Modeling international migration flows by integrating multiple data sources

Emanuele Del Fava[1], Arkadiusz Wiśniowski[2], and Emilio Zagheni[1]

[1]Max Planck Institute for Demographic Research, Rostock, Germany
[2]Social Statistics Department, University of Manchester, Manchester, UK

Abstract

Although understanding the drivers of migration is critical to enacting effective policies, theoretical advances in the study of migration processes have been limited by the lack of data on flows of migrants, or by the fragmented nature of these flows. In this work, we build on existing Bayesian modeling strategies to develop a statistical framework for integrating different types of data on migration flows. We offer estimates, as well as associated measures of uncertainty, for immigration, emigration, and net migration flows among 31 European countries, by combining administrative and household survey data from 2002 to 2018. ***Keywords***— Administrative data, Bayesian analysis, data integration, international migration, measurement error model.

1 Introduction

To gain a better understanding of the causes and consequences of international population migration movements, it is necessary to overcome the inherent limitations of the various data sources that each country uses to produce statistics on migration, and especially on migration flows.

These limitations include incompleteness and inconsistencies in the availability, definitions, and quality of the data. A variety of sources are used to produce statistics on migration, including population censuses, administrative records, and surveys. Although all of these sources contain information related to migration, most are not explicitly designed to accurately measure migration. For these reasons, we might expect to observe differences in the numbers reported by these sources. Since these limitations might hamper the use of single sources to investigate migration, a possible solution to the problem of obtaining statistics on migration flows between pairs of countries is to combine the information from all of the available data sources. The Bayesian statistical approach can be used to express the level of trust in the available information by means of probability distributions, to harmonize the data generated by different sources, and to provide measures of uncertainty for both model parameters and predictions.

In this paper, we propose extending a hierarchical Bayesian model, developed within the IMEM project (Raymer et al. 2013), by combining aggregated administrative data on the number of international migration events in Europe, measured at both the origin and the destination, - the only input used in the original IMEM project - with individual-level data on transitions to a new country drawn from national household surveys, such as the *Labour Force Survey* (LFS). The purpose of this data source integration is to obtain more realistic and accurate estimates of the flows among countries in the European Union (EU), the European Free Trade Association (EFTA), and the United Kingdom (UK), especially when the LFS data are able to capture migration flows that are not reflected in official statistics.

2 Data

Migration flow data can be obtained from different sources, each of which has its own characteristics and potential drawbacks.

Administrative data, which are derived from population registers, registers of foreigners or resident permits, sample surveys, specific statistical forms or other administrative sources capture residence changes when they occur or are declared. However, comparing the administrative data of different countries can be difficult for at least four reasons (Raymer et al. 2013). A first issue arises due to differences in the duration criteria used by countries to define international migrants. Although long-term international migrants are those who relocate from their country of usual residence to a different country for a minimum stay of 12 months (according to the UN definition and the 2007 EU directive), countries may use different duration thresholds for identifying migrants. A second issue is the undercounting bias, which is a consequence of individual choices that mainly occurs when people do not register when they in-migrate or, more likely, not deregister when they out-migrate. A third issue is due to coverage, a systematic bias in the data collection process that may exclude certain population segments, e.g., national return migrants or foreigners not being counted in the official immigration and emigration counts, respectively. Finally, the accuracy of the collection systems, i.e., the chance of making random mistakes in the registration or deregistration process.

Another data source to study migration within Europe are the Labour Force Surveys (LFS). These are large national household surveys that, although designed to measure labor migration, may also capture more general forms of migration, as information on immigration is collected for all members of the selected households. A first issue related to this source is the general lack of statistical precision, as migrants represent a tiny part of the total population of a country and the survey sample may not be large enough to capture them. Second, there may be issues of statistical bias related to how frequently the survey sample is updated and therefore to its capacity of capture new migrants over time. Third, if the survey includes only individual households and does not include collective accommodations, the size of certain migrant subgroups may be underestimated. Finally, in those countries in which survey participation is not mandatory, rates of non-response may be relatively high for a number of reasons, including a lack of interest in the survey, language barriers, or the fear of providing personal information, especially among undocumented migrants.

3 Methods

We use a hierarchical model to estimate the true latent bilateral migration flows from country i to country j in year t, Y_{ijt}, conditional on the definition of long-term migration as a relocation followed by a minimum stay of 12 months. To account for the inconsistencies among countries and data sources, we include a measurement error model, while the issue of data incompleteness is tackled using a gravity-type migration model to estimate the missing data. The statistical model that we propose combines and extends the methodology separately developed by Nowok & Willekens (2011), Raymer et al. (2013), and Wiśniowski (2017).

The specification of the measurement error model differs among the data sources to account for their specific characteristics and limitations. We assume that the number of observed migration events $y_{ijt}^{(k)}$ according to data source k is Poisson-distributed with parameter $\lambda_{ijt}^{(k)}$. The index k takes three possible values: (i) $k = 1$ is immigration by country of previous residence, from administrative sources; (ii) $k = 2$ is emigration by country of next residence, also from administrative sources; (iii) $k = 3$ is immigration by country of previous residence, from LFS.

The parameter $\lambda_{ijt}^{(k)}$ is modeled as follows:

$$\log \lambda_{ijt}^{(1)} = \log R_{ijt} - \mu_{j+t} \cdot d_m^{(j)} + \log \upsilon_j + \text{logit}^{-1}(\kappa_j) + \frac{\varepsilon_{ijt}^{(1)}}{\tau_j}, \quad (1)$$

$$\log \lambda_{ijt}^{(2)} = \log R_{ijt} - \mu_{j+t} \cdot d_m^{(j)} + \log \upsilon_i + \text{logit}^{-1}(\kappa_i) + \frac{\varepsilon_{ijt}^{(2)}}{\tau_i}, \quad (2)$$

$$\log \lambda_{ijt}^{(3)} = \log S_{ijt} + \log \frac{n_{jt}}{N_{jt}} + \log \upsilon_j + \text{logit}^{-1}(\kappa_j) + \frac{\varepsilon_{ijt}^{(3)}}{\tau_j}. \quad (3)$$

The parameter R_{ijt} denotes the number of *relocations*; if a person who has relocated in country j remains there for at least the minimum duration of stay in such country $d_m^{(j)}$ (which may differ between countries and sources), then the relocation can be considered a migration event, i.e., $R_{ijt} \exp(\mu_{j+t} d_m^{(j)})$, where $\mu_{j+t} = \sum_{i;i \neq j} \mu_{jit}$, μ_{ijt} being the true relocation rate. The parameter S_{ijt} denotes instead the number of people currently living in country j in year t, while living in country i one year before. Since we assume that one year is short enough to hypothesize that the individuals who relocated made at most one transition from country i to country j, we consider the information from the surveys and the administrative sources referring to the same migration flow. Given the probability of inclusion of an household in the sample, n_{jt}/N_{jt}, the number of migration events is calculated as $(S_{ijt} \cdot n_{jt})/N_{jt}$.

Finally, we adjust the observed number of migration events per data source for the different types of bias. First, to control for undercounting, we classify countries in two groups assuming either low or high undercounting, with the grouping changing by data source k. The associated parameter, υ, ranges between 0 and 1, with higher values indicating lower bias, and is given a Beta prior distribution informed with elicited expert opinion for $k = 1, 2$, or depending on whether survey participation is voluntary or mandatory ($k = 3$). Second, for the coverage bias, we use the parameter κ, converted on the linear scale to $\text{logit}^{-1}(\kappa)$, which ranges between 0 (very poor coverage) and 1 (excellent coverage). Countries are divided in two groups, one for excellent coverage (the Nordic countries for $k = 1, 2$, the countries that also include collective accommodations in $k = 3$), where $\text{logit}^{-1}(\kappa) = 1$, or with standard coverage, where κ is a country-specific and normally distributed random effect. Third, we assume that the random error in the data sources is normally distributed with mean 0 and precision (i.e., the reciprocal of the variance) equal to τ. Countries are classified in three groups, from high accuracy (Nordic countries for $k = 1, 2$, those with smaller relative standard error (RSE) for $k = 3$) to medium and to low accuracy. For each group, τ is given either a weakly informative prior distribution (for $k = 1, 2$) or an informed prior distribution based on the surveys' RSE.

Next, we use the stochastic process induced by the Bayesian approach to predict the missing data, and, in turn, derive the true latent migration flow Y_{ijt} (Raymer et al. 2013). Hence, we define a common log-normal migration gravity model for both R_{ijt} and S_{ijt} flows, namely, $\log(R_{ijt}) = \log(S_{ijt}) = \beta_0 + \sum_{p=1}^{P} x_p \beta_p$. The variables x_p are chosen based on both migration theory and empirical evidence and include, for each pair of countries, the yearly populations, the geographical distance, the ratio per year of the Gross National Incomes, the yearly international trade, the bilateral migrant stocks around the year 2000, an index of common language, and the yearly information on whether free movement of workers between countries is possible; moreover, we include B-splines for the time effect. We obtain the true relocation rate as a weighted average of the migration events from the administrative sources and the LFS:

$$\mu_{ijt} = w_R \frac{R_{ijt}}{N_{it}} + (1 - w_R) \frac{S_{ijt}}{N_{it}}, \quad (4)$$

where weights w_R are proportional to source's accuracy. Finally, the true migration flows, conditional on a minimum duration of stay of 12 months, is obtained as $Y_{ijt} = \mu_{ijt} N_{it} \exp(-\mu_{j+t})$ (Nowok & Willekens 2011). Our Bayesian model was developed in JAGS using R software.

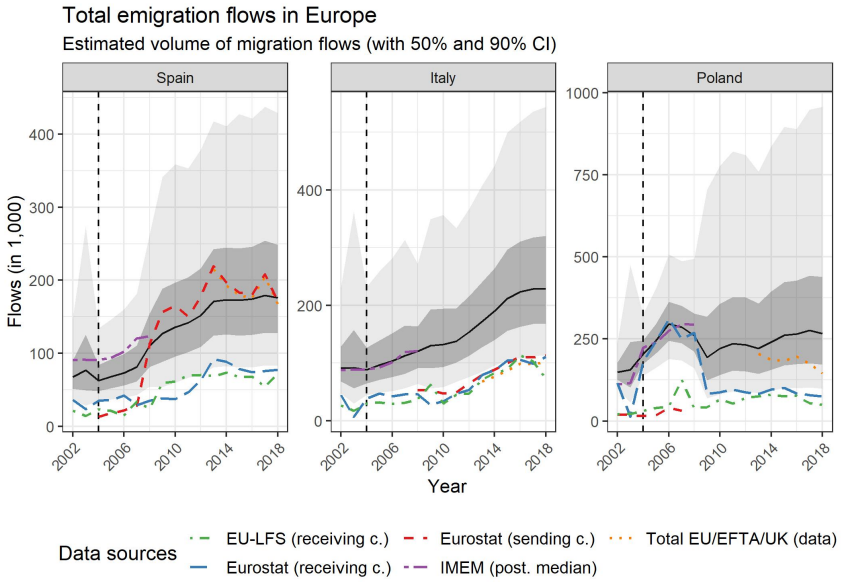

Figure 1: Posterior median (with 50% and 90% credible intervals) of the total volume of the emigration flows from Spain, Italy, and Poland to the other EU/EFTA countries from 2002 to 2018.

4 Results

As an example of our preliminary results, we show the estimated volume of the total emigration flow of three EU countries to the other EU/EFTA countries from 2002 to 2018 (Fig. 1). In addition to the posterior median of the flows and their credible intervals, we report the three used data sources and two validation data sources: (i) the flow estimated within the IMEM project (available from 2002 to 2008) and (ii) the total observed flow from Eurostat.

When complete bilateral flows from the administrative sources are available, the estimate of the true flow is mainly driven by them (as in Spain and Italy). The impact of the survey data becomes more important when administrative data are not always available (as for Poland). Our estimates are fairly consistent with those from the IMEM project until 2008.

The large uncertainty around the estimates is mainly motivated by the lack of observed data for at least one of the administrative source, as it happens with Poland and Romania, which is only partially accommodated by the migration gravity model. We plan to deal with this issue by contacting directly the national statistical offices and asking for the access to the complete bilateral time series, if available.

References

Nowok, B. & Willekens, F. (2011), 'A probabilistic framework for harmonisation of migration statistics: Harmonisation of Migration Statistics', *Population, Space and Place* **17**, 521–533.

Raymer, J., Wiśniowski, A., Forster, J. J., Smith, P. W. F. & Bijak, J. (2013), 'Integrated Modeling of European Migration', *Journal of the American Statistical Association* **108**, 801–819.

Wiśniowski, A. (2017), 'Combining Labour Force Survey data to estimate migration flows: the case of migration from Poland to the UK', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **180**, 185–202.