The Community Demographic Model International Migration (CDM-IM) Dataset: Generating Age and Gender Profiles of International Migration Flows

Raphael J. Nawrotzki
Leiwen Jiang

NCAR Technical Notes

National Center for
Atmospheric Research
P. O. Box 3000
Boulder, Colorado
80307-3000
www.ucar.edu

NCAR/TN-508+STR

NCAR TECHNICAL NOTES

http://library.ucar.edu/research/publish-technote

The Technical Notes series provides an outlet for a variety of NCAR Manuscripts that contribute in specialized ways to the body of scientific knowledge but that are not yet at a point of a formal journal, monograph or book publication. Reports in this series are issued by the NCAR scientific divisions, serviced by OpenSky and operated through the NCAR Library. Designation symbols for the series include:

EDD – Engineering, Design, or Development Reports
Equipment descriptions, test results, instrumentation, and operating and maintenance manuals.

IA – Instructional Aids
Instruction manuals, bibliographies, film supplements, and other research or instructional aids.

PPR – Program Progress Reports
Field program reports, interim and working reports, survey reports, and plans for experiments.

PROC – Proceedings
Documentation or symposia, colloquia, conferences, workshops, and lectures. (Distribution maybe limited to attendees).

STR – Scientific and Technical Reports
Data compilations, theoretical and numerical investigations, and experimental results.

The National Center for Atmospheric Research (NCAR) is operated by the nonprofit University Corporation for Atmospheric Research (UCAR) under the sponsorship of the National Science Foundation. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

National Center for Atmospheric Research
P. O. Box 3000
Boulder, Colorado 80307-3000

NCAR/TN-508+STR
NCAR Technical Note

April 2014

The Community Demographic Model International Migration (CDM-IM) Dataset: Generating Age and Gender Profiles of International Migration Flows

Raphael J. Nawrotzki
NCAR Earth System Laboratory, Climate & Global Dynamics Division, Terrestrial Sciences Section, Integrated Assessment Modeling group,
National Center of Atmospheric Research, Boulder, CO

Leiwen Jiang
NCAR Earth System Laboratory, Climate & Global Dynamics Division, Terrestrial Sciences Section, Integrated Assessment Modeling group,
National Center of Atmospheric Research, Boulder, CO

NCAR Earth System Laboratory
Climate & Global Dynamics Division

NATIONAL CENTER FOR ATMOSPHERIC RESEARCH
P. O. Box 3000
BOULDER, COLORADO 80307-3000
ISSN Print Edition 2153-2397
ISSN Electronic Edition 2153-2400

Table of Content

List of Figures

List of Tables

Acknowledgments

We gratefully acknowledge that this project received funding from the U.S. Department of Energy (DoE) grant No. 106952. We thank Ben Sanderson for his help in obtaining the raw data. We express our thanks to Jamie Jones for her careful editing and helpful suggestions. Finally, we are grateful for Bryan Jones comments on earlier drafts of this document.

1 Introduction

As the world has entered "the age of migration" (Castles and Miller 2003), increased numbers of people move across national boundaries with substantial socioeconomic and environmental impacts beyond any single nation (UNPD 2011). International migration is an increasingly important determinant of changes in the size and structure of national and regional population (Raymer et al. 2012, UNPD 2005). Policy makers require information regarding the changes in number of migrants and their demographic characteristics to improve planning of economic development and social services (Raymer et al. 2012). Moreover, to study the impacts of economic and environmental changes on human populations of different socio-demographic background, data on international migration is necessary (Brown 2008, Piguet et al. 2010). Ideally, this data would contain information on gender and age of the migrant population, allowing the analysis of migration patterns for various vulnerable sub-populations (e.g., young females).

The NCAR Community Demographic Model includes the multiregional population/urbanization projection module. This module needs to account for the impacts of international migration flows on population dynamics, and requires input of detailed information on age and gender specific international migration flows for the base year and for making assumptions about future changes. However, it is notoriously difficult to obtain high quality data on international migration for a number of reasons, such as inconsistent uses of definition and different criteria in enumerating migrants (Raymer et al. 2012), under-registration of migrants and limitation of data coverage (Nowok et al. 2006), governments' unwillingness to release data (UNECE 2012), and variations in data quality across countries (Abel 2013, Raymer et al. 2012). Censuses and surveys rarely measure migration directly. Often a comparison of places of residence at two points in time is used to approximate migration events (Raymer et al. 2012). However, differences in the years of data collection make the direct comparison of flow data between countries difficult.

Researchers usually measure the intensity of migration in terms of *migrant stock* or *migrant flow*. Migrant stock data is easier to measure and therefore more readily available than migrant flow data (Bilsborrow et al. 1997). However, migrant flow data better reflects the dynamics of the migration process and is more suitable for many demographic applications, such as projecting population behavior and change (Raymer et al. 2012). For example, the observed migrant stock constitutes a static measure of the migrant population at a particular point in time. The numbers of migrants in the stock is influenced by historical demographic processes, such as fertility, mortality, and return migration, and therefore, does not appropriately capture the true number of individuals who crossed a border during a particular time period. Obtaining a migration flow data set with global coverage would be particularly useful and would allow us to compare migration propensities across multiple countries, conduct more comprehensive testing of migration theories, and would help improve global population forecasts (Abel 2013).

International organizations and demographers have put great efforts in compiling migration data and studying the volume and characteristics of international migration, utilizing various indirect estimation methods. The UN Global Migration Database (UNGMD) (Zlotnik et al. 2010, UNDESA 2008), developed by the United Nation Population Division, is the most comprehensive international migration dataset currently available. This dataset, collected from censuses and other official statistical sources, is based on empirical records reflecting the number of international migrants by age and sex.

Based on the UNGMD database, the Development Research Group at the World Bank constructed the Global Bilateral Migration data set, constituting a table of migrant stock data for the period of 1960-2000 (Ozden et al. 2011). Using the World Bank migration stock tables in a log-linear regression framework, and adopting the correction factor approach developed by Poulain (1993, 1999), Abel (2013) estimated decennial migration flows across 191 countries to reflect the migration intensity during the period 1960 to 2000. However, his work does not provide information on gender and age composition of the migrant flows.

To study migrants' age and gender profiles, Rogers and co-authors developed the Model Migration Schedules (Rogers and Castro 1981; Rogers, Castro, and Lea 2005; Rogers, Raquillet, and Castro 1978). Employing a multi-exponential regression approach, Rogers et al. used the Model Migration Schedules to obtain age-gender profiles of the internal migrant flows across U.S. states and for international migration in several other countries (Rogers et al. 2007; Rogers and Raymer 1998; Raymer and Rogers 2007; Rogers and Wilson 1996; Rogers, Raymer and Willekens 2002).

Taking advantage of the better quality of the European migration data, two recent projects, the MIMOSA (MIgration MOdelling for Statistical Analyses) project and the IMEM (Integrated Modeling of Europe Migration) project attempted to estimate total numbers, along with age and gender profiles of international migration within Europe. The MIMOSA project, first harmonized the migration flows among 19 European countries, using Poulain's (1999) optimization procedure benchmarked to Sweden's migration flow data (de Beer et al. 2010). In a second step, the MIMOSA team estimated the missing marginal data within a hierarchical multiplicative framework (Raymer et al. 2011; Raymer 2007, 2008). In the IMEM project, Raymer et al. (2012) used a Bayesian modelling approach (Brierley et al. 2008) to estimate detailed migration flows by age and gender across European countries for the years 2002-2008.

The aforementioned modeling approaches and resulting datasets have substantially contributed to studies of international migration. However, a data set containing international migration flow information by age and gender with global coverage is missing. The current study, as part of the National Center of Atmospheric Research (NCAR) Community Demographic Model (CDM) project, attempts to substantiate this void. This paper reports on the methods and techniques used to generate the CDM International Migration (CDM-IM) dataset of age and gender specific international migrant flows around the year 2000.

2 Concepts and Data

2.1 Definitions

In this paper, we use the term “migration stream” or “stream” to define the unidirectional population migration between two countries or regions, moving from origin to destination. The “migrant stock” reflects the number of migrants “present in a given country at a particular point of time” (Abel 2013:506). In contrast, the “migration flow” adds a temporal dimension and measures the migration from an origin country to a destination country for one unit of time (e.g., per year) (Perruchoud and Redpath-Cross 2011).

2.2 Data

The raw data used in this study originates from the United Nations Global Migration Database (UNGMD); it records the migrant stock for a particular migration stream (e.g., Mexico – U.S.) at different points in time. The complete data set contains 46,431 useable files, covering 6,759 unique migration streams (about 7 files per stream). Migration streams operate at the national (n=6,193) and regional level (n=566) and report migrant stock data for various years. The migrant stock information for each recorded year is available in a separate .txt file, and typically contains information on: the country of origin and destination, the method of enumeration (country of birth vs. country of citizenship), the data source (e.g., demographic yearbook, census, register, survey, etc.), and year of enumeration. In addition, the migration stock data are disaggregated by age and gender groups. Figure 1 provides a visual depiction of a raw data file of migrant stock data for the stream from Mexico to the U.S. for the year 2000.

Figure 1: Raw data file for the migration stream from Mexico to the U.S. for the year 2000

Country or area of enumeration			United States of America
Country of birth or aggregate			Mexico
Criterion			Country of birth
Year			2000
Source			Census
Footnote			
Age Group	Male	Female	Total
Total	5,084,479	4,093,008	9,177,487
00-00	17,766	12,113	29,879
01-04	71,330	64,280	135,610
05-09	158,855	146,906	305,761
10-14	231,155	212,822	443,977
15-19	415,857	286,392	702,249
20-24	716,640	468,546	1,185,186
25-29	787,913	581,997	1,369,910
30-34	735,407	561,529	1,296,936
35-39	598,345	474,767	1,073,112
40-44	448,259	371,460	819,719
45-49	309,322	267,913	577,235
50-54	208,533	193,787	402,320
55-59	133,231	137,060	270,291
60-64	89,324	101,094	190,418
65-69	62,838	76,498	139,336
70-74	40,979	53,759	94,738
75-79	31,210	37,000	68,210
80-84	14,647	21,862	36,509
85-89	8,007	15,029	23,036
90-94	3,452	5,804	9,256
95-99	1,128	1,883	3,011
100+	281	507	788

However, the information contained in each file varies substantially between migration streams, as well as for different years of the same migration stream. For example, in some files the number of migrants is disaggregated by both age and gender, while others are disaggregated solely by either age or gender. In the most trying case, only the total number of migrants is reported. When age information is included, it frequently varies in format. Age information may be recorded evenly or unevenly in 1-year, 5-year, 10-year, 15-year, or even 20-year groups; in some cases, the recording follows the grouping method commonly used in demographic analysis (e.g., 0-4, 5-9, 10-14), while other files use unconventional groupings (e.g., 0-5, 5-10, 10-15); some files have summary groups included between regular age groups (e.g., 0-4, 5-9, 10-14, 15-19, 0-17, 20-24), but may miss age categories and may or may not have an open-ended category (e.g., 65+).

In addition, for confidentiality reasons, the UN replaced values between one and nine migrants by an asterisk (UNDESA 2008). For the same reason, countries of origin with fewer than 100 international migrants are not shown separately. As such, origins are sometimes

recorded at the regional level (e.g., Southern Europe) or as a combination of two or more countries (e.g., Greece and Italy). To overcome the difficulties caused by the irregular data format in the original files, and to generate a unified dataset, we performed several steps of data cleaning and processing described next.

3 Method

We used the programming language and statistical environment of the R project (R Core Team 2013) and adopted four steps to accomplish the data cleaning and processing: First, we assessed the overall data condition by generating an overview table from all raw data files (section 3.1); second, for each migration stream, we selected a file closest to the year 2000 with the most detailed age and gender information (Year 1), and selected a second file of the year close to, but different from Year 1 (Year 2) (section 3.2); third, we derived age and gender profiles of the migrant stock for both years (section 3.3); and finally, we used migrant stock information from the two years to compute fertility and mortality adjusted profiles of bilateral migration flows (section 3.4).

3.1 Assessment of overall data condition

From the original UNGMD data files, we extracted relevant information, including the country of destination (e.g., “Mexico”), the country of origin (e.g., “United States of America”), the year of enumeration (e.g., 2001), the data source (e.g., “Census”), the criterion of enumeration (e.g., “Country of birth”), the availability (yes=1, no=0) and quality (e.g., 20 age groups) of age categories, and the availability of gender information (yes=1, no=0). We considered the availability of five age categories, for which one age category spans on average 20 years, the bare minimum required for this data to be of any use. If, for example, only three age groups were reported (e.g., age categories 0-25, 26-64, 65+) we considered the available information as insufficient and treated this case as if no age information was available.

Based on the country and region code from the UN Statistical Division (UNSD 2013), we identified unique combinations of origin and destination for the migration streams. We only used streams for which we could unambiguously assign a UN country or region code to both origin and destination.[1] Table 1 reports an overview of the data condition.

[1] For example, files that record the origin as “China (Including Taiwan Province)” or “Chinese of Korean descent” cannot be assigned the UN code for China because of the inclusion of other populations (e.g., Taiwanese in the case “China (Including Taiwan Province)”) or the omission of important sub-populations (e.g., mainland Chinese in the case or “Chinese of Korean descent”), which would lead to biased numbers in the flow computation. We do not use such files in our analysis.

Table 1: Summary statistics of relevant information across all 46,431 raw data files of the UN Global Migration Data Base (UNGMD)

Variable	N	mean	Std.Dev.	min	max
Year of enumeration	46431	1995.39	8.7	1975	2009
Age information available	46431	0.52	0.5	0	1
Number of age categories	46431	10.94	12.8	1	102
Gender information available	46431	0.85	0.36	0	1
Criterion of enumeration	46431	0.54	0.5	0	1

Note: For criterion of enumeration, we coded "country of birth" as 1 and "country of citizenship" as 0.

Table 1 shows that the migration records ranged from 1975 to 2009. Gender differentiated migrant counts were available for 85% of the files, while disaggregation by age was less common with 52% of the files reporting five or more age categories. The reported number of age groups varied widely (from 1 to 102 age groups), with an average of 10 age groups. For the criteria of enumeration, slightly more than half of all files (54%) used "country of birth," instead of "country of citizenship" to classify individuals as migrants.

3.2 Selection of the best file for two years for each migration stream

We used a decision-tree structure and adopted five criteria to select the files with the most detailed information at two time points (Year 1 and Year 2) for each migration stream. For Year 1, the five criteria include: (1) the temporal distance from the year 2000, (2) the information content, (3) the criterion used to define migration status, (4) the quality of the age grouping, and (5) a random number to select among cases of equal quality.

To identify a file closest to the year 2000, we generated a difference measure by subtracting the year of enumeration from the year 2000, allowing us to use the absolute value as the selection criterion. We constructed a variable to capture the availability of age and gender information, which allowed us to prioritize higher information content. A dummy variable enabled us to prioritize "country of birth" (coded 0) over "country of citizenship" (coded 1). We also counted the number of age categories contained in each file to select the case with the largest number of age categories. We then used a nested sorting algorithm to choose the "best" file according to this set of selection criteria. The priority sorting mechanism is visually displayed in Figure 2.

Figure 2: Schematic depiction of employed priority sorting algorithm to select the "best" case among streams with multiple files

File	Stream ID	Year	Quality data	Criteria	Quality age prof	Random
1	840-51	2001	Gender & Age	Cntry of birth	1 year groups	1
2	840-51	2001	Gender & Age	Cntry of birth	1 year groups	5
3	840-51	2001	Gender & Age	Cntry of birth	5 year groups	
4	840-51	2001	Gender & Age	Cntry of birth	10 year groups	
5	840-51	2001	Gender & Age	Cntry of citizenship		
6	840-51	2001	Gender only			
7	840-51	2001	Age only			
8	840-51	2001	Total only			
9	840-51	1998				
10	840-51	1998				
11	840-51	1975				

Note: Hypothetical example for 12 files for the migration stream Armenia to the U.S. File 1 on top of the list was selected for total migrant stock information.

Figure 2 shows that for each criterion, the best cases were placed on top. When the algorithm was unable to produce a unique result, it chose a file based on a randomly generated number from among the cases of equally high quality.

To choose the file for Year 2, we used the same approach, except that the file corresponded to the year closest to, but different from Year 1.

3.3 Obtaining age and gender profiles

Due to the large variation in information content in the raw data, we developed a protocol to derive age and gender profiles with a standardized data format for all migration streams. We wrote a set of 21 functions for the processing of the data and compiled these in an R source package. The scripts, functions, and package are available upon request from the corresponding author.

We first assigned each file one of four treatment categories according to the information content: *category one* files contain both age and gender information; *category two* files contain gender but no age information; *category three* files contain age but no gender information; and *category four* files contain only total number of migrants and lack age or gender information.

For migration streams with data files in treatment category two, three, and four (incomplete age and gender information), additional information may be available for years other than those selected (see section 3.2). Table 2 shows the counts of migration streams for which we derived supplementary information from another file, listed as sub-category "a", "b", and "c."

For example, we obtained age information from different years for 456 (of 1,803) migration streams that have only gender, but no age information in the selected files. For all "b" and "c" categories in Table 2, we were unable to obtain complementary "real" data from within the stream; therefore, we derived age and gender profiles from regional level data (as described below).

Table 2: Treatment categories of data files selected for generating age and gender profiles for Year 1

Main category	Sub category	Description	Main (N)	Sub (N)
1		Total, gender, age	4288	
2		Total, gender, no age	1803	
2	a	Age in same stream but different year		456
2	b	No age in any year		1347
3		Total, age, no gender	73	
3	a	Gender in same stream but different year		42
3	b	No gender in any year		31
4		Total, no age, no gender	595	
4	a	Both gender and age in different years		96
4	b	Either gender or age in different years		127
4	c	Neither gender nor age in any year		372

Note: Total number of migration streams 6,759. For data files of sub-category "a", we obtained true age or gender information from another file within the same stream. For data files of sub-categories "b" and "c," we employed region-level gender and/or age profiles.

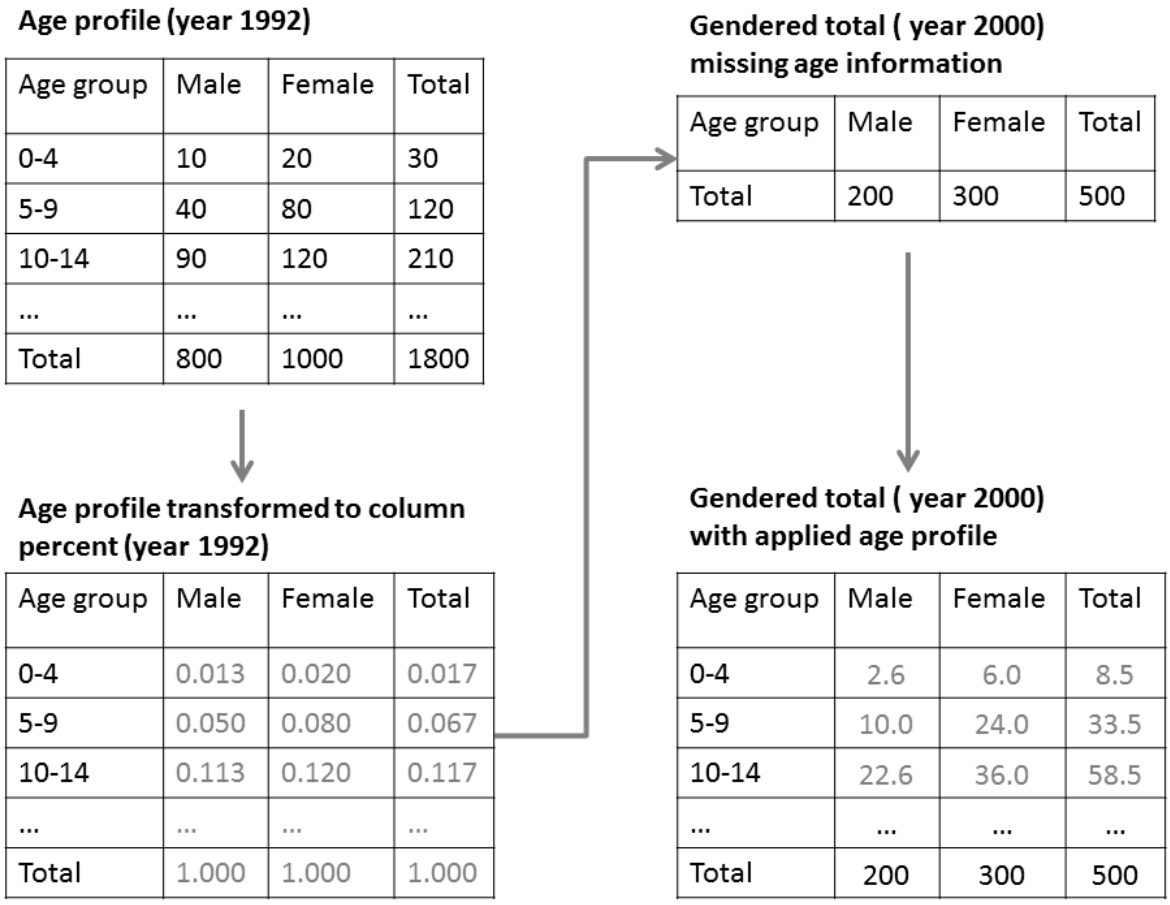
Category one files, composing the largest of the four treatment categories, have complete age and gender information. These files only require the standardization of the used age groups. We chose to standardize the data into the following sixteen age categories: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75+. As the first step in unifying the age categories of different length, we spread the values of any age group across one-year age categories. For example, if an age category 0-9 contained a value of 200 migrants, we generated 10 one-year age categories and allocated each category a value of 20. To spread the migrants for the open-ended age group, we set an upper limit of the on-year age groups to 110. As such, the value in an 85+ category, as an example, was spread across 25 one-year categories. However, in the case where the open-ended category was smaller than 75+ (e.g., 65+), we set 75 as the upper limit for the spreading to avoid biased group sizes upon reaggregation. After the spreading in one-year age categories was completed, we collapsed the data to the desired 16 five-year age groups. In addition to the standardization, a number of cleaning steps were necessary to maximize the use of the available data.

Some files (predominantly those derived from the IPUMS source) contain information for various combined age categories (e.g., age category 0-17). We use this information to derive values for age categories (e.g., age category 5-9) that are within the range of the summary category but contain a missing value in the data file.

For files in which the migrant counts between one and nine are masked by an asterisk for confidentiality reasons (e.g., male="*", female=11, total="16"), we calculated the missing migrant count manually (e.g., male=16-11=5). For cases where we were unable to compute the correct value to substitute the asterisk, we assigned a default value of 4.5 (median of possible values 1 to 9). If only total information was available we distributed this value evenly across the male and female categories.

Migration streams in the "a" categories (2a, 3a, 4a) are lacking age or gender information, but we could obtain complementary information from data files for a different year. Figure 3 visualizes the process flow for a hypothetical category 2a case, for which only gender information was available in the selected file close to 2000, but for which an age profile could be obtained from a file from an earlier year (e.g., 1992).

Figure 3: Schematic representation of working steps to derive age profile for a hypothetical category 2a case that contains total number of migrants by gender but not by age

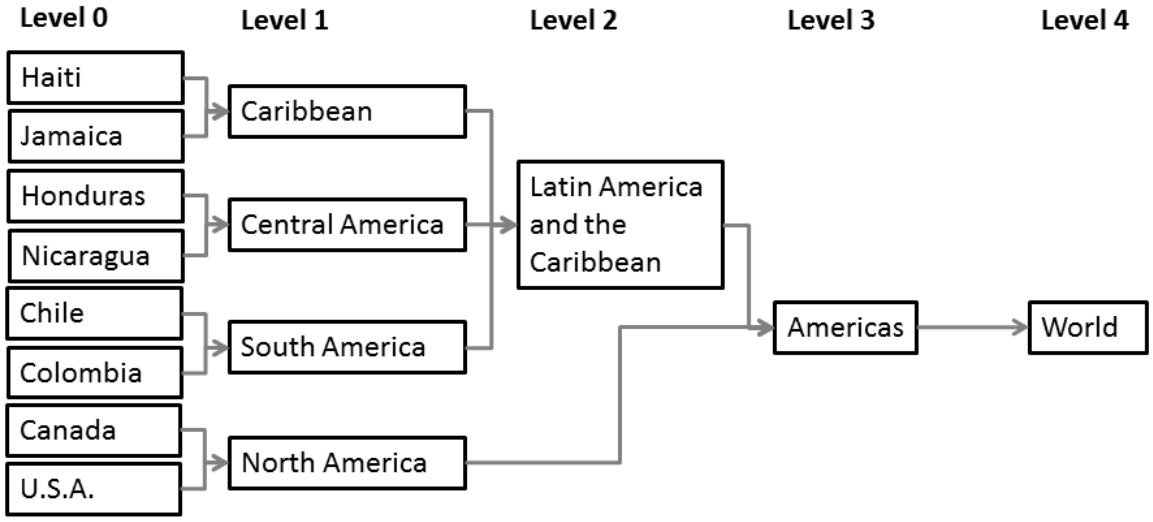


For treatment category 4a, only total information was available. In this case we derived both age and gender profiles from other years.

The data files in the "b" and "c" categories (2b, 3b, 4b, 4c) have neither age and gender information nor could we find complementary information for any other year. To maximize the use of our data we compute average age profiles of the geographic region at the next higher level (e.g., regional, continental) and apply these "artificial" age profiles to the streams lacking this information. The upper-level streams were computed for countries of origin. For example, if an age profile for the migration flow from Honduras to the U.S. was missing, we draw on information from the next higher level (e.g., region: Central America) and computed the average age profile for all migrant streams from Central American countries (e.g., El Salvador, Nicaragua, Belize, Guatemala, etc.) into the U.S. and applied this profile to the migrant stream from Honduras to the U.S.

The first step in generating the upper-level profiles was to specify a hierarchical structure of countries nested within regions. For this purpose we employed the nesting structure suggested by the UN (UNSD, 2013). In general, this hierarchy has four levels with countries (e.g., Germany, level 0) nested within regions (e.g., Western Europe, level 1), which are in turn nested within continents (e.g., Europe, level 2), which are part of the globe (e.g., world, level 3). However, the UN adds an additional level for the American continent. It combines the regions Caribbean, Central America, and South America to a super-region called "Latin America and the Caribbean." This super-region, together with North America, then forms the continent "Americas." The nesting structure is illustrated in Figure 4.

Figure 4: Hierarchy of countries, regions, continents for the example of the Americas as used by the UN (UNSD, 2013) and employed in the generation of upper-level profiles



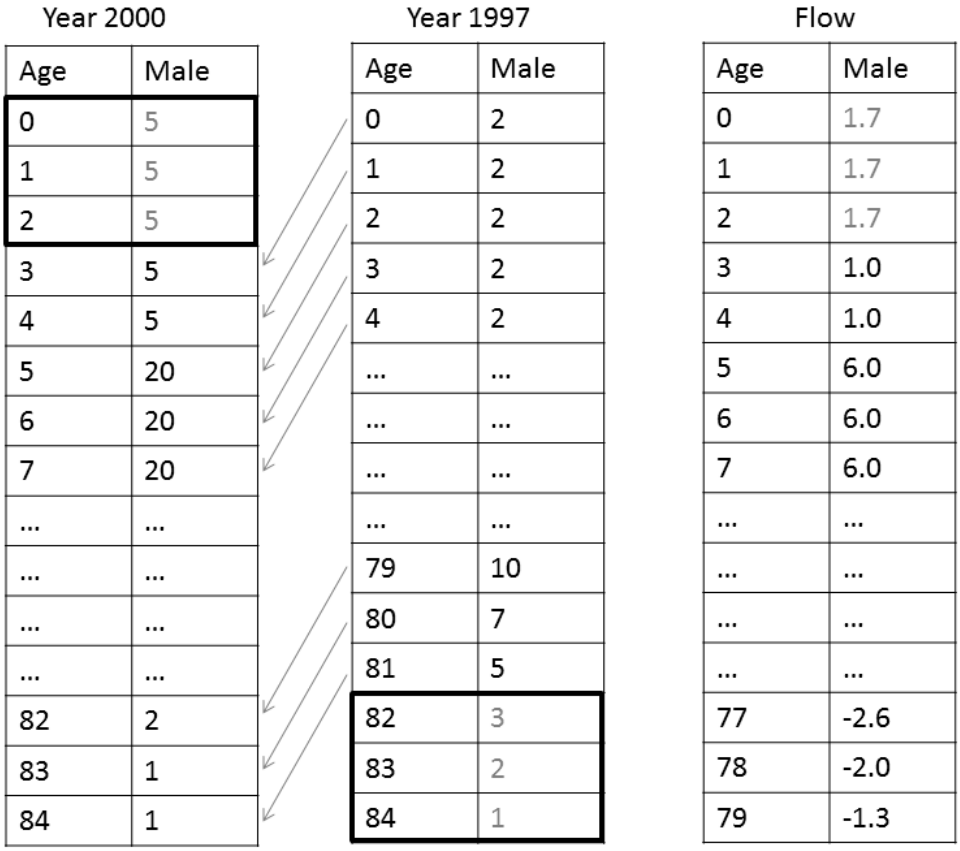
Making use of the UN hierarchy, we generated a file in which each unique migrant stream was assigned the level IDs for all possible upper-levels. This file contained only data of streams for which "real" gender and age profiles were available and had been cleaned up in prior steps. We then computed upper-level profiles by summing all country-level migrant streams within a particular region. This procedure was repeated for all levels resulting in four separate files of upper-level migrant profiles (e.g., regional, super-regional, continental, and global). For the aggregation, we only used files that had complete data (no missing data on any age group). The use of the available information was maximized by aggregating raw data operating at various levels. For example, to generate the upper-level profiles for "Central America to the U.S.A.", we summed the available profiles at the country level (e.g., "Mexico to U.S.A.", "Nicaragua to U.S.A.", "Honduras to U.S.A.", etc.) and also the profiles that already operated at higher aggregation levels (e.g., "Central America to the U.S.A.").

In a final step, we applied the region-level age and gender profiles to country-level streams lacking this information. To best represent the actual migrant characteristics we applied the profile of the lowest level, for which a full profile could be obtained. For example, if an age profile for the migrant stream from “Honduras to the U.S.A.” was missing, we used the upper-level age profile generated for “Central America to the U.S.A.” (contributing streams: n=8) instead of higher level profiles such as “Latin America and the Caribbean to the U.S.A.” (n=38), “Americas to the U.S.A.” (n=40), or “World to the U.S.A.” (n=139).

3.4 Estimation of mortality and fertility adjusted migrant flows

We computed migrant flows by subtracting the count of migrants in the earlier year (e.g., Year 1) from the count of migrants in the later year (e.g., Year 2) and divided the quantity by the period length to derive the average annual changes. For the computation of the net migrant flows, we uniformly disaggregated the migrant counts from 5-year to 1-year age groups. For the last open-ended age group (age 75+), however, we spread the migrant count across 10 years (age 75 to 84) and employed a cumulative exponential function to more closely resemble a declining migrant population at older ages. Distributing the migrant counts in one-year age groups allowed us to compute the migrant flows as conceptually displayed in Figure 5.

Figure 5: Conceptual display of computing migrant flows between two years (e.g., 1997 and 2000)



The computation of the net migration (N^x), as shown in Figure 5, can be formally described using Equation 1.

(1) $N^x = (M_t^x - M_{t-n}^{x-n})/\text{n}$

In Eq. 1, M_t^x represents the migrant stock of a particular age group (x) for the later year t (e.g., 2000) while M_{t-n}^{x-n} represents the migrant stock at the earlier year *t-n* (e.g., 1997), with *n* as the time span between the earlier and later year (e.g., 3 years). For example, the net migrant flow of male individuals in the age 7 category can be computed as the difference between the stock of migrants in the age 7 category in the later year and the stock of migrants in the age 4 category in the earlier year divided by the period as follows: 6 = (20-2)/3. For migrants of youngest ages (e.g., age 0 – 2), the net migrant flow represents the migrant count of the later year adjusted for the length of the period given that none of them were born in the earlier year. Individuals in the

earlier year (e.g., 1997) who would be 85 years or older in the later year (e.g., 2000), highlighted using a black box, are dropped from the calculation based on the assumption that the survival rate approaches zero for these ages.

However, the differences in the migrant stocks between the two time points can be, in part, attributed to deaths and births of the migrants. To obtain an accurate estimation of the net migration flow, it is necessary to exclude the impacts of mortality and fertility of the migrants, as formally described in Equation 2.

$$(2) \quad M_t^x = M_{t-n}^{x-n} + (I^{x-n} - O^{x-n}) - \left(M_{t-n}^{x-n} - \frac{O^{x-n}}{2}\right) * (1 - S^x) - (I^{x-n} - O^{x-n}) * (1 - S^x)$$

In Equation 2, the parameter S^x reflects the conditional survival probability of population from age x-n to x, while I^{x-n} and O^{x-n} are the immigrants and emigrants, respectively, of age x-n to x during the period t-n to t.[2] Therefore, $\left(M_{t-n}^{x-n} - \frac{O^{x-n}}{2}\right) * (1 - S^x)$ constitutes the number of deaths of migrants who reside in the country of destination at t-n, and $(I^{x-n} - O^{x-n}) * (1 - S^x)$ is the number of deaths of net immigrants who moved to the destination during the period of t-n to t. This equation can be applied to derive age-specific net migration flows for both male and female populations. We assume the age-specific mortality rates are the same for immigrants and emigrants.

As we do not know the number of emigrants from the data files, we use $(M_{t-n}^{x-n}) * (1 - S^x)$ to replace $\left(M_{t-n}^{x-n} - \frac{O^{x-n}}{2}\right) * (1 - S^x)$ to compute the number of deaths of non-moving migrants, and assume that the difference is relatively small. Denoting the number of net migrants of age x-n to x during the period t-n to t $as\ N^x$ ($N^x = I^{x-n} - O^{x-n}$), Equation 2 can be simplified to Equation 3:

$$(3) \quad M_t^x = M_{t-n}^{x-n} + N^x - M_{t-n}^{x-n} * (1 - S^x) - N^x * (1 - S^x)$$

$$M_t^x = M_{t-n}^{x-n} * S^x + N^x * S^x$$

$$N^x = M_t^x / S^x - M_{t-n}^{x-n}$$

[2] The conditional survival rate was computed by dividing the survival rate at age x (e.g., age 20) by the survival rate at age x-n (e.g., age 17), where n (e.g., 3) is the time difference between the two years (e.g., 2000 and 1997) for which the migrant stock data was derived.

When the UN data defined migrants by "country of birth," we calculated the number of net migrants of age 0 to n using Equation 4, where N^{-n} denotes the net migrants of age 0 to n, and M_t^{-n} represents the number of migrants of age 0 to n at time t.

(4) $$N^{-n} = M_t^{-n} + M_t^{-n} * (1 - S^n)/2$$

Alternatively, when the UN data defined migrants by "country of citizenship," government officials usually count the children who are born to female migrants in the country of destination as migrants, although they never moved. Therefore, we have to exclude the number of births given by the female migrants from the number of migrants of age 0 to n. This step is formally described as shown in the following equation.

(5) $$N^{-n} = (M_t^{-n} + M_t^{-n} * \frac{1-S^n}{2}) - \sum_{i=15}^{49}(M_{t-n,f}^{i} * b^i + M_{t-n,f}^{i} * S^i * b^{i+n}) * n/2$$

In equation 5, the parameter b^i represents the fertility of women age i. Assuming that the sex ratio of births is equal to one, we allocate the total new births evenly between the male and female groups. The UN Population Division provided the age, gender, and country-specific survival rates for the year 2000 (UNPD 2009b). We used a linear interpolation to obtain one-year survival rates from the five-year intervals. We employed the age and gender specific survival rates for the country of destination, assuming that migrant's mortality rates adjust quickly to the average of the host country. The age-specific fertility of the year 2000 came also from the UN Population Division (UNPD 2009a). The steps to compute the number of new born migrant children are conceptually depicted in Figure 6.

Figure 6: Computation of new born migrant children when "country of citizenship" was used as criterion of enumeration

Year 1997

Age group	No. female migrants
15-19	200
20-24	300
25-29	400
30-34	350
35-39	300
40-44	250
45-49	200

Fertility rates

Age group	Fertility rates
15-19	0.0633
20-24	0.1685
25-29	0.1573
30-34	0.0976
35-39	0.0518
40-44	0.0195
45-49	0.0053

Births per year

Age group	No. of births
15-19	12.66
20-24	50.55
25-29	62.92
30-34	34.16
35-39	15.54
40-44	4.88
45-49	1.06
Total	181.77

New born children across period 1997 to 2000

Birth year	Age in 2000	Female	Male
1999 - 2000	0	90.88	90.88
1998 - 1999	1	90.88	90.88
1997 - 1998	2	90.88	90.88

For our computation, we assume a constant birth rate for female migrants across the period between Year 1 and Year 2. During each year females of reproductive age give birth to a certain number of children. These children are then removed from the count of migrants in the respective age group estimated for the later year. In a final step, the migrant flow counts were re-aggregated to five-year age categories.

3.5 Region level aggregation

To derive the age-gender profiles of international migration flows for the multiregional population/urbanization projection for NCAR CDM global 31 regions, we aggregate the country-level profiles to the region-level profiles. The definition of the 31 CDM regions is presented in Appendix Table 1.

The aggregation was performed by computing the sum of migrants across country-level streams for all unique region of origin (e.g., Eastern Africa) and region of destination (Western

Europe) combinations. Only country-level streams with complete data (no missing data in any age/gender group) were used in the aggregation step.

4 Results and Discussion

4.1 Results

Using the described methodology, we were able to generate a data set of 3,713 country-level net migration flows by age and gender. Table 3 summarizes the distribution of these age and gender profiles of migration flows at the national level across 31 global regions. It also shows the proportion of available migration streams relative to the total possible number, which varied greatly between regions. For example, for regions such as Australia and the U.S.A., the dataset provides comprehensive coverage, while no data exists for the Rest of Eastern Asia, India, and Indonesia.

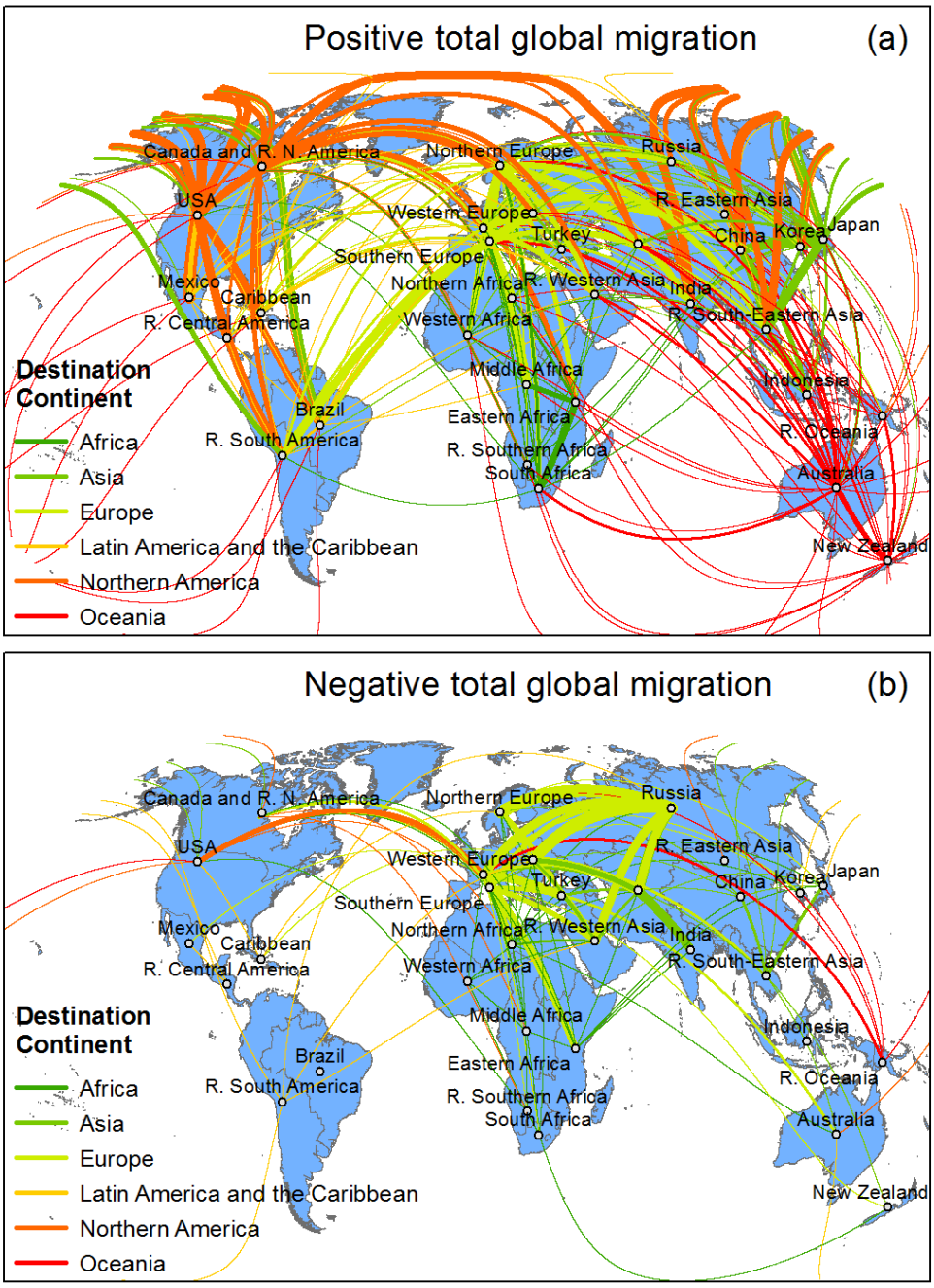
Table 3: Distribution of available age and gender profiles of migration flow data across 31 global regions

Region	Cntries	Pos. Str.	Obs. Str.	% Cov.	% High Qual.
Eastern Africa	20	4840	69	1.4	34.8
Middle Africa	9	2178	20	0.9	15
Northern Africa	7	1694	64	3.8	48.4
R. Southern Africa	4	968	21	2.2	23.8
South Africa	1	242	72	29.8	98.6
Western Africa	17	4114	71	1.7	28.2
Brazil	1	242	58	24	96.6
Canada and R. N. America	4	968	148	15.3	10.8
Mexico	1	242	54	22.3	100
R. Central America	7	1694	139	8.2	61.9
R. South America	13	3146	318	10.1	79.2
USA	1	242	190	78.5	61.1
China	4	968	15	1.5	93.3
India	1	242	-	-	-
Indonesia	1	242	-	-	-
Japan	1	242	48	19.8	0
Korea	1	242	19	7.9	100
R. Eastern Asia	2	484	-	-	-
R. South-Central Asia	13	3146	20	0.6	25
R. South-Eastern Asia	10	2420	39	1.6	25.6
R. Western Asia	17	4114	50	1.2	6
Turkey	1	242	19	7.9	0
Caribbean	29	7018	162	2.3	34.6
Northern Europe	18	4356	684	15.7	70.9
R. Eastern Europe	9	2178	76	3.5	21.1
Russia	1	242	18	7.4	0
Southern Europe	16	3872	434	11.2	26.5
Western Europe	9	2178	546	25.1	44
Australia	1	242	163	67.4	0.6
New Zealand	1	242	118	48.8	17.8
R. Oceania	23	5566	78	1.4	25.6

Note: Cntries = The number of countries located in a particular global region; Pos. Str. = The number of possible streams flowing into the particular region (Cntries * 242); Obs. Str. = The number of observed streams for which flow data by age and gender is available in the CDM-IM data set; % Cov. = The coverage of possible streams in the CDM-IM data set; % High Qual. = The percentage of age and gender profiles computed based on high quality raw data (category 1 files).

Figure 7 provides a visual depiction of total net international migration flows between global regions. It shows North America (orange lines) and Europe (light green lines) as the major receivers with large flows of positive net migration (Panel a). Other important migration destinations include: Australia and New Zealand (red lines), South Africa on the Africa continent (dark green lines), and Japan for East Asia (green lines).

Figure 7: Global representation of positive and negative net migration flows between 31 regions

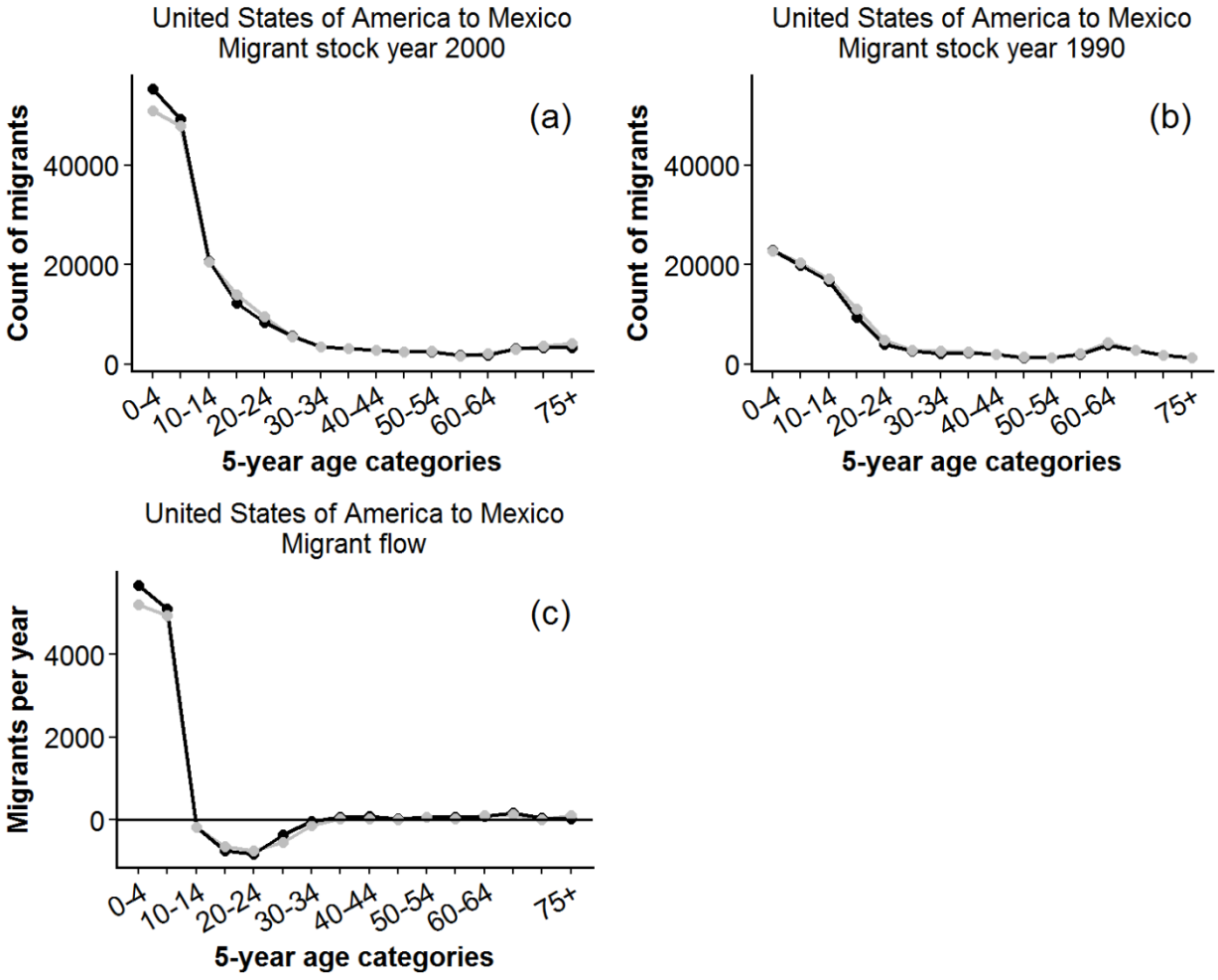


Note: Lines represent net migration flows between 31 global regions. Panel a shows positive net migration flows, while panel b depicts net negative migration flows. Line color indicates the continent of destination. For example, the orange line connecting the U.S.A. and Northern Europe in Panel a shows a positive net migration flow from Northern Europe (origin) to the U.S.A. (destination). In contrast, a green line connecting Russia with R. Western Asia in Panel b shows a negative net migration flow from R. Western Asia (origin) to Russia (destination), indicating a return of Russian migrants to their home country. Line thickness represents the size of the migration flow.

In contrast, Panel (b) shows Russia (light green lines) as the destination of the largest negative net migration flows. The return of Russian migrants from Eastern European countries to Russia during the post 1990 era after the break down of the former Soviet Union likely produced these negative flows (Pilkington 1998).

Numerous influences, including historical background, political situation, and socioeconomic conditions determine the characteristics of the observed migrant flows (Brown and Bean 2006). Three cases illustrate how the region-specific context uniquely shapes the age and gender profile of a migrant flow. First, our results suggest that the majority of U.S. immigrants residing in Mexico are children less than 19 years (Figure 8a and 8b).

Figure 8: Migrant stock data and resulting flows for the migration stream from the U.S.A. to Mexico



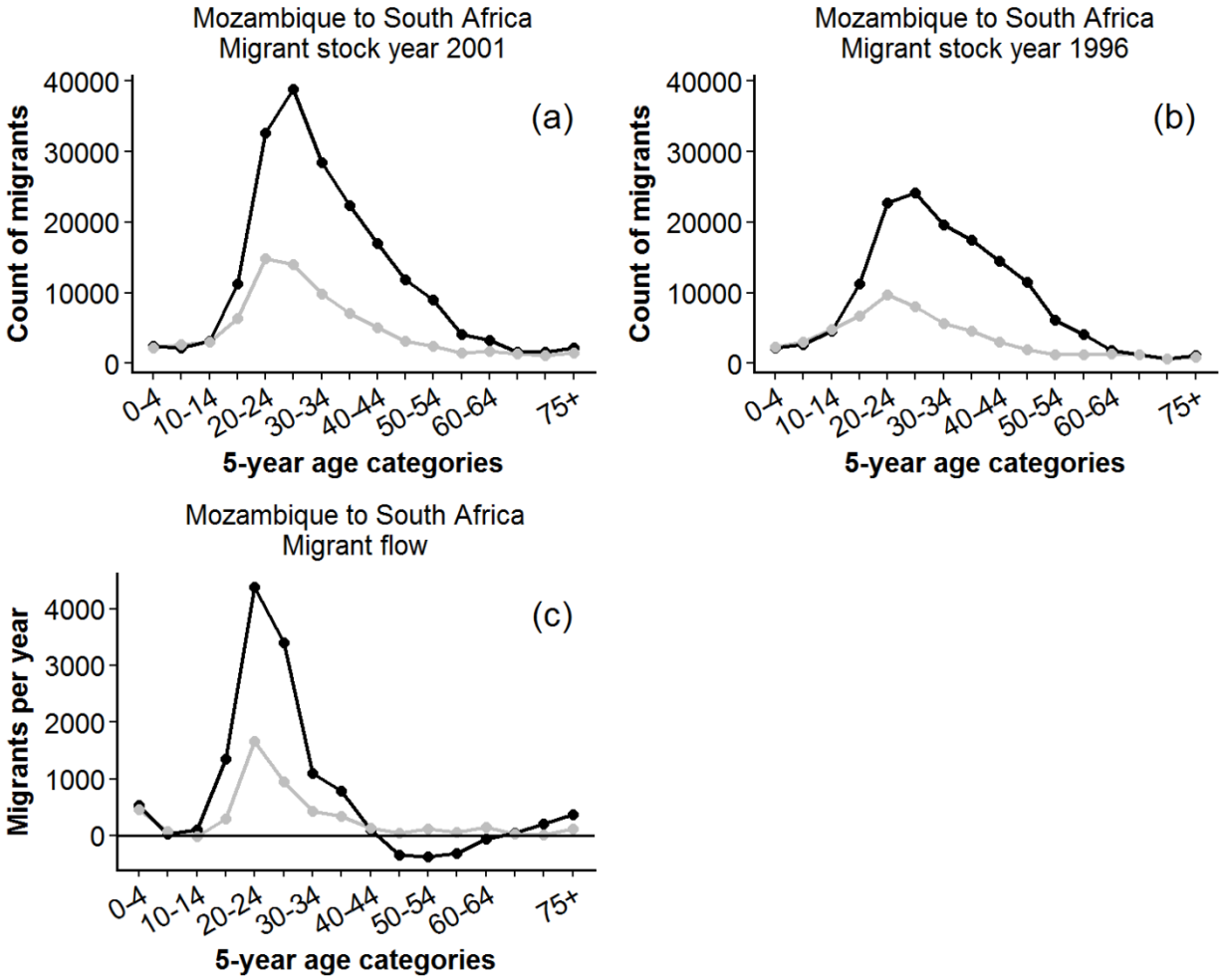
Note: Gray line represents female migrants; black line represents male migrants. Criterion of enumeration: Country of birth. Data source: Census.

These young U.S. citizens are likely the children of Mexican migrant workers born in the U.S. Most Mexican migrants are target earners who temporarily stay in the U.S. (Lindstrom 1996). Once Mexican migrants achieve a saving target, they usually return to their origin, accompanied by their U.S.-born children. Under conditions of an extended stay, the Mexican migrant might choose to send the U.S.-born child back to live with other family members in what they often perceive as a culturally more beneficial environment. In addition, the number of migrant women and movements of entire families has increased, making the event of a child birth by Mexican migrants in the U.S. more likely (Marcelli and Cornelius 2001).

The *migrant stocks* of year 2000 and 1990 clearly show the demographic features of U.S. immigrants residing in Mexico. However, an interesting phenomenon becomes visible in the age and gender specific profile of the net *migrant flow* (Figure 8c), which is not apparent from the migrant stock data: The positive net migration flow from the U.S. to Mexico only exists for young children below age 10. Between age 10 to 30, a significantly negative net-flow is apparent, meaning that U.S. immigrants living in Mexico move back to the U.S. for better education and employment opportunities (c.f., Kemper 2005). This finding echoes other studies, demonstrating an increase in the number of children from Spanish speaking families in American schools (Ruiz-de-Velasco, Fix, and Clewell 2000).

Second, analyzing the migration flow from Mozambique to South Africa reveals a different story (Figure 9). South Africa has a long history of dependence on migrant workers from Mozambique, particularly in sectors such as mining, agriculture, and construction (Crush 1999; McDonald et al. 2000). Because of the demand for manual labor, Mozambican migrants employed in those industries were traditionally male (Crush and McDonald 2001). The gender specific profiles of the migrant stocks in 2001 and 1996 demonstrate this male preference (Figure 9a and 9b). Moreover, the migrant stock data also reveal an increase in the total number of migrants between 1996 and 2001. A number of factors indicate reasons for the observed increase: (1) South Africa experienced a high influx of refugees from neighboring Mozambique during the 1990s as a result of the nation's civil war of 1983-1992 (Hargreaves et al. 2004). The South African government granted group refugee status for Mozambicans in 1993, and permanent residency became effectively available in 1999/2000 (Polzer 2007). (2) Political and policy changes in South Africa after the fall of the Apartheid system in 1994, including amnesty programs and new immigration acts, caused a surge in the admission of skilled foreign workers (Crush and McDonald 2001, Polzer 2007). (3) Concurrently the number of noncontract migrants increased as well (McDonald et al. 2000).

Figure 9: Migrant stock data and resulting flows for the migration stream from the Mozambique to South Africa

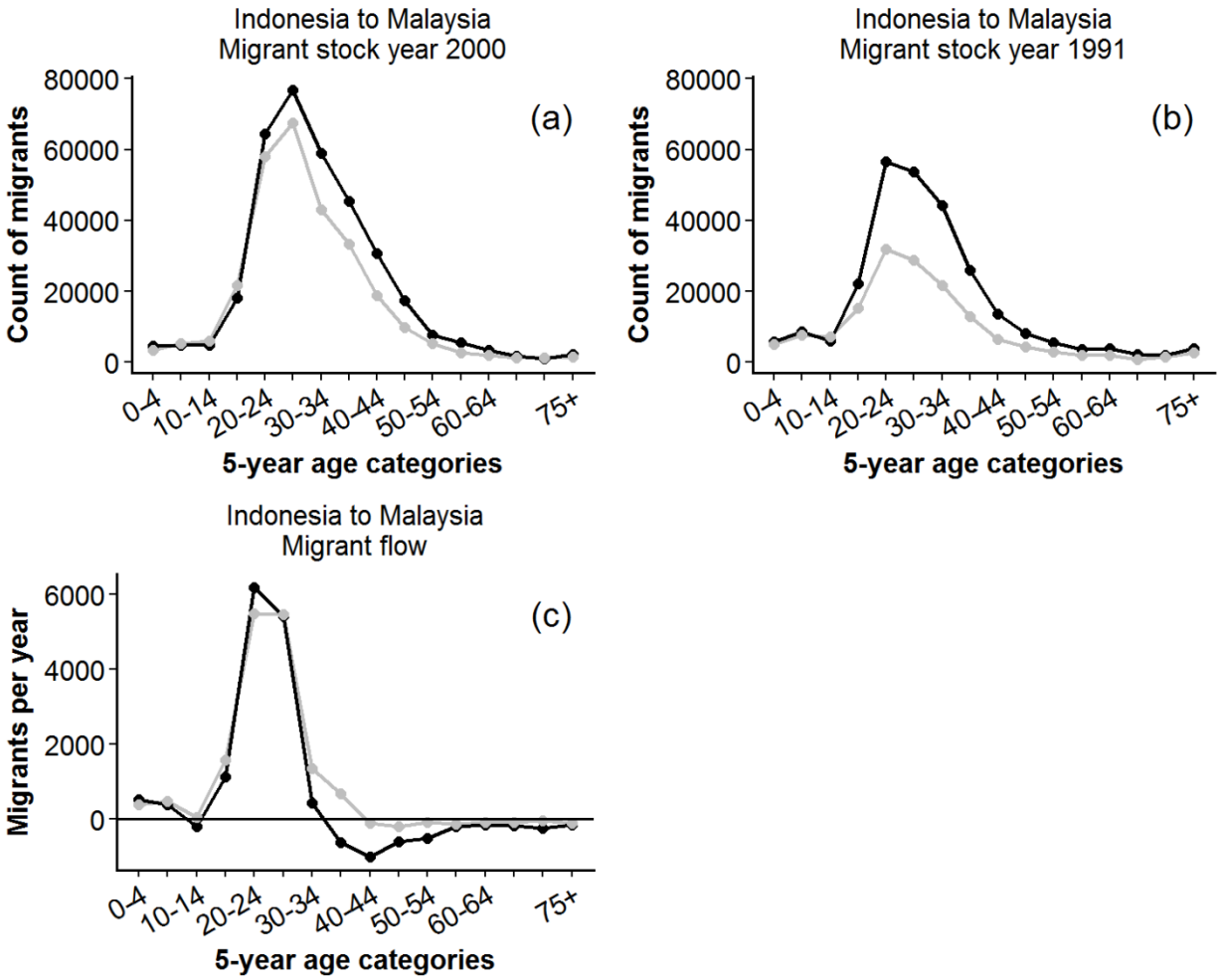


Note: Gray line represents female migrants; black line represents male migrants. Criterion of enumeration: Country of birth. Data source: Census.

Over the period of 2001 to 1996, the number of migrants age 20-30, particularly males, increased substantially, because of a positive net immigration among young migrants, as shown in Figure 9c. In contrast, an increase in return migration (both forced and voluntary), after Mozambique recovered from the civil war, likely produced the negative net migration among males age 45-64 (Polzer 2007). Surveys suggest that very few Mozambican migrants intend to settle permanently in South Africa, stressing the circular nature of the migration stream (Crush 1999; McDonald et al. 2000). Moreover, surveys of migration intentions suggest that the observed migration patterns may remain relatively stable in the future (McDonald et al. 2000), and thus the migrant flow profiles may provide a solid foundation for projecting future population changes.

Figure 10 illustrates the characteristics of the migration flow from Indonesia to Malaysia as the third and final example. The gender profiles of the migrant stock for the years 1991 and 2000 changed significantly (Figure 10a and 10b): while more males than females migrated in 1991, the gender differences in the migrant stock became less pronounced in 2000. Indonesia's migration history helps to explain this change: Labor migration has long been a remedy to the problem of unemployment and a source of foreign exchange through remittances. Especially, during the Asian financial crisis of 1997/98, the Indonesian government encouraged labor emigration (Tsai and Tsay 2004). Official statistics show that the total number of international labor migrants more than doubled during 1991-2001 (Tsai and Tsay 2004). However, the official numbers might still be underreported given that most workers choose to migrate illegally to avoid the exit tax and bureaucratic delays (Hugo 1995).

Figure 10: Migrant stock data and resulting flows for the migration stream from Indonesia to Malaysia



Note: Gray line represents female migrants; black line represents male migrants. Criterion of enumeration: Country of birth. Data source: Census.

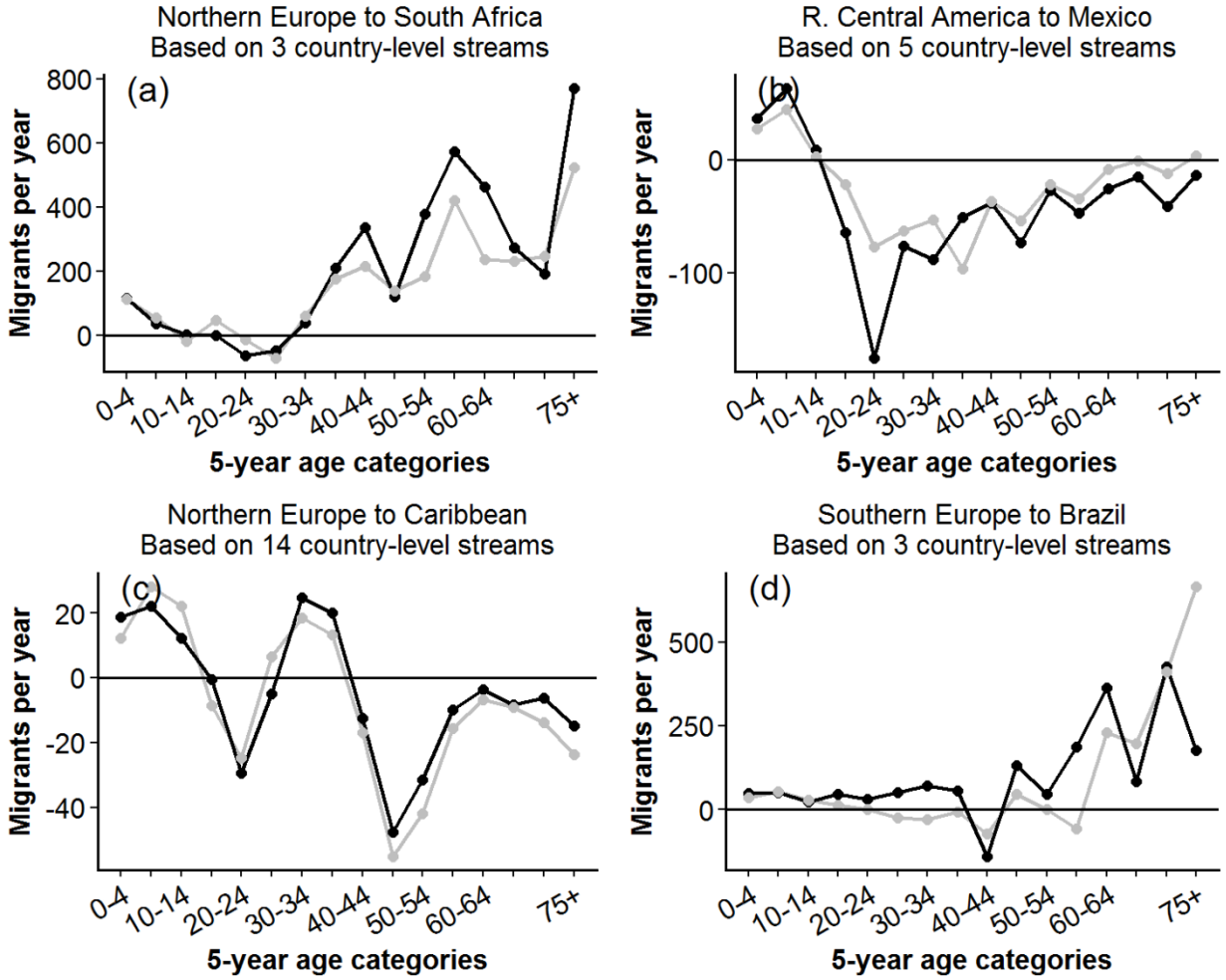
Labor migrants from Indonesia typically move to more industrialized countries such as Saudi Arabia, Singapore, and Malaysia (Tsai and Tsay 2004). Neighboring Malaysia poses a particularly interesting destination due to geographical proximity, ethnic, religious and linguistic similarities, a prosperous growing economy, comparatively high wages, and a history of reliance on migrant workers to ease local labor shortages (Tsai and Tsay 2004, Hugo 1995, 2002).

In the early 1990s, the migrant flow from Indonesia to Malaysia largely included males due to the nature of employment opportunities (low-skilled manual jobs) in the plantation, construction, and forestry sectors (Hugo 1995). However, women's migration to Malaysia has increased significantly in more recent years (Hugo 2002). A tendency of joint migration with male family members, emerging employment opportunities (e.g., factory work, prostitution), and the changing roles and social status of women may explain the increased proportion of females in the migration stream (Hugo 2002). The derived profiles of net migration flow, with similar shapes for males and females, reflect this trend (Figure 10c).

In summary, the three discussed cases demonstrate the complex context in which migration flows emerge. Changes in migrant flows over time are difficult to predict, but an increase in mobility can be expected due to various effects, such as growth in education levels and associated reorientation in values and attitudes, increase in the access and speed of global and national transport systems, tendencies towards higher levels of globalization and economic integration, and a strengthening and extension of social networks connecting destinations and origins (Hugo 1995).

Each country-level profile contributes to the gender and age specific migrant profiles at the region-level. However, region-level profiles are composed of varying numbers of country-level streams (average=5.54, range=1 to 60). Figure 11 shows a random selection of 4 region-level profiles.

Figure 11: Age and gender profiles of region-level migration streams



Note: Gray line represents female migrants; black line represents male migrants.

The region-level profiles demonstrate a trend of retirement migration from Northern Europe to South Africa (Panel a) and Southern Europe to Brazil (Panel d), with highest numbers of positive net migration among older age groups. In contrast, Mexico's stock of immigrants from the Rest of Central America declined particularly among working age adult males (Panel b), attributable to return migration. The profile migration from Northern Europe to the Caribbean (Panel c) shows largest inflows for young adults between age 25-39 and their young children age 0-14. However, only a small number of migrants characterize this flow, making the observed patterns less reliable.

Overall, the difference in the trend observed for older age groups highlights the advantage of using real data for the computation of migrant profiles. Modeling approaches frequently force a smooth deceleration of the curve to approach zero at the oldest age groups.

Our research suggests that such an assumption does not appropriately reflect reality in many cases.

4.2 Validation

To assess the quality of our data, we compare our results with two existing international migration datasets: The table of global transition flows generated by Abel (2013) and the migrant profiles computed as part of the IMEM project (Raymer et al. 2012). While both data sets can be used for comparing the *total flows*, only the IMEM project data functions to validate the migrant *age and gender profiles*. Unfortunately, the IMEM project data contains information from European countries only, which limits the number of countries that we can use for the comparison.

4.2.1 Total migration

To compare the total migration flows between the three datasets, we standardized the flow measure to reflect the average net migration *per year*. We used a seven year average (2002-2008) of the IMEM flow data instead of only one year (e.g., 2002) closest to the year 2000 due to the following reasons: (1) Computing a seven year average allows obtaining more robust estimates and reduces the influence of period effects; (2) The CDM-IM data set is computed for varying time periods; (3) Shapes of migrant profiles usually do not change much over time.

We derived net migration flows from Abel's (2013) migrant transition flow table for the period 1990 to 2000 by subtracting outflows from inflows. The *inflow* from country x to country y is represented by the cell value in Abel's (2013) transition flow table with origin (row) x and destination (column) y. The *outflow* is represented by the cell with origin y and destination x. Subsequently, we conducted an ordinary pairwise t-test to compare the size of the average flows (see Table 4).

Table 4: Ordinary pairwise two-sample mean comparison (t-test) of total migration flows, contrasting the CDM-IM, IMEM, and Abel (2013) data sets

	CDM-IM		Abel (2013)		IMEM		
DF	Mean	S.D.	Mean	S.D.	Mean	S.D.	sig.
Panel A: High quality data (category 1)							
238			49	1661	-255	4427	
238	284	2223	49	1661			
238	284	2223			-255	4427	
Panel B: High quality data (category 1) & Large flows (>100)							
55			90	2303	655	2429	
55	743	4474	90	2303			
55	743	4474			655	2429	

Note: Significance (sig.) refers to a t-test with the following p-values: * < 0.05; ** < 0.01; *** < 0.001

For the t-test, we restrict the comparisons to cases available for all three data sets, and we only use high quality cases in the CDM-IM data set (category 1 files, see Section 3.2). In addition, we present comparisons, restricting the used data further to cases composed of more than 100 migrants in order to guard against the potentially biasing influence of unstable estimates for small flows. Although the mean values differed, the t-test analysis revealed no significant differences among the three data sets. The variations in the mean values may result from the differences in the time period of data reporting, differences in the computation methods, and inconsistencies in the reporting of migration flows (c.f., Raymer et al. 2011). For the larger subset of high quality data (Panel A), our estimate of the average total migration flow ($\mu = 284$) more closely corresponds to the number produced by Abel ($\mu = 49$), compared to the IMEM project estimate ($\mu = -255$). In contrast, for the most conservative comparison of high quality data *and* large migrant flows (Panel B), our estimate ($\mu = 743$) more so reflects the IMEM project ($\mu = 655$) compared to the average total flow derived from Abel's data set ($\mu = 90$).

4.2.2 Migrant profiles by age and gender

The primary contribution of the CDM-IM data set is a comprehensive collection of country-specific migration flow profiles by age and gender with approximately global coverage. For the subset of European migration streams, we compare the CDM-IM profiles to those produced by IMEM (Raymer et al. 2012). To facilitate the comparison, we select the age categories 0-9, 20-44, and 50-69, which typically show a distinct shape of the migration profiles. We then compute the proportion of migrants in the gender-specific age categories relative to the

total number of migrants in the entire flow.[3] Because we can compute meaningful percentage values only with positive integers, we performed a reparameterization, expressing each value as the difference from the lowest value in the profile. We then employ the percentage values in a pairwise t-test to investigate the difference in the age and gender specific profiles across the two data sets (see Table 5).

Table 5: Ordinary pairwise two-sample mean comparison (t-test) of migrants' age and gender profiles in the CDM-IM and IMEM data sets

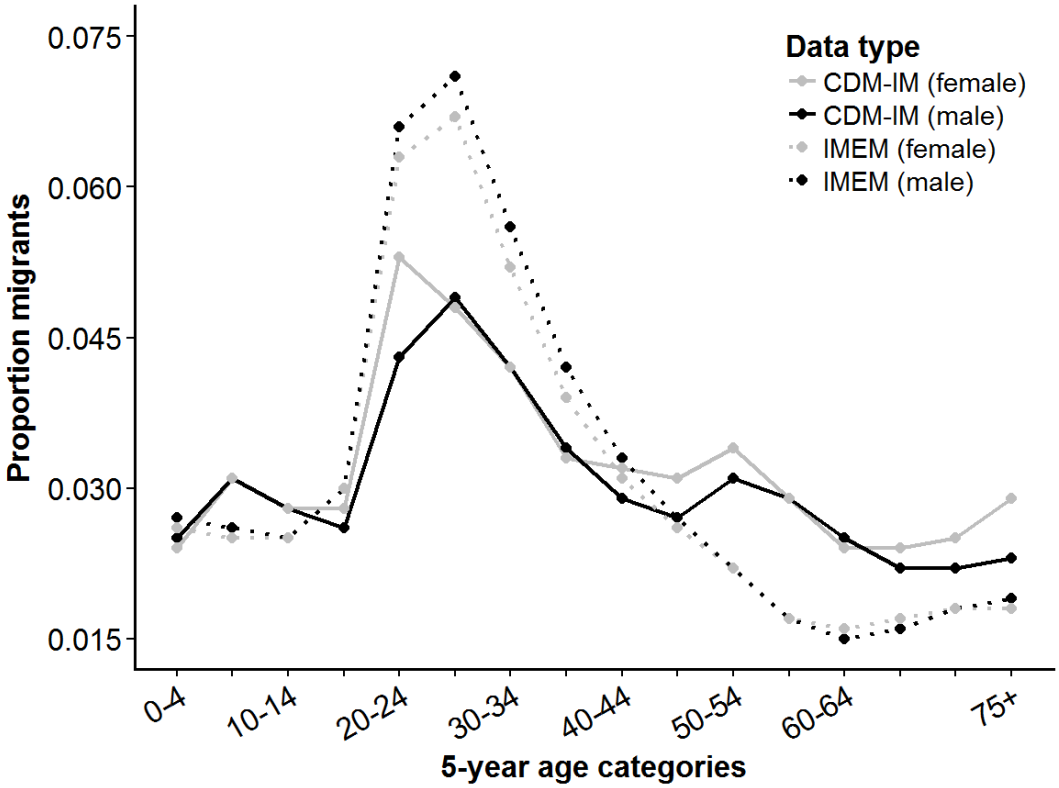
		Male			Female			Total		
Age group	DF	A	B	sig.	A	B	sig.	A	B	sig.
Panel A: High quality data (category 1)										
0-9	210	0.111	0.115		0.099	0.113		0.21	0.228	
20-44	210	0.28	0.306		0.3	0.285		0.58	0.591	
50-69	210	0.102	0.088		0.107	0.093		0.21	0.181	
Panel B: High quality data (category 1) & Large flows (>100)										
0-9	87	0.084	0.104		0.078	0.1		0.161	0.204	
20-44	87	0.282	0.332	*	0.336	0.305		0.618	0.638	
50-69	87	0.104	0.079		0.117	0.08	*	0.22	0.158	*

Note: A = CDM-IM; B = IMEM; Significance (sig.) refers to a t-test with the following p-values: * < 0.05; ** < 0.01; *** < 0.001

Table 6 demonstrates only a few significant differences. For high quality cases (Panel A), the average age and gender profiles are similar, and no statistically significant difference emerged. However, in comparing large, high quality flows (Panel B), differences become evident for the male migrants age 20-44, and for the oldest age group (age 50-69) among female migrants. For middle-aged males, IMEM predicts a larger proportion of migrants while for old females IMEM predicts a smaller proportion of migrants compared to CDM-IM. To facilitate the evaluation of these shape differences, we graphically depict the average proportion of migrants in each 5-year age category (Figure 12).

[3] By relating the count of migrants in each age group (e.g., 20 years old females) to the total migrants of the entire flow (sum of male and female migrants across all age groups) we effectively adjust for age and gender. If the count of female migrants in the 20 years age group would have only be related to the sum of migrants across all age groups in the female profile, the obtained percentages would have been only accounted for age but not for gender difference.

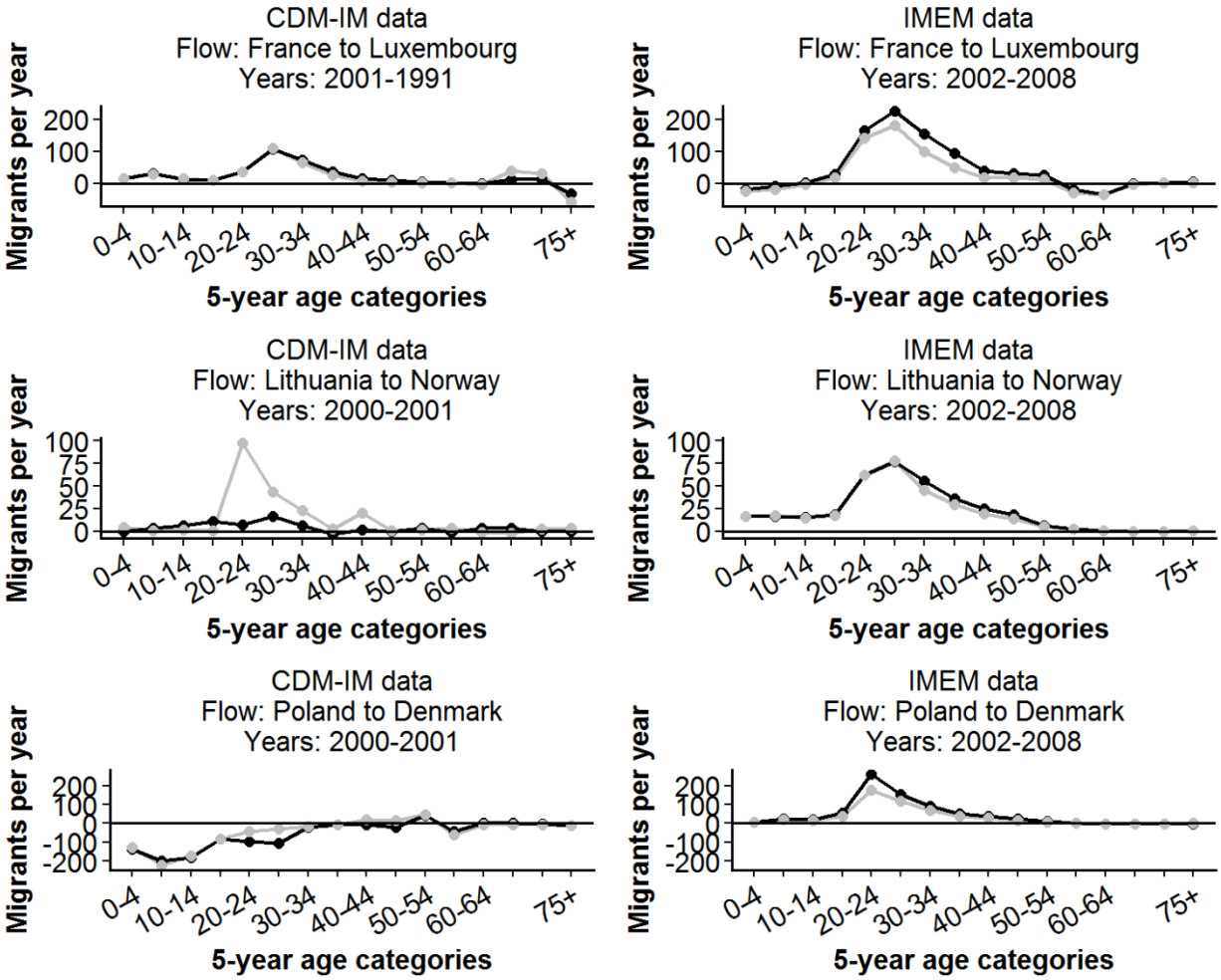
Figure 12: Average proportion of migrants by age and gender comparing the CDM-IM and IMEM data sets



Note: We computed the average proportions for flows with more than 100 migrants derived from high quality data (category 1 files).

Although the shapes of the curves show some resemblance, they are not identical. The IMEM data shows a more pointed peak around age 25-29. In contrast, the CDM-IM data is more spread in the tail for older ages with a smaller peak around age 25-29. Figure 13 shows three randomly selected flows, allowing the evaluation of differences in the shapes for specific migrant streams.

Figure 13: Comparing the age and gender profiles of migrants using three randomly selected migrant streams in the CDM-IM and IMEM datasets



Note: Gray lines represent female migrants; black lines represent male migrants.

In summary, the validation tests demonstrate that our estimates of the total migrant flow and the age and gender profiles of the migrants are similar to the results derived by other researchers using different methodologies. The few differences that we observed most likely result from the different methods used to derive the migrant flow estimates. While we calculate our results directly from the raw data, the IMEM data employs a Bayesian modeling strategy to derive the estimates (Raymer et al. 2012). As such, the CDM-IM data set constitutes a collection of "real" or observed migrant counts, while the IMEM project produced a data set of "synthetic" model output. A modeling approach can produce smoother curves, but relies on various assumptions that may or may not appropriately reflect reality. In their Bayesian model, Raymer et al. (2012) used priors to include subjective information (expert opinions) on: (1) migration undercount, (2) duration of stay, and (3) data accuracy. For example, the used prior for undercount suggests that the observed emigration and immigration data constitute, on average,

56.0% and 79.6% of the "true flows" (Raymer et al. 2012:10). To estimate the "true" migration flow, the model proportionately increases the emigration counts more than the immigration counts, leading to a reduction in the net flows. This explains, in part, why the average net migration in the IMEM data set is smaller compared to the CDM-IM data set.

In addition, regression models are sensitive to the quality and quantity of included covariates. For example, the use of a simple dummy variable to reflect contiguity (Raymer et al., 2012) might not appropriately reflect movement barriers. Instead, a matrix of physical distances between the capital cities might lead to a better model fit and different flow estimates.

Numerous sources of uncertainty, inherent in a complex modeling approach, as used by Raymer et al. (2012), may combine in cumulative ways, potentially producing estimates that differ vastly from the observed flows. Whether the *synthetic data* reflects the reality better than the *observed data* remains unclear; so we believe that the more simplistic approach of using raw data and some basic demographic adjustments, as in the CDM-IM data set, produces a more conservative picture of the true migration flows.

5 Conclusions

In this paper, we have outlined the methods used to generate the Community Demographic Model International Migration (CDM-IM) data set. The CDM-IM constitutes a novel data set of age and gender specific international migrant flows and has two main strengths: (1) It is based on real data, and (2) it contains country-level migration flows by age and gender with approximate global coverage. It provides information not only important for the multiregional population/urbanization projections of the NCAR CDM model, but is also useful for other research endeavors: First, researchers may use the counts of total migration flows and the age and gender specific migrant profiles as the base year input for models projecting national and regional population changes. Second, researchers may use the CDM-IM data set for comparative analysis of migration behaviors among various sub-populations (e.g. women, youth, and the elderly) under different socioeconomic and environmental circumstances (Feng et al. 2010). Finally, researchers may use the CDM-IM data set as the base for fitting migration model schedules (Rogers et al. 2005), or as a tool to benchmark the accuracy of synthetic data sets derived from modeling approaches.

However, this data set is not without limitations. First and foremost, the available data limit the computation of migration flows by age and gender. Although we use the most comprehensive migration data set currently available (the UNGMD), it does not provide the full scope of all possible migration streams. Second, to generate the largest number of migration streams possible, we computed flows of differing data quality. For example, we computed a small number of flows employing data based on different enumeration methods ("country of birth" versus "country of citizenship"). However, we believe that it is better to use these close approximations rather than to have no data available for the particular stream. Appendix Table 2

provides a list of data quality flags used in the CDM-IM data set. Users may select a subset of cases to fit their research needs based on these quality flags.

With these advantages and limitations in mind, we trust that the CDM-IM data set will serve as a beneficial tool for researchers in order to investigate the complex issue of population dynamics and will help to gain a deeper understanding of the causes and drivers of international population relocation in “the age of migration” (Castles and Miller 2003).

6 References

Abel, G. J. (2013). Estimating global migration flow tables using place of birth data.*Demographic Research, 28*, 505-546. doi: 10.4054/DemRes.2013.28.18

Bilsborrow, R. E., Hugo, G., Oberai, A., &Zlotnik, H. (1997). International migration statistics: Guidelines for improving data collection systems. Geneva, Switzerland: International Labour Office.

Brierley, M. J., Forester, J. J., McDonald, J. W., & Smith, P. W. F. (2008). Bayesian estimation of migration flows. In J. Raymer& F. Willekens (Eds.), *International migration in Europe: Data, models and estimates* (pp. 149-174). Chichester, U.K.: Wiley.

Brown, O. (2008). Migration and climate change. Geneva, Switzerland: International Organization for Migration.

Brown, S. K., & Bean, F. D. (2006).International Migration.In D. Posten& M. Micklin (Eds.), *Handbook of population* (pp. 347-382). New York: Springer Publishers.

Castles, S., & Miller, M. J. (2003).*The age of migration: International population movements in the modern world*. New York: The Guilford Press.

Crush, J. (1999). The discourse and dimensions of irregularity in post-apartheid South Africa.*International Migration, 37*(1), 125-151. doi: 10.1111/1468-2435.00068

Crush, J., & McDonald, D. A. (2001). Introduction to special issue: Evaluating South African immigration policy after apartheid. *Africa Today, 48*(3), 1-13.

de Beer, J., Raymer, J., van der Erf, R., & van Wissen, L. (2010). Overcoming the Problems of Inconsistent International Migration data: A New Method Applied to Flows in Europe.*European Journal of Population-Revue Europeenne De Demographie, 26*(4), 459-481. doi: 10.1007/s10680-010-9220-z

Feng, S. Z., Krueger, A. B., & Oppenheimer, M. (2010). Linkages among climate change, crop yields and Mexico-US cross-border migration.*Proceedings of the National Academy of Sciences of the United States of America, 107*(32), 14257-14262. doi: 10.1073/pnas.1002632107

Hargreaves, J. R., Collinson, M. A., Kahn, K., Clark, S. J., &Tollman, S. M. (2004).Childhood mortality among former Mozambican refugees and their hosts in rural South Africa.*International Journal of Epidemiology, 33*(6), 1271-1278. doi: 10.1093/ije/dyh257

Hugo, G. (1995). International labor migration and the family: Some observations from Indonesia. *Asian and Pacific Migration Journal, 4*(2-3), 273-301.

Hugo, G. (2002). Effects of international migration on the family in Indonesia.*Asian and Pacific Migration Journal, 11*(1), 13-46.

Kemper, R. V. (2005). Mexicans in the United States. In M. Ember, C. R. Ember & I. Skoggard (Eds.), *Encyclopedia of Diasporas*. New York: Springer.

Lindstrom, D. P. (1996).Economic opportunity in Mexico and return migration from the United States.*Demography, 33*(3), 357-374. doi: 10.2307/2061767

Marcelli, E. A., & Cornelius, W. A. (2001). The changing profile of Mexican migrants to the United States: New evidence from California and Mexico. *Latin American Research Review, 36*(3), 105-131.

McDonald, D. A., Zinyama, L., Gay, J., de Vletter, F., & Mattes, R. (2000). Guess who's coming to dinner: Migration from Lesotho, Mozambique and Zimbabwe to South Africa. *International Migration Review, 34*(3), 813-841. doi: 10.2307/2675946

Nowok, B., Kupiszewska, D., &Poulain, M. (2006). Statistics on international migration flows. In M. Poulain, N. Perrin & A. Singleton (Eds.), *THESIM: Towards HarmonisedEuropean Statistics on International Migration* (pp. 203-231). Louvain-la-Neuve: UCL Presses Universitaires de Louvain.

Ozden, C., Parsons, C. R., Schiff, M., &Walmsley, T. L. (2011). Where on Earth is Everybody? The Evolution of Global Bilateral Migration 1960-2000. *World Bank Economic Review, 25*(1), 12-56. doi: 10.1093/wber/lhr024

Perruchoud, R., &Redpath-Cross, J. (2011).*Glossary on migration* (2nd ed.). Geneva, Switzerland: International Organization for Migration.

Piguet, E. (2010). Linking climate change, environmental degradation, and migration: a methodological overview.*Climate change, 1*(4), 517-524.

Pilkington, H. (1998). *Migration, displacement and identity in post-Soviet Russia*. London: Routledge.

Polzer, T. (2007).Adapting to changing legal frameworks: Mozambican refugees in South Africa.*International Journal of Refugee Law, 19*(1), 22-50.

Poulain, M. (1993). Confrontation des statistiques de migrations intra-europeennes: Vers plus d'Harmonisation? *European Journal of Population, 9*(4), 353-381.

Poulain, M. (1999). International migration within Europe: Towards more complete and reliable data? (Vol. Working Paper No. 37). Perugia, Italy: Statistical Office of the European Communities (Eurostat).

Raymer, J. (2007). The estimation of international migration flows: a general technique focused on the origin-destination association structure. *Environment and Planning A, 39*(4), 985-995.doi: 10.1068/a38264

Raymer, J. (2008). Obtaining an overall picture of population movement in the European Union. In J. Raymer& F. Willekens (Eds.), *International migration in Europe: Data, models and estimates* (pp. 209-234). Chichester, U.K.: Wiley.

Raymer, J., de Beer, J., & van der Erf, R. (2011).Putting the pieces of the puzzle together: Age and sex-specific estimates of migration between EU/EFTA countries, 2002-2007.*European Journal of Population, 27*, 185-215.

Raymer, J., Foster, J. J., Smith, P. W. F., Bijak, J., &Wisniowski, A. (2012).Integrated modelling of European migration: Background, specification and results (pp. 1-17). London, U.K.: NORFACE Research Programme on Migration.

Raymer, J., & Rogers, A. (2007).Using age and spatial flow structures in the indirect estimation of migration streams.*Demography, 44*(2), 199-223. doi: 10.1353/dem.2007.0016

RCoreTeam. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, http://www.R-project.org/.

Rogers, A., & Castro, L. (1981). Model migration schedule (pp. 1-153). Laxenburg, Austria: International Institute for Applied Systems Analysis.

Rogers, A., Castro, L., & Lea, M. (2005). Model migration schedules: Three alternative linear parameter estimation methods. *Mathematical Population Studies, 12*(1), 17-38.

Rogers, A., Jones, B., Partida, V., &Muhidin, S. (2007). Inferring migration flows from the migration propensities of infants: Mexico and Indonesia. *Annals of Regional Science, 41*(2), 443-465. doi: 10.1007/s00168-006-0107-1

Rogers, A., Raquillet, R., & Castro, L. J. (1978).Model migration schedules and their applications.*Environment and Planning A, 10*(5), 475-502.doi: 10.1068/a100475

Rogers, A., &Raymer, J. (1998). The spatial focus of U.S. interstate migration flows. *International Journal of Population Geography, 4*, 63-80.

Rogers, A., Raymer, J., &Willekens, F. (2002).Capturing the age and spatial structures of migration.*Environment and Planning A, 34*(2), 341-359.doi: 10.1068/a33226

Rogers, A., & Wilson, R. T. (1996).Representing structural change in US migration patterns.*GeographicalAnalysis, 28*(1), 1-17.

Ruiz-de-Velasco, J., Fix, M., &Clewell, B. C. (2000). *Overlooked and underserved: Immigrant students in U.S. secondary schools*. Washington, D.C.: The Urban Institute.

Tsai, P. L., &Tsay, C. L. (2004).Foreign direct investment and international labour migration. In A. Ananta& E. N. Arifin (Eds.), *International Migration in Southeast Asia* (pp. 94-136). PasirPanjang, Singapore: Institute of Southeast Asian Studies.

UNDESA. (2008). United Nations Global Migration Database (UNGMD). New York: United Nations, Department of Economic and Social Affairs, Population Division.

UNECE.(2012). *Review of sources and quality of statistics on international migration in selected countries of the Commonwealth of Independent States*. New York: United Nations Economic Commission for Europe.

UNPD.(2005). World population prospects 2004 revision. New York: United Nations Department of Economic and Social Affairs, Population Division.

UNPD. (2009a). Age-specific fertility rates by major area, region, and country *World Population Prospects: The 2008 Revision*. New York: United Nations, Department of Economic and Social Affairs, Population Division.

UNPD. (2009b). Male life table survivors at exact age, l(x), by major area, region and country *World Population Prospects: The 2008 Revision*. New York: United Nations, Department of Economic and Social Affairs, Population Division.

UNPD. (2011). International migration report 2009: A global assessment. New York: United Nations Department of Economic and Social Affairs, Population Division.

UNSD.(2013). Composition of macro geographical (continental) regions, geographical sub-regions, and selected economic and other groupings. New York: United Nations

Statistical Division, retrieved June 18, 2013, from http://unstats.un.org/unsd/methods/m49/m49regin.htm.

Zlotnik, H., Guest, P., Hovy, B., & Henning, S. (2010). Data and analysis: Partnering to better understand and address the human development implications of migration. Geneva, Switzerland: United Nations Department of Economic and Social Affairs.

Appendix

Appendix Table 1: Definition of 31 global regions

No.	Region code	Region name	Countries in region
1	910	Eastern Africa	Burundi; Comoros; Djibouti; Eritrea; Ethiopia; Kenya; Madagascar; Malawi; Mauritius; Mayotte; Mozambique; Réunion; Rwanda; Seychelles; Somalia; South Sudan; Uganda; United Republic of Tanzania; Zambia; Zimbabwe
2	911	Middle Africa	Angola; Cameroon; Central African Republic; Chad; Congo; Democratic Republic of the Congo; Equatorial Guinea; Gabon; Sao Tome and Principe
3	912	Northern Africa	Algeria; Egypt; Libya; Morocco; Sudan; Sudan; Tunisia; Western Sahara
4	9130	R. Southern Africa	Botswana; Lesotho; Namibia; Swaziland
5	710	South Africa	South Africa
6	914	Western Africa	Benin; Burkina Faso; Cape Verde; Cote d'Ivoire; Gambia; Ghana; Guinea; Guinea-Bissau; Liberia; Mali; Mauritania; Niger; Nigeria; Saint Helena; Senegal; Sierra Leone; Togo
7	915	Caribbean	Anguilla; Antigua and Barbuda; Aruba; Bahamas; Barbados; Bonaire, Saint Eustatius and Saba; British Virgin Islands; Cayman Islands; Cuba; Curaçao; Dominica; Dominican Republic; Grenada; Guadeloupe; Haiti; Jamaica; Martinique; Montserrat; Puerto Rico; Saint-Barthélemy; Saint Kitts and Nevis; Saint Lucia; Saint Martin (French part); Saint Vincent and the Grenadines; Sint Maarten (Dutch part); Trinidad and Tobago; Turks and Caicos Islands; United States Virgin Islands; Netherlands Antilles
8	9160	R. Central America	Belize; Costa Rica; El Salvador; Guatemala; Honduras; Nicaragua; Panama
9	484	Mexico	Mexico
10	9310	R. South America	Argentina; Bolivia (Plurinational State of); Chile; Colombia; Ecuador; Falkland Islands (Malvinas); French Guiana; Guyana; Paraguay; Peru; Suriname; Uruguay; Venezuela (Bolivarian Republic of)
11	76	Brazil	Brazil
12	9050	Canada and R. N. America	Bermuda; Canada; Greenland; Saint Pierre and Miquelon
13	840	USA	United States of America
14	9210	R. South-	Kazakhstan; Kyrgyzstan; Tajikistan; Turkmenistan;

		Central Asia	Uzbekistan; Afghanistan; Bangladesh; Bhutan; Iran (Islamic Republic of); Maldives; Nepal; Pakistan; Sri Lanka
15	156	China	China; China, Hong Kong Special Administrative Region; China, Macao Special Administrative Region
16	9060	R. Eastern Asia	Democratic People's Republic of Korea; Mongolia
17	392	Japan	Japan
18	410	Korea	Republic of Korea
19	356	India	India
20	9200	R. South-Eastern Asia	Brunei Darussalam; Cambodia; Lao People's Democratic Republic; Malaysia; Myanmar; Philippines; Singapore; Thailand; Timor-Leste; Viet Nam
21	360	Indonesia	Indonesia
22	9220	R. Western Asia	Armenia; Azerbaijan; Bahrain; Cyprus; Georgia; Iraq; Israel; Jordan; Kuwait; Lebanon; Occupied Palestinian Territory; Oman; Qatar; Saudi Arabia; Syrian Arab Republic; United Arab Emirates; Yemen
23	792	Turkey	Turkey
24	9230	R. Eastern Europe	Belarus; Bulgaria; Czech Republic; former Czechoslovakia; former German Democratic Republic; Hungary; Poland; Republic of Moldova; Romania; Slovakia; Ukraine; former USSR
25	643	Russia	Russian Federation
26	924	Northern Europe	Åland Islands; Channel Islands; Denmark; Estonia; Faeroe Islands; Finland; Guernsey; Iceland; Ireland; Isle of Man; Jersey; Latvia; Lithuania; Norway; Sark; Svalbard and Jan Mayen Islands; Sweden; United Kingdom of Great Britain and Northern Ireland
27	925	Southern Europe	Albania; Andorra; Bosnia and Herzegovina; Croatia; Gibraltar; Greece; Holy See; Italy; Malta; Montenegro; Portugal; San Marino; Serbia; Slovenia; Spain; The former Yugoslav Republic of Macedonia; former Yugoslavia; Serbia and Montenegro
28	926	Western Europe	Austria; Belgium; France; Germany; former Federal Rep. of Germany; Liechtenstein; Luxembourg; Monaco; Netherlands; Switzerland
29	36	Australia	Australia
30	554	New Zealand	New Zealand
31	9090	R. Oceania	Norfolk Island; Fiji; New Caledonia; Papua New Guinea; Solomon Islands; Vanuatu; Guam; Kiribati; Marshall Islands; Micronesia (Federated States of); Nauru; Northern Mariana Islands; Palau; American Samoa; Cook Islands; French Polynesia; Niue; Pitcairn; Samoa; Tokelau; Tonga; Tuvalu; Wallis and Futuna Islands

Appendix Table 2: Variables for evaluating the quality of selected data files used to generate migration stock and flow estimates

Variable	Description	Coding
A. Indicator variables for year 1 and year 2 migrant stock data		
trcat	Identifies the treatment category of the particular profile. Category 1 files are of the highest quality because the used profiles were available directly from the particular year and stream. Categories 2a, 3a, and 4a are of slightly lower quality because we derived profile information from another year, but from the same stream. Categories 2b, 3b, 4b, 4c are of lowes quality because we derived the profiles from region-level information.	1 = total, age & gender 2 = total & gender 3 = total & age 4 = total
profyear	Identifies the year from which we derived the respective profile. Allows judgement of the time difference between profile and raw data (relevant for "b" categories).	Numeric
profID	File name of the source data from which we derived the respective profile. For "b" categories, the ID allows identifying the hierarchical level at which we generated the profile (e.g., regional, continental, global, etc.)	String
profUpSt	Indicates whether the profile was derived from upper level streams.	1 = profile from upper-level streams 0 = profile from same stream
profGendSpec	Indicates whether the derived profile has gender differentiated age groups.	1 = gender differentiated age groups 0 = age groups not gender differentiated
profCount	Identifies the number of country-level streams that contributed to the particular upper-level profile. Higher numbers suggest better representation of the region-level age and gender profiles.	Numeric
unifGendProf	Identifies whether we used a uniform gender distribution. Applies to cases where the data file provides age but no gender information.	1 = uniform gender distribution 0 = all other cases
B. Quality flags for migrant flow measures		

trcat1Both	Indicates whether we computed the flows using category 1 files for both years.	1 = flows computed from two category 1 files 0 = flow computation involves other categories
criterionSame	Indicates whether both files used the same criterion of migrant enumeration (country of birth vs. country of citizenship).	1 = files use same criterion 0 = files use different criterion
profileDif	Indicates whether both profiles came from different years.	1 = profiles from different years 0 = profiles from same year