Modeling international migration flows by integrating multiple data sources

Emanuele Del Fava, Arkadiusz Wiśniowski, and Emilio Zagheni

October 31, 2019

Abstract

Migration has become a significant source of population change at the global level, with broad societal implications. Although understanding the drivers of migration is critical to enact effective policies, theoretical advances in the study of migration processes have been limited by lack of data on flows of migrants or by their fragmented nature. In this paper, we build on existing Bayesian modeling strategies to develop a statistical framework for integrating different types of data on migration flows. We offer estimates, and associated measures of uncertainty, for both immigration flows and emigration flows among European countries, obtained from combining administrative and household survey data from 2002 to 2015. Substantively, we document the historical impact of the EU enlargement on migration flows. Methodologically, our approach improves over the Integrated Modeling of European Union (IMEM) framework and is flexible enough to be further extended to incorporate new data sources, like social media, in order to evaluate recent migration trends within a robust statistical framework.

1 Introduction

For a better understanding of the causes and consequences of international population migration movements, migration scholars, official statisticians, and policymakers must overcome the inherent limitations of the various data sources that each country uses to produce statistics on migration, especially on flows. These data limitations include incompleteness and inconsistencies in availability, definitions, and quality (Willekens et al. 2016, Willekens 2019). The sources that are used to produce statistics on migration are various, e.g., population censuses, population registers, household and passenger surveys, registers of foreigners, special statistical forms, and each one has its characteristics and limitations (Martí & Ródenas 2007, Kupiszewska & Nowok 2008). Although all these sources contain information related to migration, they are rarely explicitly designed to accurately measure migration. Therefore we might expect differences, even quite large, in the reported

numbers. Discordance among data sources regarding the bilateral migration flows is expected both within a country and between countries. Since these limitations might hamper the use of the 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 the available data sources. In order to accomplish this task, it is possible to use a statistical modeling framework to build a *synthetic database* (Willekens 1994, 2019), which integrates migration data sources and auxiliary data to model migration flows across time and estimate flows for which incomplete information is available (Raymer et al. 2013). The Bayesian statistical approach can be used to express the trust in the available information (migration and auxiliary data) in terms of probability distributions, to harmonize the data generated by different mechanisms, and to provide measures of uncertainty for both model parameters and predictions. Moreover, the Bayesian approach enables researchers to deal with the issue of data incompleteness, by borrowing information from the available migration data and the auxiliary data to estimate the missing flows (Bijak & Bryant 2016). The *Integrated Modelling of European Migration* (IMEM) project provided an excellent example of such a synthetic database. It consisted of a Bayesian hierarchical model used to estimate bilateral migration flows among the European Union (EU) and European Free Trade Agreement (EFTA) countries and with the rest of the world (Raymer et al. 2013, Wiśniowski et al. 2016). This model included two modules. First, a *measurement error model*, which adjusted the migration flow data for effects capturing different types of inconsistencies in the data. Second, a theory-driven *migration model*, which used auxiliary information on the degree of attractiveness between countries (i.e., country- and dyad-specific demographic and socio-economic covariates) to estimate the true migration flows between those countries.

In this paper, we propose to extend the hierarchical Bayesian model developed within the IMEM project, by combining official aggregated data on international out-flow and in-flow events in Europe, disseminated by Eurostat and already used in the IMEM project, with transition data from national household surveys, such as the *EU Labour Force Survey* (LFS). Our work innovates in several directions. First, we generalize previous work on using LFS data to estimate migration flows from Poland to the UK (Wiśniowski 2017) by extending it to the whole EU/EFTA area. Second, the combination of multiple migration data sources requires the specification of source-specific measurement error models that accommodate the characteristics of each data set. Third, we modify the measurement error models introduced by Raymer et al. (2013) for the migration event data and by Wiśniowski (2017) for the LFS data, by linking them through a common underlying relocation parameter, which is, in turn, informed by the specified migration model. Finally, the use of three different data sources enables us to obtain more realistic and accurate estimates, because it increases the

evidence for each dyad-specific migration flows and, for some countries, to have at least one available data source, when the IMEM had none. We apply our model to a set formed by the 28 EU countries and two EFTA countries (Switzerland and Iceland), using time series from 2002 to 2015 for both migration event data and transition data, and we estimate the true migration flows among these countries, with the associated measures of uncertainty.

The paper is organized in the following way. In Section 2, we discuss the main data sources on international migration flows that are available in the European context (population censuses, administrative data, household surveys), as well as their issues of incompleteness and inconsistencies. The Section 3 is dedicated to the description of the Bayesian methodology used to model the data. We present the results of our modeling approach in Section 4, describing with more detail the findings for three countries taken as case studies. Finally, in Section 5, we wrap up the main points of the paper and present some ideas for future research.

2 Data

Migration data, either on the stock of migrants or on flows, can be obtained from different sources, each one with its characteristics and issues.

2.1 Census data

A first data source on migration is the population census, which typically includes information on the country of birth and nationality of each respondent.

While these data can provide information on migrant stocks, their use for migration flows is problematic, since the information on the country of previous residence is usually not collected. Even when this information is collected, e.g. when asking respondents whether they have changed their usual residence compared with one or five years prior to the census, censuses are typically carried out every ten years and movements between the censuses are neglected. People may relocate more than once in the period between two censuses.

A possible solution to tackle the lack of explicit information on the flows might use demographic accounting methods to estimate the flows that are consistent with the difference in stocks between two consecutive censuses. For this purpose, it is also essential to account for the variations in births, deaths, and migration to and from countries not included in the data (Abel 2013, Abel & Sander 2014, Azose & Raftery 2019, Abel & Cohen 2019).

2.2 Administrative data

Administrative data, usually derived from population registers (but also from other sources, such as national health or insurance databases, or residence or work permits), represent a better source of information, as they can capture residence changes when they are declared. However, comparison between countries can be hindered by several issues (Abel 2010, Raymer et al. 2013, Willekens 2019).

A first problem is the under-registration of migrants. We expect to observe this issue especially for emigration and return migration, since, while immigrants can be incentivized to register as residents in the country of destination, there are hardly any incentives for people moving abroad to deregister (and subsequently to register again, in case of return). The issue of undercounting is exacerbated in surveys (such as the International Passenger Survey in the UK) as they may suffer issues of statistical precision due to small sample size and migration being a rare event in such a survey (Martí & Ródenas 2007).

A second issue deals with differences in the duration criterion used to identify international migrants. According to the UN definition (United Nations Department of Economic and Social Affairs 1978), long-term international migrants are those who relocate from their country of usual residence to a different country for a minimum stay of twelve months, while short-term migrants are those who stay three to twelve months. However, countries may use different duration thresholds for identifying migrants, often using the intended stay duration as an approximation of the actual one. For instance, Germany and Spain have no time criterion at all; therefore, all people entering the country, not for tourism or business, are required to register. In other countries, immigrants are those who register for a stay of at least three months (e.g., Austria and Slovenia), six months (e.g., Denmark and Norway), one year (e.g., Italy and Netherlands), or permanently, such as Czechia and Poland.

However, following the EU Regulation (EC) No 826/2007, European countries started the transition to the UN definition. According to the EU Regulation, the current definition of migration requires to record the actual stay for at least one year. Therefore, each country has to submit data for reference year t to Eurostat at the end of year $t + 1$, with dissemination occurring in February $t + 2$. For instance, data referring to the year 2017 were submitted to Eurostat at the end of 2018 and made publicly available from February 2019. Moreover, since 2008, countries have been using additional sources to improve the quality of the statistics transmitted to Eurostat. These sources include health insurance registers, tax registers, 2011 census data, estimation methods based on the GDP, residency index (Maasing et al. 2017), as well as the mirror flows reported by partner countries in order to solve coverage errors (personal communication with Eurostat

representative).

The third issue with register data is related to the accuracy of the collection system. On the one hand, the five Nordic countries (Denmark, Norway, Sweden, Finland, and Iceland) have very good and harmonized population registers that routinely exchange information. On the other hand, other countries use less reliable population registers or surveys, such as the UK. Finally, some countries do not use population registers at all to measure international mobility, as they do not record the country of origin of the person who changed residence (e.g., Belgium, France, and Greece).

Notwithstanding these issues, the IMEM project has shown how Bayesian hierarchical modeling can be used to harmonize migration data from population registers and to estimate the true flows, even those for which no data are available (Raymer et al. 2013, Wiśniowski et al. 2016). This accomplishment relied on a measurement error model that included parameters accounting for the issues mentioned above, and informed either by using expert opinion or by making some *ad hoc* assumptions, necessary for parameter identification.

2.3 Household survey data

Other data sources for migration are household surveys, such as the EU Labour Force Survey (LFS). Even though these surveys aim to measure, among others, labor migration, they are also able to capture more generalized migration, as they collect such information for all people within the selected household. For each participant, the LFS questionnaire contains questions that could be in principle used to produce statistics on migrant stocks (nationality and country of birth of the participant) and on migration flows (country of residence of the participant one year before the survey). However, there are some issues with using the LFS data to estimate migration without applying any data correction (Martí & Ródenas 2007).

On the one hand, data may suffer problems of statistical precision, since migrants represent a tiny part of the total population of a country and, therefore, the sample of the survey may not be large enough to capture them.

On the other hand, there may be issues of statistical bias due to the frequency of update of the survey sample. Such a drawback is related, in turn, to at least three other issues. First, the less new participants enter the survey, the less the survey will be able to capture new migrations, as the participants who remain in the sample for several waves will not be able to provide new information on migration. Second, the survey does not include collective households, which implies underestimation of the size of some migrant subgroups, e.g., military personnel, students, members of religious communities. Third, there is the possibility of non-response, at least in those countries in which survey participation is not mandatory or

due to the language barriers.

For all these reasons, the LFS data might underestimate the migration stocks and, most of all, the migration flows, unless specific adjustments are made to account for these issues.

3 Methods

In this section, we introduce the hierarchical Bayesian statistical framework for modeling the migration flows among pairs of European countries, integrating data from multiple data sources (national administrative data and household surveys). We use a migration model to estimate the true bilateral migration flows from country i to country j in year t, Y_{ijt}^{12}, conditional on the definition of long-term migration as the relocation having a minimum duration of stay of 12 months. In order to account for differences in data measurement among countries and data sources, our statistical framework includes a measurement error model. Our statistical model extends the methodology separately developed by Nowok & Willekens (2011), Raymer et al. (2013), and Wiśniowski (2017).

In the next subsections, we present in more details both the measurement error and the migration models.

3.1 Measurement error model

Our measurement error model aims at estimating the true number of relocations (change of usual residence, with no information on the duration of stay), accounting for the inconsistencies among the data sources in terms of undercounting, population coverage, and accuracy of the data collection system. For this purpose, the model specification differs among data sources to account for their characteristics and limitations. In what follows, we present the model for each type of data source.

3.1.1 Administrative data

Data on international migration flows from administrative data sources - mainly obtained from population registers at the local or central level, but also other sources such as registers of foreigners or sample surveys, and disseminated by Eurostat - consist of counts of migration *events* from country i to country j in year t. A typical distribution for count data is the Poisson distribution, which implies that the mean is equal to the variance. However, we believe that this assumption does not hold with these data, as they present high levels of uncertainty both within country and between countries because of problems of undercounting, population coverage, and accuracy of the data collection system. Hence, to account for the large uncertainty created in the measurement process of the migration data, we

assume that the number of migration events, x_{ijt}, is distributed according to a Poisson-log-normal distribution. This assumption implies that x_{ijt} is distributed according to a Poisson distribution with parameter λ_{ijt}, representing the expected number of migration events, $x_{ijt} \sim Po(\lambda_{ijt})$. For the parameter λ_{ijt}, we specify, in turn, a log-normal distribution with mean ζ^{λ}_{ijt} and dispersion τ^{λ}_{ijt}, which is equivalent to assuming a normal distribution for $\log(\lambda_{ijt})$:

$$\log \lambda_{ijt} \sim N(\zeta^{\lambda}_{ijt}, \tau^{\lambda}_{ijt}). \tag{1}$$

For the mean ζ^{λ}_{ijt}, we specify the following model:

$$\zeta^{\lambda}_{ijt} = \log Y_{ijt} + \log \upsilon_{f^{IR}(j)} - \log(1 + e^{-\kappa_j}). \tag{2}$$

With Eq. 2, we express the idea that data originate from measuring the true number of migration events, Y_{ijt}, that these measurements are biased due to undercounting and incomplete coverage, and this bias can be accommodated with specific effects depending on $\upsilon_{f^{IR}(j)}$ and κ_j, respectively. Moreover, the precision τ^{λ}_{ijt} in Eq. 1 quantifies the uncertainty around the mean ζ^{λ}_{ijt} and accounts for the accuracy of the data source.

The modeling framework introduced in Eq. 2 for the mean ζ^{λ}_{ijt} can be rewritten as a log-linear model for the parameter λ_{ijt}. This model is a standard choice for the analysis of international migration flows, and it has been previously used to parametrize the classical gravity model for mobility as a Poisson regression model under the generalized linear modeling (GLM) framework (Cohen et al. 2008). Since we consider two sources of administrative data, one on immigration by country of previous residence and one on emigration by country of next residence, we specify two different log-linear models for λ_{ijt}, namely, one for the immigration data,

$$\log \lambda^{IR}_{ijt} = \log Y_{ijt} + \log \upsilon_{f^{IR}(j)} - \log(1 + e^{-\kappa_j}) + \frac{\varepsilon^{IR}_{ijt}}{\tau_{f^{IR}(j)}}, \tag{3}$$

and one for the emigration data,

$$\log \lambda^{ER}_{ijt} = \log Y_{ijt} + \log(\upsilon_{f^{ER}(i)}) - \log(1 + e^{-\kappa_i}) + \frac{\varepsilon^{ER}_{ijt}}{\tau_{f^{ER}(i)}}. \tag{4}$$

In the following subsections, we give an explanation for each term presented in Eq. 3 and Eq. 4.

True number of migrations. The true (or underlying) number of migrations Y_{ijt} is given by the number of relocations (or movements) from i to j conditional on a minimum continuous stay in country j for a period of t_m years (Nowok & Willekens 2011). In a context with more than two countries, we can derive Y_{ijt} using the following equation:

$$Y_{ijt} = R^{R}_{ijt} \exp(-\mu_{j+t} t^{j}_{m}) = \mu_{ijt} N_{it} \exp(-\mu_{j+t} t^{j}_{m}), \tag{5}$$

where R^R_{ijt} denotes the number of relocations obtained from the population registers or other official administrative data source, μ_{ijt} denotes the relocation rate from country i to country j in year t, N_{it} denotes the population of the country of origin, and the factor $\exp(-\mu_{j+t}t^j_m)$, where $\mu_{j+t} = -\sum_{i;i\neq j}\mu_{jit}$, accounts for the survival in country j for a minimum duration of stay equal to t^j_m (Nowok 2010).

This implies that, for modeling purposes, we can rewrite the model in Eq. 3 in the following way (a similar modification can be done for Eq. 4):

$$\log \lambda^{IR}_{ijt} = \log R^R_{ijt} - \left(\sum_{i;i\neq j}\mu_{jit}\right) t^j_m + \log \upsilon_{f^{IR}(j)} - \log(1+e^{-\kappa_j}) + \frac{\varepsilon^{IR}_{ijt}}{\tau_{f^{IR}(j)}}. \quad (6)$$

The factor t^j_m, whose units are expressed in years, can be equal to zero (no time limit, therefore each relocation is a migration), 0.25 years (three months), 0.5 years (six months), one year (or twelve months, i.e., the reference period chosen by the United Nations and the EU Regulation No. 862/2007 in their recommendation for *long-term migration*), or five years or more (permanent residence) (Nowok 2007).

Undercounting. The second term in Eq. 3 and Eq. 4, based on the parameter υ^k_f, accounts for the systematic bias due to the undercounting of the migration events in the data source k (Raymer et al. 2013). The parameter υ ranges between zero and one on the linear scale, with values closer to one indicating low undercounting and values closer to zero indicating high undercounting.

Depending on the indicators $f^{IR}(j)$ (for the model based on immigration data) and $f^{ER}(i)$ (for the model based on emigration data), we differentiate the parameter between countries assumed to have high undercounting ($f = 0$) or low undercounting ($f = 1$). Hence, the parameters υ_1 and υ_2 correspond to $\upsilon_{f^{IR}(j)=1}$ and $\upsilon_{f^{IR}(j)=0}$, respectively; similarly, the parameters υ_3 and υ_4 correspond to $\upsilon_{f^{ER}(i)=1}$ and $\upsilon_{f^{ER}(i)=0}$, respectively. The classification of countries between those with low undercounting and high undercounting is based on information elicited from expert opinion (Raymer & Wiilekens 2008, Raymer et al. 2013). For instance, we expect the undercounting to be more substantial when the administrative data depend on self-declaration from the individual who relocates. Moreover, we expect υ to be closer to zero for emigration than immigration, since there usually are more incentives, in terms of personal and social benefits, to register than to deregister.

To compute the parameters $\upsilon_{f^{IR}(j)}$ and $\upsilon_{f^{ER}(i)}$, we specified a prior distribution, which is used in the Bayesian framework to express our belief or previous information in probabilistic terms. In particular, we specify the prior as a Beta distribution, which is a continuous distribution with shape parameters α and β ranging from zero to one, exactly as the parameter υ. Considering that the prior distributions based on the elicited expert opinion

(Wiśniowski et al. 2013) and used in the IMEM model proved to be, at the end of the day, rather weakly informative, because of the large uncertainty in expert opinion (Willekens 2019), we decided to specify our own priors, based on our believes and accompanied by large uncertainty. To express our belief that the undercounting is lowest with the immigration from the countries classified as having low undercounting, we define for the parameter υ_1 a Beta prior centered around 0.8 with standard deviation equal to 0.1, namely, Beta$(12, 3)$, which implies that, according to our assumption, the parameter ranges within a 95% interval between 0.57 and 0.95. On the other hand, since we believe that the undercount is highest with the emigration from the countries classified as having high undercounting, we define for the parameter υ_4 a Beta prior centered around 0.4 with standard deviation equal to 0.1, namely, Beta$(9.2, 13.8)$, which implies a parameter ranging within a 95% interval between 0.21 and 0.60. Finally, to express our uncertainty for the cases with high undercounting in immigration data (υ_2) and low undercounting in emigration data (υ_3), we define for both groups a Beta prior centered around 0.6 with standard deviation equal to 0.1, namely, Beta$(13.8, 9.2)$, which implies a parameter ranging within a 95% interval between 0.40 and 0.79.

Coverage. The third term in Eq. 3 and Eq. 4, $-\log(1 + e^{-\kappa})$, adjusts for the population coverage of the data collection system in the country of destination (model IR) or the country of origin (model ER) (Raymer et al. 2013). This parameter is assumed to range between zero (very poor coverage) to one (optimal coverage) on the linear scale. The term depends on the parameter κ, which is a country-specific and normally distributed random effect on the log-linear scale. This implies that higher positive values of κ indicate good coverage, while higher negative values indicate poor coverage.

We distinguish between countries with "standard" coverage, for which κ is distributed according to the normal distribution $N(\nu, \eta)$, and countries with "excellent" coverage (the Nordic countries plus the Netherlands), for which the term $-\log(1+e^{-\kappa})$ is assumed to be equal to one, implying perfect coverage. For the distribution of the hyperparameters ν and η, we follow the specification chosen by Raymer et al. (2013), setting $\nu \sim N(0, 0.05)$ and $\eta \sim \Gamma(4, 1)$. This implies that, for countries with standard coverage, we give κ a weakly-informative prior distribution, since it implies that ranges *a priori* $k \in [-9, 9]$, which in turn means that $\exp(-\log(1 + e^{-\kappa_i}))$ on the linear scale ranges between zero and one.

Accuracy. Finally, the term ε/τ_{fk} is an error term, where the parameter ε follows a standard normal distribution and the precision parameter τ_{fk} accounts for the accuracy of the data collection system (Raymer et al. 2013). If the precision τ_{fk} is high, the whole error term will be smaller, indicating a higher level of accuracy. On the other hand, lower values of τ_{fk} will inflate the error term, giving evidence of a lower level of accuracy.

We use prior knowledge to group the countries based on the accuracy of

the data collection system for immigration or emigration, as the precision τ_{f^k} are considered to be group-specific. The first group, characterized by the precisions $\tau_{f^{IR}(j)=1}$ and $\tau_{f^{ER}(j)=1}$, contains the Nordic countries, who are assumed to have very high accuracy, as they exchange information on migration flows among themselves. The second group, characterized by the precisions $\tau_{f^{IR}(j)=2}$ and $\tau_{f^{ER}(j)=2}$, includes countries with reliable data collection systems. Finally, the third group, characterized by the precisions $\tau_{f^{IR}(j)=3}$ and $\tau_{f^{ER}(j)=3}$, encompasses the remaining countries, which have a less reliable collection system for migration data.

For all these τ parameters, we specify weakly-informative priors, namely, $\Gamma(0.01, 0.01)$, which assume that the precision parameters are centered around one and are spread within a very large positive range. In a similar way to the undercounting parameters, the choice of weakly-informative priors is motivated by the finding that the elicited expert opinion in the IMEM project contributed very little to the prior information (Wiśniowski et al. 2013).

3.1.2 Survey data

Differently from population registers and other administrative data, household surveys like the EU LFS provide data on transitions during a given time window, usually, one year. The LFS can provide information on stocks of migrations by nationality and country of birth, but also on immigration flows, as the survey asks participants in which country they used to live one year before the survey (Martí & Ródenas 2007). We believe that one year is short enough to make the sensible assumption that individuals who relocated made at most one transition from the country of origin to the country of destination. This *naive assumption* (Schmertmann 1999) implies that the observed transitions correspond to all the relocations done by the survey participants in the given year, and therefore the number of transitions coincide with the number of events, being thus consistent with the information provided by the administrative data sources. Similarly to what assumed for the migration data from the registers in Section 3.1.1, we assume that the transition data collected with the LFS in the country of destination (IS), k_{ijt}, follow a Poisson distribution with parameter λ^{IS}_{ijt}, which is the expected number of transitions from country i to country j occurred between year $t-1$ and year t. This parameter, in turn, follows a log-normal distribution with mean ξ^{λ}_{ijt} and precision ω^{λ}_{ijt}:

$$\log \lambda^{IS}_{ijt} \quad \sim \quad N(\xi^{\lambda}_{ijt}, \omega^{\lambda}_{ijt}) \tag{7}$$

Finally, the mean ξ^{λ}_{ijt} is given the following model depending on the true number of transitions R^{S}_{ijt} and on additional effects for the bias in measurement:

$$\xi^{\lambda}_{ijt} = \log R^{S}_{ijt} + \log \frac{n_{jt}}{Njt} + \log \upsilon_{g^{IS}(j)}. \tag{8}$$

Also Eq. 8 can be rewritten as a log-linear model for the parameter λ_{ijt}^{IS}, namely,

$$\log \lambda_{ijt}^{IS} = \log R_{ijt}^{S} + \log \frac{n_{jt}}{Njt} + \log \upsilon_{g^{IS}(j)} + \frac{\varepsilon_{ijt}^{IS}}{\omega_{g^{IS}(j)}}. \tag{9}$$

The ratio between the sample size n_{jt} and the population N_{jt} in country j is the inclusion probability of the survey, i.e., the probability of an individual to be included in the sample, and is equal to the reciprocal of the design weight. In the limit case in which $n_{jt} = N_{jt}$ and the survey is actually a census, the inclusion probability is equal to one and the expected number of transitions λ_{ijt}^{IS} is equal to the true number of transitions R_{ijt}^{S} (if we do not consider the effects of the undercounting and the accuracy of the data collection system). However, usually $n_{jt} \ll N_{jt}$, thus the inclusion probability is smaller than one and the term $\log(n_{jt}/Njt)$ is negative, implying that the measured data are a subset of the whole number of transitions.

Undercounting. The third term in Eq. 9, based on the parameter $\upsilon_{g^{IS}(j)}$, accounts for the systematic bias due to the undercounting of events, as a consequence of the survey design.

We classify as *low-undercounting* those countries for which the participation to the LFS survey is mandatory by law (10 out of 31 EU/EFTA countries), while we consider as *high-undercounting* those countries for which the participation is voluntary. In case of voluntary participation, a migrant person selected to participate in the survey might likely refuse for language barriers or lack of interest in the survey. For this reason, there is a lower chance of capturing migration flows with voluntary participation than with a mandatory one. Similar to the model for the population register data, the parameter $\upsilon_{g^{IS}(j)}$ ranges between zero and one, with values closer to one indicating low undercounting, and values closer to zero indicating high undercounting. Depending on the indicator $g^{IS}(j)$, we differentiate between countries assumed to have high undercounting (with voluntary participation, $g = 0$) and low undercounting (with mandatory participation, $g = 1$). Hence, the parameters υ_5 and υ_6 correspond to $\upsilon_{g^{IS}(j)=1}$ and $\upsilon_{g^{IS}(j)=0}$, respectively.

To be able to identify these parameters $\upsilon_{g^{IS}(j)}$, we assume that the parameter for the low undercounting is larger than the one for the high undercounting. We thus specify for the low-undercounting parameter the prior Beta$(13.3, 4.4)$, which is centered around a mean of 0.75 (with standard deviation 0.1) and varies within the 95% interval between 0.53 and 0.92. Instead, for the high-undercounting parameter, we use the prior Beta$(4.4, 13.3)$, which is centered around a mean of 0.25 (with standard deviation 0.1) and varies within the 95% interval between 0.09 and 0.4.

Accuracy. The nuisance term ε/ω_{g^k} in Eq. 9 consists of an error term ε, which follows a standard normal distribution, and a precision parameter ω_{g^k}. As for the model for the register data, the larger the precision ω_{g^k}, the

smaller the whole error terms and the higher the accuracy.

We classify in each country within two groups depending on the accuracy of the survey to capture new migrants year after year, which we make depend on how frequently each country updates the sample composition. According to Martí & Ródenas (2007), each country has a certain rate of impossible answer, which is what happens when participants remain in the survey for more than one year, preventing the possibility of including new recent migrants in the sample. We consider a first group with country having a rate of impossible answer lower than 50%, and whose precision is denoted by $\omega_{g^{IS}(j)=1}$. The countries in this group change their sample composition more often and may therefore be more accurate in capturing new transitions. On the other hand, we consider the countries with an impossible answer rate ranging between 51% and 75%, and whose precision is denoted by $\omega_{g^{IS}(j)=2}$. Since these countries change less frequently their sample composition and participants remain the panel for more than on year, we expect their accuracy in capturing new migrants to be lower than for the first group.

For both $\omega_{g^{IS}(j)=1}$ and $\omega_{g^{IS}(j)=2}$ parameters, we specify priors distributions that are weakly-informative, namely, $\Gamma(0.01, 0.01)$, which assume that the parameters to have mean one and a very large variance.

3.2 Migration model

The second component of the hierarchical Bayesian model consists of a migration model, which we use to derive the true (latent, not observed) migration flows. Although we specify our model as a gravity model of migration, we do not aim the assess the effect of a set of chosen covariates on the latent migration flows, as it would happen with a regression model. Rather, we use the stochastic process induced by the Bayesian framework to predict the missing data and, by that means, derive the true latent migration flows. The variables included in the model are chosen according to both migration theory and empirical evidence, and have been shown to provide relevant information on the socio-economic and demographic factors affecting the relative attractiveness between pairs of countries (Abel 2010, Raymer et al. 2013).

Contrarily to the measurement error models, where we specifically defined a model for each data source to account for the distinguished characteristics of data, we develop a single migration model for all data sources, reflecting the fact that the true migration flow underlying the multiple data sources is the same.

We thus assume that the true latent number of relocations from the register data, R^R_{ijt}, and the true latent number of transitions from the survey data, R^S_{ijt}, originate from two log-normal distributions with a common mean μ^Y_{ijt} (because the underlying migration flows are the same) and source-specific precisions, τ^R and τ^S (because the data sources do not capture the

migration flows in the same way). In this way, we express the idea We then define the following migration model for the common μ_{ijt}^{Y}, with i denoting the country of origin, j the country of destination, and $t = 1, \ldots, T$ the year, with the period under study being from 2002 to 2015:

$$\begin{aligned} \mu_{ijt}^{Y} &= \gamma_1 + \gamma_2 \log N_{it} + \gamma_3 \log N_{jt} + \gamma_4 \log C_{ij} + \gamma_5 \log \frac{G_{jt}}{G_{it}} + \\ &\quad \gamma_6 A_{ijt} + \gamma_7 A_{it} + \gamma_8 A_{jt} + \gamma_9 \log I_{ijt} + \gamma_{10} \log M_{ij} + \\ &\quad \gamma_{11} \log M_{ji} + \gamma_{12} \log L_{ij} + \gamma_{13} F_{ijt} + \sum_{k=14}^{28} \gamma_k E_t + u_{ij}. \end{aligned} \tag{10}$$

The model in Eq. 10 contains the following variables:

1. The population of both the country of origin, N_{it}, and the country of destination, N_{jt}. (Source: Eurostat.)

2. Indicator variable for contiguity, C_{ij} that takes value one if country i and country j share a common border, zero otherwise (Mayer & Zignago 2011).

3. The ratio of the Gross National Income (GNI) per capita in the destination country, denoted as G_{jt}, to the GNI of the country of origin, denoted as G_{it}. (Source: World Development Indicators, World Bank.)

4. Three indicators variables for EU/EFTA membership status between 2002 and 2017. The first variable, A_{ijt}, takes value one if both countries i and j were members of the EU/EFTA in year t. The second indicator, A_{it}, takes value one if the origin country i was a member of the EU/EFTA in year t. The third indicator, A_{jt}, takes value one if the destination country j was a member of the EU/EFTA in year t.

5. The international trade between the country of origin i and destination j in year t, expressed as imports in EUR, I_{ijt}. (Source: Eurostat, EU/EFTA trade by SITC.)

6. The bilateral migrant stocks by country of birth around the year 2000, based on population censuses. The variable M_{ij} stands for the migrant stock born in the country of origin i and living in the country of destination j, and is used to capture the "pull effect" of the migrant network residing in the destination. The variable M_{ji} stands for the migrant stock born in the country of destination j and living in the country of origin i, and is used to capture the "push effect" from the returning migrants in the country of origin (Özden et al. 2011).

7. An 0-1 index of Common Language (CL), denoted as L_{ij}, obtained by aggregating three indices on Common Official Language (COL), Common Native Language (CNL), and Language Proximity (LP). When

its value is closer to one, it indicates a higher degree of commonality between the languages of the origin and destination countries (Melitz & Toubal 2014).

8. An indicator variable, F_{ijt}, taking value equal to one for the year t in which there is freedom of movements for workers from country i in country j, meaning that they can take up any employment on the same conditions as the nationals. (Source: European Commission.)

9. Indicator variables for the time effect for years 2002-2016, denoted as E_t with $t = 1, \ldots, 16$, to capture the time pattern of migration flows over the years. The reference year is 2017.

10. A random effect u_{ij} to smooth the data over time by capturing the dyadic effect of each pair of bilateral migration flows between two countries. These random effects are normally distributed with mean v_{ij} and precision τ_u, $u_{ij} \sim N(v_{ij}, \tau_u)$. The hyperparameters v_{ij} are constrained to be symmetric to capture the dyadic effect, namely, $v_{ij} = v_{ji}$, and are normally distributed with mean zero and precision τ_v, $v_{ij} \sim N(0, \tau_v)$. This random effect is constant across time and thus induces correlation between the yearly estimated flows within each pair of countries and allows borrowing of strength when flow data are missing. Both the precision parameters τ_u and τ_v are given weakly-informative priors $\Gamma(0.01, 0.01)$, which have mean equal to one and very large variance.

True relocation rate. Based on the estimates of the true number of relocations from the registers, R^R_{ijt}, and the true number of relocations from the survey, R^S_{ijt}, we derive the true underlying relocation rate μ_{ijt} by taking a weighted average of the relocation rates obtained using both the administrative and the survey data:

$$\mu_{ijt} = \frac{R^R_{ijt} w_R + R^S_{ijt} w_S}{N_{it}}, \tag{11}$$

where the weights w_R and w_s, with $w_R + w_S = 1$, denote the weight of the population register data source and the survey data source, respectively. To parametrize the weights w_R and w_S, we express our uncertainty on which data source should weight more by specifying for the parameter w_R the prior Beta$(12, 12)$, which is centered around a mean of 0.5 (with standard deviation 0.1) and varies within the 95% interval between 0.30 and 0.70. The parameter w_S is then determined by subtraction.

True number of migration flows. Finally, the true number of migration flows Y^{12}_{ijt}, conditional on the minimal duration of stay of 12 months (in accordance to the recommendation of the United Nations and the EU

Regulation (EC) No 862/2007), is generated using the estimate of the true relocation rate μ_{ijt} (Nowok 2010, Nowok & Willekens 2011):

$$Y_{ijt}^{12} = \mu_{ijt} N_{it} \exp\left(-\mu_{j+t} t_m^j\right) = \mu_{ijt} N_{it} \exp\left(-\mu_{j+t}\right), \tag{12}$$

where $t_m^j = 1$ and the term $\mu_{j+} = \sum_{i;i\neq j} \mu_{jit}$.

4 Results

We developed our Bayesian model in JAGS (Plummer 2003) through R software. We computed the summary statistics from the posterior distribution of the parameters with MCMC samples of 8,000 with four chains of 30,000 iterations (10,000 iterations of burn-in and a thinning of 10, i.e., we saved each $10th$ iteration). Convergence of the monitored parameters was assessed with standard diagnostics present in the package `coda` (Plummer et al. 2006).

4.1 Model results

We ran the model on a set of 30 EU/EFTA countries, excluding Lichtenstein (LI) and Malta (MT) because of the lack of LFS data (no data at all for LI, data available only from 2009 onward for MT).

We summarize the results of the measurement error model in Table 1, where we show the quantiles from the posterior distribution of each parameter.

Regarding the coverage parameters, $\text{logit}^{-1}(\kappa)$, we notice that, in case of countries without any data from the registers (CY, FR, GR, HU, LV, PT, RO), the parameter identification did not succeed, as shown by the 95% credible interval ranging between 0.01 and 1.

As concerns the undercounting parameters for the administrative data, the posterior characteristics of the parameters of the "high-undercounting" groups, namely, λ_2 and λ_4, strictly reflect our assumption for the prior distribution, while the posterior characteristics for the "low-undercounting" parameters reflect the influence of the higher quality of the data. On the other hand, fine-tuning of the priors is still deemed necessary for the parameters for the survey data, as, on the one hand, the two parameters are quite different from what assumed with the prior, and on the other hand, it appears that there is no difference between the two groups, based on "voluntary" or "mandatory" participation to the survey.

The accuracy parameters show the highest levels of accuracy in the Nordic countries for immigration and, to a lesser extent, emigration. On the other hand, we estimated particularly low values of accuracy for the countries with data collection systems classified as little reliable. The lowest values for the precision are reported for the survey data. Also in this case,

the information on the frequency of the sample update, used to classify the countries between high and low accuracy, seems to be not associated with the accuracy, as there is no difference between the parameters ω_1^{IS} snd ω_2^{IS}.

The results for the migration models are reported in Table 2, where we show the quantiles from the posterior distribution of each coefficient and of the precision parameters. Several variables showed *credible evidence for a non-zero effect* (Matthews 2019), meaning that the 95% credible interval of their coefficients did not contain the value of zero, or, if contained, this was a quite extreme value. A positive credibly non-zero effect was found for the ratio of the GNI between the destination and the origin country, implying that a richer country is more attractive as destination country. Moreover, a migration flow from i to j was found to be more likely when the origin country was a member of the EU/EFTA or, even more strongly, when in both countries there was freedom of movement for the workers, which is an indication of the positive effect of the EU enlargement and the economic freedom on the size of the intra-EU migration flows. We also found positive credibly no-zero effects for the migrant stocks in both the sending and the receiving countries, showing some evidence of both pull and push effects of the diaspora networks (Pedersen et al. 2008). The higher the level of a common language between two countries, the more likely it was to observe a migration movement between the two. The unstructured time effects, for which the reference year was 2015, were generally credibly lower than zero, indicating, on average, an increase in the numbers of migration events over time (implied by the decrease in magnitude of the yearly coefficients). Finally, we found that the precision of the variance of the random effects u_{ij} was quite large, indicating high heterogeneity among the bilateral flows. Some bilateral flows emerge particularly, such as those between DE and PL, IT and RO, and FR and UK (Figure 3), remarking their importance at the European level.

4.2 Case studies

To better present our results on the estimated migration flows, we focus on four countries, taken as case studies, i.e., France, Poland, Sweden, and United Kingdom. For each of these countries, we show the estimated total true latent flow for the immigration (Figure 1), the emigration Figure 2, the net migration Figure 3. In the appendix, we show, for nine selected European countries, the total immigration (Figure 4), emigration (Figure 5), and net migration (Figure 6) flows, as well as the bilateral flows among all of them (Figure 7).

France does not report any migration data by country of origin or destination in its administrative data, but on the other hand, has reliable immigration data from household surveys Martí & Ródenas (2007). Polish data only reflect permanent migration and are characterized, for this reason, by

Table 1: Measurement error model: Posterior characteristics of the parameters

	2.5%	25%	50%	75%	97.5%
$\text{logit}^{-1}(\kappa_{AT})$	0.83	0.86	0.88	0.90	0.93
$\text{logit}^{-1}(\kappa_{BE})$	0.95	0.98	0.99	1.00	1.00
$\text{logit}^{-1}(\kappa_{BG})$	0.19	0.22	0.23	0.24	0.27
$\text{logit}^{-1}(\kappa_{CH})$	0.95	0.98	0.99	1.00	1.00
$\text{logit}^{-1}(\kappa_{CY})$	0.01	0.41	0.82	0.97	1.00
$\text{logit}^{-1}(\kappa_{CZ})$	0.37	0.41	0.43	0.46	0.50
$\text{logit}^{-1}(\kappa_{DE})$	0.99	1.00	1.00	1.00	1.00
$\text{logit}^{-1}(\kappa_{EE})$	0.41	0.45	0.47	0.50	0.54
$\text{logit}^{-1}(\kappa_{ES})$	0.97	0.99	1.00	1.00	1.00
$\text{logit}^{-1}(\kappa_{FR})$	0.01	0.38	0.81	0.97	1.00
$\text{logit}^{-1}(\kappa_{GR})$	0.01	0.39	0.81	0.97	1.00
$\text{logit}^{-1}(\kappa_{HR})$	0.30	0.33	0.34	0.36	0.39
$\text{logit}^{-1}(\kappa_{HU})$	0.01	0.39	0.82	0.97	1.00
$\text{logit}^{-1}(\kappa_{IE})$	0.95	0.98	0.99	1.00	1.00
$\text{logit}^{-1}(\kappa_{IT})$	0.47	0.51	0.53	0.55	0.60
$\text{logit}^{-1}(\kappa_{LT})$	0.94	0.98	0.99	1.00	1.00
$\text{logit}^{-1}(\kappa_{LU})$	0.18	0.20	0.22	0.23	0.26
$\text{logit}^{-1}(\kappa_{LV})$	0.01	0.38	0.81	0.97	1.00
$\text{logit}^{-1}(\kappa_{PL})$	0.10	0.11	0.12	0.12	0.13
$\text{logit}^{-1}(\kappa_{PT})$	0.01	0.40	0.81	0.97	1.00
$\text{logit}^{-1}(\kappa_{RO})$	0.01	0.40	0.82	0.97	1.00
$\text{logit}^{-1}(\kappa_{SI})$	0.90	0.96	0.98	1.00	1.00
$\text{logit}^{-1}(\kappa_{SK})$	0.15	0.16	0.17	0.17	0.19
$\text{logit}^{-1}(\kappa_{UK})$	0.23	0.26	0.27	0.28	0.31
λ_1	0.83	0.88	0.92	0.95	0.98
λ_2	0.54	0.58	0.61	0.65	0.69
λ_3	0.79	0.83	0.87	0.90	0.93
λ_4	0.36	0.39	0.41	0.43	0.46
λ_5	0.10	0.10	0.11	0.12	0.13
λ_6	0.10	0.11	0.11	0.12	0.13
τ_1^{IR}	18.86	23.56	26.59	30.16	38.68
τ_1^{ER}	19.21	23.76	26.75	30.25	38.34
τ_2^{IR}	11.94	12.95	13.53	14.17	15.48
τ_2^{ER}	8.06	8.59	8.90	9.21	9.86
τ_3^{IR}	0.97	1.01	1.04	1.06	1.11
τ_3^{ER}	0.84	0.88	0.90	0.92	0.96
ω_1^{IS}	0.51	0.55	0.57	0.59	0.63
ω_2^{IS}	0.57	0.59	0.61	0.62	0.66

Table 2: Migration model: Posterior characteristics of the parameters

	2.5%	25%	50%	75%	97.5%
Intercept	-0.30	-0.23	-0.14	-0.06	0.02
$\log N_{it}$	-0.02	0.04	0.09	0.13	0.41
$\log N_{jt}$	-0.03	0.02	0.09	0.14	0.41
$\log C_{ij}$	-0.27	-0.18	-0.04	0.07	0.20
$\log(G_{jt}/G_{it})$	0.73	0.77	0.79	0.81	0.85
A_{ijt}	0.02	0.07	0.10	0.14	0.18
A_{it}	0.33	0.38	0.41	0.44	0.49
A_{jt}	0.07	0.14	0.16	0.19	0.26
$\log I_{ijt}$	-0.01	0.06	0.13	0.17	0.41
$\log M_{ij}$	-0.43	-0.17	-0.13	-0.06	0.02
$\log M_{ji}$	0.21	0.35	0.40	0.51	0.66
$\log L_{ij}$	0.16	0.37	0.44	0.51	0.63
F_{ijt}	1.25	1.54	1.66	2.01	2.78
E_{2002}	0.36	0.39	0.41	0.43	0.49
E_{2003}	-0.32	-0.29	-0.27	-0.25	-0.22
E_{2004}	-0.32	-0.30	-0.28	-0.26	-0.23
E_{2005}	-0.42	-0.40	-0.38	-0.36	-0.33
E_{2006}	-0.38	-0.35	-0.33	-0.31	-0.28
E_{2007}	-0.32	-0.30	-0.28	-0.26	-0.23
E_{2008}	-0.28	-0.25	-0.23	-0.22	-0.19
E_{2009}	-0.26	-0.23	-0.22	-0.20	-0.17
E_{2010}	-0.28	-0.25	-0.23	-0.22	-0.19
E_{2011}	-0.25	-0.22	-0.20	-0.19	-0.15
E_{2012}	-0.21	-0.19	-0.17	-0.16	-0.13
E_{2013}	-0.15	-0.12	-0.11	-0.09	-0.06
E_{2014}	-0.12	-0.09	-0.08	-0.06	-0.03
τ_u	5.86	6.75	7.22	7.74	8.78
τ_v	0.04	0.05	0.06	0.07	0.10

high undercounting and low accuracy. Sweden is a Nordic country with a reliable data collection system, low undercounting, excellent coverage, and high accuracy. Finally, the UK does not derive its data from population registers, but from the International Passenger Survey, a continuously running survey administered at borders (air, see, and tunnel ports) that classify respondents as long-term migrants if they intend to stay in the country for one year or more.

All four countries show increasing patterns for both immigration and emigration flows following the EU enlargement in 2004 and the freedom of movement for workers in the EU (Figure 1 and Figure 2). However, among these selected countries, only Poland shows an increasingly negative net migration pattern, reporting all the other countries highly positive net migration trends (Figure 3).

Immigration in France increased gradually over time, starting with 2004. The same occurred with the emigration, where we noticed two changing points in the increasing pattern, one in 2004 and another in 2007, corresponding to the EU enlargement and the free movement to the labor market for a number of Eastern European countries. The net migration is positive during the whole time period, increasing on from 10,000 to 30,000 people. The main in-flows in France were from the UK, Italy, and Spain (Figure 7), all increasing over time. The relative degree of attractiveness between the UK and France appeared to be similar, considering that fairly comparable migration flows could be observed between the two countries.

Poland showed an increasing trend of immigration, with an abrupt change in 2007, mainly coming from Germany and the UK. Considering that these two countries were also the leading destinations for migrants from Poland, it is likely that these inflows to Poland mainly consisted of return migrants (Figure 7). As regards the Polish emigration, it showed two abrupt increases, one in 2004 after the entrance in the EU, and one in 2011 when Polish workers are allowed to freely enter the German labor market. Looking at the net migration patterns, we notice that Poland over time generally experienced increasingly negative net migration, with turning points in the occasion of the peaks in immigration.

Sweden also showed a surge in immigration from 2004 with a peak reached between 2007 and 2008, and a gradually increasing immigration after 2004. The net migration trend was positive and also increasing over time, with turning points in 2004 and in 2009. The migration patterns for Sweden are mainly informed by the high-quality administrative data, as the estimates follow the time series provided by Eurostat quite nicely, even though the LFS are also quite informative for the immigration patterns.

Finally, the results for the UK are similar in trends to those for Sweden, except that the numbers are quite larger. The net migration was always positive, with the increasing pattern beginning in 2004, reaching a peak in 2007 and a low in 2011. The main countries of immigration to the UK were

Poland, France, Spain, and Italy (Figure 7).

Among the other five selected countries shown in the Appendix, we noticed relevant negative net migration trends for Poland, Romania, and, to a lesser extent, for Spain and Italy. Conversely, the highest positive net migration patterns were found for Germany and the UK (Figure 6).

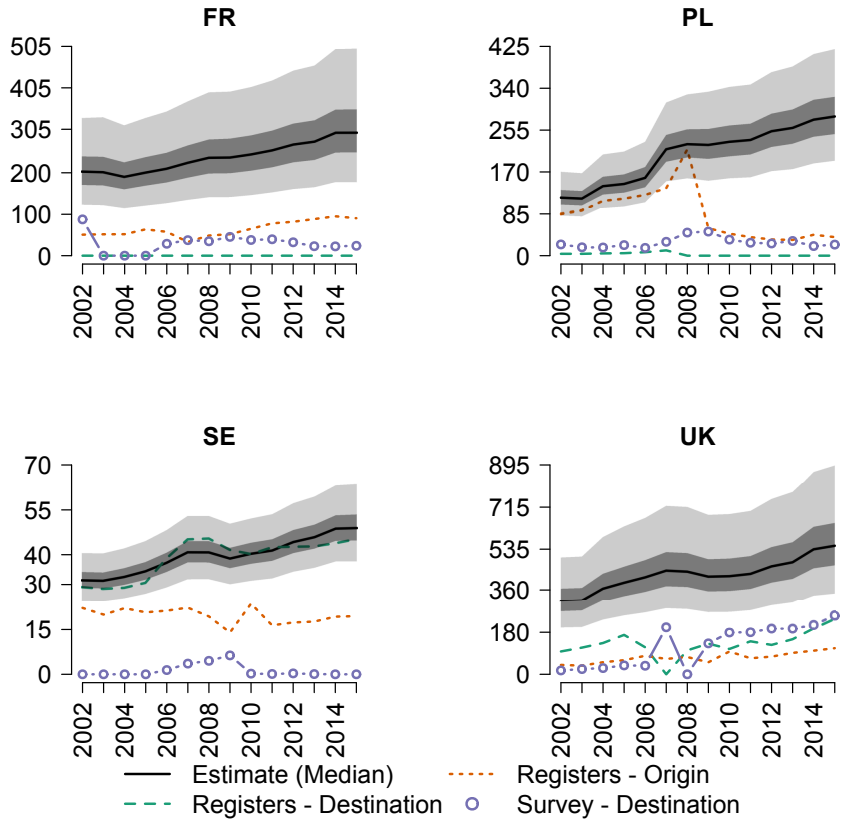

Figure 1: Immigration: Estimated total flow (per 1000) with 50% CI (dark gray) and 95% CI (light gray) for France (FR), Poland (PL), Sweden (SE), and United Kingdom (UK). In addition, total flows from immigration register data (dashed green line), from emigration register data (dotted orange lines), and from LFS data (purple circles).

5 Conclusions

We developed a hierarchical Bayesian model to estimate international migration flows based on a migration model, integrating different data sources, and adjusting for their characteristics and limitations.

First, as regards the measurement of the international migration in Europe, we showed that different data sources exist that provide information

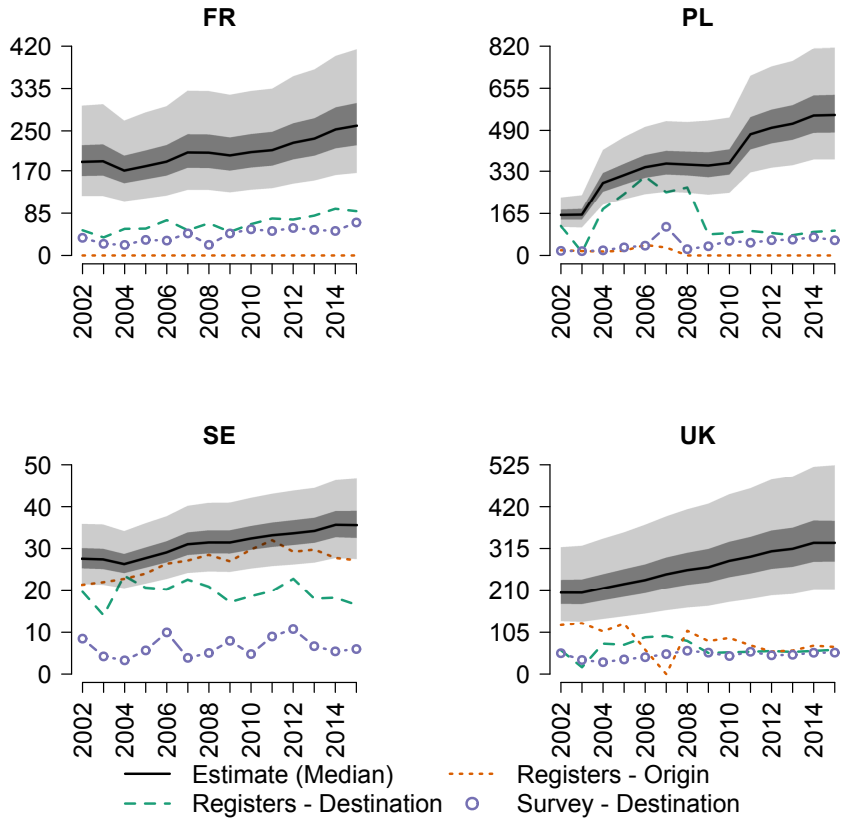

Figure 2: Emigration: Estimated total flow (per 1000) with 50% CI (dark gray) and 95% CI (light gray) for France (FR), Poland (PL), Sweden (SE), and United Kingdom (UK). In addition, total flows from immigration register data (dashed green line), from emigration register data (dotted orange lines), and from LFS data (purple circles).

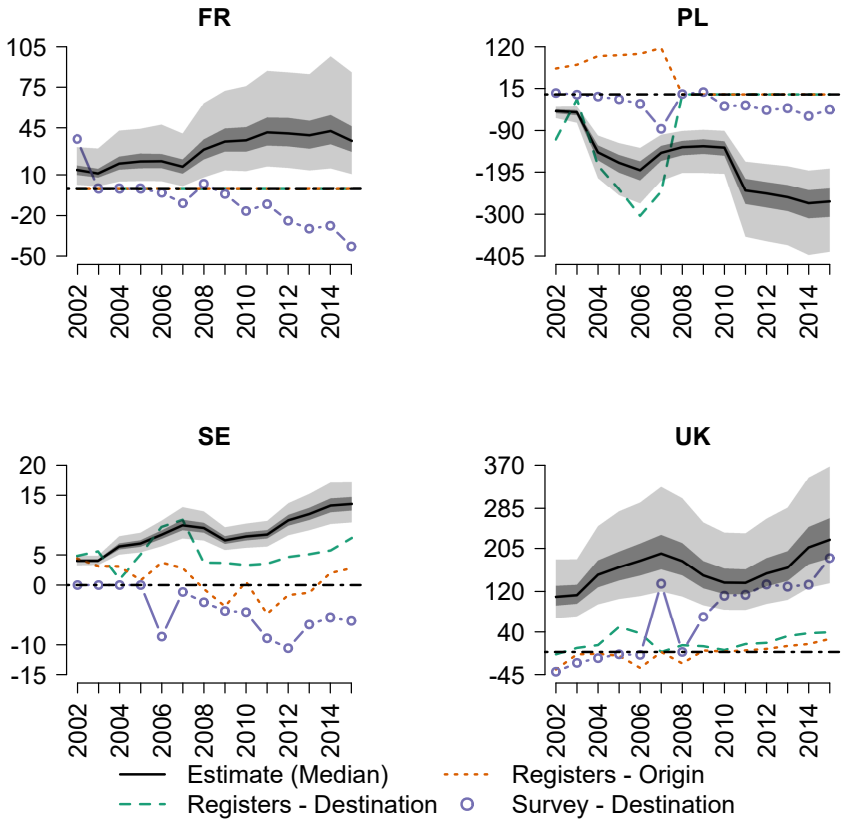

Figure 3: Net migration: Estimated total flow (per 1000) with 50% CI (dark gray) and 95% CI (light gray) for France (FR), Poland (PL), Sweden (SE), and United Kingdom (UK). In addition, total flows from immigration register data (dashed green line), from emigration register data (dotted orange lines), and from LFS data (purple circles). The horizontal black dashed-dotted line indicates the level of net migration equal to zero.

on international migration flows among European countries, namely, population registers and other administrative sources, household surveys such as the Labor Force Survey, and population census. Focusing on the first two sources (population registers and the LFS), we showed how data may suffer issues of incompleteness and inconsistency. As regards the latter, they are related to the time dimension (with some countries using different standards from the 12-month criterion for long-term migration recommended by the UN and the EU), the undercounting, the population coverage, the accuracy of the data collection system, and, for the LFS, the sampling design. Notwithstanding these issues, we showed how there is an informative gain in combining the information provided by these data sources, given that the incompleteness and the inconsistencies are properly taken into consideration.

Second, regarding the processing of data, we showed how to integrate multiple data sources in a single statistical model, developed as a hierarchical Bayesian model, accommodating also auxiliary data informing on the relative attractiveness among pairs of countries and assessments about data quality, converted into probability statements in the form of prior distributions for the parameters. Our model built upon previously published Bayesian models for administrative data and household survey data. In particular, the novelty of our model consists in the combination of the approach developed under the IMEM project (Raymer et al. 2013) with the one for the LFS data (Wiśniowski 2017), estimating from both population register and survey data the latent true relocation rate Nowok & Willekens (2011) and then using it to predict the latent true migration flows conditional on the recommended criterion of 12 months for a long-term migration event. Another aspect of innovation of our work is the extension of the time series of input data, as we use all the available administrative and LFS data from 2002 to 2015.

Third, as regards the results of our modelling framework, we found that our estimates showed the impact of the EU enlargement (2004) and the extension of the freedom of movements of workers within the EU (2007) on the intra-European migration patterns. As expected, the level of uncertainty around the estimates of the true latent migration flows highly depended on the amount of available information. Higher variability was found for France, Poland, Romania, and, to a lesser extend, Germany, as they often reported incomplete population register data, and for the UK, whose data come from surveys and not from registers. On the other hand, countries like Sweden and Netherlands reported much lower variability because of the better quality of their migration data and therefore the true latent migration flow was found to be closer to the reported data, especially those from the population registers.

As regards the future research, in the short term, we plan to perform a sensitivity analysis to assess the impact of the assumptions made to construct the priors for the parameters of interest, particularly those included

in the measurement error models. Indeed, on the one hand, there is spread criticism around the use of elicited expert opinion and other subjective beliefs for defining the prior distributions for the parameters (Willekens 2019); on the other hand, the use of weakly informative priors might lead to issues of parameter identification. It is therefore essential to assess the impact of different prior specifications on the final estimates of the true latent migration flows.

In conclusion, we believe that our Bayesian statistical framework is flexible enough to be extended with the inclusion of new data sources on migration flows or new variables for the migration model. On the one hand, we could augment data sources with data from social media, e.g., geo-referenced Twitter data (Zagheni et al. 2014) or Facebook network data (Spyratos et al. 2019), or with data from IP addresses of email service providers (Zagheni & Weber 2012). On the other hand, we could add new variables to the migration model. It would be interesting to exploire the possibility of including online search data, to be used as a proxy of migration intentions (Böhme et al. 2019), or data on income inequality and living conditions, as well as data on unemployment and social expenditure, to be used as possible push effects for migration intentions (Mayda 2010). Moreover, the model could be further extended by stratifying the migration by gender and age groups, in order to gain a better understanding of the drivers of population change and the heterogeneity within migrant groups (Wiśniowski et al. 2016).

Acknowledgements

We thank Frans Willekens, Dmitri Jdanov, Aiva Jasilioniene, and Domantas Jasilionis for their valued feedback.

References

Abel, G. J. (2010), 'Estimation of international migration flow tables in Europe: International Migration Flow Tables', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **173**(4), 797–825.

Abel, G. J. (2013), 'Estimating global migration flow tables using place of birth data', *Demographic Research* **28**(18), 505–546.

Abel, G. J. & Cohen, J. E. (2019), 'Bilateral international migration flow estimates for 200 countries', *Scientific Data* **6**(1), 1–13.

Abel, G. J. & Sander, N. (2014), 'Quantifying Global International Migration Flows', *Science* **343**(6178), 1520–1522.

Azose, J. J. & Raftery, A. E. (2019), 'Estimation of emigration, return migration, and transit migration between all pairs of countries', *Proceedings of the National Academy of Sciences* **116**(1), 116–122.

Bijak, J. & Bryant, J. (2016), 'Bayesian demography 250 years after Bayes', *Population Studies* **70**(1), 1–19.

Böhme, M. H., Gröger, A. & Stöhr, T. (2019), 'Searching for a better life: Predicting international migration with online search keywords', *Journal of Development Economics* p. 102347.

Cohen, J. E., Roig, M., Reuman, D. C. & GoGwilt, C. (2008), 'International migration beyond gravity: A statistical model for use in population projections', *Proceedings of the National Academy of Sciences* **105**(40), 15269–15274.

Kupiszewska, D. & Nowok, B. (2008), Comparability of Statistics on International Migration Flows in the European Union, *in* 'International Migration in Europe', John Wiley & Sons, Ltd, pp. 41–71.

Maasing, E., Tiit, E.-M. & Vähi, M. (2017), 'Residency index – a tool for measuring the population size', *Acta et Commentationes Universitatis Tartuensis de Mathematica* **21**(1), 129–139.

Martí, M. & Ródenas, C. (2007), 'Migration Estimation Based on the Labour Force Survey: An EU-15 Perspective', *International Migration Review* **41**(1), 101–126.

Matthews, R. A. J. (2019), 'Moving Towards the Post $p < 0.05$ Era via the Analysis of Credibility', *The American Statistician* **73**(sup1), 202–212.

Mayda, A. M. (2010), 'International migration: a panel data analysis of the determinants of bilateral flows', *Journal of Population Economics* **23**(4), 1249–1274.

Mayer, T. & Zignago, S. (2011), 'Notes on CEPII's Distances Measures: The GeoDist Database', *SSRN Electronic Journal* .

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

Nowok, B. (2007), Evolution of International Migration Statistics in Selected Central European Countries, *in* J. Raymer & F. Willekens, eds, 'International Migration in Europe', John Wiley & Sons, Ltd, Chichester, UK, pp. 73–87.

Nowok, B. (2010), Harmonization by simulation: a contribution to comparable international migration statistics in Europe, PhD thesis, University of Groningen, Groningen, Netherlands. OCLC: 955134815.

Nowok, B. & Willekens, F. (2011), 'A probabilistic framework for harmonisation of migration statistics: Harmonisation of Migration Statistics', *Population, Space and Place* **17**(5), 521–533.

Özden, c., Parsons, C. R., Schiff, M. & Walmsley, T. L. (2011), 'Where on Earth is Everybody? The Evolution of Global Bilateral Migration 1960–2000', *The World Bank Economic Review* **25**(1), 12–56.

Pedersen, P. J., Pytlikova, M. & Smith, N. (2008), 'Selection and network effects—Migration flows into OECD countries 1990–2000', *European Economic Review* **52**(7), 1160–1186.

Plummer, M. (2003), JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling, *in* 'Proceedings of the 3rd international workshop on distributed statistical computing', Vol. 1243, Vienna, Austria, p. 10.

Plummer, M., Best, N., Cowles, K. & Vines, K. (2006), 'Coda: Convergence diagnosis and output analysis for mcmc', *R News* **6**(1), 7–11.

Raymer, J. & Wiilekens, F. (2008), *International Migration in Europe: Data, Models and Estimates*, John Wiley & Sons.

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**(503), 801–819.

Schmertmann, C. P. (1999), 'Estimating Multistate Transition Hazards from Last-Move Data', *Journal of the American Statistical Association* **94**(445), 53–63.

Spyratos, S., Vespe, M., Natale, F., Weber, I., Zagheni, E. & Rango, M. (2019), 'Quantifying international human mobility patterns using Facebook Network data', *PLOS ONE* **14**(10), e0224134.

United Nations Department of Economic and Social Affairs (1978), Statistics of international migration, *in* 'United Nations Demographic Yearbook 1977', UN, pp. 3–16.

Willekens, F. (1994), 'Monitoring international migration flows in Europe', *European Journal of Population/Revue européenne de Démographie* **10**(1), 1–42.

Willekens, F. (2019), 'Evidence-Based Monitoring of International Migration Flows in Europe', *Journal of Official Statistics* **35**(1), 231–277.

Willekens, F., Massey, D., Raymer, J. & Beauchemin, C. (2016), 'International migration under the microscope', *Science* **352**(6288), 897–899.

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**(1), 185–202.

Wiśniowski, A., Forster, J. J., Smith, P. W. F., Bijak, J. & Raymer, J. (2016), 'Integrated modelling of age and sex patterns of European migration', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **179**(4), 1007–1024.

Wiśniowski, A., Bijak, J., Christiansen, S., Forster, J. J., Keilman, N., Raymer, J. & Smith, P. W. (2013), 'Utilising Expert Opinion to Improve the Measurement of International Migration in Europe', *Journal of Official Statistics* **29**(4), 583–607.

Zagheni, E., Garimella, V. R. K., Weber, I. & State, B. (2014), Inferring international and internal migration patterns from Twitter data, *in* 'Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion', ACM Press, Seoul, Korea, pp. 439–444.

Zagheni, E. & Weber, I. (2012), You are where you e-mail: using e-mail data to estimate international migration rates, *in* 'Proceedings of the 3rd Annual ACM Web Science Conference on - WebSci '12', ACM Press, Evanston, Illinois, pp. 348–351.

Appendices

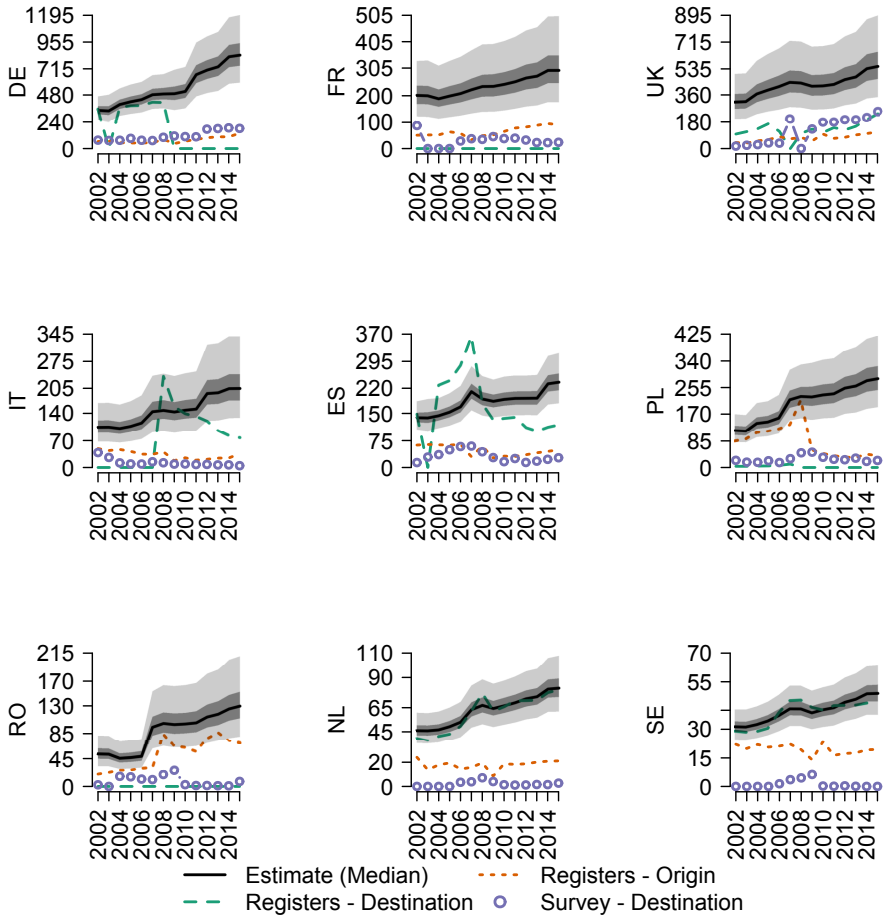

Figure 4: Immigration: Estimated total flow (per 1000) with 50% CI (dark gray) and 95% CI (light gray) for nine European countries. In addition, total flows from immigration register data (dashed green line), from emigration register data (dotted orange lines), and from LFS data (purple circles).

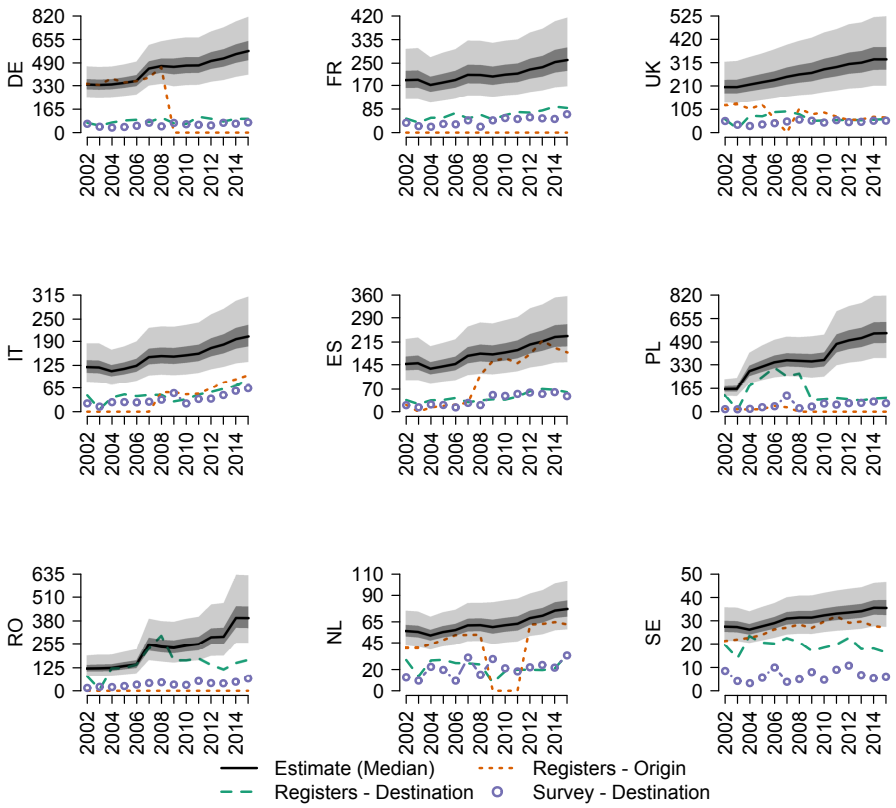

Figure 5: Emigration: Estimated total flow (per 1000) with 50% CI (dark gray) and 95% CI (light gray) for nine European countries. In addition, total flows from immigration register data (dashed green line), from emigration register data (dotted orange lines), and from LFS data (purple circles).

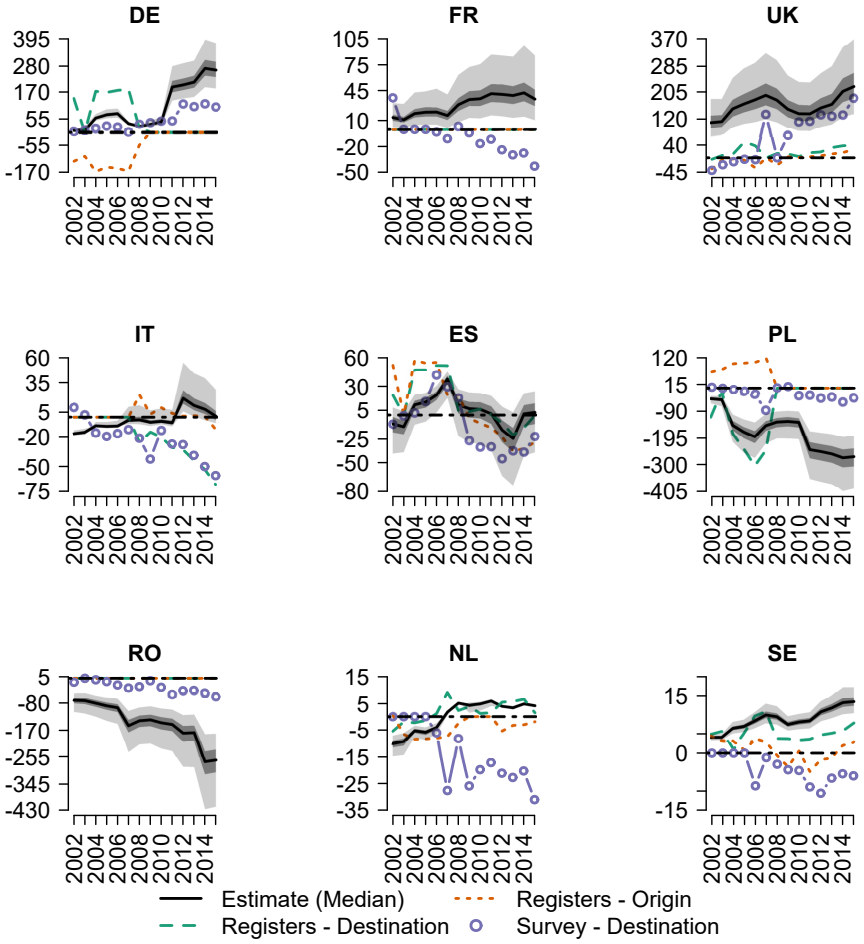

Figure 6: Net migration: Estimated total flow (per 1000) with 50% CI (dark gray) and 95% CI (light gray) for nine European countries. In addition, total flows from immigration register data (dashed green line), from emigration register data (dotted orange lines), and from LFS data (purple circles). The horizontal black dashed-dotted line indicates the level of net migration equal to zero.

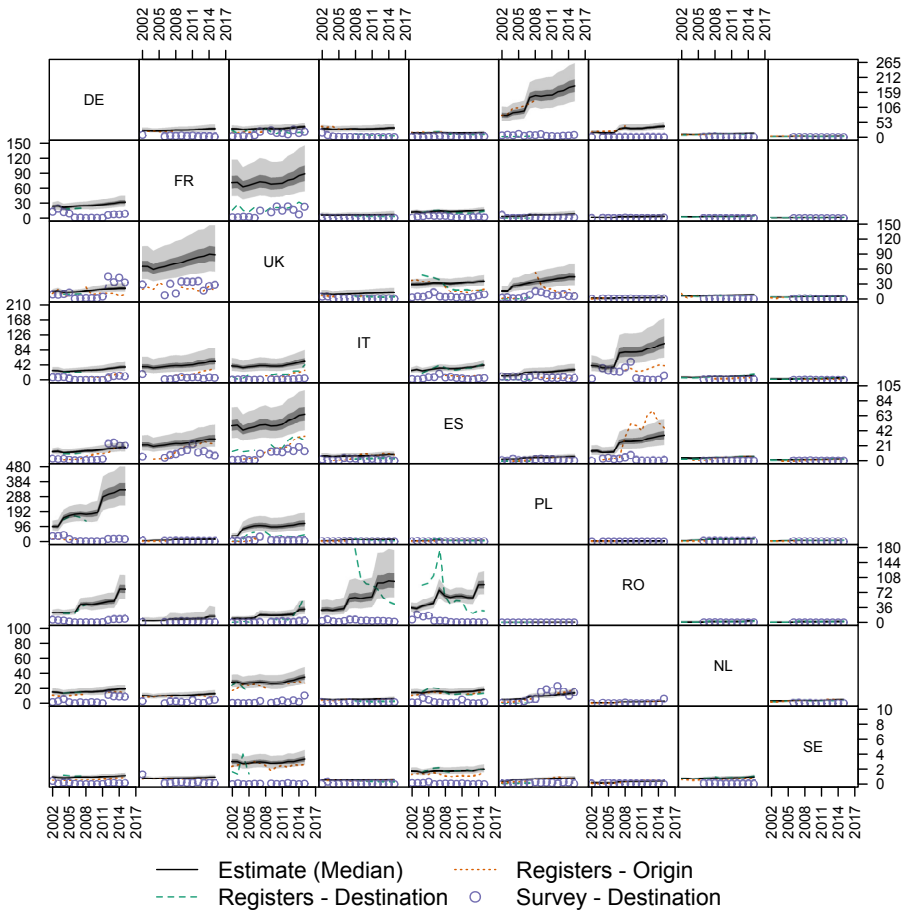

Figure 7: Estimated bilateral flows (per 1000) with 50% CI (dark gray) and 95% CI (light gray) for nine European countries. In addition, immigration flows (dashed line) and emigration flows (dotted line) from register data, and immigration flow from LFS data (circles).