SERIEs (2016) 7:221–255
DOI 10.1007/s13209-016-0138-5

SERIEs Journal of the Spanish Economic Association

ORIGINAL ARTICLE

Understanding international migration: evidence from a new dataset of bilateral stocks (1960–2000)

Joan Llull[1,2,3]

Received: 19 September 2014 / Accepted: 10 January 2016 / Published online: 13 February 2016

Abstract In this paper I present a new database of bilateral migrant stocks and I provide new evidence on the determinants of international migration. The new Census-based data are obtained from National Statistical Offices of 24 OECD countries, and they cover the total stock of immigrants in each destination country for 1960–2000, including 188 countries of origin, sometimes in grouped categories. For each census, I keep grouped categories in a raw manner, without making imputations to specific origin countries. In the empirical analysis, I give an explicit treatment to these grouped categories. Results present strong evidence of heterogeneous effects of income gains on migration prospects depending on distance. For example, a 1000$ increase in US income per capita increases the stock of Mexican immigrants in the country by

I am indebted to Manuel Arellano for his constant encouragement and advice. I wish to thank Stéphane Bonhomme, Juan J. Dolado, Manu García-Santana, Jesús Fernández-Huertas, Nezih Guner, Claudio Michelacci, Pedro Mira, Francesc Ortega, Roberto Ramos, Pedro Rey-Biel, Rob Sauer, Jim Walker, the editors (Victor Aguirregabiria and Manuel Bagues), two anonymous referees, and seminar participants at CEMFI for helpful comments and discussions. Financial support from European Research Council through Starting Grant n. 263600, and the Spanish Ministry of Economy and Competitiveness, through Grant ECO2014-59056-JIN, and Severo Ochoa Programme for Centers of Excellence in R&D (SEV-2011-0075) is gratefully acknowledged. I am thankful to the help from several statistical offices from many OECD countries in providing me with the data, and to Hugo Ferradáns and Marta Aguilera, who provided excellent research assistance.

✉ Joan Llull
joan.llull@movebarcelona.eu
http://pareto.uab.cat/jllull

1 Departament d'Economia i Història Econòmica, Facultat d' Economia, Universitat Autònoma de Barcelona, Edifici B, Campus de Bellaterra, Cerdanyola del Vallès, 08193 Barcelona, Spain

2 MOVE, Barcelona, Spain

3 Barcelona GSE, Barcelona, Spain

a percentage 2.6 times larger than the percentage increase in the stock of Chinese (8 vs. 3.1 %).

Keywords International migration · Data collection · Grouped data

JEL Classification F22 · J61 · O15

1 Introduction

International migration has increased dramatically in recent decades. Understanding the determinants of the movement of workers across international borders is crucial for immigration policy design. This paper aims to enhance our knowledge about these determinants by presenting new data on bilateral migrant stocks, a new treatment of those data in the empirical analysis, and new empirical evidence on the determinants of international migration.

To create the new database, I collected data on international migrant stocks by country of origin from National Statistical Offices of the 24 richest OECD countries. This dataset includes bilateral stocks of immigrants from 188 countries of origin into these 24 destination countries over the period 1960 to 2000. The data come from Census records at these destination countries. Given this, it covers the total amount of immigrants living in the country. Importantly, because the data sometimes appear in grouped categories, I keep track of these groups in a raw manner, without making imputations to specific countries of origin.[1] This is important because imputations, dropping grouped observations, and/or counting grouped observations as zeros may lead to important biases in the estimates.

Empirically, this paper makes two contributions. First, it gives explicit treatment to these grouped data in standard gravity regressions. Second, it presents evidence on the existence of heterogeneous effects of income gains on migration prospects depending on distance. According to a static model—the approach which mostly followed by the literature—when individuals decide whether to migrate to another country, they base their decision on *net* income gains from migration, i.e. the differential in expected wages between the two countries net of (one time) moving costs.[2] From a dynamic point of view, however, individuals may care about moving costs (distance in particular) even after having migrated. Large moving costs may reduce their flexibility to move back and forth to their home country as a consequence of income shocks;[3] and, if individuals dislike living far away from home, they may require a compensating wage

[1] These grouped categories are either residual categories or other types of groups of countries, like continents or subcontinents of origin, or former countries that later were dissolved into smaller countries, like USSR, Yugoslavia, Czechoslovakia, Rhodesia, and so on.

[2] Examples of papers using this approach include Borjas (1987), Borjas and Bratsberg (1996), Karemera et al. (2000), Chiquiar and Hanson (2005), Clark et al. (2007), Pedersen et al. (2008), Mayda (2010), and Grogger and Hanson (2011) among many others. Recent papers like Bertoli and Fernández-Huertas Moraga (2013), Bertoli et al. (2013), or Ortega and Peri (2013) estimate nested logit models that allow for different elasticities across destinations.

[3] Kennan and Walker (2011) and Lessem (2013) argue that migration is a dynamic decision, and that repeated and return migration are important in the data.

differential for living abroad that might be increasing in distance. Forward looking individuals will take these two factors into account when deciding whether to migrate in the first place. As a result, the effect of income gains on moving prospects (net of the initial moving cost) may be heterogeneous depending on distance: individuals from countries away from home would be less reactive to income fluctuations compared to individuals from closer countries. Results suggest that these heterogeneities are indeed very important. For example, a 1000$ increase in US income per capita increases the stock of Mexican immigrants in the US by a percentage that is 2.6 times larger than the percentage increase in the stock of Chinese immigrants. In other words, the effect of income on log migrant stocks is 2.6 times larger for Mexico compared to China (8 vs. 3.1 %), given that Beijing is around 2.6 times as far from Washington DC as Mexico City is. This differs from the standard gravity equation, which would predict linear effects of income gains on log migrant stocks (Beine et al. 2015). This result is relevant for immigration policy design. For example, a pull-driven immigration shock (i.e. positive income shock) may imply significant changes in the composition of immigrant population in terms of nationalities. Similarly, a negative shock to a developing country may have a much larger effect for neighboring countries than previous estimates in the literature suggest; this larger effect suggests that destination countries may want to favor neighboring countries in development assistance policies if they are interested in reducing immigrant inflows.

Collecting data on bilateral migration is, in general, a difficult task. Reliability of statistics from origin countries is low because it is difficult to keep track of the people who leave the country. Data from destination countries is more accurate. The lack of comparable cross-destination country bilateral data led many papers in the literature to follow a single destination country over time (e.g. Borjas and Bratsberg 1996; Karemera et al. 2000; Clark et al. 2007; Bertoli and Fernández-Huertas Moraga 2013). More recently, researchers and institutions have put some effort in gathering comparable bilateral migration data across destination countries. Pedersen et al. (2008) and Mayda (2010) are the first papers using cross-destination country panel data on bilateral inflows to analyze the effect of income gains and moving costs on international migration. Mayda (2010) uses a database from OECD on annual legal inflows of workers by country of origin; she uses these data to investigate the determinants of migration inflows into 14 OECD countries between 1980 and 1995. Pedersen et al. (2008) produce a similar database collecting data on issues of residence and work permits from National Statistical Offices from 1989 to 2000. They use these data to look at the effects of networks and welfare benefits on international migration. These two databases have recently been expanded by Ortega and Peri (2013) and Adserà and Pytliková (2012) respectively.[4] The four databases contain information on inflows of immigrants and, with a lower accuracy, net flows. They are based on the number of issues of residence and work permits, which is likely to produce a severe underestimation the real numbers due to illegal migration. And, acknowledged by the authors, they have an important amount of missing data and incorrect zero values (for

[4] The database in Ortega and Peri (2013) includes information for 15 destination countries and 120 countries of origin for the period 1980–2006. Adserà and Pytliková (2012) cover the period 1980 to 2009 for 30 destination countries and many countries of origin (with missing data).

countries with relatively small flows), covering, as a result, a limited fraction of total inflows (Mayda 2010, pp. 1258–1259).

Similarly to what I do in this paper, Docquier and Marfouk (2006) and Docquier et al. (2009) collect Census-based data. The aim of their databases is to gather information on stocks of immigrants by educational level, and, for this reason, they only cover two census dates, 1990 and 2000. Two papers use these data to analyze the determinants of international migration. Grogger and Hanson (2011) use them to the analyze the determinants of scale and composition of migration flows. Ortega and Peri (2014b) combine these two years of data on stocks with the OECD database on annual legal inflows used in Mayda (2010) to extrapolate stocks back to 1980 and analyze the determinants of migration flows.

Contemporaneously to this paper, a few additional datasets appeared. Özden et al. (2011) is the most similar. These authors collect bilateral stock data for the same period and from similar sources. The key difference with the current dataset is the treatment of data when bilateral information is not available. When this happens, which is often the result of grouping of data (residual categories, aggregations of countries,…), these authors try to recover the bilateral information by means of an array of different imputations. Conversely, I keep these grouping in a raw manner, giving it a specific treatment in the empirical analysis. Given the similarity, I draw some comparisons with this dataset below. The other three datasets are: United Nations (2013), which provides similar information for years 1990, 2000, 2010, and 2013; Brücker et al. (2013), who add the educational and gender dimension for the period 1980 to 2010; and, Abel and Sander (2014), who estimate inflows and outflows out of the stock data for 1990–2010.

The rest of the paper is organized as follows. Section 2 presents the database. Section 3 introduces the econometric model and explains the implications of grouped data in terms of identification of fixed effects. Section 4 shows estimation results. And Sect. 5 concludes.

2 Data

2.1 A new database on bilateral migrant stocks (1960–2000)

In this paper, I collect data from National Statistical Offices of 24 OECD countries.[5] The dataset contains stocks of immigrants by country of origin from 1960 to 2000. Data are based on destination countries' Censuses.[6] From each Census, I collect data on the stock of immigrants by country of birth or country of nationality. The dataset contains information on stocks of immigrants from 188 countries of origin—sometimes in grouped categories—into each of the destination countries.[7] Although some desti-

[5] These include: Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Italy, Japan, Korea (Rep.), Luxembourg, The Netherlands, New Zealand, Norway, Portugal, Spain, Sweden, Switzerland, United Kingdom and United States.

[6] Nordic countries replaced Censuses for continuous population registers during 1980s.

[7] Source countries include all Member States of United Nations except Andorra, Liechtenstein, Monaco, Myanmar, Marshall Islands, Nauru, San Marino, Timor-Leste, and Tuvalu (none of them are available in Penn World Tables). Additionally, it includes the dependent territories of Taiwan, Macao, Hong Kong,

nation countries carry a census every five years, most of them do it every 10 years, so data is presented at a 10-year frequency. Hence, the database is well suited for looking at long-run effects.

There are some comparability issues that are worth mentioning. First, similar to existing datasets in the literature, the definition of immigrant is different across countries. Some countries define immigrants on the basis of the place of birth whereas others do it based on nationality. This might affect the comparability of stocks across destination countries, but changes over time are reasonably comparable.[8] Second, census dates vary across destination countries—roughly a half of them are carried in even years (1960, 1970,…) and the other half in odd years (1961, 1971,…).[9] Dates are generally consistent for each country, so the difference between two census dates is usually of ten years.

Data may be grouped for several reasons. One of them is that Statistical Offices decide to group several countries into one or some *residual categories* (usually labeled as "Other countries in region X"). In some other cases, they report the stock of immigrants born in a former country that later was split into smaller countries: USSR, Czechoslovakia, Yugoslavia, Ethiopia/Eritrea, Rhodesia, and the West Indies Federation are good examples. Finally, in some cases all origin countries are grouped, either because I only observe the total stock of immigrants in the destination country (a single group), or because the data is presented in big aggregate categories (e.g. data by continent or subcontinent of origin).

Table 1 summarizes the importance of grouped data. There are several aspects to highlight from the Table. First, data are more disaggregated in recent years: the average number of countries in grouped categories decrease from 167 to 87, and the share of the total stock that they represent decreases from more than one third in 1960 to less than 10 % in year 2000. Second, even though in 1960 and 1970 the coverage of total migrant stocks by bilateral data is only of around two thirds of the stock, this coverage increases to 80 % if we exclude the destination countries for which we only observe the total migrant stock. And third, even considering only disaggregated bilateral observations, the coverage of the total stock of immigrants is pretty large. For example, it is larger than in the OECD database used in Mayda (2010). Indeed,

Footnote 7 continued
Bermuda, The Netherlands Antilles, and Puerto Rico. Montenegro and Serbia are considered as a sole country.

[8] Destination country fixed effects, and especially destination-time fixed effects, are likely to account for most if not all these differences, given the log-specification of stocks in the specification estimated below. A caveat would still remain if policies introduced by a destination country affect different origin countries differently at different points in time.

[9] The only exception is France, whose Censuses were carried in 1954, 1962, 1968, 1975, 1982, 1990, 1999 and 2006. I interpolate them linearly to fit census dates to 1961, 1971, 1981, 1991, and 2001. Some years for three additional countries have to be extrapolated as well. Disaggregated information for Denmark and Finland circa 1960 and 1970 was not available, so I exploit information on residence permits for Denmark and on main language used for Finland. For Germany, pre-unification censuses are not available, so data on legal flows into West Germany is used to extrapolate. Robustness analysis to the exclusion of 1960 and 1970 is presented in the Appendix. Finally, data for United Kingdom includes only immigrants living in England and Wales; for year 2000 they represent a 95 % of the total stock of immigrants in the UK, a percentage that was uniformly distributed across origin countries.

Table 1 Number of origin countries with grouped data across destinations

	1960		1970		1980		1990		2000	
	Number of countries	% of stock	Number of countries	% of stock	Number of countries	% of stock	Number of countries	% of stock	Number of countries	% of stock
Australia	143	7.8	134	10.0	116	8.4	46	6.2	10	0.0
Austria	187	100.0	135	49.0	125	45.9	125	42.3	110	3.5
Belgium	155	4.0	153	2.8	143	2.0	143	3.0	118	1.6
Canada	172	20.8	172	25.4	175	36.3	24	5.6	22	6.0
Denmark	170	57.2	169	26.4	24	7.2	24	6.6	22	13.1
Finland	187	100.0	187	100.0	161	21.6	24	18.1	22	46.9
France	178	15.6	177	10.7	177	14.5	156	7.9	156	11.2
Germany	187	100.0	187	100.0	187	100.0	171	65.2	108	2.6
Greece	178	17.7	178	25.7	120	2.5	120	11.3	64	0.1
Iceland	187	100.0	187	100.0	133	2.4	150	6.7	131	11.5
Ireland	181	8.6	179	8.0	177	8.3	175	10.2	150	6.6
Italy	152	11.3	187	100.0	160	23.3	137	15.3	105	2.0
Japan	184	1.8	184	3.5	184	4.5	133	2.5	151	1.1
Korea (Rep.)	179	0.5	166	0.7	166	3.6	166	7.1	169	10.4
Luxembourg	178	2.3	177	3.9	175	5.4	175	8.6	175	15.0
The Netherlands	180	46.7	180	33.4	180	28.8	180	25.8	22	5.2
New Zealand	186	89.6	186	89.5	186	89.3	139	3.2	132	3.1
Norway	169	7.2	163	11.2	161	14.1	161	22.0	155	18.1
Portugal	179	11.8	179	16.8	24	0.1	24	0.3	152	10.3
Spain	187	100.0	133	1.2	124	1.6	24	0.3	0	0.0
Sweden	148	4.0	164	10.8	24	10.1	24	11.3	5	11.3

Table 1 continued

	1960		1970		1980		1990		2000	
	Number of countries	% of stock	Number of countries	% of stock	Number of countries	% of stock	Number of countries	% of stock	Number of countries	% of stock
Switzerland	74	0.4	47	2.6	24	6.6	24	13.4	22	24.8
United Kingdom	132	9.1	147	11.0	172	47.0	157	36.6	0	0.0
United States	126	13.9	122	10.9	24	5.2	24	3.1	75	4.5
Average	167	34.6	162	31.4	131	20.4	105	13.9	87	8.7
Excluding 100 %'s	161	17.4	157	17.7	128	16.9	105	13.9	87	8.7

The first column for each year represents the number of countries that are in grouped categories in that period. The total amount of possible origin countries is 187. The second column for each year is the % from the total stock of immigrants that is in grouped categories. Each destination country may have several grouped categories. The last two rows are averages across destination countries

Mayda (2010) states that the coverage of total inflows in her database ranges from 45 % (Belgium) to 84 % (US). For the equivalent time period, the average coverage by bilateral observations here ranges from 80 to 91 %. Regarding the number of countries with disaggregate bilateral observations, Mayda (2010) and Ortega and Peri (2013) use a sample of 79 and 120 origin countries respectively—including zero flows that "are likely to correspond to very small flows rather than zero flows" (Mayda 2010); Pedersen et al. (2008), report a substantial portion of missing values among their sample of 129 countries of origin. The country coverage for these years is similar on average in Table 1, but it is much larger both if we restrict to the sample of 15 destination countries considered in Mayda (2010) and Ortega and Peri (2013), or if we consider federations of countries that were single countries at the time as ungrouped countries (e.g., Former Yugoslavia accounted for almost a half of the stock of immigrants in Austria in years between 1970 and 1990, one quarter of the stock in Switzerland in year 2000, and around a 10 % of the Swedish stock in years between 1980 and 2000, and the USSR represented between 5 and 8 % of US and Canadian stocks in years 1960 and 1970, and around a 3 % in other several destination countries).

2.2 Description of the data

Table 2 presents averages, standard deviations, and extreme values for each destination country, and the number of available observations. The left panel refers to the baseline sample, which includes all disaggregated bilateral observations plus one observation for each set of grouped countries. To compute these statistics, grouped observations are weighted by the number of countries included in the group. The right panel restricts the sample to disaggregated bilateral observations. The baseline sample includes 6,804 bilateral observations plus 625 grouped observations. These observations are not uniformly distributed across destination countries, ranging from the 55 single bilateral observations for Luxembourg (plus 26 grouped observations) to the 744 for Switzerland (plus 28 groups).

The comparison of averages across the two samples suggests that grouped observations tend to include countries with smaller stocks of migrants, which is not surprising given that some grouping occurs due to labeling like "Other countries in region X". The difference in average stock size between the two samples, however, may be exaggerated by the fact that data are more grouped in earlier years of the sample, when immigrant stocks are smaller. The fact that grouping does not occur at random highlights the importance of including grouped observations in the analysis (as opposed to dropping them from the sample).

Table 2 shows substantial variation in average migrant stocks, ranging from 46 immigrants per origin country in Iceland to 99,276 individuals per country in the United States. There is also a large variation across origin countries, as appreciated from the size of standard deviations. The extreme case is the US, with a standard deviation of 395,483 individuals, and stocks of immigrants that range from the 11 immigrants from Djibouti in 1990 to the 9,325,452 Mexicans in year 2000, but it is not the only one: Canada, Germany, France, and Japan also have sizeable standard

Table 2 Descriptive statistics for migrant stocks

	Full sample					Ungrouped observations only				
	Obs.	Mean	Sd.	Min	Max	Obs.	Mean	Sd.	Min	Max
Australia	531	15,982	78,882	1	1,104,594	486	28,975	107,671	1	1,104,594
Austria	280	1924	9938	1	132,975	253	4859	17,424	1	132,975
Belgium	261	3965	21,011	0	279,700	223	16,201	40,676	17	279,700
Canada	405	21,362	75,118	1	969,715	370	44,759	115,159	1	969,715
Denmark	555	935	3322	0	50,470	526	1444	4261	1	50,470
Finland	367	152	527	1	7887	354	218	690	1	7887
France	126	16,787	77,617	107	791,627	91	152,310	203,663	5728	791,627
Germany	108	29,159	78,535	288	1,947,938	95	95,841	227,470	366	1,947,938
Greece	301	1315	14,653	1	438,036	275	4279	26,790	8	438,036
Iceland	174	46	182	0	2456	147	204	424	1	2456
Ireland	82	1163	12,441	20	242,155	73	13,694	42,575	20	242,155
Italy	225	2214	9956	3	180,103	194	9353	20,320	15	180,103
Japan	109	4110	40,867	26	567,598	99	37,854	120,424	102	567,598
Korea (Rep.)	106	284	2272	0	47,474	89	2775	6882	1	47,474
Luxembourg	81	500	3198	0	58,657	55	7737	10,857	115	58,657
The Netherlands	208	2167	11,284	1	191,500	193	8146	23,896	1	191,500
New Zealand	138	2601	15,140	31	232,764	106	12,021	32,276	39	232,764
Norway	162	801	2691	8	33,251	126	4944	5801	26	33,251
Portugal	401	482	2619	1	37,014	377	1111	4038	1	37,014
Spain	487	2811	13,692	0	244,630	467	5157	19,080	1	244,630
Sweden	592	1993	10,369	0	181,477	570	2936	13,110	1	181,477
Switzerland	772	5411	33,453	0	583,855	744	6008	37,014	1	583,855
United Kingdom	361	17,338	56,261	3	675,870	327	39,377	90,436	3	675,870
United States	597	99,276	395,483	11	9,325,452	564	154,516	501,317	11	9,325,452
All	7429	9699	90,725	0	9,325,452	6804	26,483	162,994	1	9,325,452

The unit of observation is origin-destination-year. All figures (except the number of observations) are in individual counts. Left panel refers to the baseline sample, which includes disaggregate bilateral observations and grouped observations—grouped observations are weighted by the number of countries included in the group. Right panel restricts the sample to disaggregated bilateral observations

deviations, and they are also quite large in Greece and Ireland compared to averages. Overall, the standard deviation in the whole sample is 90,729 individuals, roughly ten times the sample average.[10]

Table 2 does not provide a sense of time series variation. Figure 1 draws the evolution of immigrant shares (i.e. stock of immigrants over population) across destination countries over the sample period. Different patterns are observed across countries:

[10] These sample standard deviations are downward biased unless the stock of immigrants from all countries in each grouped observation is the same; the underestimation of the *true* standard deviations will be larger the larger the (unobserved) dispersion within each grouped observation.

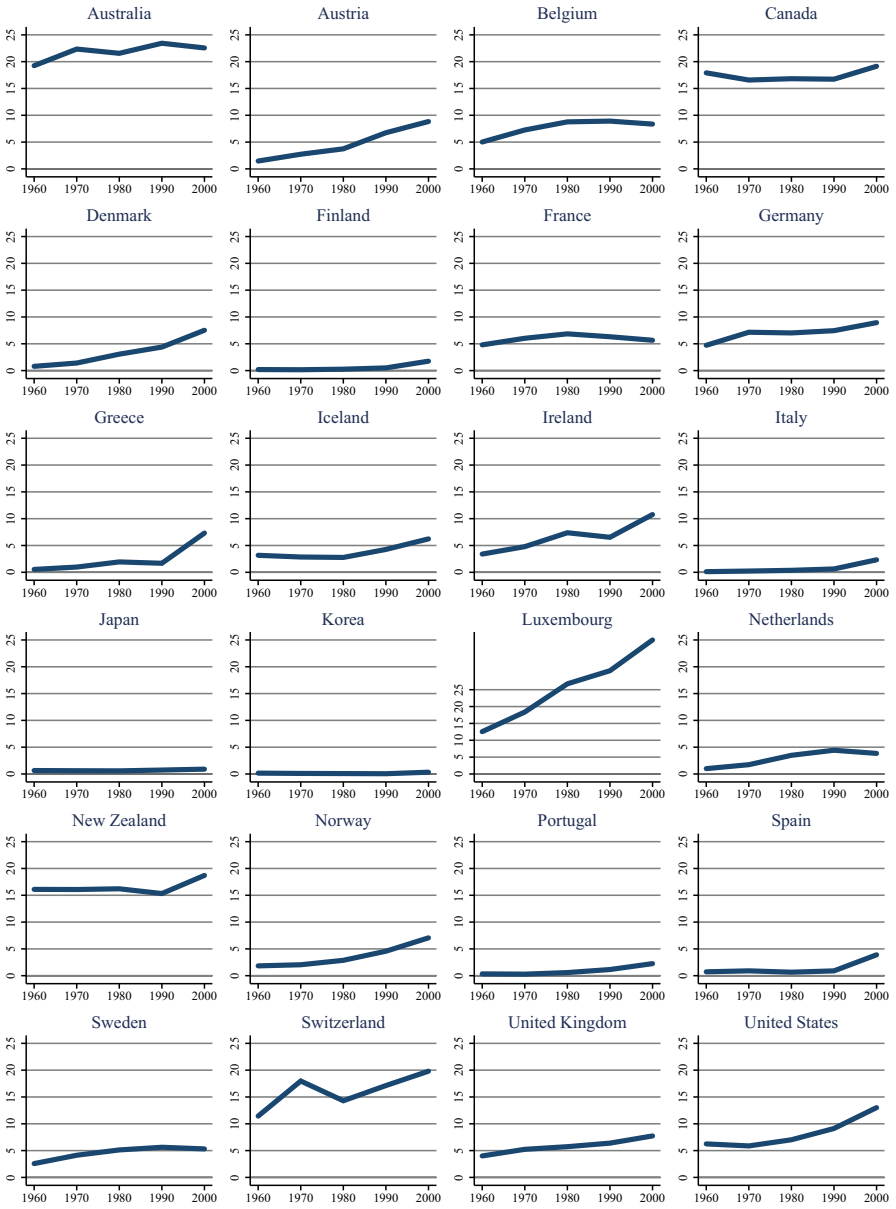

Fig. 1 Share of immigrants (%) for sample OECD countries (1960–2000). Each plot presents destination country's share of immigrants (immigrants over population). *Left axes* have a common scale, ranging from 0 to 25 %—which is compressed for Luxembourg due to its exceptionally large fraction of immigrants (36.4 % in year 2000)

stable low-immigration countries (Korea and Japan), stable high-immigration countries (Australia, Canada and New Zealand), old immigration countries with a strong increasing trend (US, Luxembourg, Switzerland, and the UK), old immigration coun-

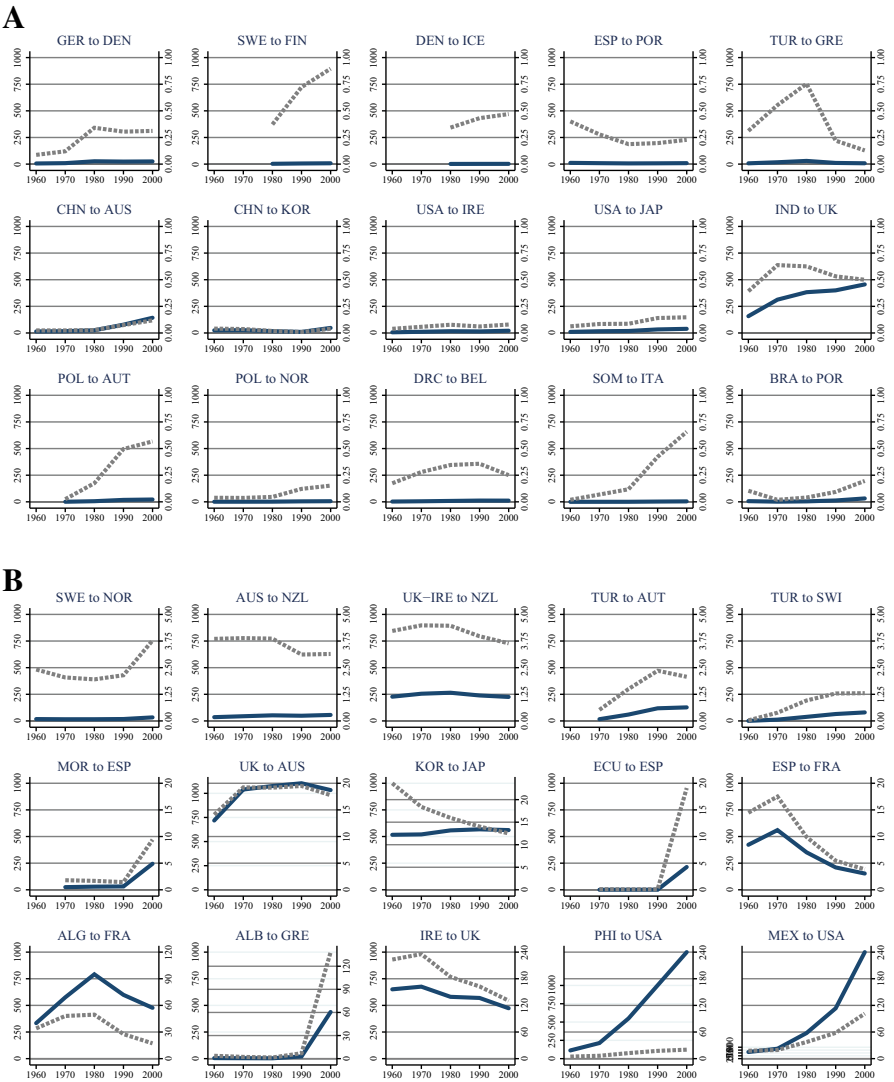

Fig. 2 Stocks (1000s) and share of population who migrated ($^0/_{00}$) for selected country pairs (1960–2000). **a** Some country pairs with low migrant rates. **b** Some country pairs with high migrant rates. *Solid lines* are bilateral migrant stocks (in 1000s, *left axis*) from the origin to the destination countries listed in each title. *Dashed lines* are migrant rates (in $^0/_{00}$, *right axis*), i.e. stock of migrants from country "X" in country "Y" over total population of country "X" (origin). *Left axis scale* is common to all country pairs ranging from 0 to 1000—which is compressed for MEX and PHI to USA (9.3 and 4.4 million respectively in year 2000). *Right axes* from *top panel* have also common scale (0 to 1 $^0/_{00}$); in the *bottom panel*, it ranges from 0 to 5 $^0/_{00}$ in the *first rows*, from 0 to 20 $^0/_{00}$ in the *second one*, and from 0 to 120 $^0/_{00}$ and 0 to 240 $^0/_{00}$ in the *last row*

tries with a slight decrease (Belgium and France), and new immigration countries (Spain, Italy, Austria, Greece, Portugal, and Nordic Countries). Figure 2 adds the country of origin layer. In particular, I plot the evolution of the stock of immi-

grants and of the bilateral migration rate for a selected group of country pairs.[11] The figure shows substantial variation across countries and over time. The top panel includes a sample of country pairs with low migration rates, which include some pairs with flat trend (e.g. North American in Ireland or Japan, Chinese in Korea), and others with important increases over the sample period (Somali in Italy, Polish in Austria, Swedish in Finland). The bottom panel includes country pairs with high migrant rates, including pairs with decreasing rates (Korean in Japan, Irish in the UK, Spanish in France), roughly constant rates (Australian and British/Irish in New Zealand, British in Australia), and sharply increasing rates (Ecuadorian in Spain, Albanian in Greece, and, most extremely, Filipino and Mexican in the US).

2.3 Comparison with Özden, Parsons, Schiff and Walmsey (2011)

Contemporaneous work by Özden et al. (2011) provides a similar dataset. These authors' approach is to impute grouped observations to specific origin countries. They do so based on the propensity of destination countries to accept migrants from a particular origin in subsequent years, and based on the propensity of a given origin country to send them abroad. These imputations may be particularly harming when one is interested in estimating, precisely, the determinants of international migration. This method may generate measurement error correlated, almost by construction, with the regressors of interest.

Table 3 reproduces Table 2 using Özden et al. (2011) dataset (generating artificially grouped observations). The comparison between the two tables is interesting. On average, their data predicts, about 2,000 extra immigrants per origin country, almost 4,000 when only origin countries with bilateral observations (in the current dataset) are considered. This gap is not homogeneous across destination countries. Some countries (Australia, Belgium, Canada, Denmark, Finland, Greece, Iceland, Ireland, Luxembourg, New Zealand, Norway, Switzerland, and United Kingdom) present very similar stocks. Others (Austria, France, Germany, Italy, Japan, The Netherlands, Portugal, Spain, Sweden, and United States) present substantially different averages both in grouped and in bilateral observations. Finally, another (Korea) is similar in the bilateral observations and differ substantially in grouped observations. Data in Özden et al. (2011) also have a larger cross-origin country variance.

To elaborate further in these differences, Fig. 3 presents histograms of the distribution of discrepancies for those observations with bilateral information available in both datasets. The majority of the observations (4,861 out of 6,804, or 71.4 %) present no discrepancies or discrepancies of less than 1000 migrants (central lines of Fig. 3c). The remaining 1943 observations (28.6 %) are distributed as follows: 1,338 (19.7 %) have discrepancies between 1000 and 10,000; 534 (7.8 %) have discrepancies of between 10,000 and 100,000 migrants, and 69 (1 %) have discrepancies above 100,000 migrants. Most of these extreme discrepancies are given by the defi-

[11] The rate is defined as country pair's stock of migrants over origin country's population.

Table 3 Descriptive statistics for migrant stocks in Özden et al. (2011)

	Full sample					Ungrouped observations only				
	Obs.	Mean	Sd.	Min	Max	Obs.	Mean	Sd.	Min	Max
Australia	531	15,611	78,667	0	1,092,182	486	28,426	107,436	0	1,092,182
Austria	280	4266	17,038	0	189,405	253	8663	28,913	0	189,405
Belgium	261	4068	21,416	2	288,899	223	16,588	41,441	5	288,899
Canada	405	21,235	76,300	0	941,217	370	44,147	117,262	0	941,217
Denmark	555	986	3234	0	31,883	526	1520	4130	0	31,883
Finland	367	320	1442	0	28,981	354	440	2255	0	28,981
France	126	27,784	118,934	228	1,493,990	91	225,080	318,615	1286	1,493,990
Germany	108	42,042	122,814	244	2,008,979	95	148,971	359,689	337	2,008,979
Greece	301	1558	14,356	0	420,838	275	4968	26,155	0	420,838
Iceland	174	41	173	0	2306	147	189	403	0	2306
Ireland	82	1111	11,926	4	227,440	73	13,105	40,820	19	227,440
Italy	225	6395	19,085	6	286,498	194	23,276	37,085	29	286,498
Japan	109	5309	50,733	21	700,574	99	48,993	148,913	126	700,574
Korea (Rep.)	106	2122	4193	0	48,165	89	2731	6856	0	48,165
Luxembourg	81	474	2549	1	41,352	55	6852	8168	139	41,352
The Netherlands	208	4250	15,273	0	178,273	193	12,971	32,005	0	178,273
New Zealand	138	2491	15,832	4	272,190	106	11,781	35,143	4	272,190
Norway	162	806	2772	8	34,109	126	5024	6020	26	34,109
Portugal	401	1570	9970	0	167,578	377	3460	15,500	0	167,578
Spain	487	4010	16,702	0	253,173	467	7504	23,105	0	253,173
Sweden	592	3438	16,593	0	250,527	570	4983	20,927	0	250,527
Switzerland	772	6203	36,556	0	590,957	744	6794	40,427	0	590,957
United Kingdom	361	17,750	58,907	0	717,774	327	40,223	95,019	0	717,774
United States	597	103,392	405,297	0	9,367,910	564	161,379	513,337	0	9,367,910
All	7429	11,551	103,308	0	9,367,910	6804	30,163	174,097	0	9,367,910

The unit of observation is origin-destination-year. All figures (except the number of observations) are in individual counts. Left panel refers to the baseline sample, which includes disaggregate bilateral observations and grouped observations—grouped observations are weighted by the number of countries included in the group. Right panel restricts the sample to disaggregated bilateral observations

nition of a migrant, like Algerian, German, Spanish or Italian in France, and Polish, Russian, and Czech in Germany (some in several periods). Özden et al. (2011) only moved out of the birthplace definition of immigrant if data by country of birth was not available for three or more periods. Otherwise, when country of birth is unavailable for only a few periods, they do imputations based on the information in the available years. Instead, in line with the spirit of this paper, I use nationality when country of birth is not consistently available for all periods, so that the data are as raw as possible.

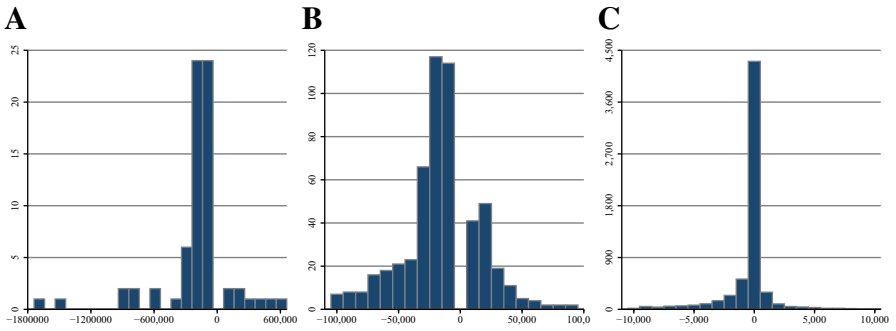

Fig. 3 Distribution of discrepancies with Özden et al. (2011) for available bilateral observations. **a** >100,000, **b** ∈ (100,000; 10,000), **c** <10,000. The *histograms* present the number of origin countries/periods with each level of discrepancies. The *left histogram* omits (absolute) discrepancies smaller than 100,000 migrants. *Center histogram* presents observations with absolute deviations between 100,000 and 10,000 migrants. And *right panel* plots observations with absolute deviations smaller than 10,000 migrants. A positive number indicates that this dataset reports more immigrants than Özden et al. (2011)

2.4 Other variables

The remaining variables used in the regression analysis below come from different sources (descriptive statistics provided in Table 4). All variables are averages over years $t-10$ to $t-1$. GDP per capita, population, and government share of GDP come from Penn World Tables (versions 6.2 and 7.0). In order to minimize the number of missing values for GDP per capita, I use Total Economy Database (Conference Board) to extrapolate backwards discontinuous Penn World Tables series. Both origin and destination countries' series are in constant international dollars of 2005 (chain). Population in origin and destination countries are in millions. Government share is public sector consumption over real GDP. Age dependence ratio at destination country—individuals older than 65 years over population of working age—is from World Development Indicators. Unemployment rate (in %) is obtained from the OECD. Geographic variables include physical distance—great circle distance between the two capitals—and dummies for having a common language, a past colonial relationship and a common border. The distance variable is based on Rose (2004) data, extended to cover the whole sample. The common language dummy is constructed using data from Alesina et al. (2003) and The World Factbook from the CIA; a pair of countries is considered to have a common language if there is a particular language that is spoken by at least a 10 % of the population in each of the two countries. Colonial relationship and common border dummies are also based on The World Factbook. War and Polity IV autocracy-democracy index are constructed with data from the Polity IV Project. The war variable measures the fraction of months over the preceding decade that the country was in any type of war. The Polity IV index ranges from −10, indicating autocracy, to 10, which indicates democracy, through values around 0, which indicate anocracy (a situation of instability emerged from the absence of a strong power and of the rule of law). The young population variable is constructed using data of total

Table 4 Descriptive statistics of the explanatory variables

	Full sample					Ungrouped observations only				
	Obs.	Mean	Sd.	Min	Max	Obs.	Mean	Sd.	Min	Max
GDPpc dest.	7429	17.85	8.33	1.59	49.94	6804	21.99	6.87	1.59	49.94
GDPpc origin	7340	8.11	8.06	0.26	215.02	6727	9.68	12.24	0.26	215.02
Log distance	7429	8.31	0.72	4.80	9.79	6804	8.11	0.95	4.80	9.79
Comm. lang.	7429	0.10	0.24	0.00	1.00	6804	0.15	0.36	0.00	1.00
Colonial rel.	7429	0.05	0.17	0.00	1.00	6804	0.07	0.26	0.00	1.00
Common border	7429	0.01	0.10	0.00	1.00	6804	0.03	0.18	0.00	1.00
Pop. origin	7429	21.68	73.19	0.01	1207.69	6804	40.89	129.93	0.01	1207.69
Pop. dest.	7429	31.27	47.62	0.16	264.74	6804	36.16	60.76	0.22	264.74
Unemp. rate dest.	7172	4.39	2.43	0.03	11.10	6608	5.29	2.34	0.03	11.10
Age dep. dest.	7429	18.45	4.75	2.32	27.66	6804	20.38	3.86	2.32	27.66
Gov. share dest.	7427	9.62	2.81	2.83	19.80	6804	9.31	2.67	2.83	19.80
War origin	7429	0.06	0.13	0.00	1.00	6804	0.08	0.22	0.00	1.00
Polity IV origin	7232	0.60	5.31	−10.00	10.00	6607	2.18	7.32	−10.00	10.00
Pop. 15–34 origin	7231	7.28	26.06	0.01	445.82	6606	14.03	46.81	0.01	445.82

The unit of observation is origin-destination-year. Left panel includes both observations with bilateral migrant data, and observations for which migrant stocks are grouped —which are grouped equivalently, weighting by the number of countries in the group. Right panel includes only observations with available bilateral stocks

population by age group from United Nations. The variable includes the population aged between 15 and 34.

3 Econometric model

3.1 Standard gravity model

In the reminder of the paper, I use the new data presented above to analyze the determinants of international migration. In particular, the data are used to estimate different types of "gravity equations" (see Beine et al. 2015 for a review of this literature). Simplest gravity equations can be derived from random utility models in which the utility of moving from home country j to country k at time t is of the form:

$$U_{ijkt} \equiv w_{kt} - c_{jk} + \varepsilon_{ijkt}, \tag{1}$$

where i indicates an individual, w_{kt} is the wage at country k and time t, c_{jk} is the moving cost from j to k (where c_{jj} is typically normalized to 0), and ε_{ijkt} is a random term that is Type-I extreme value distributed. Given the distributional assumption of the random term, the relative odds of moving from country j to country k vs. staying

in country j are equivalent to:

$$\frac{M_{jkt}}{Pop_{jt}} = \exp(w_{kt} - w_{jt}) \exp(-c_{jk}), \tag{2}$$

where M_{jkt} is the stock of migrants from country j to country k in year t, and Pop_{jt} is the (*ex-post*) population in country j at time t.

Taking logs to the above expression, using GDP per capita as a proxy for wages, and various variables to proxy for moving costs, the model can be written as:

$$\begin{aligned} \ln M_{jkt} = {} & \alpha_1 GDPpc_{kt} + \alpha_2 GDPpc_{jt} + \alpha_3 \ln dist_{jk} + \alpha_4 \mathbb{1}\{CommLang_{jk}\} \\ & + \alpha_5 \mathbb{1}\{Colony_{jk}\} + \alpha_6 \mathbb{1}\{Border_{jk}\} + \alpha_7 \ln Pop_{kt} + \alpha_8 \ln Pop_{jt} \\ & + \text{Fixed effects} + \upsilon_{jkt}. \end{aligned} \tag{3}$$

All the variables included in Eq. (3) are described in Sect. 2.4. Different specifications include different combinations of fixed effects, depending on the assumptions underlying the distribution of ε_{ijkt}. These include country of origin, destination country, year, origin × year, destination × year, and/or country pair fixed effects. Migration is expected to be positively affected by income gains (hence, α_1 is expected to be positive and α_2, negative), by having a common language, a colonial relation, and a common border, and by the population in the origin country, and negatively affected by physical distance; the expected sign of the effect of population in the destination country is ambiguous *a priori*.

Similar micro-foundation for this regression can be found in the model by Grogger and Hanson (2011), or in the survey by Beine et al. (2015), and it is comparable to the previous studies in the literature (Mayda 2010; Grogger and Hanson 2011; Ortega and Peri 2013). Bertoli and Fernández-Huertas Moraga (2013) highlight the importance of origin-time dummies combined with country pair dummies due to the "Multilateral Resistance to Migration". Beine et al. (2015) go a step further and suggest adding origin × time × nest dummies on top. As noted below, with the number of observations left due to the grouping of the data, these models are too demanding in terms of degrees of freedom to be credibly estimated.[12,13]

While the inclusion of GDP per capita in levels to approximate origin and destination country wages in Eq. (3) seems very closely connected to the underlying theoretical model described by Eqs. (1) and (2) (as noted by Grogger and Hanson 2011), there are many papers in the literature that estimate equations like (3) including GDP per capita in logs, as highlighted in Beine et al. (2015). For comparability with these studies, I run some specifications of Eq. (3) in which GDP per capita is introduced in logs.

[12] Bertoli and Fernández-Huertas Moraga (2013) propose a formal test to the "Multilateral Resistance to Migration". Such test cannot be implemented here because of the grouped data.

[13] Several papers in the literature estimated Eq. (3) using the Poisson ML estimator due to the presence of a substantial number of zero observations. This concern does not apply to this paper, as the current database includes very few zero migrant stocks (countries with few immigrants are typically in grouped categories).

3.2 Heterogeneous effects of income gains

An important implication of the model described above is that an increase in GDP per capita of a destination country would increase the stock of migrants from all origin countries by the same percentage (i.e. linear effect of GDP per capita on migrant stocks in logs). Likewise, an increase in the GDP per capita of a given country of origin would increase the stock of migrants from that country into all destinations by the same relative amount (linear effect on log migrant stocks).

However, the effect of income shocks on moving prospects might be more marked for closer countries compared countries that are farther apart. For example, large moving costs (distance) reduce the flexibility of individuals to move back and forth to their home country when income changes. As a result, in the migration decision, individuals from farther away countries may give more weight to long run income (as opposed to income shocks), whereas individuals from neighboring countries will be more prone to go back and forth to take advantage of income fluctuations. Similarly, if individuals dislike living far away from home, they might require a compensating wage differential to offset the unpleasantness of living abroad. If the disutility of being far from home increases with distance, they will require an increasing wage premium to take the decision to migrate. Hence, these compensating wage differentials would also introduce a heterogeneous effect of income gains on moving prospects depending on distance, which would make migration more reactive to income at closer distances.

As a way to micro-found these heterogeneous effects, consider the following modification of Eq. (1):

$$U_{jkt} \equiv u(w_{kt}, d_{jk}) - c_{jk} + \varepsilon_{ijkt}, \tag{4}$$

where d_{jk} is the distance between home country j and destination k, and $u(.,.)$ is a utility function with an elasticity of substitution between income and distance to be identified. The case in which wage and proximity (negative distance) are perfect substitutes is observationally equivalent to the standard utility model.

Following a similar procedure to the one used to derive Eq. (3), and approximating $u(w_{kt}, d_{jk})$ by a first order expansion around the mean we obtain:

$$\begin{aligned}\ln M_{ijt} = {} & \gamma_1 GDPpc_{it} + \gamma_2 GDPpc_{jt} + \gamma_3 \widetilde{GDPpc}_{kt} \widetilde{\ln dist}_{jk} \\ & + \gamma_4 \widetilde{GDPpc}_{jt} \widetilde{\ln dist}_{jk} + \gamma_5 \ln dist_{ij} + \gamma_6 \mathbb{1}\{CommLang_{ij}\} \\ & + \gamma_7 \mathbb{1}\{Colony_{ij}\} + \gamma_8 \mathbb{1}\{Border_{ij}\} + \gamma_9 \ln Pop_{it} \\ & + \gamma_{10} \ln Pop_{jt} + \text{Fixed effects} + \upsilon_{ijt},\end{aligned} \tag{5}$$

where $\widetilde{x} \equiv x - \bar{x}$ indicates that variables are in deviations with respect to sample means. Parameters γ_3 and γ_4 are reduced forms of the cross-partial derivative of $u(w_{kt}, d_{jk})$ evaluated at sample means. Hence, the presence of an heterogeneous response of migration to shocks to destination and/or origin country's income as a function of

distance is an indication of a complementarity (or, potentially, substitution) between income gains and proximity.

3.3 Identification of fixed effects with grouped data

A potential limitation of working with grouped data is in the identification of fixed effects in the estimation of Eqs. (3) and (5). In the simplest specifications estimated below, I introduce origin and destination country fixed effects, and year dummies. Additionally, in several specifications I introduce country-pair or country of origin × year dummies. Destination country, time, and destination × time fixed effects are identified in all cases, as grouping only affects origin countries. To identify a dummy for an origin country, we need to observe, at least, one bilateral observation from that country, or that the country appears in a unique combination of grouped observations.[14] To identify a country of origin × year dummy, this bilateral observation or unique combination of groups has to be observed in each year. And the identification of a country-pair dummy requires the bilateral observation to be observed at least once for each destination country. When one of these situations is not satisfied, a single dummy for each unique combination of groups is identified.

Figure 4 summarizes the availability of this variation in the data. The left histogram shows the number of origin countries with 0, 1, . . . , 120 (=24 × 5) country of destination × year observations. All countries of origin have between 4 and 99 destination × year observations, which is enough to identify all origin country fixed effects; in most of the cases (105 out of 188 countries, 55 % of them) we have between 20 and 40 observations. The central histogram shows the number of countries of origin × years with bilateral data for the 0, 1, . . . , 24 destination countries. The figure shows that we cannot separately identify country of origin × year dummies in 160 out of $188 \times 5 = 940$ origin × year combinations (17 %), in most of the cases, this is due to federations of countries—USSR, Yugoslavia,…—that were still federated at the given period. Finally, the right histogram shows the number of country pairs with bilateral data for the 0, 1, . . . , 5 periods. According to the figure, we cannot identify a country pair dummy for 1854 out of $24 \times 188 = 4488$ country pairs (41 %).[15] This limitation does not affect consistency of the estimates below, but it affects the precision of the estimation of the most demanding models. Richer models that include the combination of country-pair and origin × year dummies, as suggested by Bertoli and Fernández-Huertas Moraga (2013), or the even richer ones that add origin × year × nest on top, as in Beine et al. (2015) would absorb too many degrees of freedom to allow us to draw any relevant conclusion from them.

[14] For example, consider that for a destination country A we have two observations that belong to the "Rest of Europe" group. If in another destination country B one of them belongs to the "Yugoslavia" category and the other does not, I would be able to identify them separately.

[15] Additionally, for 83 countries of origin × years (9 %) and for 637 country pairs (14 %) we only have one observation. In this case, the available observation together with the grouped data for other destination countries is enough to identify the fixed effect, but such observations do not contribute to the identification of other parameters.

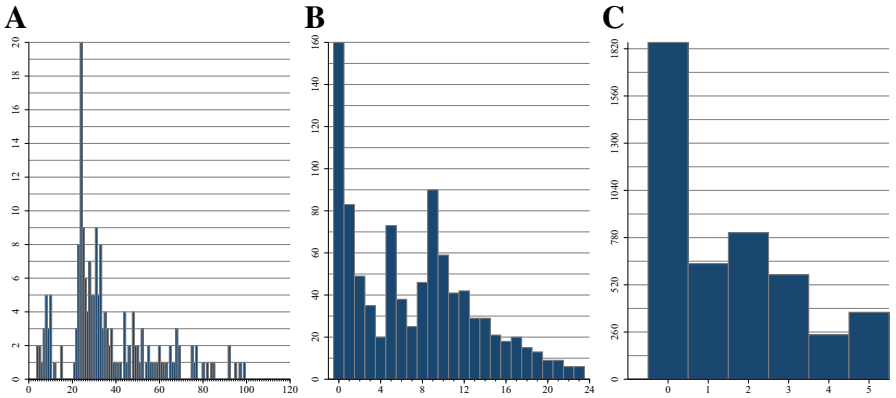

Fig. 4 Available ungrouped bilateral observations. **a** Origin countries. **b** Origin countries × periods. **c** Country pairs. *Left histogram* presents the number of origin countries with $0, 1, \ldots, 120$ destination × year observations; the total amount of origin countries is 188. *Center histogram* shows the number of origin × years with $0, 1, \ldots, 24$ destination country observations; there are $188 \times 5 = 940$ origin × year observations. *Right histogram* presents the number of country pairs with $0, 1, \ldots, 5$ yearly observations; the total amount of country pairs is $24 \times 187 = 4488$

4 Estimation results

4.1 Linear effects: standard gravity model

Table 5 presents the results for the estimation of different versions of Eq. (3). All regressions include at least origin and destination country fixed effects, and year dummies. The first column is the baseline specification. The stock of migrants is positively associated with the GDP per capita of the destination country. This result suggest that better economic opportunities in the destination country encourage migration. In particular, everything else constant, a 1000$ increase in GDP per capita of the destination country increases the immigrant stock by a 5.2 %. This magnitude is in line, for example, with Ortega and Peri (2013), who find a positive effect of a 5–6 %. According to the results in Table 5, a 10 % increase in GDP per capita of the average country of destination (which is 17,848$, see Table 4) would increase the immigrant stock by a 9.3 %.[16,17] GDP per capita in OECD countries averaged 9,101$ in 1960, and 27,341$ in year 2000. According to the results in Table 5, this 200 % increase would have increased the stock of immigrants in a 95 % (25 millions of immigrants over the OECD), more than a half of the actual increase (45 millions).

[16] This result is qualitatively in line with Mayda (2010), who finds that a 10 % increase in destination country GDP per capita increases emigration rates by a 20 %. Quantitatively, these numbers are hard to compare as her dependent variable is in flows instead of stocks.

[17] As noted by Beine et al. (2015), the inclusion of different combinations of dummies imply that estimates are consistent with random utility models that are not based on the canonical version of the gravity model. This affects the interpretation of the results in terms of elasticities.

Table 5 Determinants of bilateral migrant stocks—linear effects

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
GDPpc dest.	0.052		0.069	0.053		0.053	0.043
	(0.023)		(0.014)	(0.023)		(0.027)	(0.020)
GDPpc origin	−0.008		−0.007		−0.010	−0.014	−0.017
	(0.009)		(0.008)		(0.009)	(0.011)	(0.015)
GDPpc gap		0.023					
		(0.011)					
Log distance	−0.904	−0.903	−1.051	−0.931	−0.910		−0.832
	(0.073)	(0.074)	(0.047)	(0.077)	(0.057)		(0.069)
Common language	0.585	0.582	0.760	0.591	0.581		0.600
	(0.131)	(0.132)	(0.081)	(0.144)	(0.096)		(0.121)
Colonial rel.	2.281	2.277	2.107	2.285	2.268		2.382
	(0.146)	(0.145)	(0.099)	(0.156)	(0.122)		(0.125)
Common border	0.030	0.033	0.036	−0.004	0.026		0.235
	(0.178)	(0.178)	(0.125)	(0.187)	(0.151)		(0.164)
Log pop. origin	1.341	1.075	1.466		1.324	1.227	1.742
	(0.493)	(0.442)	(0.197)		(0.277)	(0.631)	(0.465)
Log pop. dest.	1.161	1.012	−1.902	1.082		1.100	−0.215
	(1.157)	(1.187)	(0.440)	(1.157)		(1.387)	(0.886)
Grouped obs.	Yes	Yes	No	Yes	Yes	Yes	Yes
St. devs. of controls	No	No	No	No	No	No	Yes
Time dummies	Yes	Yes	Yes	No	No	Yes	Yes
Origin dummies	Yes	Yes	Yes	No	Yes	No	Yes
Origin-time dummies	No	No	No	Yes	No	No	No
Destination dummies	Yes	Yes	Yes	Yes	No	No	Yes
Dest.-time dummies	No	No	No	No	Yes	No	No
Country-pair dummies	No	No	No	No	No	Yes	No
Obs	7340	7340	6727	7429	7340	7340	7332
$\bar{R}^2$	0.958	0.958	0.966	0.956	0.975	0.960	0.966

Standard errors, clustered at the origin-time level, in parentheses. Dependent variable: log migrant stocks. Unit of observation: origin-destination-time. Regressions include the indicated fixed effects. The p-value of a test of the null that coefficients displayed in Column (3) are jointly equal to point estimates in Column (1) is 0.000, and the corresponding p-value for Column (7) is 0.264

Theoretical predictions from models like the ones in Grogger and Hanson (2011) or in Mayda (2010) suggest that α_1 and α_2 should be similar in magnitude and of opposite sign. However, Table 5 shows a much smaller effect of origin country GDP per capita. Although it is negative (consistently in all specifications), the coefficient is far from being significantly different from zero, and point estimates are one order of magnitude smaller than destination country counterparts. This result is not new; Mayda (2010) also finds a non-significant effect, although her point estimates are indeed positive. This finding could result from an additional positive effect of origin country GDP per capita on migration prospects. Borrowing con-

straints could be a plausible explanation: if individuals from poorer countries (lower GDP per capita) are financially constrained, then, other things equal, their chances to migrate are lower; therefore, the larger the GDP per capita, the less constrained they are, and the larger is the probability that they migrate. If that were the case, one would expect that this effect should be homogeneous across all destination countries, which is in line with findings discussed in Sect. 4.2. Several papers in the literature explore this possibility, and conclude that this is likely the case (Beine et al. 2015).

Physical and cultural distance play an important role in explaining moving costs. The elasticity of the migrant stock with respect to physical distance is about 0.9. Having a common language or a colonial relationship increases importantly the stock of immigrants. A common border, however, seems less important. These results are, again, qualitatively similar to Mayda (2010), Grogger and Hanson (2011), and Ortega and Peri (2013). Finally, we can neither reject that the coefficient of log population in the origin country is equal to one, nor that the one of log population in the destination country is zero, which are the values predicted by the model outlined above.

The remaining columns of Table 5 check the stability of the estimates across different versions of the same equation. In order to obtain estimates which are fully comparable to Grogger and Hanson (2011), in Column (2) I impose the same coefficient (of opposite sign) for origin and destination countries' GDP per capita. The coefficient of income gap is 0.023 (s.e. 0.011) very close to their estimate of 0.018 (s.e. 0.029) and much more precisely estimated, given the larger coverage by the dataset presented in this paper. Additionally, the coefficients for the variables associated with moving costs are extremely similar.

The fact that these estimates are comparable to Grogger and Hanson (2011) is useful to asses the validity of the way in which grouped data is treated in this paper. Grogger and Hanson (2011) use data from Docquier and Marfouk (2006), which is census data collected in a similar manner to the one in this paper for years 1990 and 2000. Grogger and Hanson (2011) estimate their regressions using data for year 2000. The similarity in the coefficients with respect to their paper indicates that the treatment I give to the grouped data produces consistent estimates of the relevant coefficients.

In order to analyze the importance of including the 100 % of migrant stocks, I drop grouped observations in Column (3). Although qualitative results hold, point estimates are somewhat different. In particular, four out of the eight coefficients are statistically different from point estimates in Column (1), and a Wald test of the null hypothesis that all eight coefficients are equal to their counterparts in Column (1) clearly rejects it (see p-value in the note of Table 5). This differences are caused by the fact that grouped observations (which are eliminated in previous studies) are not from a random sample of countries of origin. Therefore, we can conclude that including grouped observations—so that we cover the 100 % of total migrant stocks—is very important to obtain unbiased estimates.

In Columns (4) through (6), I change the specification of fixed effects. On top of origin, destination, and time fixed effects that are included in Columns (1) though (3),

I enrich the analysis by adding destination × time, origin × time, and country-pair dummies respectively. These specifications are more demanding in terms of degrees of freedom (see discussion on Fig. 4). Estimates are very stable across specifications. This stability of the coefficients is very interesting, as each specification controls somewhat for different versions of migration policies that may affect the results. Ortega and Peri (2013) show that a specification like the one in Column (4)—which includes country of origin × year dummies—emerges from a version of the random utility model in Grogger and Hanson (2011) extended to allow for individual-specific time-invariant random effects in the specification of the idiosyncratic utility function.

A problem of having some observations aggregated in grouped categories is that, since we only observe the stock of immigrants for the group, the dependent variable is measured with error provided that the log of the average stock of the group is not equal to the average of logs of bilateral stocks. The problem with this measurement error is that it is obviously correlated with the covariates. In order to check to what extent this could be a relevant issue, in Column (7) I include as controls standard deviations of the regressors within the grouped observations (zero for bilateral observations). Given that the measurement error increases as the countries in the grouped observation differ in the stock of immigrants, these standard deviations are good proxies for the measurement error.[18] Results are again robust; none of the coefficients of the regressors of interest is statistically different from its counterpart in Column (1), and the test of the joint difference cannot reject the null hypothesis that all coefficients are (jointly) equal to point estimates in Column (1)—the p-value of the test is reported in the note of Table 5.

As noted in Sect. 3.1, several studies in the literature estimate equations similar to (3) including GDP per capita in logs. For comparability with these studies, I run the same specifications of Table 5 using log GDP per capita instead of the levels. Results are presented in Table 8 in Appendix 1. Point estimates are in line with results in Table 5 and those in the literature. Estimated elasticities of GDP per capita in destination countries are around 0.6, slightly smaller but not very different from the average elasticity implied by the coefficients in Table 5. Those for GDP per capita at origin are rather small. The coefficients of other variables are virtually unchanged. However, the precision of estimated elasticities for GDP per capita at origin and destination is substantially lower.

To keep with the comparison between the database presented here and that by Özden et al. (2011), Table 9 in Appendix 1 replicates the regressions presented in Table 5 using the dataset produced by these authors. Sample sizes are obviously larger, given that observations for grouped countries are imputed to specific countries. While qualitatively similar, point estimates are somewhat different to those in Table 5. The estimated coefficient for GDP per capita at destination, for example, is 0.015 (s.e. 0.005) instead of 0.052 (s.e. 0.023) in Table 5. The elasticity of distance is around -1.1 instead of 0.9. And so on. These differences are more likely attributable to differences in the data collected than to the grouping itself. The motivation for this belief is that the coefficients of Column (3) in each table, which are only estimated for the subsample of observations with bilateral information in both datasets, are also quite different.

[18] Given that the logarithm is a concave function, by the Jensen inequality the logarithm of the average of the group is larger than the average of the logarithm, unless all elements are equal.

As a final robustness check, I also estimate the regressions in Table 5 excluding observations for 1960, and 1960 plus 1970 (results are available upon request from the author). We have seen in Sect. 2 that data is particularly grouped in these years (especially the first), and that data reliability is slightly lower in those years (see Footnote 9). Results are again robust.

4.2 Heterogeneous effects of income gains depending on distance

Estimates for Eq. (5) are presented in Table 6. As the interacted terms in Eq. (5) are expressed in differences from sample means, the linear terms can be interpreted as effects for the average country pair (and they are comparable to estimates in Sect. 4.1). Again, all regressions include at least origin and destination country fixed effects, and year dummies.

Column (1) in Table 6 is the baseline specification. The effect of destination country GDP per capita for the average country is exactly the same as in Table 5. The effect of GDP per capita at origin country is slightly more negative (−0.014 vs. −0.008), but still not statistically different from zero. The coefficients of all other regressors that are included in Table 5 are virtually unchanged (except for the point estimate of common border, that now becomes large and significant).

As the coefficient of the interaction of destination country GDP per capita and distance suggests, the effect of income gains on moving prospects is not homogeneous across all origin countries. These coefficients are interpreted as follows: the effect of a 1000$ increase in GDP per capita of the destination country is 0.21 percentage points smaller if the distance from the origin country is a 10 percent larger than the average. To give a sense to these numbers, note that the distance between Washington DC (US) and Dublin (Ireland) is 5,448 km, roughly the average distance in the sample. On the other hand, the distance between Washington DC and Beijing (China) is 11,159 km, roughly twice as large. Therefore, a 1000$ increase in GDP per capita in the US would increase the stock of Irish living in the US by around a 5.2 %, whereas the stock of Chinese-born would only be increased by approximately 3.1 %. As an extreme example, a 1000$ increase in GDP per capita in the US would increase the stock of Mexicans by a 8 %, but the stock of Taiwanese would be increased by only a 2.8 %.

This is the main empirical result of this paper. Previous literature assumes that an income shock in a destination country increases the stock of immigrants from all origin countries by the same percentage. If that were the case, then income shocks would not affect the composition of the immigrant population. But the finding described above indicates that income shocks in a destination country have indeed very important compositional effects. This result is very important for shaping immigration policy. For example, if the policy maker is willing to preserve the ethnic mix (e.g. it was one of the goals of the US immigration policy from 1920s to mid-1960s), countermeasures will be required to compensate market forces. Additionally, if the skill composition of immigrants from a particular country of origin was not affected by changes in the size of the flow, income shocks would affect the skill composition of the immigrant workforce by changing the weight of each origin country in the total stock.

Table 6 Heterogeneous effects of income gains

	(1)	(2)	(3)	(4)	(5)	(6)
GDPpc dest.	0.052	0.046	0.062	0.044		0.050
	(0.022)	(0.022)	(0.014)	(0.022)		(0.024)
GDPpc dest. × log distance	−0.021	−0.023	−0.012	−0.019	−0.018	−0.030
	(0.005)	(0.006)	(0.004)	(0.007)	(0.005)	(0.011)
GDPpc dest. × common lang.		−0.005	−0.022	−0.020	−0.005	0.033
		(0.022)	(0.009)	(0.026)	(0.015)	(0.038)
GDPpc dest. × colonial rel.		−0.065	−0.019	−0.070	−0.028	−0.082
		(0.025)	(0.012)	(0.028)	(0.016)	(0.039)
GDPpc dest. × common border		−0.011	−0.004	−0.001	0.005	−0.049
		(0.023)	(0.017)	(0.024)	(0.020)	(0.046)
GDPpc origin	−0.014	−0.013	−0.014		−0.015	−0.020
	(0.009)	(0.009)	(0.006)		(0.008)	(0.011)
GDPpc origin × log distance	0.032	0.029	0.029	0.031	0.029	0.022
	(0.005)	(0.005)	(0.004)	(0.007)	(0.004)	(0.011)
GDPpc origin × common lang.		−0.007	−0.002	−0.001	−0.006	−0.041
		(0.012)	(0.009)	(0.012)	(0.009)	(0.030)
GDPpc origin × colonial rel.		−0.020	−0.041	−0.030	−0.022	0.004
		(0.010)	(0.008)	(0.014)	(0.009)	(0.031)
GDPpc origin × common border		−0.013	−0.009	−0.010	−0.009	−0.008
		(0.020)	(0.015)	(0.021)	(0.018)	(0.050)
Log distance	−0.936	−0.929	−1.173	−0.989	−0.944	
	(0.080)	(0.082)	(0.052)	(0.095)	(0.057)	
Common language	0.606	0.580	0.850	0.612	0.587	
	(0.128)	(0.164)	(0.095)	(0.209)	(0.135)	
Colonial rel.	2.220	2.279	2.118	2.387	2.242	
	(0.142)	(0.164)	(0.104)	(0.196)	(0.137)	
Common border	0.373	0.542	0.387	0.423	0.457	
	(0.163)	(0.256)	(0.195)	(0.276)	(0.221)	
Log pop. origin	1.382	1.449	1.214		1.348	1.498
	(0.474)	(0.472)	(0.201)		(0.258)	(0.648)
Log pop. dest.	1.404	1.367	−1.968	1.273		1.372
	(1.109)	(1.124)	(0.463)	(1.215)		(1.426)
Grouped obs.	Yes	Yes	No	Yes	Yes	Yes
Time dummies	Yes	Yes	Yes	No	No	Yes
Origin dummies	Yes	Yes	Yes	No	Yes	No
Origin-time dummies	No	No	No	Yes	No	No
Destination dummies	Yes	Yes	Yes	Yes	No	No
Dest.-time dummies	No	No	No	No	Yes	No
Country-pair dummies	No	No	No	No	No	Yes
Obs	7340	7340	6727	7340	7340	7340
$\bar{R}^2$	0.959	0.959	0.967	0.959	0.976	0.961

Standard errors, clustered at the origin-time level, in parentheses. Dependent variable: log migrant stocks. Unit of observation: origin-destination-time. Regressions include the indicated fixed effects. The p-value of a test of the null that coefficients displayed in Column (3) are jointly equal to point estimates in Column (2) is 0.000, and the p-value for a test of the null hypothesis that the interaction coefficients in Column (1) are equal in magnitude and opposite sign is 0.135

A similar story can be told for origin countries' GDP per capita. Despite linear effects are small and statistically insignificant, the interaction with distance is very important. Interestingly, the coefficient of this interaction is very similar—with the opposite sign—to the one for interaction of GDP per capita of the destination country and distance (indeed, we cannot reject statistically that their magnitudes are the same—p-value is reported in the table notes). This result, together with the small estimated coefficient for the linear term, are again suggestive of the presence of an additional effect of origin country GDP per capita on migration prospects. Following with the argument of borrowing constraints, imperfect access to credit markets in poorer countries would prevent migrants from these countries to afford the migration cost, although they would have gained from moving if they could have borrowed resources to afford it; if that were the case, credit market imperfections would increase the coefficient of the linear term (making it less negative), but would not affect the interaction term. Similarly, another positive direct effect of origin country GDP per capita could arise through immigration policies, if destination countries are more willing to accept immigrants from richer countries (which again would not affect the interaction term).[19]

The remaining columns of Table 6 check the stability of the estimates across different versions of the same equation. In Column (2) I extend Eq. (5) by including interactions of origin and destination country GDP per capita with all other measures of distance. Results are virtually unchanged. Only interactions with colonial relationship are significant. Surprisingly, both of them have a negative sign. This result, however, may be driven by policy issues as one would expect that (after controlling for having a common language) a past colonial relationship only affects migration through a special treatment by destination countries in terms of immigration policy. For example, a negative income shock would reduce the stock of immigrants from non-former colonies in a larger magnitude than from former colonies, which would receive a special treatment.

In Column (3), I check the importance of including the 100 % of migrant stocks by dropping grouped observations. As in Table 5, qualitative results hold, but point estimates are different. In particular, seven coefficients are statistically different from their counterparts in Column (2), and a Wald test of the hypothesis that all coefficients are equal to their counterparts in Column (2) clearly rejects (p-value in the table notes).

As in Table 5, in columns (4) to (6) I change the specification of fixed effects. Again, on top of origin, destination, and time fixed effects (as in columns (1) to (3)), I introduce destination $\times$ time, origin $\times$ time, and country pair dummies respectively. Once again, results are virtually unchanged.

Table 10 in Appendix 2 reproduces the regressions in Table 6 introducing GDP per capita in logs instead of levels. Again, linear coefficients and the coefficients of the cost proxies are virtually unchanged with respect to their counterparts in Table 8 (except, as in Table 6 vs. Table 5, for the coefficient of common border, and, in this case,

[19] One could argue that immigration policy is softer in destination countries in "good periods". This would tend to produce a larger linear effect of destination country GDP per capita, but it would not affect the interaction term.

also Column (6)). And again, interaction terms explain a similar story as in Table 6, similar relative magnitude compared to the linear term, and similar in size for origin and destination, with opposite sign. The interaction with colonial relationship is also significant and in the same relative size and magnitude compared to the linear term, and it is also of the same sign when interacted with GDP per capita at origin and at destination. And interactions with the other variables are not statistically significant.

I also estimate again the same regressions using the data from Özden et al. (2011). Results are presented in Table 11 in Appendix 2. As it occurred in the previous section, results are somewhat different than in Table 6. The key difference is for the destination country GDP per capita, which not only is small and insignificant in the linear term, but now also in the interaction term. Instead, results for GDP per capita at origin are qualitatively in line with those in Table 6, but with very different magnitudes. And, as it happened with Table 9, results are still very different even in the case where only observations with bilateral information in both datasets are included.

As final robustness, I estimate the same regressions excluding 1960, and excluding 1960 and 1970 (results are available upon request from the author). Results are robust.

4.3 Additional results for other *push* and *pull* factors

In Table 7, I extend Eq. (5) to control for other *push* and *pull* determinants of migration in more detail. Specifically, I add unemployment rate, age dependency ratio (older than 65 over working-age population), and government consumption share of GDP (pull factors), and wars, political regimes, and young population at origin (push factors). Finally, I also add too specifications that are very demanding because of the grouping of the data: I control for networks (stock of immigrants from a given origin in a given destination in the preceding census), and I estimate a regression in flows, computed as difference in stocks. Overall, the main results from previous sections are generally stable across specifications.

Aside from income gains, individuals value their probability of finding a job in the destination country. For this reason, higher unemployment at the country of destination reduces migration. Column (1) shows this empirically by including unemployment rate in the regression. Its effect is estimated to be negative, as expected, and very significant. Column (2) includes age dependence ratio as a regressor. Countries with older populations are more willing to admit immigrants to increase social security revenues and sustain increasingly unbalanced pay-as-you-go systems. Additionally, an older population brings in additional work opportunities for immigrants, both in terms of elderly caring services and because of a lower competition in the labor market. The coefficient of this variable has the expected positive sign, although its effect is small and statistically not different from zero. In Column (3) I include the government consumption as a share of GDP. More generous welfare state governments will spend more, and will attract more immigrants. However, larger government expenditure implies higher tax rates, and this may discourage migration. If all countries were equally efficient in their spending, the sign of the effect should depend on whether immigrants are net contributors or receivers. In that case, South-North migration should be affected positively by expenditure. However, larger expenditures in some countries may be due to lower

Table 7 Additional results

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
GDPpc dest.	0.040	0.052	0.047	0.052	0.050	0.046	0.027	0.011	0.133
	(0.016)	(0.023)	(0.024)	(0.022)	(0.022)	(0.022)	(0.017)	(0.015)	(0.032)
GDPpc dest. × log dist.	−0.021	−0.022	−0.022	−0.021	−0.020	−0.023	−0.023	−0.015	0.007
	(0.005)	(0.006)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)	(0.011)
Unemp rate dest.	−0.106						−0.085		
	(0.028)						(0.028)		
Age dep. dest.		0.015					0.044		
		(0.031)					(0.033)		
Gov. share dest.			−0.083				−0.067		
			(0.048)				(0.067)		
GDPpc origin	−0.018	−0.014	−0.014	−0.013	−0.013	−0.021	−0.021	−0.026	−0.066
	(0.010)	(0.009)	(0.009)	(0.009)	(0.009)	(0.010)	(0.010)	(0.016)	(0.024)
GDPpc origin × log dist.	0.033	0.032	0.031	0.031	0.032	0.033	0.034	0.020	0.027
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)	(0.005)	(0.011)
War origin				0.762			0.618		
				(0.178)			(0.146)		
PolityIV origin					−0.003		−0.011		
					(0.019)		(0.013)		
PolityIV^2 origin					−0.003		−0.003		
					(0.001)		(0.001)		
Log 15–34 pop.						4.583	3.388		
						(1.137)	(0.771)		
Log distance	−0.963	−0.936	−0.936	−0.937	−0.939	−0.989	−0.992	−0.624	−1.503
	(0.076)	(0.080)	(0.077)	(0.080)	(0.082)	(0.080)	(0.079)	(0.099)	(0.179)
Common language	0.635	0.606	0.605	0.606	0.615	0.608	0.630	0.416	0.810
	(0.116)	(0.128)	(0.127)	(0.128)	(0.128)	(0.126)	(0.114)	(0.131)	(0.215)

Table 7 continued

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Colonial rel.	2.196 (0.143)	2.220 (0.142)	2.218 (0.140)	2.224 (0.142)	2.155 (0.141)	2.412 (0.145)	2.264 (0.136)	1.243 (0.226)	1.186 (0.304)
Common border	0.320 (0.158)	0.373 (0.163)	0.369 (0.160)	0.367 (0.163)	0.401 (0.164)	0.311 (0.167)	0.331 (0.158)	0.108 (0.131)	−0.241 (0.337)
Networks								0.412 (0.061)	
Log pop. origin	1.234 (0.329)	1.388 (0.477)	1.392 (0.481)	1.356 (0.464)	1.368 (0.387)	−3.527 (1.314)	−2.498 (0.886)	0.063 (0.449)	−1.708 (0.917)
Log pop. dest.	0.998 (1.395)	1.475 (1.114)	0.569 (0.961)	1.399 (1.109)	1.403 (1.093)	1.417 (1.039)	0.211 (1.278)	2.925 (1.547)	7.246 (2.856)
Grouped obs.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Origin dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Destination dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Obs	7093	7340	7338	7340	7144	7149	6705	4483	1765
$\bar{R}^2$	0.965	0.959	0.957	0.959	0.959	0.961	0.966	0.978	0.959

Standard errors, clustered at the origin-time level, in parentheses. Dependent variable: log migrant stocks, except in Column (9), which is log flows. Unit of observation: origin-destination-time. Regressions include the indicated fixed effects

efficiency, which might imply that everyone becomes a net contributor, making the effect unambiguously negative. Results in Table 7 suggest that the effect is negative.

Column (4) includes a warfare measure for the origin country. This variable measures the share of months over the last decade that the country was involved in a war of any type. Armed conflicts displace a lot of people who escape from the tragedy. This fact is reflected in the estimates: a decade of war in an origin country increases the stock of immigrants from that country in a 76 %. The political regime may also be important for migration. People may be less willing to leave a good democracy (everything else constant); moreover, in a dictatorship, they are usually not allowed to escape from the country. Instead of weak central authorities (known as anocracies) may be an encouraging environment for migration. In Column (5), I introduce the Polity IV index, which ranges from −10 (autocracy) to 10 (democracy). Intermediate values (with small absolute values) indicate the presence of an anocracy. For this reason, I include a quadratic in the indicator. The quadratic term is negative and significantly different from zero. The linear term is negative but small and clearly insignificant, indicating that similarly fewer people migrate from autocracies than from democracies compared to anocracies. In particular, the stock of migrants is around 30 % lower if the origin country is a democracy or an autocracy relative to an anocracy. Column (6) introduces the log of the population at the origin country. Countries with larger young populations (relative to the total population) tend to send more migrants abroad. Specifically, holding total population constant, an extra 1 % of population of those ages increases migration by 4.6 %. Column (7) introduces all push and pull factors together without any significant change.

The remaining two columns estimate respectively a regression that includes the lagged bilateral stock as a control, and one that uses log flows as the dependent variable. The fundamental problem to estimate these equations is that, given that the grouping affects differently each census, observations need to be artificially grouped further so that groups coincide over two consecutive censuses. This reduces observations substantially, and increases the incidence of grouping. Despite that, results in Column (8) are very similar to the estimates presented above. Additionally, an extra 1 % in the stock of migrants in a country-pair in the preceding decade is associated with a 0.4 % extra stock of immigrants in the given census. Column (9), which is estimated only with 1,765 observations delivers results that are qualitatively (and, with exceptions, quantitatively) similar to previous specifications, even though precision is affected substantially.

5 Conclusions

In this paper I present a new database of bilateral migrant stocks, and I provide new evidence on the determinants of bilateral migration. The database introduced in this paper was collected from the National Statistical Offices from 24 OECD countries based on population censuses. For each destination country and census date, it covers 188 countries of origin (sometimes in a grouped category) for the period 1960 to 2000. The database fully covers the total stock of immigrants, keeping track of the residual categories reported by Statistical Offices instead of making imputations to specific countries of origin. I handle these grouped data in a raw manner in the estimation.

Empirically, I test for the existence of non-linear effects of income gains on migration prospects depending on distance. The motivation for such heterogeneity can be cost-based (individuals from closer countries can move back and forth as a consequence of income fluctuations, whereas it is more costly for individuals from farther away countries), or by means of a compensating wage differential (individuals dislike living far away from home, and require a compensating wage differential to move, that would increase with distance). Results suggest that this heterogeneity is indeed very marked. For example, a 1000$ increase in US income per capita would increase the stock of Mexican immigrants in the US by a 8 %, the stock of Irish immigrants by a 5.2 %, and the stock of Chinese-born by only a 3.1 %. This result is very robust across many different specifications.

Empirical findings in this paper suggest that income shocks have significant compositional effects, which are important for shaping immigration policy. For example, if a policy maker is willing to preserve the ethnic mix (e.g. it was one of the goals of the US immigration policy from 1920s to mid-1960s), countermeasures will be required to compensate market forces. If country of origin is a good proxy for skills of immigrants, this result would also have implications for the skill composition of migrants. Additionally, destination countries should be more concerned about income shocks in neighboring countries than what is suggested in the literature, and may want to trade off development assistance and migration policies as a result.

A few remarks need to be made on the conclusions of this paper. The first one is regarding the grouping of the data. There are 1,800 country pairs (out of $24 \times 188 = 4512$) for which I observe data in grouped categories for all years (which only allows me to identify a fixed effect for each group). Also, for a similar reason, there are 160 origin country $\times$ time dummies that cannot be individually identified (out of $188 \times 5 = 940$). And data grouping also complicates the incorporation of the role of networks in determining international migration, and the estimation of models in flows (as shown in Table 7). A second remark is that the database does not include information on educational attainment by immigrants. Such information would be useful to test whether the compositional effects that I observe with respect to nationality have important implications for skill composition of immigrants. To the best of my knowledge, Docquier and Marfouk (2006), Docquier et al. (2009), and Brücker et al. (2013) are the only databases in the literature that include such information, but they only cover a shorter period. Third, the regressions estimated in this model abstract from the role of trade. Part of migration flows can be equilibrium adjustments to trade (as in di Giovanni et al. 2014). And fourth, the results in this paper can be seen as an additional explanation to those covered in Clemens (2014) as to why the elasticity of migration with respect to GDP per capita at origin is not homogeneous across countries (he finds evidence of an inverse U-shape).

The paper also opens avenues for future research. It would be interesting to investigate how the heterogeneous effects found in this paper affect skill composition and self-selection of migrants. Likewise, the database presented in this paper can be used for a variety of cross-country migration analyses (e.g., to produce instrumental variables as in Llull 2011 or Ortega and Peri 2014a).

Appendix 1: Robustness: linear effects

See Tables 8 and 9.

Table 8 Linear effects—introducing GDP per capita in logs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Log GDPpc dest.	0.607		1.389	0.581		0.628	0.831
	(0.752)		(0.293)	(0.774)		(0.920)	(0.507)
Log GDPpc origin	0.074		−0.236		−0.169	0.160	0.020
	(0.197)		(0.089)		(0.186)	(0.282)	(0.214)
Log GDPpc gap		0.190					
		(0.330)					
Log distance	−0.907	−0.903	−1.051	−0.931	−0.909		−0.837
	(0.074)	(0.075)	(0.047)	(0.079)	(0.056)		(0.071)
Common language	0.588	0.585	0.771	0.591	0.581		0.612
	(0.129)	(0.131)	(0.081)	(0.143)	(0.096)		(0.120)
Colonial rel.	2.284	2.281	2.117	2.285	2.269		2.375
	(0.146)	(0.146)	(0.100)	(0.157)	(0.123)		(0.122)
Common border	0.025	0.033	0.044	−0.004	0.028		0.227
	(0.178)	(0.181)	(0.126)	(0.188)	(0.151)		(0.165)
Log pop. origin	1.531	1.387	1.523		1.414	1.571	1.946
	(0.487)	(0.502)	(0.158)		(0.235)	(0.635)	(0.430)
Log pop. dest.	0.834	0.880	−1.879	0.764		0.763	−0.551
	(1.068)	(1.151)	(0.426)	(1.059)		(1.243)	(0.743)
Grouped obs.	Yes	Yes	No	Yes	Yes	Yes	Yes
St. devs. of controls	No	No	No	No	No	No	Yes
Time dummies	Yes	Yes	Yes	No	No	Yes	Yes
Origin dummies	Yes	Yes	Yes	No	Yes	No	Yes
Origin-time dummies	No	No	No	Yes	No	No	No
Destination dummies	Yes	Yes	Yes	Yes	No	No	Yes
Dest.-time dummies	No	No	No	No	Yes	No	No
Country-pair dummies	No	No	No	No	No	Yes	No
Obs	7340	7340	6727	7429	7340	7340	7332
$\bar{R}^2$	0.958	0.958	0.966	0.956	0.975	0.960	0.966

Standard errors, clustered at the origin-time level, in parentheses. Dependent variable: log migrant stocks. Unit of observation: origin-destination-time. Regressions include the indicated fixed effects

Table 9 Linear effects Özden et al. (2011) data

	(1)	(2)	(3)	(4)	(5)	(6)
GDPpc dest.	0.015		0.009	0.014		0.011
	(0.005)		(0.015)	(0.005)		(0.004)
GDPpc origin	0.002		−0.008		0.002	0.002
	(0.004)		(0.006)		(0.004)	(0.004)
GDPpc gap		−0.000				
		(0.003)				
Log distance	−1.138	−1.137	−1.047	−1.101	−1.135	
	(0.039)	(0.039)	(0.051)	(0.038)	(0.039)	
Common language	0.905	0.904	0.903	0.831	0.924	
	(0.068)	(0.069)	(0.093)	(0.068)	(0.068)	
Colonial rel.	1.945	1.945	1.999	1.833	1.968	
	(0.099)	(0.099)	(0.132)	(0.095)	(0.100)	
Common border	0.375	0.375	0.046	0.543	0.354	
	(0.121)	(0.120)	(0.137)	(0.128)	(0.124)	
Log pop. origin	0.938	0.901	1.669		0.938	0.938
	(0.135)	(0.129)	(0.169)		(0.135)	(0.152)
Log pop. dest.	0.127	0.052	−0.245	0.254		−0.116
	(0.278)	(0.278)	(0.502)	(0.256)		(0.147)
Grouped obs.	Yes	Yes	No	Yes	Yes	Yes
Time dummies	Yes	Yes	Yes	No	No	Yes
Origin dummies	Yes	Yes	Yes	No	Yes	No
Origin-time dummies	No	No	No	Yes	No	No
Destination dummies	Yes	Yes	Yes	Yes	No	No
Dest.-time dummies	No	No	No	No	Yes	No
Country-pair dummies	No	No	No	No	No	Yes
Obs	20,232	20,232	6727	22,440	20,232	20,232
$\bar{R}^2$	0.906	0.906	0.953	0.899	0.914	0.949

Standard errors, clustered at the origin-time level, in parentheses. Dependent variable: log migrant stocks. Unit of observation: origin-destination-time. Regressions include the indicated fixed effects

Appendix 2: Robustness: heterogeneous effects of income gains

See Tables 10 and 11.

Table 10 Heterogeneous effects of income gains—GDP per capita in logs

	(1)	(2)	(3)	(4)	(5)	(6)
Log GDPpc dest.	0.675	0.593	1.403	0.729		0.568
	(0.755)	(0.752)	(0.299)	(0.777)		(0.918)
Log GDPpc dest. × log distance	−0.216	−0.264	−0.159	−0.171	−0.337	0.142
	(0.125)	(0.128)	(0.077)	(0.155)	(0.091)	(0.405)
Log GDPpc dest. × common lang.		−0.264	−0.518	−0.471	−0.204	0.582
		(0.273)	(0.163)	(0.371)	(0.272)	(0.534)
Log GDPpc dest. × colonial rel.		−1.215	−0.396	−1.384	−0.274	−2.176
		(0.466)	(0.194)	(0.517)	(0.267)	(1.213)
Log GDPpc dest. × common bord.		−0.128	−0.122	0.149	0.116	−1.903
		(0.450)	(0.347)	(0.497)	(0.373)	(1.785)
Log GDPpc origin	0.026	0.043	−0.344		−0.233	0.237
	(0.197)	(0.198)	(0.089)		(0.140)	(0.338)
Log GDPpc origin × log distance	0.355	0.336	0.375	0.388	0.402	−0.443
	(0.057)	(0.061)	(0.035)	(0.061)	(0.033)	(0.352)
Log GDPpc origin × common lang.		0.024	0.047	0.032	0.011	−0.303
		(0.088)	(0.058)	(0.089)	(0.060)	(0.513)
Log GDPpc origin × colonial rel.		−0.274	−0.366	−0.329	−0.251	0.290
		(0.088)	(0.063)	(0.095)	(0.067)	(0.695)
Log GDPpc origin × common bord.		−0.202	−0.223	−0.153	−0.276	0.677
		(0.270)	(0.209)	(0.275)	(0.236)	(1.602)
Log distance	−1.119	−1.098	−1.339	−1.201	−1.137	
	(0.087)	(0.090)	(0.060)	(0.090)	(0.058)	
Common language	0.599	0.638	0.942	0.684	0.634	
	(0.125)	(0.148)	(0.099)	(0.196)	(0.140)	
Colonial rel.	2.212	2.371	2.173	2.510	2.240	
	(0.143)	(0.169)	(0.108)	(0.203)	(0.139)	
Common border	0.444	0.759	0.704	0.616	0.895	
	(0.172)	(0.380)	(0.294)	(0.390)	(0.324)	
Log pop. origin	1.423	1.500	1.359		1.390	1.775
	(0.486)	(0.486)	(0.173)		(0.227)	(0.676)
Log pop. dest.	0.800	0.788	−1.754	0.560		0.674
	(0.979)	(0.953)	(0.456)	(0.913)		(1.039)
Grouped obs.	Yes	Yes	No	Yes	Yes	Yes
Time dummies	Yes	Yes	Yes	No	No	Yes
Origin dummies	Yes	Yes	Yes	No	Yes	No
Origin-time dummies	No	No	No	Yes	No	No
Destination dummies	Yes	Yes	Yes	Yes	No	No
Dest.-time dummies	No	No	No	No	Yes	No
Country-pair dummies	No	No	No	No	No	Yes
Obs	7340	7340	6727	7340	7340	7340
$\bar{R}^2$	0.959	0.959	0.968	0.959	0.977	0.961

Standard errors, clustered at the origin-time level, in parentheses. Dependent variable: log migrant stocks. Unit of observation: origin-destination-time. Regressions include the indicated fixed effects

Table 11 Heterogeneous effects Özden et al. (2011) data

	(1)	(2)	(3)	(4)	(5)	(6)
GDPpc dest.	0.017	0.007	−0.012	0.005		0.012
	(0.005)	(0.005)	(0.015)	(0.005)		(0.004)
GDPpc dest. × log distance	−0.000	−0.002	−0.019	−0.011	−0.003	0.003
	(0.003)	(0.003)	(0.005)	(0.003)	(0.003)	(0.003)
GDPpc dest. × common lang.		−0.039	−0.051	−0.050	−0.026	0.017
		(0.006)	(0.010)	(0.007)	(0.007)	(0.007)
GDPpc dest. × colonial rel.		−0.054	−0.044	−0.055	−0.044	−0.037
		(0.013)	(0.017)	(0.014)	(0.014)	(0.009)
GDPpc dest. × common border		−0.003	0.010	−0.009	−0.001	−0.015
		(0.015)	(0.017)	(0.016)	(0.017)	(0.013)
GDPpc origin	−0.004	−0.002	−0.015		−0.002	−0.003
	(0.003)	(0.004)	(0.004)		(0.004)	(0.004)
GDPpc origin × log distance	0.021	0.020	0.031	0.019	0.020	0.011
	(0.004)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)
GDPpc origin × common lang.		0.002	−0.002	0.004	0.004	−0.016
		(0.008)	(0.009)	(0.008)	(0.008)	(0.011)
GDPpc origin × colonial rel.		−0.020	−0.034	−0.020	−0.024	0.005
		(0.005)	(0.012)	(0.005)	(0.005)	(0.006)
GDPpc origin × common border		−0.024	−0.008	−0.028	−0.024	−0.016
		(0.018)	(0.017)	(0.018)	(0.019)	(0.014)
Log distance	−1.185	−1.190	−1.165	−1.178	−1.184	
	(0.038)	(0.039)	(0.057)	(0.039)	(0.039)	
Common language	0.918	0.962	1.124	0.977	0.961	
	(0.068)	(0.073)	(0.115)	(0.075)	(0.074)	
Colonial rel.	1.925	1.889	1.973	1.873	1.926	
	(0.098)	(0.098)	(0.135)	(0.099)	(0.098)	
Common border	0.574	0.739	0.353	0.769	0.721	
	(0.117)	(0.185)	(0.206)	(0.188)	(0.196)	
Log pop. origin	0.785	0.817	1.549		0.824	0.793
	(0.129)	(0.134)	(0.185)		(0.134)	(0.150)
Log pop. dest.	−0.044	0.025	0.004	0.313		−0.436
	(0.280)	(0.282)	(0.514)	(0.287)		(0.168)
Grouped obs.	Yes	Yes	No	Yes	Yes	Yes
Time dummies	Yes	Yes	Yes	No	No	Yes
Origin dummies	Yes	Yes	Yes	No	Yes	No
Origin-time dummies	No	No	No	Yes	No	No
Destination dummies	Yes	Yes	Yes	Yes	No	No
Dest.-time dummies	No	No	No	No	Yes	No
Country-pair dummies	No	No	No	No	No	Yes
Obs	20,232	20,232	6,727	20,232	20,232	20,232
$\bar{R}^2$	0.907	0.907	0.955	0.908	0.915	0.949

Note: Standard errors, clustered at the origin-time level, in parentheses. Dependent variable: log migrant stocks. Unit of observation: origin-destination-time. Regressions include the indicated fixed effects

References

Abel GJ, Sander N (2014) Quantifying global international migration flows. Science 343(6178):1520–1522

Adserà A, Pytliková M (2012) The role oflanguage in shaping international migration. NORFACE Migration Discussion Paper No. 2012-14 (2012)

Alesina AF, Devleeschauwer A, Easterly WE, Kurlat S, Wacziarg R (2003) Fractionalization. J Econ Growth 8(2):155–194

Beine M, Bertoli S, Fernández-Huertas Moraga J (2015) A practitioners guide to gravity models of international migration. World Econ 2015:1–15

Bertoli S, Fernández-Huertas Moraga J (2013) Multilateral resistance to migration. J Dev Econ 102(C):79–100

Bertoli S, Fernández-Huertas Moraga J (2013) Crossing the border: self-selection, earnings and individual migration decisions. J Dev Econ 101(C):75–91

Borjas GJ (1987) Immigrants, minorities, and labor market competition. Ind Labor Relat Rev 40(3):382–392

Borjas GJ, Bratsberg B (1996) Who leaves? The outmigration of the foreign-born. Rev Econ Stat 78(1):165–176

Brücker H, Stella C, Abdeslam M (2013) Education, gender and international migration: insights from a panel-dataset 1980–2010. Mimeo IAB

Chiquiar D, Hanson GH (2005) International migration, self-selection, and the distribution of wages: evidence from Mexico and the United States. J Polit Econ 113(2):239–281

Clark X, Hatton TJ, Williamson JG (2007) Explaining U.S. immigration, 1971–1998. Rev Econ Stat 89(2):359–373

Clemens MA (2014) Does development reduce migration? In: Lucas REB (ed) International handbook on migration and economic development. Edward Elgar Publishing, Cheltenham

di Giovanni J, Levchenko AA, Ortega F (2014) A global view of cross-border migration. J Eur Econ Assoc 13(1):168–202

Docquier F, Marfouk A (2006) International migration by educational attainment, 1990–2000, chapter 5. In: Özden Ç, Schiff MW (eds) International migration, remittances and the brain drain. Palgrave Macmillan, New York, pp 151–200

Docquier F, Lowell BL, Marfouk A (2009) A gendered assesment of highly skilled emigration. Popul Dev Rev 35(2):297–321

Grogger JT, Hanson GH (2011) Income maximization and the selection and sorting of international migrants. J Dev Econ 95(1):42–57

Karemera D, Oguledo VI, Davis B (2000) A gravity model analysis of international migration to North America. Appl Econ 32(13):1745–1755

Kennan J, Walker JR (2011) The effect of expected income on individual migration decisions. Econometrica 79(1):211–251

Lessem RH (2013) Mexico-U.S. immigration: effects of wages and border enforcement. Mimeo, Carnegie Mellon University,

Llull J (2011) Reconciling spatial correlations and factor proportions: a cross-country analysis of the economic consequences of immigration. Mimeo, CEMFI

Mayda AM (2010) International migration: a panel data analysis of the determinants of bilateral flows. J Popul Econ 23(4):1249–1274

Ortega F, Peri G (2013) The effect of income and immigration policies on international migration. Migr Stud 1(1):47–74

Ortega F, Peri G (2014) Openness and income: the roles of trade and migration. J Int Econ 92(2):231–251

Ortega F, Peri G (2014) The aggregate effects of trade and migration: evidence from OECD countries. In: Zimmerman KF, Cigno A, Tekin E, Zhang J (eds) The socio-economic impact of migration flows. Springer, Switzerland

Özden Ç, Parsons CR, Schiff M, Walmsey TL (2011) Where on earth is everybody? The evolution of global bilateral migration 1960–2000. World Bank Econ Rev 25(1):12–56

Pedersen PJ, Pytlikova M, Smith N (2008) Selection and network effects—migration flows into OECD countries 1990–2000. Eur Econ Rev 52(7):1160–1186

Rose AK (2004) Do we really know that the WTO increases trade? Am Econ Rev 94(1):98–114

United Nations (2013) Trends in international migrant stock: the2013 revision (CD-ROM edition). Population division, UnitedNations, New York