Demography (2021) 58(6):2193–2218 Published online: 9 November 2021
DOI 10.1215/00703370-9578562

A Framework for Estimating Migrant Stocks Using Digital Traces and Survey Data: An Application in the United Kingdom

Francesco Rampazzo, Jakub Bijak, Agnese Vitali, Ingmar Weber, and Emilio Zagheni

ABSTRACT An accurate estimation of international migration is hampered by a lack of timely and comprehensive data, and by the use of different definitions and measures of migration in different countries. In an effort to address this situation, we complement traditional data sources for the United Kingdom with social media data: our aim is to understand whether information from digital traces can help measure international migration. The Bayesian framework proposed is used to combine data from the Labour Force Survey (LFS) and the Facebook Advertising Platform to study the number of European migrants in the United Kingdom, with the aim of producing more accurate estimates of the numbers of European migrants. The overarching model is divided into a Theory-Based Model of migration and a Measurement Error Model. We review the quality of the LFS and Facebook data, paying particular attention to the biases of these sources. The results indicate visible yet uncertain differences between model estimates using the Bayesian framework and individual sources. Sensitivity analysis techniques are used to evaluate the quality of the model. The advantages and limitations of this approach, which can be applied in other contexts, are discussed. We cannot necessarily trust any individual source, but combining them through modeling offers valuable insights.

KEYWORDS Migration • Facebook • Bayesian methods • Integrated Model of European Migration • European migrants

Introduction

Measuring international migration is challenging (Bilsborrow et al. 1997). The lack of timely and comprehensive data about migrants, combined with the varying measures and definitions of migration used by different countries, is a barrier to accurately estimating international migration (Bijak 2010; Willekens 1994, 2019). In recent years, scholars have started using Bayesian methods to combine different sources of migration data in order to provide better estimates of the migrant stock—the total number of migrants present in a country at a certain date (Azose and Raftery 2019). In this

ELECTRONIC SUPPLEMENTARY MATERIAL The online version of this article (https://doi.org/10.1215/00703370-9578562) contains supplementary material.

article, we aim to improve estimates by complementing survey data with social media data. This is important because, when designing migration policies, it is crucial to have access to valid sources of data on international migration. We propose using a Bayesian data assessment model that combines data from the Labour Force Survey (LFS) and the Facebook Advertising Platform to assess the number of European migrants in the United Kingdom (UK). The goal is to demonstrate how such a model can produce a more accurate estimate of European migration. We use the UK as an example as it is a Western country for which the migration data are of poor quality.

We employ the Integrated Model of European Migration (IMEM), which is a Bayesian model for estimating migration. This framework was created by Raymer et al. (2013) for combining the flows reported by the sending countries with the flows reported by the receiving countries to estimate a number closer to the true value of the flows. The IMEM model with modifications has been used by Disney (2015) to combine multiple migration survey data sets in the UK, and by Wiśniowski (2017) to combine the LFS data in the case of Polish migration to the UK. More recently, Del Fava et al. (2019) expanded the model by drawing on administrative and household survey data for 31 European countries. The main feature of the IMEM approach is that it provides a framework that assesses the limitations of the available data sets in terms of the definition of migrants used. Assessments of the bias and the accuracy of these data sets are used to create appropriate prior distributions to adjust for the identified data issues.

At the same time, a new strand of research has emerged that has been repurposing digital data to complement traditional demographic data sources and to improve their coverage and timeliness of production. Since digital trace data are often geolocated, migration has received particular attention in this literature. As Cesare et al. (2018) have suggested, using digital trace data sources has advantages, such as the speed and low cost of data collection, but also limitations, with issues regarding the lack of accessibility, transparency, and representativeness. Drawing on data from the Facebook Advertising Platform and the LFS, we investigate whether the digital traces that individuals leave on Facebook can be used to estimate stocks of migrants in the UK.

This is by no means the first study that has tried to combine digital traces with survey data (Alexander et al. 2019, 2020; Zagheni et al. 2018). However, in this article, we propose for the first time an overarching framework that includes both a theoretical model that considers push and pull factors related to migration theories and a data assessment model that aims to reduce the bias from the data that enters the model. This framework provides a more context-specific model for examining migration to the UK from several sending countries. Moreover, our study provides important insights into the complex reality of international migration to the UK by shedding light on the demographics of migrants by country of origin, which are hard to obtain using currently available official statistics. The attention is limited to migrants from European countries because, in the UK context, these migrant stocks are the hardest to estimate owing to the "freedom of movement" that characterizes the European Union (EU). At least until December 2020, there has been no requirement for EU migrants in the UK to register their residence. Thus, up to now, survey data have been used to estimate the stock of migrants from the EU. We want to complement these existing, but incomplete, official estimates of migrant stocks by analyzing digital trace data. As an illustration, we produce an estimate of the total number of EU migrants for 2018 and 2019.

There are two additional reasons why it is interesting to look at the migration system of the UK. First, the UK Office of National Statistics (ONS) bases its estimates of international migration on surveys. In August 2019, the ONS reclassified their estimates as experimental statistics, emphasizing that the estimates might be inaccurate (ONS 2019a). Furthermore, the scientific literature has suggested that these surveys are affected by different sources of bias (Coleman 1983; Kupiszewska and Nowok 2008; Kupiszewska et al. 2010; Rendall et al. 2003). In Europe, the UK is an example of a country in which there is only a "bronze standard," meaning that the UK migration data sources are inferior to the "gold standard" but are of "sufficient quality for validation" (Azose and Raftery 2019). Second, although the UK has experienced a net positive increase in migration from European countries over the past two decades (Champion and Falkingham 2016), the ONS reported an undercount of 16% for the net migration estimates for the EU8 countries (Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Slovakia, and Slovenia) in 2016, suggesting that the relevant migration statistics are of insufficient quality (ONS 2019b). Using digital traces might provide insights into UK migration trends by sex and country of origin by enabling researchers to produce estimates of stocks of European migrants in the UK. Moreover, Willekens (1994, 2019) has called for the creation of a synthetic migration database that combines data from different sources. The purpose of this database would be to "create the best possible estimates of the true number of migrants" (Willekens 2019). The present article seeks to contribute to this "learning process" by answering the following research question: What can Facebook advertising data contribute to ONS migration estimates, in a context in which there is no "ground truth" data against which model estimates can be validated?

Data

Traditional Data and Their Limitations

A gold standard for migration estimates does not yet exist. In fact, Swedish register data, long considered the gold standard among demographic data sets, have been proved to overcount migrants (Monti et al. 2019). Using traditional data to estimate the migrant stock, such as data from censuses, administrative sources, and surveys, presents limitations related to the definition of migrant, the coverage of the migrant population, and the accuracy of the estimates (Willekens 2019). Moreover, traditional sources of migration data are not timely. The United Nations suggested using the following definition of an international migrant in order to harmonize data sources on migration worldwide (United Nations 1998): "person who moves from their country of usual residence for a period of at least 12 months." An individual who lives abroad for a period of 3–12 months is considered a short-term migrant.

While Europe-wide data sources follow the standard definition of an international migrant (European Parliament and Council of the European Union 2007), individual European countries use a variety of systems to track the number of international migrants living within their borders. While censuses are considered the best source of data for estimating migrant numbers, these data have at least three limitations (Willekens 1994, 2019). First is that census data are collected every 10 years, and so

they do not provide a timely picture of migration. Second, the census records immigrants living in the country, but does not account for the emigrants that have left the country. And third, the census does not ask for important data such as the individual's age at time of migration or return migration.

Administrative data sources, such as population registers, can also be used to estimate migrants. Only a handful of countries use survey data to estimate international migration. The advantage of survey data collected from migrants is that they might provide additional information that is not included in the census or administrative data sources. However, survey data might fail to adequately cover the migrant population.

In the absence of registers, the UK largely relies on a survey-based system to collect information on its migrant population. The two main sources used to estimate international migration to the UK are the International Passenger Survey (IPS) and the Labour Force Survey. The IPS has been running since 1961, and it was originally introduced to estimate levels of overseas travel and tourism. It is currently the official source of data for estimating inflows and outflows of international migrants. The ONS itself admitted that the IPS "has been stretched beyond its original purpose" (ONS 2019c) and cannot be used as the only source when seeking to estimate international migration in the UK.

The second main data source is the LFS, a Europe-wide quarterly household survey that aims to estimate labor market conditions such as employment levels. Through a boost of this survey provided by the Annual Population Survey (APS), the ONS collects data on the stocks of foreign-born and foreign citizens living in the UK at the local authority level. The APS records information on the length of time migrants have already spent in the UK. The LFS interviews 41,000 UK households per quarter (ONS 2018a) and combines these data with data from two quarterly waves of the LFS to create a sample covering 360,000 individuals and 170,000 households per year. The data are released three months after the end of the survey.

The limitations of the sampling framework, the systematic bias, and the coverage of both the IPS and LFS have been described in several studies (Coleman 1983; Kupiszewska et al. 2010; Kupiszewska and Nowok 2008; Rendall et al. 2003). In addition, the ONS has recently started a work program that aims to combine data from additional administrative sources with data from the IPS and LFS in order to obtain a comprehensive measure of migration (ONS 2018b).

Digital Traces and Their Limitations

New social media data sources might be used to improve official migration statistics, as these sources can provide information on the backgrounds and other demographic characteristics of migrants. Digital traces can be collected quickly using the Application Programming Interface (API), which links researchers, as the client, to the server, where the data we are interested in are stored in the form of a database (Cooksey 2014; Sloan and Quan-Haase 2017). The ability to know in real time how many of the users are in a specific location can help us to *nowcast* migration.

In addition, social media data can be geolocated. For example, email location data have been used to estimate international migration rates (Zagheni and Weber 2012). These data are cheap because they are collected by repurposing data sets

originally intended for advertising. Thus, by relying on these data, we no longer need to create new data infrastructures to collect data. Moreover, these new data sources can provide us with insights that will enable us to expand the definition of an international migrant. Different countries use different definitions of a migrant that vary depending on the length of time an individual must spend outside of their usual country of residence. Thus, the definition of migrant is still not harmonized worldwide (Kupiszewska and Nowok 2008; Willekens 1994). Fiorio et al. (2021) have highlighted the possibility of using geotagged Twitter data to investigate short-term mobility and long-term migration. They suggested that drawing on digital trace data could help to refine migration theory and modeling. In addition, these data can be augmented through data from dedicated surveys of populations that are too hard or too expensive to reach with a traditional sampling framework (Pötzschke and Braun 2017; Rosenzweig et al. 2020).

Nevertheless, these sources also have important limitations. In some cases, researchers do not have direct access to all these new data sets and need to create partnerships with private companies to obtain the desired level of access (Blumenstock 2012). Digital trace data from LinkedIn might provide insights into trends in highly skilled migration to the United States (State et al. 2014), while data from the Web of Science have been used to follow trends and patterns of international migration among scholars (Aref et al. 2019). However, these sources do not provide data that are representative of the entire population. Hargittai (2018) analyzed the potential bias of different platforms in the United States, including Facebook, LinkedIn, Twitter, Tumblr, and Reddit. She found that Facebook is the most representative social media platform across educational and internet skill levels, while the other social media platforms are used by smaller and more specific U.S. population groups. The work of Hargittai builds on the critique by Lazer et al. (2014) of the assumption that we can substitute traditional data sources with digital trace data by showing that using these new data sources without considering their bias is problematic. These authors have also pointed out the algorithm dynamics and the unstable characteristics of digital traces, as the companies that generate the data we are seeking to use are constantly modifying their algorithms and are in full control of the information the researchers ultimately receive.

In this article, we focus on Facebook Advertising Platform data. Facebook provides advertisers with information on its users, including each user's age, sex, level of education, and language. For this reason, Facebook has been described as a biased *digital census* (Cesare et al. 2018; Zagheni et al. 2017). Facebook's main business is advertising, and the data provided on the Facebook Advertising Platform are made available to advertisers to help them plan their online campaigns. Facebook has a strong incentive to accurately report the characteristics of its users, because its ability to do so has become the main focus of its business, as the company is aware that advertisers might change platforms if they cannot target the right audiences through their services. We are repurposing data from this advertising platform for demographic research.

The variable that is currently used to estimate international migrants is defined by Facebook as "people that used to live in country x and now live in country y." This variable was first used in Zagheni et al. (2017), where it was compared to data from the American Community Survey. Until December 2018, the variable was referred

to as "expat from country x," showing that the wording of Facebook's definition of migrant has changed over time. However, their documentation does not provide information on which individual characteristics have been used to create the variable, or whether the algorithm identifying a user as a migrant was changed along with the change in the wording of the definition in 2018. Two studies have investigated how Facebook processes this category. In the first, researchers at Facebook suggested that Facebook users are considered "expats" on the basis of the location of their hometown and the structure of their friendship networks (Herdağdelen and Marelli 2017). In the second study, Spyratos et al. (2019) ran a survey in which 114 Facebook users were asked whether Facebook's Advertising Platform identifies them as an "expat" on the targeting platform. The authors concluded that Facebook uses other types of information that are not specified in the users' profiles, including geolocation outputs. The final clue can be found in Facebook's form 10-K, which is a U.S. Securities and Exchange Commission (U.S. SEC) document that provides a summary of Facebook, Inc.'s financial performance on the stock market. In these documents, Facebook wrote that "the geographic location of our users is estimated based on a number of factors, such as user's IP address and self-disclosed location" (U.S. SEC 2019, 2020). In the current article, we additionally leverage the variable "language" from the Facebook Advertising Platform. Facebook reported that it is possible to "target people with language other than common language for a location."[1]

The Facebook marketing API provides two metrics: Daily Active Users (DAUs) and Monthly Active Users (MAUs). On *Facebook for developers*,[2] DAUs are defined as the "estimated number of people that have been active on your selected platforms and satisfy your targeting spec in the past day," while MAUs are defined as the "estimated number of people that have been active on your selected platforms and satisfy your targeting spec in the past month." The same U.S. SEC document (U.S. SEC 2019, 2020) reported estimates of the bias of MAUs in 2018 and 2019, estimating that 11% of accounts were duplicated and 5% of accounts were false. Most of these anomalies were detected in Southeast Asia. We are using the MAUs estimates, because the Facebook document makes clear that this measure is more stable than the DAUs metric. The MAUs metric does not report numbers under 1,000 to prevent the targeting of small groups of individuals. Through the Facebook marketing API, we included in the current study all Facebook users in an aggregated and anonymized format.

Comparison Between LFS Data and Facebook Data

The two main data sources we used are the LFS and the Facebook Advertising Platform. We included 20 of the EU27 countries in our study: Austria, Belgium, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, the Netherlands, Poland, Portugal, Romania, Slovakia, Spain, and Sweden. Malta and Luxembourg were excluded because of their small size, Bulgaria and Croatia were excluded because Facebook does not provide estimates of expat numbers for them, and Estonia and Slovenia were excluded for missing values in the data used as covariates.

[1] See https://developers.facebook.com/docs/marketing-api/audiences/reference/advanced-targeting/.
[2] See https://developers.facebook.com.

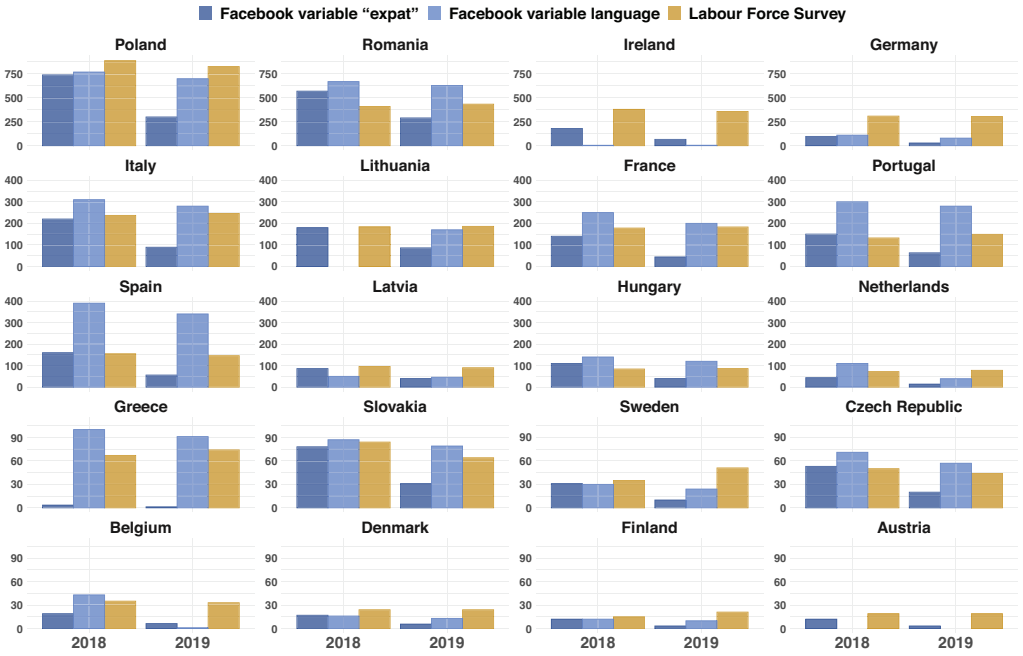

Fig. 1 Facebook's aggregated estimates (in 1,000s) for the "expat" and language variables and Labour Force Survey data of migrant stocks from 20 European countries of origin in 2018 and 2019

Moreover, Cyprus was excluded because the Facebook "expat" estimates might include all the users living there (Gendronneau et al. 2019). The aggregated estimates of European migrants from the Facebook Advertising Platform were collected in the third week of July 2018 and July 2019. We used *pySocialWatcher*, which is a Python package, to download the data (DAUs and MAUs) from the Facebook API (Araujo et al. 2017). The data from the LFS were provided by the ONS for the period of June–July 2018 and June–July 2019. For the purpose of this analysis, we have assumed that the age structure of the LFS and Facebook migrant users did not change much between 2018 and 2019.

Figure 1 shows a comparison between these two data sources for the two years included in the analysis. Three variables from three data sources are shown: the migrant variable and language variable from Facebook, and estimates of migrant stocks by country of birth from the LFS. We can see a correlation between the Facebook migrant and language variables for many countries. The correlation between these variables is 0.92 for both years, while the correlation between the Facebook migrant variable and the LFS estimates is 0.91 in 2018 and 0.88 in 2019. However, there are exceptions:

- for countries with a language that is also spoken in other countries (e.g., German in Germany, Austria, Switzerland, and Belgium; or French in France, Switzerland, and Belgium); and
- for Greece, where we notice that the "expat" variable on Facebook does not capture the Greek migrants. (The Greek language is spoken in Greece and part of Cyprus.)

Figure 1 shows a visible drop in the Facebook migrant variable estimates between 2018 and 2019. This is not due to out-migration from the UK, but rather due to an algorithm change that affected the Facebook estimates. In Figure B2 in the online appendix, we highlight the shift that happened in the middle of March 2019, which led to an average change in the estimates of 48%.

Additional Data Sources

In this analysis, additional sources are used as covariates that can help us estimate migrant stocks. We used data on inflows and outflows of migrants from and to the UK from the IPS for 2017 and 2018. We used information on the populations of the countries of origin from the projections produced by Eurostat, together with the Eurostat estimates of unemployment and gross domestic product (GDP) per capita. The population data are used for the analysis for the years 2018 and 2019, while the other two data sets are used for the analysis for the years 2017 and 2018.

Data from the UK settled and presettled status scheme are added to make an additional comparison. These data come from the UK Home Office. This scheme allows European migrants already residing in the UK to apply for presettled status if they have been living in the UK for less than five years, and for settled status if they have been living there for five years or more. The measure of applications to the scheme provides an indication of the number of Europeans who want to continue to have the right to remain in the UK after Brexit has been finalized. The data represent an estimate for the total number of applications and includes the period from August 28, 2018, to December 31, 2019.

Methodology

General Model Architecture

The aim of the IMEM framework is to estimate the true or latent flow of international migrants across sending and receiving countries by combining biased data (Raymer et al. 2013). The original IMEM model combines flows from sending and receiving countries across the EU. Our aim is to provide an estimate of the true stock of European migrants in the UK based on a combination of the LFS and Facebook Advertising Platform data. The estimate of true stock is the number of migrants who would be counted if our collection system were able to perfectly measure all migrants (Disney 2015). While the true number of migrants is not known, through the use of Bayesian methods we might estimate a probability distribution for the true number of migrants that reflects our knowledge about it. These true or latent estimates from the model incorporate all the information collected from the various data sources, as well as our prior data about the migration process. Thus, the point estimate of the true number of migrants would be a summary of this distribution (i.e., the median).

The model is divided into two parts: the Measurement Error Model (MEM) and the Theory-Based Model (TBM). In the MEM, the Facebook Advertising Platform and LFS data are combined; in the TBM, other variables are also considered in the estimation

of the true stock. In this framework, the IMEM quantifies the limitations of the data sources and provides the appropriate prior distribution in order to reduce the bias.

The limitations of the data are assessed in terms of the following (Disney 2015; Raymer et al. 2013):

- *Definition*: How closely does the international migrant measure match the United Nation's definition of an international migrant?
- *Coverage*: What proportion of the total immigration stock does the data cover?
- *Bias*: Is there any systematic bias in the data?

In Figure 2, the model is explained using a diagram that is divided into four parts: input, data assessment, model, and output. In the input column, the data sources are survey data, digital trace data, and migration theory covariates for the TBM. The data assessment is followed by a summary of the limitations of the data in terms of definition, bias, and coverage. In the model box, the true stock at the center of the figure is estimated by the TBM and the MEM, which combine the stock estimates from the LFS with those from the Facebook Advertising Platform, while incorporating considerations related to definition, bias, and accuracy. Finally, in the output, the diagnostics and results are shown.

The model is constructed as follows. The number of European migrants (stocks), z^k_{ijt}, from a certain country, i, in the UK with a certain characteristic, j, is observed. In this case the characteristic selected is sex. This is done using data from Facebook, F, and from the LFS, L, and the value k is then used to represent either L or F depending on which data are used to measure the European migrants stock (z^k_{ijt}). The year, t, in this case is 2018 and 2019. The data sets used can thus be described in the form of matrices $\mathbf{Z}^F$ (Eq. (1)) for Facebook and $\mathbf{Z}^L$ (Eq. (2)) for the LFS. The model borrows strength across the two years:

$$\mathbf{Z}^F = \begin{pmatrix} z^F_{11t} & z^F_{12t} & \cdots & z^F_{1Jt} \\ z^F_{21t} & z^F_{22t} & \cdots & z^F_{2Jt} \\ \vdots & \vdots & \ddots & \vdots \\ z^F_{I1t} & z^F_{I2t} & \cdots & z^F_{IJt} \end{pmatrix}, \tag{1}$$

$$\mathbf{Z}^L = \begin{pmatrix} z^L_{11t} & z^L_{12t} & \cdots & z^L_{1Jt} \\ z^L_{21t} & z^L_{22t} & \cdots & z^L_{2.Jt} \\ \vdots & \vdots & \ddots & \vdots \\ z^L_{I1t} & z^L_{I2t} & \cdots & z^L_{IJt} \end{pmatrix}. \tag{2}$$

For every time t, the value of $\mathbf{Y}_{ijt}$ (Eq. (3)) is the random variable estimate of the true stock. It is a matrix with dimension $I \times J$:

$$\mathbf{Y} = \begin{pmatrix} y_{11t} & y_{12t} & \cdots & y_{1Jt} \\ y_{21t} & y_{22t} & \cdots & y_{2Jt} \\ \vdots & \vdots & \ddots & \vdots \\ y_{I1t} & y_{I2t} & \cdots & y_{IJt} \end{pmatrix}. \tag{3}$$

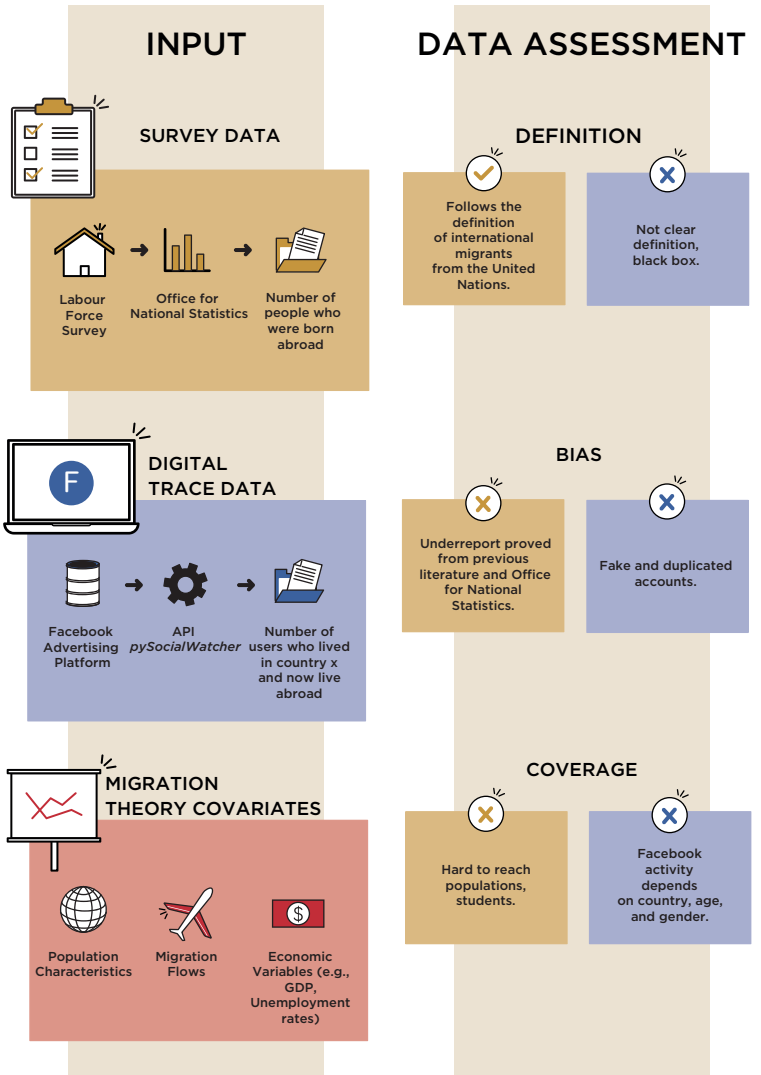

Fig. 2 Diagram illustrating the structure of the model

The value of z_{ijt}^{k} is assumed to follow a Poisson distribution (Eq. (4)). The Poisson distribution is a probability distribution of the number of times an event is expected to occur. Here, the distribution of European migrants is based on expectations from the Facebook and LFS data. The distribution is

$$z_{ijt}^{k} \sim \text{Po}(\mu_{ijt}^{k}). \tag{4}$$

Figure 3 illustrates the hierarchical structure of the model, which is explained in detail in the following section. The model is estimated using JAGS in R (Plummer et al. 2016). In JAGS, the normal distributions are defined in terms of the mean, μ, and precision (i.e., one over the variance), τ. The JAGS notation is used.

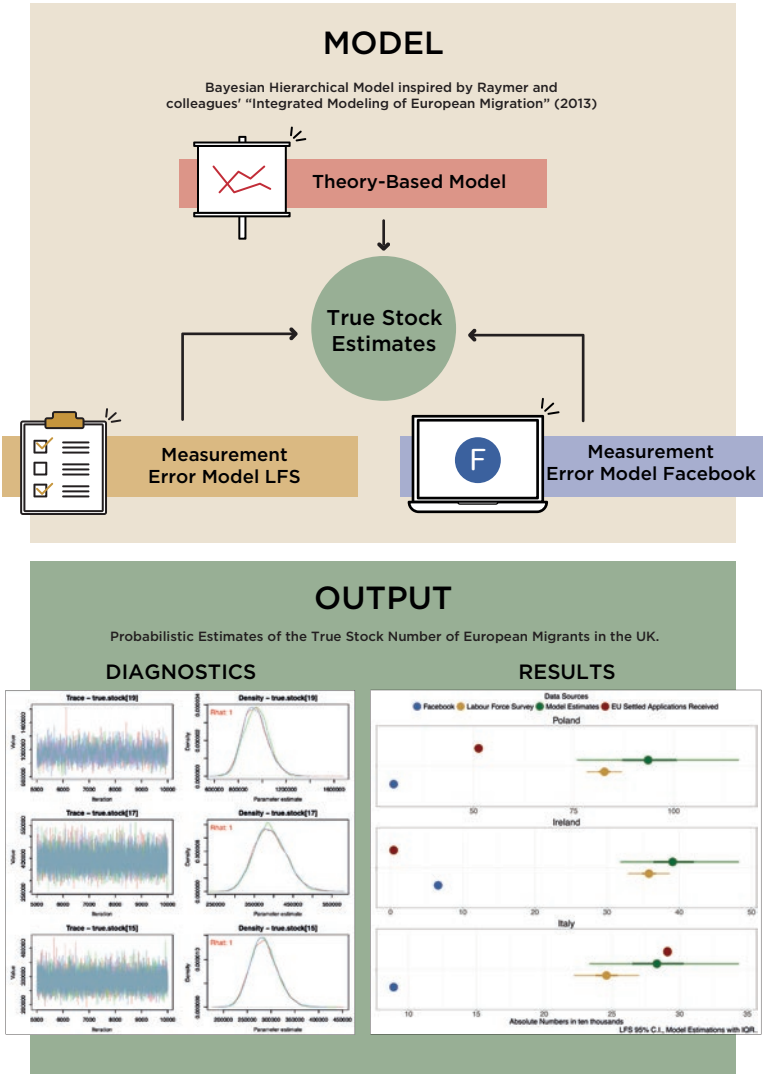

Fig. 2 (continued)

Measurement Error Models

The Measurement Error Models describe how the observed values relate to the true count. The general equation (5) of the Measurement Error Model is

$$\log \mu_{ijt}^{k} = \log y_{ijt} + \delta^{k} + \beta^{k} + \chi_{ijt}^{k} + \xi_{ijt}^{k} + \lambda_{ijt}^{k} + \varepsilon_{ijt}^{k}. \quad (5)$$

The equation is composed of five terms, δ^k, β^k, χ_{ijt}^k, ξ_{ijt}^k, and λ_{ijt}^k, which are used to convert the data from Facebook and the LFS to comply with the United Nation's definition of an international migrant, and to reduce the underestimation linked to the bias or coverage of the data. The first parameter, δ^k, captures the differences in relation to the definition of migrants. The bias in the data is captured by β^k, while the coverage

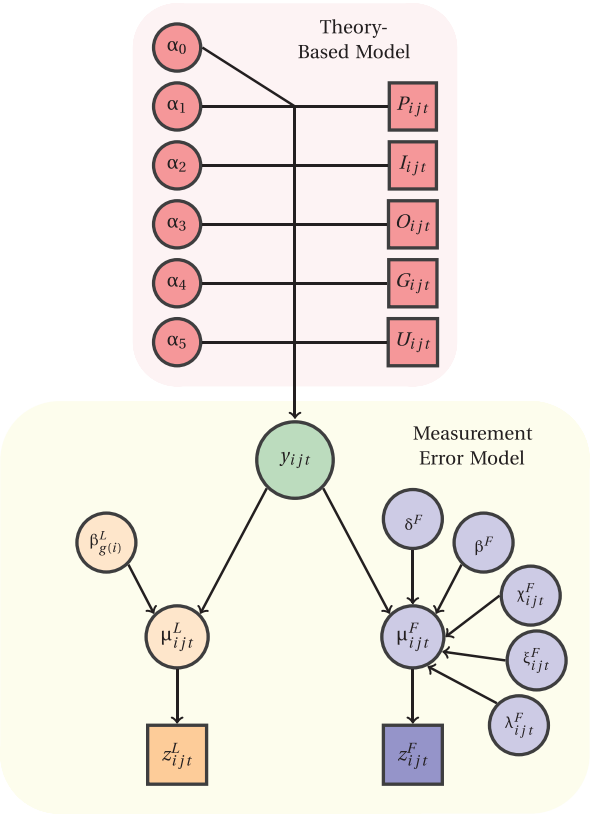

Fig. 3 Graphical representation of the adapted IMEM (diagram inspired by Raymer et al. (2013:804)). The hyperparameters are not shown for greater clarity of presentation. Indices: *i*, sending country; *j*, sex; *t*, time. Square nodes represent reported data (z^L_{ijt}, z^F_{ijt}) and covariates. Circle nodes represent parameters for the migration model and the measurement model.

of the Facebook data is considered in χ^k_{ijt}. The parameter ξ^k_{ijt} deflates the Facebook estimates of 2018 by the algorithm change that happened in 2019. The parameter λ^k_{ijt} inflates the Facebook estimates with knowledge provided by the Facebook estimates of people speaking a certain language. The term ε^k_{ijt} is the error term with normal distribution $N(0, \tau_{ijt})$, and the precision τ_{ijt} has Gamma distribution $G(100, 1)$ (where 100 is the shape parameter and 1 is the rate parameter), which has a mean equal to 100 and precision equal to 1 (e.g., variance equal to 100). Table 1 summarizes the parametrization of the model and the direction of the prior distributions.

Data Assessment of the Labour Force Survey

The LFS defines a long-term international migrant in the same way as the United Nations (ONS 2018a) and provides data on each migrant's country of birth and citizenship. For our purposes, the country of birth criterion is used because it captures individuals with a migrant background, including those who acquired citizenship

Table 1 Parameters in the Measurement Error Model for the Labour Force Survey and Facebook

Parameter	Interpretation	Labour Force Survey	Facebook
δ	Definition		Unknown definition, but with some variation
β	Bias	Inflation of the estimates + { 4% undercount low; 12% undercount medium; 30% undercount high }	Deflation of the estimates –4 fake, duplicates
χ	Coverage		± coverage by sex in the home country
ξ	Algorithm change		~ effect of an algorithm change in 2019
λ	Language parameter		~ Greek language dummy parameter

through naturalization. Because the LFS is used to estimate the stock of migrants in the UK, many researchers have investigated the quality of the survey's estimates and found that they underestimate migrants. Rendall et al. (2003), for example, reported that the 2001 LFS underreported international migrants by 26% compared with the 2001 census. Other research has shown that the bias in the LFS might be as high as 30% for nationalities with smaller stocks, such as Greeks and Lithuanians (Kupiszewska et al. 2010), and that the survey has a nonresponse rate of more than 15% (Martí and Ródenas 2007). Furthermore, the sampling framework of the LFS does not cover the entire target population (Kupiszewska et al. 2010) as students and more mobile migrants might not fully appear in the sample.

Table 2 compares data from the LFS collected between January and December 2011 with the British census that occurred on March 27, 2011. The data are aggregated for England and Wales only. The relative percentage change between the LFS and the census gives a sense of the bias between the two. It has to be stressed that the ONS has already attempted to recalibrate the LFS estimates with the results of the census. Despite this, there is still a problem with both undercounting and overcounting. The range of the bias is between –21% and 15%. This suggests that the LFS Measurement Error Equation is

$$\log \mu_{ijt}^{L} = \log y_{ijt} + \beta_{g(i)}^{L} + \varepsilon_{ijt}^{L}. \tag{6}$$

Regarding this assessment, the LFS data are deflated by only one parameter, β^{L}, which considers both the bias and the coverage of the data. A separate parameter, such as δ^{L}, is redundant as the definition of international migrant in the LFS follows the United Nations standard. The literature (Kupiszewska et al. 2010; Martí and Ródenas 2007; Rendall et al. 2003) suggests that for countries with small migrant populations in the UK, LFS migrant estimates may be around 30% lower than the true numbers. This percentage is reduced, to around 15%, for those nationalities with large populations in the UK. Table 2 provides a measure of the bias at a country level. The ONS reports that the quality of the LFS estimates decreases over time when distanced from the census year (ONS 2020). The classification relies on the literature, the data from

Table 2 Aggregated estimates of the number of EU migrants in England and Wales by country of origin according to the LFS and the census, and the relative percentage change

Country	LFS, January–March 2011	Census, March 2011	Relative Percentage Change
Austria	19,000	19,087	−0.46
Belgium	28,000	25,472	9.03
Czech Republic	37,000	37,150	−0.41
Denmark	18,000	21,445	−19.14
Finland	10,000	12,149	−21.49
France	134,000	129,804	3.13
Germany	279,000	273,564	1.95
Greece	33,000	34,389	−4.21
Hungary	44,000	48,308	−9.79
Ireland	353,000	407,357	−15.40
Italy	121,000	134,619	−11.26
Latvia	57,000	54,669	4.09
Lithuania	115,000	97,083	15.58
Netherlands	52,000	59,081	−13.62
Poland	572,000	579,121	−1.24
Portugal	83,000	88,161	−6.22
Romania	94,000	79,687	15.23
Slovakia	52,000	57,829	−11.20
Spain	69,000	79,184	−14.76
Sweden	30,000	30,694	−2.31

Notes: The relative percentage change is computed from the LFS data from January to December 2011 and the census in 2011. The LFS data available for January to March 2011 are already recalibrated through 2011 census data. LFS = Labour Force Survey.

Table 2, and the assessment from the ONS, as well as our own expertise. The LFS bias is anchored to the relative percentage change between the LFS and the census, and an increase of bias over time is also considered. As a matter of fact, the countries are divided into three groups:

- *Low*—Bias at 4%: Austria, Belgium, Czech Republic, Latvia, Sweden;
- *Medium*—Bias at 12%: France, Germany, Greece, Hungary, Lithuania;
- *High*—Bias at 30%: Denmark, Finland, Ireland, Italy, Netherlands, Poland, Portugal, Romania, Slovakia, Spain.

As a consequence, the β^L parameter is assigned according to a parameter $g(i)$, where

$$g(i) = \begin{cases} 1, & \text{if the undercount is assumed to be low;} \\ 2, & \text{if the undercount is assumed to be medium;} \\ 3, & \text{if the undercount is assumed to be high.} \end{cases} \quad (7)$$

The prior distribution is set to

$$\beta_i^L \sim \begin{cases} N(-0.04,\ 100), & \text{if the undercount is assumed to be low;} \\ N(-0.13,\ 100), & \text{if the undercount is assumed to be medium;} \\ N(-0.35,\ 100), & \text{if the undercount is assumed to be high.} \end{cases} \quad (8)$$

The means on the prior β^L are assumed to be time-invariant: they are considered as an approximation of the bias and thus small time variances are not accounted for. The term ε_{ijt}^k is the error term with normal distribution $N(0, \tau_{ijt})$, and the precision τ_{ijt} has Gamma distribution $G(100, 1)$, as previously described.

Data Assessment of the Facebook Advertising Platform

Given our earlier description of the Facebook data, a parameter was created for the data's definition, bias, and coverage. The Facebook δ^F is a priori assumed to be normally distributed with $N(0, 100)$, while β^F has a normal distribution $N(0.04, 100)$. The mean of β^F is set at 4% to deflate the Facebook estimates to account for fake and duplicate accounts. This value is lower than the 11% suggested by Facebook itself, because we assume that the percentage of fake and duplicated accounts labeled as belonging to migrants is lower in Europe. The mean of the coverage parameter χ_{ijt}^F (Eq. (9)) is the rate of non-Facebook users in the country of origin of the European migrants, since the aim is to correct by this adjustment. It is computed as

$$\chi_{ijt}^F = \log\left(1 - \frac{\text{Number of Facebook users}_{ijt}}{\text{Eurostat population size}_{ijt}}\right). \tag{9}$$

Additionally, the digital trace data are described as unstable. Indeed, it seems that Facebook reviewed its algorithm on expats in the middle of March 2019, and there was a drop in the migrant estimates after this time. The change is country- and sex-specific. For this reason, a parameter was introduced for the rate algorithm ξ_{ijt}^F (Eq. (10)), which aims to adjust the Facebook data for this bias caused by the change in the algorithm:

$$\xi_{ijt}^F = \log\left(\frac{\text{Estimates before}_{ij} - \text{Estimates after}_{ij}}{\text{Estimates before}_{ij}}\right). \tag{10}$$

A parameter was used for Greece that inflates the estimates of the Facebook expat variable (Eq. (11)), which reports a low number of "people that used to live in Greece and now live in the UK." However, the language variable, which Facebook uses to "target people with language other than common language for a location," provides some information that can be used to adjust the number of Greeks living in the UK. As the Greek language is also spoken by Cypriot migrants, the estimates are deflated by a ratio calculated using LFS data of the number of Greek and Cypriot migrants. Unfortunately, this is another sign that digital trace data are not perfect, as it seems that Facebook is not accounting for Greek migrants with the migrant variable (see also Figure A1 in the online appendix).

$$\lambda_{ijt}^F = \log\left(\frac{\text{FB Language}_{ijt}}{\text{FB Migrant}_{ijt}} \times \frac{\text{LFS Greece migrant}_{ijt}}{\text{LFS Greece migrant}_{ijt} + \text{LFS Cyprus migrant}_{ijt}}\right). \tag{11}$$

After this assessment, the Facebook Measurement Error Equation is

$$\log \mu_{ijt}^F = \log y_{ijt} + \delta^F + \beta^F + \chi_{ijt}^F + \xi_{ijt}^F + \lambda_{ijt}^F + \varepsilon_{ijt}^F. \tag{12}$$

Theory-Based Model

In this part of the model, covariates that might help to explain the true stock of European migrants in the UK are introduced:

$$\log y_{ijt} = \alpha_0 + \alpha_1 P_{ijt} + \alpha_2 I_{ijt} + \alpha_3 O_{ijt} + \alpha_4 \log G_{ijt} + \alpha_5 \log U_{ijt} + \varepsilon_{ijt}, \quad (13)$$

where $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_5)$ is a vector of parameters; α_0 is assumed to be normally distributed $\alpha_0 \sim N(0, 0.01)$, providing a weakly informative prior on the constant term, while $\alpha_{(1,\ldots,5)} \sim N(0, 1)$ is assumed to be more informative. The error term ε_{ijt} has a normal distribution $N(0, \tau_{ijt})$, with precision τ_{ijt} following a Gamma distribution $Gamma(100, 1)$.

The covariates used in the models for 2018 and 2019 include:

- P: a normalized measure of population size in the country of origin, divided by the mean of the population in the same countries considered in the model (data are from the latest estimates by Eurostat in 2018 and in 2019);
- I: a normalized measure of the inflows from European countries to the UK, divided by the mean of the inflows of migrants from the countries considered in the model (data are from the IPS in 2017 and in 2018);
- O: a normalized measure of the outflows to the European countries from the UK, divided by the mean of the outflows of emigrants from the countries considered in the model (data are from the IPS in 2017 and in 2018);
- G: ratio of GDP growth rate in the European country of origin in 2017 and in 2018, divided by the GDP growth rate in the UK (data are from Eurostat); and
- U: ratio of the unemployment rate in the European country of origin in 2017 and in 2018, divided by the unemployment rate in the UK (data are from Eurostat).

The normalized measure of the population size is a predictor of the possible number of migrants informed by a gravity model; that is, the larger the population, the larger the number of possible migrants. The normalized measures of inflows and outflows from the IPS provide an indication of the levels of fluctuation in terms of arrivals and departures for every nationality, and thus help to capture fluctuations in the stocks. The ratio of the GDP growth rate to the unemployment rate provides information on how the economy of the country of origin compares to that of the UK, and therefore is a form of economic gravity indicator. Study code and data are accessible at https://github.com/chiccorampazzo/fb-migration-uk.

Results

We present two sets of models. The first is for the total number of European migrants in the UK, and the second disaggregates the estimates by sex. They are run simultaneously by year (2018 and 2019) to borrow strength across the years. In the first model, the aim is to explain the magnitude of the undercount of the LFS data relative to the estimates produced by the model for the two years. All estimates of the models converge. Detailed results and diagnostic statistics are in the online appendix.

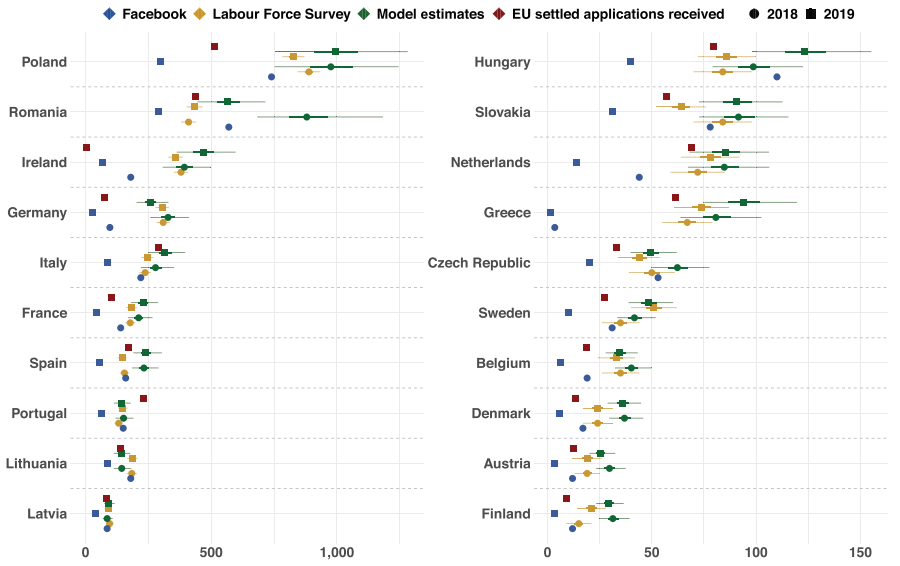

Fig. 4 Comparison of Facebook, LFS, and model estimations (absolute numbers in thousands) of European migrants aged 15 or older for the years 2018 and 2019. LFS data are shown with 95% confidence intervals, and model estimates are shown with the interquartile.

Model for Total Numbers

Figure 4 shows data from three data sets and our estimates: the Facebook advertising data are in blue, the LFS data are in yellow, the settled status application data are in red, and the model estimates are in green. The settled status data are presented for comparison only. LFS data are shown with a 95% confidence interval, while model estimates are shown with the interquartile (IQR).

There are three main messages that can be discerned from this figure. First, the differences between the Facebook data in 2018 and 2019 are readily visible, and are related to the algorithm change carried out by Facebook. However, the prior distribution on the algorithm parameter seems to fix this bias, as the differences between the 2018 and the 2019 estimates were relatively small. Second, while the LFS data are relatively consistent across the two years, a decreasing trend in the number of EU migrants in the UK is visible. Third, the model estimates are higher than the LFS estimates. In some cases, the IQR range of the model estimates includes the LFS estimates.

The parameter on Greece seems to be effective in bringing the estimates closer to the LFS values. In the online appendix, the posterior characteristics of the true stock estimates for all of the models and the $\hat{R}$ are reported, the latter being a measure that helps determine whether chains have converged depending on whether it is close to one (Gelman et al. 2013). All of the chains have converged when $\hat{R}$ is strictly equal to one (except for Romania in 2018 and Poland in 2019, where $\hat{R}$ is 1.01 as shown in the online appendix). The algorithm for estimating all of the other parameters has converged as well. Table 3 gives a comparison of the undercounted LFS estimates with the model estimates. While the ONS has estimated an undercount of 16%, the model estimates an undercount of 25% for 2018 and 20% for 2019. The undercount for 2018 has larger

Table 3 Percentage undercount of the LFS estimates in comparison with the model estimates

Year	2.5%	25%	50%	75%	97.5%
2018	13	21	25	29	37
2019	10	16	20	24	31

Note: LFS = Labour Force Survey.

intervals, likely owing to the prior on the algorithm change. Additionally, the model for 2019 estimates a higher number of migrants of certain nationalities (e.g., Polish, Italian, and Hungarian) and a lower number of migrants of other nationalities (e.g., Romanian, German, and Czech). The interquartile range of these distributions is large, highlighting the uncertainty in the estimates. However, the models for the two years indicate that the undercount and the uncertainty are in the same direction.

Model Disaggregated by Sex

In this part of the model, the estimates are disaggregated by sex, because it is important to study the age and sex differences of migrants, as there might be large differences across sexes. The model proposed works for sex disaggregation, and Figure 5 shows the estimates. In this case, the comparison with migrants who have applied for the settled status scheme is not available because the data from the Home Office are not disaggregated by sex.

Sensitivity Analysis

Some sensitivity checks of the model are provided. First, the model was run while including only the LFS data. For the model specified in this article, the undercount is estimated at 25% in 2018 and at 20% in 2019. In Table 4, the undercount of this new specification of the model is reported, estimated at a median level of 8% in 2018 and 22% in 2019. These two median levels are not close to those produced by the model that combines Facebook and LFS data, with a smaller undercount in 2018 and a larger one in 2019. Overall, the uncertainty of the undercount estimate is greater when using only LFS data. The second sensitivity check was to modify the parameters from the Facebook and LFS Measurement Error Models. In the models used here, the parameters are informed by previous research and calculations on the data, except the β^F, which is the bias parameter for Facebook. It is assumed the value is lower than the percentage of fake and duplicate accounts worldwide. In the sensitivity analysis, the Facebook bias parameter was first modified to 0%, indicating no bias in the Facebook estimates, and then to 11%.

In this paragraph, we analyze the different specification of the models looking at Table 4. The undercount with no bias attributed to the Facebook estimates is 22% for 2018 and 19% for 2019, which is slightly lower than that specified in the suggested model. The undercount with a higher β^F is 25% for 2018 and 20% for 2019. The undercounts with a β^F at 4% and at 11% are very similar.

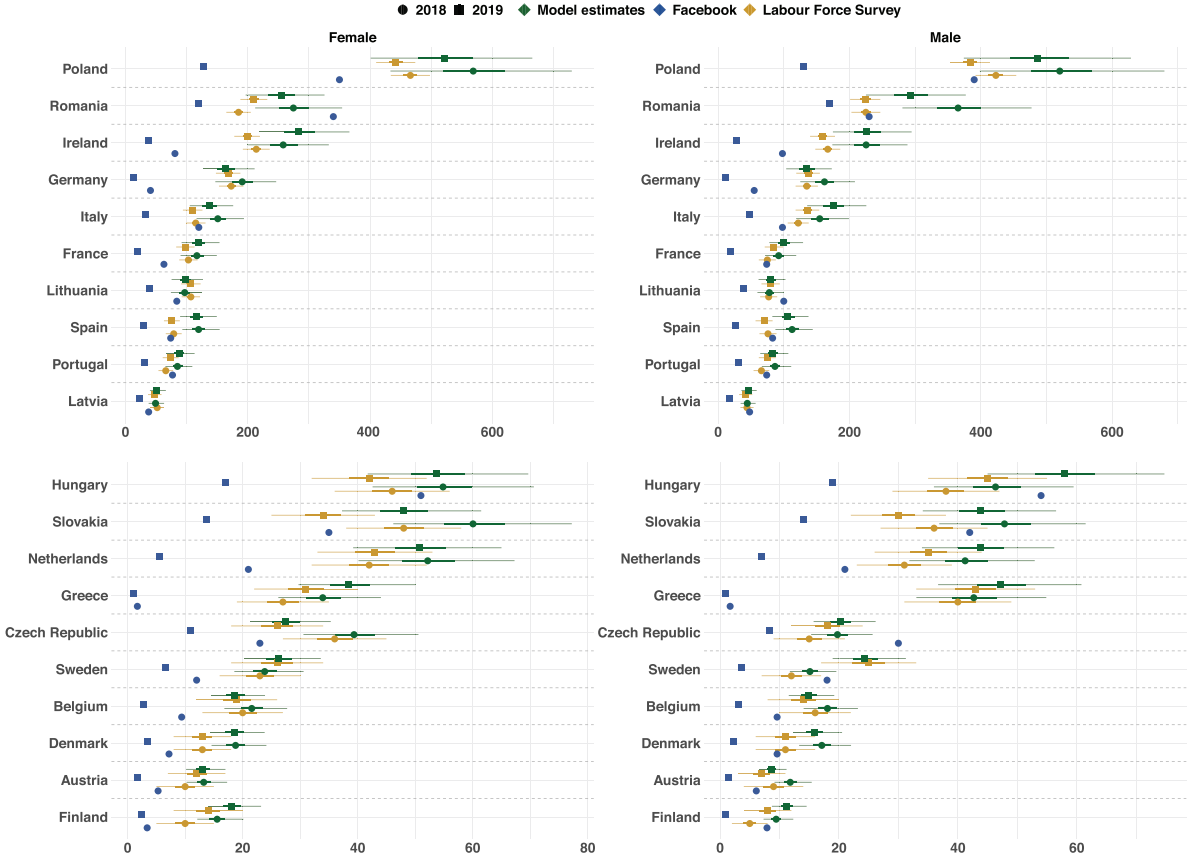

Fig. 5 Comparison of Facebook, LFS, and model estimations (absolute numbers in thousands) of European migrants aged 15 or older by sex for the years 2018 and 2019

Table 4 Percentage undercount of the LFS estimates in six different models: (1) specified only with the LFS data, (2) with the Facebook bias parameter set to 0%, (3) with the Facebook bias parameter set to 11%, (4) with the LFS bias parameter set to 4%, (5) with the LFS bias parameter set to 30%, and (6) with the *Gamma*(1,1) distribution

Model	Year	2.5%	25%	50%	75%	97.5%
Model without	2018	−77	−11	8	16	25
Facebook data	2019	−73	3	22	32	45
Model with Facebook	2018	11	18	22	26	34
bias at 0%	2019	9	15	19	22	30
Model with Facebook	2018	14	21	25	29	37
bias at 11%	2019	10	16	20	23	31
Model with LFS	2018	−9	−4	−1	2	8
bias at 4%	2019	−12	−7	−4	−1	5
Model with LFS	2018	22	29	33	37	46
bias at 30%	2019	19	26	30	34	42
Model with	2018	−9	10	21	34	65
Gamma(1,1)	2019	−15	4	15	27	55

The model is sensitive to the choice of the assumed bias of the LFS parameter. We modified the bias of the LFS to 4% (the minimum level assumed) and to 30% (the maximum level assumed) for all the countries. With the low minimum bias level assumed, the undercount reaches negative median values, while it is larger when the maximum bias level is assumed. We also tried different specifications of the precision distribution term, which is assumed to follow a *Gamma*(100, 1) in the presented model. The model was specified with a *Gamma*(1, 1), which is less informative than *Gamma*(100, 1). The gradient of the median of the undercount is similar to the one in the presented model, though the uncertainty is larger. There is some impact of the prior selection on the uncertainty of the estimates.

Finally, in Figure 6, the estimates from the model on the total estimates (model 1) are compared to the sum of the estimates from the sex disaggregation model. Though the estimates are close to each other, there are cases in which the sum from the sex disaggregation model is not completely aligned with the distribution from model 1. This is due to inconsistencies in the Facebook and LFS data disaggregated by sex. The estimates from our models seem to be stable to different prior distributions, yet the precision of those prior distributions had to be carefully chosen to ensure model convergence, while exploring reasonable areas of the parameter space with respect to the precision parameters.

Discussion

The model estimated the migrant stocks for 2018 and 2019. In the 2018 model, a prior distribution was used to account for an algorithm change that Facebook implemented in March 2019, which led to a decrease in the estimate of European migrant numbers. This algorithm change was not uniform, however, as it varied by country and sex of the migrants. This finding highlights the importance of monitoring digital traces and cautions that using digital traces alone is not sufficient to generate better

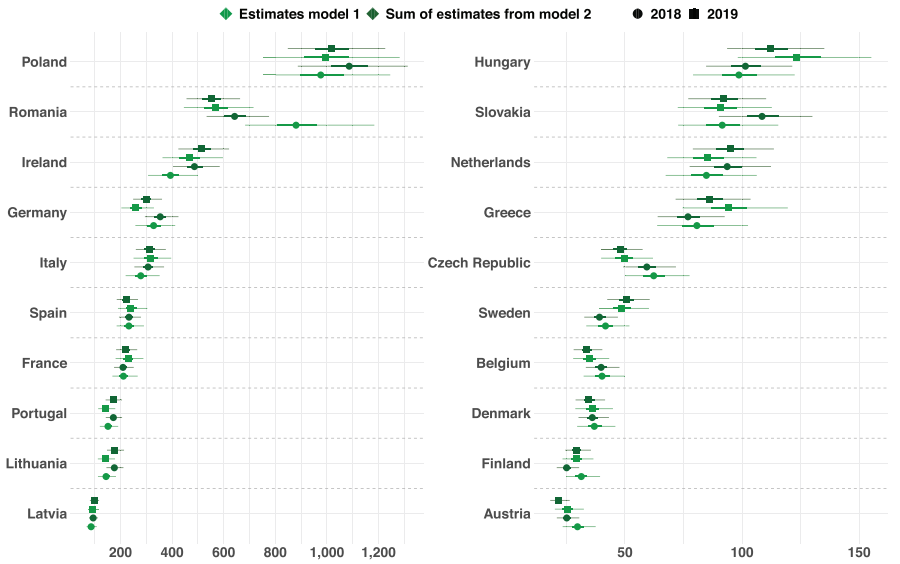

Fig. 6 Comparison between estimates from the first model and the sum of female and male migrants from the second model for 2018 and 2019 (absolute numbers in thousands)

estimates of stocks of migrants. The parameters associated with the algorithm change and the Greek factor (i.e., the factor that Greeks are underrepresented in the Facebook migrant variable) were shown to be effective in bringing the model estimates in line with the LFS estimates.

Our inclusion of the Home Office's data related to settlement and presettlement applications as an additional comparison proved interesting. For Polish migrants, the number of applicants to these schemes was lower than the LFS estimate, while for Romanian migrants, this number was the same as the LFS estimate. The number of applicants is expected to be lower than the LFS estimate of migrants because applying for the scheme before the end of the transition period is not mandatory. It was observed, however, that in some cases the settled status application number was higher than the LFS estimate but closer to the model estimates, suggesting that the model might have been producing a more accurate estimate than the LFS. For Italian migrants, for example, the number of settled status applications was close to the median estimate from the proposed model. Conversely, the model estimates for Portuguese migrants were closer to the LFS estimates and lower than the estimates of applicants for settled or presettled status. Interestingly, the results for the model estimates for Germany were also lower than the LFS estimates, but were closer to the estimates of those who filed a settlement or presettlement application. Almost no Irish nationals applied to the settled or presettled scheme owing to the bilateral agreements between the Republic of Ireland and the UK.

An estimate of the total number of European migrants by sex is also provided. The sum of the estimates from this second model was equal to the total from the first. There was uncertainty in our estimates, especially for the countries of origin with the highest number of migrants in the UK: Poland and Romania. This might suggest that for nationalities where the level of uncertainty is higher, the sample of households

and migrants interviewed should be increased. A possible solution to reduce the uncertainty would be to include a prior distribution in the model driven by expert opinion, as well as more informative priors on the Facebook and LFS data once they become available.

Moreover, the analysis showed one of the main limitations of digital trace data: the lack of transparency on how private digital companies produce their estimates. Indeed, it is not clear how exactly Facebook labels users as "people that used to live in country x and now live in country y," or how they determine which languages are spoken by the users on their platform. Furthermore, there are no details available about the algorithm change Facebook implemented in March 2019.

Conclusions

Our overarching research question was: What can Facebook advertising data contribute to ONS migration estimates in a context in which there are no "ground truth" data against which model estimates can be validated? This question was answered by exploring the two data sources and producing a probabilistic measure of European migration. Although this study found greater uncertainty in the estimates that were already known to be biased, it contributes to the "learning process" hoped for by Willekens (1994, 2019), which can lead to the extension of this framework. The obvious next step would be to expand the model to disaggregate the estimates by age and sex.

This analysis made three contributions to digital and computational demography. First, it proposed to apply a framework that is already in use in migration research to digital traces. The proposed model is a flexible framework, in which it is possible to include new information as soon as it becomes available, including additional digital trace data from other advertising platforms such as Instagram, Snapchat, and LinkedIn, as well as from other administrative sources. Second, it addressed the biases of both traditional and digital trace data. The use of a prior distribution was shown to fix these issues in a probabilistic fashion. Third, it produced an estimate of the undercount of migration levels. Overall, the model estimated an undercount of 25% for 2018 and 20% for 2019 based on the LFS data. For migrants to the UK from the EU8 countries, the ONS had estimated an undercount of 16% for March 2016. It would be possible to compute this measure using data from both the LFS and Facebook at the time of the next census (which in the UK is scheduled for 2021). In this way, the model could be used to help nowcast migration in a timely manner, thus comparing the estimates to those of the census.

Facebook's coverage of the general population varies by age and sex (self-reported by Facebook's users). A Pew Research Center report (Pew Research Center 2018) showed that while Facebook is used across all age-groups, the numbers of younger users have been declining. Facebook has, however, noted that some younger users register with an inaccurate age (U.S. SEC 2019, 2020). In addition to the age composition of Facebook users, we should consider the coverage differences between men and women. Fatehkia et al. (2018) and Garcia et al. (2018) explored patterns in the use of Facebook to describe the digital gender gap that exists even in developed countries. While the gap is growing smaller, there are still more men than women on

Facebook (Fatehkia et al. 2018). Including an age and sex disaggregation is a further step that we leave for future research.

Traditionally, demographic methods have relied on approaches like the basic demographic balancing equation, in which the terms have to add up. That may not be necessary, however, when the underlying data have different types of biases. At the same time, more and more data sources that contain important signals of change (as well as biases) are becoming available. This study contributes to the demographic literature by proposing an approach to studying migration that is able to combine and make sense of new and different data sources in a way that builds on classic demographic approaches, while repurposing them within a Bayesian statistical framework. ■

Acknowledgments Francesco Rampazzo conducted most of his research for this article while a Ph.D. student at the Department of Social Statistics and Demography at the University of Southampton with a scholarship from the Economic and Social Research Council (South Coast Doctoral Training Partnership, Project ES/P000673/1). He is now funded by a Leverhulme Trust Grant for the Leverhulme Centre for Demographic Science. We would like to thank the members of the Digital and Computational Laboratory of the Max Planck Institute for Demographic Research for their comments on a first draft of this paper, especially Emanuele Del Fava, Sofía Gil Clavel, André Grow, Daniela Negraia, and Tom Theile. In addition, we wish to thank Jason Hilton for all his help in R and JAGS, and for his comments on this paper.

References

Alexander, M., Polimis, K., & Zagheni, E. (2019). The impact of Hurricane Maria on out-migration from Puerto Rico: Evidence from Facebook data. *Population and Development Review, 45,* 617–630.

Alexander, M., Polimis, K., & Zagheni, E. (2020). Combining social media and survey data to nowcast migrant stocks in the United States. *Population Research and Policy Review.* Advance online publication. *39.* https://doi.org/10.1007/s11113-020-09599-3

Araujo, M., Mejova, Y., Weber, I., & Benevenuto, F. (2017). Using Facebook ads audiences for global lifestyle disease surveillance: Promises and limitations. In P. Boldi (Ed.), *WebSci '17: Proceedings of the 2017 ACM on web science conference* (pp. 253–257). New York, NY: Association for Computing Machinery.

Aref, S., Zagheni, E., & West, J. (2019). The demography of the peripatetic researcher: Evidence on highly mobile scholars from the web of science. In I. Weber, K. M. Darwish, C. Wagner, E. Zagheni, L. Nelson, S. Aref, & F. Flöck (Eds.), *Social informatics: Lecture notes in computer science* (pp. 50–65). Cham, Switzerland: Springer International Publishing.

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,* 116–122.

Bijak, J. (2010). *Forecasting international migration in Europe: A Bayesian view*. Dordrecht, the Netherlands: Springer Science & Business Media.

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

Blumenstock, J. E. (2012). Inferring patterns of internal migration from mobile phone call records: Evidence from Rwanda. *Information Technology for Development, 18,* 107–125.

Cesare, N., Lee, H., McCormick, T., Spiro, E., & Zagheni, E. (2018). Promises and pitfalls of using digital traces for demographic research. *Demography, 55,* 1979–1999.

Champion, T., & Falkingham, J. (2016). *Population change in the United Kingdom*. London, UK: Rowman & Littlefield International.

Coleman, D. (1983). Some problems of data for the demographic study of immigration and of immigrant and minority populations in Britain. *Ethnic and Racial Studies, 6,* 103–110.

Cooksey, B. (2014). *An introduction to APIs*. Retrieved from https://zapier.com/learn/apis/

Del Fava, E., Wiśniowski, A., & Zagheni, E. (2019). *Modelling international migration flows by integrating multiple data sources*. SocArXiv. https://doi.org/10.31235/osf.io/cma5h

Disney, G. (2015). *Model-based estimates of UK immigration* (Doctoral dissertation). Department of Social and Human Sciences, University of Southampton, Southampton, UK.

European Parliament and Council of the European Union. (2007). *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 and repealing Council Regulation (EEC) No 311/76 on the compilation of statistics on foreign workers* (No. 862/2007). Retrieved from https://www.refworld.org/docid/48abd548d.html

Fatehkia, M., Kashyap, R., & Weber, I. (2018). Using Facebook ad data to track the global digital gender gap. *World Development, 107,* 189–209.

Fiorio, L., Zagheni, E., Abel, G., Hill, J., Pestre, G., Letouzé, E., & Cai, J. (2021). Analyzing the effect of time in migration measurement using georeferenced digital trace data. *Demography, 58,* 51–74.

Garcia, D., Kassa, Y. M., Cuevas, A., Cebrian, M., Moro, E., Rahwan, I., & Cuevas, R. (2018). Analyzing gender inequality through large-scale Facebook advertising data. *Proceedings of the National Academy of Sciences, 115,* 6958–6963.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Gendronneau, C., Wiśniowski, A., Yildiz, D., Zagheni, E., Fiorio, L., Hsiao, Y., . . . Hoorens, S. (2019). *Measuring labour mobility and migration using Big Data: Exploring the potential of social-media data for measuring EU mobility flows and stocks of EU movers* (Report). Brussels, Belgium: European Commission.

Hargittai, E. (2018). Potential biases in Big Data: Omitted voices on social media. *Social Science Computer Review, 38,* 10–24.

Herdağdelen, A., & Marelli, M. (2017). Social media and language processing: How Facebook and Twitter provide the best frequency estimates for studying word recognition. *Cognitive Science, 41,* 976–995.

Kupiszewska, D., Kupiszewski, M., Martí, M., & Ródenas, C. (2010). *Possibilities and limitations of comparative quantitative research on international migration flows* (PROMINSTAT Working Paper No. 4). Luxembourg: European Commission, DG Research Sixth Framework Programme, Priority 8.

Kupiszewska, D., & Nowok, B. (2008). Comparability of statistics on international migration flows in the European Union. In J. Raymer & F. Willekens (Eds.), *International migration in Europe: Data, models and estimates* (pp. 41–71). West Sussex, UK: John Wiley & Sons.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google flu: Traps in Big Data analysis. *Science, 343,* 1203–1205.

Martí, M., & Ródenas, C. (2007). Migration estimation based on the Labour Force Survey: An EU-15 perspective. *International Migration Review, 41,* 101–126.

Monti, A., Drefahl, S., Mussino, E., & Härkönen, J. (2019). Over-coverage in population registers leads to bias in demographic estimates. *Population Studies, 74,* 451–469.

ONS. (2018a). *Labour Force Survey—User guidance* (Technical report). London, UK: Office for National Statistics.

ONS. (2018b). *Migration statistics transformation update: May 2018* (Report). London, UK: Office for National Statistics. Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/migrationstatisticstransformationupdate/2018-05-24

ONS. (2019a, August 21). *Statement from the ONS on the reclassification of international migration statistics* [Media statement]. London, UK: Office for National Statistics. Retrieved from https://www.ons.gov.uk/news/statementsandletters/statementfromtheonsonthereclassificationofinternationalmigrationstatistics

ONS. (2019b). *Understanding different migration data sources: August progress report*. London, UK: Office for National Statistics. Retrieved from https://www.ons.gov.uk/releases/understandingdifferentmigrationdatasourcesaugustprogressreport

ONS. (2019c). *Update on our population and migration statistics transformation journey: A research engagement report*. London, UK: Office for National Statistics. Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/updateonourpopulationandmigrationstatisticstransformationjourneyaresearchengagementreport/2019-01-30

ONS. (2020). Population and migration statistics system transformation—Overview. London, UK: Office for National Statistics. Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/transformationofthepopulationandmigrationstatisticssystemoverview/2019-06-21

Pew Research Center. (2018). *Social media use 2018: Demographics and statistics* (Report). Retrieved from http://www.pewinternet.org/2018/03/01/social-media-use-in-2018/

Plummer, M., Stukalov, A., & Denwood, M. (2016). *Package 'rjags'*. Retrieved from https://cran.r-project.org/web/packages/rjags/index.html

Pötzschke, S., & Braun, M. (2017). Migrant sampling using Facebook advertisements: A case study of Polish migrants in four European countries. *Social Science Computer Review, 35,* 633–653.

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.

Rendall, M. S., Tomassini, C., & Elliot, D. J. (2003). Estimation of annual international migration from the Labour Force Surveys of the United Kingdom and the continental European Union. *Statistical Journal of the United Nations Economic Commission for Europe, 20,* 219–234.

Rosenzweig, L., Bergquist, P., Pham, K. H., Rampazzo, F., & Mildenberger, M. (2020). *Survey sampling in the Global South using Facebook advertisements*. SocArXiv. https://doi.org/10.31235/osf.io/dka8f

Sloan, L., & Quan-Haase, A. (Eds.). (2017). *The SAGE handbook of social media research methods*. London, UK: Sage Publications.

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,* e0224134. https://doi.org/10.1371/journal.pone.0224134

State, B., Rodriguez, M., Helbing, D., & Zagheni, E. (2014). Migration of professionals to the U.S. In L. M. Aiello & D. McFarland (Eds.), *Social informatics* (Vol. 8851, pp. 531–543). Cham, Switzerland: Springer International Publishing.

United Nations. (1998). *Recommendations on statistics of international migration: Revision 1* (Statistical Papers, Series M, No. 58, Rev. 1). New York, NY: United Nations, Department of Economic and Social Affairs, Statistics Division.

U.S. SEC. (2019). *Facebook, Inc.: 2018 annual report, form 10-K*. Retrieved from https://www.sec.gov/Archives/edgar/data/1326801/000132680119000009/fb-12312018x10k.htm

U.S. SEC. (2020). *Facebook, Inc.: 2019 annual report, form 10-K*. Retrieved from https://sec.report/Document/0001326801-20-000013/fb-12312019x10k.htm

Willekens, F. (1994). Monitoring international migration flows in Europe: Towards a statistical data base combining data from different sources. *European Journal of Population, 10,* 1–42.

Willekens, F. (2019). Evidence-based monitoring of international migration flows in Europe. *Journal of Official Statistics, 35,* 231–277.

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.

Zagheni, E., Polimis, K., Alexander, M., Weber, I., & Billari, F. C. (2018, April). *Combining social media data and traditional surveys to nowcast migration stocks*. Paper presented at the annual meeting of the Population Association of America, Denver, CO.

Zagheni, E., & Weber, I. (2012). You are where you e-mail: Using e-mail data to estimate international migration rates. In N. Contractor (Ed.), *WebSci '12: Proceedings of the 4th annual ACM web science conference* (pp. 348–351). New York, NY: Association for Computing Machinery.

Zagheni, E., Weber, I., & Gummadi, K. (2017). Leveraging Facebook's advertising platform to monitor stocks of migrants. *Population and Development Review, 43,* 721–734.

Francesco Rampazzo (corresponding author)
francesco.rampazzo@sbs.ox.ac.uk

Rampazzo • Saïd Business School, Leverhulme Centre for Demographic Science, and Nuffield College, University of Oxford, Oxford, UK; Centre for Population Change, University of Southampton, Southampton, UK; https://orcid.org/0000-0002-5071-7048

Bijak • Department of Social Statistics and Demography, University of Southampton, Southampton, UK; https://orcid.org/0000-0002-2563-5040

Vitali • Department of Sociology and Social Research, University of Trento, Trento, Italy; https://orcid.org/0000-0003-0029-9447

Weber • Qatar Computing Research Institute, Doha, Qatar; https://orcid.org/0000-0003-4169-2579

Zagheni • Max Planck Institute for Demographic Research, Rostock, Germany; https://orcid.org/0000-0002-7660-8368