Using Digital Traces to Measure European Migrant Population in the UK by combining the Labour Force Survey and Facebook Advertising Data

Francesco Rampazzo[1,2], Jakub Bijak[1], Agnese Vitali[3], Ingmar Weber[4], and Emilio Zagheni[2]

[1]University of Southampton

[2]Max Planck Institute for Demographic Research

[3]Università di Trento

[4]Qatar Computing Research Institute

April 2019

This is a preliminary draft. Please do not cite without permission of the authors.

1 Introduction

Measuring international migration is challenging. The lack of timely and comprehensive data about migrants, and varying measures and definitions used by countries are a barrier to understanding international migration (Willekens (1994); Kupiszewska and Nowok (2008); Bijak (2010)). In this paper, we aim to complement traditional data sources with social media data. We are proposing to use a Bayesian data assessment model to combine the data from the Labour Force Survey (LFS) and Facebook Advertising Platform to study the number of European migrants in the UK, aiming to produce estimates of European migrants closer to their *true stock* number. In recent years, Bayesian methods have started to be used to combine different sources of migration data in order to provide a better estimate of the number of migrants (Bijak (2010); Azose and Raftery (2019)).

The Integrated Model of European Migration (IMEM) is the Bayesian model that we are aiming to use. This framework has been created by Raymer and colleagues (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. This model has been applied by Disney (2015) to combine multiple migration survey datasets in the United Kingdom (UK), and by Wiśniowski (2017) to combine the LFS data for Polish migration to the UK. The main feature of the IMEM is that it provides a framework which assesses the limitations of the datasets in terms of the definition of migrants used; the bias and the accuracy are also considered to create an appropriate prior distribution, which could adjust these data issues.

In parallel, a new strand of recent research has been re-purposing digital data to complement traditional demographic data sources and improve their coverage and timeliness of production. Migration has received particular attention, since digital traces data are mostly geo-located. As suggested by Cesare et al. (2018), digital traces data sources bring advantages, such as speed and low cost of data collection, but have limitations as well, for example accessibility and lack of representativeness. In this study, we aim to describe the features of Facebook Advertising Platform in order to use it to improve estimates of European migration in the UK. IN the UK long-term migration estimates are based on information collected through survey data, which are well-known for their sampling bias.

In order to inform migration policies, it is crucial to have access to valid sources of data on international migration. In this paper, we investigate whether the digital traces that individuals leave on Facebook can be used to estimate stock of migrants in the UK, using the Facebook Advertising Platform and the Labour Force Survey. This is not the first example of research which has tried to combine digital traces with survey data (2018), however, our approach differs in the Bayesian model used, the IMEM model which embeds a theoretical model considering push and pull factors related to migration theories. Moreover, our study also differs due to the specificities of the UK context and its existing migration data. In this study, we limit our attention to migrants from European countries, since in the British context they are the hardest groups to estimate. Since, at least until 2019, there is no requirement for registration for EU migrants in the UK, survey data are used to estimate the stock of migrants from the EU. The aim of the paper is thus to complement existing, but incomplete, official estimates of migrant's stocks' through digital traces.

2 Migration to the United Kingdom

In 1979, the UK first recorded a positive net migration, i.e. the number of immigrants exceeded the number of emigrants (Champion and Falkingham (2016)). From the 1980s, the UK has changed from a country of emigration to a country of immigration. The UK story of immigration is linked to two political organization: The Commonwealth, and the European Economic Community (EEC)/European Union (EU). The UK received many migrants from its former colonies such as India and Pakistan. However, in the last two decades migration to the UK from Europe has increased (Alfano, Dustmann, and Frattini (2016)).

The UK entered the EEC in 1973 (Hix and Høyland (2011)), and joining this political organisation agreed to the Treaty of Rome (1957), which in its article 3 established the freedom of movements of people across member states.

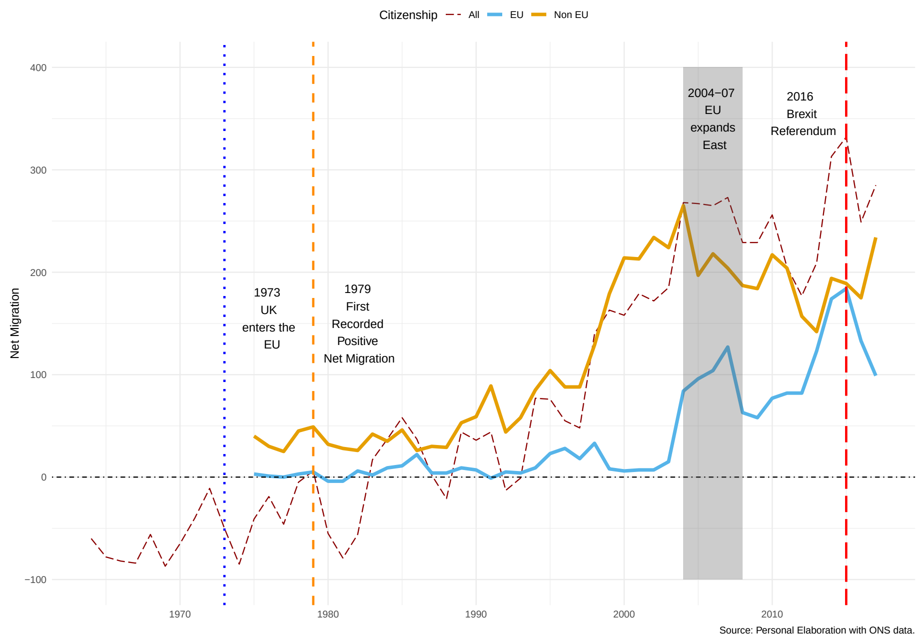

Figure 1: Net Migration estimates of long-term migrants since 1964.

Boswell and Geddes (2010) define European mobility as the migration of Europeans within Europe's borders. Once the Eastern European countries became part of the European Union, mobility within Europe increased. In 2004, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Slovenia and Slovakia (EU8) became part of the EU. By 2008, the most numerous foreign nationalities in the UK were Polish. In 2017, the Office of National Statistics (ONS) reported 9 million non-British born residents in the UK, 39.49%, of which were born in a European country. The second European nationality in the UK was Romanian. During 2017 Romanian nationals increased by 83,000 to 411,000. Figure 1 represents the overall trend of net migration since 1964, and from EU and non-EU countries since 1974. This data suggests that EU migration was increasing at the time of the Brexit Referendum (23rd June 2016), net migration may be still positive, but has significantly declined.

Since the UK triggered Article 50 of the Lisbon Treaty on March 2017, the British Government is trying to conclude an agreement with the EU on the process of leaving the EU, which would impact the European migrants currently residing in the UK and those of UK migrants elsewhere in the EU. Rights of EU migrants are expected to change after the UK departure (*Brexit*). If the withdrawal agreement is concluded, the rights and status of EU migrants are expected to stay the same until 30th June 2021, whereafter the UK will have a new migration policy for the EU migrants. All this process is still uncertain, because it is not clear when or if the UK will leave the EU. The bureaucratic burden for EU migrants in the UK has already started changing before Brexit: in fact, EU migrants are already allowed to register for a "*settle status*" in the UK, to prove they were living in the UK for five years continuously.

3 Migration Data in the United Kingdom

3.1 Survey-based migration data

British migration data is fragmentary: different data sources measure different migrant populations or migration events. In the absence of register data, the UK largely relies on a survey-based system to collect information on its population. The two main sources to estimate international migration to the UK are the International Passenger Survey (IPS), and, secondly, the Labour Force Survey (LFS).

The IPS has been running since 1961, and it was originally introduced to estimate overseas travel and tourism. In addition, it also provides estimates of inflows and outflows of international migrants. Nevertheless, the ONS suggests that the IPS *has been stretched beyond its original purpose* (ONS (2019)), and cannot be used as the only source to estimate international migration into the UK. As a matter of fact, the IPS measures the intention of the respondents to stay in the UK, and not the actual stay, which subsequently has to be adjusted for (Kupiszewska, Kupiszewski, et al. (2010)). This is not the best measure to estimate migration to a country, because the intention might change through the stay in the UK. The question of intention might lead to underestimation or overestimation of migrants. The IPS interviews travellers for 362 days a year (except Christmas Eve, Christmas Day, and Boxing Day) and has a coverage of 90% of the passengers travelling to and from the UK (ONS (2014)). The interviews happen in nineteen airports, eight ports and the Channel tunnel. The sample is usually made of 700,000-800,000 interviews, of these only 4,000 interviews per year are of long-term migrants. The response rate is 74.5%, and there is a lag of 11 months to have access to the data. The limited sample size implies that estimates of migrants by a cross-classification of country of origin, sex, age and possibly other basic demographic characteristics is difficult to obtain.

The second main source of data is the LFS, a Europe-wide quarterly household survey, which aims to estimate labour market conditions, including employment and unemployment. Through a boost of this survey, the Annual Population Survey (APS), the ONS collects data on the stocks of foreign born and foreign citizen in the UK at local area level. Further, the APS does not ask about intention to stay in the UK, but the period that has been already spent in the UK. The LFS interviews 41,000 households per quarter (ONS (2018a)), and combines two quarterly waves of the LFS, with a sample covering 360,000 individuals and 170,000 households per year. The data are released 3 months after the end of the survey.

At the moment, the ONS complements the above-mentioned IPS survey data with data from the Northern Ireland Migration Office, asylum seeker data from the Home Office, and visitor switchers (i.e., individuals that stay shorter or longer time than the intended 12 months). These corrections are applied to IPS, in order to arrive at the LTIM - long-term international migration estimates. The limitations of the sampling framework, the systematic bias, and the coverage of both the IPS and APS have been described by several researchers (Coleman (1983); Rendall, Tomassini, and Elliot (2003); Kupiszewska and Nowok (2008); Kupiszewska, Kupiszewski, et al. (2010)). The ONS is aware of the limitations of its approach and has recently started an ambitious plan that aims to complement additional administrative data sources with the IPS, and LFS to obtain a comprehensive measure of migration (ONS (2018b)). This process is hoped be completed by 2020.

3.2 New Migration Data

Exploring the digital traces that we leave as queries on search engines and as posts on social media is a new trend in the social sciences Cesare et al. (2018). Being a data driven discipline, demography is one of the fields of research which can benefit the most from the abundance of digital data. Moreover, Billari and Zagheni (2017) express their hope that the Data Revolution currently in progress will lead to studies at smaller granularities and on topics not yet explored by demography. They stress how important it will be to use this data, in formal demography, borrowing modelling techniques from computational disciplines.

New data sources are a *gold mine* for migration studies because they can contribute to the lack of information relating to this field of research. Digital traces from social media are fast to collect using Application Programming Interface (API), which is an access point to an app that can access a database [1]. Through Twitter or Facebook's APIs, it is possible to know in real time how many of the users are in a specific location. This feature has the potential to contribute to *nowcasting* migration. Other forms of digital traces that can be geo-located, such as e-mails, have been used to estimate international migration rates (Zagheni and Weber (2012)). Data from digital traces sources is cheap because by *repurposing* datasets originally intended for other purposing we no longer need to create new data infrastructure. Besides, new data sources can also add insights to expand the definition of international migrant. The definition of migrants is different among different countries, they depend

[1] https://www.howtogeek.com/343877/what-is-an-api/

on the time of stay outside of the country of usual residence, but they are still not harmonized worldwide (Kupiszewska and Nowok (2008); Willekens (1994)). Fiorio et al. (2017) highlights the potential of geo-tagged Twitter data to investigate short-term mobility and long-term migration. They suggest that digital traces data can contribute to refine migration theory and modelling. In addition, this data can be augmented through survey: it is possible to survey hard-to-reach populations, that with a traditional sampling framework would be too hard and expensive to interview. Indeed, Polish migrants have been interviewed on Facebook in Austria, Ireland, Switzerland, and the UK, with the intention of supplementing existing cross-national surveys (Pötzschke and Braun (2017)). In a period of 4 weeks, 1,100 Polish migrants were interviewed for a total budget of 500 euro. Nevertheless, there are also important limitations of these sources. Researchers do not have direct access to all these new datasets, and need to create partnerships with private companies instead. For example, Blumenstock (2012), in partnership with Rwanda's primary telecommunications operator, was able to obtain mobile phone records of 1.5 million mobile subscribers between 2005 and 2008 to study internal migration within the country. New companies, like LinkedIn, are providing access to their user's data through the *Economic Graph Challenge*, in which researchers can submit a proposal and, if selected, use LinkedIn data to research computational social science topics. However, this usually limit the research to a topic in line with the company's interests.

In addition, these data sources are non-representative of the entire population. Hargittai (2018) analyses the potential bias of different platforms in the US: Facebook, LinkedIn, Twitter, Tumblr and Reddit. She found that Facebook is the most *representative* social media across education levels, and internet skills, while the other social media are used by smaller and more selected groups of the population. The work of Hergittai builds on Lazer and colleagues (2014) critique of the assumption that we can substitute traditional data sources with *big data*, which is problematic without considering the bias of this new data sources. The authors also make the point about algorithm dynamics, which means that the companies whose data we are trying to use are constantly modifying their algorithms, and are in full control of the information the researchers ultimately receive.

3.3 Facebook in migration studies

The main contribution of Facebook data to demography so far has been the work by Zagheni et al. (2017), who proposed that migration can be estimated in the US by combining Facebook's Advertising Platform data with data from the American Community Survey (ACS), a high-quality survey. New data sources are compared with traditional data sources in order to understand the bias between the re-purposed dataset i.e., the Facebook Advertising Platform, and the official statistics. Building on this research, Zagheni et al. (2018) have tried to combine Facebook data with the ACS data to nowcast migration in the US. They use a Bayesian approach to combine the two data sources. Since they trust the data from the ACS, they created weights to apply to the Facebook data comparing Facebook to the ACS. In this way, they can combine the two data sources. Facebook might be used to have a more timely picture where the ACS takes longer to produce.

Facebook data have been used also in studying integration of migrants in Germany and the US (Dubois et al. (2018); Stewart et al. (2019)). Also, official statistics producers are investigating the use of *big data*. The European Commission have published two reports that investigate the effectiveness of inferring migrations through a combination of traditional and more recent data sources, e.g. mobile phone, social media, and other 'Big Data'. In the first report, the authors describe the different data sources and the ways to reduce bias through calibration techniques (Hughes et al. (2016)). The second reviews the possibility of using Facebook data to study migration and inform policy (Spyratos et al. (2018)). Therefore, it is still too early to judge the usefullness of digital traces, as the methodology that would pass the official statistics standards is not yet available.

4 Methodology

There is no perfect system capable to estimate international migrants. Recently, the Swedish register data, considered as the gold standard among the demographic datasets, have been proved to overcount migrants (Monti et al. (2018)). The IMEM builds on this consideration, that migration data might be biased, and through a Bayesian hierarchical model estimates the *true flow* of international migrants across sending and receiving countries (Raymer et al. (2013)). The original IMEM combines flows from sending and receiving countries across all EU. Wiśniowski (2017) uses the IMEM to combine the British and Polish LFS to estimate the flows across the two countries. Disney (2015) uses the IMEM model to combine different data sources, which measures different migrants' population, but for a single country, the UK.

In this study, we aim to measure the *true stock* of European migrants in the UK, which is the quantity of migrants known if our collection system were able to perfectly measure all the migrants (Disney (2015)), combining the LFS data with Facebook's Advertising Platform data. The definition of the true stock of migrants follows the United Nation definition, which defines an international migrant as a *person who moves from their country of usual residence for a period of at least 12 months* (UN (1998)). 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 together, while in the TBM other variables are considered in the estimation of the true stock. In its framework, the IMEM quantifies the limitations of the data sources and provides appropriate prior distribution to reduce the bias.

The data are assessed in terms of their limitations in respect to the UN definition in regards of (Raymer et al. (2013); Disney (2015)):

- *definition*: how close does the international migrant variable match the UN definition of international migrant?
- *coverage*: what proportion of the total immigration stock does the data

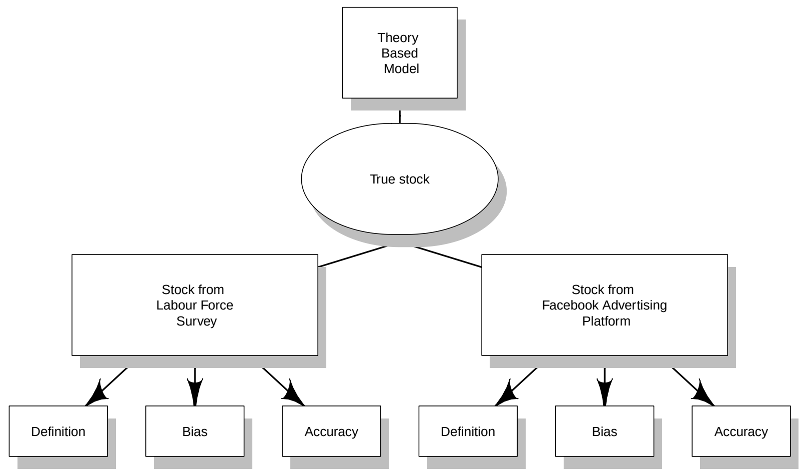

Figure 2: Diagram describing the structure of the model.

cover?

- *bias*: is there any systematic bias in the data?

In Figure 2, the model is described with a diagram. The true stock is at the center of the graph, and it is estimated by the TBM, and by the MEM, which combines the stock from the LFS with the stock from Facebook Advertising Platform, after considerations on their definition, bias, and accuracy.

The data model is constructed as follows. We observe the number of European migrants (stocks) Z^k from country i to the UK and with certain characteristics j, which are age and sex, from Facebook, F, and from the LFS, L, where $k \in (L, F)$. The datasets we are using can be described in the form of matrices Z^F for Facebook, and Z^L for the LFS.

$$Z^F = \begin{pmatrix} z_{11}^F & z_{12}^F & \dots & z_{1J}^F \\ z_{21}^F & z_{22}^F & \dots & z_{2J}^F \\ \vdots & \vdots & \ddots & \vdots \\ z_{ni1}^F & z_{I2}^F & \dots & z_{IJ}^F \end{pmatrix} \tag{1}$$

$$Z^L = \begin{pmatrix} z_{11}^L & z_{12}^L & \dots & z_{1J}^L \\ z_{21}^L & z_{22ij}^L & \dots & z_{2J}^L \\ \vdots & \vdots & \ddots & \vdots \\ z_{I1}^L & z_{I2}^L & \dots & z_{IJ}^L \end{pmatrix} \tag{2}$$

The value Y_{ij} is the random variable describing the number of the true stock we are aiming to estimate. It is a matrix with dimension IJ.

$$Y = \begin{pmatrix} y_{11} & y_{12} & \dots & y_{1J} \\ y_{21} & y_{22} & \dots & y_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ y_{I1} & y_{I2} & \dots & y_{IJ} \end{pmatrix} \tag{3}$$

The value of z_{ij}^k are assumed to follow a Poisson distribution:

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

4.0.1 Measurement Error Model

The general equation of the measurement error model is:

$$\log \mu_{ij}^k = \log(y_{ij}) + \delta_{ij}^k + \beta_{ij}^k + \chi_{ij}^k + \epsilon_{ij}^k \tag{5}$$

The equation is composed by three terms, δ_{ij}^k, β_{ij}^k, and χ_{ij}^k, which are used to convert the data from Facebook and the LFS to comply with the UN definition of international migrant and reduce the underestimation linked to the bias or coverage of the data. The first parameter, δ_{ij}^k, captures the differences in relation to the definition of migrants. The bias in the data is captured by β_{ij}^k, while the coverage is considered in χ_{ij}^k. The term ϵ_{ij}^k is the error term with normal distribution $N(0, \tau_{ij})$, the variance τ_{ij} has Gamma distribution $G(1, 0.01)$, with mean equal to 1 and variance equal to 0.01. The model is estimated using *Jags* in *R*. In *Jags*, the normal distributions are defines in terms of the mean, μ, and precision, τ. In the notation of the paper we express the probabilistic distributions in terms of mean, μ, and variance, σ.

Data Assessment of the Labour Force Survey The LFS defines an international migrant as it is suggested by the UN ONS (2018a). The LFS provides data on country of birth and citizenship. We prefer to use country of birth data since this should include all the individuals that are supposed to have a migrant background. Using the data on citizenship, we might lose all those individuals who have naturalised. Since the LFS is used to estimate the stock of migrants in the UK, many researchers have investigated the quality of its estimates. Rendall, Tommasini, and Elliot (2003) find that in 2001 LFS undereports 26% of the international migrants compared to the 2001 census. The bias might be over 30% for small nationalities, such as Greek and Lithuanians (Kupiszewska, Kupiszewski, et al. (2010)). The survey has over 15% of non-response (Martí and Ródenas (2007)). The sampling framework does not cover the entire target population (Kupiszewska, Kupiszewski, et al. (2010)), students and more mobile migrants might not appear in the sample. The response rate with imputation is 52.1% in 2015 (ONS (2015)).

The LFS measurement error equation is:

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

As for this assessment, the LFS data are deflated only by a parameter, β_{ij}^L, which consider both the bias and the coverage of the data. We consider the parameter δ_{ij}^L redundant, since the definition of international migrant in the LFS follows the one from the UN. For those countries with a small migrant population in the UK, we know from the literature (Rendall, Tomassini, and Elliot (2003); Kupiszewska, Kupiszewski, et al. (2010); Martí and Ródenas (2007)) that those numbers are 30% lower than the real number, this percentage is lower, 15%, for the country with a big population in the UK. As a consequence, the $\beta^{L}_{(ij)}$ parameter is assigned according to a parameter $g(ij)$, where:

$$g(ij) = \begin{cases} 1, & \text{if undercount assumed low.} \\ 2, & \text{if undercount assumed high.} \end{cases} \tag{7}$$

Generally, we assumed the undercount to be low for big European migrants country, and high for small European migrants country in the UK. For this reason,

$$\beta^{L}_{ij} \sim \begin{cases} N(-0.3, 0.01), & \text{if undercount assumed low.} \\ N(-0.15, 0.01), & \text{if undercount assumed high.} \end{cases} \tag{8}$$

Data Assessment of Facebook Advertising Platform Facebook provides their advertisers with information on its users' age, sex, level of education and et cetera, for this reason Facebook has been described as a bias *Digital Census* (Cesare et al. (2018)). The Facebook data are collected using pySocialWatcher (Araujo et al. (2017)). The variable that we are using to estimate international migrants is defined on Facebook as *People that were used to live in country x and now live in country y*. Until December 2018, the variable was defined as *Expat from country x*. Facebook's documentation does not provide any details on which information are used to create the variable. As a consequence, this is not a comprehensive definition since it does not cover any specification on the timing of the migration. Two studies have tried to infer how Facebook processes this category. In the first, researchers at Facebook suggest that Facebook users are considered in that category based on information from where they locate their hometown, and from the structure of their friendship network (Herdağdelen et al. (2016)). In the second, Spyratos et al. (Spyratos et al. (2018)) run a survey of 114 Facebook users asking whether Facebook's Advertising Platform identifies them as *expat*. They conclude that Facebook uses other information that are not specified in their profiles, and they suggest that this piece of information is an output of the geo-location. The final proof is given by a U.S Securities and Exchange Commission document, in which the Facebook company writes 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* (Commision (2018)).

Facebook marketing API provides two metrics: Daily Active Users (DAUs), and Monthly Active Users (MAUs). Facebook does not provide an explanation of how these measures are computed. Game metrics suggests that the DAUs measures the *unique users per day, usually calculated over the last 7 days*, while the MAUs *is an aggregate of the DAU over a month* (Fiels (2013)). In the same U.S Securities and Exchange Commission document (Commision (2018)) are reported estimates of the bias of the MAUs, which are 11% of duplicates accounts, and 5% of false accounts. Most of these anomalies are estimated in South East Asia. We are using the MAU estimates since from the Facebook document it is clear that this measure is more stable than the DAU. However, the MAU does not report numbers under 1000 for avoiding people to target too small groups.

Through Facebook Marketing API, we covered all Facebook users in an aggregated and anonymised format, but the coverage depends on age and gender. A Pew Research Center report (2018) showed that Facebook is used across all the age groups, with lower percentages of users at 50+ years old, and highlighted a decline in the numbers of younger users on Facebook, however, Facebook suggests that younger users register on Facebook with an inaccurate age, which is generally older than their actual age (Commision (2018)). In addition to the representation of Facebook across ages, we should consider the coverage between genders. Fatehkia and colleagues (2018), and Garcia and colleagues (2018) explored the use of Facebook to describe the digital gender gap: even in developed countries, the gap is reducing, but there are still more men than women on Facebook.

The measurement error model equation for Facebook is:

$$\log \mu_{ij}^F = \log(y_{ij}) + \delta_{ij}^F + \beta_{ij}^F + \chi_{ij}^F + \epsilon_{ij}^F \tag{9}$$

After this data assessment, we decided to have a parameter for the definition, the bias, and the coverage in the Facebook data. The Facebook δ_{ij}^F is normally distributed with $N(0, 0.1)$, while the β_{ij}^F has normal distribution $N(0.04, 0.1)$. The mean of β_{ij}^F is equal to 4% in order to deflate the FB estimates from fake and duplicate accounts. The mean of the coverage parameter χ_{ij}^F is the rate of non-Facebook users in the country of origin of the European migrants, since we are aiming to inflate by this bias. It is computed as:

$$\text{non-FB rate}_{ij} = 1 - \left(\frac{\text{Number of Facebook Users}_{ij}}{\text{Eurostat Population Size}_{ij}} \right) \tag{10}$$

The variance is equal to 0.1.

$$\chi_{ij}^F \sim N(-\text{non-FB rate}_{ij}, 0.1) \tag{11}$$

4.0.2 Theory Based Model

In this part of the model, we are introducing covariates that might help in explaining the true stock of European migrants in the UK.

$$\log y_{ij} = \alpha_0 + \alpha_1 P_{ij} + \alpha_2 I_{ij} + \alpha_3 O_{ij} + \alpha_4 \log G_i + \alpha_5 \log U_{ij} + \epsilon_{ij} \tag{12}$$

where $\alpha = (\alpha_0, \ldots, \alpha_5)^T$ is a vector of parameters; α_0 is normally distributed $\alpha_0 \sim N(0, 0.01)$, the constant is weakly informative, while $\alpha_{(1,\ldots,5)} \sim N(0, 100)$ are not weakly informative, since the variance is larger. The term ϵ_{ij} is the error term with normal distribution $N(0, \tau_{ij})$, the variance τ_{ij} has Gamma distribution $G(1, 0.001)$, with mean equal to 1 and variance equal to 0.001.

Here, a list of the covariates included in the model:

- P: Population size in the country of origin of the countries considered in the model. The data are the latest estimates from Eurostat (2019).
- I: Inflows from European conuntries to the UK; data from the International Passenger Survey 2017.
- O: Outflows to the European countries from the UK; data from the International Passenger Survey 2017.
- G: GDP growth rate in the European country of origin in 2017; data from Eurostat.
- U: Unemployment rate in the European country of origin in 2017; data from Eurostat.

5 Results

Out of the 27 European countries, 22 European countries are considered in the study: Austria, Belgium, Czech Republic, Denmark, France, Finland, Germany, Greece, Hungary, Latvia, Lithuania, Luxembourg, Ireland, Italy, Netherlands, Poland, Portugal, Romania, Slovenia, Slovakia, Spain, and Sweden. The 5 countries excluded, due to limitations in their data, are: Bulgaria, Croatia, Cyprus, Estonia, and Malta. We estimated the model only on the total number of European migrants in the UK, in a second model we included the dimension of sex, and finally, we measured 5 years age group estimates by sex for the ten most numerous countries. We decided to include only the numerous nationalities, since Facebook Advertising Platform gives back a counterfeit 1000 for all those groups that have a potential target between 0 and 1000.

In a first model, the true stock of European migrants in the UK is estimated for individuals older than 15 years old. The results from this model are presented in Figure 3; the estimates from the LFS are presented with confidence interval at 50%, while the estimates of the true stock are showed with interquintiles at 25% and 75%. The model tends to provide an estimate of the true stock higher than the LFS estimate. The model estimates more Italians than Germans, inverting the ranking from the LFS estimates. This might be caused by the wide difference across LFS and Facebook estimates of German migrants. The parameter χ_{ij}^{F} , which considers the coverage of Facebook in the sending country, might be the parameter at play in the German case. In fact, Germany has one of the lowest percentage of Facebook users in Europe, which is 40%, while the rest of Europe is over 50%. In only four countries, Belgium, Germany, Greece, and Ireland, the true stock is lower than the LFS estimates. In addition, in many cases, the LFS estimates are included in the interquintile of the true stock estimates. In the case of Belgium, Ireland, Netherlands, and other countries, the median of the true stock falls in the confidence interval of the LFS estimates. This happens mostly in small countries. In Table 1, the estimates from the LFS, Facebook, and the true stock are presented, alongside the mean percentage differences between the LFS and Facebook estimates on the one hand, and median of the true stock estimates on the other. From this measure, we notice the gradient of the difference between the true stock and the two data sources, and in which directions the true stock changed. We observe that the mean percentage difference between the LFS and the median has the highest negative values for Greece (-61.72%), and Germany (-17.92%). On the other hand, these two countries has the highest mean percentage difference between the Facebook estimates and the median, which is 632.71% for Greece, and 161.47% for Germany.

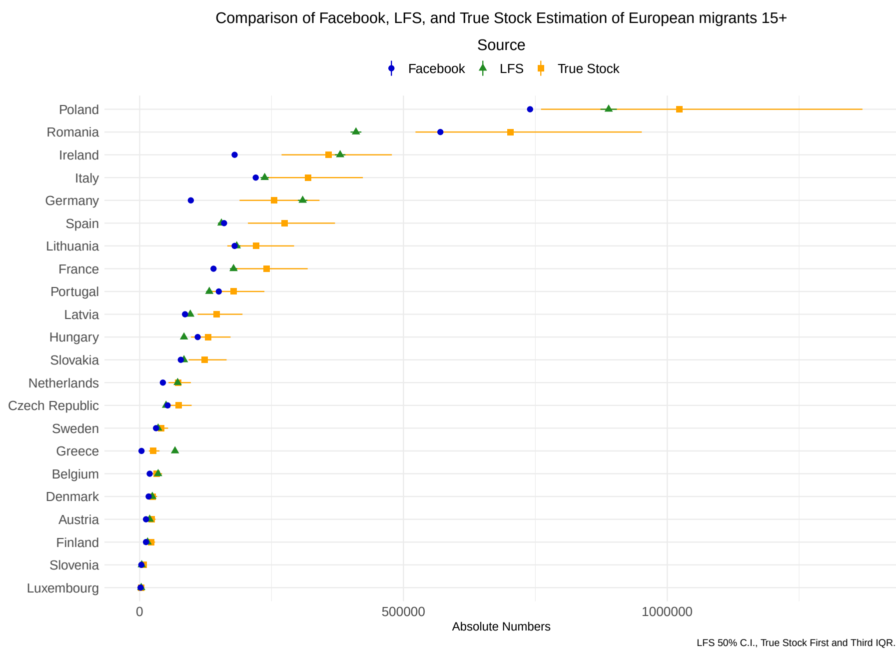

Figure 3: Comparison of the LFS, Facebook, and True Stock estimates of the total numbers of European migrants in the UK in 2018.

Table 1: Comparison of the LFS, Facebook, and True Stock estimates of the total numbers of European migrants in the UK in 2018.

	Data		True Stock Estimates			Mean Percentage Difference	
country	LFS	Facebook	q25%	Median	q75%	LFS-Median	Facebook-Median
Poland	889000	740000	755046	1018649	1372732	14.58	37.66
Romania	410000	570000	518786	698851	935921	70.45	22.61
Ireland	380000	180000	268308	357852	476238	-5.83	98.81
Italy	237000	220000	239096	317024	419071	33.77	44.10
Spain	155000	160000	207460	275369	367781	77.66	72.11
Germany	309000	97000	188600	253628	339820	-17.92	161.47
France	178000	140000	180047	240406	319816	35.06	71.72
Lithuania	184000	180000	166013	221074	292667	20.15	22.82
Portugal	132000	150000	134043	177554	234598	34.51	18.37
Latvia	96000	86000	110097	146693	194283	52.81	70.57
Hungary	84000	110000	97722	129321	170590	53.95	17.56
Slovakia	84000	78000	93599	123880	163619	47.48	58.82
Czech Republic	50000	53000	55790	74607	98458	49.22	40.77
Netherlands	72000	44000	54286	72435	96276	0.60	64.63
Sweden	35000	31000	30525	40580	53394	15.95	30.91
Belgium	35000	19000	24151	32139	42683	-8.17	69.16
Greece	67000	3500	17375	25644	37647	-61.72	632.71
Denmark	24000	17000	18868	25151	33338	4.80	47.95
Austria	19000	12000	17111	22721	29960	19.59	89.35
Finland	15000	12000	16532	21901	29063	46.01	82.51
Slovenia	4000	3200	5748	7716	10466	92.92	141.15
Luxembourg	3000	1700	2311	3069	4089	2.33	80.58

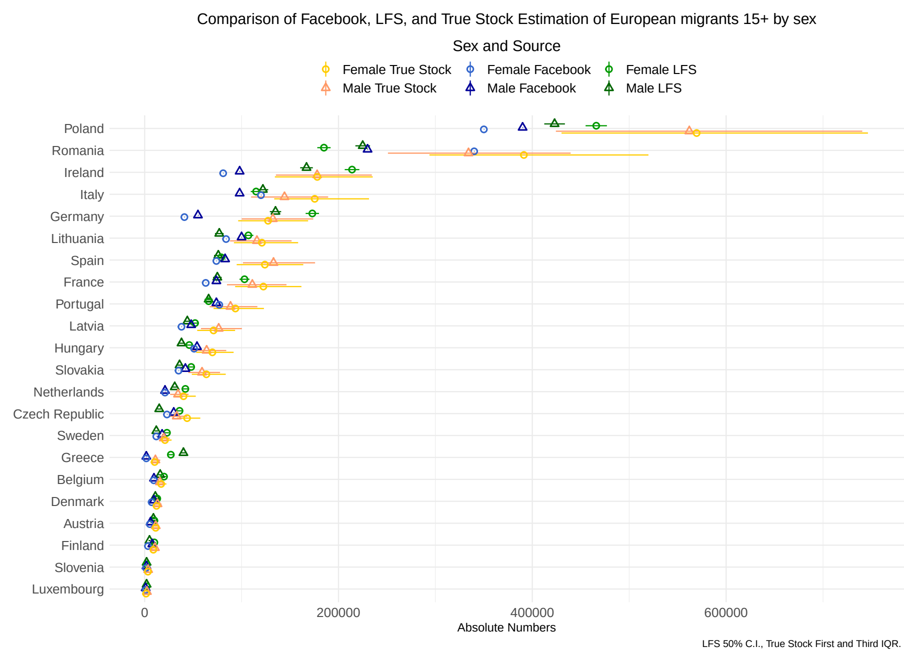

Figure 4: Comparison of the LFS, Facebook, and True Stock estimates of the total numbers of European migrants in the UK in 2018.

In a second model, we dissagregate the total estimates by sex. Adding this dimension, the true stock estimates remains consistent with the one estimated in the absence of this variable. In Figure 4, we can observe the differences across the estimates. The key main result from the model is that the true female stock in the first four most numerous nationalities, which are Poland, Romania, Ireland, and Italy, is higher than for male. The gap between female and male is wider for Romania and Italy. Table 2 reports the estimates from the LFS, Facebook, and the true stock.

Table 2: Comparison of the LFS, Facebook, and True Stock estimates of the total numbers of European migrants in the UK in 2018 by Sex.

	Female					Male				
	Data		True Stock Estimates			Data		True Stock Estimates		
Country	LFS	Facebook	q25%	Median	q75%	LFS	Facebook	q25%	Median	q75%
Poland	466000	350000	431243	569360	751498	423000	390000	425993	562080	742005
Romania	185000	340000	294255	392635	522911	167000	98000	135701	178389	234378
Ireland	214000	81000	134815	178564	236185	225000	230000	254109	338206	449325
Italy	115000	120000	133376	175624	231300	122000	98000	110752	145544	191430
Germany	173000	41000	95650	127270	168495	135000	55000	100352	132031	173154
Spain	79000	74000	94173	123874	163712	75000	74000	84595	110977	145940
France	103000	63000	93627	122869	161812	76000	83000	101230	133246	176060
Lithuania	107000	84000	91764	120917	158258	77000	100000	87733	115731	152351
Portugal	66000	77000	71457	94338	124038	66000	74000	67075	88233	116058
Latvia	52000	38000	54291	71561	93591	44000	48000	58051	76349	100390
Hungary	46000	51000	53328	69753	91255	38000	54000	48461	63964	84421
Slovakia	48000	35000	48457	63682	83796	40000	1700	7604	11082	16203
Czech Republic	36000	23000	33226	43810	57533	31000	21000	26414	34669	45613
Netherlands	42000	21000	30122	39900	52376	36000	42000	45246	59485	78216
Sweden	23000	12000	15956	21042	27526	15000	30000	25120	33208	44168
Belgium	20000	9400	12868	17065	22579	12000	18000	15043	19764	25995
Denmark	13000	7200	9353	12304	16195	11000	9600	10095	13258	17386
Austria	10000	5300	8633	11294	14836	5000	7900	7972	10567	14018
Greece	27000	1700	7324	10341	14713	16000	9600	11932	15760	20713
Finland	10000	3400	6766	8978	11844	9000	6100	8730	11451	14982
Slovenia	2000	1300	2550	3345	4424	2000	1800	2719	3579	4718
Luxembourg	2000	1000	1104	1466	1941	2000	1000	1760	2317	3058

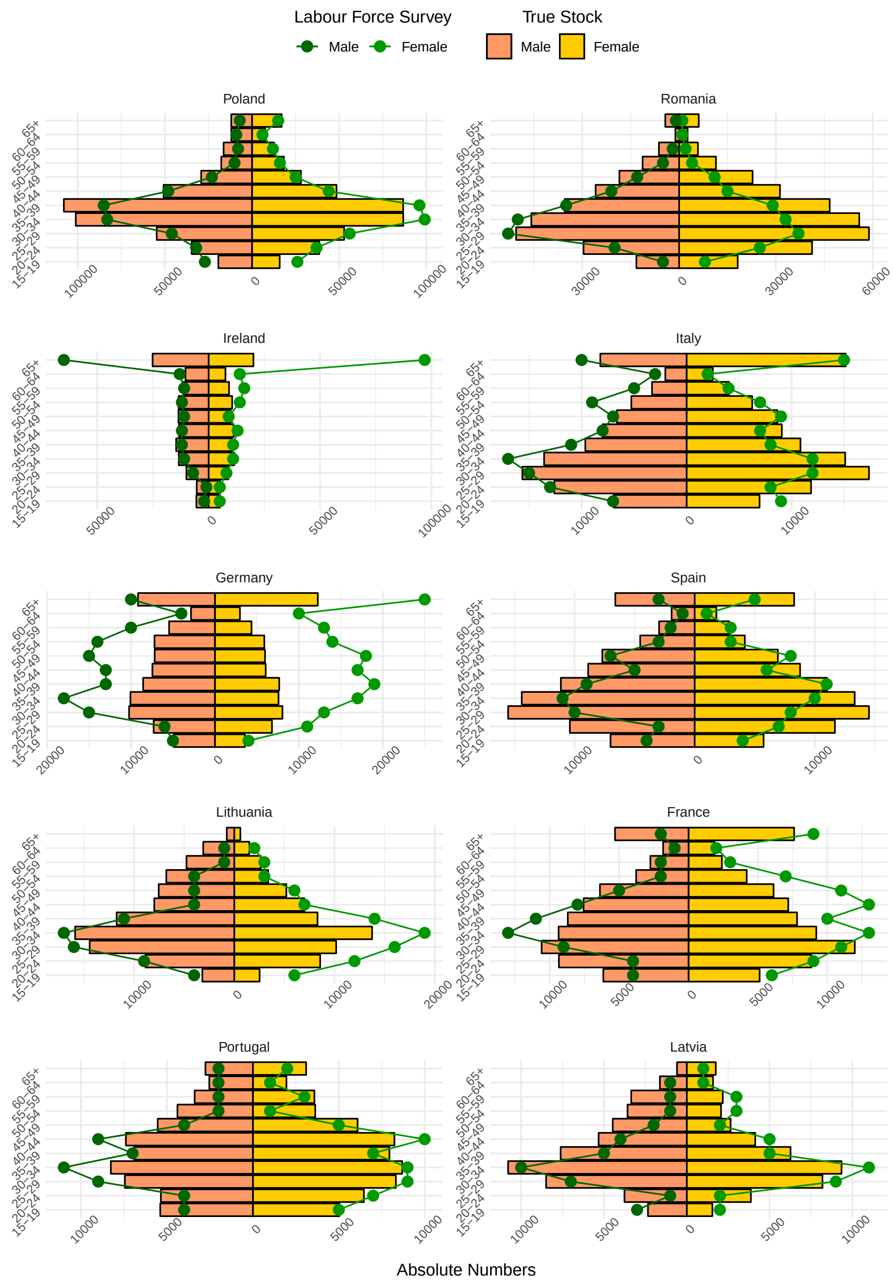

Figure 5: Comparison of the LFS, and True Stock estimates of the ten most numerous European migrants nationalities in the UK in 2018 by age and sex.

Finally, a third model is estimated only on the ten most numerous countries by sex and 5 years age groups. In Figure 5, the true stock estimates are portrayed in a population pyramid format. Over the population pyramids, the LFS estimates are showed though lines. We discard the Facebook estimates from this comparison since we think it is redundant to show the Facebook estimates; at this point, the comparison should be between the LFS and the true stock estimates. The shapes of the true stock population pyramids are more 'normal', and they follow the pattern of the Rogers and Castro's (1981) shape observed for migration flows, allowing for the ageing of the migrant populations themselves.

The model was also able to estimate true stocks for cells in which the LFS data are missing: for example, for the male Romanian between 60-64 years old, the Lithuanians older than 65+, and the male Latvian older than 65+. Observing the population pyramids, it seems there are problems in the estimates of the true stocks of Germany, which are consistently lower than the LFS. In addition, there are discrepancies at older ages for Ireland.

6 Conclusions

The model presented in the results session combines the LFS estimation of European migrants in the UK with data from Facebook's Advertising Platform. In addition, a migration theory model introduces covariates, which might have an impact on the estimation of the true stock. From the results presented, it seems that the true stock estimates are a reasonable combination of the different data sources without giving too much power to one or the other dataset. The results from the first model, which estimates the true stock for European total migrant population in the UK for people over 15 years old, are consistent with the results from the second model, in which a disaggregation by sex is included. The second model highlights a *feminization* of European migration to the UK (Kofman (2000)). European women might have higher education than male, and might find employment in occupations with higher female segregation. Moreover, for the ten most numerous European countries of immigration, the true stock is estimated by sex and five years age group. These results might be of interest for the ONS in order to provide better estimates of European migrants at the younger ages, or in cases when the LFS estimates are equal to zero. There are limitations to this approach at older ages, as we can notice in the Irish case (Figure 5). An additional complication of Ireland is the cultural closeness to the British culture and language, which might be difficult for Facebook algorithm to identify. Germany's true stock estimation are not congruent with the LFS estimates. There might be two reasons of this discrepancy. First, the percentage of Facebook users in Germany is the lowest in Europe, around 40%, while the majority of European countries are over 50%. This might suggest that there is an effect of the percentage of Facebook users in the population. Secondly, it might be that the ONS is over-reporting Germans living in the UK. The IMEM model used in this paper can be expanded to include more traditional data sources from the ONS, and also additional social media platforms. Moreover, the framework can be applied to other countries in which data sources are limited. We show the potential use of the model in cases in which post-stratification is not sufficient since the only traditional data sources available are biased. Nevertheless, in order to conclusively understand the approach a partnership with Official Statistical Offices is needed to comprise the most up-to-date data and evaluate the performance of the model.

References

Alfano, Marco, Christian Dustmann, and Tommaso Frattini (2016). *Immigration and the UK: Reflections after Brexit*. en. SSRN Scholarly Paper ID 2900373. Rochester, NY: Social Science Research Network. URL: https://papers.ssrn.com/abstract=2900373 (visited on 03/12/2019).

Araujo, Matheus et al. (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. event-place: Troy, New York, USA. ACM: New York, NY, USA, pp. 253–257. ISBN: 978-1-4503-4896-6. DOI: 10.1145/3091478.3091513. URL: http://doi.acm.org/10.1145/3091478.3091513 (visited on 03/10/2019).

Azose, Jonathan J. and Adrian E. Raftery (2019). "Estimation of emigration, return migration, and transit migration between all pairs of countries". en. In: *Proceedings of the National Academy of Sciences* 116(1), pp. 116–122. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1722334116. URL: https://www.pnas.org/content/116/1/116 (visited on 03/06/2019).

Bijak, Jakub (2010). *Forecasting International Migration in Europe: A Bayesian View*. en. Google-Books-ID: XGiApQtIt2QC. Springer Science & Business Media. ISBN: 978-90-481-8897-0.

Billari, Francesco C. and Emilio Zagheni (2017). "Big data and population processes: A revolution". In: *Statistics and Data Science: New Challenges, New Generations*, pp. 167–178.

Blumenstock, Joshua E. (2012). "Inferring patterns of internal migration from mobile phone call records: evidence from Rwanda". In: *Information Technology for Development* 18(2), pp. 107–125. ISSN: 0268-1102. DOI: 10.1080/02681102.2011.643209. URL: https://doi.org/10.1080/02681102.2011.643209 (visited on 03/11/2019).

Boswell, Christina and Andrew Geddes (2010). *Migration and Mobility in the European Union*. en. Google-Books-ID: FjEdBQAAQBAJ. Macmillan International Higher Education. ISBN: 978-1-137-28548-5.

Cesare, Nina et al. (2018). "Promises and Pitfalls of Using Digital Traces for Demographic Research". en. In: *Demography* 55(5), pp. 1979–1999. ISSN: 1533-7790. DOI: 10.1007/s13524-018-0715-2. URL: https://doi.org/10.1007/s13524-018-0715-2 (visited on 03/06/2019).

Champion, Tony and Jane Falkingham (2016). *Population Change in the United Kingdom*. en. URL: https://www.rowmaninternational.com/buy-books/product-details/ (visited on 03/08/2019).

Coleman, David (1983). "Some problems of data for the demographic study of immigration and of immigrant and minority populations in Britain". In:

Ethnic and Racial Studies 6(1), pp. 103–110. ISSN: 0141-9870. DOI: 10.1080/01419870.1983.9993401. URL: https://doi.org/10.1080/01419870.1983.9993401 (visited on 03/07/2019).

Commision, US SEC (2018). *FORM 10-K*. URL: https://www.sec.gov/Archives/edgar/data/1326801/000132680119000009/fb-12312018x10k.htm (visited on 03/10/2019).

Disney, George (2015). "Model-based estimates of UK immigration". en. phd. University of Southampton. URL: https://eprints.soton.ac.uk/374938/ (visited on 03/06/2019).

Dubois, Antoine et al. (2018). "Studying Migrant Assimilation Through Facebook Interests". In: *arXiv:1801.09430 [cs]*. arXiv: 1801.09430. URL: http://arxiv.org/abs/1801.09430 (visited on 03/17/2019).

Fatehkia, Masoomali, Ridhi Kashyap, and Ingmar Weber (2018). "Using Facebook ad data to track the global digital gender gap". In: *World Development* 107, pp. 189–209. ISSN: 0305-750X. DOI: 10.1016/j.worlddev.2018.03.007. URL: http://www.sciencedirect.com/science/article/pii/S0305750X18300883 (visited on 03/18/2019).

Fiels, Timothy Victor (2013). "Game Industry Metrics Terminology and Analytics Case Study". en. In: *Game Analytics: Maximizing the Value of Player Data*. Ed. by Magy Seif El-Nasr, Anders Drachen, and Alessandro Canossa. Springer-Verlag: London. ISBN: 978-1-4471-4768-8. URL: https://www.springer.com/us/book/9781447147688 (visited on 03/10/2019).

Fiorio, Lee et al. (2017). "Using Twitter Data to Estimate the Relationship between Short-term Mobility and Long-term Migration". en. In: *Proceedings of the 2017 ACM on Web Science Conference - WebSci '17*. ACM Press: Troy, New York, USA, pp. 103–110. ISBN: 978-1-4503-4896-6. DOI: 10.1145/3091478.3091496. URL: http://dl.acm.org/citation.cfm?doid=3091478.3091496 (visited on 03/11/2019).

Garcia, David et al. (2018). "Analyzing gender inequality through large-scale Facebook advertising data". en. In: *Proceedings of the National Academy of Sciences* 115(27), pp. 6958–6963. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/

pnas.1717781115. URL: https://www.pnas.org/content/115/27/6958 (visited on 03/18/2019).

Hargittai, Eszter (2018). "Potential Biases in Big Data: Omitted Voices on Social Media". en. In: *Social Science Computer Review*, p. 089443931878832. ISSN: 0894-4393, 1552-8286. DOI: 10.1177/0894439318788322. URL: http://journals.sagepub.com/doi/10.1177/0894439318788322 (visited on 03/17/2019).

Herdağdelen, Amaç et al. (2016). "The Social Ties of Immigrant Communities in the United States". In: *Proceedings of the 8th ACM Conference on Web Science*. WebSci '16. event-place: Hannover, Germany. ACM: New York, NY, USA, pp. 78–84. ISBN: 978-1-4503-4208-7. DOI: 10.1145/2908131.2908163. URL: http://doi.acm.org/10.1145/2908131.2908163 (visited on 03/10/2019).

Hix, Simon and Bjørn Høyland (2011). *The Political System of the European Union*. en. Google-Books-ID: lk5dDwAAQBAJ. Macmillan International Higher Education. ISBN: 978-1-349-27531-1.

Hughes, Christina et al. (2016). *Inferring migrations, traditional methods and new approaches based on mobile phone, social media, and other big data: feasibility study on inferring (labour) mobility and migration in the European Union from big data and social media data*. en. OCLC: 960398851. Publications Office: Luxembourg. ISBN: 978-92-79-59742-8. URL: http://bookshop.europa.eu/uri?target=EUB:NOTICE:KE0216632:EN:HTML (visited on 03/17/2019).

Kofman, Eleonore (2000). "The invisibility of skilled female migrants and gender relations in studies of skilled migration in Europe". en. In: *International Journal of Population Geography* 6(1), pp. 45–59. ISSN: 1099-1220. DOI: 10.1002/(SICI)1099-1220(200001/02)6:1⟨45::AID-IJPG169⟩3.0.CO;2-B. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291099-1220%28200001/02%296%3A1%3C45%3A%3AAID-IJPG169%3E3.0.CO%3B2-B (visited on 04/03/2019).

Kupiszewska, Dorota, Marek Kupiszewski, et al. (2010). "Possibilities and limitations of comparative quantitative research on international migration flows". en. In: p. 57.

Kupiszewska, Dorota and Beata Nowok (2008). “Comparability of statistics on international migration flows in the European Union”. en. In: *International Migration in Europe: Data, Models and Estimates*. John Wiley & Sons, pp. 41–71. ISBN: 978-0-470-98554-0.

Lazer, David et al. (2014). “The Parable of Google Flu: Traps in Big Data Analysis”. en. In: *Science* 343(6176), pp. 1203–1205. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1248506. URL: http://science.sciencemag.org/content/343/6176/1203 (visited on 03/16/2019).

Martí, Mónica and Carmen Ródenas (2007). “Migration Estimation Based on the Labour Force Survey: An EU-15 Perspective”. In: *The International Migration Review* 41(1), pp. 101–126. ISSN: 0197-9183. URL: https://www.jstor.org/stable/27645654 (visited on 03/10/2019).

Monti, Andrea Fuentes et al. (2018). *Over-coverage in Swedish Population Registers Leads to Bias in Demographic Estimates*. DOI: 10.17045/sthlmuni.7000907.v2. URL: https://su.figshare.com/articles/Over-coverage_in_Swedish_Population_Registers_Leads_to_Biasin_Demographic_Estimates/7000907 (visited on 03/10/2019).

ONS (2014). *International passenger survey (IPS) methodology*. eng. Text User Guide (Volume 1). URL: https://webarchive.nationalarchives.gov.uk/20160115212418/http://www.ons.gov.uk/ons/guide-method/method-quality/specific/travel-and-transport-methodology/international-passenger-survey/index.html (visited on 03/07/2019).

ONS (2015). *Performance and Quality Monitoring Report - Labour Force Survey*. URL: https://webarchive.nationalarchives.gov.uk/20160106001313/http://www.ons.gov.uk/ons/guide-method/method-quality/specific/labour-market/labour-force-survey/index.html.

ONS (2018a). *Labour Force Survey – user guidance - Office for National Statistics*. Tech. rep. URL: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/methodologies/labourforcesurveyuserguidance#labour-force-survey-lfs-user-guides (visited on 03/08/2019).

ONS (2018b). *Migration statistics transformation update - Office for National Statistics*. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/migrationstatisticstransformationupdate/2018-05-24 (visited on 03/07/2019).

ONS (2019). *Update on our population and migration statistics transformation journey: a research engagement report - Office for National Statistics*. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/updateonourpopulationandmigrationstatisticstran 2019-01-30 (visited on 03/07/2019).

Pötzschke, Steffen and Michael Braun (2017). "Migrant Sampling Using Facebook Advertisements: A Case Study of Polish Migrants in Four European Countries". en. In: *Social Science Computer Review* 35(5), pp. 633–653. ISSN: 0894-4393. DOI: 10.1177/0894439316666262. URL: https://doi.org/10.1177/0894439316666262 (visited on 03/10/2019).

Raymer, James et al. (2013). "Integrated Modeling of European Migration". In: *Journal of the American Statistical Association* 108(503), pp. 801–819. ISSN: 0162-1459. DOI: 10.1080/01621459.2013.789435. URL: https://amstat.tandfonline.com/doi/abs/10.1080/01621459.2013.789435 (visited on 03/06/2019).

Rendall, Michael S., Cecilia Tomassini, and David J. Elliot (2003). "Estimation of annual international migration from the Labour Force Surveys of the United Kingdom and the continental European Union". en. In: *Statistical Journal of the United Nations Economic Commission for Europe* 20(3,4), pp. 219–234. ISSN: 0167-8000. URL: https://content.iospress.com/articles/statistical-journal-of-the-united-nations-economic-commission-for-europe/sju00560 (visited on 03/07/2019).

Research, Pew (2018). *Social Media Use 2018: Demographics and Statistics — Pew Research Center*. en-US. URL: http://www.pewinternet.org/2018/03/01/social-media-use-in-2018/ (visited on 03/18/2019).

Rogers, A. and L. J. Castro (1981). *Model Migration Schedules*. en. Monograph. URL: http://pure.iiasa.ac.at/id/eprint/1543/ (visited on 04/01/2019).

Spyratos, S et al. (2018). *Migration data using social media a European perspective.* en. Tech. rep. OCLC: 1059244090.

Stewart, Ian et al. (2019). “Rock, Rap, or Reggaeton?: Assessing Mexican Immigrants’ Cultural Assimilation Using Facebook Data”. In: *arXiv:1902.09453 [cs]*. arXiv: 1902.09453. URL: http://arxiv.org/abs/1902.09453 (visited on 03/17/2019).

UN, ed. (1998). *Recommendations on statistics of international migration*. en. Statistical papers. Series M no. 58, rev. 1. United Nations: New York. ISBN: 978-92-1-161408-4.

Willekens, Frans (1994). “Monitoring international migration flows in Europe”. en. In: *European Journal of Population / Revue européenne de Démographie* 10(1), pp. 1–42. ISSN: 1572-9885. DOI: 10.1007/BF01268210. URL: https://doi.org/10.1007/BF01268210 (visited on 03/06/2019).

Wiśniowski, Arkadiusz (2017). “Combining Labour Force Survey data to estimate migration flows: the case of migration from Poland to the UK”. en. In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180(1), pp. 185–202. ISSN: 1467-985X. DOI: 10.1111/rssa.12189. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssa.12189 (visited on 03/06/2019).

Zagheni, Emilio, Kivan Polimis, et al. (2018). “Combining Social Media Data and Traditional Surveys to Nowcast Migration Stocks”. In: *Annual Meeting of the Population Association of America*.

Zagheni, Emilio and Ingmar Weber (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. event-place: Evanston, Illinois. ACM: New York, NY, USA, pp. 348–351. ISBN: 978-1-4503-1228-8. DOI: 10.1145/2380718.2380764. URL: http://doi.acm.org/10.1145/2380718.2380764 (visited on 03/10/2019).

Zagheni, Emilio, Ingmar Weber, and Krishna Gummadi (2017). “Leveraging Facebook’s Advertising Platform to Monitor Stocks of Migrants”. en. In: *Population and Development Review* 43(4), pp. 721–734. ISSN: 1728-4457.

DOI: 10.1111/padr.12102. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/padr.12102 (visited on 03/06/2019).