Migration Flows: An Overall Look

Matteo Poirè
Institutional ID: 529713
Email:
matteo.poire01@universitadipavia.it

***Abstract*— This report aims to achieve two main objectives: first, to provide a comprehensive overview of the global migration flow network by analyzing inter-country migration flows across a broad set of nations from 1990 to 2020 in 5-year intervals; and second, to explore potential dependencies between migration flows and economic, geographical, linguistic, and historical factors. To address the first objective, centrality measures were computed from the migration network's adjacency matrix to identify key hubs, authorities, most relevant countries and bridges. Correlations between centrality measures and economic indicators were analyzed, and community structures within the network were detected.
For the second objective, various regression models were developed, ranging from simple linear models to more complex quadratic and gravity models, to assess the relationships between economic indicators and both centrality measures and migration flows. The network analysis identified countries playing significant roles in global migration dynamics and demonstrated correlations between centrality measures and certain economic indicators. However, the regression analysis revealed a lack of significant linear dependence between economic indicators and either migration flows or centrality measures. A fixed effects model highlighted the influence of other factors as primary drivers of migration flows, and the gravity model further emphasized the necessity of other terms to better explain the migration flow patterns.

Introduction

Migration is a fundamental aspect of human history, shaping societies and economies worldwide. Understanding the dynamics of migration flows is crucial for policymakers, researchers, and international organizations as they address challenges related to population distribution, labor markets, and social integration. However, migration is a complex phenomenon influenced by a variety of factors, including economic conditions, geographical proximity, linguistic ties, and historical relationships.

In recent decades, globalization has intensified migration flows, making it increasingly important to analyze the underlying structures and dynamics of these flows. Network science offers a powerful framework to model and study migration as a complex, interconnected system, where countries act as nodes and migration flows as directed, weighted edges.

This report investigates the global migration flow network from 1990 to 2020, using a network science approach. The analysis encompasses a wide range of countries and explores how migration patterns evolve over time. By leveraging centrality measures, the report identifies key players in the migration network, highlighting countries that act as hubs, authorities, and bridges. Additionally, it examines the network community structure to uncover clusters of closely connected countries.

A secondary focus of this study is to explore potential dependencies between migration flows and economic indicators, such as GDP per capita, Human Development Index (HDI), Gini coefficient and unemployment rates. Using regression models the report seeks to determine whether these indicators can explain migration dynamics. The study also considers geographical, linguistic, and historical factors, such as shared borders and colonial ties, to provide a comprehensive analysis.

Despite the intuitive link between economic conditions and migration, the findings suggest that migration flows cannot be fully explained by linear relationships with economic indicators. A fixed effects model further reveals the existence of other, more complex drivers behind migration patterns. This underscores the multifaceted nature of migration and the need for interdisciplinary approaches to better understand its underlying causes.

The remainder of this report is structured as follows:

- Section 2 describes the datasets and their preparation.
- Section 3 outlines the methodologies applied, including network and regression analysis.
- Section 4 presents the results and a discussion of the findings
- Section 5 concludes with key takeaways and directions for future research

Data

This project required multiple datasets, primarly sourced from Our World In Data website and supplemented with data from the CEPII institute. Three main datasets were used:

Dataset A – Migration Flows

This dataset, retrieved from the Our World In Data Website (Fiona Spooner, 2022), represents the core dataset. It includes 1659 rows and 484 columns providing information about all the emigrants and immigrants of each one of the 237 countries from 1990 to 2020 with a step interval of 5 years. The origin country for immigrants and the destination country for emigrants was specified with integer values,

precisely positive values for immigrants and negative ones for emigrants. The following is the list of key attributes:

- *Country*
- *Year*
- *Immgrants_from_**
- *Emigrants_to_**

Dataset B – Economic Indicators

This dataset, represented by the so named 'economyT' table in the project, gathers different economic indicator for each country in the same time interval specified in the migration flow dataset. The dataset is the result of a join operation of a group of datasets retrieved from the Our World In Data website, each one specifying a different economic indicator.

The list of the indicators considered and the respective datasets is the following:

- *GDP per capita* (Bank, 2023): GDP per capita based on purchasing power parity (PPP). PPP GDP is gross domestic product converted to international dollars using purchasing power parity rates. An international dollar has the same purchasing power over GDP as the U.S. dollar has in the United States. Data are in constant 2017 international dollars.
- *Gini coefficient* (Data, 2024): the Gini coefficient is the most commonly used measure of inequality. It measures inequality on a scale from 0 to 1. Higher values indicate higher inequality. Depending on the country and year, the data relates to income measured after taxes and benefits, or to consumption, per capita.
- *Human Development Index* (UNDP, 2024): the Human Development Index (HDI) is a summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable and having a decent standard of living. The HDI is the geometric mean of normalized indices for each of the three dimensions.
- *Unemployment Rate* (Bank, Unemployment Rate, 2024): unemployment refers to the share of the labor force that is without work but available for and seeking employment.

Other than the economic indicators, such dataset was characterized by 3 additional information for each country, i.e. the population, the country code and the continent.

Dataset C – GeoDist Dataset

The dataset was downloaded from the CEPII website (Mayer, 2011), a French institute for research in international economics. The main aim of this dyadic dataset (dataset including variables valid for pairs of countries) was that of providing useful variables for building a significative gravity model. The dataset comprises the following columns:

- *Distance*: Geodesic distances are calculated following the great circle formula, which uses latitudes and longitudes of the most important cities/agglomerations (in terms of population)
- *Contig*: it indicates whether the two countries are contiguous
- Smctry: whether two countries are/were the same
- *Language*: it indicates whether two countries share a common language
- *Colony*: have ever had a colonial link

Precisely, the variables 'Language' and 'Colony' encodes all the information carried by columns referring to language and colonial relationships. In detail, 'Language' encodes information about official and secondary languages using the following formula:

$$Language = 0.8 * officialLang + 0.2 * ethnoLang$$

Where:

- ethnoLang: secondary languages

'Colony', instead, encodes information about colonial relationships in the following way:

$Colony$ = 0.6 * current_colony + 0.2 * colonizer_after45 + 0.15 * col_relation_after45 + 0.05 * colonial_link

Where:

- current_colony: countries currently in a colonial relationship
- colonizer_after45: colonizer after 1945
- col_relation_after45: countries being colony after 1945
- relationship after 1945
- colonial_link: being ever in a colonial relationship

Data Preparation

In order to make information in each dataset consistent, it was performed some common and specific data pre-processing step:

- *Data Cleaning*

First of all, datasets were processed removing from each one of them all the rows containing missing values.

- *Filtering Rows*

Dataset rows were filtered to match the same years in the Migration Flow dataset. Afterwards, it was derived a list of common countries present in each dataset, representing the countries for which all the information (migratortion flows, economic indicators and distance values) was available, and each dataset was filtered against this list. The result was a group of datasets providing information about the same set of countries with no missing value.

- *Fill of missing Continent values*

The GDP per capita dataset carried information about country continents too, however some country entry was missing. A simple extraction of the available Continent entries and a matching with the respective countries solved the problem filling all the missing entries

- *Code, Total_Immigrants, Total_Immigrants columns integration*

Given the presence of a 3-characters country code in the economic indicators dataset, also the Migration Flow and

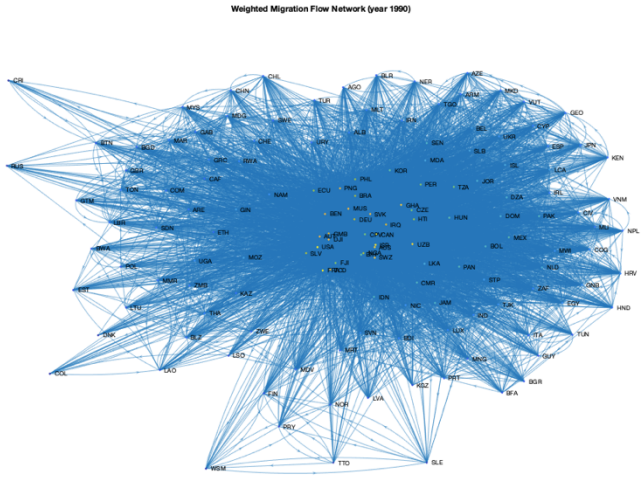
Geodist datasets were integrated with such column. Furthermore, to make simpler the information retrieval of the total number of immigrants and emigrants, it was added 2 columns (Total_Immigrants, Total_Emigrants) to the Migration Flow dataset obtained as the simple sum of immigrants and emigrants respectively for each country by year.

Methodology

This section outlines the methodologies employed to analyze the global migration flow network and explore the relationship between migration dynamics and economic, geographical, linguistic, and historical factors. The analysis was conducted in two stages: network analysis and regression modeling.

Network Analysis

The migration flow network was represented as a 3D weighted, directed adjacency matrix (third dimension represents the adjacency matrix of each year), obtained from the Migration Flow dataset, where nodes represent countries, and edges represent migration flows between them.



The analysis focused on extracting centrality measures and detecting community structures over time.

Centrality Measures

To identify the most influential countries in the migration network, other than the simple in- and out-degree/strength network features, the following centrality measures were computed for each 5-year interval from 1990 to 2020:

- *Betweenness Centrality*: measures the extent to which a country acts as a bridge in the shortest migration paths between other countries.
- *In-Closeness and Out-Closeness Centrality*: quantifies the proximity of a country to others based on inward and outward migration flows, respectively.
- *Eigenvector Centrality*: assesses a country's influence based on the influence of its neighbors.
- *PageRank*: a variant of the eigenvector centrality, it evaluates the importance of a country considering both direct and indirect migration flows, favoring countries connected to fewer, but influent, nodes.
- *Hubs and Authorities*: Identifies countries that are sources (hubs) and destinations (authorities) of migration.
- *Assortativity*: a correlation coefficient between the strengths of all nodes on two opposite ends of a link. A positive assortativity coefficient indicates that nodes tend to link to other nodes with the same or similar strength.

Community Detection

To measure the tendency of nodes in a network to form tightly-knit groups or clusters, it was measured local clustering coefficients:

- *In-Clustering*:

$$C_{in} = \frac{\sum_{j,k}\left(w_{ij}w_{jk}w_{ki}\right)^{1/3}}{k_{in}(i) \cdot \left(k_{in}(i)-1\right)}$$

where w_{ij} is the weight of the directed edge from i to j, and $k_{in}(i)$ is the in-degree of node i

- *Out-Clustering*:

$$C_{out} = \frac{\sum_{j,k}\left(w_{ji}w_{ik}w_{kj}\right)^{1/3}}{k_{out}(i) \cdot \left(k_{out}(i)-1\right)}$$

where $k_{out}(i)$ is the out-degree of node i

While to uncover clusters of closely connected countries, the network was transformed into an undirected adjacency matrix and community detection algorithms were applied. Precisely, it was applied the *Louvain community detection algorithm*, which detects node communities trying to maximize the modularity value.

Regression Modeling

The second stage of the analysis aimed to explore the relationship between migration flows, centrality measures, and economic indicators. Several regression models were developed and evaluated:

- *Linear Regression*

Simple linear regression models were constructed to assess the relationship between centrality measures, migration flows and individual economic indicators in each specific year:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \epsilon$$

Where Y represents the centrality measure or migration flow, β_0 represents the intercept, X_i an economic indicator and β_i its estimated coefficient

- *Step-wise Regression*

To capture which model better represents the relationships between economic indicators and centrality measures or migration flows, a step-wise regression was employed. With such approach it was

possible to find the best fitting regression model, even non-liner, and to understand which economic indicator played a key role to define the relation and which other could be ignored according to the *p*-value for an *F*-test of the change in the sum of squared error that results from adding or removing a term. Precisely, the tested models were:

- *Constant*

$$Y = \beta_0 + \epsilon$$

- *Linear* (the same at the previous point)
- *Interactions*: a model containing an intercept, a linear term for each predictor, and all products of pairs of distinct predictors

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \theta_1 X_1 X_2 + \cdots + \theta_n X_n X_{n-1} + \epsilon$$

- *Purequadratic*: a model containing an intercept term, a linear and squared terms for each predictor

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \theta_1 {X_1}^2 + \cdots + \theta_n {X_n}^2 + \epsilon$$

- *Quadratic*: a model contains an intercept term, linear and squared terms for each predictor, and all products of pairs of distinct predictors

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \theta_1 X_1 X_2 + \cdots + \theta_n X_n X_{n-1} + \zeta_1 {X_1}^2 + \cdots + \zeta_n {X_n}^2 + \epsilon$$

- *Polynomial*: a model with all terms up to a specified degree for each predictor. Example with two predictors, the first one up to degree *i* and the second up to degree *j*:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \theta_{1,1} X_1 X_2 + \cdots + \theta_{i,j} {X_1}^i {X_2}^j + \zeta_{1,1} {X_1}^2 + \cdots + \zeta_{1,i} {X_1}^i + \zeta_{2,1} {X_2}^2 + \cdots + \zeta_{2,j} {X_2}^j + \epsilon$$

- *Fixed Effects Model*

  To control for unobserved heterogeneity across countries and years, a fixed effects model was applied:

$$F_{i,t} = \alpha_i + \beta_{i,t} X_{i,t} + \mu_t + \epsilon_{i,t}$$

  where:

  - $F_{i,t}$ represents the migration flow of country *i* at time *t*
  - α_i represents country-specific effects
  - μ_t represents time-specific effects

- *Gravity Model*

  A gravity model, commonly used in migration studies, was constructed to evaluate the influence of economic, geographical, linguistic and historical factors on migration flows:

$$F_{i,j} = \left[\frac{(GDP_i \cdot GDP_j)^{\alpha} \cdot (HDI_i \cdot HDI_j)^{\gamma}}{D^{\beta}} + \frac{(1-Gini_i)^{\delta} \cdot (1-Pop_i)^{\lambda} \cdot (1-Unem_i)^{\lambda}}{D^{\beta}} \right] \cdot e^{\theta \cdot Lang + \phi \cdot Col + \psi \cdot Cont + \zeta Hist}$$

  where:

  - $F_{i,j}$ represents the migration flow from country *i* to country *j*
  - GDP_i represents the GDP per capita of country *i*
  - HDI_i represents the Human Development Index of country *i*
  - $Gini_i$ represents the Gini coefficient relative to country *i*
  - Pop_i represents the population of country *i*
  - $Unem_i$ represents the Unemployment Rate of country *i*
  - D represents the distance between country *i* and *j*
  - $Lang$ represents the language relationship between country *i* and *j*
  - Col represents the colonial relationship between country *i* and *j*
  - $Cont$ represents if country *i* and *j* are neighbors countries
  - $Hist$ represents whether countries *i* and *j* are/were the same

Model Evaluation

Each model was evaluated based on standard performance metrics, including:

- *R-squared (R^2) and Adjusted R^2:* measures the proportion of variance explained by the model
- *Mean Absolute Error (MAE):* assesses the average magnitude of errors
- *Significance of Coefficients:* determines the statistical significance of the predictors

Analysis Description and Results

This section presents the results of the network analysis and regression modeling, followed by a discussion of their implications in understanding global migration flows and their relationship with economic, geographical, linguistic and historical factors.

Given that the analysis was conducted in two stages, as stated before, also results and discussions will be presented in two different sub-sections.

Network Analysis Results

The network analysis provided insights into the structure and dynamics of global migration flows over time.

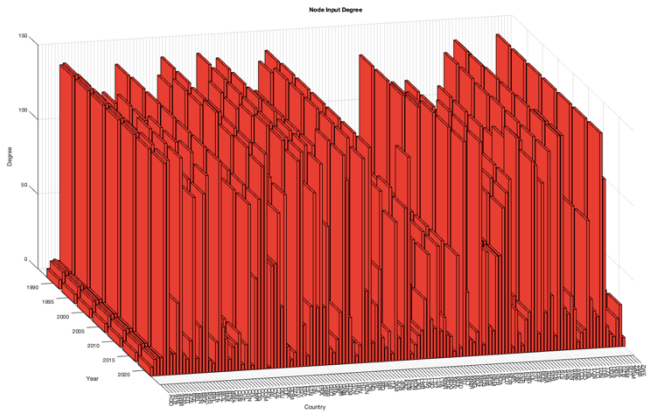
Such insights derived from some basic network characteristic like node degrees and strenghts, but also from the computation of some specific centrality measure and modularity.

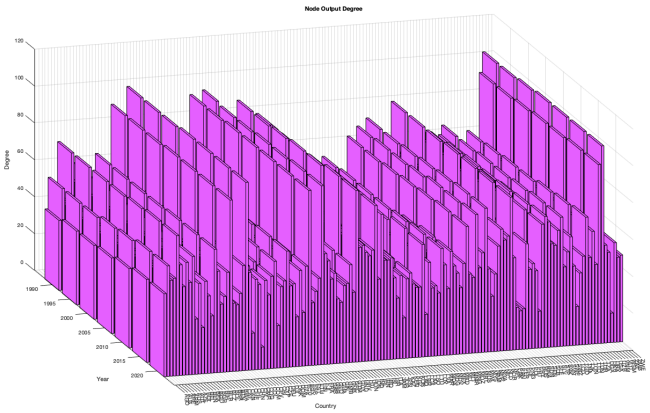
Node Degree and Strenght

Node degree and node strength can be easily derived from the 3D adjacency matrix representing the network. They help to identify the number of direct connections a country has with other countries and the magnitude of such connections respectively.

Since the target network is a directed one, it was possible to to retrieve some more meaningful information of the generic total degree and total strength considering the edges directions and computing in- and out- characteristics.

Starting from the in- and out-degree measures, these two measures allow to identify the number of countries sending migrants to a specific country (red histogram) and number of countries receiving migrants from a specific country (magenta histogram) respectively.





The histograms reveal a well-connected network whose connections are quite stable across years, indeed country in- and out-degree distributions change very rarely across years. The same conclusion can be derived listing the top 5 countries having the largest number of incoming and outgoing connections and observing how the rankings are dominated by the same few countries in both situations:

Top 5 countries having more incoming edges:

	Rank1	Rank2	Rank3	Rank4	Rank5
1990	"France"	"United Kingdom"	"Austria"	"Denmark"	"Italy"
1995	"France"	"Italy"	"United Kingdom"	"Austria"	"Denmark"
2000	"France"	"Italy"	"United Kingdom"	"Austria"	"Sweden"
2005	"France"	"Italy"	"United Kingdom"	"Austria"	"Czechia"
2010	"Belgium"	"France"	"Italy"	"Netherlands"	"Norway"
2015	"Austria"	"Belgium"	"France"	"Italy"	"Netherlands"
2020	"Belgium"	"Austria"	"Denmark"	"France"	"Italy"

Top 5 countries having more outgoing edges:

	Rank1	Rank2	Rank3	Rank4	Rank5
1990	"United States"	"China"	"Germany"	"France"	"United Kingdom"
1995	"United States"	"China"	"Germany"	"France"	"United Kingdom"
2000	"United States"	"China"	"France"	"Germany"	"United Kingdom"
2005	"United States"	"China"	"Germany"	"United Kingdom"	"France"
2010	"United States"	"China"	"Germany"	"United Kingdom"	"France"
2015	"United States"	"China"	"Germany"	"United Kingdom"	"France"
2020	"United States"	"China"	"Germany"	"United Kingdom"	"France"

However, it must not confuse the number of connection with the actual number of immigrants and emigrants received and sent by a country, a quantity which is detected by the node in- and out-strength (green and cyan histograms respectively). These two measures define how much each country contributes to the actual migration flow and reveal the importance of the edges of the previously listed countries.

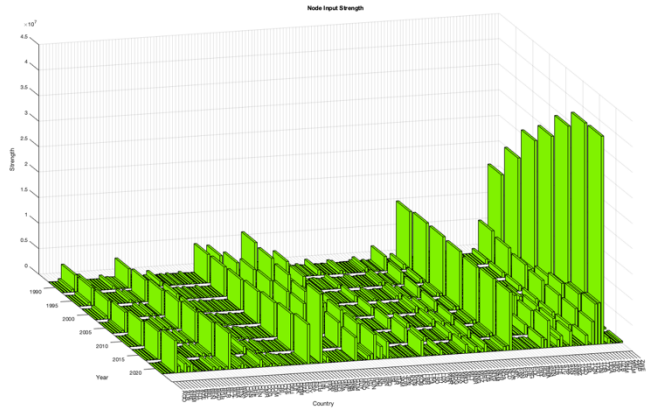
Top 5 most popular migration destinations:

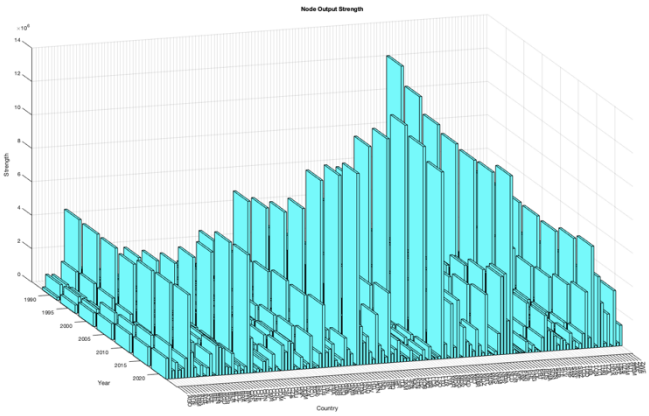
	Rank1	Rank2	Rank3	Rank4	Rank5
1990	"United States"	"Russia"	"India"	"Ukraine"	"France"
1995	"United States"	"Russia"	"India"	"Germany"	"France"
2000	"United States"	"Russia"	"Germany"	"India"	"France"
2005	"United States"	"Russia"	"Germany"	"France"	"India"
2010	"United States"	"Russia"	"Germany"	"France"	"United Arab Emirates"
2015	"United States"	"Russia"	"Germany"	"United Kingdom"	"United Arab Emirates"
2020	"United States"	"Germany"	"Russia"	"United Kingdom"	"United Arab Emirates"

Top 5 countries sources of immigrants:

	Rank1	Rank2	Rank3	Rank4	Rank5
1990	"Russia"	"Ukraine"	"India"	"Bangladesh"	"Mexico"
1995	"Russia"	"Mexico"	"Ukraine"	"India"	"Bangladesh"
2000	"Russia"	"Mexico"	"India"	"Ukraine"	"Bangladesh"
2005	"Mexico"	"Russia"	"India"	"Ukraine"	"China"
2010	"Mexico"	"Russia"	"India"	"China"	"Ukraine"
2015	"Mexico"	"India"	"Russia"	"China"	"Ukraine"
2020	"India"	"Mexico"	"Russia"	"China"	"Ukraine"

From the in- and out-strength ranking we notice as Austria, Belgium, Czechia, Denmark, Italy, Netherlands, Norway, Sweden, which appeared among the top 5 countries for incoming connections, are not among the top 5 popular destinations; and the same can be said about France, Germany, United Kingdom, United States, which appeared as the top 5 countries for outgoing connections, but do not appear among the top 5 countries sources of immigrants.





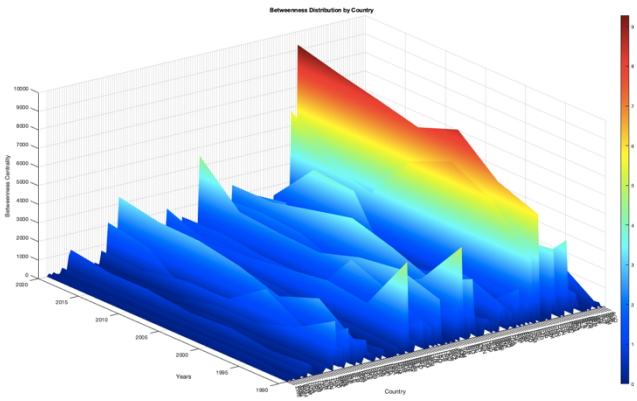
These results suggest the presence of few very relevant nodes and the consequent presence of few hubs and authorities, structure insights collected through centrality measures.

Centrality Measures

The network structure has been analysed in detail by calculating different centrality measures.

- Betweenness

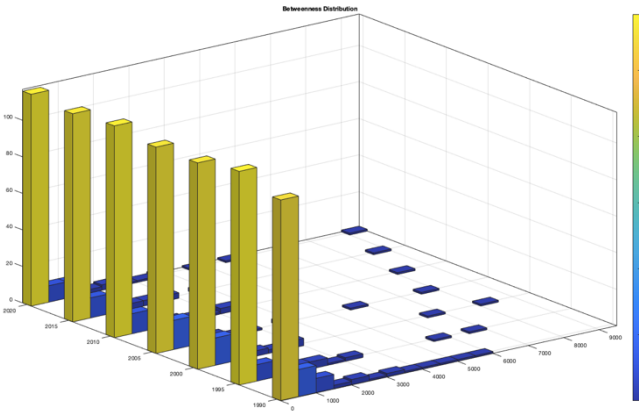
Betweenness helped quantifying how many times migrants must pass thorugh a country to reach their destination, if they had to follow the shortest possible path. From the calculus it emrged that the most busiest countries across years, used by migrants to reach their final destinations, are: Costa Rica, being always the first country across all period considered, Cyprus, being always the second in the ranking except for year 1990 and year 2010 where it positioned at the 3rd place as most popular country, Central African Republic, being always among the top 5 choices except for year 2000, Latvia, Iceland and Saint Lucia. While it resulted that countries like United States, United Arab Emirates, Uzbekistan and Zimabawe act more as destination or source countries for migrants.



A further analysis revealed also the absence of a significant change in network betweenness. Indeed, a t-test between the averages by country computed between periods 1990-2000 and 2000-2010, and still between periods 2000-2010 and 2010-2020 resulted in unsignificant changes (i.e. p-values > 0.05).

This means that the positive trend across years presented by many different countries (e.g. Austria, Belarus, Canada, Denmark, Malta) is not big enoguh to become crucial point of exchange for migrants.

This condition repeats very commonly also for almost all the centrality measures, whose present almost the same right skewed distribution pattern.

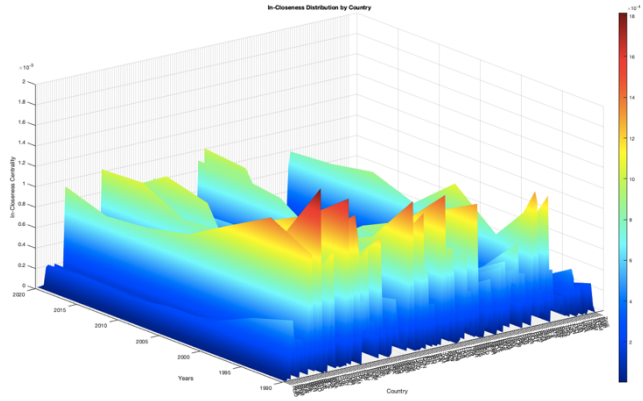


A situation which underlines what already suspected by the node and strenght degree analysis, i.e. a network which changes its shape very rarely and tends to be stable.

- In- and Out-Closeness

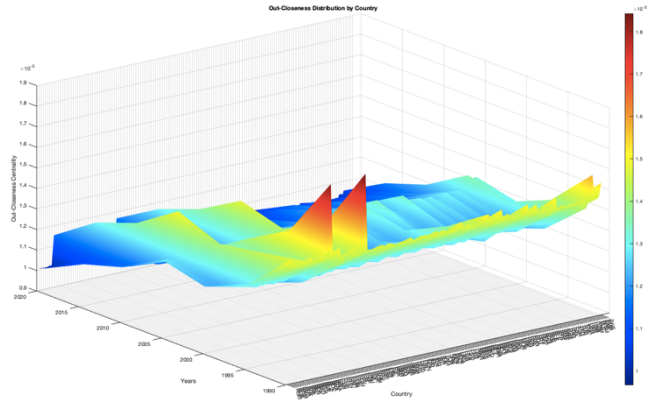
In and out-closeness measures help understanding how easily a country can receive, or send respectively, migrants from/to other countries, based on the shortest paths in the network. Consequently they measure how well connected a country is to receive migrants or how well positioned it can be to send them.

The analysis revealed that countries like Costa Rica, always in the first three ranks, Bulgaria, Cyprus, not appearing in the top 5 ranking just in year 2010 and Latvia are attractive places or common destinations for migrants, while countries like Djibouti, always being the first or second country, Uzbekistan, for all the time period and Bangladesh, with an increasing importance across the years, are predominant sources of migrants.



Also for in- and out-closeness the network results to be quite stable with no significant changes, except for Albania between periods 1990-2000 and 2000-2010 for both the measures and between periods 2000-2010 and 2010-2020 for out-closeness. In particular, showing a positive trend for in-closeness and a negative trend for out-closeness, so becoming always more important as receiving node.

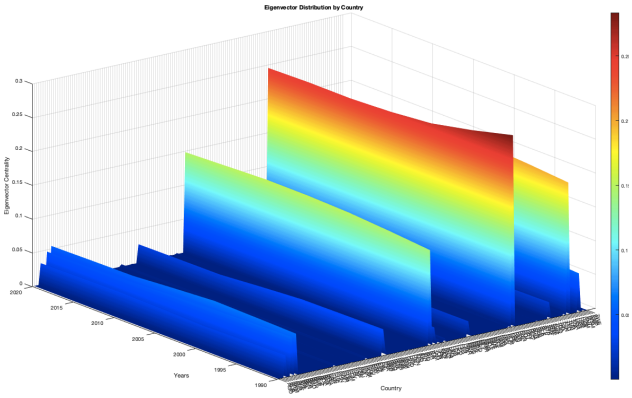
Another peculiarity, derived from the out-closeness measure, is that there is no country showing a positive trend in it. This demonstrates how the migrant flows are decreasing across the years.



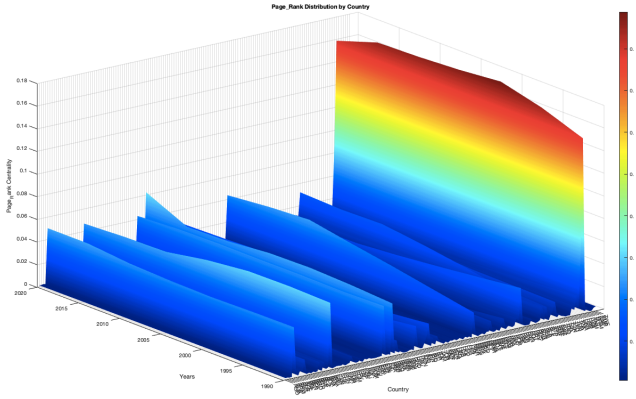
- Eigenvector and Page Rank

Both eigenvector centrality and page rank measures are used to evaluate the importance of nodes in a network. However page rank, differently from eigenvector centrality, takes also in consideration the number and type

of connections of a node penalizing nodes that link to many others and favoring those linked from fewer but important nodes.
So, according to the eigenevctor measure, it is possible to state that the most relevant countries are: Azerbaijan, Belarus, Kazakhstan, Russia, Ukraine, Uzbekistan. These countries dominate the top 5 ranking across all years, maintaing also the same relevance apart from 2000 where Uzbekistan changes its position in the ranking from the 5th to the 4th despite Belarus.



Nevertheless, if we analyse the results obtained from the execution of the page rank algorithm, we notice a completely different ranking dominated from the following countries: United States, always at the first position, Canada, with a negative trend across years which caused the drop from the 2nd position in 1990 to the 5th position in 2020, United Kingdom, showing a solid 2nd place, and Germany, varying its positioning from the 4th to the 5th and finally to the 3rd place.

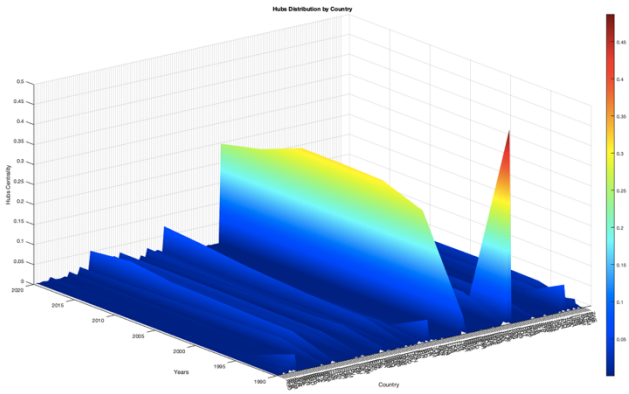


Despite the presence of positive and negative trends by many countries for what it concerns eigenvector centrality and page rank, there were not still significant changes comparing the situation in the 3 different decades 1990-2000, 2000-2010 and 2010-2020.

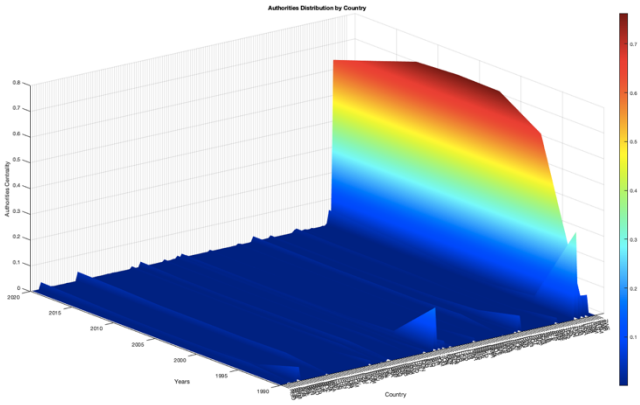
- Hubs and Authorities

Finally, it was computed hubs and authorities scores to have an overall idea of what countries played a role of origins and destinations respectively for migrants.

From this analysis it emerged that the countries acting as most common sources of migrants from 1995 onwards are: India, Philippines and China, these last three countries share the 2nd, 3rd and 4th positions except in 1990, Mexico, conquering the first position in the ranking, and Vietnam, showing its presence as the most common 5th country source of migrants. In 1990 the ranking is represented by: Russia, Ukraine, Kazakhstan, Mexico and Belarus.



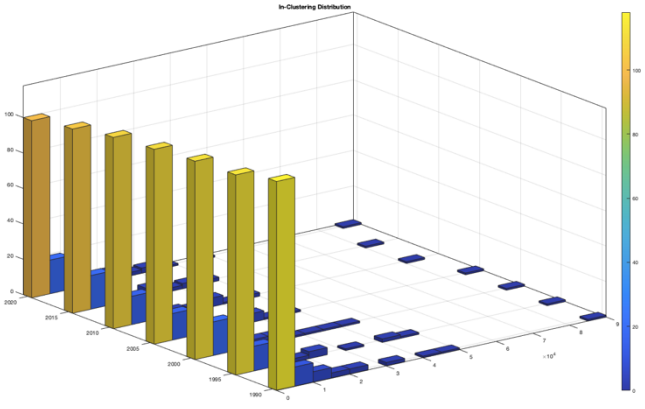
While the countries acting as the most common destinations are: United States, being the most popular from 1995 onwards, Canada, showing in 2nd and 3rd placement, United Arab Emirates, becoming particularly relevant from 2010, but also Russia and Ukraine in 1990-1995, Pakistan in 2000-2010 and Australia in 2015-2020.

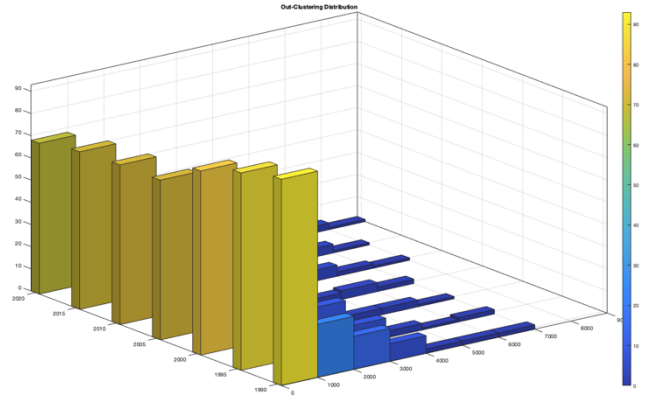


Community Detection

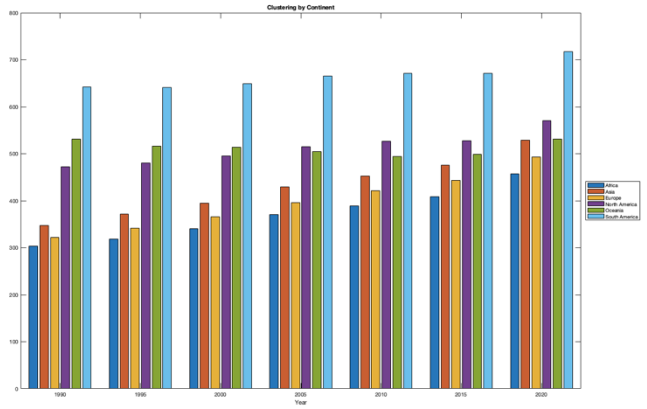
Since from the first observations the network resulted to be well-connected, it decided to delve more this characteristic focusing on the presence of communities.

Nevertheless, already from the first measures of country in- and out-clutering coefficients, the network looked not to be characterized by tightly-knit groups and the distributions clearly show this feature.

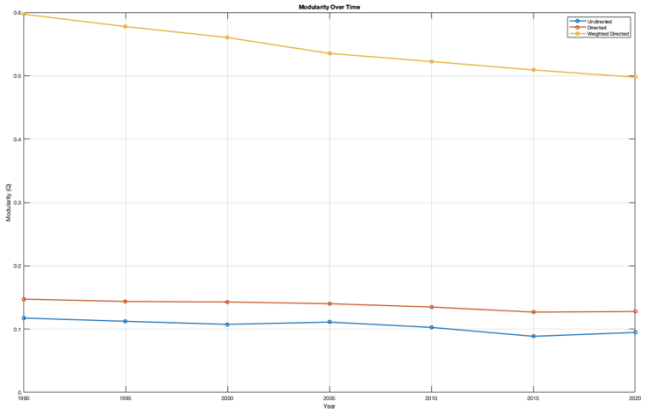




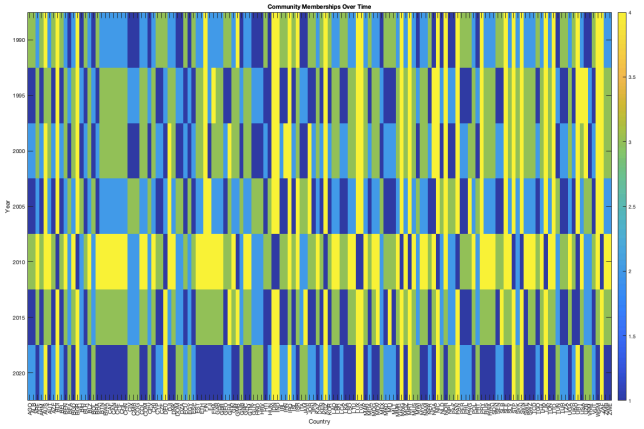
It was also possible to broadening the view and look at the level of country interconnections from a continental point view. This was achieved calculating an overall clustering coefficient by country and computing the mean of such values by continent across years. The result shows three possible scenarios: a strong interconnection among countries in South America, a similar but less strong interconnection among countries in North America and Oceania and a even weaker interconnection for countries in Europe, Asia and Africa. However, the difference between the last two groups looks to evanish across the years, arriving in 2020 with the only south american continent playing as outlier.

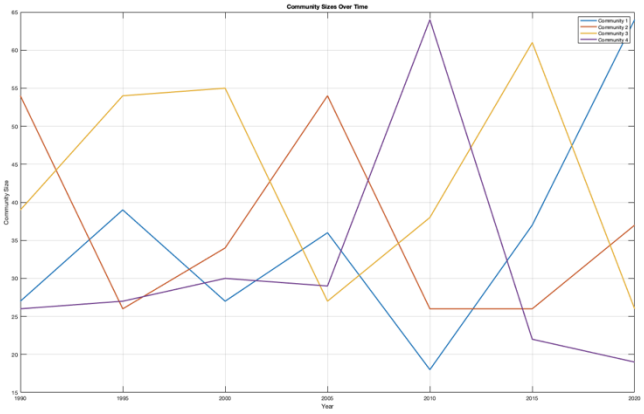


Despite the discouraging results obtained from the clustering coefficient calculi, a final research about the existence of communities was done computing the network modularity. This measure was calculated using the Louvain community detection algorithm considering the network as undirected, directed and weighted.



The constant small modularity value, when the network is considered both as undirected and directed, and the its negative trend, when it takes in consideration weigths of network links, suggests a considerable difficulty to divide the network into communities. Indeed a further analysis revealed as communities are completely unstable, their sizes change sharply and the country memberships change almost every year.





Regression Results

The regression analysis aimed to quantify the relationship between migration flows, centrality measures, and economic indicators. The analysis can be divided into three sections: linear and non-linear regression, fixed effects model and gravity model.

Linear and non-linear regression

To study if economy factors can well describe the existance of a pattern in migration flows, it was fitted a linear regression model with some key economic indicators: GDP per capita, Gini coefficient, HDI, Unemployment Rate and Population. A model was built per year, so that it was possible to better detect how the evolution centrality measures and the total amount of incoming and outgoing migrants could depend from economic factors.

The analysis shows how it is impossible to define the existence of a strong relationship between country economic indicators and their centrality measures. Indeed, the centrality measure evolutions look to be explained better by other events than the simple economic indicators, thus suggesting the need of additional variables to the model.

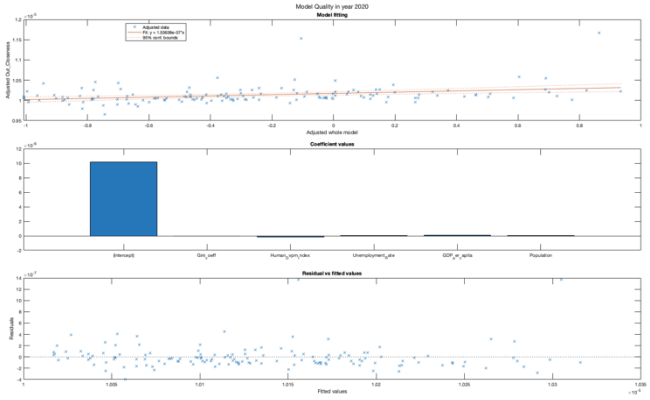
The coefficients estimates are not significant, in fact p-values associated to the t-statistic, where for each coefficient

is tested the null hypothesis that the corresponding coefficient is zero against the alternative that it is different from zero given the other predictors in the model, confirm that there is no predictor impacting with 95% of confidence level.

Even the other statistics like Adjusted R-squared (Adjusted R^2) and Root Mean Square Error (RMSE) confirm the poor relationship: the first, with very small values, shows that the model is not able to explain more than the 4% (worst case) to 25% (best case) of the data variability, while the second shows a very large standard deviation of the error distribution.

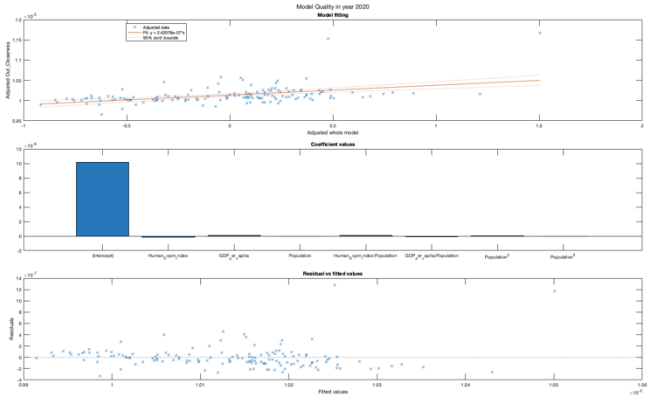
The best linear model according to RMSE is the model fitting Out-Closeness measure data in year 2020. As stated before, no economic indicator coefficient results significative from the point of view of t-statistic test. Most of the variability can be simply explained by a model characterized by the only Intercept.

```
Linear regression model:
    Out_Closeness ~ 1 + Gini_coeff + Human_Dvpm_Index + Unemployment_Rate + GDP_per_capita + Population

Estimated Coefficients:
                        Estimate        SE          tStat       pValue
                        ________    __________    _______    __________

    (Intercept)         1.0172e-05    2.305e-08     441.31    5.9008e-222
    Gini_coeff         -1.0594e-08    1.4973e-08   -0.70757        0.48039
    Human_Dvpm_Index   -1.361e-07     6.5258e-08    -2.0856        0.038827
    Unemployment_Rate   1.3391e-08    1.5283e-08     0.87619       0.38243
    GDP_per_capita      6.9097e-08    6.0638e-08      1.1395       0.25644
    Population          3.7561e-09    1.1676e-08      0.3217       0.74816


Number of observations: 146, Error degrees of freedom: 140
Root Mean Squared Error: 2.17e-07
R-squared: 0.102,  Adjusted R-Squared: 0.0702
F-statistic vs. constant model: 3.19, p-value = 0.00931
```



Even comparing the results of more complex regression models, which use different combinations of the predictors to better explain the variability of migration flows, the scenario doesn't change, although better results are obtained. In this case the best model built still refers to Out-Closeness centrality data of year 2020.

```
1. Adding Human_Dvpm_Index:Population, FStat = 5.2121, pValue = 0.023947
2. Adding Population^2, FStat = 4.9674, pValue = 0.027446
3. Adding GDP_per_capita:Population, FStat = 4.4033, pValue = 0.037705
4. Adding Population^3, FStat = 6.081, pValue = 0.014909
5. Removing Unemployment_Rate, FStat = 0.84159, pValue = 0.36057
6. Removing Gini_coeff, FStat = 2.0647, pValue = 0.15303
Linear regression model:
    Out_Closeness ~ 1 + Human_Dvpm_Index*Population + GDP_per_capita*Population + Population^2 + Population^3

Estimated Coefficients:
                                  Estimate        SE          tStat       pValue
                                  ________    __________    _______    __________

    (Intercept)                   1.0136e-05    2.57e-08     394.41    1.6423e-212
    Human_Dvpm_Index             -1.6191e-07    6.3924e-08    -2.5329        0.012432
    GDP_per_capita                8.9087e-08    5.9612e-08     1.4944        0.13735
    Population                   -6.2714e-10    1.9606e-08   -0.031988        0.97453
    Human_Dvpm_Index:Population   1.2322e-07    4.0258e-08     3.0608       0.0026538
    GDP_per_capita:Population    -9.4138e-08    3.7331e-08    -2.5217        0.012815
    Population^2                  1.9552e-08    5.5468e-09     3.5249       0.0005752
    Population^3                 -4.6576e-09    2.0873e-09    -2.2314        0.027266


Number of observations: 146, Error degrees of freedom: 138
Root Mean Squared Error: 2.05e-07
R-squared: 0.209,  Adjusted R-Squared: 0.169
F-statistic vs. constant model: 5.2, p-value = 2.76e-05
```

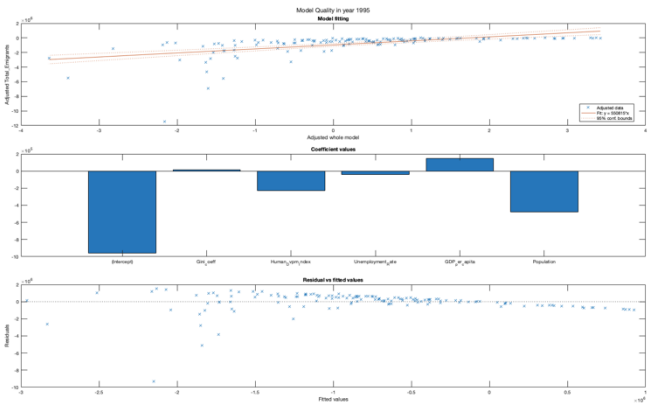


A similar scenario occurs when a linear model fits to total immigrants and emigrants data. Even in this case the economic indicators alone seem not to explain completely the evolution of the migration flows presenting worse results. Models across years are able to explain only between 18% and 30% percent of the data variability according to the Adjusted R^2 measures, characterized also by very large MSEs.

The main difference from centrality measures models is that, as expected, the indicator of the amount of population in a country seems to achieve some relevance in the understanding of migration flows patterns.

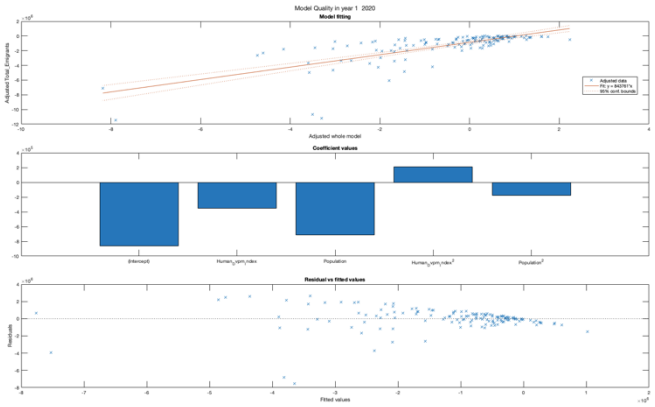
The model corresponding to the pictures below is the one trying fitting the total amount of emigrants from a country in year 1995.

```
Estimated Coefficients:
                        Estimate        SE          tStat       pValue
                        ________    __________    _______    __________

    (Intercept)        -9.5906e+05    1.1153e+05    -8.5987    1.4523e-14
    Gini_coeff             13858         60311      0.22978        0.8186
    Human_Dvpm_Index   -2.2901e+05    2.0318e+05    -1.1271       0.26162
    Unemployment_Rate      -43077         66867    -0.64422       0.52049
    GDP_per_capita      1.4515e+05    2.0456e+05     0.70959       0.47914
    Population         -4.7732e+05         63021    -7.5739    4.46e-12


Number of observations: 146, Error degrees of freedom: 140
Root Mean Squared Error: 1.19e+06
R-squared: 0.306,  Adjusted R-Squared: 0.281
F-statistic vs. constant model: 12.3, p-value = 6.36e-10
```



Differently from the previous case, the direct explanation of the amont of immigrants and emigrants through a more complex model shows a significant improvement underlined by the larger Adjusted R^2 measure equal to 0.52.

```
1. Adding Population^2, FStat = 37.3739, pValue = 9.29273e-09
2. Adding Human_Dvpm_Index^2, FStat = 8.0351, pValue = 0.0052777
3. Removing Unemployment_Rate, FStat = 0.19287, pValue = 0.66122
4. Removing GDP_per_capita, FStat = 0.42173, pValue = 0.51715
5. Removing Gini_coeff, FStat = 2.271, pValue = 0.13407
Linear regression model:
    Total_Emigrants ~ 1 + Human_Dvpm_Index + Population + Human_Dvpm_Index^2 + Population^2

Estimated Coefficients:
                          Estimate        SE        tStat       pValue
                          ________    _________    _______    _________

    (Intercept)           -8.6195e+05    1.8392e+05    -4.6867    6.4771e-06
    Human_Dvpm_Index      -3.5376e+05    1.0872e+05    -3.2538     0.0014253
    Population            -7.1556e+05         68663    -10.421    3.2866e-19
    Human_Dvpm_Index^2     2.0967e+05         84639     2.4773      0.014421
    Population^2          -1.7548e+05         31243    -5.6165    1.0029e-07


Number of observations: 146, Error degrees of freedom: 141
Root Mean Squared Error: 1.31e+06
R-squared: 0.539,  Adjusted R-Squared: 0.526
F-statistic vs. constant model: 41.3, p-value = 7.23e-23
```



Fixed effects model

Fixed effects allow to control for unobservable characteristics that are constant over time but may vary across observed units. In other words, fixed effects capture unobserved heterogeneity across units that might affect the relationship between the independent and dependent variables. In this project fixed effects are defined as:

- Individual fixed effects: they capture the impact of unobserved characteristics (e.g. culture, geographic location) that are constant over time for each country. This allows to correct for factors that might affect the dependent variable in a systematic way.
- Time fixed effects: they capture the impact of events or changes (e.g. global economic crises, changes in international migration policies) over time that affect all units equally.

Considering the contribution of economic indicators keeping fixed country-specific and global events provided more significant results with similar considerations.

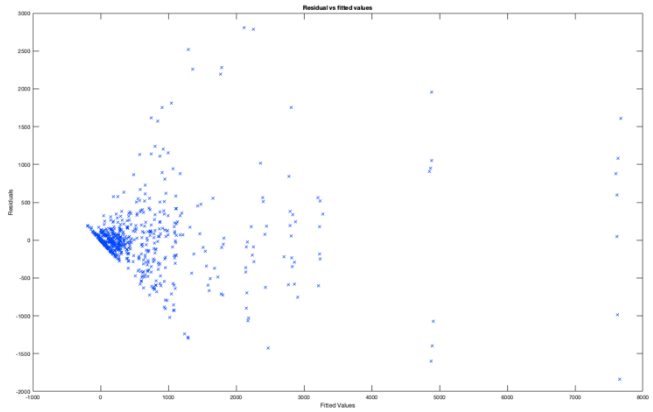
Focusing, for instance, to the fixed-effects model for betweenness centrality, which in the previous linear analysis represented the worst case scenario, a much larger data variability is explained and predictors like Gini coefficient, Unemployment Rate and Population become statistically significant. Precisely, we have that inequality and the total population have significant negative effects on betweenness centrality, while higher unemployment rates are associated with increased values of betweenness.

Observing, instead, the random effects covariance parameters we find that:

- The model finds substantial variability in the response due to country-level effects. These effects account for systematic differences in the centrality measure across countries, independent of other predictors
- The model finds that year-level effects have little to no impact on the response
- There is a significant amount of unexplained variability. This could indicate that there are other unmodeled factors influencing centrality, or that the model predictors do not fully capture the relationship

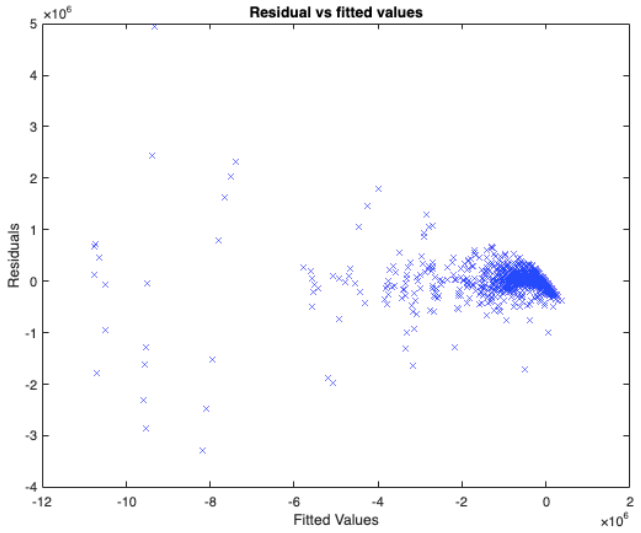
```
Model fit statistics:
    AIC      BIC      LogLikelihood    Deviance
    15687    15731    -7834.3          15669

Fixed effects coefficients (95% CIs):
    Name                  Estimate    SE        tStat      DF      pValue        Lower      Upper
    {'(Intercept)'     }     469.03    75.804     6.1874    1016    8.8537e-10     320.28     617.78
    {'Gini_coeff'      }    -43.455    15.953    -2.7239    1016     0.0065614    -74.76     -12.15
    {'Human_Dvpm_Index' }    -25.66     50.631    -0.50679   1016     0.61241     -125.01     73.694
    {'Unemployment_Rate'}     40.493    16.563     2.4448    1016     0.014663      7.9914    72.994
    {'GDP_per_capita'  }      66.696    60.614     1.1003    1016     0.27145     -52.247    185.64
    {'Population'      }    -102.76     44.46     -2.3113    1016     0.021016    -190       -15.516

Random effects covariance parameters (95% CIs):
Group: Country (146 Levels)
    Name1              Name2              Type          Estimate    Lower     Upper
    {'(Intercept)'}    {'(Intercept)'}    {'std'}       901.81      863.55    941.76

Group: Year (7 Levels)
    Name1              Name2              Type          Estimate      Lower    Upper
    {'(Intercept)'}    {'(Intercept)'}    {'std'}       3.0172e-06    NaN      NaN

Group: Error
    Name           Estimate    Lower     Upper
    {'Res Std'}    399.15      382.21    416.83
```

Conclusions which can also be derived from the fitted vs residual values plot.



Similarly, also the fixed-effects model applied to understand the behaviour of total incoming and outgoing migrants produces largely better results. Both the models can well explain the data variability, as the Adjsted R-squared measure suggests, but the main difference between the two is the role of the GDP per capita economic indicator. Indeed, this predictor shows relevance only in the model fitting the total amount of emigrants by country, suggesting that the average annual income of countries can play a key role in determing the outgoing flow, but being not decisive enough to predict the incoming one. The random effects covariance parameters, instead, reveal a general behaviour similar to that one already described for the fixed-effects model fitting betwennes centrality.

```
Formula:
    Total_Emigrants ~ 1 + Gini_coeff + Human_Dvpm_Index + Unemployment_Rate + GDP_per_capita + Population + (1 | Country) + (1 | Year)

Model fit statistics:
    AIC      BIC      LogLikelihood    Deviance
    30002    30046    -14992           29984

Fixed effects coefficients (95% CIs):
    Name                  Estimate      SE          tStat      DF      pValue        Lower         Upper
    {'(Intercept)'     }    -1.8049e+06    1.0911e+05    -9.2097    1016    1.8216e-19    -2.1219e+06    -1.9077e+05
    {'Gini_coeff'      }       -31791         17098     -1.8593    1016     0.063267      -65341         1760.2
    {'Human_Dvpm_Index' }       -14628         56476     -0.25901   1016     0.79568     -1.2545e+05      96195
    {'Unemployment_Rate'}       -2720.2        17778     -0.15301   1016     0.87842       -37606         32166
    {'GDP_per_capita'  }    -3.955e+05         67403     -5.8678    1016    5.9731e-09    -5.2777e+05    -2.6324e+05
    {'Population'      }    -5.2708e+05        60752     -8.676     1016    1.6043e-17    -6.463e+05     -4.0787e+05

Random effects covariance parameters (95% CIs):
Group: Country (146 Levels)
    Name1              Name2              Type          Estimate      Lower         Upper
    {'(Intercept)'}    {'(Intercept)'}    {'std'}       1.3071e+06    1.1608e+06    1.4719e+06

Group: Year (7 Levels)
    Name1              Name2              Type          Estimate    Lower         Upper
    {'(Intercept)'}    {'(Intercept)'}    {'std'}       6576.2      9.3415e-09    4.6296e+15

Group: Error
    Name           Estimate      Lower         Upper
    {'Res Std'}    4.1989e+05    4.0052e+05    4.402e+05
```



Gravity Model

A gravity model is a popular framework used to predict and analyze the interaction between two entities, two countries in this context. The model used in this analysis tries to highlight how the following factors may influence migration flows in both the origin and destination countries:

- *Economic Attractiveness*:
  - GDP and HDI: both often have symmetrical roles in migration. High GDP or HDI in a destination country (e.g., better economy, higher quality of life) makes it attractive, while low GDP or HDI in the origin country can push people to migrate. So, pairing GDP and HDI values captures this push-pull dynamic.
- *Origin-Specific Factors*:
  - Gini coefficient and Population: these indicators primarily act as push factors. High inequality (Gini) or population in the origin country might increase migration pressure, while their values in the destination country might not strongly influence a person's decision to migrate.
- *Destination-Specific Factors*:
  - Unemployment Rate: this is generally a pull factor for migration. High unemployment in the destination country can deter migrants, while low unemployment might attract them. Therefore, only the destination's unemployment rate is relevant in influencing migration.
- *Geographical Factors*:
  - Geographical distance and Contiguity: this is also generally a pull factor. Indeed, the further away the destination country, the less likely migrants will move there
- *Cultural and historical Factors*:
  - Same Country, Language and Colonial relationships: sharing the same language or having historical and colonial relationships with the destination country are usually pushing factors.

To estimate parameters following the formula presented in the previous sub-section Regression Modeling, the model was transformed into its corresponding log-linear form:

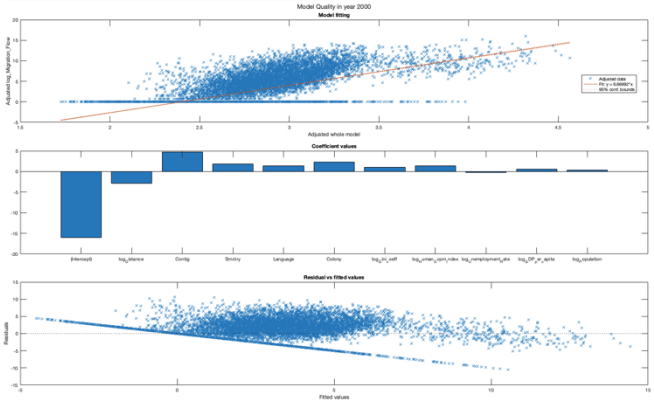
$$log(F_{ij,t}) = \alpha \cdot log(GDP_{i,t} \cdot GDP_{j,t}) - \beta \cdot log(D_{ij}) + \gamma \cdot log(HDI_{i,t} \cdot HDI_{j,t}) + \delta \cdot log(1 - Gini_{i,t}) + \lambda \cdot log(1 - Pop_{i,t}) + \eta \cdot log(1 - Unem_{j,t}) + \theta \cdot Lang_{ij} + \phi \cdot Col_{ij} + \psi \cdot Contig_{ij} + \zeta \cdot Hist_{ij}$$

Before running the model it was also checked the presence of feature collinearity computing the variance inflation factor and running consequently the model only for variables not exceeding a threshold equal to 0.7.

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	−16.001	0.61596	−25.977	1.7244e−146
log_Distance	−2.9013	0.072513	−40.01	0
Contig	4.7189	0.12962	36.407	1.3429e−281
Smctry	1.818	0.17539	10.365	4.0907e−25
Language	1.3729	0.057914	23.705	1.2269e−122
Colony	2.3177	0.31298	7.4054	1.3559e−13
log_Gini_coeff	0.98313	0.075738	12.981	2.2033e−38
log_Human_Dvpm_Index	1.3411	0.098331	13.639	3.539e−42
log_Unemployment_Rate	−0.18829	0.021652	−8.6964	3.6695e−18
log_GDP_per_capita	0.54077	0.025397	21.293	1.4516e−99
log_Population	0.38404	0.0067456	56.931	0

Number of observations: 21316, Error degrees of freedom: 21305
Root Mean Squared Error: 2.43
R-squared: 0.434, Adjusted R-Squared: 0.434
F-statistic vs. constant model: 1.64e+03, p-value = 0



From the first plot we can observe how the fitted values align reasonably well with the actual data, especially for higher adjusted flows. However, some deviations are visible for lower migration flows, where the model appears less accurate. The second plot, concering the coefficient values, reveals the relevance of the geographical contiguity factor, showing a significant effect for neighboring countries, as migrants are more likely to move between contiguous nations. Nevertheless, it is possible to observe also that, differently from the previous regression analysis, all the predictors are statistically significant at 95% confidence levels, with Population and Distance closer to the significance threshold. This confirms the importance of each variable in explaining migration flows within the model.

Finally, the quality of the model is represented by the last plot showing the distribution of the residuals against the fitted values. The picture shows that residuals scatter symmetrically around zero for most fitted values, indicating no significant systematic errors. However, for higher fitted values, the residuals show slight negative trends, suggesting the model slightly overpredicts migration flows at these levels.

In conclusion we can say that coefficients align well with theoretical expectations from gravity models. The negative coefficient for distance and positive coefficients for language, contiguity, and colonial ties highlight how these factors drive migration flows. The economic indicators' coefficients are small but directionally consistent, indicating their influence is secondary but meaningful. So, the results suggest log-linear gravity model provides a robust framework for explaining migration flows but might benefit from additional refinement to capture outliers or extreme cases better.

Conclusion

This analysis started revealing the network structrure of migration flows, identifying the existence of a well-connected and stable network across years, characterized by few important hubs and authorities. Moreover, a further research about the existance of communities showed a small network modularity coefficient suggesting the impossibility to identify well-defined communities.

Subsequently, it was showed how the simple economic characteristics of nations are not enough to determine the migration flow patterns in a reliable way.

A fixed-effects regression analysis highlighted the presence of other events better influencing migration, although also in this case the linear dependence struggles to explain the flow well.

Finally, the introduction of further terms in a gravity model, as distance between countries, language, colonial and historical relationships, improved the predictive qualities making it more robust, but still necessitating additional refinements and suggesting a better interpretation if non-linear models were used.

Further research

The goal of this analysis was the ambitious one to have an overall look of migration flows and interactions with economic indicators among the largest possible number of countries. However, one of the main problems was the lack of data: timeseries span from 1990 to 2020 with a time interval of 5 years, which means to have just 7 observation per country, too few to provide qualitative results.

If there was the possibility to collect more migration flows data to enrich the dataset, it could be interesting to explore the interdependencies between migration flows and economic indicators over time. For instance, how a change in GDP or unemployment in one period affects future migration patterns and vice versa.

Other possible analysis could include the use of non-linear gravity models, to provide a robust estimation when migration flows vary significantly or include many zero values, or the use of spatial econometric models to incorporate spatial dependencies, which are relevant given that migration flows can be influenced by neighboring countries' economic conditions or policies.

Works Cited

Fiona Spooner, T. A.-O. (2022). *Migration.* OurWorldinData.org.

Bank, W. (2023). *GDP per capita (in constant 2017 internation $).* World Bank.

Data, W. B. (2024). *Gini Coefficient – World Bank.* World Bank.

UNDP. (2024). *Human Development Report.* Our World In Data.

Bank, W. (2024). *Unemployment Rate.* Our World In Data.

Mayer, T. &. (2011). *Notes on CEPII's distances measures: the GeoDist Database.* CEPII.