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

Francesco Rampazzo[†1], Jakub Bijak [‡2], Agnese Vitali[§3], Ingmar Weber[¶4], and Emilio Zagheni[‖5]

[1]Saïd Business School, Leverhulme Centre for Demographic Science, and Nuffield College, University of Oxford
[2]Centre for Population Change, University of Southampton
[3]Max Planck Institute for Demographic Research
[4]University of Trento
[5]Qatar Computing Research Institute

*Acknowledgement: FR conducted most of his research for this article as a PhD student at the Department of Social Statistics and Demography of the University of Southampton with a scholarship from the Economic and Social Research Council (South Coast Doctoral Training Partnership; Project: ES/P000673/1). FR 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.

[†]Corresponding author. Email: francesco.rampazzo@sbs.ox.ac.uk
[‡]Email: j.bijak@soton.ac.uk
[§]Email: agnese.vitali@unitn.it
[¶]Email: iweber@hbku.edu.qa
[‖]Email: zagheni@demogr.mpg.de

Abstract

An accurate estimation of international migration is hampered by a lack of timely and comprehensive data, with different definitions and measures of migration adopted by different countries. Thus, 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 in the Integrated Model of European Migration is used to combine data from the Labour Force Survey (LFS) and the Facebook Advertising Platform in order to study the number of European migrants in the UK, aiming to produce more accurate estimates 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 also discussed. We cannot necessarily trust any individual source, but combining them through modelling offers valuable insights.

Keywords: Migration, Facebook, Bayesian

1 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, are barriers 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 paper, we aim to improve estimates by complementing survey data with social media data. This is important as, 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 aim is to demonstrate how such a model can produce a more accurate estimate of European migration. The UK is used as an example in this study as it is a Western country for which the migration data is of poor quality.

In this paper we use the Integrated Model of European Migration (IMEM), 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 datasets 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) have 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 which assesses the limitations of the available datasets in terms of the definition of migrants used. Assessments of the bias and the accuracy of these datasets are used to create appropriate prior distributions in order to adjust for the identified data issues.

At the same time, a new strand of research has emerged recently 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 in the lack of accessibility, transparency, and representativeness. Drawing on data from the Facebook Advertising Platform and the LFS, we investigate in this paper 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 (Zagheni et al. 2018; Alexander et al. 2019, 2020). However, in this paper, 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 due to the EU's "freedom of movement". At least until December 2020, there is no requirement for EU migrants in the UK to register their residence. Thus, up to now, survey data has been used to estimate the stock of migrants from the EU. The aim of our paper is to complement these existing, but incomplete, official estimates of migrant stocks by analysing digital trace data. As an illustration, an estimate of the total number of EU migrants for 2018 and 2019 is produced.

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, emphasising 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). Secondly, 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 in 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, combining 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). This paper 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?

2 Data

2.1 Traditional data and their limitations

A gold standard for migration estimates does not yet exist. In fact, Swedish register data, long considered as the gold standard among demographic datasets, 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 harmonise data sources on migration worldwide (UN 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 three to twelve 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, this data has at least three limitations (Willekens 1994, 2019). First is that census data is collected every ten years, and so they do not provide a timely picture of migration. Secondly, the census records immigrants living in the country, but does not account for the emigrants that have left the country. And lastly, 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 (LFS). 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 which aims to estimate labour 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 this 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 is 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 previous studies (Coleman 1983; Rendall et al. 2003; Kupiszewska and Nowok 2008; Kupiszewska et al. 2010). In addition, the ONS has recently started a work programme 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).

2.2 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 us, as the client, to the server where the data we are interested in is 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 has been used to estimate international migration rates (Zagheni and Weber 2012). This data is cheap, because it is collected by repurposing datasets originally intended for advertising. Thus, by relying on this 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 to be classified as a migrant. Thus, the definition of migrant is still not harmonised 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 modelling. In addition, this data can be augmented through data from dedicated surveys of populations who 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 datasets 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 US (State et al. 2014), while data from the Web of Science has been used to follow trends and patterns of international migration among scholars (Aref et al. 2019). However, these sources do not provide data that is representative of the entire population. Hargittai (2018) analysed the potential bias of different platforms in the United States of America (US), 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 US 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 paper, we focus on Facebook Advertising Platform data. Facebook provides advertisers with information on its users, including on each user's age, sex, level of education, and language. For this reason, Facebook has

been described as a biased *digital census* (Zagheni et al. 2017; Cesare et al. 2018). Facebook's main business is advertising, and the data provided on the Facebook Advertising Platform is 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 Facebook. In this study, we are repurposing data from this advertising platform for demographic research.

Facebook defines the variable that is 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 by Zagheni et al. (2017), where it was compared to data from the American Community Survey. Until December 2018, the variable was called as *"Expat from country x"*, showing that the wording of Facebook's definition of migrant has changed over time. However, Facebook's 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" based on 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". They 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 US Securities and Exchange Commission 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"* (US SEC 2018, 2019). In the current paper, 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 US Securities and Exchange Commission document (US SEC 2018, 2019) 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 South East 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 1000 to prevent the targeting of small groups of individuals.

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

[2] https://developers.facebook.com

Through the Facebook Marketing API, we included in the current study all Facebook users in an aggregated and anonymised format.

2.3 Comparison between LFS data and Facebook data

In this paper, 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, the Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, the Netherlands, Poland, Portugal, Romania, Slovakia, Spain, and Sweden. Malta and Luxembourg have been excluded because of their small size, while Bulgaria and Croatia have been excluded because Facebook does not provide estimates of expat numbers for them, and Estonia and Slovenia have been excluded for missing values in the data used as covariates. 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 was 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. Looking at the figure, we can see a correlation between the Facebook migrant variable and the Facebook language variable for many countries. The correlation between the Facebook "expat" variable and the Facebook language variable 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;
- 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 because of an algorithm change that affected the Facebook estimates. In Figure B2 in the Supplementary Materials, we highlight the shift that happened in the middle of March 2019, which led to an average change in the estimates of 48%.

2.4 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

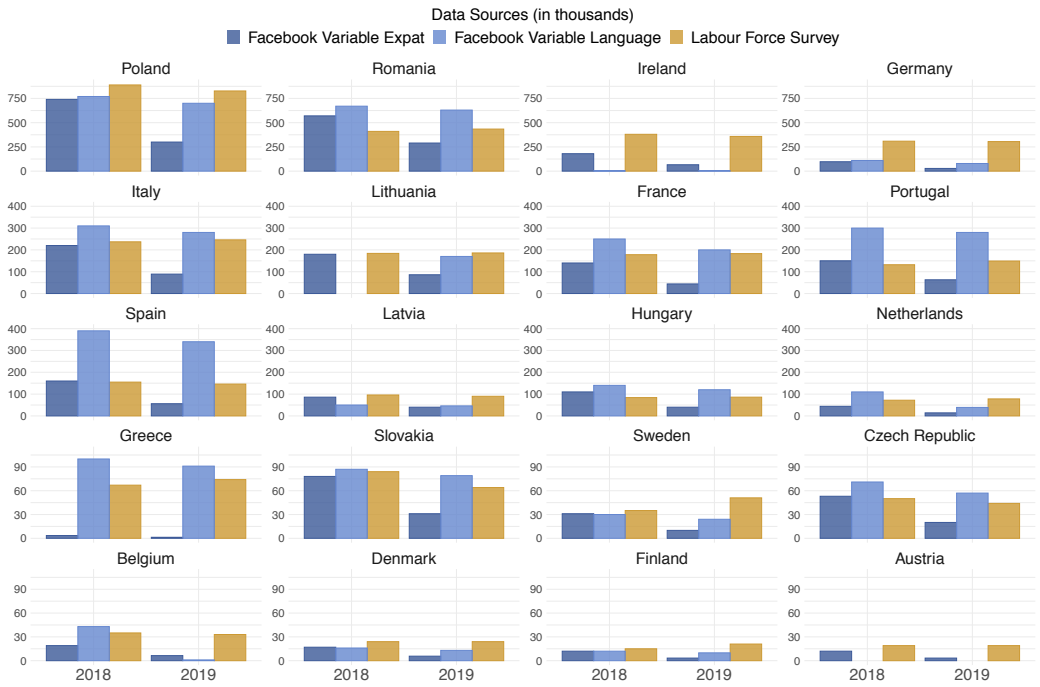

Figure 1: Facebook's aggregated estimates for the "expat" and language variables and Labour Force Survey data of migrant stocks from 20 European countries of origin in 2018 and 2019.

Eurostat, together with the Eurostat estimates of unemployment and Gross Domestic Product (GDP) per capita. The population data is used for the analysis for the years 2018 and 2019, while the other two datasets are used for the analysis for the years 2017 and 2018.

Data from the UK settled and pre-settled status scheme is added to make an additional comparison. This scheme allows European migrants already residing in the UK to apply for pre-settled status if they have been living in the UK for less than five years, and for settled status if they have been living in the UK 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 finalised. The data represents an estimate for the total number of applications, and includes data from 28^{th} August 2018 to 31^{st} December 2019.

3 Methodology

3.1 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. In this study, the 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; while 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 (Raymer et al. 2013; Disney 2015):

- **Definition**: how closely does the international migrant measure match the UN'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 presented as being survey data, digital traces, and migration theory covariates for the TBM. The data assessment is followed by a summary

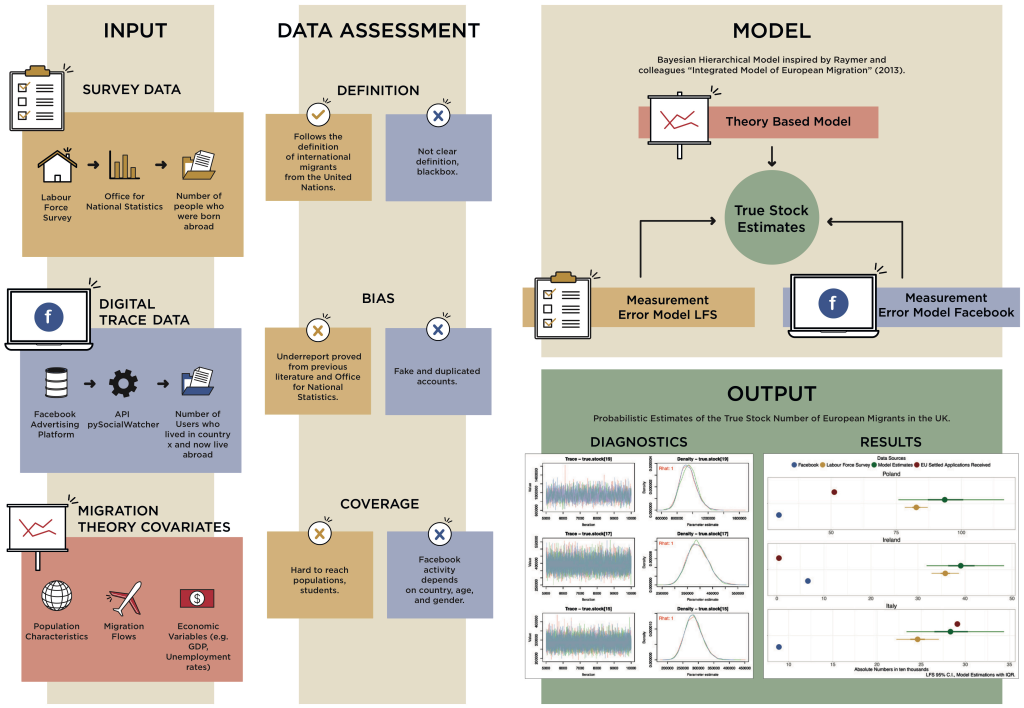

Figure 2: Diagram describing the structure of the model.

of the limitations of the data in terms of definition, bias, and coverage. In the model box, the true stock at the centre 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_{ijt}^k, 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 is used to measure the European migrants stock (z^k). The year, t, in this case is 2018 and 2019. The datasets used can thus be described in the form of matrices Z^F (Eq. 1) for Facebook, and Z^L (Eq. 2) for the LFS. The model borrows strength across the two years.

$$Z^F = \begin{pmatrix} z^F_{11t} & z^F_{12t} & \cdots & z^F_{1Jt} \\ z^F_{21t} & z^F_{22t} & \cdots & z^F_{2Jt} \\ \vdots & \vdots & \ddots & \vdots \\ z^F_{i1t} & z^F_{I2t} & \cdots & z^F_{IJt} \end{pmatrix} \tag{1}$$

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

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

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

The value of z^k_{ijt} 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:

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

Figure 3 illustrates the hierarchical structure of the model. In the next section, the model is explained in detail. 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.

3.2 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:

$$\log \mu_{ijt}^{k} = \log y_{ijt} + \delta^{k} + \beta^{k} + \chi_{ij}^{k} + \xi_{ijt}^{k} + \lambda_{ijt}^{k} + \epsilon_{ijt}^{k} \tag{5}$$

The equation is composed of five terms, δ^{k}, β^{k}, χ_{ij}^{k}, ξ_{ijt}^{k}, and λ_{ijt}^{k} which are used to convert the data from Facebook and the LFS to comply with the UN's definition of an international migrant, and to reduce the underestimation linked to the bias or coverage of the data. The first parameter, δ^{F}, captures the differences in relation to the definition of migrants. The bias in the data is captured by β^{F}, while the coverage of the Facebook data is considered in χ_{ij}^{F}. The parameter ξ_{ijt}^{F} deflates the Facebook estimates of 2018 by the algorithm change that happened in 2019. The parameter λ_{ijt}^{k} inflates the Facebook estimates with knowledge provided by the Facebook estimates of people speaking a certain language. The term ϵ_{ijt}^{k} is the error term with

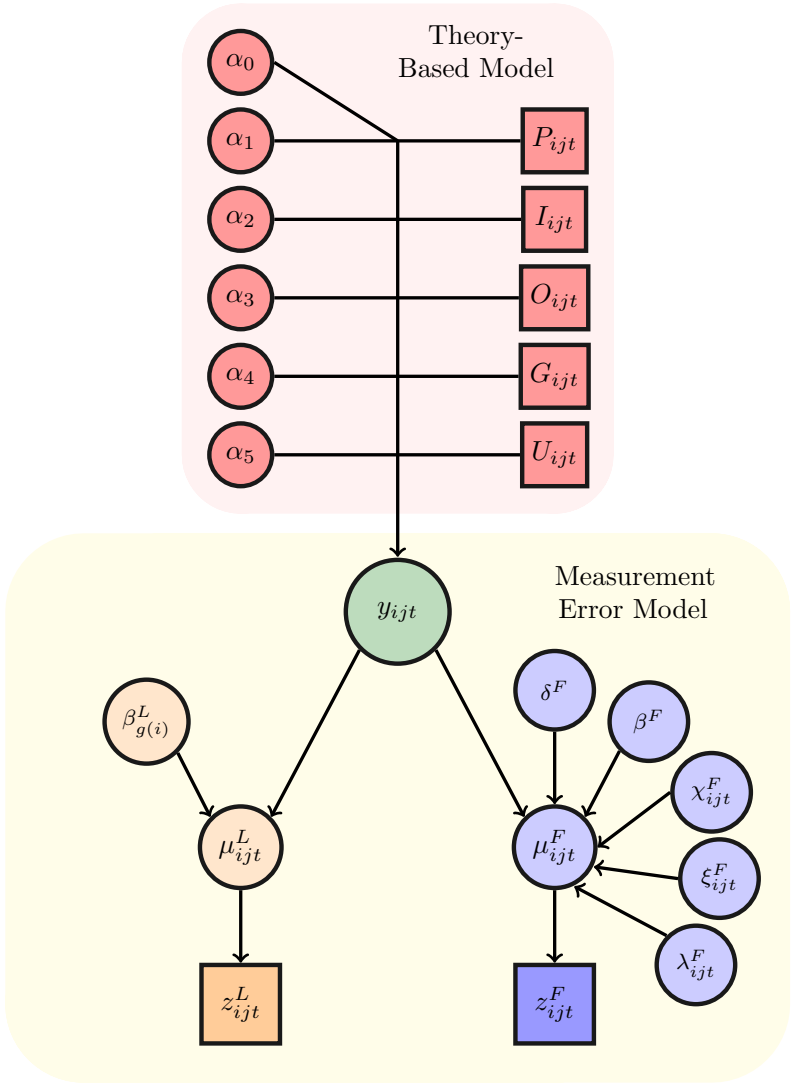

Figure 3: Graphical representation of the adapted IMEM (diagram inspired by Raymer et al. (2013, p. 804)). The hyperparameters are not shown for greater clarity of presentation. Indices: i, sending country; j, sex; t, time. Square nodes represent reported data (z^L_{ijt}, z^F_{ijt}) and covariates. Circle nodes represent parameters for the migration model (see Section 3.2) and the measurement model (see Section 3.3).

normal distribution $N(0, \tau_{ijt})$, the precision τ_{ijt} has Gamma distribution $G(100, 1)$, (where 100 is the shape parameter and one is the rate parameter) which has a mean equal to 100 and precision equal to 1 (e.g variance equal to 100). The term ϵ^k_{ijt} is the error term with normal distribution $N(0, \tau_{ijt})$, the precision τ_{ijt} has Gamma distribution $G(100, 1)$, with a mean equal to 100 and a precision equal to 1 (e.g variance equal to 100). Table 1 summarises the parametrisation of the model and the direction of the prior distributions.

Table 1: Table summarising the parameters in the measurement error model for the Labour Force Survey and Facebook.

Measurement Error Model			
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		$\pm$ coverage by sex in the home country
ξ	Algorithm Change		$\sim$ effect of an algorithm change in 2019
λ	Language Parameter		$\sim$ Greek language dummy parameter

3.2.1 Data Assessment of the Labour Force Survey

The LFS defines a long-term international migrant in the same way as the UN (ONS 2018a), and provides data on each migrant's country of birth and citizenship. For the purposes of this paper, the country of birth criterion is used because it captures individuals with a migrant background, including those who acquired citizenship through naturalisation. Since 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 have found that they underestimate migrants. Rendall et al. (2003), for example, reported that the 2001 LFS under-reported international migrants by 26% compared to 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 non-response rate of over 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 27^{th} March 2011. The data is aggregated for England and Wales only. It reveals the relative percentage change between the LFS and the census. The relative percentage change gives a sense of the bias between the LFS and the census. 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 issue suggests the LFS Measurement Error Equation (6) to be:

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

As for this assessment, the LFS data is deflated only by 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 UN standard. The literature (Rendall et al. 2003; Kupiszewska et al. 2010; Martí and Ródenas 2007) 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, at 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, as well as assessment from the ONS and 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:

1. **Low** - Bias at 4%: Austria, Belgium, Czech Republic, Latvia, Sweden;
2. **Medium** - Bias at 12%: France, Germany, Greece, Hungary, Lithuania;
3. **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

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

	LFS January - March 2011	Census March 2011	Relative Percentage Change
Austria	19000	19087	-0.46
Belgium	28000	25472	9.03
Czech Republic	37000	37150	-0.41
Denmark	18000	21445	-19.14
Finland	10000	12149	-21.49
France	134000	129804	3.13
Germany	279000	273564	1.95
Greece	33000	34389	-4.21
Hungary	44000	48308	-9.79
Ireland	353000	407357	-15.40
Italy	121000	134619	-11.26
Latvia	57000	54669	4.09
Lithuania	115000	97083	15.58
Netherlands	52000	59081	-13.62
Poland	572000	579121	-1.24
Portugal	83000	88161	-6.22
Romania	94000	79687	15.23
Slovakia	52000	57824	-11.20
Spain	69000	79184	-14.76
Sweden	30000	30694	-2.31

Note: 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 is already recalibrated through 2011 census data.

$g(i)$, where:

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

The prior distribution is set to:

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

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.

3.2.2 Data assessment of the Facebook Advertising Platform

Given the description of the Facebook data in Section 2.3, a parameter was created for both the definition, bias, and coverage of the Facebook data. 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 in order to account for fake and duplicate accounts. This value is lower than the 11% suggested by Facebook themselves, because it is assumed that the percentage of fake and duplicated accounts labelled 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:

$$\chi_{ijt} = \log\left(1 - \frac{\text{Number of Facebook Users}_{ijt}}{\text{Eurostat Population Size}_{ijt}}\right) \tag{9}$$

Additionally, the digital trace data is described as unstable. Indeed, it seems that Facebook reviewed its algorithm on expats in the middle of March 2020, 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_{ij}^{F} (Eq. 10), which aims to adjust the Facebook data for this bias caused by the change in the algorithm.

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

A parameter was used for Greece that inflates the estimates of the Facebook expat variable (Eq. 11). The Facebook expat variable 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 is not perfect, as it seems that Facebook is not accounting for Greek migrants with the migrant

variable (see also Figure A1 in the Supplementary Materials).

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

After this assessment, the Facebook Measurement Error Equation (12) is:

$$\log \mu_{ij}^{F} = \log y_{ij} + \delta^{F} + \beta^{F} + \chi_{ij}^{F} + \xi_{ij}^{F} + \lambda_{ij}^{F} + \epsilon_{ij}^{F} \tag{12}$$

3.3 Theory-based model

In this part of the model (Eq. 13), covariates that might help to explain the true stock of European migrants in the UK are introduced.

$$\log y_{ij} = \alpha_0 + \alpha_1 P_{ij} + \alpha_2 I_{ij} + \alpha_3 O_{ij} + \alpha_4 \log G_{ij} + \alpha_5 \log U_{ij} + \epsilon_{ij} \tag{13}$$

Where $\alpha = (\alpha_0, \ldots, \alpha_5)$ 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 an Gamma distribution $Gamma(100, 1)$.
The covariates used in the models for 2018 and 2019 include:

- $\boldsymbol{P}$: a normalised measure of population size in the country of origin, divided by the mean of the population in the same countries considered in the model. The data is from the latest estimates by Eurostat in 2018 and in 2019;

- $\boldsymbol{I}$: a normalised 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. The data is from the IPS in 2017 and in 2018;

- $\boldsymbol{O}$: a normalised 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. The data is from the IPS in 2017 and in 2018;

- $\boldsymbol{G}$: 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. The data is from Eurostat;

- $\boldsymbol{U}$: ratio of the unemployment rate in the European country of origin in 2017 and in 2018, divided by the unemployment rate in the UK. The data is from Eurostat.

The normalised measure of the population size is a predictor of the possible number of migrants informed by a gravity model; i.e. the larger the population, the larger the number of possible migrants. The normalised 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.

4 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. The two sets of models 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. Finally, all the estimates of the models converge. Detailed results and some diagnostic statistics are included in the Supplementary Materials.

4.1 Model for Total Numbers

Figure 4 shows data from three datasets and our estimates: the Facebook Advertising data is in blue, the LFS data is in yellow, the settled status application data is in red, and the model estimates are in green. The settled status data is used for comparison, and is not used in the analysis. LFS data is shown with a 95% confidence interval (CI), while model estimates are shown with the interquartile (IQR). The data for the two years is identified by a circle for 2018, and by a square for 2019.

There are three main messages that can be discerned from this figure. Firstly, 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 is that, while the LFS data is relatively consistent across the two

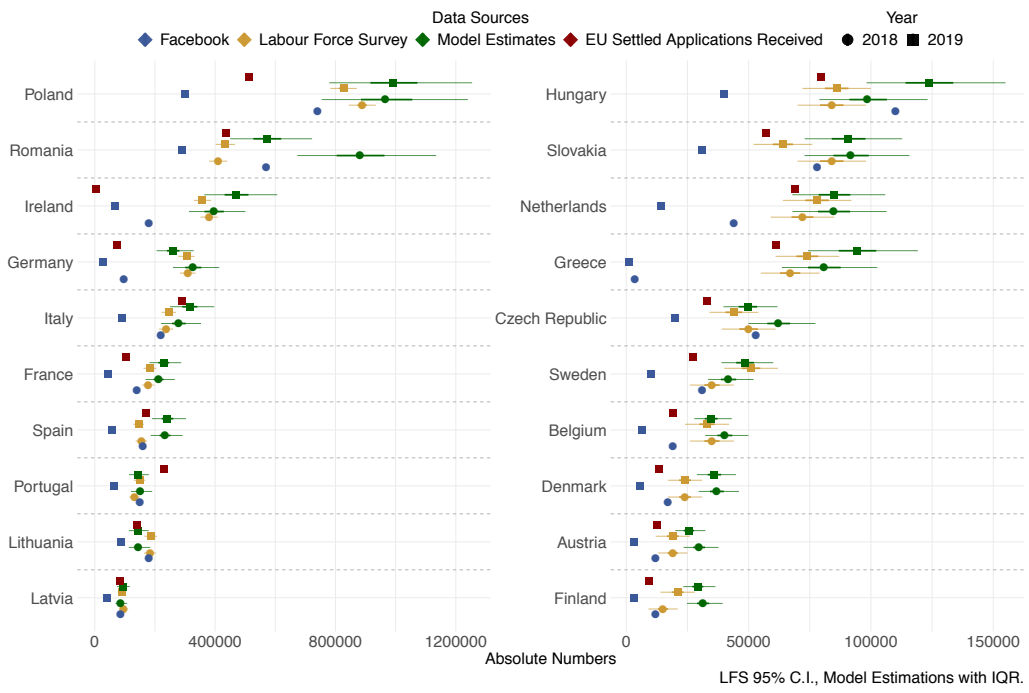

Figure 4: Comparison of Facebook, LFS, and model estimations of European migrants aged 15+ for the years 2018 and 2019.

years, a decreasing trend in the number of EU migrants in the UK is visible. Thirdly, the model estimates are higher than the LFS estimates. In some cases, the IQR range of the model estimates includes the LFS estimates.

In Figure 4 the estimates for the second group of countries are also shown. The parameter on Greece seems to be effective in bringing the estimates closer to the LFS values. In the Supplementary Materials, the posterior characteristics of the true stock estimates for all of the models and the $\hat{R}$ are reported, 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 Supplementary Materials). The algorithm for estimating all of the other parameters has converged as well.

In Table 3, a comparison of the undercounted LFS estimates with the model estimates is presented. While the ONS has estimated an undercount of 16%, the model estimates an undercount of 25% for 2018 and 20% for 2019.

Table 3: Undercount of the LFS estimates in comparison with the model estimates.

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

The undercount for 2018 has larger intervals, likely due 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.

4.2 Model disaggregated by sex

In this part of the model the estimates are disaggregated by sex. It is important to study the age and sex differences of migrants. 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 is not disaggregated by sex.

4.3 Sensitivity Analysis

Some sensitivity checks of the model are provided. First, the model was run while only including the LFS data. For the model specified in this paper, 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 included in

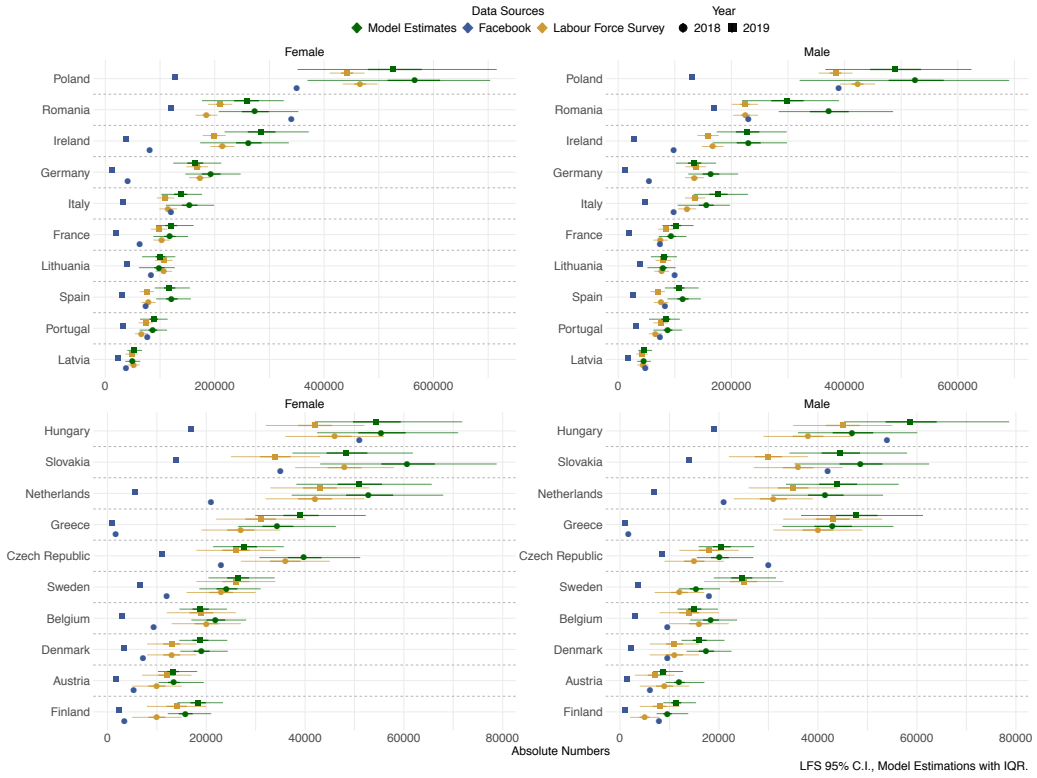

Figure 5: Comparison of Facebook, LFS, and model estimations of European migrants aged 15+ by sex for the years 2018 and 2019.

this paper, 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 Table 4 the undercount value of the new specifications of the model is reported. The undercount with no bias attributed to the Facebook estimates

Table 4: Undercount of the LFS estimates in three different models 1) the model specified only with the LFS data, 2) the model with the Facebook bias parameter set to 0%, 3) the model with the Facebook bias parameter set to 11%, 4) the model with the LFS bias parameter set to 4%, 5) the model with the LFS bias parameter set to 30%, and 6) the model with the $Gamma(1,1)$ distribution.

		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%

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 undercount with a β^F at 4% and at 11% are very similar.

The model is sensitive to the choice of the assumed bias of the LFS parameter. In Table 4 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 assumed. We also tried different specifications of the precision distribution term, which is assumed to follow a $Gamma(100, 1)$ in the presented model. In Table 4, 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.

Additionally, 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. While 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. While the estimates from our models seem to be stable to different prior distributions, the precision of those prior distributions had to be carefully chosen to to ensure model convergence, while exploring reasonable areas of the parameter space with respect to the precision parameters.

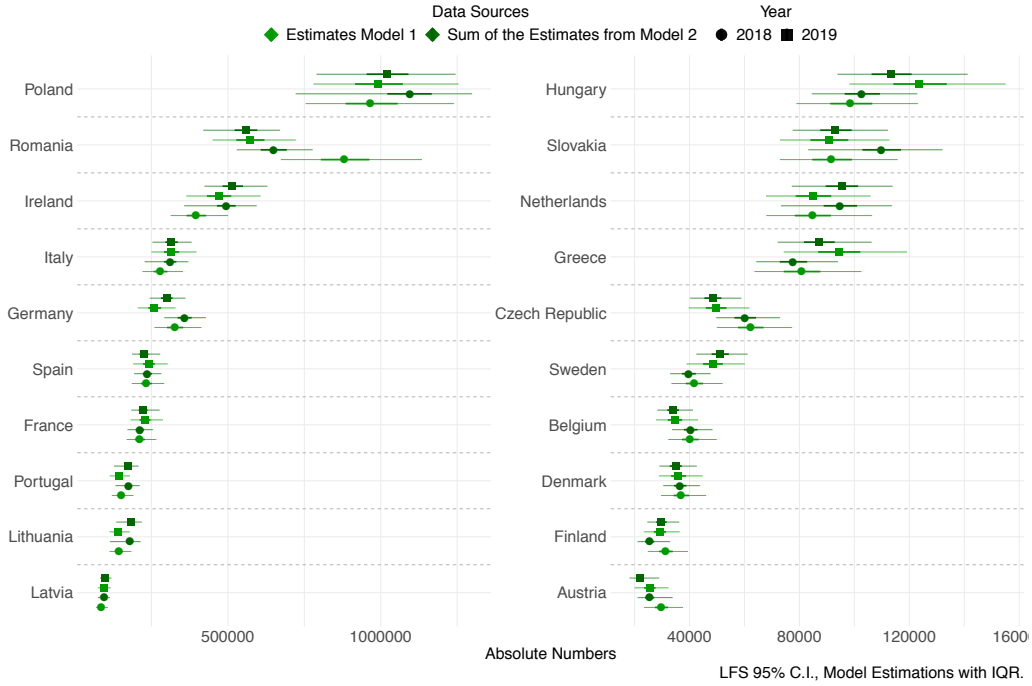

Figure 6: Comparison between estimates from the first model and the sum of female and male migrants from the second model for 2018 and 2019.

5 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 that using digital traces alone is not sufficient to generate better 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.

Including the Home Office's data related to settlement and pre-settlement 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 as 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 pre-settled 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 pre-settlement application. Almost no Irish nationals applied to the settled or pre-settled scheme due 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 were equal to the total from the first. There was uncertainty in our estimates, greatest 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 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 the users on their platform are able to speak. Furthermore, there are no details available about the algorithm change Facebook implemented in March 2019.

6 Conclusions

The overarching research question of this paper was: 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? This question has been answered by exploring the two data sources and producing a probabilistic measure of European migration. Although it has found greater uncertainty in the estimates that were already known to be biased, this research contributes to the "*learning process*" hoped for by Willekens (1994, 2019) which can lead to the extension of this framework. The obvious next step for this research would be to expand the model to

disaggregate the estimates by age and sex.

This analysis has made three contributions to digital and computational demography. First, it has 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 such as from other advertising platforms like Instagram, Snapchat, and LinkedIn, as well as from other administrative sources. Second, it has addressed the biases of both traditional and digital trace data. The use of a prior distribution has been shown to fix these issues in a probabilistic fashion. Third, it has 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 based on 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, 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 2018) showed that while Facebook is used across all age groups, the numbers of younger users on Facebook have been declining. Facebook has, however, noted that some younger users register on Facebook with an inaccurate age (US SEC 2018, 2019). 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 which 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 has 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 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.

References

Alexander, M., Polimis, K., and Zagheni, E. (2019). The Impact of Hurricane Maria on Out-migration from Puerto Rico: Evidence from Facebook Data. *Population and Development Review*, 45(3):617–630.

Alexander, M., Polimis, K., and Zagheni, E. (2020). Combining social media and survey data to nowcast migrant stocks in the United States. *arXiv:2003.02895 [stat]*.

Araujo, M., Mejova, Y., Weber, I., and Benevenuto, F. (2017). Using Facebook Ads Audiences for Global Lifestyle Disease Surveillance: Promises and

Limitations. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, pages 253–257, New York, NY, USA. ACM.

Aref, S., Zagheni, E., and West, J. (2019). The Demography of the Peripatetic Researcher: Evidence on Highly Mobile Scholars from the Web of Science. In Weber, I., Darwish, K. M., Wagner, C., Zagheni, E., Nelson, L., Aref, S., and Flöck, F., editors, *Social Informatics*, Lecture Notes in Computer Science, pages 50–65, Cham. Springer International Publishing.

Azose, J. J. and 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. (2010). *Forecasting International Migration in Europe: A Bayesian View*. Springer Science & Business Media.

Bilsborrow, R. E., Hugo, G., Zlotnik, H., and Oberai, A. S. (1997). *International Migration Statistics: Guidelines for Improving Data Collection Systems*. International Labour Organization.

Blumenstock, J. E. (2012). Inferring patterns of internal migration from mobile phone call records: Evidence from Rwanda. *Information Technology for Development*, 18(2):107–125.

Cesare, N., Lee, H., McCormick, T., Spiro, E., and Zagheni, E. (2018). Promises and Pitfalls of Using Digital Traces for Demographic Research. *Demography*, 55(5):1979–1999.

Champion, T. and Falkingham, J. (2016). *Population Change in the United Kingdom*. Rowman & Littlefield.

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(1):103–110.

Cooksey, B. (2014). An Introduction to APIs. https://zapier.com/learn/apis/.

Del Fava, E., Wiśniowski, A., and Zagheni, E. (2019). Modelling International Migration Flows by Integrating Multiple Data Sources. Preprint, SocArXiv.

Disney, G. (2015). *Model-Based Estimates of UK Immigration*. PhD thesis, University of Southampton.

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 (Text with EEA relevance).

Fatehkia, M., Kashyap, R., and 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., and Cai, J. (2021). Analyzing the Effect of Time in Migration Measurement Using Georeferenced Digital Trace Data. *Demography*, (8917630).

Garcia, D., Kassa, Y. M., Cuevas, A., Cebrian, M., Moro, E., Rahwan, I., and Cuevas, R. (2018). Analyzing gender inequality through large-scale Facebook advertising data. *Proceedings of the National Academy of Sciences*, 115(27):6958–6963.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition*. CRC Press.

Gendronneau, C., Wiśniowski, A., Yildiz, D., Zagheni, E., Fiorio, L., Hsiao, Y., Stepanek, M., Weber, I., Abel, G., and 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*. Publications Office of the European Union.

Hargittai, E. (2018). Potential Biases in Big Data: Omitted Voices on Social Media. *Social Science Computer Review*, page 089443931878832.

Herdağdelen, A. and Marelli, M. (2017). Social Media and Language Processing: How Facebook and Twitter Provide the Best Frequency Estimates for Studying Word Recognition. *Cognitive Science*, 41(4):976–995.

Kupiszewska, D., Kupiszewski, M., Martí, M., and Ródenas, C. (2010). Possibilities and limitations of comparative quantitative research on international migration flows. Technical Report Promoting Comparative Quantitative Research in the Field of Migration and Integration in Europe (PROMINSTAT), Project funded by the European Commission, DG Research Sixth Framework Programme, Priority 8.

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

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science*, 343(6176):1203–1205.

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

Monti, A., Drefahl, S., Mussino, E., and Härkönen, J. (2019). Over-coverage in population registers leads to bias in demographic estimates. *Population Studies*, 0(0):1–19.

ONS (2018a). Labour Force Survey – user guidance - Office for National Statistics. Technical report, Office for National Statistics.

ONS (2018b). Migration statistics transformation update - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/migrationstatisticstransformationupdate/2018-05-24.

ONS (2019a). Statement from the ONS on the reclassification of international migration statistics - Office for National Statistics. https://www.ons.gov.uk/news/statementsandletters/statementfromtheonsonthereclassificationofinternationalmigrationstatistics.

ONS (2019b). Understanding different migration data sources: August progress report - Office for National Statistics. https://www.ons.gov.uk/releases/understandingdifferentmigrationdatasourcesaugustprogressreport.

ONS (2019c). Update on our population and migration statistics transformation journey: A research engagement report - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/

populationandmigration/internationalmigration/articles/updateonourpopulationandmigrationstatisticstransformationjourneyaresearchengagem 2019-01-30.

ONS (2020). Population and migration statistics system transformation – overview - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/transformationofthepopulationandmigrationstatisticssystemoverview/2019-06-21.

Pew Research (2018). Social Media Use 2018: Demographics and Statistics | Pew Research Center. http://www.pewinternet.org/2018/03/01/social-media-use-in-2018/.

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

Pötzschke, S. and Braun, M. (2017). Migrant Sampling Using Facebook Advertisements: A Case Study of Polish Migrants in Four European Countries. *Social Science Computer Review*, 35(5):633–653.

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

Rendall, M. S., Tomassini, C., and 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(3,4):219–234.

Rosenzweig, L., Bergquist, P., Pham, K. H., Rampazzo, F., and Mildenberger, M. (2020). Survey sampling in the Global South using Facebook advertisements. Technical report, SocArXiv.

Sloan, L. and Quan-Haase, A. (2017). *The SAGE Handbook of Social Media Research Methods*. SAGE.

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

State, B., Rodriguez, M., Helbing, D., and Zagheni, E. (2014). Migration of Professionals to the U.S. In Aiello, L. M. and McFarland, D., editors, *Social Informatics*, volume 8851, pages 531–543. Springer International Publishing, Cham.

UN (1998). *Recommendations on Statistics of International Migration*. Number no. 58, rev. 1 in Statistical Papers. Series M. United Nations, New York.

US SEC (2018). Facebook Inc. 2018 Annual Report 10-K. https://www.sec.gov/Archives/edgar/data/1326801/000132680119000009/fb-12312018x10k.htm.

US SEC (2019). Facebook Inc. 2019 Annual Report 10-K. 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):1–42.

Willekens, F. (2019). Evidence-Based Monitoring of International Migration Flows in Europe. *Journal of Official Statistics*, 35(1):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(1):185–202.

Zagheni, E., Polimis, K., Alexander, M., Weber, I., and Billari, F. C. (2018). Combining Social Media Data and Traditional Surveys to Nowcast Migration Stocks. In *Annual Meeting of the Population Association of America*.

Zagheni, E. and Weber, I. (2012). You Are Where You e-Mail: Using e-Mail Data to Estimate International Migration Rates. In *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12, pages 348–351, New York, NY, USA. ACM.

Zagheni, E., Weber, I., and Gummadi, K. (2017). Leveraging Facebook's Advertising Platform to Monitor Stocks of Migrants. *Population and Development Review*, 43(4):721–734.

Supplementary Materials

Figure A1 shows the number of Greek migrants across European countries with Eurostat data from 2018. We compare the Eurostat data with Facebook Advertising data from 2020 estimating the number of Greek migrants, *"People that used to live in Greece and now live abroad"* and the number of people speaking Greek on Facebook. The latter variable seems to approximate better the number of Greek migrants living abroad. For the majority of the countries, except the UK, the Netherlands, France, Germany, Portugal, and Spain, the variable of Greek migrants from Facebook does not account for any Greek migrants.

Figure B2 shows the shift that happened in the middle of March 2019, which led to a change in the Facebook estimates of migrants, that was country-, age-, and sex-specific.

Tables A1, B2, C3, D4 reports the posterior characteristics of the coefficients y (the true stock estimates) for models 1 and 2 respectively for the two years of analysis. The tables report $\hat{R}$ and $\hat{n}_{eff}$, which is the effective number of simulation draws (Gelman et al. 2013); it is reported as an additional measure to show the series converge.

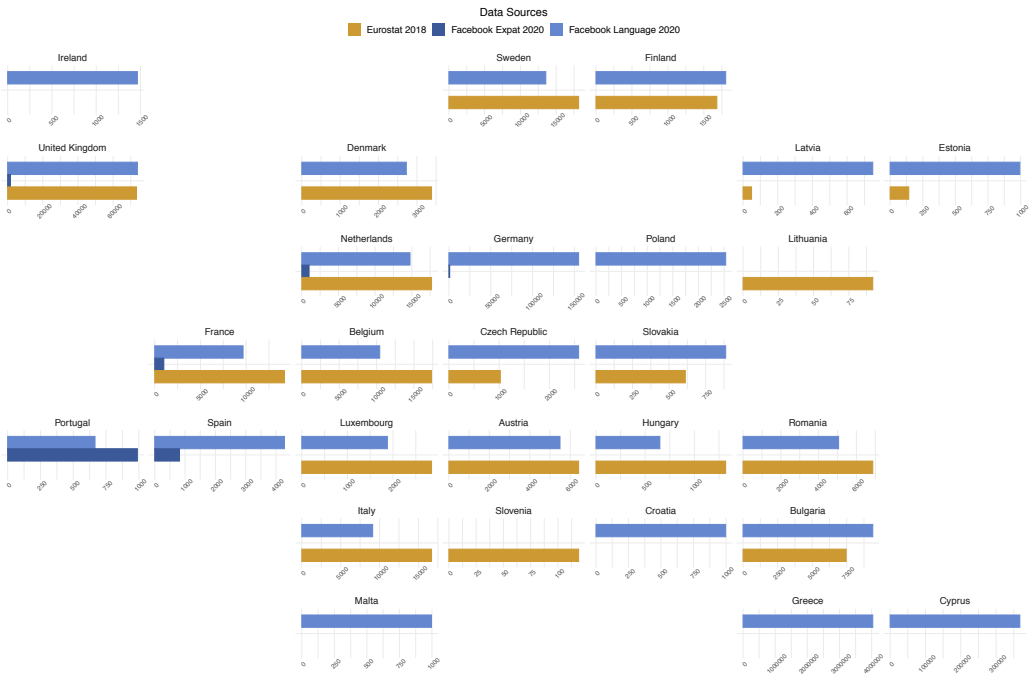

Fig. A1: The number of Greek migrants in European countries based on Facebook Advertising data and Eurostat data, and the number of Greek-speaking people on Facebook.

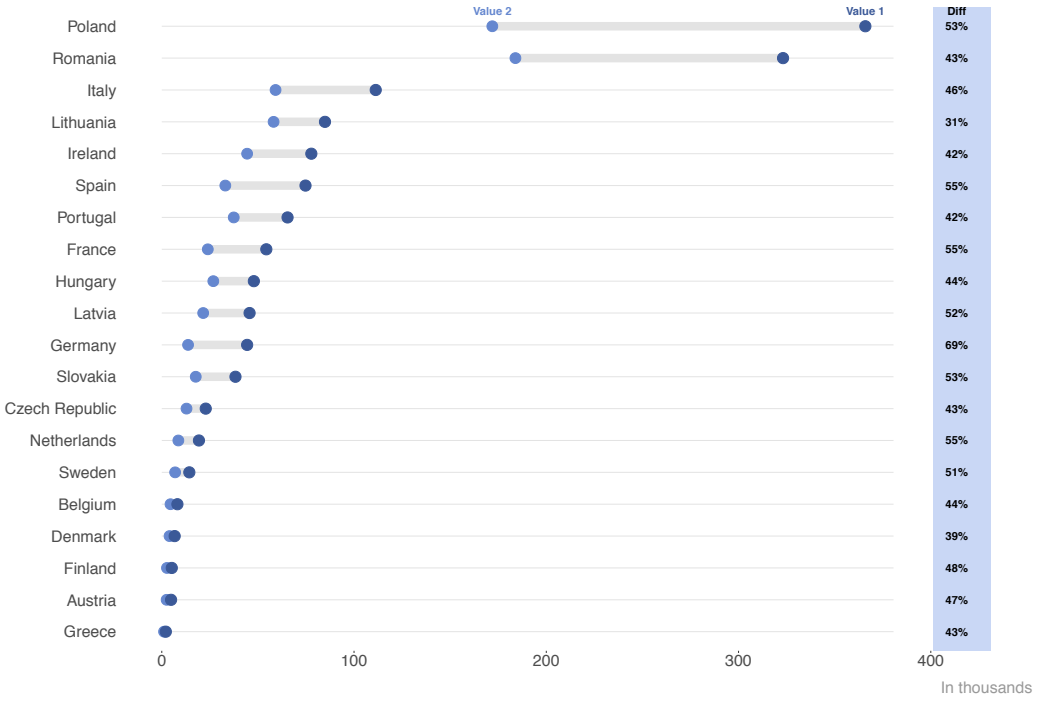

Fig. B2: Change in the Facebook algorithm. Magnitude of the decline in Facebook's estimates of European migrant stocks in the UK in the middle of March 2019 from "*value 1*" to "*value 2*".

Tab. A1: Posterior characteristics of the coefficients of the true stock estimates, y, in the first model for 2018 with $\hat{R}$ and $\hat{n}_{eff}$.

Country	2.5%	25%	50%	75%	97.5%	$\hat{R}$	$\hat{n}_{eff}$
y_{Poland}	754235	885039	965000	1055382	1239939	1.00	501
$y_{Romania}$	673195	804057	880670	962875	1134917	1.01	727
$y_{Ireland}$	312864	364849	395657	429117	500759	1.00	1652
$y_{Germany}$	259740	300986	326180	353982	413632	1.00	1808
y_{Italy}	220539	256977	278584	302263	353103	1.00	4180
y_{Spain}	185601	215405	233228	251625	291926	1.00	9009
y_{France}	168883	195798	211531	228483	265994	1.00	7752
$y_{Lithuania}$	112917	132389	143880	156557	184292	1.00	5418
$y_{Portugal}$	120372	140140	151386	163779	190961	1.00	7379
$y_{Hungary}$	78897	91221	98500	106521	123178	1.00	10746
y_{Latvia}	68940	79709	86057	92924	107735	1.00	11243
$y_{Slovakia}$	72859	84677	91594	99108	115761	1.00	9274
y_{Greece}	63629	74283	80764	87615	102582	1.00	16990
$y_{Netherlands}$	67927	78328	84728	91484	106404	1.00	10194
$y_{CzechRepublic}$	49938	57645	62146	66973	77340	1.00	16447
y_{Sweden}	33423	38668	41669	44923	52033	1.00	21890
$y_{Belgium}$	32269	37250	40173	43293	49904	1.00	22783
$y_{Denmark}$	29612	34224	36910	39855	46047	1.00	17382
$y_{Finland}$	24837	28932	31330	33885	39424	1.00	15248
$y_{Austria}$	23440	27391	29691	32192	37647	1.00	14945

Tab. B2: Posterior characteristics of the coefficients of the true stock estimates, y, in the first model for 2019 with $\hat{R}$ and $\hat{n}_{eff}$.

Country	2.5%	25%	50%	75%	97.5%	$\hat{R}$	$\hat{n}_{eff}$
y_{Poland}	779380	916113	990514	1072395	1254251	1.01	467
$y_{Romania}$	450243	527361	571965	619605	722326	1.00	1846
$y_{Ireland}$	363765	431661	469397	510745	606800	1.00	1417
y_{Italy}	249730	290940	315117	341046	397192	1.00	6175
y_{Spain}	190068	222428	241081	261064	303308	1.00	8012
y_{France}	181922	211406	228603	247112	287564	1.00	6703
$y_{Lithuania}$	112908	131401	142220	154012	179159	1.00	6229
$y_{Hungary}$	98188	114171	123593	133643	155016	1.00	8388
$y_{Germany}$	205372	239835	259900	281465	329158	1.00	2778
$y_{Portugal}$	113657	132024	142772	154320	179694	1.00	7192
y_{Latvia}	74570	86315	93229	100511	116579	1.00	15066
y_{Greece}	74272	86836	94267	102147	119193	1.00	17624
$y_{Slovakia}$	72918	83982	90602	97767	112757	1.00	11584
$y_{Netherlands}$	67854	78629	84866	91572	105893	1.00	10611
$y_{CzechRepublic}$	39718	45921	49563	53469	61799	1.00	16812
y_{Sweden}	38887	44922	48366	52147	60092	1.00	18602
$y_{Belgium}$	27785	32074	34616	37336	43119	1.00	19519
$y_{Denmark}$	28844	33294	35960	38726	44826	1.00	18468
$y_{Austria}$	20026	23495	25495	27658	32360	1.00	12320
$y_{Finland}$	23310	27001	29170	31460	36424	1.00	20909

Tab. C3: Posterior characteristics of the coefficients of the true stock estimates, y, in the second model for 2018 with $\hat{R}$ and $\hat{n}_{eff}$.

	Male							Female						
Country	2.5%	25%	50%	75%	97.5%	$\hat{R}$	$\hat{n}_{eff}$	2.5%	25%	50%	75%	97.5%	$\hat{R}$	$\hat{n}_{eff}$
y_{Poland}	320559	477960	524554	575046	690686	1.04	178	369704	516059	565957	612160	703561	1.03	139
$y_{Romania}$	283698	338753	371860	407291	485793	1.00	1915	207491	249883	273068	298882	352976	1.01	839
$y_{Ireland}$	167058	209512	230216	251800	298402	1.02	367	173389	239221	261547	285607	335446	1.01	266
y_{Italy}	105093	142617	155733	168867	197530	1.01	520	110627	139769	153586	168624	198769	1.02	509
y_{Spain}	87083	104373	114033	124460	146371	1.00	1698	92747	110951	121152	132253	156458	1.01	2065
y_{France}	88006	107389	117530	128473	151311	1.00	1129	71741	86015	93757	102104	120618	1.00	1553
$y_{Lithuania}$	51780	72912	79581	86686	101405	1.00	779	61734	89446	98364	108163	126934	1.00	356
$y_{Germany}$	110627	139769	153586	168624	198769	1.02	509	123821	149258	163671	178653	212033	1.00	1206
$y_{Hungary}$	35969	42980	46928	51155	60065	1.00	3011	42451	50775	55386	60359	71008	1.00	3348
$y_{Portugal}$	62638	79095	86845	94961	112879	1.01	1063	62178	80506	87900	95811	113192	1.01	1176
y_{Latvia}	33730	41306	45043	49155	57817	1.00	2280	36500	45481	49719	54357	63861	1.00	2441
y_{Greece}	32806	39281	42952	46886	55250	1.00	4145	26431	31348	34306	37610	46214	1.00	5264
$y_{Slovakia}$	35277	44365	48566	53032	62466	1.00	1924	43060	55427	60666	66277	78794	1.00	1529
$y_{Netherlands}$	30607	37973	41481	45192	53167	1.00	2054	37294	48306	52857	57825	67966	1.00	1369
$y_{CzechRepublic}$	15532	18400	20085	21974	26951	1.00	4766	30755	36498	39762	43331	51159	1.00	4731
y_{Sweden}	11873	14077	15375	16804	20222	1.00	7462	18591	22084	24101	26295	31000	1.00	5418
$y_{Belgium}$	14190	16797	18327	19992	23667	1.00	8612	16981	20069	21875	23829	28073	1.00	7658
$y_{Denmark}$	13451	15929	17386	19001	22541	1.00	11027	14744	17448	18998	20685	24345	1.00	8651
$y_{Austria}$	9240	10942	11964	13109	17069	1.00	985	10340	12247	13392	14676	19492	1.00	519
$y_{Finland}$	7401	8758	9578	10491	13803	1.00	848	12205	14455	15784	17250	20993	1.00	5536

Tab. D4: Posterior characteristics of the coefficients of the true stock estimates, y, in the second model for 2019 with $\hat{R}$ and $\hat{n}_{eff}$.

	Male							Female						
Country	2.5%	25%	50%	75%	97.5%	$\hat{R}$	$\hat{n}_{eff}$	2.5%	25%	50%	75%	97.5%	$\hat{R}$	$\hat{n}_{eff}$
y_{Poland}	365990	444945	488992	535045	624067	1.03	288	351423	480011	525332	578674	715557	1.03	199
$y_{Romania}$	218962	270479	297414	327576	390218	1.00	700	176715	235467	258367	280952	326206	1.00	415
$y_{Ireland}$	174020	208079	227567	249378	297762	1.00	671	218035	260421	284984	311080	372042	1.01	631
y_{Italy}	133524	161014	176434	193334	229348	1.00	2067	102507	125706	137513	149684	176917	1.01	1192
y_{Spain}	82502	98202	107286	117211	142247	1.00	3088	90227	107347	117276	128029	154496	1.00	1993
y_{France}	71741	86015	93757	102104	120618	1.00	1553	88006	107389	117530	128473	151311	1.00	1129
$y_{Lithuania}$	57967	73911	80838	88179	103855	1.00	912	67220	91676	100569	109516	128038	1.01	498
$y_{Germany}$	102159	123303	134707	146858	173111	1.00	976	124190	149989	163840	178943	212056	1.01	1434
$y_{Hungary}$	45277	53665	58552	63996	78637	1.00	2909	41996	49700	54270	59378	71828	1.00	3828
$y_{Portugal}$	54437	76528	84020	92046	109020	1.00	588	63376	80994	88461	96419	113941	1.00	1343
y_{Latvia}	35605	42064	45887	50108	60042	1.00	3134	40088	47679	52090	56891	67215	1.00	3068
y_{Greece}	36592	43647	47650	52047	61242	1.00	4928	29925	35621	38999	42752	52287	1.00	4081
$y_{Slovakia}$	34248	40740	44438	48489	58010	1.00	4997	37429	44366	48349	52609	61827	1.00	3671
$y_{Netherlands}$	33542	40242	43946	47898	56335	1.00	2879	38231	46609	50970	55572	65645	1.01	1862
$y_{CzechRepublic}$	15872	18754	20475	22373	27128	1.00	7002	21381	25403	27717	30250	35698	1.00	5447
y_{Sweden}	18943	22496	24555	26744	31529	1.00	5942	20434	24220	26349	28661	33848	1.00	5120
$y_{Belgium}$	11643	13774	15044	16441	19799	1.00	8000	14556	17217	18789	20506	24204	1.00	8867
$y_{Denmark}$	12421	14699	16047	17528	21122	1.00	6850	14527	17184	18744	20446	24265	1.00	9191
$y_{Austria}$	6692	7926	8665	9498	12764	1.00	625	10211	12098	13212	14464	18170	1.00	2930
$y_{Finland}$	8755	10367	11320	12377	15358	1.00	4723	14135	16770	18285	19908	23389	1.00	8566

References

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition*. CRC Press.