Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis BSc^2 Double Bachelor Econometrics & Economics

Determining International Migration Flows by a Regression of Exchangeable Relational Arrays

Yvette Spoor (503334)

Supervisor:	Wan, P.
Second assessor:	Archimbaud, A.
Date final version:	9th July 2023

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Abstract

Understanding international country-to-country migration flows is a complex challenge due to the multifaceted nature of migration and the heterogeneity of countries. This thesis presents an innovative approach to modeling international migration flows, leveraging the exchangeable relational array structure to gain insights into the underlying patterns of migration by using a set of newly introduced standard error estimators. This study focuses on seven countries by examining their pairwise migration flows over the period 1996-2004. Modified gravity mean models are constructed for both inflows and outflows of migration, incorporating distinct 'push' and 'pull' factors to account for their varying characteristics. The exchangeable relational array model, with its ability to capture dependencies in dyadic data, provides a robust analytical framework for this task. The evaluation of the model is based on R-squared and Mean Squared Predection Error statistics, revealing that the exchangeable estimator significantly outperforms the ordinary least squares estimator. However, further testing is advised due to possible overfitting. By advancing our understanding of migration patterns and offering more accurate predictions, this research contributes to the field of econometrics, and it provides valuable insights for policymakers and researchers studying international migration. The findings highlight the importance of considering the exchangeable relational array model for analyzing migration flows and suggest further research in this area.

Contents

1 Introduction

In an era of globalization, international migration is an increasingly important consequence, with an estimated number of 281 million international migrants in 2020, being over three times what it was in 1970 (International Organization for Migration, 2020). The movement of people across borders gives rise to questions related to various topics, such as economic growth, international relations, and labor markets. Accurately determining the flows of migrants between countries can give valuable insights to policymakers, international organizations, and researchers studying the causes and consequences of migration. Yet, the modeling of these country-to-country migration flows presents a significant challenge. Migration is a complex process influenced by a multitude of factors, including socio-economic, political, and environmental ones (Massey et al., 2001). Moreover, these flows are inherently interdependent, with the migratory patterns of individuals often being contingent upon the actions and experiences of others (Munshi, 2003). To unravel this complexity, this thesis applies a new methodological approach based on the regression of exchangeable relational arrays by Marrs, Fosdick and McCormick (2023), whereby migration flow data is represented as a relational array. Exchangeability is a statistical property that allows for a permutation of indices without altering the joint probability distribution (Hoover, 1979; Aldous, 1981). In the context of modeling migration flows, this property enables to create models that are invariant to the specific ordering of countries. By doing so, the interconnected nature of global migration flows can be captured, while also accounting for the impact of influencing factors. Based on Marrs et al. (2023), migration flow data can be represented as a relational array in the following way

$$y_{i,j,r} = \beta^T x_{i,j,r} + \xi_{i,j,r}(i, j = 1, ..., m; i \neq j; r = 1, ...R), \tag{1}$$

where $y_{i,j,r}$ represents the number of migrants from country i to j, $x_{i,j,r}$ is the $(p \times 1)$ vector of covariates for the relation between i and j, and $\xi_{i,j,r}$ represents the random error for the relation between i and j. The paper of Marrs et al. (2023), which is the basis of this thesis, proposes a new estimator for the coefficients and standard errors of regressions of relational arrays including exchangeability. This leads to the following research question: "Can the modeling of international country-to-country migration flows be improved using exchangeable relational array structures?"

While there is an extensive amount of research exploring which economic factors have an impact on international migration flows, the role of non-economic factors received less attention (Massey et al., 1993). This is better captured by the pairwise country-to-country (C2C) structure of the model, which allows the inclusion of factors between specific pairs of countries. Most existing papers use aggregate measures of migration flows per country, hence either the migration of several origin countries to one destination country or vice versa. These singly comparative researches cannot capture the pairwise variation between origin and destination factors and therefore possibly miss relevant drivers of migration (de Haas, 2011).

This research is relevant in different ways. First, it adds to the scientific literature by evaluating the performance of the new pairwise model of Marrs et al. (2023), which has only been tested on international trade data. This approach allows for the capturing of nuanced relationships that traditional aggregate measures might overlook. Also, by comparing the new structure of the model to more traditional models used for pairwise migration data, it can possibly be further verified that the model allowing for exchangeability is appropriate in different contexts.

On a practical level, insights from this research are helpful for policymakers and international organizations that deal with migration. If the model turns out to be effective in identifying factors for migration, more effective policies on international migration can be taken, for example by introducing programs on cultural integration, or by strategically anticipating the inflow or outflow of people. Businesses that operate in different countries may also benefit from the research on migration patterns by being able to anticipate trends in the labor market and to adapt accordingly.

The results of the research highlight the potential of the exchangeable estimator, both in the replication of the simulation and trade study of Marrs et al. (2023), as well as in the migration context. A modified gravity mean model is constructed for migration inflows of seven countries, and outflows of six countries, based on the country-to-country DEMIG (2015) dataset. Extensive data pre-processing and the collection of relevant country-specific parameters influencing migration was required. The modified gravity mean models are estimated by making out-of-sample forecasts of the migration inflows and outflows for the years 1996-2004. The predictive ability of the exchangeable relational array estimator is determined by comparing the R^2 and Mean Squared Prediction Error to that of a least squares estimator. The exchangeable estimator outperforms the least squares estimator significantly for both statistics, highlighting its potential in determining migration flows. However, due to possible overfitting, further testing with larger datasets is recommended.

This thesis is structured as follows. First, an extensive literature review is provided in Section 2, which discusses migration theories, as well as the link to exchangeable relational arrays. Section 3 explains the data sources used, and the pre-processing that was required on the data. Then, in Section 4, the methodology is explained, first highlighting the replication of Marrs et al. (2023) and then continuing with the migration extension. After this, the results are presented in Section 5. Finally, in Section 6 a conclusion, discussion, and directions for further research.

2 Literature Review

The topic of international migration is a complex and multifaceted issue, shaped by a lot of factors spanning from individual to global levels. This literature review aims to explore the main theories and methodologies that have been used to study international migration, from both an inflows and outflows perspective. Finally, this literature review discusses the innovative use of exchangeable relational arrays and their potential to make a comprehensive understanding of migration patterns.

2.1 Overview of International Migration

In today's interconnected society, international migration is impacting almost all parts of the world (Massey et al., 2001). On one side, people are forced to leave their homes due to conflict, poverty, and disparity, seeking an improved future elsewhere. But simultaneously, for example due to improved modern transportation methods, it has become easier and cheaper for individuals to move looking for work, education, opportunities, or improved quality of life. To give a proper definition of migration, we have that according to the United Nations (2017), a person is considered a migrant if he or she moves to a country other than his or her birth or citizenship country, for a period of at least a year, so that the country of destination effectively becomes his or her new country of usual residence.

The study of international migration is diverse, covering various disciplines, but also yielding wide-ranging findings, trends, and patterns (Massey et al., 2001). Previous research highlights the following factors to have the largest effects on migration. First, economic factors, such as income differentials and unemployment rates, are often trivial in shaping migration decisions (Ortega & Peri, 2013). Research consistently shows that the potential for higher earnings and better economic opportunities can draw individuals to certain countries (Borjas, 1989). Moreover, political conditions, governance quality, and human rights significantly influence migration flows. Studies have shown that political instability, poor governance, and human rights abuses in origin countries can drive individuals to leave. One such example is the exodus observed in Syria due to the ongoing conflict and severe human rights violations (Betts & Collier, 2017). Conversely, democratic political systems, good governance, and respect for human rights attract migrants to such countries. Next to the political conditions, the political position towards immigration, integration, and asylum policies in destination countries also shapes migration flows. Countries that have welcoming immigration policies, such as Germany during the refugee crisis of Syria, have significantly higher inflows of migrants (Betts & Collier, 2017; Hatton, 2017). Likewise, according to Massey et al. (2001) social networks are crucial for migration as well, as these provide resources and support for migrants, for example through information on job opportunities and helping with integration. They argue that once the migration from country A to country B has begun, there tends to be a self-perpetuating flow through the formation of social networks in country B, which help subsequent migrants. More recently, research has started to investigate the increasing role of environmental factors on migration, such as climate change and natural disasters (Black et al., 2011; Piguet, Pécoud & de Guchteneire, 2011).

2.2 Theoretical Framework

Based on the previous mentioned drivers of migration, there are numerous theories serving as tools to understand the dynamics of migration, of which a few central theories are highlighted below. Currently, there is no single theory that individually captures the whole picture of migration, and many theories have been written in isolation from each other, sometimes segmented by different disciplines (Massey et al., 1993). The theories reflect the multidimensional nature of migration, where factors are having an influence on individual (micro), familial (meso), societal (macro) and global levels, often interrelating (Massey et al., 2001; Kuhnt, 2019). Due to this complex varied structure of migration, many theories exist, having their own strenghts, but collectively contributing to the greater understanding of migration.

2.2.1 Push-Pull Theory

The Push-Pull Theory is one of the earliest nowadays relevant migration models, as it still remains influential in understanding which forces drive migration. The first to introduce the idea was Ravenstein (1885, 1889), but Lee (1966) formalized the push-pull theory further. According to Lee (1966), migration is driven by negative factors in the origin country that “push” people away, and positive factors in the destination country that “pull” people towards it. Examples of push factors include worsening economic conditions, political unrest, natural disasters, and poverty, whereas examples of pull factors are better economic conditions or political freedom (Simpson, 2022).

The Push-Pull Theory has been used in empirical studies to analyze migration flows, often using economic variables as a measure of push and pull factors. Specific variables that are commonly used in analyses related to this theory could include GDP per capita, unemployment rates, scores for political stability, and levels of violence or conflict.

Despite being intuitively appealing and widely used, the push-pull theory received criticism for its simplicity. It oversimplifies the migration process by identifying factors as either “pushing” people out of their home countries or “pulling” them into another country, and hence not taking other drivers into account. Lee (1966) recognizes this himself by concluding the paper with: “It is to be expected that many expectations will be found, since migration is a complex phenomenon and the often necessary simplifying condition – all other things being equal – is impossible to realize”. Also, it assumes that the choice of migration is solely an economic decision, whereas there are many other factors driving migration, such as social networks, policies, and personal ambitions, as mentioned before. Moreover, it considers migration to be voluntary, overlooking forced migration situations.

2.2.2 Neoclassical Economics Migration Theory

The Neoclassical Economics Migration Theory builds upon the concepts of the Push-Pull Theory by incorporating economic principles and was introduced in the work of Lewis (1954), Harris and Todaro (1970) and Borjas (1989). It treats migration as an individual cost-benefit decision, where the choice of migration is based on maximizing personal utility. A trade-off is made between the wages at home and the prospect of possibly higher wages elsewhere. This theory has played a large role in interpreting migration flows from developing to developed countries, as differences in wages and employment opportunities between regions or countries are believed to drive people from low-wage, high-unemployment regions to high-wage, low-unemployment regions. As a result, it assumes migration to lead to an equilibrium situation in which wage differences no longer exist.

Empirical studies using this theory often focus on wage differentials between countries. Researchers might analyze differences in average wages or wage distributions, unemployment rates, and relative cost of living.

However, critiques to the Neoclassical approach are that it assumes potential migrants to have perfect information on wages in the destination country. Also, similar to the Push-Pull theory, it neglects non-economic variables like social networks and migration policies (Massey et al., 1993). Moreover, the ideal equilibrium outcome without wage differences is critiqued, as reality also exhibits migration flows without wage differences.

2.2.3 New Economics of Labor Migration

Due to the shortcomings of the existing models, the New Economics of Labor Migration was developed at the end of the 20th century. This theory shifts away from an individual decision-making perspective, towards a more collective, household-centered perspective. It assumes that migration choices are made by the whole family or household to maximize not only income but to also minimize and spread risks as a group (Stark & Bloom, 1985; Taylor, 1999). Groups of individuals are able to control the risk associated with their income by diversifying some parts of the household to different countries. As more developed countries have minimized risk on household income through insurance or government programs, and these programs might be absent in more developing countries (Massey et al., 1993), this theory could especially be relevant for people living in developing countries.

To apply this theory, empirical studies often focus on household income as the main variable of interest. Additional variables could include the diversity of income sources within a household and measures of economic uncertainty or risk.

Despite the innovative approach of the New Economics of Labor Migration, critiques are that it assumes altruistic household or family members, which might not be the case, neglecting the importance of the individual agents. There could be conflicts or dynamics that the framework does not account for, and individuals might have motivations that are not in line with the household. Furthermore, this theory still overlooks non-economic variables, just like the other theories discussed.

2.2.4 Dual Labor Market Theory

Following the New Economics of Labor Migration, another noteworthy theory is the Dual Labor Market Theory, developed by Piore (1979). It assumes that migration is predominantly demand-driven by the labor needs of industrial countries, indicating that migration is caused by pull factors, and not by the push factors of the home country. Certain elements in developed economies require a permanent supply of low-skilled labor, which is often not met by the local population and hence attracts migrant labor.

Empirical studies tested this theory and found evidence supporting the Dual Labor Market Theory, showing that the labor demand of countries indeed significantly influences the migration flows towards it (Constant & Zimmermann, 2005). When applying this theory, common variables to use are labor market variables such as wages, unemployment rates, and the number of job openings in different sectors.

However, this theory is also argued for downplaying the influence of non-economic factors on migration. Moreover, it fails to consider the motivations and individual agency of migrants, because it primarily focuses on the macroeconomics of the destination country.

2.2.5 World Systems Theory

The World Systems Theory, primarily developed by Wallerstein (2011) in 1974, takes a more global perspective and argues that the world is divided into three types of countries/regions: 1) developed core countries that dominate trade, have advanced technologies and benefit from the capitalist world, 2) semiperipheral countries, which are intermediary and often provide core countries with manufactured goods, and 3) peripheral countries which are less developed due to weak economies, which supply raw materials and human labor force, and rely on core countries for capital. It argues that migration flows are largely shaped by the economic imbalances between countries, where core countries attract migrants from peripheral or semi-peripheral countries in search of better economic conditions (Portes, 1978).

Empirical studies often use measures of economic development or inequality when using this theory. These could include GDP per capita, measures of inequality, or indices of economic development.

However, a critique of this theory is that it inadequately accounts for individual or household decision-making, by painting migration as an inevitable response to globalization and capitalism (Massey et al., 1993). Also, it overlooks political factors by not taking the significance of policies into account.

2.2.6 Migration Networks/Social Capital Theory

Lastly, the Migration Networks or Social Capital Theory is discussed. It moves away from the macro-perspective that many theories have and is focused on the interpersonal micro-level networks that individuals have. It discusses that the networks, communities, and relationships that migrants have can arouse further migration by reducing costs, risks, and efforts for people in their networks (Massey, 1990; Massey et al., 1993, 2001). Munshi (2003) analyzes this for Mexican migrants in the U.S. and finds that the same individual is more likely to be employed and hold a higher salary when his or her network is significantly larger, and Palloni, Massey, Ceballos, Espinosa and Spittel (2001) also show that social networks have a significant impact on migration decisions.

When this theory is applied in empirical studies, researchers often focus on variables that measure social connections, such as the size of existing migrant communities in the destination country, or measurements of social network sizes or densities.

However, criticism of this theory includes limited consideration of macroeconomic variables and individual choices, and that it is potentially overemphasizing the effect of networks on migration (Garip, 2008).

2.2.7 Summary of Theories

This review of migration theories indicates a variety of factors at different levels - individual, household, national, and global - that can affect migration decisions and patterns. Each theory contributes in its own way, but especially the combined insights provide a strong foundation for investigating international migration.

The Push-Pull Theory and Neoclassical Economics Migration Theory offer insights on an individual level and navigate towards using variables such as GDP per capita, wage levels, and unemployment rates. The New Economics of Labor Migration extends this perspective to the household level, suggesting that variables related to household income and risk could be relevant. The Dual Labor Market Theory and World Systems Theory highlight the importance of macro-level economic factors that represent labor market conditions and global economic differences, which could be shown by means of industrialization levels or labor force participation rates. And finally, the Migration Networks or Social Capital Theory underlines the importance of social networks in migration decisions. The diaspora size, which is the dispersion or spread of a people from their original homeland, or measures of social cohesion could be relevant to consider.

Given the complexity of the migration structure and the limited time of this thesis, an eclectic approach will be used, by implementing insights from the different theories to construct a model of migration flows. It is noticeable that a significant number of these theories do not incorporate non-economic factors, a gap this research intends to bridge. By making use of pairwise in- and outflow data, it is crucial to remember that each pair of countries has unique patterns influenced by different factors. Therefore, the impact of the previously mentioned theories will likely differ across the country pairs.

2.3 Inflows and Outflows of Migration

When modeling international migration, it is important to note that the concepts of inflows and outflows of a country are fundamentally different. They both have their own set of influencing factors and implications and should therefore be modeled with separate approaches. Understanding these differences can improve the accuracy of migration models.

Migration inflows refer to the arrival of individuals into a specific country from another country. This is driven by a complex interplay of factors, which is already outlined in the theories. Reasons that people move into a new country can include various factors, such as economic opportunities, political stability, social networks, and environmental conditions (Massey et al., 2001; Ortega & Peri, 2013; Borjas, 1989; Betts & Collier, 2017; Hatton, 2017; Black et al., 2011; Piguet et al., 2011). For modeling, considering the "pull" factors of the destination country will likely have a large impact on the inflow of migrants. For an inflow model, factors like the state of the economy, job opportunities, political conditions, and migration policies can all significantly affect the inflows (Lee, 1966; Simpson, 2022). Additionally, the perception of the destination country from the origin country can impact the flows. Countries with positive reputations regarding immigrants draw more immigrants towards them (Betts & Collier, 2017).

On the other hand, migration outflows represent the movement of people out of a particular country into a different one. This is largely influenced by the "push" factors of the origin country, such as political instability, lack of economic opportunities, and environmental threats (Lee, 1966; Simpson, 2022). Moreover, outflows are also influenced by factors such as family structure, or forced migration due to conflicts (Betts & Collier, 2017; Hatton, 2017).

However, outflows are to a certain extent also influenced by the "pull" of destination countries and inflows by the "push" factors of other countries. This leads to an overlap between the factors of in- and outflows, as someone's decision to leave a country is often tied to the perceived opportunities in the destination country (Lee, 1966). An overview of the differences between inflows and outflows is given in Table 1.

	Inflows	**Outflows**
Definition	The arrival of individuals into a country	The movement of people out of a country
Key Influencing Factors	In the destination country: • Economic opportunities • Political stability • Social networks • Environmental conditions The perception of the destination country from the origin country	In the origin country: • Political instability • Lack of economic opportunities • Environmental threats Family structure and forced migration also play a role
Theoretical Basis	Associated with the "pull" factors as explained by the Push-Pull theory.	Associated with the "push" factors as posited by the Push-Pull theory.

Table 1: Comparison of Migration Inflows and Outflows

2.4 Econometric Approaches for Migration Modeling

Migration studies often utilize a range of methodological approaches, shaped by the availability of data, the objectives of the study, and the specific focus of the migration (whether it is internal or international, in- or outflows, or other differentiations). Here, the focus is on common econometric methods used to model international migration in- and outflows. A widely used approach in the empirical literature on migration is the Gravity Model of Migration, which is a special case of the spatial interaction model, first developed by Tinbergen (1962). This model derives from Newton's law of gravity, stating that the interaction (in this case, migration flows) between two places is proportional to the product of their masses (usually represented by population size or GDP) and inversely proportional to the distance between them (Anderson, 2011; Lewer & Van den Berg, 2008). Various specifications of the gravity model include variables that represent the 'push' and 'pull' factors of migration. These factors could include economic variables such as unemployment rates, wage differences, and GDP per capita, as well as socio-political factors like political stability and quality of institutions, among others (Mayda, 2010; Ortega & Peri, 2013).

Ordinary Least Squares (OLS) is a commonly used estimation method in these models, due to its simplicity and ease of interpretation (Lewer & Van den Berg, 2008). However, as migration datasets are often over-dispersed and zero-inflated because there are many zero values for country pairs that do not have migration between them, count data models such as Poisson or Negative Binomial models are also frequently employed (Silva & Tenreyro, 2006; Burger, Oort & Linders, 2009). More recent research has started to use panel data methods to account for unobserved time-invariant country characteristics that might influence migration, and to handle issues of endogeneity. These models might use fixed effects or random effects specifications, and might also employ instrumental variable methods to handle endogeneity (Beine & Parsons, 2015). Despite the rise of these new specifications for modeling gravity models of migration, Ordinary Least Squares remains the most commonly used technique. Then, logarithms and other transformations are commonly taken to deal with over-dispersion and zero-inflated data.

2.5 Migration Flows and Exchangeable Relational Arrays

As discussed in the previous sections, migration flows consist of intricate networks of inflows and outflows, driven by a multitude of factors, from economic opportunities and political stability to environmental conditions. These flows are not isolated events, but interconnected processes forming a complex system of relational arrays. Understanding the dynamics of migration flows hence requires a framework that is able to capture the dependencies and relational structures inherent in migration data. Conventional approaches often fall short in accounting for these complexities, leading to a loss of insights. This can also be seen in the theoretical framework, where it became clear that many of the proposed migration methods have their limitations with regard to taking all drivers of migration into account.

Structuring migration in- and outflows as exchangeable relational arrays is a promising outcome. It offers a more analytical approach that can accommodate the interconnectedness of migration flows. In this way, migration flows between pairs of countries are represented as multidimensional data structures, where each entry corresponds to a specific relational flow, corresponding to a sending country, an incoming country, and a year. Specifically, as the aim is to separately create a model for in- and outflows, there are two options. For the inflows model, there is a reporting ingoing country, an outgoing country, and for the outflows model there is a reporting outgoing country, and an ingoing country.

Furthermore, exchangeability in probability theory assumes that the order in which migration flows are observed does not have an impact on the joint probability of the migration flows. The concept of exchangeability originally introduced by De Finetti (1937) for univariate sequences, and has been generalized to array data by Hoover (1979) and Aldous (1981). In the context of relational models, exchangeability assumes that the probability distribution of the error array is invariant under any simultaneous permutation of rows, columns, and secondary permutations of the third dimension (Kallenberg, 1997; Marrs et al., 2023). This essentially means that the labels or the order of the array indices do not provide any meaningful information about the distribution of the errors. In other words, the estimation of migration flows stays the same if the order of observations, hence the order of country pairs, changes. What matters is the set of observations, not the order they are listed in.

The primary advantage of utilizing exchangeable relational arrays lies in their ability to capture the relational structure which is inherent in migration data. By treating migration flows as entries within multidimensional data structures, not only the size of flows can be explained, but also the relational context, since patterns and dependencies that may be lost or obscured in conventional data representations can be better captured.

Secondly, exchangeability contributes to a more robust modeling approach. As long as the observed covariates are sufficiently informative, the relational error array's probability distribution stays unchanged regardless of the order of observations, providing a significant advantage in the analysis of migration data, which inherently lacks a natural ordering (Marrs et al., 2023). The assumption of exchangeability also simplifies the covariance structure of the relational error array, which gives a covariance matrix with a finite number of unique entries, which further aids in data analysis and model estimation (Li & Loken, 2002; Hoff, 2005; Kallenberg, 1997; Marrs et al., 2023). This leads to the introduction of new exchangeability covariance matrices, with ten possible nonzero entries, which are elaborated on more in the Methodology. This form of covariance matrix, known as an exchangeable covariance matrix, is particularly suitable for modeling exchangeable relational arrays (Marrs et al., 2023; Hoff, 2008; Bickel & Chen, 2009).

To conclude, the combination of exchangeable relational arrays and the redefined covariance matrix by Marrs et al. (2023) has the potential of providing a more nuanced understanding of migration dynamics, allowing for more accurate and robust statistical analyses and predictions.

2.6 Research Gap

While there exists a substantial body of literature that illustrates the various drivers of migration at both individual and macro levels, much of the existing research tends to separate economic from non-economic factors. Such analyses often fall short of providing a thorough understanding of migration, which is a complex phenomenon. Furthermore, the approaches that take into account multiple factors often do so in isolation, failing to fully consider the intricate relations between them.

Moreover, there is little research on the use of relational data structures, such as exchangeable relational arrays, in understanding migration flows. This approach could provide valuable insights into the complex dynamics and interconnectedness of global migration, especially considering the paired structure of international migration data, where each pair of countries has unique patterns driven by different factors.

In addition, migration theories often overlook the impact of non-economic variables such as political conditions, environmental changes, and social networks. Given the large amount of evidence on the significance of these factors (de Haas, 2011), this research intends to contribute to the existing literature by integrating both economic and non-economic factors in the analysis of international migration flows, utilizing exchangeable relational arrays as an innovative data structure.

3 Data

3.1 Trade Data

To replicate the trade model of Marrs et al. (2023), yearly international trade data involving 58 countries over six years is used. These data, originally analyzed and made available by Westveld and Hoff (2011), use a modified gravity mean model (Ward & Hoff, 2007) to represent the logarithm of yearly trade between each pair of countries as a linear function of seven covariates, from 1981 to 2000. The covariates include the log GDP of each country, the log geographic distance between the countries, a measure of cooperation in conflict, and a measure of democracy.

3.2 Migration Data

Data with a similar pairwise structure to the trade data of Westveld and Hoff (2011) are collected on international migration flows. These data are cleaned and formatted to match the requirements of the analysis. The dataset used is the publicly available DEMIG-C2C (country-to-country) dataset by DEMIG (2015), containing comprehensive bilateral migration flows for 34 reporting countries over the period 1946-2011. It includes 29 OECD countries (excluding Estonia, Ireland, Japan, Korea and Turkey), and 5 non-OECD countries (Argentina, Brazil, Czechoslovakia, South Africa and Uruguay). The data were collected from both archival and electronical sources of the respective national statistical offices after World War II. Almost all data reported by the national statistical offices are included, only some countries with minimum reports have been removed to reduce the list of countries in the dataset (Vezzoli, Villares-Varela & de Haas, 2014). The dataset has separate information on inflow, outflow, and net flow totals per country per year, is disaggregated by gender when possible, and also includes whether a country is less or more developed (Vezzoli et al., 2014). In total, it includes 236 countries by citizenship, for the 34 countries reporting their in- and outflows. This enables to see whether a citizen or foreigner left or entered a specific reporting country. The full edition of the dataset cannot be released publicly due to copyright reasons, hence the limited online version provided online (DEMIG, 2015) is used in this research.

The usage of longitudinal data in this dataset involves studying the migration flows of multiple countries over a period of multiple decades. This, together with looking at country-to-country flows, having data on gender, and the inclusion of non-OECD countries makes the dataset unique according to Vezzoli et al. (2014). Migration data is often discussed to have limitations in availability, completeness, consistency, accuracy and comparability (Vezzoli et al., 2014). This is caused by differences in standards reporting by statistical offices, but also by the unrecorded migration of illegals. Also, short-term migration of for example seasonal workers or international students might be misinterpreted (Herrera & Kapur, 2017). Since there is no general data collection method, and this not likely to emerge quickly, it is therefore most common to use this second-best approach where a combination is made of available data from a large number of national statistical offices. However, due to the extensive data collection, combination, and optimization procedure, it turned out that there was less scarcity in migration flow data than is often assumed (Vezzoli et al., 2014).

3.2.1 Data Pre-processing

Given the limitations and complexity of the DEMIG-C2C dataset, and the pairwise structure of the model, comprehensive preprocessing was an essential step in preparing the data for analysis. The preprocessing primarily aimed to adjust the dataset to the needs of the exchangeable relational array models and to select a representative subset of countries and years for studying international migration flows, while preserving as much data as possible. To begin, entries where the same country was recorded as both the origin and the destination of migration flows were removed, as these instances do not contribute to our understanding of international migration and are not compatible with exchangeable relational array models. Next, the inherent pairwise structure of the relational arrays was addressed by filtering the entries based on country pairs. Given that the list of reporting countries (34 in total) was significantly smaller than the list of countries for which in- and out-migration flows were recorded (236 countries), I removed all entries involving countries not included in the list of reporting countries. Furthermore, I ensured that each country pair had a corresponding counterpart pair in the dataset (e.g., for each France-Germany entry, there was a corresponding Germany-France entry), further adhering to the pairwise structure of the model. Lastly, for the remaining set of countries, the next step was to ensure that all countries are full pairs, meaning that they reported in or outflows to all the countries in the set, which reduced the set of countries even further.

The dataset recorded migration flows based on three variables: country of residence, country of citizenship, and country of birth. Country of residence is the most suitable for this research as it allows to track where people are living at a given point in time, regardless of where they were born or what their citizenship status is. It is particularly useful for understanding patterns of movement, including both permanent and temporary migration. Given the interest of this thesis in understanding the overall patterns of migration (both inflows and outflows), the country of residence was selected as the primary variable due to its data availability and relevance to the research, subsequently discarding the other variables. Moreover, to deal with the instances where countries reported separate in and outflows based on gender and citizenship status (citizens and 'both', which includes citizens and foreigners), only the total flow entries were retained, and the gender-specific and citizen-specific entries were removed in order to maintain consistency and avoid duplication, leading to a minimal loss of data.

Lastly, the aim is to identify the most representative time period for the study. After considering the total number of entries per year, a 20-year span providing an optimal balance was found from 1991-2010. Within this period, the nine-year span from 1996 to 2004 contained the highest number of country pairs. To account for potential missing data, I allowed country pairs to miss at most 2 years of data within this nine-year period. Missing values were imputed through linear interpolation, using the second or second-to-last value for instances where the first or last value was missing. Following these pre-processing steps, the refined dataset includes 7 countries for inflows and 6 countries for outflows, giving a total of 42 and 30 country pairs respectively over a nine-year period. These countries are Austria, Denmark, Germany, New Zealand, South Africa, Spain, and Sweden (where Spain is only included in the inflows model). Through these measures, a clean and representative dataset is ensured, suitable for the application of exchangeable relational array models in studying international migration flows.

3.3 Country-specific Parameters

Next to the country in- and outflows data from the DEMIG-C2C dataset, data on country-specific parameters are collected from the World Development Indicators of the World Bank (World Bank, 2023). These indicators provide information on various aspects of a country's economy, political stability, demographics, and labor force characteristics. For the remaining seven countries and nine years in the dataset, the following five indicators are collected:

1. **GDP per Capita (current US$):** This is a measure of the economic output that accounts for the relative cost of living and the inflation rates of the country. It is widely used as an indicator of the standard of living in a country. The measure in US$ is used to make valid comparisons between countries in case of other currencies.

2. **Political Stability and Absence of Violence/Terrorism; Estimate:** This indicator measures the likelihood that the government will be destabilized or overthrown by unconstitutional or violent means, including politically-motivated violence and terrorism. Estimates are a score in units of the standard normal distribution (approximately ranging from -2.5 to 2.5).

3. **Population, total:** This indicates the total population of a country in a specific year, which is important to normalize the migration flows and to understand the relative impact of migration on a country's demographics.

4. **Labor force with advanced education (% of total working-age population with advanced education):** This measures the percentage of the working-age population with tertiary education, serving as a proxy for the skill level of a country's workforce. These values are missing for New Zealand only, hence the World Development Indicator 'Labor force participation rate, total (% of total population, ages 15-64) (modeled ILO estimate) is used as a proxy for this, by multiplying it by 1.1, as a similar but approximately 10% higher trend is observed for the other countries between the two parameters in the dataset.

5. **Unemployment, total (% of total labor force) (national estimate):** This is the

percentage of the labor force that is without work but available for and seeking employment. This parameter provides insight into the job market conditions in a country.

In case of missing values for the country-specific parameters, linear interpolation is used in the same manner as with the migration flows, hence when the first or last value was missing, the second or second-to-last value is used.

For distinguishing between in- and outflows, one factors will be added to inflows and outflows respectively. For inflows, this is the percentage of the labor force with advanced education, as a high value of this can serve as a pull-factor when determining inflows. Other factors on the attractiveness of a country could also be included, such as an index on job vacancy rates, or an indicator of social inclusion and equality like the GINI index which measures economic equality (World Bank, 2023), or a variable on life expectancy. But since these data are difficult to retrieve, and to avoid multicollinearity in the country-specific parameters, the labor force with advanced education is the only factor added to the model right now. For outflows, factors affecting the push factors from a country are helpful. This involves indicators of economic hardship, like poverty or inflation rates, or measures of social unrest. For this reason, the unemployment percentage is included as a push factor in the outflows model.

Moreover, one last parameter is retrieved from a different source, namely a parameter on the distance between country pairs. It is retrieced from the GeoDist database of CEPII (Mayer & Zignago, 2011; CEPII, 2023) which is a French research center in international economics. They provide a large database containing bilateral distances among 225 countries. The geographical distance between countries is a key determinant of migration flows since it is acting as a physical barrier to movement. The greater the distance between the origin and destination countries, the higher the cost (both financially and psychologically) associated with migration, which may deter potential migrants. As such, it is expected that all else being equal, countries that are geographically closer together will have higher migration flows than those that are further apart, which is a common feature of gravity models (Lewer & Van den Berg, 2008).

A summary of the mean and standard deviations of the country-specific parameters is given in Appendix A.

4 Methodology

In this section, I first elaborate on the parts of the methodology of Marrs et al. (2023) that are replicated. This includes a discussion of the least squares framework, an explanation of the dyadic clustering and exchangeable estimator, and an overview of the trade application. Then, the techniques used for the migration application are discussed, and finally, some details about the programming code are given.

4.1 Linear Regression Estimation for Relational Arrays

4.1.1 Least Squares estimation of regression coefficients

A least squares framework is performed to estimate the β coefficients in Equation (1), which is used to perform inference on β in the relational regression model of Equation (1). The least squares estimator $\hat{\beta}_{GLS} = (X^T X)^{-1} X^T y$ is the best estimator if the residuals have constant variance and are uncorrelated, so when the covariance matrix Ω is an identity matrix. However, dependence is expected in pairwise relational data. For example, consider migration flows out of country i towards country j and k, which are both dependent on country i and therefore have a high chance of being correlated. Hence, a generalized least squares framework is considered, which gives the following estimator for β if Ω is known (Aitkin, 1935; Heij et al., 2004; Marrs et al., 2023)

$$\hat{\beta}_{GLS} = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} y. \tag{2}$$

However, since Ω is unknown and has to be estimated we replace Ω in Equation (2) by $\hat{\Omega}$ and obtain the feasible generalized least squares estimator. When $\hat{\Omega}$ is consistent, feasible generalized least squares is asymptotically efficient for β. For inference of β, an estimator of Ω is required, regardless of the estimation of β. These are commonly denoted as sandwich estimators (Huber, 1967; White, 1980). For generalized least squares, the sandwich estimator is given by the following formula

$$\text{var}(\hat{\beta}|X) = (X^T \tilde{\Omega}^{-1} X)^{-1} X^T \tilde{\Omega}^{-1} \Omega \tilde{\Omega}^{-1} X (X^T \tilde{\Omega}^{-1} X)^{-1}, \tag{3}$$

where $\tilde{\Omega}$ is the final estimate of Ω from the generalized least squares procedure. For the purpose of estimating $\tilde{\Omega}$, the paper by Marrs et al. (2023) proposes a dyadic clustering estimator and an exchangeable covariance estimator to compare their performances, which will be presented in the following two sections.

4.1.2 Dyadic Clustering Estimator

The dyadic clustering estimator, as discussed by Fafchamps and Gubert (2007); Cameron and Trivedi (2011); Aronow, Samii and Assenova (2015); and Tabord-Meehan (2018), is a versatile standard error estimator for relational regressions. The fundamental assumption of this estimator is that two relational entries (i, j, r) and (k, l, s) are independent as long as the actor pairs (i, j) and (k, l) do not share an actor. This premise implies that the covariance $\text{cov}(y_{ijr}, y_{kls}|X) = \text{cov}(\xi_{ijr}, \xi_{kls}|X) = 0$ when the pairs of relations do not overlap. However, it does not impose any constraints on the covariance elements for overlapping pairs of relations. In case of migration flows, this indicates that two country pairs that do not share an actor, hence a country, are independent irrespective of the year they are in. In practice, there might be other factors or interdependencies that this structure does not capture. For instance, global events could influence migration flows in multiple country pairs simultaneously, even though they do not share an actor. This might not be captured by the dynamics of the dyadic clustering estimator.

The covariance matrix of ξ is denoted by Ω_{DC}, subject to the independence assumption for non-overlapping pairs. As suggested Fafchamps and Gubert (2007), each non-zero element of Ω_{DC} can be estimated by the product of residuals, that is, $e_{ijr} e_{iks}$ for estimating $\text{cov}(\xi_{ijr}, \xi_{iks})$, where e_{ijr} is the residual $y_{ijr} - \hat{\beta}^T x_{ijr}$. The estimator $\hat{\Omega}_{DC}$ calculates the empirical covariance of the residuals as defined by ee^T (where e is a vector with all residuals e_{ijr}), and introduces zeros to uphold the non-overlapping pair independence assumption. To estimate the variance of the beta coefficients given X, var$(\hat{\beta}|X)$, in the following equation, Fafchamps and Gubert (2007) propose a sandwich variance estimator based on $\hat{\Omega}_{DC}$, which is formulated as

$$\hat{V}_{DC} = (X^T X)^{-1} X^T \hat{\Omega}_{DC} X (X^T X)^{-1}, \tag{4}$$

and is referred to as the dyadic clustering estimator $\hat{V}_{DC}$, and it traces its origin to the extensive literature on cluster-robust standard error estimators (Marrs et al., 2023).

The dyadic clustering estimator has appealing features: it is asymptotically consistent across a variety of error-dependence structures and it is computationally efficient. However, because $\hat{\Omega}_{DC}$ estimates $O(R^2n^3)$ non-zero covariance elements separately based on $O(R^2n^2)$ dependent observations, $\hat{V}_{DC}$ inherently exhibits substantial variability. The dyadic clustering method is optimally efficient only when there is significant heterogeneity in the true covariance structure; otherwise, it may be less efficient.

4.1.3 Exchangeable Covariance Estimator

Exchangeability, as already discussed in Section 2.5, is a commonly adopted modeling assumption for relational and array structured errors. Within this context, the errors are considered jointly exchangeable if the probability distribution of the error array $\Xi = (\xi_{ijr})$ remains unchanged under any simultaneous permutation of the rows and columns, and secondary permutation of the third dimension (Marrs et al., 2023). In mathematical terms, this means $pr(\Xi) = pr\{\Pi(\Xi)\}$, where $\Pi(\Xi) = \{\xi_{\pi(i)\pi(j)\nu(r)}\}$ is the error array with its indices reordered according to permutation operators π and ν. Exchangeability in the regression context implies that the observed covariates are sufficiently informative such that the labels of the rows and columns in the error array are uninformative.

Specific random effects models for $R = 1$ have been described by Li and Loken (2002) and Hoff (2005). Although the corresponding error covariance matrices can have different entries depending on the model, all covariance matrices have at most six unique entries. A significant contribution of Marrs et al. (2023) is to formalize and extend this observation, demonstrating that any jointly exchangeable model for relational array Ξ results in a Ω of the same form, with at most six unique terms when $R = 1$ and at most twelve unique terms when $R > 1$. If a probability model for a directed relational array Ξ is jointly exchangeable and has finite second moments, then the covariance matrix of Ξ contains at most twelve unique values. These twelve unique entries in Ω correspond to the twelve distinguishable configurations of relation pairs (i, j, r) and (k, l, s) with unlabeled actors. These configurations can be separated into two sets of six identical configurations of relations (i, j) and (k, l) with unlabeled actors, as depicted in Figure 1, each set corresponding to $r = s$ and $r \neq s$ respectively. Similar to the dyadic clustering estimator, the exchangeable covariance estimator assumes that non-overlapping relation pairs are independent, such that $cov(\xi_{kls}, \xi_{ijr}) = 0$ for any s and r when (i, j, k, l) are distinct. This assumption sets two of the twelve parameters in Ω to zero (Marrs et al., 2023), hence ten parameters remain.

A new class of covariance matrices is introduced by Marrs et al. (2023), which contains these ten potentially non-zero entries, $\phi_0^{(\eta)}, \phi_a^{(\eta)}, \phi_b^{(\eta)}, \phi_c^{(\eta)}, \phi_d^{(\eta)}$, $(\eta = 1, 2)$. The separation of covariances by $r = s$ and $r \neq s$ implies that Ω consists of blocks of Ω_1 and Ω_2, each consisting of five non-zero terms for $r = s$, $(\eta = 1)$ and $r \neq s$, $(\eta = 2)$, respectively. An exchangeable covariance matrix is defined as any covariance matrix of this form and denoted by Ω_E (Marrs et al., 2023). The proposed estimator of $var(\beta|X)$ is then given by

$$\hat{V}_E = (X^T X)^{-1} X^T \hat{\Omega}_E X (X^T X)^{-1}, \quad \hat{\Omega}_E = \sum_{\eta=1}^{2} \sum_{u=0}^{d} \hat{\phi}_u^{(\eta)} S^{(\eta)}, \tag{5}$$

where $\hat{\Omega}_E$ is a specific form of the covariance matrix. $S^{(\eta)}$ is a binary matrix of dimension $\{Rn(n-1) \times Rn(n-1)\}$ with 1's in the entries corresponding to relation pairs of type $(u = 0, a, b, c, d; \eta = 1, 2)$. The ten parameters in Ω are estimated by averaging the residual products sharing the same index configurations. For instance, the estimate of $cov(\xi_{kls}, \xi_{ijr})$, corresponding to $u = b$ and $\eta = 2$, is given by

$$\hat{\phi}_b^{(2)} = \binom{R}{2}^{-1} \frac{1}{n(n-1)(n-2)} \sum_{r \neq s} \sum_{i} \sum_{j \neq i} e_{ijr} (\sum_{k \neq j} e_{iks} - e_{ijs}), \tag{6}$$

and the estimators of the other nine parameters of the covariance structure for $(s = 0, a, \ldots, e;$ $\eta = 1, 2)$ are defined similarly in Appendix B, each reflecting another dependency. They can be interpreted as the projection of $\hat{\Omega}_{DC}$ into the vector space over symmetric matrices that have the form of Ω_E.

The theoretical comparison of the dyadic clustering estimator and exchangeable estimator by Marrs et al. (2023) proves that the dyadic clustering intuitively that the dyadic clustering estimator is biased downwards and that the bias is more than twice the bias of the exchangeable estimator, hence it tends to improve smaller estimates. When the exchangeability assumption is satisfied, the exchangeable estimator is consistent and more efficient than the dyadic clustering estimator. This will be tested in a simulation study as well to compare the bias and 95% confidence intervals for the two estimators.

4.1.4 Simulation

To reproduce the simulation output and figures of Marrs et al. (2023), the following steps are required. Data are generated in line with a linear regression model using three different types of errors, namely independent and identically distributed (IID) errors, exchangeable errors (by means of a bilinear mixed effects model), and non-exchangeable errors. The regression model used in the paper by Marrs et al. (2023) is

$$y_{ij} = \beta_1 + \beta_2 1_{(x_{2i} \in C)} 1_{(x_{2j} \in C)} + \beta_3 |x_{3i} - x_{3j}| + \beta_4 x_{4ij} + \xi_{ij}, \tag{7}$$

where β_1 represents an intercept, and β_2, β_3, and β_4 are coefficients that are associated with different types of covariates. The true coefficients were fixed to $\beta = (1, 1, 1, 1)^T$. The coefficient β_2 is associated with x_{2ij}, where x_{2ij} follows a Bernoulli(1/2) distribution. If for a pair (i, j) it occurs that $x_{2i} = x_{2j}$, one of the two is randomly flipped to either 0 or 1 to ensure variability. All x_{3ij} and $x_{4i,j}$ are drawn randomly from a standard normal distribution, and β_3 is calculated using the absolute difference between a pair $|x_{3i} - x_{3j}|$.

The error term ξ_{ij} is manipulated to have the same error setting to maintain the same total variance in each specification for comparability of the models, such that the total variance becomes $3n(n-1)$ for each specification. Hence, it mirrors the regression mean model $\beta^T x_{i,j}$.

Three types of errors get generated. First, the independent and identically distributed setting has errors structured as $\xi_{ij} \sim_{iid} \mathcal{N}(0,3)$ for all pairs (i,j). Second, to generate non-exchangeable errors, a mean-zero random effect was incorporated to the upper left quadrant of var(ξ) under independent and identically distributed errors. The errors are hence defined as

$$\xi_{ij} = \tau 1_{(i \leq \lfloor n/2 \rfloor)} 1_{(j \leq \lfloor n/2 \rfloor)} + \epsilon_{ij}, \tau \sim \mathcal{N}(0, \frac{9n}{4\lfloor n/2 \rfloor}), \epsilon_{ij} \sim_{idd} \mathcal{N}(0, 3/4), \tag{8}$$

where $1_{(i \leq \lfloor n/2 \rfloor)}$ indicates whether the index i is less or equal than the floor of $n/2$.

Thirdly, the exchangeable errors are distributed following the bilinear mixed effects model defined by Hoff (2005). It is a generalized form of the "additive common shock" error structure which is used in simulation studies to justify the dyadic clustering estimator. This gives the following equations:

$$\begin{aligned} y_{ij} = \beta^T x_{ij} + \xi_{ij}, \quad \xi_{ij} = a_i + b_j + z_i^T z_j + \gamma_{(ij)} + \epsilon_{ij}, \\ (a_i, b_i) \sim \mathcal{N}(0_2, \Sigma_{ab}), \quad \Sigma_{ab} = \begin{bmatrix} \sigma_a^2 & \rho_{ab}\sigma_a\sigma_b \\ \rho_{ab}\sigma_a\sigma_b & \sigma_b^2 \end{bmatrix}, \\ z_i, z_j \sim \mathcal{N}_d(0, \sigma_z^2 I_d), \quad \gamma_{(ij)} = \gamma_{(ji)} \sim \mathcal{N}(0, \sigma^2\gamma), \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2\epsilon), \end{aligned} \tag{9}$$

where a_i, b_j, z_i, z_j, and ϵ_{ij} are independent. The dimension of the latent space is $d = 2$, the correlation between sender and receiver effects equals $\rho_{ab} = 1/2$, and the sender variance is twice the receiver variance ($\sigma_a^2 = 2\sigma_b^2$). Moreover, $\sigma_z = \sigma_\gamma = \sigma_b$, and $\sigma_\epsilon^2 = 3/4$.

The data generation function `Generate.data` in R implements the generation of these errors based on the provided 'model' parameter. Additionally, the covariate matrix was created using the `Generate.X.bv` function to meet the requirements of the simulation. This is done based on the code provided by Marrs et al. (2023).

The simulation generates 500 random realizations of the covariates for relational datasets of varying sizes ($n = 10, 20, 30, 40$) - a modification to the paper by Marrs et al. (2023) due to running time constraints). For each covariate realization, 100 random error realizations are generated for all three error settings. Each covariate and error realization pair is used to simulate a dataset, which is then fit to the regression model using ordinary least squares. Standard errors are estimated using the exchangeable and dyadic clustering sandwich variance estimators from the previous sections. Confidence interval coverage, bias, and variance of the standard error estimators are calculated and analyzed.

4.2 Forecasting Trade and Migration Using Relational Arrays

4.2.1 Trade Model

To replicate the trade application of Marrs et al. (2023), two models are compared. An OLS model, which assumes independence between yearly trade is used as a baseline and compared to the exchangeable approach, which builds upon the assumptions of the OLS model but also accounts for the error covariance parameters. It estimates the regression coefficients assuming that errors are jointly exchangeable. The complete trade model of Marrs et al. (2023) is a modified gravity mean model, adapted from the gravity model by Tinbergen (1962), and has the following form:

$$\begin{aligned}\log(\text{Trade}_{ijt}) = \beta_{0t} + \beta_{1t}\log(\text{GDP}_{it}) + \beta_{2t}\log(\text{GDP}_{jt}) + \beta_{3t}\log(D_{ijt}) \\ +\beta_{4t}\text{Pol}_{it} + \beta_{5t}\text{Pol}_{jt} + \beta_{6t}\text{CC}_{ijt} + \beta_{7t}(\text{Pol}_{it} \times \text{Pol}_{jt}) + \varepsilon_{ijt},\end{aligned} \tag{10}$$

where $\log(\text{Trade}_{ijt})$ is the log volume of trade sent from country i to country j in year t; $\log(\text{GDP}_{it})$ and $\log(\text{GDP}_{jt})$ are the log Gross Domestic Product of countries i and j, respectively; $\log(D_{ijt})$ is the log geographic distance between the countries; CC_{ijt} is the measure of cooperation in conflict, which is coded as +1 if the countries were on the same side of a dispute and as -1 if they were on opposing sides; and Pol_{it} and Pol_{jt} are the polity measures for i and j, respectively, where polity ranges from 0 (highly authoritarian) to 20 (highly democratic).

4.2.2 Migration Model

To model international bilateral migration flows, a variant of the gravity model proposed by Tinbergen (1962) is adapted to incorporate variables of interest, which capture characteristics of countries that are likely to affect migration decisions. This approach is analogous to the modeling of trade flows by Marrs et al. (2023) which is given in Equation (10), but with adjustments to reflect the differences in the mechanisms underlying trade and migration. The models for inflows and outflows have the following forms respectively:

$$\begin{aligned}\log(\text{Inflow}_{ijt}) = \alpha_{0t} + \alpha_{1t}\log(\text{GDP/capita}_{it}) + \alpha_{2t}\log(\text{GDP/capita}_{jt}) + \alpha_{3t}\text{Pol}_{it} \\ +\alpha_{4t}\log(\text{Pop}_{it}) + \alpha_{5t}\log(\text{Pop}_{jt}) + \alpha_{6t}\text{Edu}_{it} + \alpha_{7t}\log(D_{ijt}) + \varepsilon_{ijt},\end{aligned} \tag{11}$$

$$\begin{aligned}\log(\text{Outflow}_{ijt}) = \delta_{0t} + \delta_{1t}\log(\text{GDP/capita}_{it}) + \delta_{2t}\log(\text{GDP/capita}_{jt}) + \delta_{3t}\text{Pol}_{it} \\ +\delta_{4t}\log(\text{Pop}_{it}) + \delta_{5t}\log(\text{Pop}_{jt}) + \delta_{6t}\log(\text{Unemp}_{it}) + \delta_{7t}\log(D_{ijt}) + \varepsilon_{ijt},\end{aligned} \tag{12}$$

where $\log(\text{Inflow}_{ijt})$ and $\log(\text{Outflow}_{ijt})$ respectively represent the natural logarithm of migration inflows and outflows from country i to country j in year t; $\log(\text{GDP/capita}_{it})$ and $\log(\text{GDP/capita}_{jt})$ represent the natural logarithm of the GDP per capita in countries i and j; Pol_{it} is a measure of political stability in country i; $\log(\text{Pop}_{it})$ and $\log(\text{Pop}_{jt})$ are the natural logarithms of the populations of countries i and j; and $\log(D_{ijt})$ is the log of the distance between countries i and j. The following parameter is specific for the inflows model: Edu_{it} represents the percentage of the labor force with high education in countries i and j as a percentage of the total working-age population with advanced education. Finally, the logarithm of the unemployment percentage of country i is added to the outflows model only, represented as $\log(\text{Unemp}_{it})$.

A log transformation of the in- and outflows, as well as the GDP per capita, the population, the distance, and the unemployment percentage is applied to address the positively skewed nature of these data. Such skewedness is common in migration data as migration flows are always positive and can vary significantly across different country pairs and over time, resulting in a right-skewed distribution. The log transformation then helps to normalize the data distribution and makes the distribution of the error terms more likely to satisfy the normality assumption. Another key feature of this modeling approach is the distinction between 'push' and 'pull' factors in migration. The inflow model, which captures immigration into a country, emphasizes 'pull' factors such as political stability in the inflowing country and educational attainment, which attract immigrants. The outflow model, on the other hand, is based on 'push' factors such as political instability in the outgoing country and high unemployment, which drive emigration.

4.2.3 Trade and Migration Forecasting

First, the regression coefficients β are initialized using ordinary least squares and are then re-estimated with generalized least squares using the exchangeability estimator and covariance matrix from equations (5) and (6), with $\Omega = \hat{\Omega}$ in equation (2). The model is iteratively estimated by means of generalized least squares until convergence of the exchangeable estimator, which is defined as when an absolute change in weighted residual inner product is less than a tolerance level of $\epsilon = 10^{-6}$ for trade and migration inflows, in line with Marrs et al. (2023), and $\epsilon = 10^{-2}$ for migration outflows due to running time limitations. This occurs when

$$\begin{gathered} \left|Q^{\gamma} - Q^{\gamma-1}\right| < \epsilon, \\ Q^{\gamma} = (y - X\hat{\beta}^{\gamma})^T (\hat{\Omega}^{\gamma})^{-1} (y - X\hat{\beta}^{\gamma}), \quad (\gamma = 1, 2, \dots) \end{gathered} \tag{13}$$

where $\hat{\beta}^{\gamma}$ and $\hat{\Omega}^{\gamma}$ are the estimators at iteration γ of the regression coefficients and error covariance matrix, respectively. After convergence, $\Omega = \hat{\Omega}$ is set in equation (3) to get the exchangeable standard errors.

To examine the ability of the exchangeable model compared to the regular model, the out-of-sample predictive performance of the models is compared. For both the exchangeable and the ordinary least squares model, the regression coefficients β are estimated using the initial four years of trade or migration data. These estimates are then used to predict trade values for the subsequent five years, and migration flows for the subsequent four years. The estimates are generated by computing the conditional expectation $E(y_T | \{y_t\}_{t=1}^{T-1})$, based on the assumption by Marrs et al. (2023) that y_T and (y_r) for $r = 1, \dots, T-1$ are jointly normal.

For ordinary least squares, the estimator of trade values or migration flows in year T is based on the coefficients from the previous time period and assumes independent and identically distributed years, it hence becomes $E(y_T | \{y_t\}_{t=1}^{T-1})_{OLS} = X_T \hat{\beta}_{T-1}$.

Secondly, for the exchangeable model, $\hat{\beta}_T = \hat{\beta}_{T-1}$ is set. Also, according to Marrs et al. (2023), the variance $\mathrm{var}(y_t) = \Omega_1$ for all $t = 1, \dots, T$ and the covariance $\mathrm{cov}(y_t, y_{t+h}) = \Omega_2$ for all h. The precision corresponding to the concatenated vector $z_{T-1}^T = (y_1^T, y_2^T, \dots, y_{T_1}^T)$ has the same pattern as the variance $\mathrm{var}(z_{T-1})$, having Ω_1 on the diagonals and Ω_2 on the off-diagonals (Marrs et al., 2023). For $\mathrm{var}(z_{T-1})^{-1}$, the diagonals are defined as Ψ_1 and the off-diagonals as Ψ_2. When the relations $\{y\}_{t=1}^{T-1}$ follow a normal distribution, we have the following estimator for the exchangeable procedure:

$$E\left(y_T \mid \{y\}_{t=1}^{T-1}\right)_E = X_T \hat{\beta}_{T-1} + \Omega_2 \left(\Psi_1 + (T-2)\Psi_2\right) \sum_{t=1}^{T-1} (y_t - X_t \hat{\beta}_t). \tag{14}$$

For the comparative analysis, two main statistical tools are applied: R-squared (R^2) and Mean Squared Prediction Error (MSPE). These measures enable to effectively compare the performance of the different models to the actual trade and migration data of the years after $t = 4$, where the results of the ordinary least squares model will be compared to the exchangeable model estimator. The trade performance will be estimated for $t = 5$ until $t = 10$ (corresponding to the period 1985-1990), and the migration performance for $t = 5$ until $t = 9$ (corresponding to the period 2000-2004).

A statistical measure used to quantify the goodness of fit of a model is R-squared (R^2). In the context of predictions, particularly for out-of-sample data, R^2 provides an understanding of how much of the variance in the dependent variable can be predicted from the independent variables. The formula for R^2 is

$$R^2 = 1 - \frac{\left|E\left(y_T \mid y_{t=1}^{T-1}\right) - y_T\right|_2^2}{\left|y_T - \overline{y}\right|_2^2} \tag{15}$$

In this equation, y_T represents the actual value of the dependent variable (trade or migration), $E\left(y_T \mid y_{t=1}^{T-1}\right)$ represents the predicted value for the model, estimated from the previous data. The double bars, $|\cdot|_2$, denote the L2 norm, which computes the squared differences and sums them. The mean of the dependent variable is given by $\overline{y}$, and n is the number of observations. The term in the numerator, $\mid E\left(y_T \mid y_{t=1}^{T-1}\right) - y_T \mid_2^2$, is known as the sum of squared residuals, it captures the unexplained variability in the data after the model has been fitted. The denominator $|y_T - \overline{y}|_2^2$ represents the total sum of squares, which measures the total variability in the data. In the context of out-of-sample prediction, an R^2 close to 1 indicates a model with a strong predictive performance, while a value close to 0, or even negative, suggests poor generalization to unseen data. In this study, R^2 is used for the out-of-sample prediction performance of the trade model and migration inflow and outflow models.

On the other hand, the Mean Squared Prediction Error (MSPE) provides a more direct measure of a model's predictive accuracy. The MSPE is calculated using the following formula:

$$MSPE = \frac{1}{n(n-1)}\left|E\left(y_T \mid \{y\}_{t=1}^{T-1}\right) - y_T\right|_2^2 \tag{16}$$

where n is the number of observations, y_T is the actual value, and $E\left(y_T \mid \{y\}_{t=1}^{T-1}\right)$ is the predicted value for the ordinary least squares or exchangeable estimator respectively, estimated from the previous data (Marrs et al., 2023). The double bars, $|\cdot|_2$, denote the L2 norm, effectively squaring and summing the differences between predicted and actual values. A lower MSPE, closer to 0, indicates a model with superior predictive accuracy, while higher MSPE values suggest less accurate predictions. The MSPE is exclusively used for evaluating the predictive performance of the migration inflow and outflow models. By utilizing both R^2 and MSPE, a comprehensive view of each model's predictive performance is made, taking into account both the proportion of variance they can predict (R^2) and their predictive accuracy for new, unseen data (MSPE).

4.3 Programming Software

The programming software R will be used to implement the models. The files `function_file_thesis.R`, `reproduce_simulations_thesis.R`, and `reproduce_trade_example_thesis.R` are adapted from the original versions `function_file_.R`, `reproduce_simulations.R`, `reproduce_trade_example.R` of Marrs et al. (2023). The package `netregR`, also created by Marrs et al. (2023) is used. I created the files `migration_extension_testing_inflows.R` and `migration_extension_testing_outflows.R`, which are required to pre-process the data. To run the migration application, I created the files `migration_running.R`, `function_file_thesis_migration_inflows.R`, and `function_file_thesis_migration_outflows.R` to run the migration model.

5 Results

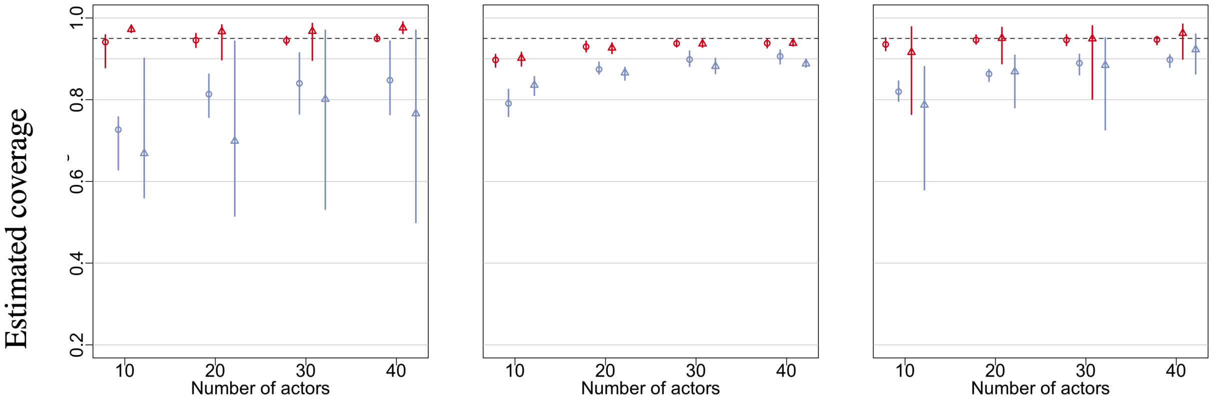
Here, I first provide the results of the replication of the simulation study that compares the performance of the exchangeable and dyadic clustering estimator in Section 5.1. Then, the results of the comparative analysis of trade and migration models with exchangeable and ordinary least squares estimators are given in Section 5.2 and 5.3 respectively.

5.1 Simulation Results

This simulation study is designed to replicate the findings of (Marrs et al., 2023) and to compare the bias and 95% confidence interval coverage when using the exchangeable and dyadic clustering estimators. Simulations are made from a model with three covariates (one binary, one positive real, and one real-valued) with exchangeable and non-exchangeable error models, for $R = 1$ and $n = 10, 20, 30, 40$. Marrs et al. (2023) used a similar approach but had larger numbers of covariates and random error realizations. Figure 1 exhibits the mean coverage and 95% confidence intervals around the mean coverage across various realizations of X covariates. Consistent with the results reported by Marrs et al. (2023), the findings reveal that the estimated mean coverage of the exchangeable estimator is closer to the nominal 0.95 level than that of the dyadic clustering estimator. The exchangeable estimator invariably approaches the nominal 0.95 level more closely than the dyadic clustering estimator. This difference becomes most notable in the case of the binary covariate, where the signal-to-noise ratio is lower than in the other covariates.

In line with the findings reported in the original study (Marrs et al., 2023), the average bias of the dyadic clustering estimator in this replication study is typically more than four times that of the exchangeable estimator under the exchangeable error model. This higher bias could be driving the comparatively poorer coverage performance of the dyadic clustering estimator.

Although the overall patterns in my results mirror those of Marrs et al. (2023), some distinctions arise. Notably, the standard errors in this simulation tend to be somewhat smaller, particularly for the non-exchangeable model (triangles in the figures). These differences are likely due to the modifications made in the replication, which involved fewer random error realizations ($x = 100$ versus $x = 1000$) and smaller dataset sizes ($n = 10, 20, 30, 40$ versus $n = 20, 40, 80, 160, 320$). Future studies with more extensive simulations might yield results more closely aligned with the original study.



(a) binary

(b) positive-real

(c) real-valued

Figure 1: Estimated probability that the true coefficient is in the 95% confidence interval for the three covariates (binary, positive-real, and real-valued), when errors are generated from exchangeable (circles) and non-exchangeable (triangles) models. Points denote mean estimated coverage and lines denote the 95% confidence interval, for exchangeable (red) and dyadic clustering (blue) estimators.

5.2 Trade Results

In the out-of-sample prediction study, the regression coefficients are estimated using the first four years of trade data, and the exchangeable and ordinary least squares estimators are used to predict trade values in the following five years (instead of sixteen years in the replication paper). Similar to Marrs et al. (2023), we found that the exchangeable estimator outperforms the ordinary least squares estimator in all periods, and a similar trend is observed in Figure 2. In the original paper, the gap decreases as the number of years decreases, but this occurs outside of the current range of 10 years which is predicted here in Figure 2.

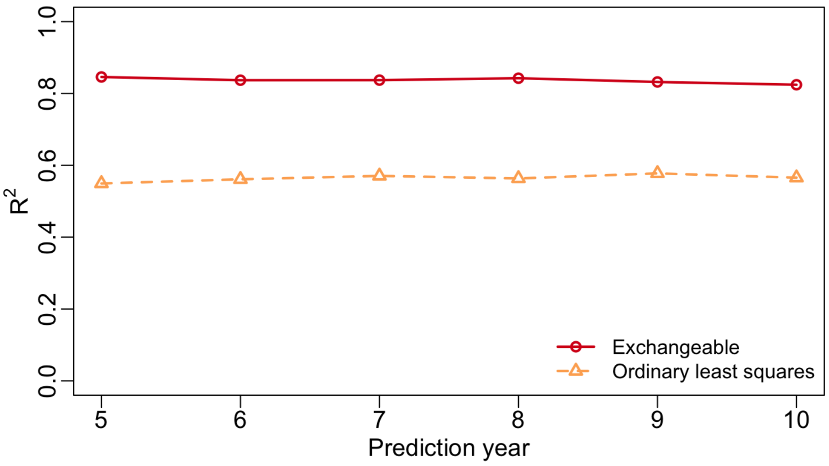


Figure 2: The R^2-values for predicting one-year-ahead trade flows using exchangeable and ordinary least squares approaches

Hence, in line with (Marrs et al., 2023), these results suggest that the exchangeable model provides a more accurate estimation of the trade relationships between countries compared to the ordinary least squares model.

5.3 Migration Results

The regression coefficients are first initialized using ordinary least squares and then fitted using the exchangeability estimator and covariance matrix. The models are iteratively estimated until convergence, which was defined as an absolute change in the weighted residual inner product less than a tolerance level of 10^{-6} for inflows and 10^{-2} for outflows. Next, the out-of-sample predictive performance of the ordinary least squares and the exchangeable models is compared using R^2 and the MSPE. The prediction accuracy for each year was assessed by comparing the predicted trade values with the actual values. Below, the results of migration inflows and outflows are presented separately.

5.3.1 Migration Inflows

In this section, the predictive performance of the exchangeable and least squares estimators for migration inflows is compared, starting with an examination of the out-of-sample R^2 outcomes. In Figure 3 it can be seen that across all the time periods ($t = 5$ until $t = 9$) the model with exchangeable estimator consistently outperforms the ordinary least squares model in terms of R^2. This suggests that the exchangeable estimator can explain a larger proportion of the variance in the dependent variable (total migration inflows per country per year) compared to the least squares estimator when applied to unseen data. For example, at time $t = 5$, the exchangeable model delivers an R^2 of 0.978, whereas the least squares model records an R^2 of 0.792. This pattern persists every year, indicating the robust performance of the exhangeable model in terms of explaining variance in out-of-sample data. However, it is essential to note that the R^2 values nearing 98% could also indicate potential overfitting of the model, or other issues related to multicollinearity, specification errors, or selection bias in the data. In case of overfitting, the model excessively adjusts to the specifics of the training data, where it loses its predictive ability on unseen data. This might result in poor performance when applied to other datasets with more unseen data.

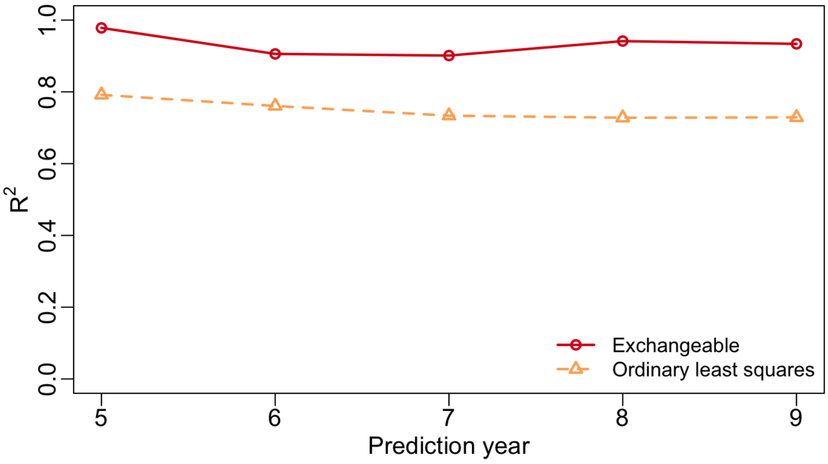


Figure 3: The R^2-values for predicting one-year-ahead migration inflows using exchangeable and ordinary least squares approaches

Secondly, for the MSPE outcomes, which are presented in Figure 4, the superiority of the

GEE approach becomes even more evident. Across all periods, the exchangeable estimator provides lower MSPE values, implying greater accuracy in its predictions compared to the least squares approach. At time $t = 5$, for instance, the exchangeable MSPE has a value of 0.102, while the ordinary least squares model yields an MSPE of 0.981. This trend continues across all time periods, solidifying the exchangeable estimator its dominance in terms of prediction accuracy. Here again, low MSPE values raise concerns of overfitting.

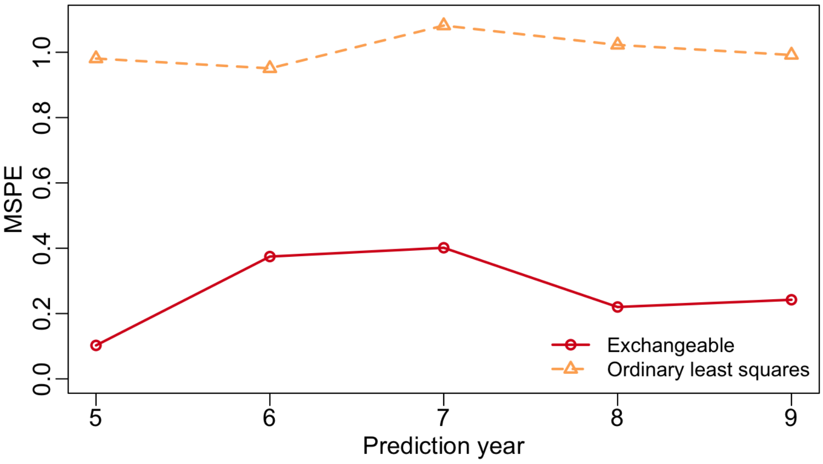


Figure 4: The MSPE values for predicting one-year-ahead migration inflows using exchangeable and ordinary least squares approaches

In conclusion, when predicting migration inflows, the exchangeable estimator outperforms the ordinary least squares estimator, both in terms of explaining a higher percentage of variance in out-of-sample data (as evident by higher R^2 values) and delivering more accurate predictions (as evident by lower MSPE values). However, it is essential to remain vigilant towards the potential overfitting concerns.

5.3.2 Migration Outflows

Turning the attention to migration outflows, a more intriguing pattern is visible compared to the case of migration inflows. As before, the exchangeable estimator consistently outperforms the least squares estimator in terms of R^2 and MSPE across all time periods, which can be seen in Figure 5 and 6, suggesting that the exchangeable model can explain a larger proportion of the variance in the dependent variable when applied to unseen data (R^2), and that it is able to deliver more accurate predictions (MSPE).

However, it is notable that both models experience a significant decline in their performance at $t = 8$ (corresponding to the year 2003), with the R^2 of the least squares model even turning negative, indicating a very poor fit for this particular year's data. Interestingly, there is a considerable spike in the MSPE values for both models at the same period. While the exchangeable estimator model still outperforms ordinary least squares in terms of MSPE (with a statistic of 0.616 and 3.663 respectively), both models exhibit less predictive accuracy during this year.

Given that this anomaly appears for outflows but not inflows, one potential explanation could be an event or series of events in 2003 that significantly affected the migration patterns for the countries under consideration (Germany, Austria, Denmark, New Zealand, South Africa, and

Sweden). Such events could include economic changes, political instability, or social unrest in the source countries that triggered significant outflows, but not necessarily reciprocal inflows. The migration inflows results namely showed a stable trend, with almost the same set of countries (only Spain is extra in the inflows data). It is important to remember that the inflow and outflow of migrants might be driven by different factors. It is hence plausible that an event in 2003 had a more substantial impact on the outflows from these countries than on the inflows. However, further research would be needed to pinpoint the exact cause of these patterns.

For instance, this was a period of increased international tensions due to events like the onset of the Iraq War and the spread of the Severe Acute Respiratory Syndrome (SARS) epidemic, which both could have potentially caused atypical migration patterns. Moreover, specific events related to the countries included in the study, such as Germany and Austria's labor market reforms or Sweden's changing asylum policies, could have induced unusual outflows. However, it is important to note that the raw data did not suggest any large differences in migration outflows around this time period, indicating that the anomaly could also be attributed to model overfitting.

After the year 2003, the performance of both models appears to stabilize, and the superiority of the exchangeable estimator reasserts itself. The consistent performance of the exchangeable model, even in the face of anomalous events, further solidifies its reliability and robustness in predicting migration outflows. Therefore, when predicting migration outflows, despite the anomaly at $t = 8$, the exchangeable approach still provides a more reliable and accurate predictive tool than the OLS approach.

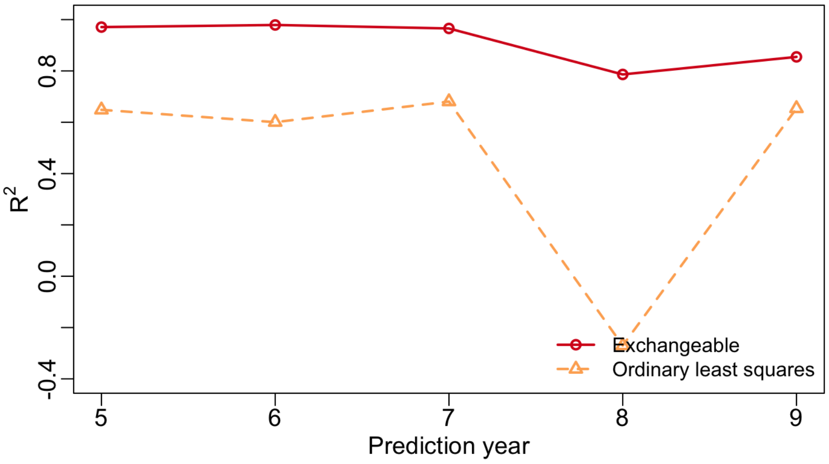


Figure 5: The R^2-values for predicting one-year-ahead migration outflows using exchangeable and ordinary least squares approaches

5.3.3 Concluding Remarks on Migration Results Comparative Evaluation

In summary, the analyses of migration inflows and outflows yield some important insights. The exchangeable estimator model demonstrates consistently superior performance over the ordinary least squares model in predicting both migration inflows and outflows across all considered time periods. This superiority is evidenced by higher R^2 values, indicating a better fit to unseen data, and lower MSPE values, suggesting more accurate predictions.

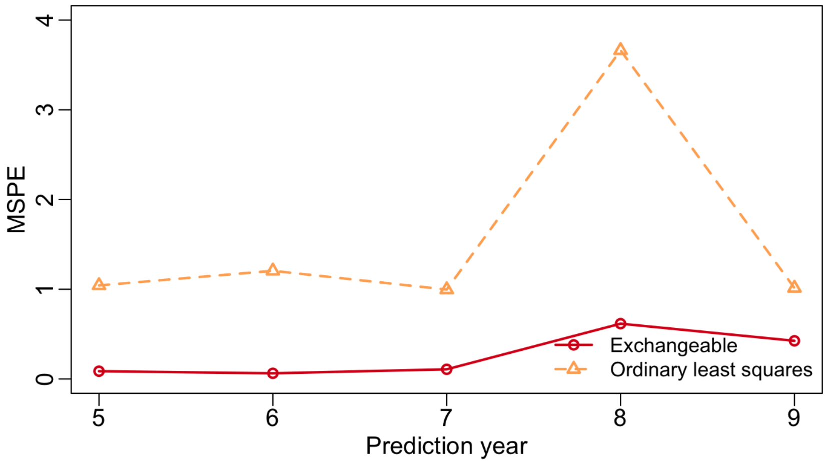


Figure 6: The MSPE values for predicting one-year-ahead migration outflows using exchangeable and ordinary least squares approaches

However, it is important to note that despite the favorable results, the potential risk of overfitting should be considered carefully. While the exchangeable estimator exhibits superior performance, overfitting can occur if the model is excessively complex or if there is limited data available for training. Further testing and validation should be conducted to ensure the robustness and generalizability of the exchangeable estimator.

Moreover, it is worth mentioning that the performance of both models experiences a significant decline at $t = 8$ (corresponding to the year 2003) when predicting migration outflows. The potential reasons for this anomaly, which could include geopolitical events, economic changes, or public health crises, underline the complexity of migration patterns and suggest areas for further investigation.

Taking these factors into account, the exchangeable estimator model consistently outperforms the ordinary least squares model in terms of explanatory power and prediction accuracy for migration patterns. However, further analysis, extensive testing, and validation are needed to confirm its superiority and assess its performance under different scenarios. Future studies should consider the exchangeable estimator as a valuable approach for predicting migration inflows and outflows.

6 Conclusion and Discussion

This study aimed to adress the question, “Can the modeling of international country-to-country migration flows be improved using exchangeable relational array structures?” Grounded in an extensive literature review, it was evident that a comprehensive approach was needed to capture the complex and multi-faceted nature of migration, which has drivers at individual, household, national and global levels. Despite the depth of existing research, many studies offered a fragmented view of migration, focusing either on economic or non-economic factors and failing to explore their interdependencies. For the analysis, data was used from the DEMIG C2C dataset (DEMIG, 2015), covering seven countries for inflows and six for inflows over the period 1996-2004. Building on the foundational work by Marrs et al. (2023), the exchangeable relational array model was adopted and adapted in this study to capture the unique migration patterns between different country pairs. This structure permits each country pair, or dyad, to exhibit unique migration patterns influenced by different factors, with corresponding country-specific parameters. These parameters were derived from data provided by the World Bank (World Bank, 2023). By utilizing the property of exchangeability, the model accommodates the interconnectedness inherent to migration data, taking into account that the order of observations does not influence the joint probability of the migration flows. Thus, it offers a significant advantage in analyzing migration data, which inherently lacks a natural order. Furthermore, this means the labels or the order of array indices do not yield any meaningful information about the distribution of errors.

First, the results of Marrs et al. (2023) were replicated by means of a simulation study that compares the performance of the dyadic clustering estimator and the exchangeable covariance estimator. Similar results to the original study are found, where the exchangeable estimator outperforms the dyadic clustering estimator, with better mean coverage and smaller standard errors. The replication of the trade results also gave similar results, with the exchangeable estimator outperforming the ordinary least squares estimator in all periods. Moving on, the main goal was to test the appropriateness of an exchangeable estimator for modeling international pairwise migration in- and outflows. The empirical evidence from the analyses consistently highlights the potential advantages of using the exchangeable estimator over the ordinary least squares estimator in the modified gravity models used, giving attention to the unique dyadic structure of migration data reflecting specific country-pair effects. The performance metrics highlighted the fit of the model, with the R^2 values of the exchangeable estimator models nearing 98% for both migration inflows and outflows. While these exceptionally high R^2 values may point towards an excellent fit of the model, they also raise questions concerning potential overfitting or other issues related to multicollinearity or specification errors. On the other hand, the Mean Squared Prediction Error statistics were consistently lower for the exchangeable estimator than for the ordinary least squares estimator, providing additional support to its superior predictive performance.

This study is the first to adopt the model of exchangeable relational arrays to migration research. By accounting for the multi-dimensional nature of migration, it offers a more comprehensive perspective than traditional economic theories. It also addresses the dyadic structure of migration data and emphasizes the importance of examining complex interdependencies of various factors influencing migration flows. Consequently, it contributes significantly to the understanding of international migration patterns and proposes an effective modeling approach that can be utilized in related research and policy-making.

Despite the promising results, this study is not impervious to limitations. The anomaly at t=8 for the outflows model remains puzzling, particularly given that there was no anomaly present in the results of the inflows model despite having almost identical countries in the analysis. This calls for further research in investigating this deviation. Secondly, the dataset encompassed a restricted set of countries over a relatively short period. The data pre-processing required for the modeling approach led to the loss of many countries due to the pairwise structure of the data that is required. This potentially restricts the generalizability of the findings, given the constrained temporal and geographical coverage of the data. Furthermore, due to the exclusion of countries with zero migration flows or non-reporting countries, the results might also be susceptible to selection bias. While this decision might have enhanced the data quality, it might have also introduced a bias in the model, as it could disproportionately represent certain types of countries and thereby skew the results. Future research should explore strategies to mitigate this potential bias. Expanding the dataset to include a more extensive array of countries and a longer time period could enhance the robustness and generalizability of the findings. This would also allow researchers to investigate more diverse migration patterns and contexts, deepening the breadth and depth of the understanding of international migration. Moreover, this research did not dive deep into testing the optimal country-specific factors influencing migration in- and outflows due to time constraints and the complexity arising from the multicollinearity among country factors due to their interdependence. After all, the main goal was to compare the performance of the exchangeable estimator opposed to more general methods used, and not specifically to make the best possible models for migration in- and outflows. Due to this, only a limited number of factors was included in the model equations, with only one factor differing in the equations. Future studies should devote more attention to determining which factors are most relevant for both models. For example, factors such as cultural and historical ties between countries, which are currently not accounted for in our model, could be included in future studies to further refine the model's predictive power. This points back to the complex multi-faceted nature of migration flows, which was extensively discussed in the literature review by comparing the aforementioned economic theories regarding migration. The need to account for these interdependencies while avoiding the pitfalls of multicollinearity remains an ongoing challenge for migration research. Lastly, minor differences in the results of this study and that of Marrs et al. (2023) highlight the need for more sophisticated simulations to test and validate the robustness of the models. For this, it is also interesting to apply the model to other multidimensional datasets, such as social network analysis where the relationships between individuals can be modeled, bioinformatics for gene-gene interactions, environmental science for modeling interactions between different species, and in an economic analysis to interdependencies between different market factors. Essentially, any scenario requiring the modeling of complex, interconnected, and multi-faceted relational structures can potentially benefit from the use of exchangeable relational array models.

In conclusion, this research strived to enhance the modeling of international country-to-country migration flows by using exchangeable relational array structures. The findings have presented a nuanced understanding of migration, integrating many country-specific factors within a single framework and demonstrating the utility of the exchangeable estimator in capturing the unique dyadic structure of migration data. This study found that the exchangeable estimator provides better estimates for modeling migration flows based on the results of the R^2 and MSPE statistics, but further testing is required. While there are some limitations to be addressed, this research presents a significant addition to the literature on international migration and lays a foundation for future work in this field.

References

Aitkin, A. (1935). On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh*, *55*, 42–48.

Aldous, D. (1981). Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, *11*(4), 581–598.

Anderson, J. E. (2011). The gravity model. *Annual Review of Economics*, *3*(1), 133-160.

Aronow, P. M., Samii, C. & Assenova, V. A. (2015). A general method for detecting interference between units in randomized experiments. *Sociological Methods & Research*, *48*(1), 426–456.

Beine, M. & Parsons, C. (2015). Climatic factors as determinants of international migration. *The Scandinavian Journal of Economics*, *117*(2), 723–767.

Betts, A. & Collier, P. (2017). *Refuge: Transforming a broken refugee system.* Penguin UK.

Bickel, P. J. & Chen, A. (2009). A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, *106*(50), 21068-21073.

Black, R., Adger, W. N., Arnell, N. W., Dercon, S., Geddes, A. & Thomas, D. (2011). The effect of environmental change on human migration. *Global Environmental Change*, *21*, S3-S11.

Borjas, G. J. (1989). Economic theory and international migration. *International Migration Review*, *23*(3), 457-485.

Burger, M., Oort, F. & Linders, G.-J. (2009). On the specification of the gravity model of trade: Zeros, excess zeros and zero-inflated estimation. *Spatial Economic Analysis*, *4*(2), 167-190.

Cameron, A. C. & Trivedi, P. K. (2011). *Microeconometrics using stata.* Stata Press.

CEPII. (2023). *Geodist database.* Retrieved from `http://www.cepii.fr/CEPII/en/bdd_modele/presentation.asp?id=6`

Constant, A. & Zimmermann, K. F. (2005). Immigrant performance and selective immigration policy: A european perspective. *National Institute Economic Review*, *194*, 94–105.

De Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut henri poincaré* (Vol. 7, pp. 1–68).

de Haas, H. (2011). The determinants of international migration: conceptualizing policy, origin and destination effects. *IMI Working Papers*.

DEMIG. (2015). *Demig c2c, version 1.2, limited online edition.* Oxford. Retrieved from `http://www.migrationdeterminants.eu`

Fafchamps, M. & Gubert, F. (2007). The formation of risk sharing networks. *Journal of Development Economics*, *83*(2), 326–350.

Garip, F. (2008). Social capital and migration: How do similar resources lead to divergent outcomes? *Demography*, *45*, 591–617.

Harris, J. R. & Todaro, M. P. (1970). Migration, unemployment and development: A two-sector analysis. *The American Economic Review*, *60*(1), 126–142.

Hatton, T. J. (2017, 08). Refugees and asylum seekers, the crisis in Europe and the future of policy. *Economic Policy*, *32*(91), 447-496.

Heij, C., de Boer, P., Franses, P., Kloek, T., van Dijk, H. & Rotterdam, A. (2004). *Econometric methods with applications in business and economics.* OUP Oxford.

Herrera, Y. M. & Kapur, D. (2017). Improving data quality: Actors, incentives, and capabilities. *Political Analysis*, *15*(4), 365–386.

Hoff, P. D. (2005). Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association*, *100*(469), 286–295.

Hoff, P. D. (2008). Modeling homophily and stochastic equivalence in symmetric relational data. In *Neural information processing systems* (pp. 657–664).

Hoover, D. N. (1979). *Relations on probability spaces and arrays of random variables* (Tech. Rep.). Princeton, NJ: Institute for Advanced Study.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 221–233).

International Organization for Migration. (2020). *World migration report 2020.* Retrieved from `https://publications.iom.int/system/files/pdf/wmr_2020.pdf`

Kallenberg, O. (1997). Stationary processes and ergodic theory. In *Foundations of modern probability* (p. 168-170). Springer.

Kuhnt, J. (2019). Literature review: drivers of migration. why do people leave their homes? is there an easy answer? a structured overview of migratory determinants..

Lee, E. S. (1966). A theory of migration. *Demography*, *3*(1), 47–57.

Lewer, J. & Van den Berg, H. (2008). A gravity model of immigration. *Economics Letters*, *99*(1), 164-167.

Lewis, W. A. (1954). Economic development with unlimited supplies of labour. *The Manchester School*, *22*(2), 139-191.

Li, H. & Loken, E. (2002). A unified theory of statistical analysis and inference for variance component models for dyadic data. *Statistica Sinica*, 519–535.

Marrs, F., Fosdick, B. & McCormick, T. (2023). Regression of exchangeable relational arrays. *Biometrika*, *110*(2), 513–529.

Massey, D. (1990). Social structure, household strategies, and the cumulative causation of migration. *Population Index*, *56*(1), 3–26.

Massey, D., Arango, J., Hugo, G., Kouaouci, A., Pellegrino, A. & Taylor, J. (1993). Theories of international migration: A review and appraisal. *Population and Development Review*, *19*(3), 431–466.

Massey, D., Arango, J., Hugo, G., Kouaouci, A., Pellegrino, A. & Taylor, J. (2001). *Worlds in motion: Understanding international migration at the end of the millennium* (Vol. 77).

Mayda, A. M. (2010). International migration: A panel data analysis of the determinants of bilateral flows. *Journal of Population Economics*, *23*(4), 1249–1274.

Mayer, T. & Zignago, S. (2011). *Notes on cepii's distances measures: The geodist database* (Working Paper No. 2011-25). CEPII.

Munshi, K. (2003). Networks in the modern economy: Mexican migrants in the u. s. labor market. *The Quarterly Journal of Economics*, *118*(2), 549–599.

Ortega, F. & Peri, G. (2013, 01). The effect of income and immigration policies on international migration. *Migration Studies*, *1*(1), 47-74.

Palloni, A., Massey, D. S., Ceballos, M., Espinosa, K. & Spittel, M. (2001). Social capital and international migration: A test using information on family networks. *American Journal of Sociology*, *106*(5), 1262–1298.

Piguet, E., Pécoud, A. & de Guchteneire, P. (2011). Migration and climate change: An overview. *Refugee Survey Quarterly*, *30*(3), 1–23.

Piore, M. J. (1979). *Birds of passage: Migrant labor and industrial societies.* Cambridge University Press.

Portes, A. (1978). Migration and underdevelopment. *Politics & Society*, *8*(1), 1-48.

Ravenstein, E. G. (1885). The laws of migration. *Journal of the Statistical Society of London*, *48*(2), 167–227.

Ravenstein, E. G. (1889). The laws of migration: Second paper. *Journal of the Royal Statistical Society*, *52*(2), 241–305.

Silva, J. M. C. S. & Tenreyro, S. (2006, November). The Log of Gravity. *The Review of Economics and Statistics*, *88*(4), 641-658.

Simpson, N. (2022). Demographic and economic determinants of migration. *IZA World of Labor*.

Stark, O. & Bloom, D. E. (1985). The new economics of labor migration. *The American Economic Review*, *75*(2), 173–178.

Tabord-Meehan, M. (2018). Inference with dyadic data: Asymptotic behavior of the dyadic-robust t-statistic. *Journal of Business & Economic Statistics*, 1–10.

Taylor, E. J. (1999). The new economics of labour migration and the role of remittances in the migration process. *International Migration*, *37*(1), 63-88.

Tinbergen, J. (1962). *Shaping the world economy; suggestions for an international economic policy.* Twentieth Century Fund.

United Nations. (2017). *International migration report 2017: Highlights.* Retrieved from `https://www.un.org/en/development/desa/population/migration/publications/migrationreport/docs/MigrationReport2017_Highlights.pdf`

Vezzoli, S., Villares-Varela, M. & de Haas, H. (2014). *Uncovering international migration flow data: Insights from the demig databases.* Retrieved from `https://www.migrationinstitute.org/publications/wp-88-14`

Wallerstein, I. (2011). *The modern world-system I: Capitalist agriculture and the origins of the european world-economy in the sixteenth century* (1st ed.). University of California Press. (Originally published in 1974)

Ward, M. D. & Hoff, P. D. (2007). Persistent patterns of international commerce. *J. Peace Res.*, *44*(2), 157–175.

Westveld, A. H. & Hoff, P. D. (2011). A mixed effects model for longitudinal relational and network data, with applications to international trade and conflict. *Ann. Appl. Stat.*, 843–872.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, *48*(4), 817–838.

World Bank. (2023). *World development indicators.* Retrieved 2023-06-10, from `http://`

data.worldbank.org/data-catalog/world-development-indicators

A Mean and Standard Deviation Statistics of Country-Specific Parameters

In the tables below, an overview is given of the mean and standard deviation of the country-specific parameters for the seven different countries, over the period 1996-2004.

Table 2: Mean and Standard Deviation of Country-Specific Parameters for Austria

	Mean	Std. Dev.
GDP per capita ($)	28439.46	3986.51
Political Stability indicator	1.13	0.23
Population	8036152.89	74574.33
Unemployment %	4.99	0.56
Labor force advanced education %	78.55	3.13

Table 3: Mean and Standard Deviation of Country-Specific Parameters for Denmark

	Mean	Std. Dev.
GDP per capita ($)	35221.10	5140.27
Political Stability indicator	1.36	0.18
Population	5338167.78	48592.53
Unemployment %	5.10	0.80
Labor force advanced education %	84.62	0.69

Table 4: Mean and Standard Deviation of Country-Specific Parameters for Germany

	Mean	Std. Dev.
GDP per capita ($)	27601.25	3465.16
Political Stability indicator	1.04	0.35
Population	82244156.00	235028.81
Unemployment %	9.11	0.99
Labor force advanced education %	79.38	0.60

Table 5: Mean and Standard Deviation of Country-Specific Parameters for New Zealand

	Mean	Std. Dev.
GDP per capita ($)	17562.38	3946.32
Political Stability indicator	1.34	0.10
Population	3884977.78	115962.05
Unemployment %	5.94	1.18
Labor force advanced education %	83.59	0.57

Table 6: Mean and Standard Deviation of Country-Specific Parameters for South Africa

	Mean	Std. Dev.
GDP per capita ($)	3574.70	765.10
Political Stability indicator	-0.31	0.14
Population	46725345.67	1304651.14
Unemployment %	26.37	1.84
Labor force advanced education %	84.80	1.25

Table 7: Mean and Standard Deviation of Country-Specific Parameters for Spain

	Mean	Std. Dev.
GDP per capita ($)	17288.68	3535.74
Political Stability indicator	0.18	0.23
Population	40946333.22	1035218.87
Unemployment %	14.96	4.53
Labor force advanced education %	81.91	1.09

Table 8: Mean and Standard Deviation of Country-Specific Parameters for Sweden

	Mean	Std. Dev.
GDP per capita ($)	32418.28	4788.39
Political Stability indicator	1.40	0.05
Population	8893410.56	54467.54
Unemployment %	7.10	2.11
Labor force advanced education %	82.93	2.34

B Parameters of the Exchangeable Covariance Structure

The main text gives an example for the parameter estimator of $\hat{\phi}_b^{(2)}$. Empirical mean estimates used in the exchangeable estimator are defined as follows (Marrs et al., 2023):

$$\hat{\phi}_0^{(1)} = \frac{1}{Rn(n-1)} \sum_r \sum_i \sum_{j \neq i} e_{ijr}^2,$$

$$\hat{\phi}_a^{(1)} = \frac{1}{Rn(n-1)} \sum_r \sum_i \sum_{j \neq i} e_{ijr} e_{jir},$$

$$\hat{\phi}_b^{(1)} = \frac{1}{Rn(n-1)(n-2)} \sum_r \sum_i \sum_{j \neq i} e_{ijr} \left(\sum_{k \neq i} e_{ikr} - e_{ijr} \right),$$

$$\hat{\phi}_c^{(1)} = \frac{1}{Rn(n-1)(n-2)} \sum_r \sum_i \sum_{j \neq i} e_{ijr} \left(\sum_{k \neq j} e_{kjr} - e_{ijr} \right),$$

$$\hat{\phi}_d^{(1)} = \frac{1}{2Rn(n-1)(n-2)} \sum_r \sum_i \sum_{j \neq i} e_{ijr} \left(\sum_{k \neq i} e_{kir} + \sum_{k \neq j} e_{jkr} - 2e_{jir} \right),$$

$$\hat{\phi}_0^{(2)} = \binom{R}{2}^{-1} \frac{1}{n(n-1)} \sum_{r \neq s} \sum_i \sum_{j \neq i} e_{ijr} e_{ijs},$$

$$\hat{\phi}_a^{(2)} = \binom{R}{2}^{-1} \frac{1}{n(n-1)} \sum_{r \neq s} \sum_i \sum_{j \neq i} e_{ijr} e_{jis}$$

$$\hat{\phi}_b^{(2)} = \binom{R}{2}^{-1} \frac{1}{n(n-1)(n-2)} \sum_{r \neq s} \sum_i \sum_{j \neq i} e_{ijr} \left(\sum_{k \neq i} e_{iks} - e_{ijs} \right),$$

$$\hat{\phi}_c^{(2)} = \binom{R}{2}^{-1} \frac{1}{n(n-1)(n-2)} \sum_{r \neq s} \sum_i \sum_{j \neq i} e_{ijr} \left(\sum_{k \neq j} e_{kjs} - e_{ijs} \right),$$

$$\hat{\phi}_d^{(2)} = \binom{R}{2}^{-1} \frac{1}{2n(n-1)(n-2)} \sum_{r \neq s} \sum_i \sum_{j \neq i} e_{ijr} \left(\sum_{k \neq i} e_{kis} + \sum_{k \neq i} e_{jks} - 2e_{jis} \right).$$

Where $\hat{\phi}_0^{(1)}$ is an estimator of $\text{var}(\xi_{ijr})$, $\hat{\phi}_a^{(1)}$ is an estimator of $\text{cov}(\xi_{ijr}, \xi_{jir})$, $\hat{\phi}_b^{(1)}$ is an estimator of $\text{cov}(\xi_{ijr}, \xi_{ikr})$, $\hat{\phi}_c^{(1)}$ is an estimator of $\text{cov}(\xi_{ijr}, \xi_{kjr})$, and $\hat{\phi}_d^{(1)}$ is an estimator of $\text{cov}(\xi_{ijr}, \xi_{kir})$. The estimator $\hat{\phi}_i^{(2)}$, for $i \in \{0, a, b, c, d\}$, is the analogous estimator to $\hat{\phi}_i^{(1)}$, only when $r \neq s$ (Marrs et al., 2023).