Georgios Aristotelous, Peter W. F. Smith, Jakub Bijak

Technical report: Estimation methodology

Deliverable 6.3

QuantMig has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 870299. Website: **www.quantmig.eu**

History of changes

Version	Date	Changes
1.0	13 May 2022	Issued for Consortium Review
1.1	28 June 2022	First version submitted as official deliverable to the EC

Suggested citation

Aristotelous G, Smith P. W. F, Bijak J (2022) Technical report: Estimation methodology. QuantMig Project Deliverable D6.3. Southampton: University of Southampton.

Dissemination level

PU Public

Acknowledgments

Comments made by Jon Forster and Michaela Potancokova are gratefully acknowledged. This document reflects the authors' view and the Research Executive Agency of the EC are not responsible for any use that may be made of the information it contains.

Cover photo: iStockphoto.com/Guenter Guni

Technical report: Estimation methodology

QuantMig Deliverable 6.3 [*]

Georgios Aristotelous[†], Peter W.F. Smith[‡] and Jakub Bijak[§]

[*] QuantMig is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 870299.

[†] Department of Social Statistics and Demography, University of Southampton, UK, email: G.Aristotelous@soton.ac.uk

[‡] Department of Social Statistics and Demography, University of Southampton, UK, email: P.W.Smith@soton.ac.uk

[§] Department of Social Statistics and Demography, University of Southampton, UK, email: J.Bijak@soton.ac.uk

Contents

1 Introduction

This document is Deliverable 6.3 (D6.3) of the project Quantifying Migration Scenarios for Better Policy (QuantMig) and reports on work undertaken as part of Task 6.3 of Work Package 6 (WP6). The project proposal of QuantMig describes the aim of the project as ' ... to produce comprehensive, multi-perspective and robust quantitative migration scenarios to support various areas of European migration policy' and proceeds to add that ' ... the scenarios will be based on a bespoke set of statistical estimates derived from a distinctive and comprehensive set of harmonised data on migration and its drivers.'

The aim of WP6 is, based on the available data, to develop a model for estimating migration flows within Europe and into and out of Europe, with uncertainty assessment, and to apply it to create a custom-made, harmonised dataset based on reconciling secondary data from different sources. Task 6.3 in particular concerns the design of that model and its application to the available data. The present document, D6.3, is a technical report, providing a detailed description of the methodology related to the estimation model, as well as some indicative results from the model.

The described model produces estimates of migration flows by origin, destination and year. We consider a closed system of countries and regions consisting of 32 European countries, North Macedonia, and 8 rest of the world (RW) regions. These 32 European countries are the 28 EU-member countries[1], and the 4 EFTA countries, namely Iceland, Liechtenstein, Norway and Switzerland. We use the abbreviation EU+ to refer to these 32 European countries. The 8 RW regions are Other Europe, North Africa, Sub-Saharan Africa, West Asia, East Asia, South-Southeast Asia, North America and Oceania and Latin America. North Macedonia was originally included in the Other Europe region but was subsequently modelled on its own as part of an externally funded consultancy work (Aristotelous et al., 2022b). Nonetheless this makes no difference to the methodology behind the model. Our considered time period is the years 2009 to 2019. Our model is based on the model developed in the Integrated Model of European Migration (IMEM) project (see Raymer et al. (2013)) and it extends the time horizon of migration estimates from 2002-2008 to 2009-2019, while additionally providing a segmentation of the flows between EU+ countries and the RW by considering specific RW regions.

This report is structured as follows. Section 2 provides some background information highlighting some of the challenges involved in the estimation of migration flows while it also introduces some of terminology and notation used in the report. Section 3 gives a detailed description of all components of the model. In Section 4, we provide some results from the model which help illustrate how the model works. Finally, in Section 5 we give some additional perspectives and commentary on the work described in this report.

[1] For our considered time period, 2009-2019, the United Kingdom was an EU-member country and it is therefore included in the 28 EU-member countries.

2 Background

2.1 Challenges in estimating migration flows

There are many difficulties inherent in estimating migration flows. For example, countries may use different definitions of migration. Now, most European countries use the United Nations' (UN's) 12-month duration of stay definition[2]. However, this was not always the case and some countries still report using different duration critera. Another issue is that some countries exclude certain subgroups, such as refugees or asylum seekers, which should be counted but are not, or fail to report some of the migrants, such as those without a legal status in their country of residence. That is to say, migration is typically undercounted. Also, countries may use different data collection systems, for example, surveys or population registers, which means that the accuracy at which migration data are collected may vary with country (Mooyaart et al., 2021).

To illustrate these issues, we present a small segment of the reported data on migration flows, for some selected countries and the North America and Oceania (NAO) region, in 2009. The data are presented in Table 1, in the form of what is called a double-entry matrix (Kelly, 1987; Poulain, 1999). Each cell corresponds to a migration flow, where the row is the origin/sending country and the column is the destination/receiving country. A given flow can potentially have two reports, one by the sending and one by the receiving country. However, these two reports will not necessarily agree.

For example, for the Italy to Spain flow, Italy reports sending 4479 migrants to Spain, whereas Spain reports receiving 10561 migrants from Italy, which is over twice as many. This highlights the issue of undercounting; in this particular example Italy undercounts emigrations. Another example, where the discrepancy is even larger, is flows from Slovakia to Italy, where Slovakia reports 62 emigrations and Italy 1089 immigrations, which is over 17 times as many. One reason for this is that Slovakia uses a permanent definition for migration (where the individual should intend to settle permanently in the destination country), whereas Italy uses the UN's 12-month duration definition. Note also that some countries do not report some or all of their flows. For example, Poland does not report any flows. Similarly, no flows are reported by any of the rest of the world regions, as seen by the example of NAO. This results in some flows only being reported once, e.g. Poland to United Kingdom (UK), and some not at all, e.g. Poland to NAO.

2.2 Terminology and notation

We use the notation EU+↔EU+ to denote flows between each of the EU+ countries. Recall that the abbreviation EU+ is used to refer to all of the 32 European countries, the 28 EU-member countries and the 4 EFTA countries. Similarly, the notation EU+↔MK will denote flows between an EU+ country

[2] Since 2009 the use of a 12-month criterion across the EU has been stipulated by the law, i.e. the Regulation (EC) No 862/2007 of the European Parliament and of the Council of 11 July 2007 on Community statistics on migration and international protection, OJ L 199, 31.7.2007, p. 23–29 (http://data.europa.eu/eli/reg/2007/862/oj).

Table 1: Double-entry matrix of reported flows for selected countries/regions for the year 2009.

		Destination					
Origin		Italy	Poland	Slovakia	Spain	UK	NAO
Italy	S	-	2253	579	4479	7762	5463
	R	-	NA	244	10561	10148	NA
Poland	S	NA	-	NA	NA	NA	NA
	R	9334	-	382	3654	35016	NA
Slovakia	S	62	30	-	21	104	112
	R	1089	NA	-	NA	NA	NA
Spain	S	8644	7063	NA	-	15093	8761
	R	2999	NA	119	-	22331	NA
UK	S	5144	27600	NA	14720	-	NA
	R	4760	NA	279	20361	-	NA
NAO	S	NA	NA	NA	NA	NA	-
	R	5493	NA	222	7558	NA	-

S = sending country's reported flow;
R = receiving country's reported flow;
NA = no reported data available.

and North Macedonia, while EU+↔RW and MK↔RW will respectively denote flows between an EU+ country and a RW region and between North Macedonia and a RW region.

Throughout this report, we respectively use the subscripts i, j and t, to index the three dimensions, origin, destination and time, by which we break down the flows. Similarly, we use the superscripts S and R to denote quantities associated with the reporting of sending and receiving country, respectively. So, for example, z^S_{ijt} denote the data reported by the sending country, for the flow from origin i to destination j, at year t. Finally, we note that in many instances thoughout this report, to simplify wording, we commonly refer to countries and regions as countries.

3 Methodology

To address the challenges illustrated above, we perform our task of estimating migration flows using a Bayesian hierarchical model. At a general level, the Bayesian approach to estimation works as follows. We have a prior distribution for the model parameters, a likelihood function, describing the joint probability of the observed data as a function of the model parameters, and a posterior distribution for the model parameters. The prior distribution can be thought of as expressing one's beliefs regarding the model parameters, before seeing the data, while the likelihood function represents the information from the observed data. The posterior distribution combines the information from the prior distribution and the likelihood and represents our beliefs about the model parameters after we have observed the data, and it is the basis of our estimation.

Bayesian models have many appealing features, more specific to our setting being that they have the ability to correct for the inadequacies in the available data, reconcile the differences between reports of the same flow, and provide estimation for flows that are completely missing, all while providing coherent measures of uncertainty. Our Bayesian model is based on the IMEM model (see Raymer et al. (2013)), extended to the time horizon of 2009-2019 and generalized to consider specific RW regions. First, in Section 3.1, we provide an overview of the modelling framework. Subsequently, in Sections 3.2 to 3.4, we provide a detailed description of each component of the model.

3.1 Modelling framework

The full model is consisted of three submodels: the *data model*, the *measurement model*, and the *migration model*. The data model:

- Incorporates the information from the reported flow data, reported by the sending and the receiving country.
- Models (some of) the variability in the reported data, that which corresponds to Poisson variability.

The measurement model:

- Corrects for bias in the reported flow data arising from differences in the *duration* of stay criterion (e.g. 12-month criterion, permanent criterion) used by the reporting country and from the effect of *undercount* in the reporting of data.
- Accounts for differences in the *accuracy* between different data collection systems, such as registers or surveys, so that reports from countries with more accurate data collection systems are associated with less uncertainty and also carry more weight in the estimation of flows.
- Models the variability in the data that is additional to Poisson variability and so it accounts for the effect of overdispersion (which is a typical feature of migration data).

The migration model:

- Adds information from economic, demographic and geographic explanatory variables of migration (i.e. migration drivers). We commonly refer to these variables as *migration covariates*.
- Provides smoothing and helps estimate flows for which flow data are not reported.

As can been seen from the features of these three submodels, our estimation procedure utilizes and combines the following three sources of information: flow data reported by sending and receiving country, the measurement features of a country with respect to its reporting of flow data, and migration covariate information. The model could also incorporate additional pertinent information via the prior distribution of its parameters, a point we return to in Section 3.8.

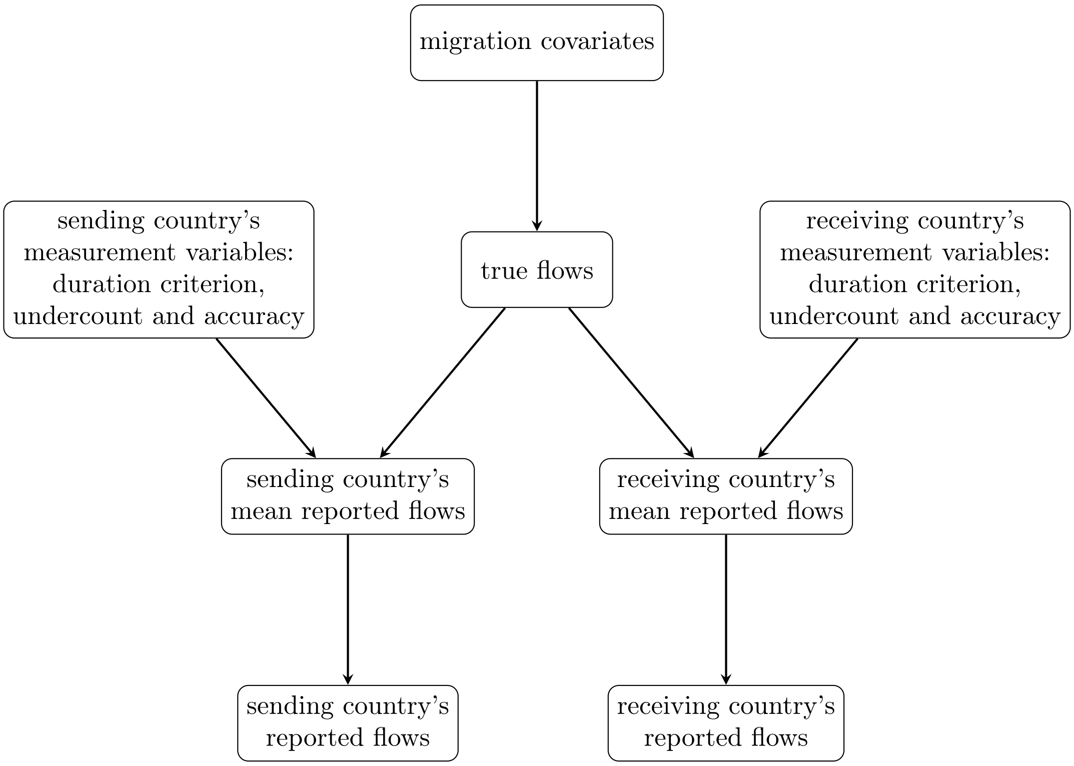

migration covariates
sending country's measurement variables: duration criterion, undercount and accuracy
true flows
receiving country's measurement variables: duration criterion, undercount and accuracy
sending country's mean reported flows
receiving country's mean reported flows
sending country's reported flows
receiving country's reported flows

Figure 1: Conceptual modelling framework (after Raymer et al. (2013)).

To illustrate the general idea of the model, we use a graphical representation, presented in Figure 1. We assume that a true unobserved flow follows a model of migration, which relates the true flow to the migration covariates. In turn, we assume that the two reported flows, one for the sending and one for the receiving country, are on average perturbed versions of the true flow, i.e. that the sending and receiving countries' mean reported flows are perturbed versions of the true flow. This pertubation induces both bias and variance, with the amount of each being determined by the measurement variables of the reporting country: the duration criterion used and the extent of uncercount determine the bias; while the accuracy of the data collection system determines the variance. Lastly, we assume that the reported flows are (unbiased) Poisson variations of the corresponding mean reported flows.

3.2 Data model

Consider a flow from country i to country j at year t. As previously mentioned, we generally may have two reports on that flow, one by the sending country and one by the receiving country. We denote these two reported flows as z^S_{ijt} and z^R_{ijt}, respectively. As highlighted in Section 2.2, the notational conventions of using the subscripts i, j and t to respectively denote sending country, receiving country and year, and of using the superscripts S and R for quantities associated with the reporting of sending and receiving country, respectively, are conventions we maintain throughout this report. Following Raymer et al. (2013), we assume that z^S_{ijt} and z^R_{ijt} are realizations of Poisson random variables with corresponding means μ_{ijt}^S and μ_{ijt}^R. That is, we assume that:

$$z_{ijt}^S \sim \text{Pois}(\mu_{ijt}^S) \tag{3.1}$$

$$z_{ijt}^R \sim \text{Pois}(\mu_{ijt}^R). \tag{3.2}$$

The migration flow data are sourced from the Eurostat database[3], which relies on the annual Joint Questionnaire on Migration Statistics collected from all national statistical agencies in the European Union. This questionnaire is coordinated by Eurostat, and is sent out on behalf of the Council of Europe, the United Nations Statistical Division, the United Nations Economic Commission for Europe, and the International Labour Organization. We sourced data regarding the flows among 33 countries, the 32 EU+ countries and North Macedonia, and regarding flows from (to) these countries to (from) the 8 RW regions, for the years 2009 to 2019. After being sourced, the data underwent a process of cleaning before being input into the model. Both the sourcing and cleaning processes are described in Aristotelous et al. (2020). We note that flows involving a RW region are calculated by summing over the flows of each country in the region. For example, if i is Italy and j is Latin America, z_{ijt}^S, the flow from Italy to Latin America, reported by Italy, is calculated by summing the flows that Italy reports sending to each country in the Latin Ametica region.

3.3 Measurement model

We now proceed to describe the measurement model which is a modified version of the measurement model used in Raymer et al. (2013). We provide the specification of the model for the case of EU+↔EU+ flows and explain how this is modified for EU+↔MK and EU+↔RW flows further below in Section 3.3.4. The model for EU+↔EU+ flows is as follows:

$$\log \mu_{ijt}^S = \log y_{ijt} + \delta_{g(i)} + \log \lambda_{f(i)}^S + \omega_i + \varepsilon_{ijt}^S, \tag{3.3}$$

$$\log \mu_{ijt}^R = \log y_{ijt} + \delta_{g(j)} + \log \lambda_{f(j)}^R + \omega_j + \varepsilon_{ijt}^R, \tag{3.4}$$

where $\varepsilon_{ijt}^S \sim N(0, \tau_{h(i)}^S)$ and $\varepsilon_{ijt}^R \sim N(0, \tau_{h(j)}^R)$, with $N(\mu, \tau)$ denoting a normal distribution with mean μ and precision (invserse variance) τ, throughout this report. As for the data model, we have two equations, (3.3) and (3.4), the first modelling the measurement features associated with the reporting of the sending country and the latter with that of the receiving country. In the above equations, and throughout this report, y_{ijt} denotes the true flow of migration from country i to country j at year t, the main focus of our estimation. The δ parameters correct for any differences in the duration of stay criterion whereas the λ and ω parameters correct for the effect of undercount. The ε terms are error terms accounting for the different levels of accuracy related to the reporting of countries.

[3] `https://ec.europa.eu/eurostat/estat-navtree-portlet-prod/BulkDownloadListing` (files: `migr_imm5prv` and `migr_emi3nxt`)

Functions $g(k)$, $f(k)$ and $h(k)$ respectively indicate the group of country k with respect to duration criterion, undercount and accuracy. To perfom this grouping we used the extensive metadata information provided as part of the work undertaken in Mooyaart et al. (2021), combined with information we extracted by our own investigations of the migration flow data. Table 2 lists the measurement variable groupings for all countries. Below we provide details of these groups as well as on how the δ, λ, ω and ε parameters are specified.

Table 2: Country groupings for the measurement model variables.

country	accuracy	duration	undercount
Austria	good	12-month	low
Belgium	good	12-month	excellent
Bulgaria	low	12-month	high
Croatia	low	12-month	high
Cyprus	-	-	-
Czechia	-	-	-
Denmark	excellent	12-month	excellent
Estonia	low	12-month	low
Finland	excellent	12-month	low
France	good	12-month	low
Germany	-	-	-
Greece	-	-	-
Hungary	-	-	-
Iceland	excellent	12-month	low
Ireland	low	12-month	low
Italy	good	12-month	high
Latvia	low	12-month	high
Liechtenstein	good	12-month	low
Lithuania	good	12-month	low
Luxembourg	-	-	-
Malta	-	-	-
Netherlands	excellent	12-month	excellent
Norway	excellent	12-month	low
Poland	low	12-month	high
Portugal	-	-	-
Romania	low	12-month	high
Slovakia	low	permanent	high
Slovenia	good	12-month	low
Spain	good	12-month	low
Sweden	excellent	12-month	low
Switzerland	excellent	12-month	excellent
United Kingdom	low	12-month	low
North Macedonia	low	12-month	high

accuracy: excellent=excellent registers, good=other good registers, low=less reliable registers or surveys;
undercount: excellent=none to very low undercount, low=low undercount, high=high undercount;
the '-' entries correspond to countries that do not report any bilateral flow data;
the RW regions are not listed since we do not have any data being reported by them.

Equations (3.3) and (3.4) capture the idea that the reported flows are distorted versions of the true flows, the distortion coming from the effect of the measurement features (duration criterion, undercount and accuracy), and in linear scale they can loosely be interpreted as:

$$\text{mean reported flow} = \text{true flow} \times \text{duration} \times \text{undercount} \times \text{error}. \tag{3.5}$$

3.3.1 Duration

As can be seen in Table 2, with the only exception of Slovakia, all other countries are assumed to report data to Eurostat using the UN's 12-month duration of stay criterion, for years 2009-2019. Therefore, the groups for duration are:

$$g(k) = \begin{cases} \text{12-m} & \text{if country } k \text{ uses the 12-month duration criterion} \\ \text{perm} & \text{if country } k \text{ uses the permanent duration criterion,} \end{cases} \tag{3.6}$$

and the duration parameters are specified as $\delta_{\text{12-m}} = 0$ and $\delta_{\text{perm}} = -d$, where d is an auxiliary parameter such that $d > 0$. To gain an understanding of how the parameter for permanent durarion criterion δ_{perm} is interpreted we can consider equation (3.5) ignoring the effect of the other measurement variables:

$$\text{mean reported flow} = \text{true flow} \times \exp(\delta_{\text{perm}}). \tag{3.7}$$

Since $\exp(\delta_{\text{perm}}) \in (0, 1)$ it can be interpreted as the proportion by which one multiplies a true flow, obeying the UN's 12-month duration criterion, to obtain the corresponding mean reported flow, under a permanent duration criterion, or, equivalently, $\exp(-\delta_{\text{perm}}) > 1$ can be interpreted as the factor by which one multiplies a flow, reported under a permanent duration criterion, in order to harmonize it to the UN's 12-month criterion. Notice that this specification of δ_{perm} constrains a permanent duration flow to be smaller than the corresponding 12-month duration flow, because fewer migrants will meet the permanent criterion compared to the 12-month one.

3.3.2 Undercount

Assuming that a country reports migration with the UN's 12-month criterion, the extent of its total undercounting of migration may be attributed to two factors. The first is the failure to register (for immigration) or deregister (for emigration) some of the migrants and the second is some subpopulations not being covered by the reporting country, something that is commonly referred to as lack of coverage. In our model we make no attempt to seperately model or estimate the amount of undercount associated with each of these factors since there is no information in the data to do so. Instead, we directly model and estimate the total amount of undercount associated with the reporting of each country.

We assume that the total undercount associated with the reporting of a country is a result of what we call a group undercount, a level of undercount which is common among countries belonging to the same undercount group, and a country-specific undercount, an undercount that is additional to the group undercount and may be different for each country. The group undercount effects are captured by the λ parameters and the country-specific undercount effects by the ω parameters, the precise manner we now proceed to describe. Explanation of why we decompose the total undercount into group and country-specific undercount is postponed until the end of Section 3.3.2.2.

3.3.2.1 Group undercount As can be seen in Table 2 we consider three undercount groups, namely excellent, low and high. These groups respectively represent the cases that a country is assumed to report data with none to very little undercount, low undercount and high undercount. Consequently, we specify the undercount group function $f(k)$ as:

$$f(k) = \begin{cases} E & \text{if country } k \text{ is in the excellent undercount group} \\ L & \text{if country } k \text{ is in the low undercount group} \\ H & \text{if country } k \text{ is in the high undercount group.} \end{cases} \tag{3.8}$$

As seen from equations (3.3) and (3.4), we allow different group undercount parameters for the receiving and sending data case and thus we have in total six group undercount parameters, λ_E^R, λ_L^R and λ_H^R, being the excellent, low and high group undercount parameters for the receiving case and λ_E^S, λ_L^S and λ_H^S the corresponding ones for the sending case. All group undercount parameters take values in $(0, 1)$ with higher values of λ implying less undercount and with the value of 1 meaning that there is no undercount.

Considering equation (3.5), and ignoring the effect of the other measurement variables, we can see that a group undercount parameter λ (we momentarily drop the subscrips and superscrips to simplify notation) is such that:

$$\text{mean reported flow} = \text{true flow} \times \lambda, \tag{3.9}$$

and so λ can be interpreted as the proportion by which one multiplies a true flow to get the corresponding mean reported flow. For example, a value of $\lambda = 0.8$ would imply that the reported flow is on average 0.8 times the value of the corresponding true flow.

The group undercount parameters are specified via a set of auxilliary parameters, the p parameters, as follows:

$$\begin{aligned} \lambda_E^R &= 1 \\ \lambda_L^R &= p_{EL}\lambda_E^R \\ \lambda_H^R &= p_{EL}p_{LH}\lambda_E^R \\ \lambda_E^S &= p_E^{RS}\lambda_E^R \\ \lambda_L^S &= p_L^{RS}p_{EL}\lambda_E^R \\ \lambda_H^S &= p_H^{RS}p_{EL}p_{LH}\lambda_E^R. \end{aligned} \tag{3.10}$$

As can be seen, λ_E^R is fixed at 1. This assumption is made to achieve identification of the rest of the group undercount parameters and it implies that there is no group undercount for countries in the excellent undercount group, for the case of reporting immigration data. We note that this does not mean that these countries have no undercount at all, since they may still have country-specific undercounts a point we return to in Section 3.3.2.2. The p parameters all take values in $(0,1)$ and act as multiplying proportion factors by which λ_E^R is related to all other group undercount parameters. More precisely, p_{EL} is the proportion by which λ_E^R is multiplied to give λ_E^R. In turn, $p_{EL}p_{LH}$ is the proportion by which λ_E^R is multiplied to give λ_H^R, or, equivalently, p_{LH} is the proportion by which λ_L^R is multiplied to give λ_H^R. The group parameters for the sending data case are acquired in a similar way as for the receiving case, only that one additionally multiplies by p_E^{RS}, p_L^{RS} and p_H^{RS} for excellent, low and high undercount groups, respectively.

Notice that, the above specification of the λ parameters imposes the constrains:

$$\begin{aligned}
&\lambda_E^R > \lambda_L^R > \lambda_H^R \\
&\lambda_E^S > \lambda_L^S > \lambda_H^S \\
&\lambda_E^R > \lambda_E^S,\ \lambda_L^R > \lambda_L^S,\ \lambda_H^R > \lambda_H^S.
\end{aligned} \tag{3.11}$$

These constrains are easy to justify in the sense that it is natural to assume that, for both receiving and sending data, the extent of undercount will be the highest in the high undercount and the lowest in the excellent undercount group. Similarly, it is reasonable to assume that, for any given undercount group, excellent, low or high, the extent of undercount will be higher in the sending data case since there is usually much less incentive for migrants to deregister from a country they leave from, compared to registering in a country they arrive to. From an inference standpoint, specifying the λ parameters via the p parameters, as above, is particularly important since it makes it possible to identify and estimate all λ parameters using essentially only information from the data, without needing to use informative prior distributions, a point we return to in Section 3.8.1.

To choose the undercount group of each country we rank all countries with respect to their extent of undercount, by considering all possible pairwise comparisons on data reportings of the same flow, so that countries which systematically report a lower number of migrants are assigned to a higher undercount group. The way we do this is by using a Bradley-Terry (BT) model (Bradley and Terry, 1952). Broadly speaking, the BT model assumes that in a 'contest' between any two 'players', say player i and player j, the odds that i 'beats' j are a_i/a_j, where a_i and a_j are positive-valued parameters, which can be though of as representing the 'ability' of i and j, respectively, where $a_i > a_j$ means that player i has higher ability than player j.

In our context, the 'contest' is the reporting of a given flow, from country i to country j, at a time t. The 'players' are the sending country i and the receiving country j, and we consider that i 'beats' j in the case that $z_{ijt}^S > z_{ijt}^R$, that is the case that the flow data count reported by the sending country i is greater than that reported by the receiving country j. Under this formulation, the 'ability' parameters of countries i and j, a_i and a_j, are representative of their extent of undercount, in the sense that if $a_i > a_j$ then country i is less likely to undercount compared to j. In this model, we also include a sending/receiving effect, denoted as γ, to account for the fact that undercounting of emigration is generally higher than that of immigration. The model can be expressed in a logistic regression form:

$$\text{logit}(P(z^S_{ijt} > z^R_{ijt})) = \log a_i - \log a_j - \gamma, \tag{3.12}$$

where $\text{logit}(\pi) = \log(\frac{\pi}{1-\pi})$, $\pi \in (0, 1)$, and by assuming independence of all contests (over i, j and t) one can estimate its parameters via maximum likelihood estimation, using standard software for generalized linear models. We conduct this estimation in the statistical programming languange R Core Team (2020), as described in Turner and Firth (2012). The estimated country ability parameters are ranked to produce a ranked list of countries with respect to their extent of undercount. We then use this list and accordingly split the countries into the three undercount groups, excellent, low and high, as presented in Table 2.

3.3.2.2 Country-specific undercount As already mentioned, in addition to the group undercount, a country k is assumed to have another source of undercount that is specific to k, captured by the parameter ω_k. For the three countries, which the BT model ranks as first within each of the three undercount groups, namely Switzerland in the excellent undercount group, Ireland in the low undercount group and Croatia in the high undercount group, ω_k is fixed at 0. This means that these three countries are assumed not to have any additional source of undercount besides that of their group. This assumption is natural from a modelling standpoint in the sense that these three countries are ranked as having the lowest undercount in their respective undercount group, and therefore further undercount effects should only be applied to the rest of the countries within their group and not to them. From an inference standpoint, this assumption allows us to identify and estimate the ω_k parameters for the rest of the countries.

For any other country k, besides Switzerland, Ireland and Croatia, the country-specific undercount ω_k is specified as:

$$\omega_k = -\log(1 + e^{-\kappa_k}), \tag{3.13}$$

where κ_k are country-specific random effects, $\kappa_k \sim N(\mu_\kappa, \tau_\kappa)$. To highlight how a country-specific undercount acts in the model, alongside the group-undercount, we momentarily drop any subscripts and superscripts to simplify notation and consider equation (3.5), ignoring the effect of the other measurement variables:

$$\text{mean reported flow} = \text{true flow} \times \lambda \times \exp(\omega). \tag{3.14}$$

For the case of Switzerland, Ireland or Croatia, we have that $\exp(\omega) = 1$ and therefore there is no country-specific undercount effect as discussed above. For the case of any other country, $\exp(\omega) \in (0, 1)$ and thus $\exp(\omega)$ is a proportion, an additional one to the group undercount proportion λ, by which one multiplies a true flow to get the corresponding mean reported flow. For instance, if $\lambda = 0.8$ and $\exp(\omega) = 0.6$, the total undercount would be equal to $0.8 \times 0.6 = 0.48$, meaning that the reported flow would be on average 0.48 times the value of the corresponding true flow.

One might wonder why we choose to model the total undercount using group undercounts in conjuction with country-specific undercounts, as opposed to solely using country-specific undercounts. The reason for this is the following. For the cases that a country reports very little data, the undercount group structure allows the model to apply to the country in question the amount of undercount which is associated with what is assumed to be its undercount group. This would not be possible in the absense of an undercount group structure. Another thing to note is that we assume that the country-specific undercount effects are the same when measuring emigration and immigration, as can be seen by equations (3.3) and (3.4). This assumption allows borrowing of strength between these two sources of data and helps us estimate the ω_k parameters. Finally, we note that the only case for which no undercount is assumed is the case of Switzerland, for the reporting of immigration, being the only case for which both the group and the country specific undercounts are equal to 1 on the linear scale.

3.3.3 Accuracy

Regarding the accuracy of data collection, we use three groups, namely excellent, good and low, that is we specify the accuracy group function $h(k)$ as:

$$h(k) = \begin{cases} E & \text{if country } k \text{ is in the excellent accuracy group} \\ G & \text{if country } k \text{ is in the good accuracy group} \\ L & \text{if country } k \text{ is in the low accuracy group.} \end{cases} \tag{3.15}$$

Excellent accuracy refers to excellent registers which we assume to be those of the Nordic countries, and those of Netherlands and Switzerland. Good accuracy refers to registers which are still considered fairly reliable but not as reliable as the high accuracy registers, for example those of Austria and Italy. Lastly, the low accuracy group refers to data collection using less reliable registers or surveys, examples being Bulgaria and the UK. The accuracy group of each country is given in Table 2. As already mentioned, the categorisation of countries with respect to accuracy was guided by the metadata information provided in Mooyaart et al. (2021).

The model parameters that capture accuracy are the τ precision parameters featuring in the ε error terms of equations (3.3) and (3.4). As evident from these equations, for a given accuracy group, we consider different precision parameters for emigration and immigration and so we have in total six precision parameters, τ_E^S, τ_G^S and τ_L^S, respectively the precision parameters for the excellent, good and low accuracy group for the case of emigration, and τ_E^R, τ_G^R and τ_L^R, the corresponding ones for immigration. As can be seen in equation (3.5), these precision parameters control the errors in the measurement model, where the higher the precision the smaller the error. In addition to that, the precision parameters play a less obvious but equally crucial role, by acting as weights in the estimation of flows, a point we discuss in more detail in Section 3.6.

3.3.4 Flows outside the EU+ system

For non EU+↔EU+ flows, that is EU+↔MK and EU+↔RW flows, EU+ countries have more rigorous registration requirements (e.g. visa requirements and residence permits) for migration compared to EU+↔EU+ flows. This is because there are typically more incentives to record migrants originating from or departing to countries outside the EU+ system. To reflect this, we consider an upgrade in the measurement features of EU+ countries when it comes to the reporting of flows outside the EU+ system. For undercount, we assume no country-specific undercounts for all countries. In addition, countries that are in the high undercount group have their group undercount parameter upgraded to that of the low undercount group. Similarly, for accuracy, countries that are in the low and good accuracy groups have their precision parameters upgraded to that of the good and excellent accuracy groups, respectively. The remaining parameters remain as for the EU+↔EU+ flows case. Specifically, the measurement model equations related to the reporting by EU+ countries of EU+↔MK and EU+↔RW flows are as follows:

$$\log \mu^S_{ijt} = \log y_{ijt} + \delta_{g(i)} + \log \lambda^S_{f^U(i)} + \varepsilon^S_{ijt}, \tag{3.16}$$

$$\log \mu^R_{ijt} = \log y_{ijt} + \delta_{g(j)} + \log \lambda^R_{f^U(j)} + \varepsilon^R_{ijt}, \tag{3.17}$$

where $\varepsilon^S_{ijt} \sim N(0, \tau^S_{h^U(i)})$, $\varepsilon^R_{ijt} \sim N(0, \tau^R_{h^U(j)})$, with f^U and h^U respectively being the undercount and precision parameter upgrade functions for EU+↔MK and EU+↔RW flows, given by:

$$f^U(k) = \begin{cases} E & \text{if country } k \text{ is in the excellent undercount group} \\ L & \text{if country } k \text{ is in the low or the high undercount group,} \end{cases} \tag{3.18}$$

and

$$h^U(k) = \begin{cases} E & \text{if country } k \text{ is in the excellent or the good accuracy group} \\ G & \text{if country } k \text{ is in the low accuracy group.} \end{cases} \tag{3.19}$$

The undercount and accuracy groups are given in Table 2. Equation (3.16) corresponds to the case that the reporting EU+ country is the sending country (i.e. in equation (3.16) i is an EU+ country and j is a RW region or North Macedonia) whereas equation (3.17) to the case that it is the receiving country (i.e. in equation (3.17) j is an EU+ country and i is a RW region or North Macedonia). For clarity, Table 3 collects the total undercount and precision parameters corresponding to the reporting of an EU+ country, when it is the sending and when it is the receiving country, and for each of the EU+↔EU+ flow and non EU+↔EU+ flow cases.

Table 3: Total undercount and precision parameters corresponding to the reporting of an EU+ country.

Flow	Sending country		Receiving country	
	Total undercount	Precision	Total undercount	Precision
EU+↔EU+	$\lambda^S_{f(i)}\omega_i$	$\tau^S_{h(i)}$	$\lambda^R_{f(j)}\omega_j$	$\tau^R_{h(j)}$
non EU+↔EU+	$\lambda^S_{f^U(i)}$	$\tau^S_{h^U(i)}$	$\lambda^R_{f^U(j)}$	$\tau^R_{h^U(j)}$

Note that we are upgrading parameters rather than introducing new group undercount and precision parameters for the EU+↔RW flows, since there is no information in the data to allow such new parameters to be estimated. This is because we have at most one report, that of the EU+ country, for these flows.

3.4 Migration model

Recall from Section 3.1 that the migration model relates the true flows to a set of migration covariates, incorporating economic, demographic and geographic information on migration. We consider two migration models, one for EU+↔EU+ and EU+↔MK flows and the other for EU+↔RW and MK↔RW flows. A description of the covariates included in the two models is provided after presentation of the model equations.

For EU+↔EU+ and EU+↔MK flows, the migration model is specified as follows:

$$\begin{aligned}\log y_{ijt} &= \beta_1 + \beta_2 \log P_{it} + \beta_3 \log P_{jt} + \beta_4 B_{ij} + \beta_5 \log(G_{it}/G_{jt}) + \beta_6 \log T_{ijt} \\ &\quad + \beta_7 M_{it} M^c_{jt} + \beta_8 M^c_{it} M_{jt} + \beta_9 M^c_{it} M^c_{jt} + \beta_{10} \log S_{ij} + \beta_{11} \log S_{ji} \\ &\quad + \beta_{12} L_{ij} + \beta_{13} A_{ijt} + \beta_{14} C_{ij} + \beta_{15} Y_{2,t} + \cdots + \beta_{24} Y_{11,t} + u_{ij} + \varepsilon_{ijt}.\end{aligned} \tag{3.20}$$

We refer to this migration model as M1. In the above equation, $\beta = (\beta_1, \beta_2, \ldots, \beta_{24})$ is a vector of regression parameters, $\varepsilon_{ijt} \sim N(0, \tau^{M_1})$ is an error term, and u_{ij} are i-to-j-flow-specific, constant over time random effects, $u_{ij} \sim N(v_{ij}, \tau_u)$, where $v_{ij} \sim N(0, \tau_v)$ and $v_{ji} = v_{ij}$. This random effect specification serves two purposes. First, for a given flow i to j, it induces correlation and smoothing across time, that is among the $y_{ijt_1}, y_{ijt_2} \ldots y_{ijt_{11}}$. Second, for a given i, j and t, it induces correlation between y_{ijt} and y_{jit}, capturing the idea that if a flow in one direction is larger (or smaller) than explained by the covariates, then we expect the flow in the opposite direction to exhibit similar behaviour. Both of these correlation structures allow borrowing of strength which helps estimate missing flows.

For EU+↔RW and MK↔RW flows, we consider a different migration model, referred to as M2 and specified as:

$$\begin{aligned}\log y_{ijt} &= \alpha_1 + \alpha_2 \log P_{it} + \alpha_3 \log P_{jt} + \alpha_4 \log D_{ij} + \alpha_5 \log(G_{it}/G_{jt}) \\ &\quad + \alpha_6 \log S_{ij} + \alpha_7 \log S_{ji} + \alpha_8 L_{ij} + \alpha_9 C_{ij} \\ &\quad + \alpha_{10} Y_{2,t} + \cdots + \alpha_{19} Y_{11,t} + \xi_{ij} + \epsilon_{ijt},\end{aligned} \tag{3.21}$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_{19})$ is a vector of regression parameters, $\epsilon_{ijt} \sim N(0, \tau^{M_2})$ is an error term, and ξ_{ij} are random effects such that $\xi_{ij} \sim N(\psi_{ij}, \tau_\xi)$, where $\psi_{ij} \sim N(0, \tau_\psi)$ and $\psi_{ji} = \psi_{ij}$. The random effect specification of ξ and ψ in M2 (equation (3.21)) is identical to that of u and v in M1 above (equation (3.20)) and serves the same purposes.

Both migration models M1 and M2 are adaptations of the migration models used in Raymer et al.

(2013). The covariates featuring in models M1 and M2 are:

- P_{it} and P_{jt}: The mid-year populations in sending country i and receiving country j, respectively, at year t. Featuring in: Both M1 and M2. Note: The mid-year population of a RW region is calculated by summing over the mid-year populations of each country in the region. Source: Eurostat database[4] for EU+ countries; World Bank database[5] for North Macedonia; United Nation database[6] for RW regions.

- B_{ij}: An indicator variable indicating whether sending and receiving countries, i and j, share a common border; $B_{ij} = 1$ if yes, $B_{ij} = 0$ if no. Featuring in: Only M1. Source: Mayer and Zignago (2011).

- G_{it} and G_{jt}: The Gross National Income (GNI) per capita in sending country i and receiving country j, respectively, at year t. Featuring in: Both M1 and M2. Note: The GNI of a RW region is calculated by taking a population weighted average over the GNI of each country in the region. Source: World Bank database[7].

- T_{ijt}: The international trade in goods between sending and receiving countries, i and j, reported as imports, at year t. Featuring in: Only M1. Source: Eurostat database[8] for EU+ countries; United Nations Commodity Statistics database[9] for North Macedonia.

- M_{it}, M_{jt}, M^c_{it} and M^c_{jt}: Indicator variables indicating whether sending country i and receiving country j are in the EU+ system, at year t; $M_{it} = 1$ if yes, $M_{it} = 0$ if no; $M_{jt} = 1$ if yes, $M_{jt} = 0$ if no; $M^c_{it} = 0$ if yes, $M^c_{it} = 1$ if no; $M^c_{jt} = 0$ if yes, $M^c_{jt} = 1$ if no. Featuring in: Only M1. Note: At year t, $M_{it}M^c_{jt} = 1$ if i is in the EU+ system but j is not, $M^c_{it}M_{jt} = 1$ if i is not in the EU+ system but j is, and $M^c_{it}M^c_{jt} = 1$ if neither i nor j are in the EU+ system, with the case of both i and j are in the EU+ system, being the reference case. Source: Barker (2021).

- S_{ij} and S_{ji}: Stocks of individuals born in the sending country i and living in the receiving country j (pull factors) and stocks of individuals born in the receiving country j and living in the sending country i (push factors), respectively, at the year 2017. Featuring in: Both M1 and M2. Note: The stock of a RW region is calculated by summing over the stocks of each country in the region. Source: World Bank database[10].

- L_{ij}: A common languange index measuring the commonality between the languanges of sending and receiving countries, i and j; L_{ij} takes values in $[0, 1]$ where the closer the value is to 1 the higher the commonality. Featuring in: Both M1 and M2. Note: A common languange index involving

[4] https://ec.europa.eu/eurostat/estat-navtree-portlet-prod/BulkDownloadListing (file: migr_pop3ctb)

[5] https://data.worldbank.org/indicator/SP.POP.TOTL

[6] https://population.un.org/wpp/Download/Standard/CSV/ (file: WPP2019_TotalPopulationBySex)

[7] https://data.worldbank.org/indicator/NY.GNP.PCAP.CD

[8] https://ec.europa.eu/eurostat/web/international-trade-in-goods/data/database (EU/EFTA trade by SITC)

[9] https://comtrade.un.org/data

[10] https://www.worldbank.org/en/topic/migrationremittancesdiasporaissues/brief/migration-remittances-data (Bilateral Migration Matrix 2017)

a RW region is calculated by taking a population weighted average over the common languange indeces of each country in the region. Source: Melitz and Toubal (2014).

- A_{ijt}: An indicator variable indicating whether migrants from the sending country i can take up any employment in the receiving country j, under the same conditions as those that apply to nationals of the receving country, at year t; $A_{ijt} = 1$ if yes, $A_{ijt} = 0$ if no. Featuring in: Only M1. Source: Barker (2021).

- C_{ij}: An indicator variable indicating whether sending and receiving countries, i and j, have ever had a colonial link; $C_{ij} = 1$ if yes, $C_{ij} = 0$ if no. Featuring in: Both M1 and M2. Note: A colonial link indicator involving a RW region is calculated by taking a population weighted average over the colonial link indicators of each country in the region and so in such cases C_{ij} takes values in $[0, 1]$ where the closer the value is to 1 the higher the colonial link. Source: Mayer and Zignago (2011).

- D_{ij}: The distance between sending and receiving countries, i and j, a population weighted average between the distances of the 25 most populated cities of country i and j. Featuring in: Only M2. Note: A distance involving a RW region is calculated by taking a population weighted average over the distances of each country in the region. Source: Mayer and Zignago (2011).

- $Y_{k,t}$, $k = 2, 3, \dots 11$: Indicator variables indicating the year t; For each $k = 2, 3, \dots 11$, $Y_{k,t} = 1$ if $t = k$, $Y_{k,t} = 0$ if $t \neq k$. Featuring in: Both M1 and M2. Note: Reference year is 2009.

The population, trade, stock and distance covariates were divided by their mean. To the stock covariates we added one to remove zero entries.

3.5 Partly-covered flows

As already mentioned in Section 3.2, reported flows involving a RW region are calculated by summing over the flows of each country in the region. There are cases though where not all flows of the countries in a region are reported. These cases occur in the reporting of Spain and the United Kingdom, from and to some of the RW regions. For instance, for the Latin America to Spain flow, Spain reports the number of migrants it receives from some of the countries in the region, but not for all. We refer to such flows as *partly-covered flows*, as the reported data for these flows only cover a part of the region. For modelling purposes, one approach, and arguably the simplest one, would be to treat such flows as missing, a consequence of the fact that they are sums for which some of the addends are missing. However, such an approach would discard valuable information in the reported data, especially for the cases where large parts of the region are covered. For example, for the Latin America to Spain flow, the aggregated population of countries for which migration data are reported, that is the population of the covered part of the Latin America region, corresponds to a proportion of 92% of the total population of the region.

Evidently, treating such entries as missing seems like a suboptimal solution.

We propose an alternative approach for modelling partly-covered flows, which is as follows. Consider a EU+$\leftrightarrow$RW flow, from i to j, at year t, such that it is partly-covered. As thoughout this report, the true unknown total flow (corresponding to the whole of the region) from i to j, at t, is denoted by y_{ijt}. For exposition purposes we assume that the RW region is the origin i and the EU+ country is the destination j but we note that the case that the RW region is the destination j and the EU+ country is the origin i is handled in the same manner, simply by exchanging j and i and receiving data R with sending data S. We split the total flow into two subflows, one corresponding to the covered part of the region, denoted as y_{ijt}^{cov}, and one corresponding to the non-covered part, denoted as $y_{ijt}^{\text{non-cov}}$, so that $y_{ijt} = y_{ijt}^{\text{cov}} + y_{ijt}^{\text{non-cov}}$. We then model y_{ijt}^{cov} and $y_{ijt}^{\text{non-cov}}$ in the same way as we model any other EU+$\leftrightarrow$RW flow, that is by using a data model, a measurement model and a migration model.

For the covered flow, these three model equations respectively are:

$$z_{ijt}^{R,\text{cov}} \sim \text{Pois}(\mu_{ijt}^{R,\text{cov}}), \tag{3.22}$$

$$\log \mu_{ijt}^{R,\text{cov}} = \log y_{ijt}^{\text{cov}} + \delta_{g(j)} + \log \lambda_{q(j)}^{R} + \varepsilon_{ijt}^{R}, \tag{3.23}$$

and

$$\begin{aligned}\log y_{ijt}^{\text{cov}} = {} & \alpha_1 + \alpha_2 \log P_{it}^{\text{cov}} + \alpha_3 \log P_{jt} + \alpha_4 \log D_{ij}^{\text{cov}} + \alpha_5 \log(G_{it}^{\text{cov}}/G_{jt}) \\ & + \alpha_6 \log S_{ij}^{\text{cov}} + \alpha_7 \log S_{ji}^{\text{cov}} + \alpha_8 L_{ij}^{\text{cov}} + \alpha_9 C_{ij}^{\text{cov}} \\ & + \alpha_{10} Y_{2,t} + \cdots + \alpha_{19} Y_{11,t} + \xi_{ij}^{\text{cov}} + \epsilon_{ijt}^{\text{cov}},\end{aligned} \tag{3.24}$$

where $z_{ijt}^{R,\text{cov}}$ are the reported data for the covered part of the region, $\mu_{ijt}^{R,\text{cov}}$ is the corresponding Poisson mean and $\epsilon_{ijt}^{\text{cov}} \sim N(0, \tau^{M_2})$. For the non-covered flow we only need a migration model equation since that flow does not correspond to any reported data, as we explain further in Section 3.6. The equation is:

$$\begin{aligned}\log y_{ijt}^{\text{non-cov}} = {} & \alpha_1 + \alpha_2 \log P_{it}^{\text{non-cov}} + \alpha_3 \log P_{jt} + \alpha_4 \log D_{ij}^{\text{non-cov}} + \alpha_5 \log(G_{it}^{\text{non-cov}}/G_{jt}) \\ & + \alpha_6 \log S_{ij}^{\text{non-cov}} + \alpha_7 \log S_{ji}^{\text{non-cov}} + \alpha_8 L_{ij}^{\text{non-cov}} + \alpha_9 C_{ij}^{\text{non-cov}} \\ & + \alpha_{10} Y_{2,t} + \cdots + \alpha_{19} Y_{11,t} + \xi_{ij}^{\text{non-cov}} + \epsilon_{ijt}^{\text{non-cov}},\end{aligned} \tag{3.25}$$

where $\epsilon_{ijt}^{\text{non-cov}} \sim N(0, \tau^{M_2})$. The cov and non-cov supercripts, added in the notation of the covariates in equations (3.24) and (3.25), are to indicate covariate data corresponding to the covered and non-covered part of the region respectively, so that for example P_{it}^{cov} is the population of the covered part of region i at time t and, accordingly, $P_{it}^{\text{non-cov}}$ that of the non-covered. The random effect terms in the two equations, ξ_{ij}^{cov} and $\xi_{ij}^{\text{non-cov}}$, are specified as in M2 (equation (3.21)), that is as $\xi_{ij}^{\text{cov}} \sim N(\psi_{ij}^{\text{cov}}, \tau_{\xi})$, $\psi_{ij}^{\text{cov}} \sim N(0, \tau_{\psi})$, $\psi_{ji}^{\text{cov}} = \psi_{ij}^{\text{cov}}$ and $\xi_{ij}^{\text{non-cov}} \sim N(\psi_{ij}^{\text{non-cov}}, \tau_{\xi})$, $\psi_{ij}^{\text{non-cov}} \sim N(0, \tau_{\psi})$, $\psi_{ji}^{\text{non-cov}} = \psi_{ij}^{\text{non-cov}}$. The remaining parameters featuring in the above set of equations are as previously specified (for the δ,

λ and ε parameters see Section 3.3 and for the α parameters see Section 3.4).

3.6 How flows are determined by the model

In the case of missing flow data, the corresponding data and measurement model contributions can be analytically marginalized out from the posterior distribution of the model. This marginalizaiton is relatively straighforard to perform and it is essentially only utilizing the fact that probability density (mass) functions integrate (sum) to 1. The marginal posterior distribution, resulting after the marginalization, is the same as the posterior distribution of a model for which for the cases of missing data, the corresponding data model and measurement model equations are not specified in the first place. What this essentially means is that for example, for a flow such that z^S_{ijt} is reported and z^R_{ijt} is missing, we only need to specify a migration model, and data and measurement models for the sending data case, since for the receiving data case it makes no difference if the data and measurement models are first specified and subsequently marginalized out from the posterior distribution, or, if they remain unspecified. Similarly, if a flow is such that both z^S_{ijt} and z^R_{ijt} are missing, then the data and measurement model contributions for both sending and receiving data can be marginilized out of the posterior distribution, or, equivalently, only the migration model needs to be specified.

The above are very useful in explaining how flows are estimated by the model. At most, there are three contributing sources providing information for a true flow y_{ijt}:

- The data reported by the sending country (data model) taking into account the measurement features of the sending country (measurement model).
- The data reported by the receiving country (data model) taking into account the measurement features of the receiving country (measurement model).
- The migration model.

The migration model contribution is always present whereas the other two contributing factors, relating to sending and receiving data respectively, are only present when the corresponding data are reported. In the presence of all three sources of information, an intuitive way of thinking how the model works is that it first corrects for any sources of bias in the sending and receiving country reported data, and it then combines the information from the bias-corrected sending data, the bias-corrected receiving data and the migration model, to produce an estimation for the flow.

All these can be made more precise by looking at the form of the distribution of $\log(y_{ijt})$, given the other parameters and data, that is the full conditional distribution of $\log(y_{ijt})$. Specifically, the full conditional distribution of $\log(y_{ijt})$ is $N(B/A, A)$ where B and A are such that:

$$B = \begin{cases} \eta_S m_S + \eta_R m_R + \eta_M m_M & \text{if both sending and receiving data are reported} \\ \eta_S m_S + \eta_M m_M & \text{if only sending data are reported} \\ \eta_R m_R + \eta_M m_M & \text{if only receiving data are reported} \\ \eta_M m_M & \text{if neither sending nor receiving data are reported,} \end{cases} \tag{3.26}$$

and

$$A = \begin{cases} \eta_S + \eta_R + \eta_M & \text{if both sending and receiving data are reported} \\ \eta_S + \eta_M & \text{if only sending data are reported} \\ \eta_R + \eta_M & \text{if only receiving data are reported} \\ \eta_M & \text{if neither sending nor receiving data are reported.} \end{cases} \tag{3.27}$$

In the above equations, m_S, m_R and m_M can loosely be thought of as central 'estimates' of $\log(y_{ijt})$, respectively corresponding to the measurement model for sending data, the measurement model for receiving data and the migration model, in the sense that, m_S, m_R and m_M are the quantities that you obtain if you solve (ignoring the error terms) these three model equations with respect to $\log(y_{ijt})$. Similarly, η_S, η_R and η_M, respectively are the precisions of the error terms in these three equations. For example, for an EU+↔EU+ flow, $m_S = \log \mu^S_{ijt} - \delta_{g(i)} - \log \lambda^S_{f(i)} - \omega_i$ (from equation (3.3)), $m_R = \log \mu^R_{ijt} - \delta_{g(j)} - \log \lambda^R_{f(j)} - \omega_j$ (from equation (3.4)), $m_M = \beta_1 + \beta_2 \log P_{it} + \cdots + \beta_{24} Y_{11} + u_{ij}$ (from equation (3.20)), $\eta_S = \tau^S_{h(i)}$ (from equation (3.3)), $\eta_R = \tau^R_{h(j)}$ (from equation (3.4)) , and $\eta_M = \tau^{M_1}$ (from equation (3.20)).

The forms of the mean B/A and the precision A of the full conditional distribution of $\log(y_{ijt})$, presented in equations (3.26) and (3.27), reveal which of the three sources of information (sending data, receiving data and migration model information) are contributing into the estimation of y_{ijt} under each pattern of observed data. They also provide a precise explanation of how this information is combined, as well as of how the extent of the information is controlled by the precision parameters. For example, for the case that both sending and receiving data are reported, all three sources of information are contributing into the estimation. The mean B/A is a weighted average of the three central 'estimates', m_S, m_R and m_M, with the corresponding weights being the precisions η_S, η_R and η_M. The precision A is equal to the sum of the three precisions, that is $A = \eta_S + \eta_R + \eta_M$. For the case that one data source is missing, say for instance the receiving country data are missing, we can see that contributing information is only coming from the sending data and the migration model, with B/A being a weighted average of m_S and m_M, where once again the weights are the corresponding precisions η_S and η_M. The precision A is a sum of the precisions of the two contributing sources of information, that is $A = \eta_S + \eta_M$. For the case that neither sending nor receiving data are available, $B/A = m_M$ and $A = \eta_M$. This illustrates how in the case that no data are reported for a given flow, the model relies on information from the migration model to estimate that flow. A final thing to observe from the form of the full conditional distribution of $\log(y_{ijt})$ is how the precision A reduces as data sources become missing, which is a quantification of the idea that the uncertainty in the estimation of flows is generally higher when less data are available.

Visual illustration of the points made in this section are provided in Section 4.

Lastly, it is worth noting that besides facilitating the conditions for explaining how flows are estimated by the model, another direct gain from working with the marginalized posterior distribution of the model described above is that it helps with the inference procedure since it makes the target parameter space smaller and therefore easier for our Markov chain Monte Carlo (MCMC) algorithm to explore. For example, all Poisson mean parameters corresponding to missing data are marginalized out and are not part of the MCMC updating scheme. Such parameter reduction techniques find use in many different Bayesian inference contexts (see e.g. Neal and Roberts (2005)).

3.7 A constraint on the precision parameters

Recall from Section 3.3.3 that τ_E^S, τ_G^S and τ_L^S, respectively are the precision parameters for the excellent, good and low accuracy group for emigration, and τ_E^R, τ_G^R and τ_L^R, the corresponding ones for immigration. Recall also from Section 3.4 that τ^{M_1} is the precision parameter of the migration model for EU+↔EU+ and EU+↔MK flows, while τ^{M_2} that of the migration model for EU+↔RW and MK↔RW flows. We impose the following constraint on the precision parameters:

$$\tau_E^S, \tau_G^S, \tau_E^R, \tau_G^R > \tau^{M_1}, \tau^{M_2}. \tag{3.28}$$

This constraint relates to the role of the precision parameters as weights in the estimation of flows, described in detail in Section 3.6 above, and is to ensure that data reported by countries considered to be of excellent or good accuracy carry more weight in the estimation of flows compared to the migration model. One way to think about the migration model is as a prior distirbution for flows, that is a prior belief about the flows before observing the data, and so in this sense it is meaningful to desire that the model should significantly update that belief in the presence of reliable data.

The above constraint serves also a second purpose, which is to allow identification of the precision parameters. More precisely, while it is possible to estimate the total variance in the measurement and migration models, there is no information in the data to distinguish how much of this variance should be attributed to each of these models. By using the above constraint this becomes possible. We note that it is also possible to achieve identification of the precision parameters by using informative prior distributions for some of the precision parameters. This approach, followed in Raymer et al. (2013), was something we initially considered and to this end, as part of work undertaken in Keilman and Aristotelous (2020), we elicited expert-based prior distributions for the measurement model precision parameters using a Delphi survey. However, we concluded that using the expert-elicited prior distirbutions in the model would be unneccesary since it transpired that the use of the constraints (3.28) is enough on its own to achieve identification of these parameters, without needing to use informative prior distributions.

3.8 Prior distribution

3.8.1 Measurement model parameters

For parameters related to the measurement model we set their prior distributions as follows. Parameter $d = -\delta_{\text{perm}}$, the auxilliary parameter controlling the effect for the permament duration criterion (see Section 3.3.1), is assigned a prior distribution as $d \sim logN\left(\log(\log(2.26)), 100\right)$ where $logN(\mu, \tau)$ denotes a log-normal distribution with parameters μ and τ such that μ is the mean and τ is the precision of the corresponding Normal distribution. This prior distribution is quite informative and is to help estimate d, since the effect for the permament duration criterion cannot be identified only from the data. The intention behind this assignment is so that $\exp(-\delta_{\text{perm}})$, the multiplying factor by which one multiplies a flow, reported under a permanent duration criterion, in order to harmonize it to the UN's 12-month criterion, has a prior distribution that is similar to its posterior distribution from Raymer et al. (2013). Specifically, the 0.05, 0.25, 0.5, 0.75 and 0.95 quantiles of $\exp(-\delta_{\text{perm}})$ under the above prior distribution are 2.00, 2.14, 2.26, 2.39 and 2.61, respectively, with the corresponding values under the posterior distribution of Raymer et al. (2013) being 1.96, 2.12, 2.26, 2.38 and 2.58. The prior belief that the multiplying factor for the permament duration criterion for the years 2009 to 2019 (the time period our model is fitted to) should be similar to what it was estimated to be for the years 2002 to 2008 (the time period that the model in Raymer et al. (2013) was fitted to) is a reasonable one, in the sense that there are no reasons to believe that this factor would change much over time.

To the p parameters, namely p_{EL}, p_{LH}, p_E^{RS}, p_L^{RS} and p_H^{RS}, the auxilliary proportion parameters via which the group undercount parameters λ are specified (see Section 3.3.2), we assign uninformative $U[0, 1]$ prior distributions, where $U[a, b]$ denotes a uniform distribution with support $[a, b]$. A point worth mentioning here is that as part of Keilman and Aristotelous (2020) we have elicited expert-based prior distributions for the group undercount parameters λ by conducting a Delphi survey. This approach, also followed in Raymer et al. (2013), is based on an impression that the undercount of countries cannot be identified solely from the data and thus informative prior distributions would be needed for the undercount parameters. However, as already mentioned in Section 3.3.2, we subsequently discovered that the undercount of countries can actually be identified from the data, without the requirement of informative prior distributions, if one specifies the λ parameters via the p parameters as in equation (3.10). For these reasons we decided that it would be preferrable not to use the expert-elicited prior distributions for the λ parameters and instead we opted to model them via the p parameters and allow their estimation to be driven by the data.

Parameters μ_κ and τ_κ, the mean and variance of the country-specific random effects κ_k, which control the country-specific undercounts $\exp(\omega_k)$, $\omega_k = -\log(1 + e^{-\kappa_k})$ (see Section 3.3.2), are assigned prior distributions as $\mu_\kappa \sim N(0, 0.5)$ and $\tau_\kappa \sim G(4, 1)$, where $G(\nu, \rho)$ denotes a gamma distribution with shape parameter ν and rate parameter ρ. This assignment corresponds to the prior distribution of a country-specific undercount $\exp(\omega_k)$ being very close to a $U[0,1]$ distribution (with slightly lighter tails) and it is in this sense rather uninformative. More precisely, the 0.05, 0.25, 0.5, 0.75 and 0.95 quantiles of $\exp(\omega_k)$ under the above prior distribution are 0.07, 0.26, 0.50, 0.74 and 0.93, respectively.

As explained in Section 3.7, due to the precision constraint which we impose in inequality (3.28) the prior distributions of the precision parameters related to the measurement model can be set to be rather uninformative. To this end, τ_E^S, τ_G^S, τ_L^S, τ_E^R, τ_G^R and τ_L^R are all assigned $G(0.001, 0.001)$ prior distributions.

3.8.2 Migration model parameters

The prior distributions of migration model parameters are set as follows. For the covariate parameters $\beta = (\beta_1, \beta_1, \ldots, \beta_{24})$ of migration model M1 we assume that they are a priori independent and the assignment of the prior distribution is done marginally as $\beta_k \sim N(0, 10^{-4})$, $k = 1, 2, \ldots 24$. The same assumption is made for $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_{19})$, the covariate parameters featuring in migration model M2. That is, we assume that $\alpha_k \sim N(0, 10^{-4})$, $k = 1, 2, \ldots 19$. These assignments are rather uninformative with a $N(0, 10^{-4})$ distribution having variance equal to 10^4 and being quite dispersed.

As for the measurement model, the precision parameters of the migration models M1 and M2, τ^{M_1} and τ^{M_2}, are set to have $G(0.001, 0.001)$ prior distributions. The same prior distribution is assigned to the precisions of the random effect terms in these models, τ_u and τ_v for M1 and τ_ξ and τ_ψ for M2.

4 Results

In this section we present some indicative results from the model. The results are based on a sample of size 5000 from the posterior distribution of the model, obtained using an MCMC algorithm, with a thinning of 50 (i.e. every 50th iteration was stored), after a burn-in of 100000 iterations. The code for the algorithm was written by the first author in the statistical programming languange R Core Team (2020).

We note that our sole purpose here is simply to illustrate how the model works and not to provide or discuss full results for all parameters of the model. Full results from the model will be provided in Aristotelous et al. (2022a), which will be a database assembling all flow estimates, prepared for further use within the project and for external dissemination. In particular, we aim to provide a visual illustration of the points already made in Section 3.6, about flows being determined by contributions from the migration model and (if available) from data reported by the sending and receiving countries taking into account their measurement features, that is from the contributions from the measurement models for sending data and receiving data.

Figure 2 presents posterior medians along with bounds of uncertainty (0.05 and 0.95 quantiles) for some selected flows, against time. In all plots, we additionally present any data reported by the sending and/or the receiving country, as well as the posterior median of the migration model. In addition, in Table 4, we provide the total undercount and precision parameters related to the measurement of sending and receiving countries, as well as the precision parameter of the corresponding migration model, for each of the considered flows. For reference, posterior summaries of these parameters are given in Table 5. Before looking at the plots, it is helpful to recall that the precision parameters, besides quantifying accuracy, play a crucial role in the estimation of flows, by being the weights of the corresponding model contributions, a point we explained in detail in Section 3.6.

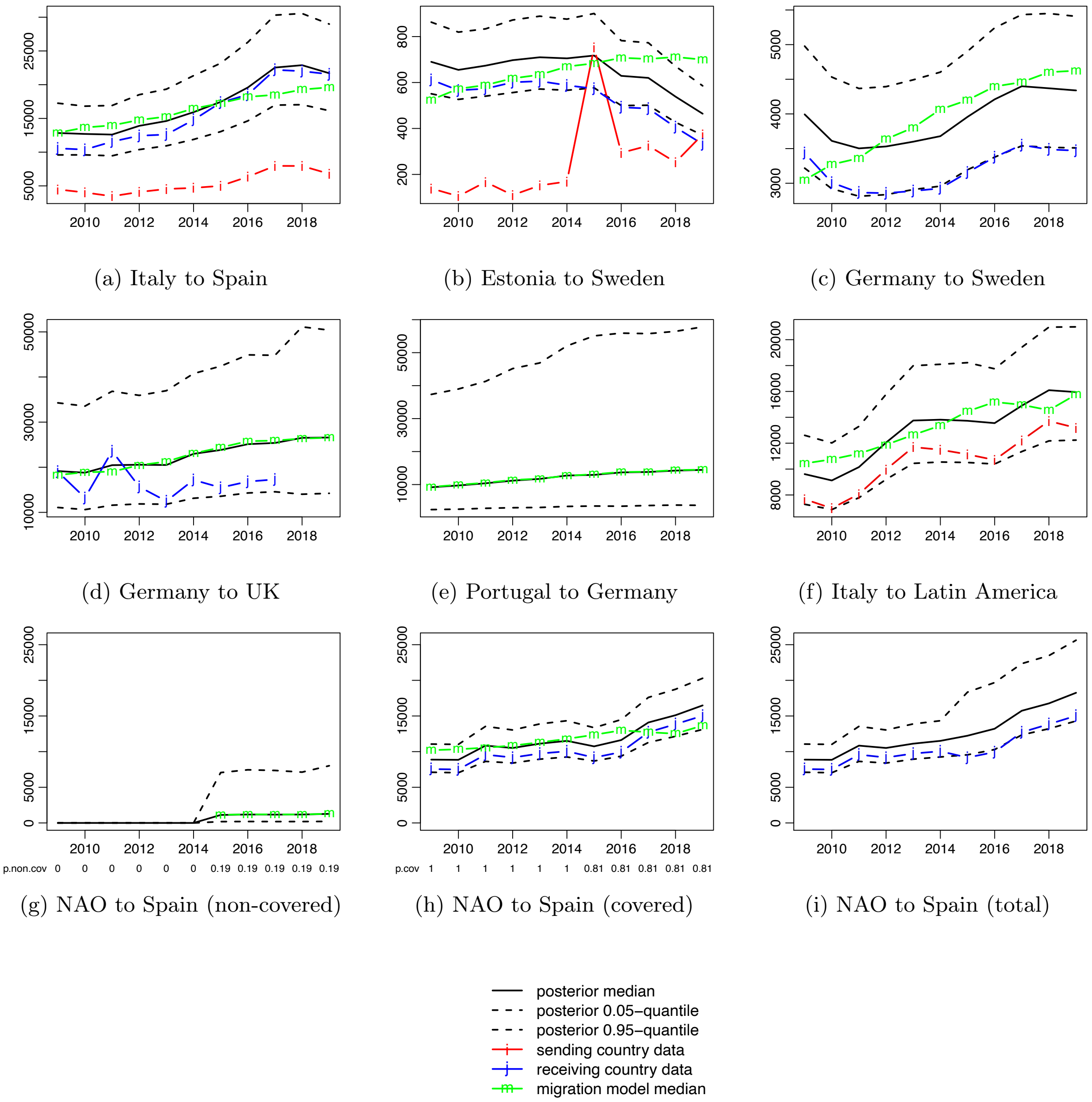
Table 4: Total undercount and precision parameters for sending and receiving countries for each of the selected flows featuring in Figure 2. Posterior summaries of the parameters are given in Table 5.

Flow	Sending country		Receiving country		Migration model
	Total undercount	Precision	Total undercount	Precision	Precision
Italy to Spain	$\lambda_H^S \omega_{IT}$	τ_G^S	$\lambda_L^R \omega_{ES}$	τ_G^R	τ^{M_1}
Estonia to Sweden	$\lambda_L^S \omega_{EE}$	τ_L^S	$\lambda_L^R \omega_{SE}$	τ_E^R	τ^{M_1}
Germany to Sweden	-	-	$\lambda_L^R \omega_{SE}$	τ_E^R	τ^{M_1}
Germany to UK	-	-	$\lambda_L^R \omega_{UK}$	τ_L^R	τ^{M_1}
Portugal to Germany	-	-	-	-	τ^{M_1}
Italy to Latin America	λ_L^S	τ_E^S	-	-	τ^{M_2}
NAO to Spain (NC)	-	-	-	-	τ^{M_2}
NAO to Spain (C)	-	-	λ_L^R	τ_E^R	τ^{M_2}

NC = non-covered
C = covered

The first example in Figure 2 is the flow from Italy to Spain. This is an example where both sending and receiving countries report data. As seen in Tables 4 and 5, Italy is estimated to be highly undercounting EU+↔EU+ flows, whereas Spain has relatively low undercount, so the model accordingly corrects for the effect of undercount in the two countries. The trend in the data reported by Spain is similar to that in the data reported by Italy, and since both countries are considered to report EU+↔EU+ flow data with good accuracy, the two reported flows together are given more weight than the migration model. (Posterior summaries for the related precision parameters are given in Table 5.)

Next, we look at the flow from Estonia to Sweden. Once again, both countries report data though in this example the accuracy of Sweden is considered to be excellent whereas Estonia's accuracy is considered to be low. As a result, the Estonian data carry very little weight compared to the Swedish data, and the estimated flow is largely determined by the Swedish data, after the appropriate correction for undercount is applied (see Tables 4 and 5). This example is a good illustration of the extent to which the reportings of the same flow might differ between two countries and on how the model deals with this



(a) Italy to Spain

(b) Estonia to Sweden

(c) Germany to Sweden

(d) Germany to UK

(e) Portugal to Germany

(f) Italy to Latin America

(g) NAO to Spain (non-covered)

(h) NAO to Spain (covered)

(i) NAO to Spain (total)

Figure 2: Plots of selected migration flows for the years 2009 to 2019.

by taking into account the differences in the measurement features of the countries.

The third and fourth flows are the flows from Germany to Sweden and from Germany to the UK. In both of these flows we only have one of the countries reporting data, the receiving country. Thus information for the estimation of these flows is only coming from two sources, from the receving country data and the migration model. Nonetheless, the way that these two sources of information are weighted by the model in each of the two flows is quite different. For the Germany to Sweden flow, the model gives

Table 5: Posterior summaries for the total undercount and precision parameters featuring in Table 4.

	Parameter	Median (95% equal-tailed credible interval)
Undercount	$\lambda_H^S \omega_{IT}$	0.33 (0.31, 0.35)
	$\lambda_L^R \omega_{ES}$	0.85 (0.81, 0.89)
	$\lambda_L^S \omega_{EE}$	0.41 (0.37, 0.45)
	$\lambda_L^R \omega_{SE}$	0.81 (0.77, 0.84)
	$\lambda_L^R \omega_{UK}$	0.76 (0.62, 0.87)
	λ_L^S	0.82 (0.78, 0.87)
	λ_L^R	0.88 (0.83, 0.93)
Accuracy	τ_E^S	27.5 (23.4, 33.2)
	τ_G^S	11.2 (10.6, 11.9)
	τ_L^S	1.4 (1.3, 1.5)
	τ_E^R	49.3 (37.6, 70.8)
	τ_G^R	11.3 (10.7, 12.2)
	τ_L^R	2.1 (1.9, 2.3)
	τ^{M_1}	11.2 (10.6, 11.8)
	τ^{M_2}	11.1 (10.5, 11.7)

much more weight to trends in the data compared to the migration model, since Sweden is considered to have excellent accuracy (see Tables 4 and 5).

On the other hand, for the Germany to UK flow, the model gives more weight on the migration model and is only slightly affected by the trends in the UK data, because the UK is considered to report EU+↔EU+ flow data with low accuracy (see Tables 4 and 5). We note though, that information coming from the UK data for the overall magnitude of the flow over the whole time period is still utilized by the model through the i-to-j-flow-specific (constant over time) random effect, featuring in the migration model. This can be appreciated by noticing that the overall (over the whole time period) magnitude of the estimated flow is about the same as that of the observed data, after accounting for the undercount associated with the reporting of the UK. A last thing to note on this flow is that the UK does not report data for the last two years of the period. For these two years the model relies only on information from the migration model while it also accounts for the additional uncertainty, as can be seen by the uncertainty bounds getting wider.

The next flow is that from Portugal to Germany. This is an example where neither sending nor receiving countries report data and thus the flow is solely determined by the migration model. Characteristically, it can be seen from the plot that the posterior median of the flow coincides with that of the migration model. Since there is only one source of information contributing to this flow, the uncertainty in the posterior distribution of the flow is relatively high. For example, when comparing the uncertainty of the

Portugal to Germany flow with that of the Italy to Spain flow, for which the magnitude is comparable and data are reported by both countries, we can see that it is much higher for the Portugal to Germany flow.

The last two flows presented in Figure 2 are flows involving RW regions, namely Italy to Latin America and North America and Oceania (NAO) to Spain. As previously mentioned, no data are ever reported by RW regions and so in these cases we only have one data source available, at most. For the Italy to Latin America flow, Italy does report emigration data covering the whole of the Latin America region, for all considered years. For this flow, there is much more weight placed on the reported data than on the migration model, as can be seen by the trend of the estimated flow following that of the data. This is because Italy is assumed to report data regarding flows outside the EU+ system with excellent accuracy (see Tables 4 and 5). What is also worth noticing here is that the undercount correction applied to Italy's emigration data is much smaller for the Italy to Latin America flow compared to the Italy to Spain flow (see Tables 4 and 5). This is because we assume that the extent of Italy's undercount will be lower when it comes to the reporting of flows outside the EU+ compared to within the EU. The Italy to Latin America flow is a good example of how this upgrade in the measurement features of EU+ countries regarding EU+↔RW flows, described in Section 3.3.4, works in practice.

Finally, the NAO to Spain flow provides an example of a partly-covered flow. As previously described in Section 3.5, partly-covered flows are EU+↔RW flows for which the EU+ country reports data covering only a part of the RW region, and we model these flows by splitting them into two subflows, one corresponding to the covered part of the region and one corresponding to the non-covered part, and then summing them to obtain the total flow. For the flow of NAO to Spain, Spain reports immigration data covering the whole of the NAO region for years 2009 to 2014, but the data reported for 2015 to 2019 correspond to covering a 82% of the population of the region whilst the remaining 0.18% remains non-covered. We first look at the non-covered flow, and specifically the years 2015 to 2019, since for the previous years all of the region is covered and thus the non-covered flow is deterministically equal to 0. By definition, the non-covered flow is one for which no data ara reported and thus it is estimated using only information from the migration model, as can be seen from the plot with the posterior median of the migration model being on top of that of the of flow. As noted in Section 3.5, the covariate data featuring in this migration model correspond to the non-covered part of the region. That is to say, the way that the model estimates a non-covered flow is identical to the way that it estimates any other flow for which no data are available, by relying on the covariate data information in its corresponding migration model. For the covered flow from NAO to Spain, again the model works in the same way as for any other flow for which one source of data are available, by combining information from the data and the migration model. In this instance, the weight of the latter source of information is much more, since Spain is considered to record these data with excellent accuracy (see Tables 4 and 5). A final thing to note here is that when looking at the total flow from NAO to Spain, one can see the additional uncertainty for years 2015 to 2019, the years for which the flow is partly-covered and there is added uncertainty coming from the non-covered part of the flow.

5 Discussion

The example flows presented in Figure 2 illustrate how the assumptions we make for the measurement features of countries are determining the estimation of flows. For example, for the Estonia to Sweden flow, our assumption that Sweden reports data with excellent accuracy and Estonia with low accuracy, led to the model giving much more weight to the Swedish data compared to the Estonian data. If we were to assume that Sweden was a low accuracy country and Estonia an excellent accuracy country, then the estimation of the flow in question would be quite different, since it would then be driven by the Estonian data. Having said that, we are quite confident in these assumptions, since, as mentioned earlier, they are based on a combination of thorough metadata information (see Mooyaart et al. (2021)), our own analysis of the data (see Section 3.3.2 for the pairwise comparisons analysis determining the undercount groups), as well as internal discussions with demography experts across the QuantMig project. It is also worth noting that the model can easily accommodate changes in the assumptions concerning measurement features of countries.

As previously mentioned, full results for all flows of the model will be provided in Aristotelous et al. (2022a). It is our hope that by the time that Aristotelous et al. (2022a) is produced, more migration data will become available. In particular, we are hoping that we will be able to obtain flow data reported by Germany, something that will produce an enhanced set of estimates.

References

Aristotelous, G., Smith, P. W. F., and Bijak, J. (2020). Database: Flows, stocks and quality for modelling. QuantMig Deliverable D5.3, University of Southampton, Southampton.

Aristotelous, G., Smith, P. W. F., and Bijak, J. (2022a). (forthcoming) Database: Migration estimates. QuantMig Deliverable D6.4, University of Southampton, Southampton.

Aristotelous, G., Smith, P. W. F., and Bijak, J. (2022b). (unpublished) Migration estimates for North Macedonia by using mirror statistics. Report for the UNFPA Office in North Macedonia, University of Southampton, Southampton.

Barker, E. R. (2021). The expansion of the European Labour Market. QuantMig Deliverable D9.5, University of Southampton, Southampton.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324.

Keilman, N. and Aristotelous, G. (2020). Expert opinion on migration data. QuantMig Deliverable D6.1, University of Oslo and University of Southampton, Oslo and Southampton.

Kelly, J. J. (1987). Improving the comparability of international migration statistics: Contributions by the Conference of European Statisticians from 1971 to date. *International Migration Review*, 21(4):1017.

Mayer, T. and Zignago, S. (2011). Notes on CEPII's distances measures: The GeoDist database. Working Papers 2011-25, CEPII.

Melitz, J. and Toubal, F. (2014). Native language, spoken language, translation and trade. *Journal of International Economics*, 93(2):351–363.

Mooyaart, J., Danko, M., Costa, R., and Boissonneault (2021). Quality assessment of European migration data. QuantMig Deliverable D6.2, Interdisciplinary Demographic Institute (NIDI-KNAW)/University of Groningen, The Hague: Netherlands.

Neal, P. and Roberts, G. (2005). A case study in non-centering for data augmentation: Stochastic epidemics. *Statistics and Computing*, 15(4):315–327.

Poulain, M. (1999). International migration within europe: Towards more complete and reliable data? Conference of European Statisticians, Statistical Office of the European Communities, Perugia, Italy.

R Core Team (2020). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Raymer, J., Wiśniowski, A., Forster, J. J., Smith, P. W. F., and Bijak, J. (2013). Integrated modeling of European migration. *Journal of the American Statistical Association*, 108(503):801–819.

Turner, H. and Firth, D. (2012). Bradley-Terry Models in R: The BradleyTerry2 Package. *Journal of Statistical Software*, 48(9).