rights	D4: Org. Rights
E: Rule of Law	E3: Due Process	F: Personal Autonomy
F3: Movement Rights	G1: Control of Corruption	-

The model was retrained using this cleaner, more focused dataset. The results of this updated variable importance graph is discussed in the final ranking table in the results section.

4 Results

Distance to Kyiv, border adjacency, and labor participation again dominate, but now several Freedom House indicators break into the visible range. While they do not outperform economic factors, their inclusion now enhances multidimensional fidelity across all of the four employed refugee migration drivers. Despite strong performance within this case study, caution should be exercised when generalizing to other refugee contexts. Different crises may elevate different drivers. For instance, opportunity dominated Venezuelan secondary migration, while accessibility was paramount for Syrians. The model's strength lies in its adaptability and modular construction, but replication in new contexts would require re-tuning to reflect local dynamics. Thus, this study is best understood as a transferable framework, not a universal pannacea.



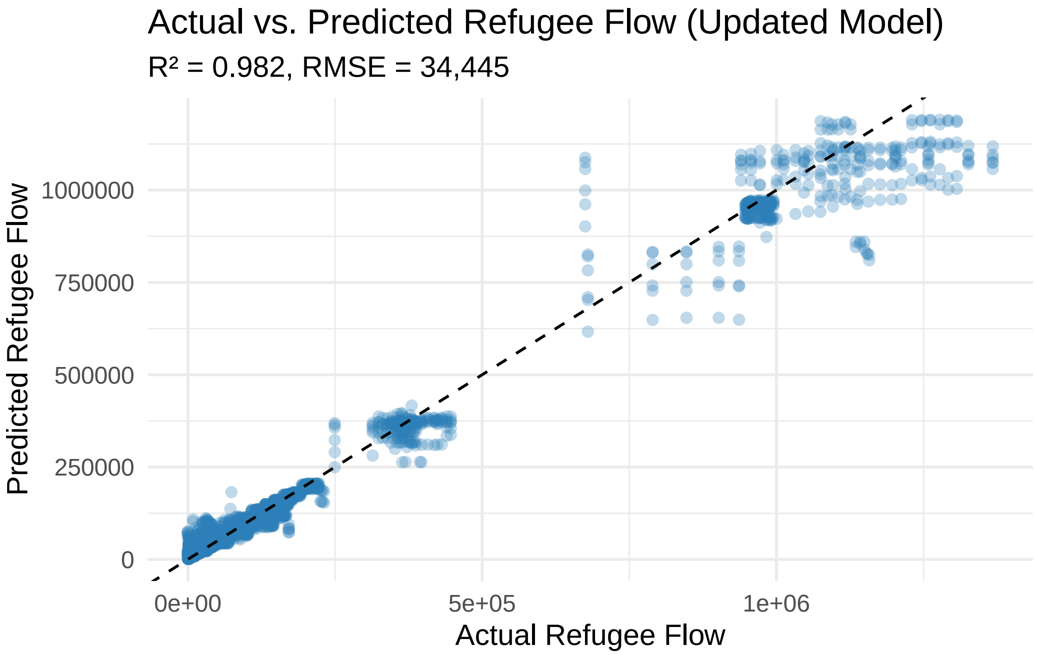
This chart displays the 15 most influential predictors from the updated Random Forest model, ranked by impurity importance. The full model included a larger set of predictors, but this view is intentionally truncated to emphasize the strongest contributors to explaining total refugee presence. By focusing on the top variables, this figure highlights the structural and economic factors that most consistently drive the model's predictive accuracy, while filtering out lower-impact variables that add little explanatory power. These results serve as a key step in identifying durable predictors for integration into broader ensemble forecasting frameworks.

The full list of predictors can be found in this GitHub repository (MacLeod 2025). Upon review, 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).

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.

Again, 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.



The scatterplot compares predicted refugee presence against observed values for each country-month in the dataset with the dashed line representing perfect prediction.

The updated Random Forest model achieved an R^2 of 0.982 and a Root Mean Square Error (RMSE) of 34,445, indicating strong overall fit. While predictions align closely with actual values in countries with large refugee populations, dispersion increases for smaller or more policy-restricted destinations, reflecting the model's sensitivity to scale and context-specific constraints. While the updated model achieved a low aggregate RMSE, this alone does not confirm accuracy. High aggregate fit can mask large relative errors in smaller-host countries, especially where predictions are more sensitive to policy ceilings, infrastructure limits, or sparse migration links.

In these cases, the model's structural accuracy at scale may obscure volatility at the margins. Evaluating performance requires both country-level error analysis and summary metrics to identify where predictive accuracy breaks down. This approach also helps reveal the contextual factors, such as geographic proximity, absorptive infrastructure, and prior migration ties, that drive those differences.

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 vs Actual Ukrainian Refugee Presence by Country

Country	Predicted	Actual	% Error	Country	Predicted	Actual	% Error
Germany	1190396	1306505	8.9	Cyprus	124996	22955	444.5
Poland	1119924	1367555	18.1	Ireland	123035	111020	10.8
Czechia	438120	446695	1.9	Norway	120661	79625	51.5
Romania	265469	182310	45.6	Finland	119363	71110	67.9
Slovakia	238688	131990	80.8	Latvia	115642	48935	136.3
Spain	220820	231260	4.5	Portugal	102937	56695	81.6
Bulgaria	206230	171555	20.2	Switzerland	101557	67415	50.6
Italy	176692	166935	5.8	Sweden	98237	47595	106.4
Hungary	171277	40240	325.6	Malta	97275	2305	4120.2
Netherlands	143115	122075	17.2	France	96280	69510	38.5
Belgium	142831	88270	61.8	Iceland	93343	4090	2182.2
Lithuania	136049	80715	68.6	Denmark	82439	38595	113.6
Croatia	135893	26440	414.0	Greece	78712	35180	123.7

Estonia	135285	39655	241.2	Luxembourg	70372	4520	1456.9

The decision to model total refugee presence, rather than monthly inflows, stemmed from established literature and theoretical logic. While inflow data is more immediately reactive, it is naturally noisier. By contrast, presence reflects the cumulative retention of displaced persons, a more stable, policy-relevant target for long-term planning in housing, employment, schooling, and public health.

The final Random Forest model achieved an R^2 of 0.982. These figures suggest a good model fit in aggregate. Indeed, the model correctly ordered many countries by magnitude of refugee burden, identifying Germany, Poland, and Czechia as consistent high-retention destinations. These outputs aligned with geographic proximity, transit capacity, and existing Ukrainian diaspora presence. The model's most important predictors were economic opportunity, border accessibility, and physical proximity, in line with the three modeled theoretical pillars of refugee migration: accessibility, opportunity, and gravity.

However, the addition of actual observed values by country presents a striking contrast. Though the model ranked countries well, its absolute predictions varied widely, with percent errors ranging from under 10% to over 2,000%. Some cases, like Czechia (1.9% error), Spain (4.5%), and Germany (8.9%), demonstrate strong fidelity between predicted and actual values, especially among countries with large refugee footprints. This suggests that the model captured the macro-scale structural forces reasonably well. But in many cases, especially among smaller or peripheral countries, the model faltered. Iceland (2182% error), Malta (412.0%), Luxembourg (1456.9%), and Slovenia (1231.7%) reveal dramatic overestimates, underscoring the model's inability to incorporate policy ceilings, logistical bottlenecks, or small-scale absorption limits.

This lends insight on how the model functions and the purpose it served. The model's failure to scale predictions to extreme ends of the target distribution is a referendum on the limitations of high-level modeling when applied to target variables of disparate outcomes. It highlights the balance between modeling for generalization and modeling for precision. No model can do both perfectly when the data contains vast disparities, and here, the model's strength in ordering burden came at the cost of predictive specificity for countries with small or policy-bound refugee totals. With this context established, the analysis transitions from model performance to its broader implications for policy and forecasting, emphasizing how identified predictors can inform anticipatory planning in future displacement crises. These performance patterns set the stage for interpreting the model's broader implications for policy and forecasting.

5 Conclusion

This project began by asking the 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?*" The model revealed that political conditions inside Ukraine are not a sustained important variable when modeling secondary migration. It also found little evidence that political freedom in destination countries plays a strong role in shaping where refugees ultimately stay. Once safety is assured, economic opportunity and geographic accessibility, not ideology, define long-term presence.

Choosing to model total presence rather than inflow was both a technical and conceptual decision. Presence reflects infrastructure burden, community permanence, and the stabilization of diaspora networks, critical factors for housing, education, and labor policy. This orientation produced fruitful aggregate results: the final Random Forest model achieved a strong fit, and it consistently highlighted economically vibrant, accessible states as key absorbers.

The findings also carry practical weight. Countries seeking to retain or repel refugee populations cannot rely on abstract freedoms alone. Material structure, employment, wages, and proximity will dominate decisions once a refugee crosses a border. For crisis modelers, the lesson is equally clear: focus not just on predicting numbers, but on understanding which variables are most important. To this end we should observe that model accuracy was highest for countries geographically closer to Ukraine, which tended to have lower percentage errors. This suggests that physical proximity, beyond being a strong pull factor, may also enhance the model's predictive reliability. Future research should investigate why distance amplifies predictive accuracy and explore modeling approaches that explicitly weight proximity-based effects.

These findings reinforce that the model's greatest contribution lies in identifying durable, high-importance predictors rather than generating flawless point estimates. By exposing where errors concentrate, particularly in smaller or more isolated states, the analysis signals where future modeling efforts must incorporate policy constraints, infrastructure limits, and proximity-based effects to improve reliability.

This study shows that high aggregate accuracy can obscure important variance at the country level, while clarifying which national characteristics consistently shape refugee retention. In displacement forecasting, the central task is to identify the durable drivers of presence, those that persist into secondary migration trends. It offers a framework for policy-aware modeling that aligns empirical insight with operational relevance. By identifying the structural and economic factors most consistently tied to long-term refugee presence, this model offers a framework adaptable to similar crises where initial displacement has stabilized. Future applications should test the framework in regions with greater safety variance and data scarcity.

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.

MacLeod, Daniel. 2025. “DataAnalyticsUkraineCapstone.” https://github.com/DanMacCode/DataAnalyticsUkraineCapstone.

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.

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