UNU-CRIS
WORKING PAPER SERIES

#04
May
2022

Where Did They Come From, Where Did They Go?

Bridging the Gaps in Migration Data

Samuel Standaert
Glenn Rayp

www.cris.unu.edu

UNITED NATIONS UNIVERSITY
UNU-CRIS
Institute on Comparative Regional Integration Studies

in alliance with
VUB VRIJE UNIVERSITEIT BRUSSEL
UNIVERSITEIT GENT
Flanders State of the Art
west-vlaanderen de gedreven provincie
BRUGGE

About the authors:

Samuel Standaert is a Professorial Fellow at UNU-CRIS and Assistant Professor at Ghent University. In this role, he works on various topics, including the effects and determinants of regional integration agreements, the gravity model of international trade, sustainable development, Bayesian econometrics, and migration policy. He also coordinates the work at UNU-CRIS on RIKS 2.0. At Ghent University, he supervises PhD students working on various projects in the field of regional integration agreements and teaches several courses on economics.

Contact: sstandaert@cris.unu.edu

Glenn Rayp is Professorial Fellow at UNU-CRIS and Professor of International Economics at Ghent University. He coordinates the trade and investment cluster of UNU-CRIS. His current research focuses on the impact of international trade on productivity and labour demand and on the efficiency and impact of regional integration.

Contact: grayp@cris.unu.edu

Abstract

Many research analyses monitoring the patterns and evolution of international migration would benefit from high-frequency data on a global scale. However, the presently existing databases force a choice between the frequency of the data and the geographical scale. Yearly data exist but only for a small subset of countries, while most others are only covered every 5 to 10 years. To fill in the gaps in the coverage, the vast majority of databases use some imputation method. Gaps in the stock of migrants are often filled by combining information on migrants based on their country of birth with data based on nationality or using 'model' countries and propensity methods. Gaps in the data on the flow of migrants, on the other hand, are often filled by taking the difference in the stock, which the 'demographic accounting' methods then adjust for demographic evolutions. This paper proposes a novel approach to estimating the most likely values of missing migration stocks and flows. Specifically, we use a Bayesian state-space model to combine the information from multiple datasets on both stocks and flows into a single estimate. Like the demographic accounting technique, the state-space model is built on the demographic relationship between migrant stocks, flows, births and deaths. The most crucial difference is that the state-space model combines the information from multiple databases, including those covering migrant stocks, net flows, and gross flows. The result of this analysis is a global, yearly, bilateral database on the stock of migrants according to their country of birth. This database contains close to 2.9 million observations on over 56,000 country pairs from 1960 to 2020, a tenfold increase relative to the second-largest database. In addition, it also produces an estimate of the net flow of migrants. For a subset of countries –over 8,000 country pairs and half a million observations– we also have lower-bound estimates of the gross in- and outflow.

Keywords: Bilateral migration data, Stock, Imputation, State-Space model

Contents

1 Introduction

Recent contributions to the empirical study of international migration have underscored an urgent need for more comprehensive data on migration stocks and flows. This data can be divided into four main categories. Bilateral migrant stocks count the total number of migrants from a particular country of origin currently living in a particular destination country. The net migration flows describe the yearly change in migrant stocks and equal the number of departures (gross outflow) minus the number of arrivals (gross inflow) from a particular origin country. Empirical analyses that wish to determine the size, causes and consequences of migration need this data to be ample and accurate, particularly if this research is to serve policymakers.

Over the past decades, the need for more detailed migration data has given rise to some important initiatives, most initiated by international organisations such as the United Nations, the World Bank and the OECD. The databases built by these organisations have provided researchers and policymakers with global-level data for the past 60 years. While such a long timespan corresponds roughly to other such databases, e.g. the UN Comtrade data on international trade flows, the sources upon which the migration data are based are much more irregular, and the data comes with many gaps. Overall, there is a direct trade-off between geographical reach and data frequency with migration data. For example, the World Bank's Global Bilateral Migration database covers over 200 origin and destination countries from the 1960s until today, but only every ten years. The OECD's International Migration Database, on the other hand, has yearly data but only covers 36 destination countries. Its coverage is still far from complete, even with the restriction in destination countries, as more than two-thirds of the data is missing. Direct measurements of the in, out or net flow of migrants are even less numerous and limited to a few dozen destination countries.

Many research analyses monitoring the patterns and evolution of international migration would benefit from high-frequency data on a global scale. However, the presently existing databases force a choice between the frequency of the data and the geographical scale. This problem is likely to persist as the primary sources for bilateral migration data are population censuses and registers, which have been already fully mapped. From this perspective, the sources for new empirical data on international migration have been exhausted. The only way to get more data, particularly for non-OECD destination countries, is through an imputation or estimation procedure.

Many of the databases on migrant stocks contain some imputed data flows. The most often used solution is to combine information on migration stock based on

the migrant's country of birth with that based on the migrant's nationality. Other techniques include the use of 'model' countries and propensity methods. The latter is used in particular when the migrants' origin was not registered to a specific country but a more general region (see section 2.1).

When data on the flow of migrants is missing, the changes in the stock are a valuable source of information. In his review of how to get migration flow data from stocks, Abel and Cohen (2019) distinguished between three methods: stock differencing, migration rates and demographic accounting. The first one entails simply taking the difference of migration stocks at different times, often with a five or even 10-year gap. Decreases in stocks are then either dropped or treated as reverse migration flows as in Beine and Parsons (2015). In the second approach, migration flow rates are computed as the origin-specific stock of migrants in the total stock of migrants from all origin and destination countries (i.e., all foreign-born populations). The flows are then computed as the worldwide flow of migrants multiplied by these migration flow rates. The last approach, labelled 'demographic accounting' by Abel and Cohen (2019), uses the fact that the total population born in a particular country has to be living either at home or in one of the destination countries. As such, changes in the stock are combined with information on births and deaths to estimate the global migration flows.

We propose a novel approach to estimating the most likely values of missing migration stocks and flows. Similar to Abel and Cohen (2019), our model is built on the demographic relationship between migrant stocks, flows, births and deaths. However, we use a Bayesian state-space model to combine the information from multiple datasets on both stocks and flows. There are several noteworthy differences between the approach followed here and previous approaches. The use of different databases allows us to get a better sense of the reliability of some of the sources.

Second, unlike migration rates or demographic accounting, each country-pair is considered separately. As we do not need to build contiguity tables, we can run the regression for an origin country as soon as one origin-destination combination is available. This enables us to increase the scale of the estimations considerably. Our final dataset covers the bilateral migration stocks and flows (based on the country of birth) yearly for 60 years, over 240 countries, 56,000 country pairs and almost three million observations.

In contrast to most stock differencing approaches, our net migration flow can be negative when the stock decreases over time, i.e., the number of people migrating to a destination country is lower than those leaving. Unlike Beine and Parsons (2015), we make a distinction between migrants leaving the country (either to re-

turn to their origin country or to move to a different destination country), which we call gross migration outflow, and people from the destination country that migrate to the origin country.

Finally, the resulting imputed stock and flow data are consistent: the yearly change in the stock of migrants is equal to the net flows minus the migrant deaths. The net flows, in turn, equal the difference between the gross in- and outflow. This cannot be said, e.g., of the OECD's international migration database. The correlation between change in the OECD's stock and its net flows (inflow-outflow) is practically zero (-0.012), and that of the change in the stock and inflows is even lower (-0.14).

The remainder of this paper is as follows. The following section discusses the data sources used to run the estimation model. Section 3 discusses the imputation algorithm and section 4 shows some of the results, after which we conclude.

2 Data sources on migration flows and stocks

2.1 Data on international migrant stocks

Three databases allow a worldwide comparison of the stock of migrants identified by country of origin and destination. Table 1 provides an overview.

The first database, Trends in International Migrant Stock (TIMS), has the most comprehensive coverage of countries, keeping track of the number of migrants in 232 origin and destination countries. It is published by the Population Division of the Department of Economic and Social Affairs of the United Nations Secretariat. TIMS tracks Migrants according to their country of birth, supplemented with data on citizenship when the former was unavailable. The data is available on a 5-yearly basis from 1990 to 2020. It is mainly based on census data augmented with information from population registers and representative surveys. The place of birth criterion is used to identify international migrants whenever available. This information was lacking for one in five destination countries, and the citizenship criterion was used instead. In addition to resolving differences in the definition of origin countries, the authors of the dataset also estimated the number of migrants for countries for which information was incomplete or lacking, using several estimation methods. In the absence of any data source, figures were estimated using information from 'model' countries (UNDESA, 2020). This is the case for six destination countries. Interpolation and extrapolation were used when two or more data points were available. If data were lacking for this, then the growth rate of the most adjacent period was applied. The most recent versions of the TIMS dataset keep track of whether the data measures migration based on country of birth or

nationality and whether it was imputed. This allowed us to separate the UN data into two indicators: migrant stock based on country of birth and migrant stocks based on nationality or both.[1] In addition, we discarded those observations that were entirely imputed.

The second database is the Global Bilateral Migration database (GBM), composed by Özden et al. (2011) and published by the World Bank. This was the first attempt to create a bilateral international migration matrix at the global level. It covers slightly fewer countries than the UN's database, but the main difference in coverage is over time. While GBM only provides ten-yearly estimates of migrant stocks, it covers the period from the 1960s to the 2000s. The World Bank has published two additional bilateral migration matrices for 2010 and 2018. The primary data source used to compose this dataset is the Global Migration Database compiled by the United Nations (UNDESA, 2008). Like TIMS, GMDB tracks migrants by country of birth but augments this with information on country of residence when the former is unavailable.[2] It also tries to fill in gaps in the dataset using interpolation and estimations based on regional shares and propensities to migrate, of which interpolation is found to work the best (Özden et al., 2011). The number of observations for which interpolation and estimation techniques were required is fairly substantial. Depending on the year considered, observations derived from interpolation represent between 20 and 40% of the total, and 40% of the data comes from aggregated categories separated using propensity methods.

The third and last database is the International Migration Database (IMDB), published by the OECD. It only contains information on 36 (OECD) destination countries and 226 origin countries, but unlike the TIMS and GMDB, it offers yearly data on the stock of migrants. The database starts in 1975 and for some origin countries has data on migration up to 2020. The data is sourced from national statistical offices, Eurostat and the United Nations. The types of data used to compile migration stock (and flow) data are the population censuses, population registers, residence permits and labour force surveys. Another notable difference with the GMDB (and to a lesser extent TIMS) is that the OECD has kept the data based on country of birth entirely separate from the data based on citizenship.

[1] Whether or not the stock data also included refugees was ignored as it only pertained a small number of observations.

[2] The population division of the United Nations hence publishes two databases on international migration: TIMS and GMB. The main difference is that the first only provides aggregate bilateral data at five-year intervals starting in 1990. In contrast, the second provides a breakdown of migration stocks by gender. This seemingly affects data availability, which is much more irregular in GMB than TIMS. Moreover, the gender breakdown is less relevant in this data imputation approach that focuses on the origin-destination aggregates.

2.2 Data on international migrant flows

In addition to data on the stock of migrants, several data sources measure the yearly net and gross flow of migrants.

Firstly, the OECD's IMDB dataset provides information on the gross migrant in- and outflow. This database covers about the same number of origin and destination countries as the stock data in IMDB during the same period and uses the same sources.

The UN also has a database on International Migration Flows to and from Selected Countries (IMFSC). The latest version, the 2015 revision, provides yearly in- and outflow data on 45 destination countries and 217 origin countries. The data source depends on the origin country but includes (municipal) population registers, arrival and departure cards at airports and statistical forms collected by the government. In 23 countries, the migration data are collected from population registers, 15 countries from residence permits, and seven from statistical forms when changing residency. The database identifies migrants' origin countries based on country of birth, citizenship and residence, in particular, the two last criteria for migration inflows (residence data available for 43 countries, citizenship data for 36 and only one by country of birth) UNDESA (2015)

The DEMIG C2C database was constructed as part of the Determinants of International Migration project. The migration data was scoured from both online sources and national historical archives (DEMIG, 2015). Unlike the other databases, the data is retained as reported by the national statistical sources and the data type and source are extensively detailed. Among other characteristics, DEMIG distinguishes between the type of flows (gross inflow, gross outflow and net flows), flows based on country of birth, origin and nationality, and the coverage of the dataset (foreigners, citizens, or both). As these are kept separate, the DEMIG C2C database supplies most of the indicators included in our analysis, even if many only contain a couple of hundred observations. To somewhat limit this problem, we ignored the distinction between gender and the collection method for this analysis.[3]

[3] For gender, we used the totals when available and the sum of men and women otherwise. For the collection method, we combined the different sources by using whichever was first available in the following order: 1) population register, 2) population register 12 months and over, 3) residence permit, 4) multiple registers, 5) border statistics, 6) work permits, 7) other registrations and 8) passenger surveys.

Table 1: Availability statistics of the migration data

Abbr.	Name	Source	Periodicity	Start	End	Origin	Dest.
Stock data							
TIMS	Trends in International Migrant Stock	UN	5-yearly	1990	2020	232	231
GBM	Global Bilateral Migration	World Bank	10-yearly	1960	2010	228	226
T2010	Bilateral Migration Matrix 2010	World Bank	–	2010	2010	215	215
T2018	Bilateral Migration Matrix 2018	World Bank	–	2018	2018	214	214
IMDB	International Migration Database	OECD	yearly (with gaps)	1975	2020	36	226
Flow data							
IMDB	International Migration Database	OECD	yearly (with gaps)	1995	2020	36	226
DEMIG	Determinants of International Migration	DEMIG	yearly (with gaps)	1932	2011	34	240
IMFSC	International Migration Flows to and from Selected Countries	UN	yearly (with gaps)	1980	2013	45	217

2.3 Supporting databases

In order to complete our demographic model, we have to account for the death of migrants in the destination country, which is unfortunately not available for the vast majority of countries. However, we can proxy it using the information on the destination-country specific death rates combined with the current stock of migrants. As the age profile of migrants can be markedly different from that of the total population of the destination country, using the raw mortality rates risks overestimating migrant deaths. One solution is to use the age-standardized death rates published by World Health Organization. Their main downside is that they are available for far fewer countries and periods. Specifically, the raw mortality rates have four times as many observations and cover five times as many of the destination countries.

2.4 Combined database

Putting these different sources together, we have a total of 41 indicators of migration flows, the full details of which can be found in Appendix A. Our combined

database distinguishes between indicators based on their source, flow type, coverage, and criterion. The type of flow refers to whether the indicators measure the migration stock (5), the net migration flow (8), the gross inflow (14) or gross outflow (15). The coverage of an index indicates whether the measured flow refers to foreigners, citizens, or both. The interpretation of the coverage depends on which country is reporting the data. If the data comes from Australia, the flow of citizens counts the number of Australians, while foreigners track people from the rest of the world DEMIG (2015, p. 33). Finally, the criterion distinguished between those indicators that count migrants based on their country of birth, their nationality or their latest country of residence.

Some of the indicators in the dataset directly capture what we are trying to measure. For example, both S^1 and S^3 capture the stock of (foreign) migrants, N^1, the net flow, and I^7 and O^8 the gross inflow and outflow based on country of birth. Even if these indicators were available more frequently, combining them directly –adding the flow data to the stocks– would still be difficult as they are often not consistent. Firstly, different sources give different values even when measuring the same type of flow, coverage and criterion. Secondly, the change in the stock is not equal to the sum of the flow data, even for data that comes from the same source. For example, the difference in the stock in the OECD data rarely equals the difference between inflow and outflow. For this reason, our estimation model includes almost all indicators with some form of measurement error.

Together with the information on the type of data, Appendix A also summarises the availability of each indicator. The most appropriate indicators only cover a fraction of the other indicators. For example, I^7 covers fewer than a sixth of the country pairs and observations of the IMDB's inflows. For this reason, we also include the indicators that use different a criterion or coverage as they capture either part of the flow that we are interested in or are likely to be correlated with it. The relationship between these indicators and our main variables of interest can be relatively complex. Take, for example, indicator I^{13}, which counts the inflow of both foreigners and citizens based on country of birth. If we look at the data reported by France and referring to Belgium, this indicator is the sum of people born in Belgium moving to France and the people born in France returning from Belgium. While the former corresponds to our measure of inflow from Belgians into France, the returning migrants are not equal to the outflow of French from Belgium, as they do not necessarily return to their country of birth. For this reason, our error model allows for scale differences and country-pair fixed effects and heteroskedasticity.

In total, there are 56,084 country pairs and 650,711 observations for which we have at least one indicator available. The database starts in 1932 and ends in

2020, with a gap in the coverage during World War II. Once we fill in the gaps between the data, we have a dataset with almost three million observations. As we will show in section 4, this database covers four-fifths of all possible country-pair combinations each year, with an additional 400,000 observations for countries before they gained independence. As can be expected from the description in the previous paragraph, the availability of data can be widely different depending on the country pair in question. Specifically, we distinguish between five groups. These groups are identified and separated from each other in the following order:

1. **Only zeros.** The first group consists of those country pairs where all available stock and flow indicators indicate that no migration is taking place between both countries. There is migration stock and flow information available for these country pairs and is consistently equal to zero. This group is surprisingly large, consisting of close to a fifth of all non-missing data (124,784) and two-fifths of all observations (1,106,295) and country pairs (23,152). While it contains all destination and origin countries, the smaller island nations (e.g., Niue or Tuvalu) are the most frequent.

2. **Not enough data.** By far the smallest group are those country pairs where we have fewer than five years with data available (66 country pairs) or where only information on migration flows is available (143 country pairs). Both groups were left out of the analysis. The former has too few data points to identify the model's parameters reliably. For the latter, without any information on the (initial) stock of migrants, the imputations would make little sense.

3. **Only migration stock data** The third group consists of those country pairs where only information on the stock of migrants is available. This group is similar to the first one, covering 36% of observations (1,066,462). It covers slightly fewer country pairs (19,899) but contains more observations per country pair and more of the non-missing observations (147,534).

4. **Only migration stock and net migration flows** We have access to both stock and flow data for the next group, but only the net flows. While the size of this group in terms of the non-missing data is very similar to the first and third group (101,952), it covers only 7% of the country pairs (3,676) and overall observations (218,533) as the available data per country pair is much higher.

5. **Migration Stocks and at least one gross flow** The final group contains all other country pairs, namely those where we have information on migration stocks and at least one gross migration flow. Almost all country pairs have some info on the gross outflows, half cover the gross inflows and two fifths also have information on the net flows. In terms of the non-missing data, this last group is the largest one containing 42% of the available data (272,338).

As the availability per country pair is again much higher, this corresponds to only 18% of the overall dataset (516,697) and 16% of the country pairs (8,706).

3 Imputation algorithms

A key issue that had to be overcome when designing the imputation algorithm was the paucity of the migration data. For example, the country pair with the highest data availability (Fins migrating to Germany) does not have information on migrant stocks for more than a third of the years from any source, and two-thirds of the individual stock and flow indicators are missing. The two main ways we dealt with this were by adjusting the algorithm based on which data was available and anchoring the results to one of the stock variables.

By anchoring the data, we mean that one of the migration data sources was chosen as the correct value. Any deviations in other sources are interpreted either as measurement errors or structural differences with our anchor. Failing to pick an anchor would often lead to nonsensical results or even the algorithm crashing. While this seems like a strong assumption, this is in line with the rest of the literature as, e.g., Beine and Parsons (2015) or Abel and Cohen (2019) also trust the available data. As their imputations rely on a single data source, this assumption is not specified as overtly as it is here.

As anchors, we chose the UN's TIMS stock data or, if that was unavailable, the OECD's IMDB stock data, both of which are based on the country of birth. These particular sources were chosen because a lot of the stock data comes from census data which tends to be of higher quality than the information from population registers. Both sources were preferred over the World Bank's GBM data. The latter does not track whether their migrant stock data is based on country of birth or nationality, nor does it keep track of whether the migrant stock of a particular country is imputed.

The second way we dealt with the overabundance of missing data is by simplifying the imputation algorithm when not enough data was available for that particular country pair. If we only have information on the migration stocks, then modelling the relationship between the net and gross flows only adds uncertainty to the imputations. Similarly, if all of the available information on migration stocks and flows is zero, it makes little sense to estimate a complex state-space model. In these cases, the value for the intervening years was also set to zero.

The remainder of this section describes the different imputation models when at least one indicator was strictly positive, starting with those country pairs where we

only have information on the stock of migrants.

3.1 Only stock data

The state equation is built on a demographic identity. Namely, the only way the stock of migrants based on country of birth can change is if migrants enter the country, leave the country, or die.[4] If $S_{ij,t}$ is the Stock of migrants from i in j at time t, $N_{ij,t}$ are the net flows from i to j and $D_{ij,t}$ is number of immigrants from i in j that have died in year t, this gives us the following equation:

$$S_{ij,t} \equiv S_{ij,t-1} + N_{ij,t} - D_{ij,t} \tag{1}$$

For the vast majority of countries, the information on how many migrants have died per origin country is not available. For this reason, we assume that the deaths equal to the stock of migrants already in the country multiplied by a destination-country-specific death rate.

$$D_{ij,t} = \delta_{i,t}\ S_{ij,t-1}, \tag{2}$$

As many of the variables that influence migration flows are highly persistent (e.g., the size of the migrant population or that of the origin country), we also want to allow for this persistence in the net migration flows. To that end, we model this variable as an autoregressive process with one lag process. The level of persistence in these flows is estimated within the model.

$$\begin{aligned} N_{ij,t} &= \tau_{ij}^{N}\ N_{ij,t-1} + \mu_{ij,t}^{N} \\ \mu_{ij,t}^{N} &\sim N(0, \sigma_{ij}^{N}) \end{aligned} \tag{3}$$

The measurement equation is a simple linear error model with country pair fixed effects. If $\hat{S}_{ij,t}^{k}$ is the k^{th} variable that measures the stock of migrants based on their country of birth, the measurement equation is as follows:

$$\begin{aligned} \hat{S}_{ij,t}^{k} &= c_{ij}^{Sk} + z^{Sk} S_{ij,t} + \epsilon_{ij,t}^{Sk} \\ \epsilon_{ij,t}^{Sk} &\sim N(0, \Omega_{ij}^{Sk}) \end{aligned} \tag{4}$$

The intercept term z captures the overall difference between the actual migrant stock $S_{ij,t}$ and the specific data source $\hat{S}_{ij,t}^{k}$. This allows for, e.g., the differences

[4] Depending on the legal system, babies born from immigrant mothers are counted as an increase in the domestic population or the net migration flow. Either way, the births are already incorporated.

between migrant flow based on country of birth versus nationality. In addition, each country pair has a specific constant term c_{ij}^{Sk}, that can capture persistent measurement errors due to, e.g., the use of different sources or a lower quality of data collection in a particular destination country. The error term $\epsilon_{ij,t}^{k}$ accounts for any stochastic deviation and is corrected for potential heteroskedasticity and the county-pair level. A second reason for allowing the constant and the variance of the error term to differ over country pairs is that the magnitude of the flow and stock of migrants can be very different depending on the countries in question.

Putting these equations together, we get the following state-space model:

$$\begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} S_{ij,t} \\ N_{ij,t} \end{bmatrix} = \begin{bmatrix} 1-\delta_{j,t} & 0 \\ 0 & \tau^{N} \end{bmatrix} \begin{bmatrix} S_{ij,t-1} \\ N_{ij,t-1} \end{bmatrix} + \begin{bmatrix} 0 \\ \mu_{ij,t}^{N} \end{bmatrix} \tag{5}$$

$$\begin{bmatrix} \hat{S}_{ij,t}^{1} \\ \vdots \end{bmatrix} = \begin{bmatrix} c_{ij}^{S1} \\ \vdots \end{bmatrix} + \begin{bmatrix} z^{S1} & 0 \\ \vdots & \vdots \end{bmatrix} \begin{bmatrix} S_{ij,t} \\ N_{ij,t} \end{bmatrix} + \begin{bmatrix} \epsilon_{ij,t}^{S1} \\ \vdots \end{bmatrix} \tag{6}$$

where $\left[\hat{S_{ij,t}^{1}}, \ldots\right]'$ is a vector containing all variables that measure the stock of migrants based on country of birth or nationality.

3.2 Only stock and net flows

When we also have information on the net flow of migrants, the changes to the model are only minimal, as it only requires the addition of a measurement equation for the flow data. Similar to Equation 4, we model this as a simple linear error model:

$$\begin{aligned} \hat{N}_{ij,t}^{k} &= c_{ij}^{Nk} + z_{ij,t}^{Nk} + \epsilon_{ij,t}^{Nk} \\ \epsilon_{ij,t}^{Nk} &\sim N(0, \Omega_{ij}^{Nk}) \end{aligned} \tag{7}$$

This changes equation 6 to

$$\begin{bmatrix} \hat{S}_{ij,t}^{1} \\ \vdots \\ \hat{N}_{ij,t}^{1} \\ \vdots \end{bmatrix} = \begin{bmatrix} c_{ij}^{S1} \\ \vdots \\ c_{ij}^{N1} \\ \vdots \end{bmatrix} + \begin{bmatrix} z^{S1} & 0 \\ \vdots & \vdots \\ 0 & z^{N1} \\ \vdots & \vdots \end{bmatrix} \begin{bmatrix} S_{ij,t} \\ N_{ij,t} \end{bmatrix} + \begin{bmatrix} \epsilon_{ij,t}^{S1} \\ \vdots \\ \epsilon_{ij,t}^{N1} \\ \vdots \end{bmatrix} \tag{8}$$

3.3 Stock and at least one gross flow

Once we have information on either the gross inflow or gross outflow of migrants, we can further separate our net migration flows into their constituent flows:

$$N_{ij,t} \equiv I_{ij,t} - O_{ij,t} \tag{9}$$

where $I_{ij,t}$ is the migrant flow from i to j and $O_{ij,t}$ is the return migration returning to country i having stayed in country j.

While the migrant stocks constrain the net flows between two countries, this does not apply to the gross flows.[5] A given net migration flow can be the result of an infinite combination of gross in and outflows. For the regression model to provide sensible results, we had to take the gross outflow of migrants as a given. If no data was available on the gross flows, they were set to zero.

Similar to how the net flows were modelled (Equation 3) and for the same reason, we model the gross migration inflow as being autoregressive. The extent to which it depends on its previous values is endogenously determined.

$$\begin{aligned} I_{ij,t} &= \tau^I \, I_{ij,t-1} + \mu^I_{ij,t} \\ \mu^I_{ij,t} &\sim N(0, \sigma^I_{ij}) \end{aligned} \tag{10}$$

The relationship between the observed flows and the estimated flows is once again modelled as a linear error model with fixed effects. For the inflow data this equation is:

$$\begin{aligned} \hat{I}^k_{ij,t} &= c^{Ik}_{ij} + z^{Ik} I_{ij,t} + \epsilon^{Ik}_{ij,t} \\ \epsilon^{Ik}_{ij,t} &\sim N(0, \Omega^{Ik}_{ij}) \end{aligned} \tag{11}$$

Only the DEMIG dataset had information on net migrant flows, $\hat{N}^k$, which was almost always accompanied by information on gross outflows, $\hat{O}^k$. By adding both, we get the following measurement equations:

$$\begin{aligned} \hat{N}^k_{ij,t} - \hat{O}^k_{ij,t} &= c^{NOk}_{ij} + z^{NOk} I_{ij,t} + \epsilon^{NOk}_{ij,t} \\ \epsilon^{NOk}_{ij,t} &\sim N(0, \Omega^{NOk}_{ij}) \end{aligned} \tag{12}$$

[5] With fixed beginning and ending values of the migrant stocks, the autoregressive model ensures that the imputed values remain sensible, as long as the autocorrelation term τ^N is strictly positive.

Combining all of the above gives us the following state-space model:

$$\begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} S_{ij,t} \\ I_{ij,t} \end{bmatrix} = \begin{bmatrix} -O_{ij,t} \\ 0 \end{bmatrix} + \begin{bmatrix} 1-\delta_{j,t} & 0 \\ 0 & \tau^I \end{bmatrix} \begin{bmatrix} S_{ij,t-1} \\ I_{ij,t-1} \end{bmatrix} + \begin{bmatrix} 0 \\ \mu^I_{ij,t} \end{bmatrix} \quad (13)$$

$$\begin{bmatrix} \hat{S}^1_{ij,t} \\ \vdots \\ \hat{I}^1_{ij,t} \\ \vdots \\ \hat{N}^1_{ij,t} - \hat{O}^1_{ij,t} \\ \vdots \end{bmatrix} = \begin{bmatrix} c^{S1}_{ij} \\ \vdots \\ c^{I1}_{ij} \\ \vdots \\ c^{NO1}_{ij} \\ \vdots \end{bmatrix} + \begin{bmatrix} z^{S1} & 0 \\ & \vdots \\ 0 & z^{I1} \\ & \vdots \\ 0 & z^{NO1} \\ & \vdots \end{bmatrix} \begin{bmatrix} S_{ij,t} \\ I_{ij,t} \end{bmatrix} + \begin{bmatrix} \epsilon^{S1}_{ij,t} \\ \vdots \\ \epsilon^{I1}_{ij,t} \\ \vdots \\ \epsilon^{NO1}_{ij,t} \\ \vdots \end{bmatrix} \quad (14)$$

4 Results

The estimation model ran for 3000 iterations, of which the first 2000 were discarded as burn in.[6] The remaining iterations were used to compute the most likely values of the migration stocks and flows.

Overall, the imputed database covers over 56,000 country pairs, from 1960 to 2020, for a total of 2.89 million observations. This is a ten-fold increase relative to the second most abundant database, GBM, which makes sense as the latter is available only every decade. More than 700,000 observations cover origin or destination countries before they officially gained independence, although it should be noted that almost three-quarters of those are zero. Take, for example, Timor-Leste, where we have stock data from the 1960s, which predates its independence from Portugal (1975) and Indonesia (2001). However, almost 90% of this stock data is zero.

In order to compare the geographic coverage of the data, Table 2 compares the probability of the availability of stock data depending on the continent of the origin and destination countries. The numbers give the percentage of country pairs covered over the 1960 to 2020 time span for each combination. To help identify the geographic pattern in the data, the colour of each cell indicates relative abundance per source (darker red is less abundant). Unsurprisingly, the OECD's IMDB database has a higher coverage for European and North American destination countries, and a similar pattern can be seen in the UN's TIMS database. The

[6] We used uninformative priors and checked the model's convergence using a visual inspection of the parameters plots, autocorrelation function and CUMSUM graphs.

World Bank's GBM database, in contrast, is evenly distributed over almost all continents, except for Oceania. This pattern is repeated in the imputed data, although with higher coverage rates. Given that the World Bank data is available only every ten years, this ten-fold increase in the coverage matches our expectations.

Table 2: Percentage of country-pairs where stock data is available for each source, split up by continent.

Destination → / Origin ↓	Africa	Asia	Europe	Latin America	Middle East	North America	Oceania	Africa	Asia	Europe	Latin America	Middle East	North America	Oceania
	GBM (World Bank)							TIMS (UN)						
Africa	9.6%	9.3%	9.5%	8.9%	9.3%	10.0%	6.1%	2.7%	0.2%	5.8%	0.9%	1.4%	3.0%	0.7%
Asia	9.4%	9.4%	9.4%	8.9%	9.3%	9.8%	6.1%	0.9%	2.2%	5.6%	1.4%	2.0%	3.7%	1.5%
Europe	9.0%	8.9%	8.9%	8.6%	8.8%	9.3%	5.8%	1.1%	1.0%	7.2%	2.4%	1.4%	4.4%	0.9%
Latin America	8.8%	8.6%	8.7%	8.4%	8.5%	9.0%	5.5%	0.3%	0.1%	4.5%	3.9%	0.2%	6.1%	0.6%
Middle East	9.5%	9.3%	9.5%	9.0%	9.5%	9.9%	6.0%	1.2%	0.8%	6.8%	1.7%	6.9%	4.6%	0.7%
North America	9.5%	9.3%	9.3%	9.1%	9.2%	9.3%	6.0%	1.7%	1.9%	4.5%	5.6%	1.6%	8.4%	1.5%
Oceania	6.0%	6.0%	5.8%	5.5%	5.8%	5.9%	3.6%	0.2%	0.7%	1.3%	0.5%	0.2%	0.8%	1.2%
	IMDB (OECD)							Imputed						
Africa	0.0%	0.8%	11.5%	0.1%	0.3%	6.1%	2.2%	91.7%	89.4%	87.0%	82.6%	88.1%	87.9%	54.6%
Asia	0.0%	1.2%	12.0%	0.1%	0.2%	8.0%	2.5%	89.5%	88.7%	85.9%	81.4%	87.0%	86.6%	54.6%
Europe	0.0%	0.9%	14.3%	0.1%	0.7%	8.6%	2.4%	85.0%	83.9%	83.5%	78.0%	82.4%	82.9%	52.0%
Latin America	0.0%	0.6%	8.3%	0.1%	0.3%	7.1%	1.6%	82.3%	80.8%	78.8%	77.6%	79.0%	83.3%	49.3%
Middle East	0.0%	0.8%	12.1%	0.1%	0.6%	7.5%	2.4%	89.9%	88.5%	86.2%	81.5%	90.0%	87.0%	53.3%
North America	0.0%	1.0%	9.4%	0.1%	0.7%	4.8%	1.7%	87.2%	85.4%	83.0%	83.5%	84.1%	89.9%	52.2%
Oceania	0.0%	0.3%	3.3%	0.0%	0.1%	1.7%	0.7%	54.4%	54.3%	53.9%	49.4%	52.2%	52.2%	33.1%

The percentage of the country-pairs for which we have stock data between 1960 and 2020. The values are shaded to indicate the probability that a combination of origin and destination countries is covered (darker red = less likely).

Similar to Table 2, Table 3 counts the number of country pairs where there is data on net flows and gross in and outflows. A cell value of 100% would indicate that all three are available for the entire period. All three sources focus on European and North American destination countries, with the UN also covering the European origin countries. In contrast, the Imputed dataset covers most parts of the world more equal, although European origin and destination countries are almost twice as likely to be covered.

Figure 1 compares the source data with the imputed values for four country pairs where only data on migrant stocks are available. They are arranged in increasing size, from the less-than-a-handful migrants from Tajikistan living in Mali (panel a) to the tens of thousands of migrants from Vietnam living in Bangladesh (panel d). Only one data source is available for the first two country pairs (GBM), while the last two also include information from the TIMS database. In these last two panels, the GBM data (downward triangles) says migration stocks are zero until 2000, while the TIMS data (upward triangles) indicates a stock of thousands if not tens of thousands of migrants. While the overall correlation between both databases is relatively high (0.78), the examples illustrate that the inconsistencies can be considerable for a given country pair.

Table 4 shows the correlation between imputed data and the different sources, split up over the three main sections of the data. As the algorithm was set to prefer

Table 3: Percentage of country-pairs where flow data is available for each source, split up by continent.

Destination → / Origin ↓	Africa	Asia	Europe	Latin America	Middle East	North America	Oceania	Africa	Asia	Europe	Latin America	Middle East	North America	Oceania
	C2C - DEMIG							IMFSC (UN)						
Africa	0.3%	0.0%	8.7%	0.1%	0.1%	4.6%	2.9%	0.0%	0.0%	4.6%	0.0%	0.0%	5.4%	1.8%
Asia	0.2%	0.0%	8.0%	0.2%	0.2%	4.9%	3.4%	0.0%	0.2%	4.3%	0.0%	0.0%	5.2%	1.8%
Europe	0.5%	0.0%	12.4%	0.5%	0.5%	5.8%	3.6%	3.9%	3.8%	9.7%	3.1%	4.3%	7.9%	2.5%
Latin America	0.1%	0.0%	7.0%	0.4%	0.2%	5.0%	2.4%	0.0%	0.0%	3.4%	0.0%	0.0%	4.4%	1.6%
Middle East	0.3%	0.0%	9.4%	0.3%	0.2%	5.1%	2.8%	0.0%	0.0%	4.8%	0.0%	0.0%	5.5%	1.8%
North America	0.5%	0.0%	11.2%	0.7%	0.6%	4.5%	3.3%	0.0%	0.1%	3.5%	0.0%	0.0%	2.9%	1.6%
Oceania	0.1%	0.0%	3.0%	0.1%	0.1%	2.2%	1.9%	1.8%	1.8%	2.5%	1.6%	1.7%	3.2%	2.0%
	IMDB (OECD)							Imputed						
Africa	0.0%	0.7%	6.1%	0.1%	0.1%	4.2%	1.7%	31.3%	29.6%	44.0%	27.3%	29.3%	29.0%	21.7%
Asia	0.0%	0.8%	6.2%	0.2%	0.0%	4.4%	2.2%	30.5%	30.0%	43.2%	27.1%	29.2%	29.8%	22.8%
Europe	0.0%	0.7%	6.8%	0.2%	0.3%	4.2%	2.0%	54.0%	53.0%	66.6%	45.9%	58.5%	49.8%	25.2%
Latin America	0.0%	0.5%	4.4%	0.1%	0.2%	3.4%	1.3%	28.0%	26.8%	37.7%	25.8%	26.7%	27.3%	19.3%
Middle East	0.0%	0.7%	6.3%	0.2%	0.1%	4.4%	2.1%	30.6%	29.2%	43.5%	27.2%	30.3%	28.6%	21.4%
North America	0.0%	0.5%	4.6%	0.1%	0.4%	2.4%	1.3%	29.3%	28.2%	41.9%	28.1%	30.3%	29.4%	21.6%
Oceania	0.0%	0.2%	1.6%	0.0%	0.1%	0.8%	0.4%	24.6%	24.7%	26.0%	22.3%	23.8%	23.2%	17.0%

The percentage of the country-pairs for which we have data on the net flows or gross in and outflows between 1960 and 2020. A value of 100% indicates that all three flows are available. The values are shaded to indicate the probability that a combination of origin and destination countries is covered (darker red = less likely).

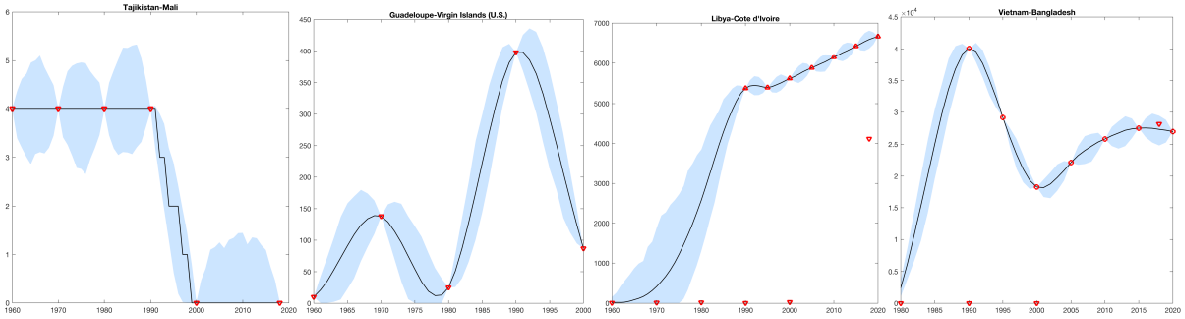
the TIMS stock data based on country of birth, the correlation with this source is always one. More surprising is the low correlation with the OECD's IMDB data when only stock data is available, especially as the correlation for the nationality-based data is higher than that for country of birth. This can be explained by a very low correlation between the OECD and UN data for this set of country pairs (0.38 and 0.47 for country of birth and nationality, respectively). Moreover, less than 5% of IMDB data is available. For the other two groups in the database, the correlation with the IMDB stock data (based on country of birth) is in excess of 96%.

Table 4: correlation between imputed and stock data

	TIMS		IMDB		GBM
	birth S^1	nat S^2	birth S^3	nat S^4	birth/nat S^5
Only Stock					
Imputed	1.00	1.00	0.43	0.74	0.93
	(25,600)	(6,947)	(3,001)	(9,637)	(113,807)
Stock and Net Flow					
Imputed	1.00	1.00	1.00	0.98	0.99
	(19,162)	(21)	(22,620)	(23,983)	(23,745)
Stock and at least one gross flow					
Imputed	1.00	0.64	0.96	0.81	0.97
	(29,163)	(2,803)	(45,226)	(55,739)	(52,448)

number of observations in brackets

Figure 2 plots the imputed values for two of the country pairs that have data on the stock and net flow of migrants. The left-hand panel compares the values for the migrant stocks and the right-hand panel for the net flows. Panel (a) shows



(a) From Tajikistan to Mali

(b) From Guadeloupe to Virgin Islands

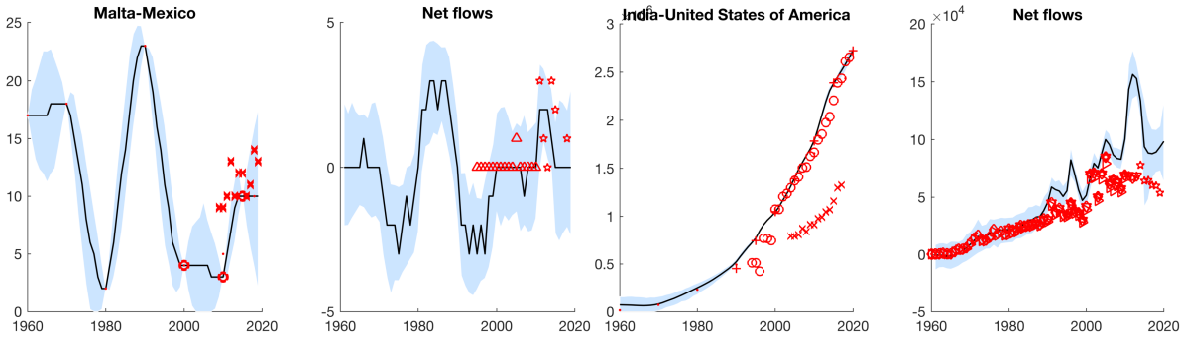
(c) From Libya to Côte d'Ivoire

(d) From Vietnam to Bangladesh

Figure 1: Imputations using only the stock values

Imputed values of migration stocks (black lines) and their 95% confidence intervals (blue shaded area). Available source data is also plotted using different symbols for each source.

the result for a country pair with minimal flows: the number of Maltese living in Mexico is less than 30. In contrast, the number of Indian migrants in the United States (panel b) numbers in the hundreds of thousand. Panel (a) illustrates the data smoothing by the state-space model. In the absence of more information, the large swings in the number of Maltese migrants between 1970 and 2000 are spread out. However, when data on the net flows are available, like in 2010-2020, the number of migrants can suddenly increase from one year to the next.



(a) From Malta to Mexico

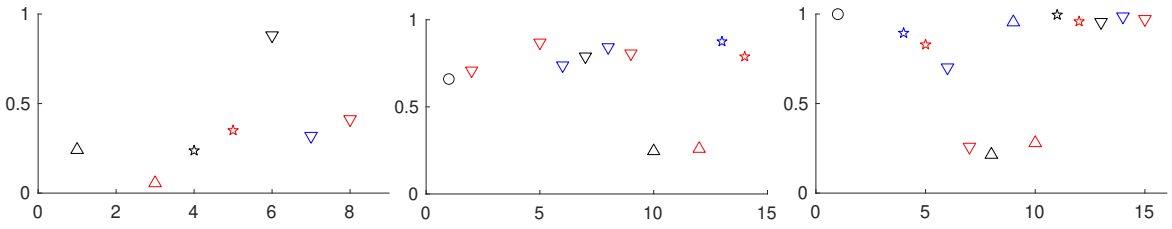
(b) From India to United States of America

Figure 2: Imputations using only the stock and net flows

Black lines represent the imputed values of migration stocks (left-hand panel), net flows right-hand panel) and their 95% confidence intervals (blue shaded area). Available source data is also plotted using different symbols for each source.

There can be considerable inconsistencies between the different data sources for many of the country pairs. For example, the nationality-based IMDB data in the stock plot of panel (b) in Figure 2, marked by x, lies well below the other data. One possible explanation could be that the IMDB data measures migration based on nationality. However, other nationality-based sources are much more in line with

the rest of the database. Moreover, the sum of the flows can differ markedly from the change in the stocks. While the flow and stock of Indian migrants are consistent in the first half of the sample, from the 1990s onward, there are considerable differences between the imputed flow data –which has to be consistent with the stocks– and the measured migration flows. Overall, the correlation between the imputed and measured net flows is positive but very low except for N[6], the DEMIG data based on country of birth, where the correlation is 0.88 (panel a, Figure 3).



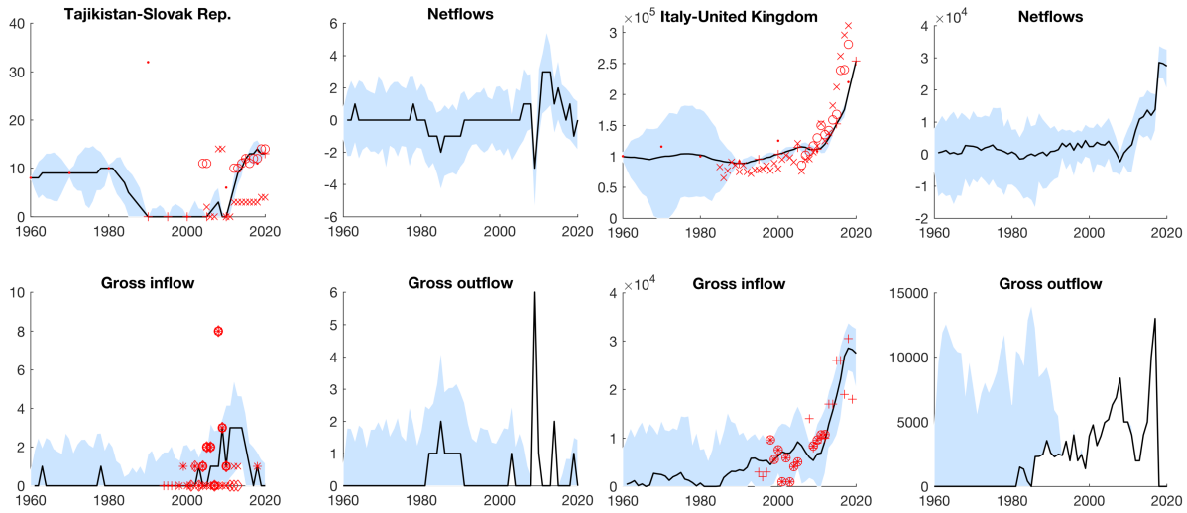
(a) Net flows (b) Gross inflows (c) Gross outflows

Figure 3: Correlations between the imputed in, out, and net-flows and the respective sources.

The plot identifies indicators based on country of birth (black), nationality (blue) and residence (red), and those based on citizens(△), foreigners (▽) or both (⋆), or unidentified (○).

Lastly, Figure 4 shows the correspondence between the imputed and source data for the group of countries for which we have the most data. In addition to the stock and net flows, the bottom panels show the gross in and outflows. Overall, the correlation between the imputed stocks and the source data is much higher than was the case for the first group of country pairs (Table 4). A likely explanation is that the data quality and availability have improved, resulting in a much higher consistency for the last two groups.

While the gross flows are internally consistent with the net flows and total stocks, in most cases, it more accurately represents the lower bound on the gross flows. Due to identification issues, a number of simplifying assumptions were made (cf. supra), especially for those years when no information on the gross flows was available. Specifically, in those cases, negative net flows are assigned to the gross outflow and inflows are set to zero (e.g., the migrant flows from Tajikistan to the Slovak Republic from 1980 to 1990 in panel a). Vice versa, positive net flows are assigned to the gross inflow and outflows are set to zero (e.g., the Italian migrants in the UK between 1960 and 1980 in panel b). Any positive integer can be added to both gross flows without affecting the net flows or stocks, so the gross flows represent a lower bound. Notwithstanding, the correlation between the gross inflow and the source data ends up being very high for the vast majority of indicators (panel b, Figure 3).



(a) From Tajikistan to Slovak Republic (b) From Italy to United Kingdom

Figure 4: Imputations using gross flows

Black lines represent the imputed values of migration stocks (upper left), net flows (upper right), gross inflow (bottom left) and gross outflow (bottom right) and their 95% confidence intervals (blue shaded area). Available source data is also plotted using different symbols for each source.

4.1 Out-of-sample validation

While the results outlined so far indicate a high correlation between the imputed data and the respective sources. However, these comparisons are all in-sample –comparing the imputed data with its source material– which biases the results in favour of the model. In this last section of the paper, we compare the imputed results with data not included in our model to get a better sense of the quality of the imputations.

First, we compare our imputed flow data with the dataset provided by Abel and Cohen (2019) which includes stock differences, migration rates and demographic accounting. The first difference is the size of both databases: Abel's imputed migrant flow data is available every five years between 1990 and 2015, while we have yearly data from 1960 to 2020. Abel has a square migration matrix with 201 origin and destination countries each year, and the demographic accounting technique focuses on filling in all of the cells in this matrix. In contrast, the state-space approach aims to fill the gaps between the measurements within a single country pair. As a result, Abel has information on several country pairs not covered in our dataset, but this only amounts to about 7% of his dataset. On the other hand, as we do not require this square migration table, our dataset still covers a quarter more country pairs than Abel (40,393 vs 56,084).

In order to compare the values of the imputed flow data with those of Abel and

Cohen (2019), we first have to put both on an equal footing. As such, we computed the sum of our yearly flow data over the preceding five years. We also compared both datasets to the UN's (IMFSC) and OECD (IMDB) measured flow data. The correlations between these figures are described in Table 5,

Table 5: Pariwise correlations with Abel and Cohen (2019)

	$\sum_{t=1}^{5}$ net flow$_t$	$\sum_{t=1}^{5}$ gross inflow$_t$	$\sum_{t=1}^{5}$ IMDB$_t$	$\sum_{t=1}^{5}$ IMFSC$_t$
Imputed data				
$\sum_{t=1}^{5}$ net flow$_t$	1	0.9276	0.8471	0.8496
$\sum_{t=1}^{5}$ gross inflow$_t$	0.9276	1	0.9171	0.9164
Stock difference				
Drop negative	0.6022	0.6092	0.5626	0.5685
Reverse negative	0.602	0.612	0.5673	0.5708
Migration rates				
	0.5457	0.7189	0.7611	0.7577
Demographic accounting				
Minimisation open	0.6194	0.644	0.6086	0.6135
Minimisation closed	0.4825	0.4492	0.5276	0.5361
Pseudo Bayesian closed	0.5572	0.6391	0.7081	0.7109

Overall, the correlations between our imputed data and the various techniques described in Abel and Cohen (2019) are positive but not extremely high, fluctuating around 0.6. Surprisingly, the correlation with the imputed net flows tends to be lower than that of the gross flows. However, this is likely due to a sample selection effect. There are fewer observations of the gross flows, and those tend to come from countries with higher quality data. Comparing both imputations to the measured migration flows from OECD and the UN, our imputations seem to follow the measured migration flows more closely. However, this should not be too surprising as those measured flows are incorporated in the state-space model estimations.

Secondly, we re-estimated our model with a reduced dataset and compared the results to the indicators that we left out. To that end, we ran two separate robustness checks, one in which we excluded the UN stock data (S^1 and S^2) and one in which we excluded both the OECD stock and flow data (S^3, S^4, I^1 and O^1). In both cases, we limited the sample to the group of countries where we have the highest data availability: those that have information on the migration stocks and gross flows (group 5, cf infra). Leaving out the TIMS data was likely to significantly affect the stability of the estimations as this source is used as the anchor for our baseline estimations. To compensate, we further restricted the dataset to a subset of origin and destination countries with the highest data availability: the twenty founding

members of OECD.[7] Table 6 compares these robustness check to our baseline results.

Table 6: Correlation with out-of-sample robustness checks

Robustness check 1: No UN stock data

Stock	*Overall*		*Within*	
	TIMS (S^1)	Stock	TIMS (S^1)	Baseline
Baseline	0.9999	–	0.9985	–
No TIMS ($S^{-[S^1,S^2]}$)	0.8714	0.8727	0.1554	0.1484
Inflow	*Overall*		*Within*	
	IMFSC (I)	Baseline	IMFSC (I)	Baseline
Baseline	0.8192	–	0.42	–
No TIMS ($I^{-[S^1,S^2]}$)	0.749	0.8633	0.3869	0.7484

Robustness check 2: No OECD stock or flow data

Stock	*Overall*		*Within*	
	IMDB (S^2)	Baseline	IMDB (S^2)	Baseline
Baseline	0.962	–	0.783	–
No OECD ($S^{-[S^3S^4I^1O^1]}$)	0.955	0.9954	0.771	0.981
Inflow	*Overall*		*Within*	
	IMDB (I^1)	Baseline	IMDB (I^1)	Baseline
Baseline	0.706	–	0.380	–
No OECD ($I^{-[S^3S^4I^1O^1]}$)	0.657	0.919	0.342	0.846

Internal consistency OECD data

OECD	*Overall*		*Within*	
	Δ Stock (S^2)	Inflow (I^1)	Δ Stock (S^2)	Inflow (I^1)
Inflow (I^1)	0.464	–	0.254	–
Net flow ($I^1 - O^1$)	0.157	0.297	0.069	0.23

The out-of-sample comparisons were obtained by rerunning the estimation model while excluding specific variables, indicated by the minus sign in the superscript. In the first robustness check (no TIMS), the sample was limited to migration between the OECD founding countries. The second robustness check (no OECD) used all country pairs for which data on the gross flows. The within correlation is computed as the correlation after the average value per country pair is subtracted.

Leaving out the UN stock data gives broadly similar results as the correlation between the baseline and the TIMS data is 87%. However, both models disagree quite considerably on how the stock of migration changes over time. The within correlation –i.e., the correlation after the mean value of each country-pair is subtracted– is only 15%. That being said, the correlation between the gross migration inflow

[7] This subset includes: Austria, Belgium, Canada, Denmark, France, Germany, Greece, Iceland, Ireland, Italy, Luxembourg, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, Turkey, United Kingdom and the United States of America. As noted above, the average data availability of the entire dataset is only 1.6%, and for the country pairs where gross migration flow data is available, this rises to 6.2%. When further limiting the dataset to the OECD founding countries, data availability increases by 2.5 to 15.6%.

and the UN IMFSC flow data remains high. Overall, even for this more restricted sample of countries with better data availability, keeping the imputed data anchored to the TIMS stock data has a considerable impact on the imputed changes in the stock data.

As expected, removing the OECD data from our dataset had a much smaller impact. The overall correlation with the excluded OECD stocks is very high (95%). The within correlations are slightly lower (77%) but still indicate a very close match between both data series. The match between the inflow data is noticeably worse, especially when considering the within correlation (34%). While it is difficult to pinpoint the exact source of the deviations, this is likely due to the fact that the imputations enforce consistency between the stock and flow data. As can be seen in the lower panel of Table 6, the OECD's IMDB stock data is not consistent with its flow data. The change in the stocks is barely correlated with the net flows, especially when considering the within correlation. The correlation with the IMDB's gross inflow is slightly higher but still well below what we find for the *no OECD* robustness check.

5 Conclusion

In this paper, we describe a new technique to better estimate migration stocks and flows. We analysed its reliability and compared its accuracy to other available methods.

Our technique consists of using a Bayesian state-space model. At the heart of the state-space model lies the demographic relationship between the stock of migrants on the one hand and the (gross) flows, deaths and births on the other. The model amalgamates information dispersed over widely differing datasets and fills in the gaps between reference years, which tend to be manifold in the case of migration data.

We found a strong correlation when comparing our imputed data to the source material. Moreover, our method consistently outperforms other imputation techniques that were put to the test.

The result of our work is publicly available on the Regional Integration Knowledge System platform, a web-based information and learning platform on regional integration processes worldwide.[8] It contains almost three million observations on migration between 1960 and 2020. It covers the majority of possible origin-destination combinations, providing an even distribution across the globe.

Potential users for this database can be found both in academia and in policymaking. Anyone interested in getting a more accurate picture of inflows, outflows and migration stocks between any given pair of countries in any given year over the past 60 years will find it a helpful tool. One way in which the database has already proved its worth is by mapping out the extent to which migration flows are stimulated in regional integration agreements. Another critical usage can be found in gravity model analysis of migration flows, especially those suffering from sample selection bias. Finally, the database could also be of particular use to policymakers, e.g., aiding in the understanding of a particular country's diaspora or providing an insight into potential partners for bilateral migration agreements.

[8] https://riks.cris.unu.edu

References

G. Abel and J. E. Cohen. Bilateral international migration flows for 200 countries. Scientific Data, 6(82):1–13, 2019. https://doi.org/10.1038/s41597-019-0089-3.

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

DEMIG. DEMIG C2C, version 1.2, limited online edition, 2015. Oxford: International Migration Institute, University of Oxford. www.migrationinstitute.org.

c. Özden, C. R. Parsons, M. Schiff, and T. L. Walmsley. Where on earth is everybody ? the evolution of global migration 1960-2000. World Bank Economic Review, 25(1):12–56, 2011. doi:10.1093/wber/lhr024.

UNDESA. United nations global migration database, 2008.

UNDESA. International migration flows to and from selected countries: the 2015 revision., 2015. United Nations, Department of Economic and Social Affairs, Population Division, POP/DB/MIG/Flow/Rev.2015.

UNDESA. International migrant stock 2020. documentation, 2020. United Nations, Department of Economic and Social Affairs, Population Division, POP/DB/MIG/Stock/Rev.2020.

A List of indicators included in the imputations

Name	Source	Country of	Measuring	Observations	panels	years
			Stocks			
S^1	TIMS	Birth		74,058	11154	1990-2020
S^2	TIMS	Nationality		9,786	1532	1990-2020
S^3	IMDB	Birth		72,215	5753	1981-2020
S^4	IMDB	Nationality		91,181	6153	1975-2020
S^5	GBM	Birth / Nat		312,154	55842	1960-2018
			Net Flows			
N^1	DEMIG	Birth	Citizens	1,488	199	1960-2011
N^2	DEMIG	Nationality	Citizens	421	17	1960-2011
N^3	DEMIG	Residents	Citizens	32,432	1592	1960-2011
N^4	DEMIG	Birth	For./Cit.	4,941	462	1960-2011
N^5	DEMIG	Residents	For./Cit.	59,365	2630	1960-2011
N^6	DEMIG	Birth	Foreingers	2,266	232	1960-2011
N^7	DEMIG	Nationality	Foreingers	44,281	2485	1960-2011
N^8	DEMIG	Residents	Foreingers	31,810	1500	1960-2011
			Inflows			
I^1	IMDB			104,412	6430	1995-2019
I^2	IMFSC	Residents	For./Cit.	79,434	5499	1980-2013
I^3	IMFSC	Residents	Citizens	31	2	1998-2013
I^4	IMFSC	Nationality	Citizens	615	33	1980-2013
I^5	IMFSC	Residents	Foreingers	11,604	572	1980-2013
I^6	IMFSC	Nationality	Foreingers	93,726	5067	1980-2013
I^7	DEMIG	Birth	Foreingers	17,462	879	1960-2011
I^8	DEMIG	Nationality	Foreingers	58,092	3426	1960-2011
I^9	DEMIG	Residents	Foreingers	44,576	1980	1960-2011
I^{10}	DEMIG	Birth	Citizens	2,543	200	1960-2011
I^{11}	DEMIG	Nationality	Citizens	464	19	1960-2011
I^{12}	DEMIG	Residents	Citizens	32,456	1559	1960-2011
I^{13}	DEMIG	Birth	For./Cit.	9,219	673	1960-2011
I^{14}	DEMIG	Residents	For./Cit.	61,664	2620	1960-2011
			Outflows			
O^1	IMDB			69,897	4457	1995-2019
O^2	IMFSC	Nationality	Citizens	540	32	1980-2013
O^3	IMFSC	Residents	Citizens	21	1	1991-2013
O^4	IMFSC	Nationality	For./Cit.	3,937	322	1990-2013
O^5	IMFSC	Residents	For./Cit.	71,793	4740	1980-2013
O^6	IMFSC	Nationality	Foreingers	70,147	4346	1980-2013
O^7	IMFSC	Residents	Foreingers	12,226	530	1980-2013
O^8	DEMIG	Birth	Citizens	1,210	188	1960-2011
O^9	DEMIG	Nationality	Citizens	508	20	1960-2011
O^{10}	DEMIG	Residents	Citizens	32,198	1749	1960-2011
O^{11}	DEMIG	Birth	For./Cit.	5,720	580	1960-2011
O^{12}	DEMIG	Residents	For./Cit.	57,648	2587	1960-2011
O^{13}	DEMIG	Birth	Foreingers	1,661	208	1960-2011
O^{14}	DEMIG	Nationality	Foreingers	43,558	2539	1960-2011
O^{15}	DEMIG	Residents	Foreingers	30,452	1487	1960-2011

UNITED NATIONS
UNIVERSITY

UNU-CRIS

Institute on Comparative Regional Integration Studies

in alliance with

The United Nations University Institute on Comparative Regional Integration Studies (UNU-CRIS) is a research and training institute of the United Nations University, a global network engaged in research and capacity development to support the universal goals of the United Nations and generate new knowledge and ideas. Based in Bruges, UNU-CRIS focuses on the provision of global and regional public goods, and on processes and consequences of intra- and inter-regional integration. The Institute aims to generate policy-relevant knowledge about new patterns of governance and cooperation, and build capacity on a global and regional level. UNU-CRIS acts as a resource for the United Nations system, with strong links to other United Nations bodies dealing with the provision and management of international and regional public goods.

The mission of UNU-CRIS is to contribute to generate policy-relevant knowledge about new forms of governance and cooperation on the regional and global level, about patterns of collective action and decision-making.

UNU-CRIS focuses on issues of imminent concern to the United Nations, such as the 2030 Development Agenda and the challenges arising from new and evolving peace, security, economic and environmental developments regionally and globally. On these issues, the Institute will develop solutions based on research on new patterns of collective action and regional and global governance. The Institute endeavours to pair academic excellence with policy-relevant research in these domains.

For more information, please visit www.cris.unu.edu

UNU-CRIS

Potterierei 72

8000 Bruges

BELGIUM