The evolution of scientific research on the economics of migration

Bruno Paese*

1 Introduction

Migration is most certainly not something new. Migratory movements have helped shape the world as we know today, besides adding diversity and complexity to societies. Throughout history, people have had various reasons to migrate. Many had to escape poverty, famine, conflicts, or even religious persecution. Some were moved by curiosity and the will to explore the world. However, behind all these motives stands one common reason, which is the search for better life conditions (King, 2012; **?**). If we want to have an idea of how significant the phenomenon has been, we can turn to a few 21st-century numbers. From 2000 to 2024, the estimated number of international migrants jumped from 150 to 281 million; the estimated proportion of the world population considered migrants increased from 2.8% to 3.6%; and the total value of international remittances increased by approximately 550%, rising from 128 to 831 billion US dollars (**?**).

The study of migration has drawn the attention of many areas. Economics was among the first disciplines to be interested in the subject, therefore being one of the cornerstones of migration studies as a research field[1](Greenwood and Hunt, 2003; Levy et al., 2020; Scholten et al., 2022).

2 Literature review

Even though migration emerged as a field of study in the late 19th, early 20th centuries, the interest of economists for migration began to take form only in the 1930s. The Depression of the 1930s caused unemployment rates to reach notable high levels in important industrialized countries, which fuelled more rural-to-urban movements, with masses of unemployed people from rural areas moving toward cities in search of job opportunities and better living conditions, which in turn contributed to an increase in the already high unemployment levels. This scenario raised numerous policy concerns among economists, fostering their interest in migration research (Lucas, 1997; Greenwood, 1997; Greenwood and Hunt, 2003). However, despite the increasing interest in migration in the first half of the 20th century, actual

*Contact: bruno.paese7@gmail.com. This dissertation project is supervised by Prof. Dr. Claire Naiditch.

[1]For an overview of the field of migration studies, see Appendix A.

research faced many obstacles, such as the lack of proper theoretical frameworks (Greenwood and Hunt, 2003).

It is of great importance for the purpose of this study that we understand how the literature on the economic aspects of migration developed. Therefore, in this section, we review the main contributions to the economic research on migration, from the early days of migration studies until approximately the end of the 20th century. Given the purpose of our work, we present our contributions following the traditional classification between internal and international migration. For organizational reasons, we have divided the review into two parts, presenting first the review of the theoretical literature, and then moving on to the review of the empirical literature.

2.1 Theoretical literature

2.1.1 The influence of mainstream economic theory and Sjaastad's human capital model

In 1932, Hicks wrote that "recent researches are indicating more and more clearly that differences in net economic advantages, chiefly differences in wages, are the main causes of migration" (Hicks, 1932, p. 76). According to Greenwood (1975), this concisely states the position of mainstream (or orthodox) economic theory on migration[2] until at least the 1970s, when most, if not all, migration research done in economics adopted the neoclassical framework[3] to overcome the lack of a suitable theory of migration. Under this theoretical framework, also called *disequilibrium system* (Greenwood, 1997), migration is driven by interregional wage differentials, which, from the perspective of rational individuals that maximize utility, represent opportunities for utility gains (Greenwood, 1975; Lucas, 1997; Greenwood, 1997).

It wasn't until the 1960s that important theoretical advances made it possible to talk about an *economics of migration*. Sjaastad (1962) came up with what is arguably the first proper economic model in migration by applying the notion of investment in human capital to the decision to migrate (Greenwood, 1997). The author proposed to identify important costs and returns (or benefits) to migration, reducing the decision to migrate to a simple cost-return analysis. In this setting, wage differentials over space account for private money returns to migration, that is, the possibility of obtaining higher earnings in a different location is seen as a return to migration, thus impacting the decision to migrate[4]. Therefore, in Sjaastad's model, migration helps individuals seek the highest returns on their human capital (Sjaastad,

[2] Actually, Greenwood (1975) used the term "geographical mobility of labor" instead of "migration", because in the neoclassical framework, every individual is seen as a worker, who is a potential migrant, and so migration is treated a movement of workers, which from a macro perspective means a supply of labor. That's the reason why sometimes the terms "migration", "labor migration", and "labor mobility" are used interchangeably in the economic literature on migration.

[3] The neoclassical framework, which dominated mainstream economic theory for most of the 20th century, is based on assumptions of individual utility maximization and rational choice, according to which rational individuals seek to maximize their (expected) utility. From the point of view of an individual, or worker, maximizing utility meant maximizing income, and so an utility-maximizing behaviour would be seeking a higher income.

[4] Greenwood notes that the influence of the neoclassical tradition is evident in Sjaastad's model, since "economic opportunity differentials represent potential for household utility gains that can be arbitraged by migration" (Greenwood, 1997, p. 670).

1962). For this reason, some authors referred to it as the *human capital model* of migration (Shields and Shields, 1989; Greenwood, 1997).

Sjaastad's prominent work was so influential for the economic research on migration that Greenwood wrote that after its publication "migration research by economists really began to blossom" (Greenwood, 1997, p. 669). Although it was not a theoretical framework with a sophisticated mathematical foundation, it provided a solid conceptual basis on which other economists could build, and did build.

2.1.2 Internal migration in developing countries: the Harris-Todaro model

Until the mid-1960s, most efforts had been directed toward studying migration in developed, industrialized countries, since it was the spread of urbanization in these countries that had stimulated research on migration, as previously stated. However, during the 1960s, many so-called developing countries began to undergo the same processes of urbanization and industrialization, drawing the attention of scholars to the relationship between internal migration, especially the rural-urban migration characteristic of urbanization processes, economic growth, and economic development (Lewis, 1954; Ranis and Fei, 1961; Todaro, 1980). Therefore, in the economic literature on migration, migration and economic development walked side by side from the second half of the 1950s until at least the mid-1970s, motivated by an interest in the role of labor supply in the process of economic development and growth in developing countries (Lewis, 1954; Ranis and Fei, 1961). In this context, internal migration was seen as a beneficial, even desirable process, according to which cheap labor from agricultural sectors would flow to industrial, urban regions, fuelling industrialization, and consequently development and growth (Todaro, 1980). During this period, notable models based on the foundations established by Sjaastad (1962) were proposed (Lucas, 1997).

In the late 1960s, Todaro (1969) developed the *Todaro migration model*, which, according to Stark (1991), kicked off the movement of intensive research in the field of rural-urban migration in developing countries. The model consists of a two-sector framework–one industrialized urban sector and one underdeveloped rural sector–and it is fundamentally a behavioral model of labor supply. Among the main features is that migration is primarily stimulated by rational economic considerations of relative costs and benefits (Todaro, 1980), showing off the influence of Sjaastad's human capital model (Shields and Shields, 1989; Lucas, 1997). Todaro (1969) introduces the novel idea that the decision to migrate is conditioned by the urban-rural differences in *expected* income[5], which are made possible by what he defined as *expected* urban earnings. According to Todaro, "the fundamental premise is that as decision-makers migrants consider the various labor-market opportunities available to them as, say, between the rural and urban sectors, choosing the one that maximizes their "expected" gains from migration" (Todaro, 1980, p. 364), meaning that potential rural migrants are maximizers of *expected* utility (Todaro, 1980). Therefore, as long as there are regional differences in *expected* earnings, that is, the expected earnings in

[5]For this reason, Shields and Shields (1989) called it the *expected income model*.

the urban sector exceed the actual earnings in the agricultural sector, there will be rural-urban migration, despite the levels of urban unemployment[6].

Following Todaro (1969), Harris and Todaro (1970) presented their seminal *Harris-Todaro model* of migration, which is an extended version of the basic Todaro model where potential rural migrants continue to behave as maximizers of expected utility. However, Harris and Todaro (1970) present a comprehensive two-sector general equilibrium model, integrating both sectors into a cohesive analytical that investigates the migration impact on production and welfare (Harris and Todaro, 1970; Todaro, 1980). The Harris-Todaro model of migration was as influential as Sjaastad's human capital model for the economic research on migration. It has been ever since the standard neoclassical two-sector model of migration, contributing with important public policy recommendations for developing countries (Harris and Todaro, 1970; Todaro, 1980). Furthermore, like previous models, this one has also been subject to modifications and extensions, such as that made by Corden and Findlay (1975), who introduced capital mobility between the rural and urban sectors, and the one proposed by Fields (1975), which was praised by Todaro (1980) as the most extensive and useful modification of Harris-Todaro model.

2.1.3 Theoretical frameworks on international migration

All of the aforementioned models were originally conceived to be applied to cases of internal migration, given the factors already presented that motivated the study of migration in economics. Even though the work considered by many to be the first on international migration was published in 1927 (see Appendix A), the theoretical literature on international migration did not developed as prolifically as that on internal migration. As a matter of fact, the most basic model for studying international migration was built borrowing from models of internal migration[7], especially Sjaastad's human capital model and the Todaro migration model. According to this basic framework, international migration would be driven by differences in the expected earnings between sending and receiving countries (Lalonde and Topel, 1997).

The main contributions to the theoretical literature ended up being developed with a focus on analyzing the economic impacts of immigrants on the native populations, or the receiving country as a whole. As pointed out by Card (2005), one of the main approaches to estimate the impact of immigration on native workers is the so-called *local labor markets approach*, which was pioneered by Grossman (1982). As stated by Card, this approach "relates differences in the relative structure of wages in different local labor markets to differences in the relative supply of immigrants" (Card, 2005, p. F302), which is closely related to other works on internal migration and wage structures, including Sjaastad (1962). Among notable studies that adopted this local labor markets approach is the work of Borjas (1987),

[6] As we can see, an important difference between Todaro's model and Sjaastad's model is that, for Sjaastad, migrants take into consideration the *actual* rural-urban wage differentials in their migration decision (Lucas, 1997).

[7] Apparently, this was not unique to the economics of migration. As shown by Massey et al. (1993) and King (2012), other social sciences also adopted and adapted theoretical frameworks and models initially developed by economists for the study of internal migration to international migration. In fact, it seems that economic theories and models have been prominent in migration studies (Massey et al., 1993; King, 2012).

which developed his *self-selection* model, a model of selection of immigrants on the basis of unobserved characteristics[8]–such as underlined abilities and productivities–built on the classic work of (Roy, 1951). Later on, Borjas (1991) expanded the self-selection model to account for observed characteristics, such as education (Borjas, 1994).

2.1.4 The New Economics of Labor Migration

As we have shown, theoretical frameworks based on neoclassical economic theory simply dominated migration research during the 20th century, and not only in economics. However, this class of theories and models also had their critics. For instance, Stark and Bloom (1985) and Stark (1991) questioned the assumptions and conclusions of neoclassical theories, arguing that the axiomatic of neoclassical models caused them to overlook empirical evidence, leading to conflicting or fallacious views between rural-urban migration and topics such as fertility, education, and urban employment, among others.

These critiques were part of a broader movement that proposed a new theoretical framework for migration. This new theoretical framework came to be known as *New Economics of Labor Migration* (NELM), or just *New Economics of Migration*, which is a paradigm shift in theories of migration (Stark and Bloom, 1985; Stark, 1991). The most significant change introduced by this new framework concerns the unit responsible for the decision to migrate, which under the neoclassical tradition is a single individual, that happens to be a rational utility-maximizer. According to Stark, "migration decisions are often made jointly by the migrant and by some group of non-migrants. Costs and returns are shared, with the rule governing the distribution of both spelled out in an implicit contractual arrangement between the two parties" (Stark, 1991, p. 25). Therefore, proponents of the NELM highlight that migration decisions are not individual decisions, but joint decisions instead, taken within the ambit of the household or family (Stark and Bloom, 1985; Stark and Lucas, 1988; Stark, 1991).

Just like other models previously mentioned, the NELM primarily addressed questions and issues regarding rural-urban migration[9] in least developed countries (LDCs). Economists noted that migration played a key role in mitigating risks in this group of countries. Therefore, contrary to what neoclassical tradition said, the incentive for rural-urban migration was not so much the interregional wage differentials, but rather the possibility of diversifying the family's income-generating activities, leading to the minimization of the risks to which the household was exposed, such as crop failures. In other words, migration–in this case, sending a family member to work in the urban center–would provide the household with a reliable source of liquidity, in the form of *remittances*, which could be used to offer income insurance, or to finance new production technologies, inputs and activities (Stark and Levhari, 1982; Stark and Bloom, 1985; Stark and Lucas, 1988; Stark, 1991; Taylor, 1999).

[8] Borjas (1987) was careful to note that his model adopts the cornerstone hypothesis of income maximization of Sjaastad's human capital model, once again making clear the influence of internal migration models on theoretical frameworks in international migration.

[9] It is worth noting that the theoretical framework of NELM would also end up being adapted for the analysis of international migration, as shown by Massey et al. (1993) and King (2012).

2.2 Empirical literature on internal migration

2.2.1 Early contributions and modified gravity models

From the early days of migration studies until the end of the 1950s, nearly all research efforts on migration was directed towards a better understanding of the causes, or determinants, of migration (Sjaastad, 1962; Greenwood, 1975; Greenwood and Hunt, 2003). Besides, although some studies did deal with flows of immigrants (see Appendix A), internal migration dominated the agenda, which is consistent with the motivation behind migration research, as previously discussed. However, the lack of theories and appropriate data limited the scope of studies, which were mainly descriptive in style and rather informal in technique. Among the possible determinants considered in early contributions were distance and population size, which are well-known gravity variables, and a couple of personal demographic characteristics, such as age, education, race, income, and marital status (Greenwood, 1997).

The fact that two of the determinants mentioned above were gravity variables was not a mere coincidence. Gravity models are considered by some to be the first formal class of models applied to migration (Greenwood and Hunt, 2003; Cushing and Poot, 2004). These models were limited to analysis from a macro perspective, focusing on the role of space in driving migration, and relied on aggregate data, which was more accessible at the time (Greenwood, 1975, 1997; Greenwood and Hunt, 2003; Cushing and Poot, 2004). According to Greenwood and Hunt (2003), the intuition of gravity models of migration can be traced back to Ravenstein's *laws of migration* (Ravenstein, 1885, 1889). However, the adaptation required to their use in the social sciences, and particularly in migration studies, only occurred in the 1940s, when Zipf (1942, 1946) and Stewart (1947) "elaborated the model and applied it to migration and other human spatial interactions" (Greenwood and Hunt, 2003, p. 26).

The emergence of the first theoretical frameworks during the 1960s motivated the "modification" of gravity models, with the addition of variables expected to influence the decision to migrate, and the redefinition of other variables, which were given behavioral content. This new class of gravity models, called *modified gravity models*, have played a remarkably important role in migration research at least until around the mid-1970s, in a period when virtually all empirical studies were specified using them (Greenwood, 1975, 1997; Greenwood and Hunt, 2003). As an example, Sjaastad's human capital model "provided an appealing rationale for the presence of income variables in modified gravity models, as well as in other models of migration" (Greenwood, 1997, p. 670).

In a context in which modified gravity models prevailed, Greenwood (1975) provided a comprehensive survey of economists' main empirical contributions to the literature on internal migration in the United States from 1960 onwards, focusing on the determinants and consequences of migration. Concerning the determinants, Greenwood states that "one of the clearest implications of the related literature is that gross migration declines perceptibly with increased distance" (Greenwood, 1975, p. 410). Additionally, there was sufficient evidence in the literature that personal characteristics, such as income, employment

status, age, level of education, and race, were important in determining migration. In what concerns the consequences of migration, Greenwood (1975) noted the lack of studies focusing on them, as was already the case before 1960, as pointed out by Sjaastad (1962). According to Greenwood, it was likely that economists gave more attention to the determinants than to the consequences of migration because investigating the consequences would require overly complex models and types of data that were not widely available at the time (Greenwood, 1975).

2.2.2 Improvements in data quality: the role of personal characteristics and life-cycle and family factors

The lack of adequate data was one of the biggest, if not the biggest, limiting factor to the development of economic research on (internal) migration at the time, since virtually all applied migration research was necessarily based on aggregate data, as stated before. As Greenwood himself would write more than two decades later, the aggregate data on which modified gravity models were based on embodied "a number of shortcomings that prevented the study of many important issues bearing on migration" (Greenwood, 1997, p. 707). For these reasons, the improvement in the quality of migration data is arguably the most critical change in migration research in the last two decades of the 20th century.

Particularly since the 1970s, the increasing availability and use of other types of data, especially microdata and longitudinal data, but also time series data, allowed for refinements in econometric techniques, and consequently a better overall understanding of migration processes. This increased availability of data had a considerable impact in several areas. For instance, it allowed a better understanding of how personal characteristics–such as employment status, earnings, education, accumulated skills and training, job tenure, age, sex, and health–and life-cycle and family factors–among which are marriage, divorce, birth and ageing of children, completion of schooling, military service, and retirement–influence the decision to migrate (Greenwood, 1985, 1997).

One relationship that had long troubled economists was that between employment status, or unemployment, and migration[10]. About this, scholars identified three possible "channels"–as defined by Greenwood (1997)–through which unemployment can impact migration: (i) regional unemployment, that is, a region's unemployment rate relative to other regions; (ii) personal unemployment; and (iii) aggregate, or national, unemployment rates (Greenwood, 1997). Among these channels, personal unemployment was the one that benefited most from the increased availability of different types of data. According to Herzog et al. (1993), the notable works of Navratil and Doyle (1977), possibly the first to use microdata to study the relationship between personal unemployment and migration[11], and DaVanzo (1978), found evidence that unemployed individuals are considerably more likely to migrate[12].

[10] As Greenwood had noted on his survey, "one of the most perplexing problems confronting migration scholars is the lack of significance of local unemployment rates in explaining migration" (Greenwood, 1975, p. 411).

[11] Navratil and Doyle also found evidence indicating that the impact of personal and place characteristics on migration, as measured by previous works using aggregate data, are likely biased. Nevertheless, the authors did find that personal characteristics display patterns consistent with past evidences (Navratil and Doyle, 1977).

[12] For a thorough survey of the empirical literature regarding the relationship between personal unemployment and

Another area that benefited significantly from the improvement in the quality of migration data, as mentioned earlier, is the greater emphasis given to life-cycle and family factors as possible determinants of migration. For instance, the advances made in this area are illustrated by the prominent works of Sandell (1977), which identifies the importance of a wife's employment in the family's migration decision, and Mincer (1978), whose study demonstrates that family ties represent negative externalities for individuals, which tend to discourage migration.

Besides allowing for a better understanding of how personal characteristics and life-cycle and family factors influence the decision to migrate, the improvement in data quality also made it possible to better understand the unit responsible for the decision to migrate (individual, family, or household), allowed for a better accountability of the experience of migrants, and paved the way for the study of other types of migration, such as return and repeat migration (Greenwood, 1985, 1997). Still, most of the progress mentioned kept focusing on the determinants of migration, meaning that the increasing availability of other types of data seen in the last two decades of the 20th century was not enough to stimulate the literature on the consequences of internal migration in developed countries.

2.3 Empirical literature on international migration

As noted by LaLonde and Topel, "immigration to the United States during the 1970's and 1980's was greater than in any decade since the 1920's" (LaLonde and Topel, 1991a, p. 297). The economic problems related to these movements of immigrants toward the US caught the attention of economists, prompting a series of empirical studies, aimed mainly at understanding the impact of these immigrants on the US labor market, as well as the problems related to their assimilation, that is, their ability to adapt to the receiving country[13] (Greenwood and McDowell, 1986; Lalonde and Topel, 1997).

2.3.1 Labor market adjustments

What are the effects of immigration in wage rates and employment in the receiving country? The debate surrounding this question is among the oldest ones concerning international migration, largely due to its policy concerns (Greenwood and McDowell, 1986). From a classic perspective, immigration flows represent an adjustment in the labor markets of both the origin and receiving countries. Particularly with respect to the receiving country, it represents an increase in the labor supply, which can potentially affect wages and displace native workers from jobs, in addition to having significant distributional effects (Greenwood and McDowell, 1986; LaLonde and Topel, 1991a,b; Lalonde and Topel, 1997).

Perhaps the first attempt to summarize the empirical literature on the economic impacts of immigration on the labor market of a receiving country (the US, in this case) is the work of Greenwood and McDowell (1986). In this opportunity the authors had not found any robust evidence, stating that

migration, see Herzog et al. (1993).

[13] As pointed out in Appendix A, here we can see another example of how an academically dominant country, like the US, has the power to dictate an entire research agenda.

"empirical conclusions regarding the effects of immigration on US workers have frequently been based on circumstantial rather than on direct evidence" (Greenwood and McDowell, 1986, p. 1767). LaLonde and Topel (1991b) evaluated the effects of immigration to the US by estimating relative wage adjustments among five immigrant cohorts, as well as among young black and Hispanic natives, and concluded that the effects on wages and employment were modest. When surveying the literature, Lalonde and Topel (1997) showed that most studies found only small effects of immigration on labor market outcomes of natives.

2.3.2 Assimilation of immigrants

A classic definition of assimilation of immigrants that was widely used by economists, however simplistic, is based on their relative earning power. There are country-specific skills, such as language, institutional knowledge, job-related skills in particular occupations, and culture, which new immigrants naturally lack compared to similar natives. As stated by LaLonde and Topel, "this means that new immigrants arrive with a human capital 'deficit' that reduces their earning power relative to ethnically similar natives" (Lalonde and Topel, 1997, p. 828). Thus, the process of assimilation of immigrants involves narrowing the skills gap between immigrants and native workers, which would lead to a convergence in earnings. One of the most influential studies of immigrant assimilation is Chiswick (1978), which found that immigrants' relative wages rose with time spent in the US. Chiswick's study also suggests that immigrants acquire a significant amount of country-specific skills during their first decade in the receiving country (Chiswick, 1978; Lalonde and Topel, 1997).

3 The economics of migration

Economics is one of many disciplines interested in migration, and sometimes the multidisciplinary nature of the field of migration studies makes it difficult to recognize specific disciplinary contributions. As far as we could observe, there is no clear and formal definition in the literature of the research field of the *economics of migration*, and we do not intend to provide one. However, considering the purpose of this study, it is necessary that we at least establish the boundaries of the field with which we are working. Therefore, considering what we have seen in the literature, in this study we understand scientific works belonging to the economics of migration to be any work, whether theoretical or empirical, that involves at least one economic aspect of migration, namely: (i) economic determinants of migration; (ii) economic consequences of migration; (iii) economic policies on migration; and (iv) economic theories and models in migration.

Considering what we presented in Section 2, it is safe to say that, for economics in the 20th century, migration was internal migration, and not just any type of migration, but labor migration[14]. Economic

[14] As noted by Lucas, "indeed in general terms it is probably fair to say that economists have been largely preoccupied with the migration of labor" (Lucas, 1997, p. 786).

research on migration has devoted much more effort to studies on internal migration, both theoretical and empirical. For example, from the review of the theoretical literature, we showed that the most widely used and applied theoretical frameworks and models, which have even broken down disciplinary barriers, were developed for internal migration and were later adapted for use in international migration. Regarding the empirical literature, we also noticed that the literature on internal migration developed more extensively and fruitfully, since interest in international migration seems to have come later. For instance, until the publication of their survey in 1997, LaLonde and Topel had not found any robust evidence regarding the determinants of international migration, a clear contrast with the literature on internal migration, which is an indicative that the empirical research on international migration was lagging behind (Lalonde and Topel, 1997).

As pointed out by Cushing and Poot, "recent years have seen a marked shift in migration research from population redistribution within national borders (internal migration) to movement across borders (international migration)" (Cushing and Poot, 2004, p. 318). In fact, we were able to see from the literature review that the greatest advances in international migration research began to occur at the end of the 20th century, specifically from the mid-1980s onwards. Therefore, it is natural to hypothesize that the increasing interest in international migration observed at the end of the 20th century has continued throughout the 21st century, and thus that the *economics of migration* has devoted more attention to international migration issues, as highlighted by Cushing and Poot (2004).

4 Methodology

This study proposes an investigation on how the scientific research on the *economics of migration* has evolved in terms of its topical composition, with the spotlight on the 21st century. Following the discussion in Section 3, our hypothesis is that topics related to internal migration, which were prominent in the 20th century, have lost importance, making way for topics related to international migration to increase their participation in the 21st century. Therefore, we expect to see greater participation in topics related to international migration, to the detriment of those related to internal migration. Ultimately, this change would represent a greater diversification in the field of *economics of migration*, that is, a greater diversity in the field's topical structure.

Intuitively speaking, our problem is that we have a corpus of documents, each consisting of a mixture of topics, but we do not know a priori which topics the documents are about, and that's mainly what we want to find out, so we can proceed to analyse the evolution of the topical composition in our collection of documents. To identify the latent topics in the corpus, we decided to use topic modelling, which is an unsupervised machine learning method of text analysis (Evans and Aceves, 2016). Topic modelling is a strategy through which we can fit a topic model, which in turn are generative Bayesian probabilistic models that allow us to uncover the semantic structure underlying a collection of documents (Blei et al., 2003; Blei and Lafferty, 2009). Among the many types of topic models, one of the most widely used

	Word 1 (w_1)	Word 2 (w_2)	$\cdots$	Word n (w_n)	$\cdots$	Word V (w_V)
Document 1 ($\mathbf{w}_1$)	1	2	$\cdots$	4	$\cdots$	0
Document 2 ($\mathbf{w}_2$)	0	4	$\cdots$	3	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Document m ($\mathbf{w}_m$)	2	2	$\cdots$	0	$\cdots$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Document M ($\mathbf{w}_M$)	5	0	$\cdots$	1	$\cdots$	2

Table 1: Hypothetical example of a document-term matrix with M documents and V words, adapted from Ponweiser (2012). The cells indicate how often a word shows up in a document.

with discrete text data, and at the same time one of the simplest and most straightforward, is Latent Dirichlet Allocation (LDA), which we briefly present in Appendix B. Motivated by Pisarevskaya et al. (2020), we will fit an LDA topic model on the abstracts of scientific works obtained from the Web of Science (WoS) database.

4.1 LDA topic modelling

In a descriptive way, our entire process of performing LDA topic modelling can be summarized as follows: we started with a corpus of M abstracts of scientific works, each of them made up of a bunch of words. However, LDA extracts the underlying semantic structure of documents, and many of our words do not carry any semantic content; hence, we had to pre-process our text data[15]. Pre-processing involves steps such as (i) tokenization, which is the process of breaking down a text into smaller units called tokens, which can be words or other characters; (ii) the removal of stop words, punctuation, and other unused words; (iii) the adjustment of words with different spellings (British and American, for example), synonyms, and acronyms; (iv) and lemmatization, which is a process that helps standardize words that appear in different grammatical and orthographical forms. After pre-processing our text data, we were left with the words that have a semantic content only, which are the ones that form our fixed vocabulary $\{1, \cdots, V\}$. The final step in our data preparation was representing our corpus of M documents as a so-called document-term matrix (DTM)[16], which represents every document in terms of frequencies of words in our fixed vocabulary V. Also, after creating the DTM, we can trim less frequent words using an arbitrary threshold. Given that LDA topic modelling is quite intensive, computationally speaking, removing less frequent words reduces the size of our fixed vocabulary V, which makes computations easier and more efficient.

Once we have the document-term matrix, we can proceed to estimate the LDA topic model using Gibbs sampling[17]. All the computations were done using the statistical software R[18]. Packages *quanteda*

[15] This is a common step in natural language processing (NLP). For a detailed explanation, see (Ponweiser, 2012, p. 9-12).

[16] This is only possible under the bag-of-words assumption (Ponweiser, 2012).

[17] For the sake of brevity, we will omit any technical discussion of Gibbs sampling, but its use when estimating an LDA model is a standard practice. Another common alternative method is variational expectation-maximization (VEM) (Grün and Hornik, 2011).

[18] Our code to implement LDA topic modelling was built following Wouter van Atteveldt and Kasper Welbers (2020).

and *textstem* were used in the text pre-processing, while the function *LDA* from the package *topicmodels*[19] was used to fit our LDA model. There are three elements on the model that need to be defined by us, namely, the hyperparameters η and α—which are responsible for the sparsity of the distributions of words over topics and topics over documents[20], respectively—and the number of latent topics K, which was chosen with the help of the *ldatuning* package[21]. It's worth pointing out that there are no single values that are the correct ones for any of these three parameters, so their definition is thus a research decision. Since LDA is a probabilistic generative process, the fitted LDA topic model produced as outcomes two matrices of probability distributions, the *matrix of per-topic word probabilities* and the *matrix of per-document topic probabilities*.

The *matrix of per-topic word probabilities* displays the posterior probabilities of words belonging to topics, which will help us understand the content of the topics on the *economics of migration*, and consequently label them[22]. In essence, each topic is a list of all the words in the fixed vocabulary, but some words have a much higher probability of belonging to them. Pisarevskaya and colleagues say that "the 20-30 most probable words for each topic can be helpful in understanding the content of the topic" (Pisarevskaya et al., 2020, p. 460), but we concluded that 10-15 is enough, so this is the number of words per topic we will display to assess them.

The second outcome is the *matrix of per-document topic probabilities*, which exhibits the posterior probabilities of documents belonging to topics. Each document is associated with all the topics with some probability, but some topics have a much higher likelihood to be the actual topics that the documents deal with. Therefore, these probabilities show the proportion of each topic in a document (remember that a document can in fact have multiple topics). In the end, these topic proportions were used to perform a *trend topic analysis*, allowing us to follow the trends in topics over time. By grouping the posterior probabilities by topic and year, and averaging them within each year over all of our abstracts, we computed the yearly mean posterior probabilities for each topic, which are considered to be the topic proportions of documents we can expect to have over the years, allowing us to investigate how each topic has changed over time.

To test for our hypothesis that we expect to see a greater diversification in the field of *economics of migration* recently, we used mainly trend topic analysis. As previously stated, the trend topic analysis indicates how the topic proportions changed over time; hence, given our hypothesis, we expect to see a higher proportion of topics related to theories and economic modelling of migration in the late 20th

The data repository with all the replication files is available in: .

[19] This is an R package that provides basic tools for fitting topic models. For a detailed presentation, see Grün and Hornik (2011).

[20] That is, lower values of both η and α lead LDA to fit models with sparse distributions. Usually, the benchmark for what is considered low is anything below 1. Therefore, if $\eta, \alpha < 1$, then each topic will be composed by a smaller number of very high-probability words, and each document will be composed by a smaller number of very high-probability topics.

[21] The *ldatuning* package computes four metrics—Griffiths2004 (Griffiths and Steyvers, 2004), CaoJuan2009 (Cao et al., 2009), Arun2010 (Arun et al., 2010), and Deveaud2014 (Deveaud et al., 2014)—whose results are the statistically optimal number of topics, whereby the researcher may use all of them to make their decision, or only those they deem appropriate.

[22] Remember that LDA is an unsupervised technique, meaning that it learns from the unlabeled data we provide, but it's unable to assess and interpret topics qualitatively (Evans and Aceves, 2016). For a discussion on this, see Chang et al. (2009).

century, but as time goes by we expect to see a decrease in the share of these topics, while other topics increase theirs, showing a greater proportion. In the end, the greater diversity in the field's topical structure that we are hypothesizing would be seen by a convergence in topic proportions.

4.2 Diversity indices

Additionally, following Pisarevskaya et al. (2020), we decided to present two diversity indices to supplement our trend topic analysis with the use of quantitative measures of diversity. Two of the most famous measures of diversity that have been used in many sciences are the Gini-Simpson index and the Shannon-Wiener index (Jost, 2006). The Gini-Simpson index was first proposed by Simpson (1949), and it can be used to measure the probability of two entities representing different types or species. The formula of the Gini-Simpson index adapted to our study, here referred to as GS, is given by:

$$GS = 1 - \sum_{i=1}^{T} (p_i)^2, \tag{1}$$

where T is the number of topics, and p_i is the topic proportion, with $GS \in [0, 1)$. Applied to our case, the index measures the probability that two abstracts of scientific documents selected at random represent different topics. Therefore, the higher the value of GS, the more diverse is our corpus because the higher is the probability that two documents randomly selected are from different topics.

The Shannon-Wiener index, also called the Shannon entropy, is a measure of diversity related to uncertainty. It was first presented by Shannon (1948), and it measures the uncertainty of identifying an entity at random in a certain group or collection. The index, here referred to as H, has the following formula:

$$H = -\sum_{i=1}^{T} p_i \ln p_i, \tag{2}$$

where T is the number of topics, and p_i is the topic proportion, with the value of $H \in [0, \infty)$, where the higher H is, the more diverse our corpus, since we have a greater uncertainty about which topic the document is about when choosing a document randomly in a very diverse collection of documents.

It's worth mentioning that there is a discussion proposed by Jost (2006), where the author points out that indices like Gini-Simpson and Shannon-Wiener are actually entropies and not diversities, and follows to propose variations that truly measure diversity. However, Jost also admits that "entropies are reasonable indices of diversity" (Jost, 2006, p. 363), leaving it clear that the discussion is much more conceptual. Following this discussion, we decided to present only the indices defined above. In any case, acknowledging the differences, we took the decision to call our indices *Gini-Simpson entropy index* and *Shannon-Wiener entropy index*.

5 Data

Regarding our data, we gathered scientific works from the WoS database from January 1, 1983 to December 31, 2024. We performed two searches in all editions of the Web of Science Core Collection, which provides the most complete bibliographic information that can be extracted from the WoS database and allows us to export every file in BibTex extension, which is one of the most recommended to handle bibliographic data. The first search gave us access to data with an excellent time span, while the second search was more extensive in terms of the keywords/terms considered and also gave us access to data of higher bibliographic quality.

Even though in the end the goal was to retrieve documents from the same field, the two searches were carried out using different strategies with the goal of complementing each other, thus forming the most complete and representative database. In the first search, we adopted a less restrictive search strategy, using a smaller set of keywords/terms and also being less strict regarding the quality of bibliographic information. This strategy led us to retrieve and add documents of lower bibliographic quality, that is, missing one or more elements from fields such as abstract or keywords, but that could still be useful. In the second search, we aimed at filling in the gaps left by the first search, retrieving and adding works that have been left out previously by using a broader list of keywords/terms, and adopting a search strategy that would retrieve only documents of the highest bibliographic quality, that is, works that contained almost all of the most important bibliographic elements, which is of the utmost importance for our topic modelling.

Unquestionably, the main concern during the data collection process was having at hand a dataset in which the documents belonged to the field of *economics of migration*. In practical terms, to fulfil this condition, we had to make sure that (i) the documents belonged to the field of migration, and that (ii) those works were dealing with economic aspects of migration. In order to retrieve only documents on migration, we used a set of representative keywords/terms of migration studies in combination with different fields in the WoS advanced search tool[23]. After that, to keep only works dealing with economic aspects of migration, we applied filters in the Research Area and Web of Science Category fields in the WoS advanced search tool; specifically, we only considered scientific works that were simultaneously classified in the Research Area "Business & Economics" and in the Web of Science Category "Economics", although not solely[24]. Additional filters were applied in both searches to consider documents of all types but meeting abstracts, corrections, notes, letters, discussions, bibliographies, bibliographical-items, news

[23] A file detailing how we conducted our searches in the WoS database, including all keywords and terms used, can be found in the data repository.

[24] Documents can belong to multiple Research Areas and Web of Science Categories. Here, we consider documents classified in those areas, solely or not, meaning that a document classified in multiple Research Areas and Web of Science Categories is suitable for us given that one of the Research Areas is "Business & Economics" and one of the Web of Science Categories is "Economics". We think this is the right approach, considering that migration is multidisciplinary by nature, and has recently become more interdisciplinary, as shown in Appendix A. Hence, if we had considered documents belonging only to the Research Area and Web of Science Category of interest, we would have been too strict, leaving out many documents that would fit our definition of *economics of migration*.

items, reprints, and data papers[25], and documents written in English only.

Eventually, the first search retrieved 4,890 scientific works[26], out of which we added 4,696 to our dataset[27]. The second search retrieved 6,056 documents that were all added to the dataset. Summing up the two searches, we had a total of 10,752 documents, with 2,425 works common to both searches. After excluding the duplicates, we end up with the 8,327 documents. However, the WoS database do not include the abstracts of 971 of these documents, making them useless for our topic modelling strategy.

Thus, after removing the documents that did not have their abstracts included, we ended up with 7,386 scientific works from 1991 to 2024 (all the documents from 1983 to 1990 did not have their abstracts included), which constitute our final dataset. Below, we present some bibliographic information about our dataset, in order to show the main characteristics of the scientific research on the *economics of migration*, such as scientific production, most frequent keywords, and most productive countries.

5.1 Metadata

Regarding the document types, the vast majority are articles (both assigned to a final issue and early access) and proceedings papers. In addition, we have a total of 773 sources (journals, books, etc) and 11,131 different authors. Concerning the scientific production, Figure 1 displays the quantity of scientific works published each year. As we can see, from 1991 until 2005, the production stayed below 100 documents per year, accelerating starting from 2006, and reaching a peak production of 671 documents in 2024.

The twelve most relevant authors' keywords in our dataset, by the absolute number of occurrences, can be seen in Figure 2. The most frequent one was originally "migration" with 1,686 appearances, but we decided to remove it because it does not add any semantic value to our analysis. For instance, if we intend to use the list of most relevant authors' keywords to gain an understanding of the topics that have been studied the most, the keyword "migration" does not provide any useful information because it can refer to literally any topic we are dealing with. Also, in the list we have "J61", which is the Journal of Economic Literature (JEL) code for "Geographic Labor Mobility, Immigrant Workers". As we can see, the three most common words are directly related to international migration, with "immigration" and "remittances" ahead of the rest by a large margin. As a matter of fact, seven out of the twelve words are directly related to international migration: "immigration", "remittances", "international migration", "immigrants", "refugees", "J61", and "brain drain". This may indicate a shift in the *economics of migration* research, which, after focusing much more on internal migration throughout the last century, as seen in Section 2, may now be turning its attention to international migration.

[25] For a complete description of all document types available in the Web of Science Core Collection, see: https://webofscience.help.clarivate.com/en-us/Content/document-types.html.

[26] This was the amount recorded the last time the search was performed before the database was exported. Attempted replications using the same settings may generate different results as the WoS databases are constantly being updated.

[27] The difference of the amount retrieved by the search and the amount added to our dataset, of 193 documents, is due to works that we decided to drop after individually reviewing and screening each document retrieved by the first search, a method applied to minimize the presence of off-topic documents.

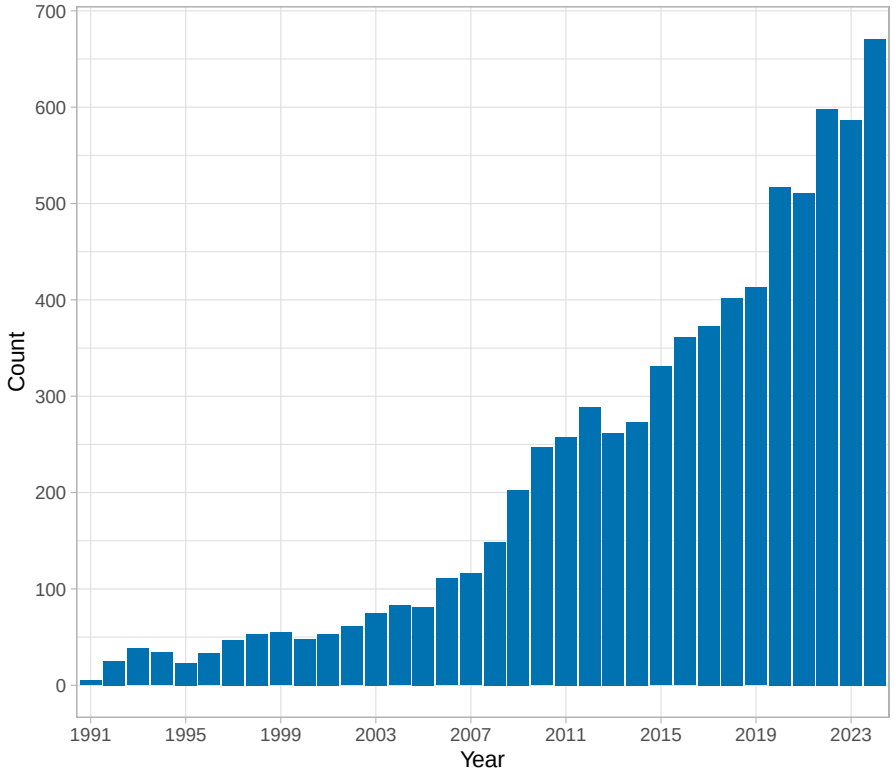


Figure 1: Yearly production of scientific works, 1991-2024.

Table 2 shows the ten most productive countries[28] in terms of five metrics: number of documents, single country publications (SCP), multiple country publications (MCP), total number of citations, and average citation per document. From the figures shown, we highlight the United States, which leads in all metrics but the average document citation, where the United Kingdom is in the lead; and the relative low number of total citations in scientific papers linked to Chinese institutions, leading to a low average citation value per document. Overall, this table shows that research on the economics of migration is heavily concentrated in the United States, Canada, and Europe, aligning with what is seen in the field of migration studies, as shown in Appendix A.

6 Results

6.1 LDA topic model

As discussed in Section 5, we ended up with a corpus of 7,386 abstracts. During the pre-processing of our text data, we removed some neutral words that came up frequently, such as "migration", "migrant", "migrants", "mobility", "article", "paper", "study", "research", and "analysis". Additionally, we defined

[28] To build Table 2, we used one of the main bibliometric measures returned by the function *biblioAnalysis* called *Countries*, which according to Aria and Cuccurullo (2017), it's the affiliation countries' frequency distribution of all co-authors for each paper.

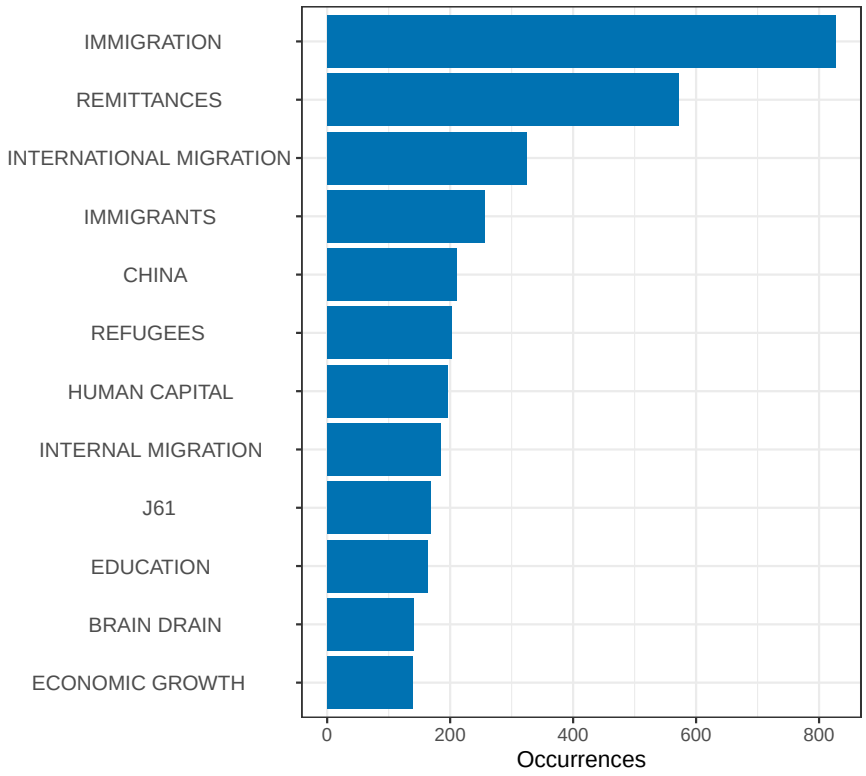


Figure 2: The twelve most relevant authors' keywords in the whole dataset, by the absolute number of occurrences.

the following multi-word expressions we wanted to treat as a single term: "human capital", "brain drain", "brain gain", "foreign direct investment", "european union", "united states", "united kingdom", "climate change", "health care", "asylum seeker". This is an important step, as these multi-word expressions should be considered as they are, because the words that compose them, separately, even if they appear together in the topics, might make their interpretation challenging.

After pre-processing our data and creating the document-term matrix (DTM), we decided trim the terms that appeared less than 10 times, in order to keep only the most essential words that could be identified as being part of a topic. In the end, we were left with 3,748 words that made up our fixed vocabulary. Thus, our DTM–as illustrated in Table 1–is defined as a matrix of 7,386 rows and 3,748 columns. With that in hand, we estimate our LDA topic model with $K = 26$ topics, a number validated by the package *ldatuning*, and $\alpha = \frac{1}{K} \Leftrightarrow \alpha = 0.038$, which is a value that tells the algorithm that each document will be composed by a smaller number of very high-probability topics. For η[29], we used the default value of 0.1 given by the *topicmodels*'s function *LDA*. Tables 3 and 4 present the fifteen most likely terms from our 26 topics, which were taken from the matrix of per-topic word probabilities. As mentioned in Section 4.1, LDA topic modelling is unable to assess and interpret topics qualitatively, meaning that it is one of our tasks to label the topics. To do that, we used a joint strategy of checking

[29] Just for the record, *topicmodels* uses β to define what we refer to here as η. In fact, this change in the notation of this hyperparameter is not unusual, as seen in Ponweiser (2012).

Country	Number of Documents	SCP	MCP	Total citations	Average document citation
USA	1,696	1,244	452	57,517	33.91
CHINA	607	423	184	9,016	14.85
UNITED KINGDOM	599	329	270	23,805	39.74
GERMANY	492	276	216	9,175	18.65
ITALY	341	215	126	6,397	18.76
CANADA	247	172	75	4,741	19.19
FRANCE	246	146	100	6,273	25.50
SPAIN	237	172	65	3,946	16.65
AUSTRALIA	236	137	99	5,722	24.25
NETHERLANDS	154	95	59	3,882	25.21

Table 2: The ten most productive countries in terms of total number of documents published, collaboration indices SCP and MCP, total number of citations, and average citation per document, 1991-2024.

the twenty words with the highest probability in each topic and checking the ten works that are most likely to belong to each topic. We present the labels for the 26 topics in Table 5.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13
labor	tax	country	health	policy	labor	social	education	entrepreneur	immigration	country	immigrant	population
market	model	economic	immigrant	european_union	wage	policy	child	firm	effect	trade	social	europe
unemployment	policy	development	increase	state	worker	understand	school	business	immigrant	flow	network	age
rate	welfare	labor	united_states	europe	effect	development	student	knowledge	increase	effect	cultural	year
price	income	factor	program	immigration	immigration	literature	effect	innovation	find	international	ethnic	economic
house	public	process	policy	country	model	economy	high	activity	native	destination	country	change
effect	equilibrium	social	immigration	member	skill	economic	parent	technology	impact	origin	integration	rate
shock	country	problem	effect	asylum	increase	focus	graduate	expatriate	evidence	export	language	increase
increase	government	main	health_care	right	market	process	find	new	crime	bilateral	diversity	country
employment	show	international	undocumented	integration	sector	context	university	development	result	find	effect	growth
wage	economy	european_union	find	national	impact	study	impact	employee	local	result	group	century
find	may	population	insurance	border	native	approach	result	diaspora	datum	immigration	host	new
high	cost	aim	service	political	productivity	global	human_capital	local	use	fdi	community	period
supply	fiscal	level	illegal	united_kingdom	economy	issue	increase	skill	share	increase	different	time
impact	rate	people	use	economic	unskilled	explore	age	result	exploit	use	origin	early

Table 3: Fifteen most likely terms from topics 1 to 13.

Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19	Topic 20	Topic 21	Topic 22	Topic 23	Topic 24	Topic 25	Topic 26
attitude	region	rural	city	remittance	model	refugee	household	woman	immigrant	environmental	country	worker
individual	economic	china	area	country	use	covid-19	remittance	gender	native	effect	return	labor
survey	flow	urban	move	growth	datum	pandemic	income	labor	wage	climate_change	skill	work
intention	spatial	agriculture	location	economic	estimate	host	poverty	female	earnings	increase	human_capital	employment
information	growth	area	population	financial	effect	crisis	effect	family	difference	pollution	emigration	job
social	internal	land	house	impact	result	syrian	impact	household	gap	natural	home	market
risk	factor	city	individual	development	variable	country	consumption	man	find	shock	education	employer
find	income	labor	large	effect	measure	war	use	work	high	air	develop	condition
factor	labor	rural-urban	local	use	method	impact	food	male	group	impact	brain_drain	informal
result	effect	impact	destination	inflow	approach	force	increase	employment	labor	agriculture	model	sector
use	across	increase	decision	result	empirical	large	inequality	market	market	household	destination	train
preference	population	household	urban	develop	base	turkey	result	find	education	find	selection	formal
datum	model	local	state	positive	test	aid	international	likely	use	disaster	origin	wage
life	interregional	farm	datum	relationship	panel	economic	datum	participation	country	climate	level	construction
well-being	difference	effect	choice	significant	estimation	conflict	family	marriage	among	result	high	industry

Table 4: Fifteen most likely terms from topics 14 to 26.

We will not address all the topics, but some comments are in order. In all cases where we label the topic simply with "migration", without specifying whether it refers to internal or international migration, it's because both forms are considered within the topic. Topics 2, 6, and 19 are strictly about theoretical frameworks and economic modelling. Considering the importance these topics had throughout the 20th century, it is expected that they will play a greater role in this period. Topics 7, 9, 10, 11, 18, 23, and 25 are exclusively about international migration, while Topic 21 has a very strong international migration content[30]. Hence, following our hypothesis, we expect to see an increase in the participation of these topics. Topics 3 and 7 were labelled as "selected topics" because neither his words nor the works related to them seem to define any specific topic. Topic 4 has words related to irregular migration and health-related issues (some papers belonging to this topic are even related to both), and it seems to be heavily focused on the US. Topics 8 and 16 are heavily skewed toward China, with the latter related to many works dealing with Chinese rural-urban migration (the 18th most likely word in this topic is "hukou"). Topic 20 can be found in the literature as "forced migration", but this terminology is outdated. Topic 21 is related to many works dealing with income inequality and poverty. Finally, the works related to Topic 25 make many references to return migration.

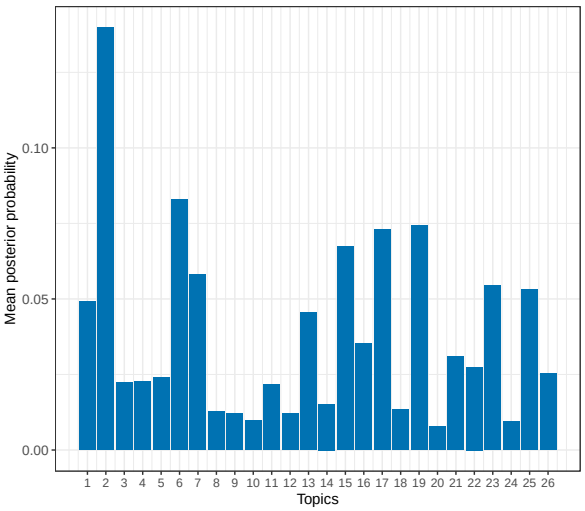
As defined in Section 4, to investigate how the field of *economics of migration* has evolved in terms of its topical composition, we used trend topic analysis with the topic proportions. First, we visualize the trends in the topics by aggregating the average posterior probabilities over three periods, as shown in Figure 3. The periods defined are 1991-2002, 2003-2013, and 2014-2024. We can see from Figure 3a that Topics 2, 6, and 19 were the three with the highest participations in the period 1991-2002, with Topic 2 representing a proportion of approximately 14%; this means that, if a scientific work on the field of *economics of migration* were to be written during this period, there's a probability of 14% that this work would be about Topic 2, which is a huge value in a pool of 26 topics. This is a results that we expected to see given the evidence provided in our literature review in Section 2. However, the second period already shows a convergence in our topic proportions, although Topic 2 remains the most prevalent. By our last period, between 2014 and 2024, 23 of our 26 topics have proportions below 5%, with the five topics with the highest participation in the literature being Topics 23, 10, 3, 7, and 18, the latter with a proportion of approximately 6.9%. Out of these, only Topic 3 is not exclusively about international migration, which again confirms our expectations.

The convergence of topic proportions we see in Figure 3 can be better visualized plotting the mean posterior probabilities in a continuous time, like in Figures 4 and 5. This convergence means in the mean posterior probability means that, if a scientific paper in the field of *economics of migration* were to be produced today, it would be equally likely to be written on any of our 26 topics, which is a result that is only possible because the corpus we use to fit our LDA topic model is currently fairly diverse. However, our trend topic analysis shows that this was not always the case, especially in the period 1991-2002.

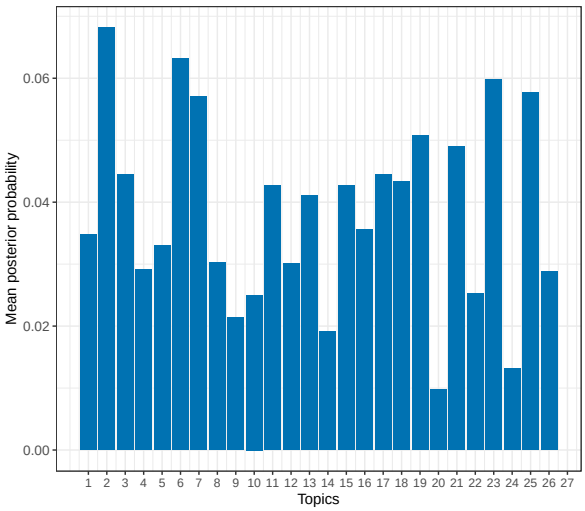
[30]The only reason why we do not classify this topic as being exclusively about international migration as well is because even though most remittances are sent abroad, we found cases that investigated domestic remittances.

Topics	Label
Topic 1	Internal Migration and Labor and Housing Markets
Topic 2	Economic Models of Migration and Public Finance
Topic 3	Selected Topics in Migration
Topic 4	Irregular Migration and Welfare Benefits
Topic 5	European Migration Policy and Governance
Topic 6	Economic Models of Migration and Labor Markets
Topic 7	Selected Topics in International Migration
Topic 8	Parental Migration and Children's Education
Topic 9	Business Economics and Expatriate Workers
Topic 10	Immigration, Social Perception, and Politics
Topic 11	International Migration, FDI, and Trade-Related Subjects
Topic 12	Migration-Related Diversities and Social Networks
Topic 13	Migration and Demography
Topic 14	Preferences, Intentions, and Attitudes Towards Migration
Topic 15	Regional and Spatial Migration
Topic 16	Rural Households and Rural-Urban Migration
Topic 17	Internal Migration and Urban-Related Issues
Topic 18	Macroeconomic Effects of Remittances
Topic 19	Quantitative Models and Research Methods of Migration
Topic 20	Humanitarian Migration
Topic 21	Remittances, Households Expenditure, and Development
Topic 22	Family and Gender Migration
Topic 23	Effects of Immigration on Labor Markets
Topic 24	Environmental and Climate-Related Migration
Topic 25	Brain Drain and Human Capital Formation
Topic 26	Migrant's Working Conditions

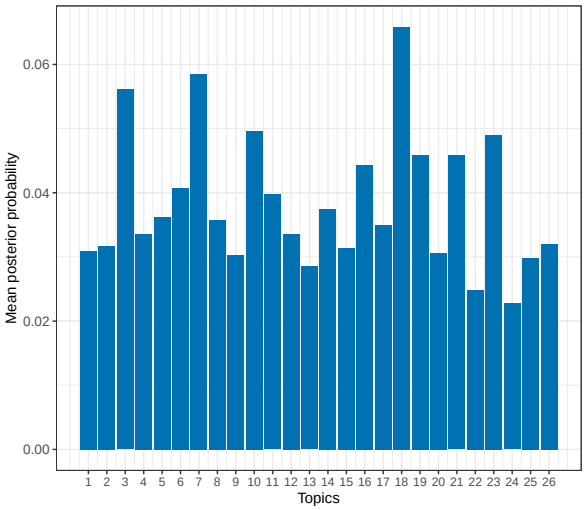
Table 5: Labels for the 26 topics of our LDA topic model.



(a) Mean posterior probabilities for the period 1991-2002.



(b) Mean posterior probabilities for the period 2003-2013.



(c) Mean posterior probabilities for the period 2014-2024.

Figure 3: Mean posterior probability of the 26 topics for three periods.

Therefore, we see this convergence as an increase in the diversity in the field's topical structure, which confirms our hypothesis.

6.2 Diversity indices

In order to validate the results of our trend topic analysis, we present two diversity indices, the Gini-Simpson and the Shannon-Wiener entropy indices, where the higher the values of both indices, the greater the diversity. As can be seen in Figures 6a and 6b, both of our indices show that diversity in our corpus increased considerably between 1991 and 2009, reaching high levels of diversity, which have remained relatively stable since then. This result contrasts with that found by (Pisarevskaya et al., 2020,

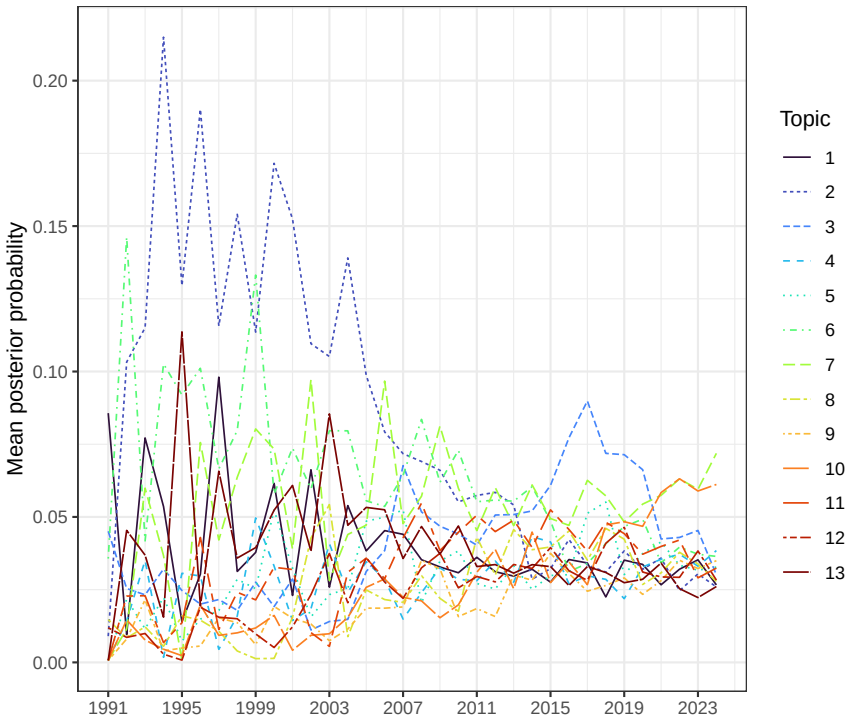


Figure 4: Trends of topics 1 to 13 over the period 1991-2024.

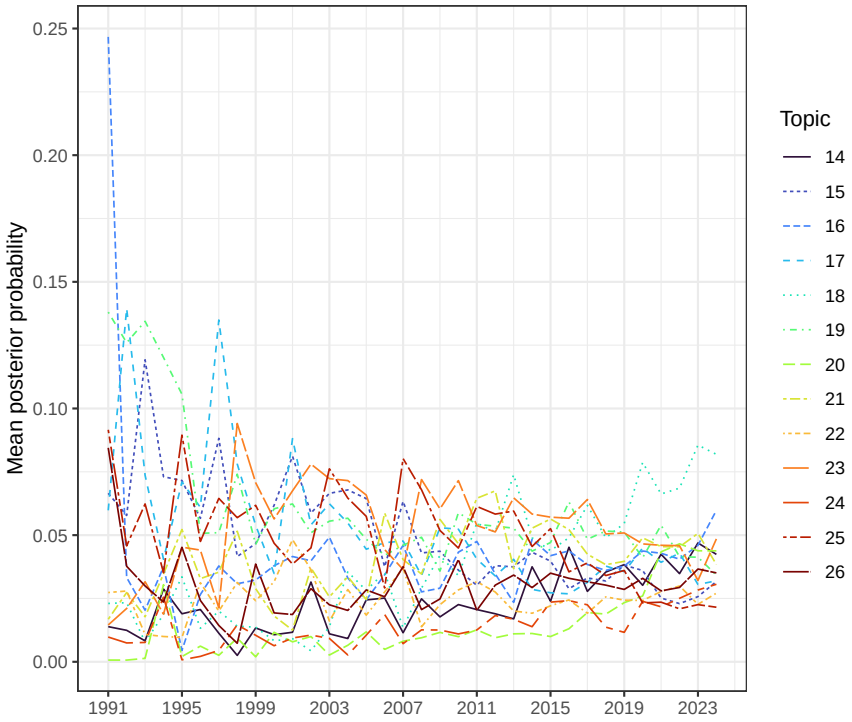
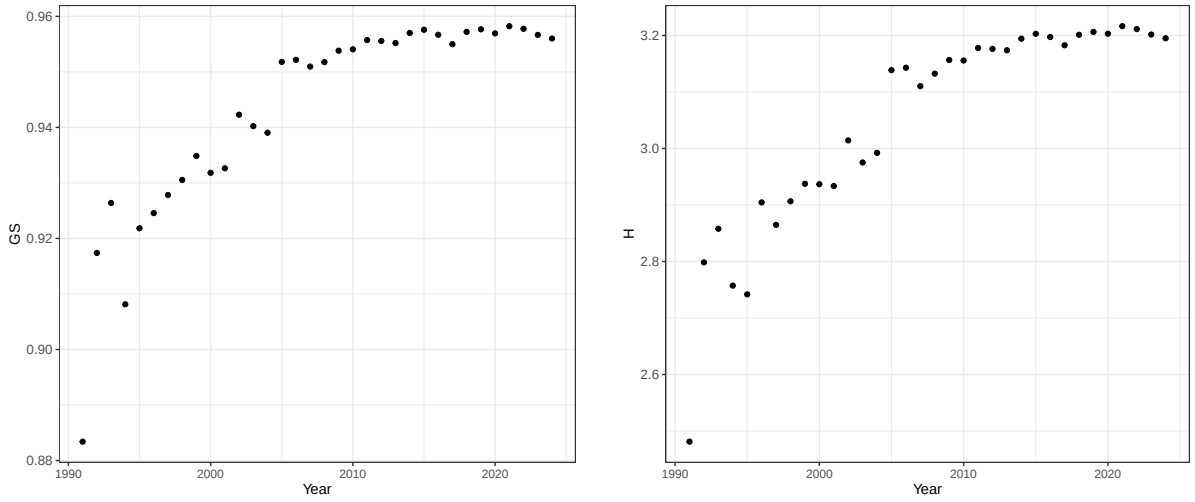


Figure 5: Trends of topics 14 to 26 over the period 1991-2024.

p. 467) in their analysis of diversity in the field of migration studies.



(a) Gini-Simpson entropy index.

(b) Shannon-Wiener entropy index.

Figure 6: Diversity measures.

7 Conclusion

References

Aria, M. and C. Cuccurullo (2017). *bibliometrix*: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics 11*(4), 959–975.

Arun, R., V. Suresh, C. E. Veni Madhavan, and M. N. Narasimha Murthy (2010). On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In *Lecture Notes in Computer Science*, pp. 391–402. Berlin, Heidelberg: Springer Berlin Heidelberg. ISSN: 0302-9743, 1611-3349.

Blei, D. M. and J. D. Lafferty (2009). Topic Models. In *Text mining: classification, clustering, and applications*, Chapman & Hall/CRC data mining and knowledge discovery series, pp. 71–93. Boca Raton, FL: CRC Press.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research 3*(null), 993–1022.

Borjas, G. J. (1987). Self-Selection and the Earnings of Immigrants. *The American Economic Review 77*(4), 531–553. Publisher: American Economic Association.

Borjas, G. J. (1991). Immigration and Self-Selection. *NBER Chapters*, 29–76. Publisher: National Bureau of Economic Research, Inc.

Borjas, G. J. (1994). The Economics of Immigration. *Journal of Economic Literature 32*(4), 1667–1717. Publisher: American Economic Association.

Brettell, C. B. and J. F. Hollifield (2015). Introduction. In *Migration Theory: Talking across Disciplines* (3rd ed.)., pp. 344. New York: Routledge.

Cao, J., T. Xia, J. Li, Y. Zhang, and S. Tang (2009). A density-based method for adaptive LDA model selection. *Neurocomputing 72*(7-9), 1775–1781. Publisher: Elsevier BV.

Card, D. (2005). Is the New Immigration Really so Bad? *The Economic Journal 115*(507), F300–F323. Publisher: [Royal Economic Society, Wiley].

Castles, S., H. d. Haas, and M. J. Miller (2014). Migration in Europe since 1945. In *The age of migration: international population movements in the modern world* (Fifth edition ed.). New York: Guilford Press. OCLC: 846540484.

Chang, J., S. Gerrish, C. Wang, J. Boyd-graber, and D. Blei (2009). Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in Neural Information Processing Systems*, Volume 22. Curran Associates, Inc.

Chiswick, B. R. (1978). The Effect of Americanization on the Earnings of Foreign-born Men. *Journal of Political Economy 86*(5), 897–921. Publisher: University of Chicago Press.

Cohen, R. (1996). Introduction. In *Theories of Migration*, The International Library of Studies on Migration, pp. 544. Cheltenham: Edward Elgar Publishing Limited.

Cohen, R. (2008). *Global Diasporas: An Introduction* (2 ed.). London: Routledge.

Corden, W. M. and R. Findlay (1975). Urban Unemployment, Intersectoral Capital Mobility and Development Policy. *Economica 42*(165), 59–78. Publisher: [London School of Economics, Wiley, London School of Economics and Political Science, Suntory and Toyota International Centres for Economics and Related Disciplines].

Cushing, B. and J. Poot (2004). Crossing boundaries and borders: Regional science advances in migration modelling. *Papers in Regional Science 83*(1), 317–338.

DaVanzo, J. (1978). Does Unemployment Affect Migration? Evidence from Micro Data. *The Review of Economics and Statistics 60*(4), 504–514. Publisher: The MIT Press.

De Haas, H. (2021). A theory of migration: the aspirations-capabilities framework. *CMS 9*(1). Publisher: Springer Science and Business Media LLC.

Deveaud, R., E. SanJuan, and P. Bellot (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique 17*(1), 61–84. Publisher: Lavoisier.

Evans, J. A. and P. Aceves (2016). Machine Translation: Mining Text for Social Theory. *Annu. Rev. Sociol. 42*(1), 21–50. Publisher: Annual Reviews.

Fields, G. S. (1975). Rural-urban migration, urban unemployment and underemployment, and job-search activity in LDCs. *Journal of Development Economics 2*(2), 165–187.

Greenwood, M. J. (1975). Research on Internal Migration in the United States: A Survey. *Journal of Economic Literature 13*(2), 397–433. Publisher: American Economic Association.

Greenwood, M. J. (1985). Human Migration: Theory, Models, and Empirical Studies. *Journal of Regional Science 25*(4), 521–544. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9787.1985.tb00321.x.

Greenwood, M. J. (1997). Internal migration in developed countries. In *Handbook of Population and Family Economics*, Volume 1, pp. 647–720. Elsevier.

Greenwood, M. J. and G. L. Hunt (2003). The Early History Of Migration Research. *International Regional Science Review 26*(1), 3–37.

Greenwood, M. J. and J. M. McDowell (1986). The Factor Market Consequences of U.S. Immigration. *Journal of Economic Literature 24*(4), 1738–1772. Publisher: American Economic Association.

Griffiths, T. L. and M. Steyvers (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences 101*(suppl_1), 5228–5235. Publisher: Proceedings of the National Academy of Sciences.

Grossman, J. B. (1982). The Substitutability of Natives and Immigrants in Production. *The Review of Economics and Statistics 64*(4), 596–603. Publisher: MIT Press.

Grün, B. and K. Hornik (2011). topicmodels: An R Package for Fitting Topic Models. *J. Stat. Soft. 40*(13). Publisher: Foundation for Open Access Statistic.

Harris, J. R. and M. P. Todaro (1970). Migration, Unemployment and Development: A Two-Sector Analysis. *The American Economic Review 60*(1), 126–142. Publisher: American Economic Association.

Herzog, H. W., A. M. Schlottmann, and T. P. Boehm (1993). Migration as spatial job-search: a survey of empirical findings. *Reg Stud 27*(4), 327–340.

Hicks, J. R. (1932). *The Theory of Wages*. London: Palgrave Macmillan UK.

Jost, L. (2006). Entropy and diversity. *Oikos 113*(2), 363–375. _eprint: https://nsojournals.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2006.0030-1299.14714.x.

King, R. (2002). Towards a new map of European migration. *International Journal of Population Geography 8*(2), 89–106. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ijpg.246.

King, R. (2012). Theories and typologies of migration: an overview and a primer. report, University of Sussex. ISSN: 2344-6544.

King, R. and R. Skeldon (2010). 'Mind the Gap!' Integrating Approaches to Internal and International Migration. *Journal of Ethnic and Migration Studies 36*(10), 1619–1646. Publisher: Informa UK Limited.

LaLonde, R. J. and R. H. Topel (1991a). Immigrants in the American Labor Market: Quality, Assimilation, and Distributional Effects. *The American Economic Review 81*(2), 297–302. Publisher: American Economic Association.

LaLonde, R. J. and R. H. Topel (1991b). Labor Market Adjustments to Increased Immigration. *NBER Chapters*, 167–199. Publisher: National Bureau of Economic Research, Inc.

Lalonde, R. J. and R. H. Topel (1997). Economic impact of international migration and the economic performance of migrants. In *Handbook of Population and Family Economics*, Volume 1, pp. 799–850. Elsevier.

Lee, E. S. (1966). A theory of migration. *Demography 3*(1), 47–57. Publisher: Duke University Press.

Levy, N., A. Pisarevskaya, and P. Scholten (2020). Between fragmentation and institutionalisation: the rise of migration studies as a research field. *Comparative Migration Studies 8*(1), 24.

Lewis, W. A. (1954). Economic Development with Unlimited Supplies of Labour. *The Manchester School 22*(2), 139–191. Publisher: Wiley.

Lucas, R. (1997). Internal migration in developing countries. In *Handbook of Population and Family Economics*, pp. 721–798. Elsevier. ISSN: 1574-003X.

Mabogunje, A. L. (1970). Systems Approach to a Theory of Rural-Urban Migration. *Geographical Analysis 2*(1), 1–18. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1538-4632.1970.tb00140.x.

Martiniello, M. and J. Rath (2012). An Introduction to International Migration Studies: European Perspectives. In *An Introduction to International Migration Studies*, pp. 15–22. Amsterdam University Press.

Massey, D. S., J. Arango, G. Hugo, A. Kouaouci, A. Pellegrino, and J. E. Taylor (1993). Theories of International Migration: A Review and Appraisal. *Population and Development Review 19*(3), 431. Publisher: JSTOR.

Massey, D. S., J. Arango, G. Hugo, A. Kouaouci, A. Pellegrino, and J. E. Taylor (1994). An Evaluation of International Migration Theory: The North American Case. *Population and Development Review 20*(4), 699. Publisher: JSTOR.

Mincer, J. (1978). Family Migration Decisions. *Journal of Political Economy 86*(5), 749–773. Publisher: University of Chicago Press.

Navratil, F. J. and J. J. Doyle (1977). The Socioeconomic Determinants of Migration and the Level of Aggregation. *Southern Economic Journal 43*(4), 1547–1559. Publisher: Southern Economic Association.

Pedraza-Bailey, S. (1990). Immigration Research: A Conceptual Map. *Social Science History 14*(1), 43.

Piore, M. J. (1979). *Birds of Passage: Migrant Labor and Industrial Societies*. Cambridge: Cambridge University Press.

Pisarevskaya, A., N. Levy, P. Scholten, and J. Jansen (2020). Mapping migration studies: An empirical analysis of the coming of age of a research field. *Migration Studies 8*(3), 455–481.

Ponweiser, M. (2012). Latent Dirichlet Allocation in R. Medium: application/pdf.

Portes, A., S. A. McLeod, and R. N. Parker (1978). Immigrant Aspirations. *Sociology of Education 51*(4), 241–260. Publisher: [Sage Publications, Inc., American Sociological Association].

Ranis, G. and J. C. H. Fei (1961). A Theory of Economic Development. *The American Economic Review 51*(4), 533–565. Publisher: American Economic Association.

Ravenstein, E. G. (1885). The Laws of Migration. *Journal of the Statistical Society of London 48*(2), 167–235. Publisher: [Royal Statistical Society, Oxford University Press].

Ravenstein, E. G. (1889). The Laws of Migration. *Journal of the Royal Statistical Society 52*(2), 241–305. Publisher: [Oxford University Press, Royal Statistical Society].

Roy, A. D. (1951). Some Thoughts on the Distribution of Earnings. *Oxford Economic Papers 3*(2), 135–146. Publisher: Oxford University Press.

Sandell, S. H. (1977). Women and the Economics of Family Migration. *The Review of Economics and Statistics 59*(4), 406–414. Publisher: The MIT Press.

Schiller, N. G., L. Basch, and C. Blanc-Szanton (1992). Transnationalism: A New Analytic Framework for Understanding Migration. *Annals of the New York Academy of Sciences 645*(1), 1–24. _eprint: https://nyaspubs.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1749-6632.1992.tb33484.x.

Scholten, P., A. Pisarevskaya, and N. Levy (2022). An Introduction to Migration Studies: The Rise and Coming of Age of a Research Field. In P. Scholten (Ed.), *Introduction to Migration Studies: An Interactive Guide to the Literatures on Migration and Diversity*, pp. 3–24. Cham: Springer International Publishing.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal 27*(3), 379–423. Publisher: Institute of Electrical and Electronics Engineers (IEEE).

Shields, G. M. and M. P. Shields (1989). The Emergence of Migration Theory and a Suggested New Direction. *Journal of Economic Surveys 3*(4), 277–304. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-6419.1989.tb00072.x.

Simpson, E. H. (1949). Measurement of Diversity. *Nature 163*(4148), 688–688. Publisher: Nature Publishing Group.

Sjaastad, L. A. (1962). The Costs and Returns of Human Migration. *Journal of Political Economy 70*(5), 80–93. Publisher: The University of Chicago Press.

Stark, O. (1991). *The migration of labor*. Cambridge (Mass.) Oxford: B. Blackwell.

Stark, O. and D. E. Bloom (1985). The New Economics of Labor Migration. *The American Economic Review 75*(2), 173–178. Publisher: American Economic Association.

Stark, O. and D. Levhari (1982). On Migration and Risk in LDCs. *Economic Development and Cultural Change 31*(1), 191–196. Publisher: University of Chicago Press.

Stark, O. and R. E. B. Lucas (1988). Migration, Remittances, and the Family. *Economic Development and Cultural Change 36*(3), 465–481. Publisher: University of Chicago Press.

Stewart, J. Q. (1947). Empirical Mathematical Rules concerning the Distribution and Equilibrium of Population. *Geographical Review 37*(3), 461–485. Publisher: Taylor & Francis, Ltd.

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. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-2435.00066.

Thomas, W. I. and F. Znaniecki (1927). *The Polish Peasant in Europe and America: A Classic Work in Immigration History*. University of Illinois Press.

Todaro, M. P. (1969). A Model of Labor Migration and Urban Unemployment in Less Developed Countries. *The American Economic Review 59*(1), 138–148. Publisher: American Economic Association.

Todaro, M. P. (1980). Internal Migration in Developing Countries: A Survey. In R. Easterlin (Ed.), *Population and economic change in developing countries*. Chicago: Univ. of Chicago Pr. OCLC: 1052756278.

Zipf, G. K. (1942). The Unity of Nature, Least-Action, and Natural Social Science. *Sociometry 5*(1), 48–62. Publisher: [American Sociological Association, Sage Publications, Inc.].

Zipf, G. K. (1946). The P1 P2/D Hypothesis: On the Intercity Movement of Persons. *American Sociological Review 11*(6), 677–686. Publisher: [American Sociological Association, Sage Publications, Inc.].

Appendix

A Overview of migration studies

The study of migration has been of interest for many areas, such as economics, sociology, geography, demography, and anthropology, among others (Cohen, 1996; Brettell and Hollifield, 2015). This characteristic makes migration studies a research field that builds from knowledge and insights from multiple disciplines, and draws on different methods, theories, and definitions (Scholten et al., 2022; Brettell and Hollifield, 2015; King, 2012). Ever since the field formally emerged in the late XlX and early XX centuries, scholars have put effort to define and structure it in a uniform and cohesive way, but without forgetting its plurality and diversity (Scholten et al., 2022). This appendix does not aim to address all the nuances of this overly complex field, but rather seeks to highlight the main aspects of its development from a historical perspective, with the aim of helping us contextualize the study of the *economics of migration*, which is an integral part of the field of migration studies. Largely based on (Scholten et al., 2022, p. 9-21)[31], we identified that the development of migration studies as a research field can be roughly divided into three periods.

The first period is defined from 1885 until around the 1950s, with the appearance of the first scientific works and the increasing interest of the academic community on the subject. Migration studies as a research field emerged in the late 19th and early 20th centuries, stimulated by rapid urbanization, which was largely caused by rural-urban migration, in industrialized parts of the world, such as the United States (US), the United Kingdom (UK), and Western Europe. The first two works that most likely mark the starting point of the field of migration studies are Raventein's *"The Laws of Migration"* (Ravenstein, 1885, 1889), considered by many scholars to be the first scientific study on internal migration (Scholten et al., 2022; King, 2012; Greenwood and Hunt, 2003; Cohen, 1996), and Thomas and Znaniecki's work *"The Polish Peasant in Europe and America"*, which is probably the work that started the studies on international migration (Thomas and Znaniecki, 1927; Scholten et al., 2022). Nevertheless, most studies in this early period of migration research, like the ones mentioned, lacked formality and adequate techniques. That said, it is also true that there were no appropriate theories or data available (Greenwood, 1975).

Among the first disciplines to be interested in migration research, sociology, demography, and geography stand out. In fact, sociologists, geographers, and demographers had already made important contributions to the field when economists began to take an interest in migration. As discussed in Section 2, economists only began to take a greater interest in migration after the Depression of the 1930s. Coincidentally or not, it was during the 1930s that migration research really picked up, including with

[31] Their work combines the bibliometric analyses carried out by Levy et al. (2020) and Pisarevskaya et al. (2020), with "a strong historical awareness of how the field has developed and how specific topics, concepts, and methods have emerged" (Scholten et al., 2022, p. 5).

the emergence of more formal models. At this moment, migration research was mainly focused on internal migration, especially rural-to-urban movements fuelled by urban growth and urbanization in the so-called developed, or industrialized, countries (Greenwood and Hunt, 2003).

The second period happens between around 1950s and 1980s, and it's characterized by the field's formalization and expansion. Many important theoretical contributions to the field were made during this period (Massey et al., 1993; De Haas, 2021), like Lee's *theory of migration* (Lee, 1966), Harris and Todaro's *neoclassical migration theory* (Todaro, 1969; Harris and Todaro, 1970), Mabogunje's *migration systems theory* (Mabogunje, 1970), and Piore's *dual labour-market theory* (Piore, 1979). As pointed out by King and Skeldon (2010), earlier theorisations focused on internal migration, being adapted to analyze international migration later on. In addition, even though the interest in more formal models had already appeared in the 1920s and 1930s, it was only during the 1960s that the field acquired a more formal tone, especially with the use of modified gravity models (Greenwood and Hunt, 2003). Regarding the field's expansion, it is reflected by the increase in both the quantity of articles published yearly in migration studies, which jumped from just under 350 in 1975 to almost 750 in 1989 (Levy et al., 2020; Scholten et al., 2022), and journals focused on migration and migration-related diversity, which can be seen as the first steps towards the field's institutionalization; according to Pisarevskaya et al. (2020), the number of journals grew from zero to 15 in the period between 1959 and 1988.

Two of the main changes that occurred in this period in relation to the approach to migration studies were: (i) the shift towards a greater focus on international migration, in detriment of internal migration (Scholten et al., 2022), driven by the population and economic dynamics of the post-World War II era (Castles et al., 2014), (ii) and a greater attention to issues of ethnic and racial relations involving immigrants, prompted by the civil rights movements in the US in the 1950s and 1960s[32] (Portes et al., 1978; Pedraza-Bailey, 1990); this change was part of a context of greater interest in questions of assimilation and integration of these immigrants in the host societies (Scholten et al., 2022).

Another important characteristic of the second period is the increase in the number of disciplines interested in the topic of migration, which led to the development of the multidisciplinary field as we know it today. However, the disciplines were mostly working isolated from one another, within their own conceptual frameworks and methodologies (Levy et al., 2020; Scholten et al., 2022). Thus, although by the end of the 1980s the research field of migration studies was indeed multidisciplinary, it was definitely not interdisciplinary[33].

The third and final period in our historical analysis is what Castles et al. (2014) have called the *Age of Migration*, which begins in the 1990s and continues to the present day[34], during which migration

[32]This is just another example of how academically dominant country, or region, such as the US, UK and Europe, have the power to influence, and ultimately dictate, the research agenda.

[33]Regarding the concepts of multidisciplinary and interdisciplinary, we follow King's way of thinking, which stated that "cross- and multidisciplinary implies two or more different disciplines working alongside each other, in parallel so to speak; interdisciplinarity implies a fusing of disciplines in an integrated analysis" (King, 2012, p. 10).

[34]Levy and colleagues seem to indicate a new developmental phase by writing that "since 2005, a new age of migration studies has emerged" (Levy et al., 2020, p. 17). Therefore, what we are defining as the third period would be broken down in two, one from 1990 to 2005, and the other from 2005 to the present. In any case, we decided to stick to our initial

has become more diverse, globalized, politicized, and feminized[35] (King, 2012; Castles et al., 2014). The defining characteristic of the *Age of Migration* is the political prominence of the topic of migration[36]. As Castles and colleagues have stated, what is distinctive in recent migratory movements "is their global scope, their centrality to domestic and international politics and their considerable economic and social consequences" (Castles et al., 2014, p. 6). The unprecedented speed of communications and transportation—combined with the social, economic, and political transformations that the world underwent after the end of the Cold War—have intensified globalization impulses, and, consequently, the movement of people across borders. In this context, politicians and policy-makers have seen this phenomenon as a threat to the sovereignty of states and their ability to regulate and control immigrant flows, which has contributed to the rise of anti-immigration sentiment (Martiniello and Rath, 2012; Castles et al., 2014).

Regarding the scientific production in the third period, the field has seen a notable acceleration. If the number of migration studies published in 1989 had been slightly under 750, this number increased to 1,000 in 1995, and reached the remarkable figure of 3,000 in 2017 (Levy et al., 2020; Scholten et al., 2022). A noteworthy methodological change in the field since the early 1990s noted by King (2012) was the shift from quantitative to qualitative research, which was influenced by new approaches, insights, and perspectives arising mainly from qualitative sociology, anthropology, human geography and cultural studies. This shift reflected the widespread moment in the social sciences that King called "cultural turn", which according to him "did not so much re-make theories of the causes of migration as enrich our understanding of the migrant experience" (King, 2012, p. 25). As pointed out by Scholten et al. (2022), this meant that migration studies were shifting from migration to migrants, that is, there was an increase in the importance of "research on the subjective experiences of migrants, perceptions of migrants' identity and belonging, as well as attention to the cultural (super)diversity of societies" (Scholten et al., 2022, p. 21).

According to King, "the 'Age of Migration' has seen a proliferation of new types of migration and international mobility which form important elements of the increasingly complex global map of population movements" (King, 2012, p. 9)], such as *transnationalism* (Schiller et al., 1992) and *diasporas* (Cohen, 2008), which were added to the conceptual frameworks of migration studies (King, 2012). The typological discussion of these types of migration and migrants has been a source of controversy because some definitions involve complications related to the way migration unfolds in time and space, and also because of many conceptual differences found in the social sciences[37]. As a matter of fact, the differences between the social sciences—not only in terms of concepts, but also theories and methodologies—in their

division.

[35] Castles et al. (2014) actually point out to the fact that women have played an increasing role in labor migration since the 1960s, which falls into our second period. In any case, after 1990s, women started forming the majority in multiple movements of workers.

[36] As stated by King and Skeldon (2010), during this time migration came to be understood as international migration. For Scholten and colleagues, this was due to "a move beyond a strong focus on the national dimension of migration and diversities" (Scholten et al., 2022, p. 16).

[37] For a more detailed discussion, see King (2012), p. 7-9.

approach to migration studies began to bother some scholars, who observed the problem of a lack of interdisciplinarity in the field, as eloquently expressed by Massey and colleagues:

> Social scientists do not approach the study of immigration from a shared paradigm, but from a variety of competing theoretical viewpoints fragmented across disciplines, regions, and ideologies. As a result, research on the subject tends to be narrow, often inefficient, and characterized by duplication, miscommunication, reinvention, and bickering about fundamentals and terminology (Massey et al., 1994, p. 700-701).

Those scholars identified that solving all the differences and challenges within the field would require more integration from the disciplines. King (2002) and King (2012) called for an "interdisciplinary synthesis", while Brettell and Hollifield (2015) proposed a "talk across disciplines" in an effort to bring them closer together. Moving away from the sphere of mere rhetoric, important contributions were made to seek a convergence in terminology. For instance, Cohen presented his nine *dyads*[38], which according to him were "building blocks to a theory of migration" (Cohen, 1996, p. xvi). King (2002) called for a deconstruction of the traditional dichotomies presented by Cohen, arguing that migration's motivations and forms have become much more diverse. He offered new geographies and typologies of (international) migration applied to the European context, while stressing the need for a more integrated approach that "brings together and integrates a range of perspectives, frameworks, theoretical stances and methodologies in order to study migration (or the various forms of migration)" (King, 2002, p. 90).

However, some studies provide quantitative evidence that migration studies may have been more interdisciplinary than it is perceived to be. The investigation into the field's topical composition performed by Pisarevskaya et al. (2020) showed that, overall, the field's topics have been well-connected at least since the late 1980s, which could suggest the existence of a shared conceptual and theoretical language. Another study conducted by Levy, Pisarevskaya, and Scholten found a growing coherence between the epistemic communities (or co-citation networks), which, according to them, suggests that "migration studies evolved from a multidisciplinary field (with various but very distinct disciplines) to a more interdisciplinary field (with various and linked disciplines)" (Levy et al., 2020, p. 22-23). In fact, Levy et al. (2020) concluded that not only the field has become more interdisciplinary, but it has also increased in terms of internationalization—with a growing number of countries collaborating in the scientific production, even though if still in an uneven way—and self-referentiality, which is evidence that the research field of migration studies has institutionalized. Another factor that points to the field's institutionalization is the growing number of journals, which jumped from 15 in 1988 to 45 in 2018, accompanying the accelerated growth in scientific production.

[38] Also called dichotomies or binaries (King, 2012).

B Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA)[39] was first proposed by Blei et al. (2003), and it's build upon the assumption that documents exhibit multiple topics in different proportions, where a topic is a probabilistic distribution over the words in a fixed vocabulary. The basic unit of an LDA model is a *word*, or *term*, w, which is an item from the fixed vocabulary V; a *document* $\mathbf{w} = (w_1, \cdots, w_n, \cdots, w_N)$ is a sequence of N words[40], where w_n is the nth word in the sequence; and a *corpus* $D = \{\mathbf{w}_1, \cdots, \mathbf{w}_m, \cdots, \mathbf{w}_M\}$ is a collection of M documents, where $\mathbf{w}_m$ is the mth document in the collection.

Other elements of the LDA model are: the number of topics $K = \{1, \cdots, k\}$; the parameters α and η, which follow Dirichlet distributions; $\beta_k \sim \text{Dirichlet}(\eta)$, which is defined for a given η and indicates the probability of a word occurring given a topic k, and follows a Dirichlet distribution; $\theta \sim \text{Dirichlet}(\alpha)$, which is defined for a given α and gives the topic proportions per document $\mathbf{w}_m$, and that also follows a Dirichlet distribution; and the topic $z_n \sim \text{Mult}(\theta)$, which is a topic assigned for a given θ and follows a Multinomial distribution (Blei et al., 2003; Blei and Lafferty, 2009; Grün and Hornik, 2011). An illustration of how an LDA model works is presented in Figure 7. In a simplified way, LDA applies the following probabilistic generative process (Blei and Lafferty, 2009; Grün and Hornik, 2011; Ponweiser, 2012):

1. For each topic k: draw a word distribution β_k;
2. For each document $\mathbf{w}_m$ in corpus D:
   - 2.1 Draw a topic distribution θ;
   - 2.2 For each of the N words w_n ;
     - i. Choose a topic z_n;
     - ii. Choose a word w_n from a multinomial probability $p(w_n|z_n, \beta_k)$, conditioned on both the topic z_n from step (i) and the word distribution β_k from step (1).

C Per-topic word probabilities

[39] Regarding terminology and notation, we follow Blei et al. (2003).

[40] LDA is based on the bag-of-words assumption, according to which words in a document are exchangeable, and so the order of words in a document can be neglected (Blei et al., 2003).

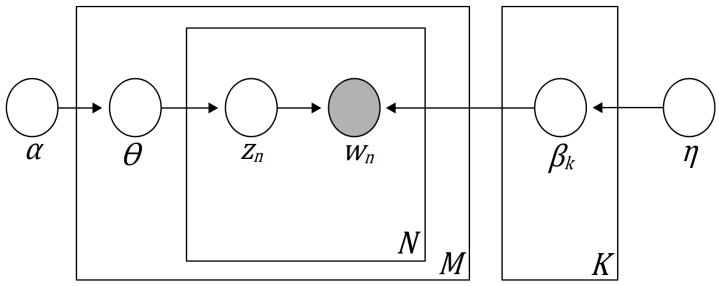


Figure 7: A graphical model representation of LDA, adapted from Blei et al. (2003) and Blei and Lafferty (2009).

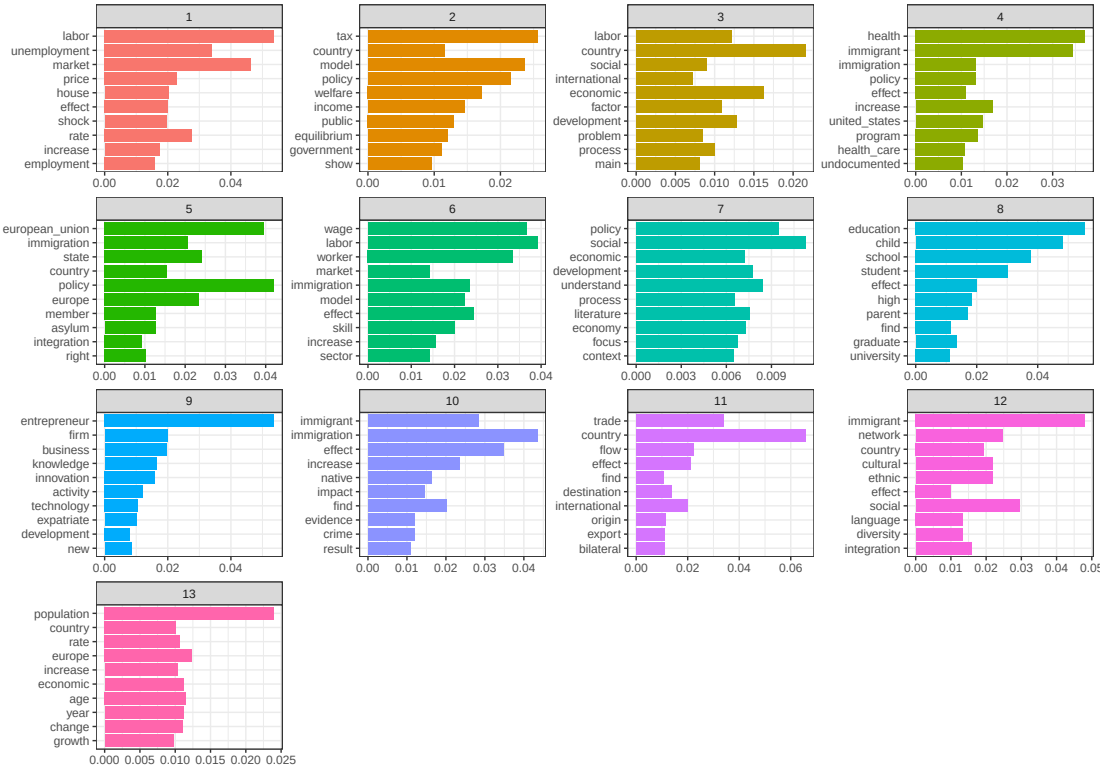


Figure 8: Posterior probabilities of words belonging to topics. Ten most likely words over topics 1 to 13.

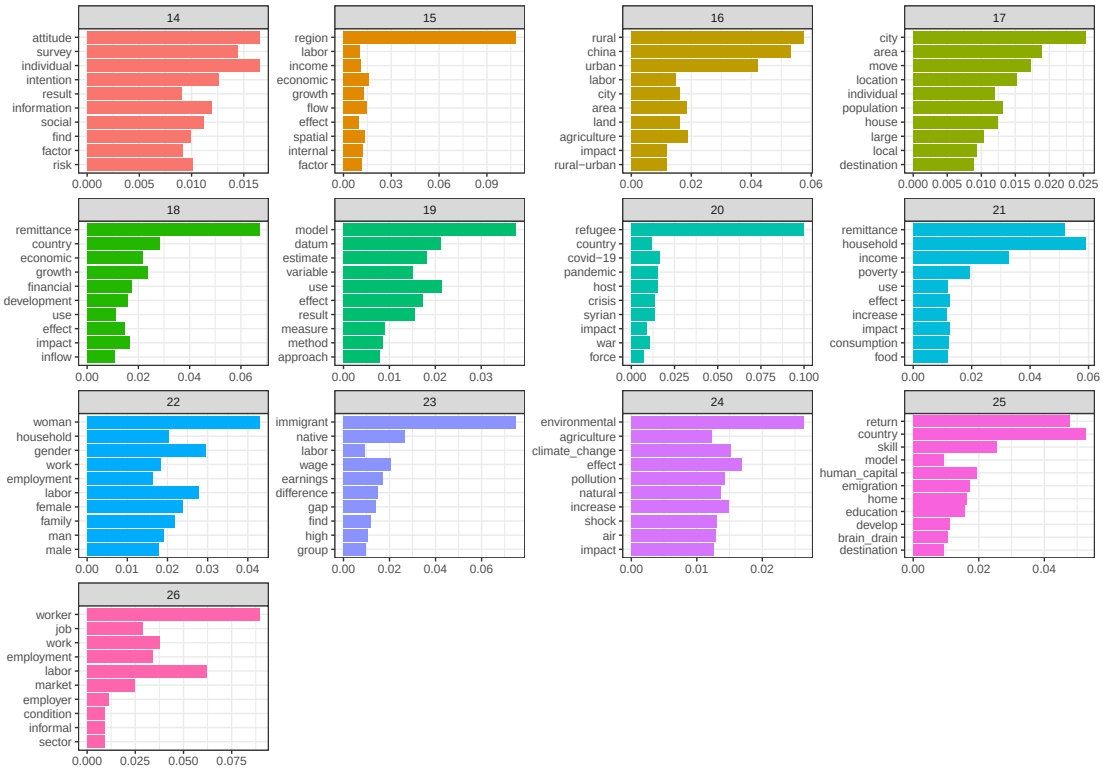


Figure 9: Posterior probabilities of words belonging to topics. Ten most likely words over topics 14 to 26.