Capstone Paper Template

Collin Paschall

This file is a template for your capstone project. You should not need to change any settings in the YAML front matter or any of the LaTex commands at the very top of the document (the ragged right, etc. etc.). The table of contents will be automatically generated based on the headers you specify in the document.

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: *"What causes certain countries to receive more refugees than others during crises, and how can these causal patterns be modeled to forecast future flows?"* 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

This is an example citation (Paschall 2023). Here's another (Bachrach and Baratz 2017). Here, there are two citations supporting this sentence. (Bachrach and Baratz 2017; Paschall 2023). Notice that the period follows the parentheses. If you want to put the citation in text it looks like Paschall (2023).

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) 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 diasporas5. 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 percent of the rise in applications lodged there, with a corresponding 7.9 percent 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 modelling 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. 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 12 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 responds directly to those gaps by proposing an ensemble modeling framework that integrates the five most consistently observed causal themes, accessibility, safety, familiarity, opportunity, and gravity, into a predictive system. Designed to be transferable across crises and usable by host nations for anticipatory planning, the model combines structural gravity logic and PPML with dynamic features modelled with XGBoost. The goal is to bridge the gap between explanatory insight and operational forecasting, contributing to both academic theory and practical policy application.

3 Data and Methods

Make sure all of your tables and figures are labeled, numbered, and have descriptive titles. Make sure they are aesthetically pleasing, with plain English variable names, legends, etc. Generate all of your tables and figures in the same Quarto document in which you write your paper, using code blocks. You should hide the code that generates the figures and tables - you only want the tables and figures themselves in the final PDF for submission. **NO RAW CODE**.

You can choose to either put your tables and figures in-line with your document, or you can include them at the end of your main text as an Appendix of Tables and Figures. It's preferable that you put them in-line, but sometimes this is a formatting pain because Quarto has very strong opinions about where figures go on the page. So, the path of least resistance (and what is consistent with how you'd prepare an article for publication) is to include placeholders in between paragraphs that say something like:

freedom house data is the freedom house data just for ukrian accross theyears.

eu capitols is the distance each capitol city is from kyiv. aggregated data 2 has vairous data accrosss various ecenomic and social criteria and its done annually. aggregated data 3 is similar but shows unemployment numbers and HICP (you know what that is right?) that one is done month over month. aggregated data is labelled and pretty straight forward: it is the flow of migrants into various eu countries by month and also existing diaspora in thsoe countries. we want to stress that to fulfill the hypothesis we need to address both flow of migration as well as the net migration.

[Table 1 about here]

How do aspects of political and freedom culture in Ukraine affect refugee migration to EU countries, given the aspects of those countries and the five main refugee drivers?

At the end of this document, you'll find a few resources for and examples of tables of and figures. Make tables and figures a reasonable size and use your space efficiently.

"I have finished cleaning and aligning my datasets for a refugee destination analysis project. I now have multiple datasets and and predictors from multiple sources (refugee counts, economic indicators, Freedom House data, etc.) . The datasets pulls variables from different original files,.

The overall goal is to train a machine learning model — possibly Random Forest or XGBoost — to predict refugee flows based on these predictors. I understand that even though the data came from multiple sources, the model function (like train() or xgboost()) requires a single, unified dataframe as input.

Please walk me through:

Finalizing the structure of the modeling dataframe

Preprocessing steps needed (handling NAs, scaling, etc.)

Example R code for training the model using this structure

Any model evaluation or interpretation steps you'd recommend

Assume we are building off the logic you previously explained about aligning multi-source predictors before modeling.

4 Results

Talk about the results of your analysis. Make sure you include tables and figures!

5 Conclusion

Summarize and wrap up.

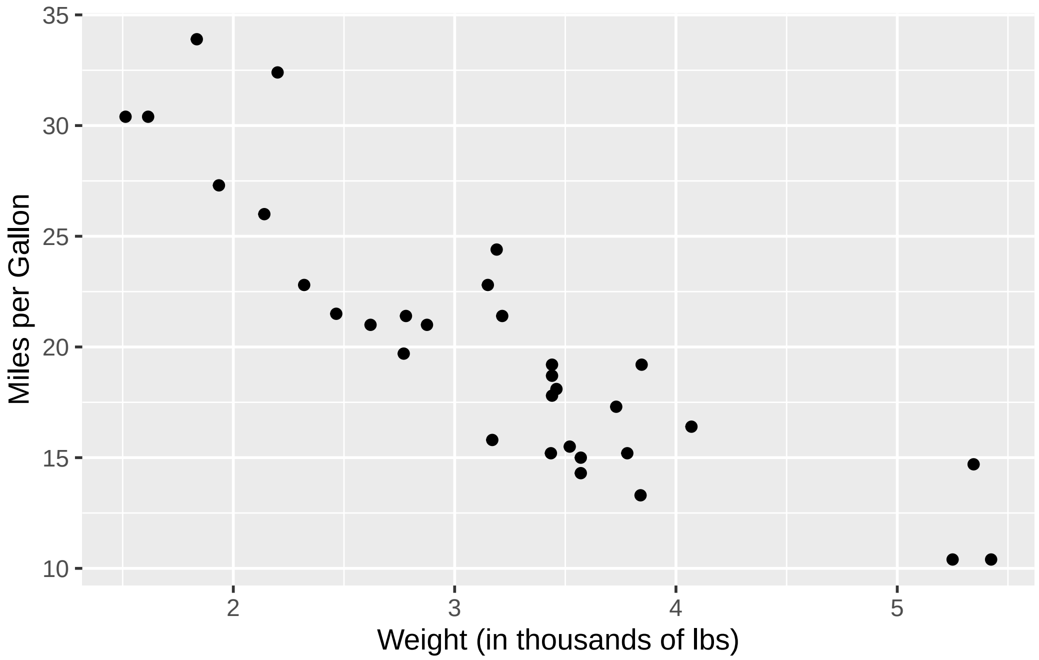
6 Appendix of Tables and Figures

For more details about how to work with figures in Quarto, check the Quarto documentation (or just follow this example). Make sure all your figures are appropriately scaled, easy to read/meaningful, and have axis labels with plain English names of variables

For tables, you can make tables manually in Quarto, but it's often better to use an R package that generates nice looking tables automatically. There are lots of options for this, but a good choice is `kable` in conjunction with `kableExtra`. For PDF documents that are rendered using LaTex, here is the specific documentation.

There are similarly lots of options for making nice-looking regression output. A good choice is `jtools`. You will want to install the packages `kableExtra` and `huxtable` along with `jtools`. Stargazer is another common choice.

Figure 1: Fuel Efficiency and Weight



Note: if you look at the R code, there is an option in the `kable_styling()` function - the "HOLD__position" value. This is important because it specifies that your table will go exactly where you want it in the document.

Table 1: Example Table

	mpg	cyl	disp	hp
Mazda RX4	21.0	6	160	110
Mazda RX4 Wag	21.0	6	160	110
Datsun 710	22.8	4	108	93
Hornet 4 Drive	21.4	6	258	110
Hornet Sportabout	18.7	8	360	175
Valiant	18.1	6	225	105

Table 2: Regression Example

	Model 1
(Intercept)	-39.96 ***
	(5.92)
imdb_rating	12.80 ***
	(0.49)
log(us_gross)	0.47
	(0.31)
genre5Comedy	6.32 ***
	(1.06)
genre5Drama	7.66 ***
	(1.08)
genre5Horror/Thriller	-0.73
	(1.51)
genre5Other	5.86
	(3.25)
N	831
R2	0.55

*** p < 0.001; ** p < 0.01; * p < 0.05.

7 References

[Note that references will be included here at the end of the document automatically if you use a .bib file and RStudio's citation tools.]

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.

Bachrach, Peter, and Morton S Baratz. 2017. "Two Faces of Power." In *Paradigms of Political Power*, Routledge, 118–31.

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.

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.

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.

Paschall, Collin. 2023. *My Awesome Book*. Johns Hopkins Press.

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.”