DS 3001 Final Project Paper

Henry Allen [*1] **Kyra Lim** [*1] **Emma Wunderly** [*1]

Abstract

Our Project explores refugee migration rates and how we can predict such rates based on origin country and asylum country freedom indexes.

1. Data

The data in our project consists of refugee migration statistics from United Nations High Commission for Refugees and Freedom Indexes from the Freedom House. With this Data we plan to investigate refugee rates from countries around the world, and eventually create a model that predicts migration rates based on the freedom indexes of ones home and host country.

The Freedom in the World dataset, published annually by Freedom House, provides scores and ratings from 2013 to 2021 for countries and territories. The data include measures such as political rights ratings, civil liberties ratings, and aggregate scores for subcategories like electoral process, political pluralism and participation, rule of law, and freedom of expression and belief. Each country-year entry also includes an overall status of Free, Partly Free, or Not Free. These variables allow us to capture the political and civil environment in both origin and asylum countries.

The UNHCR Persons of Concern dataset contains annual records beginning in 2004, organized by year, country of origin, and country of asylum. For each country pair, the dataset reports counts of various displaced groups, including refugees, asylum-seekers, internally displaced persons (IDPs), stateless persons, and others of concern. These values reflect the scale of displacement across borders and within states. These values provide the dependent variable in our analysis, which is the amount of refugee movement from one country to another in a given year.

[*]Equal contribution [1]School of Data Science, University of Virginia, Charlottesville, VA, United States of America. Correspondence to: Henry Allen <jpg7hy@virginia.edu>, Kyra Lim <smy2qe@virginia.edu>, Emma Wunderly <tyw3nq@virginia.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

The unit of analysis in our study is the country-year pair, which links refugee flows from an origin country to an asylum country in a given year with their corresponding freedom index values. This allows us to test hypotheses about how political repression, democratic governance, and civil liberties shape refugee migration. While the UNHCR data captures migration outcomes, the Freedom House data provides measures of the political conditions that may act as push and pull factors.

These datasets do consist limitations, such that refugee statistics often underrepresented the quantities of undocumented displacement, and the Freedom House scores, while widely used, are based partly on subjective assessments. Despite this, their examination provides a unique opportunity to connect quantitative migration patterns with qualitative assessments of political freedom.

1.1. Data Cleaning

The data cleaning process for the refugee and freedom index data involved several steps to ensure the data was ready for analysis. For the persons-of-concern.csv file, which contains the data regarding refugee migration, the initial loading and inspection revealed no immediate issues with missing values or data types, so no further cleaning was required. For the All-data-FIW-2013-2021.xlsx file, which contains the data regarding freedom indexes, initial attempts to load the data resulted in incorrect headers due to metadata at the beginning of the sheet. This was corrected by specifying the correct sheet name ('FIW13-21') and header row (row 0) during loading. The columns were then renamed for clarity, specifically 'C/T?' to 'Country Type', 'Country/Territory' to 'Country', and 'Total' to 'Freedom Total Score'. Finally, the first row, which contained descriptive information rather than data. We removed this so data handling could be smoother. Then relevant columns were selected and converted to appropriate numeric types. Finally we checked for null values and confirmed that there were no missing values in the selected columns after these steps.

1.2. Data Dictionary

This table provides a data dictionary for the Freedom in the World (FIW) dataset, defining the various columns and their meanings.

Table 1. Data Dictionary

Column Name	Description
Year	The year of the data record.
Country of Asylum	The name of the country where a person is seeking or has received protection.
Country of Origin	The name of the country from which a person has been displaced.
Country of Asylum ISO	The ISO 3166-1 alpha-3 code for the country of asylum.
Country of Origin ISO	The ISO 3166-1 alpha-3 code for the country of origin.
Refugees	The total number of refugees.
Asylum-seekers	The total number of asylum-seekers.
IDPs	The total number of Internally Displaced Persons.
Other people in need of international protection	The total number of people in need of international protection who do not fall into the other categories.
Stateless persons	The total number of stateless persons.
Host community	The total number of individuals in the host community.
Others of concern	The total number of individuals who are not included in the main categories but are of concern to humanitarian organizations.
Country/Territory	The name of the country or territory.
Region	The geographic region of the country or territory.
C/T	Indicates if the entry is a country (c) or territory (t).
Edition	The year of the report.
Status	The overall freedom status: F (Free), PF (Partly Free), or NF (Not Free).
PR rating	The Political Rights Rating, on a scale of 1 (least free) to 7 (most free).
CL rating	The Civil Liberties Rating, on a scale of 1 (least free) to 7 (most free).
PR	The aggregate score for the Political Rights category, based on sub-scores from A, B, and C.
A1 to A3	Sub-scores for the A. Electoral Process subcategory.
A	The aggregate score for the A. Electoral Process subcategory.
B1 to B4	Sub-scores for the B. Political Pluralism and Participation subcategory.
B	The aggregate score for the B. Political Pluralism and Participation subcategory.
C1 to C3	Sub-scores for the C. Functioning of Government subcategory.
C	The aggregate score for the C. Functioning of Government subcategory.
CL	The aggregate score for the Civil Liberties category, based on sub-scores from D, E, F, and G.
D1 to D4	Sub-scores for the D. Freedom of Expression and Belief subcategory.

2. Methods and Results

3. Conclusion

Impact Statement

References

Freedom House. Freedom in the world. Technical report, Freedom House, Washington, DC, 2025. URL https://freedomhouse.org/report/freedom-world.

UNHCR. Refugee statistics – operational data portal. Technical report, United Nations High Commissioner for Refugees, Geneva, Switzerland, 2025. URL https://www.unhcr.org.

World Bank. World development indicators. Technical report, The World Bank, Washington, DC, 2025. URL https://databank.worldbank.org/source/world-development-indicators.

A. Appendix

You can have as much text here as you want. The main body must be at most 8 pages long. For the final version, one more page can be added. If you want, you can use an appendix like this one.

The `\onecolumn` command above can be kept in place if you prefer a one-column appendix, or can be removed if you prefer a two-column appendix. Apart from this possible change, the style (font size, spacing, margins, page numbering, etc.) should be kept the same as the main body.