European Commission

JRC TECHNICAL REPORTS

Migration Data using Social Media

a European Perspective

Spyratos S., Vespe M., Natale F.,
Weber I., Zagheni E., Rango M.

2018

EUR 29273 EN

Contact information
Name: Spyridon Spyratos
Address: TP 266, Via E.Fermi 2749, 21027 Ispra (VA), Italy
Email: spyridon.spyratos@ec.europa.eu
Tel.: +39 033278 5024

JRC Science Hub
https://ec.europa.eu/jrc

JRC112310

EUR 29273 EN

PDF ISBN 978-92-79-87989-0 ISSN 1831-9424 doi:10.2760/964282

Luxembourg: Publications Office of the European Union, 2018

Spyratos, S., Vespe, M., Natale, F., Ingmar, W., Zagheni, E. and Rango, M., Migration Data using Social Media: a European Perspective, EUR 29273 EN, Publications Office of the European Union, Luxembourg, 2018, ISBN 978-92-79-87989-0, doi:10.2760/964282, JRC112310.

Contents

Acknowledgements

We would like to thank Tintori Guido and Pietro Argentieri for their contribution to this report.

Authors

Spyratos Spyridon[1], Vespe Michele[1], Natale Fabrizio[1], Weber Ingmar[2], Zagheni Emilio[3], Rango Marzia[4]

[1] European Commission, Joint Research Centre, Knowledge Centre on Migration and Demography
[2] Qatar Computing Research Institute, Social Computing Group
[3] Department of Sociology, University of Washington
[4] International Organization for Migration, Global Migration Data Analysis Centre (GMDAC)

Abstract

Migration is a top political priority for the European Union (EU). Data on international migrant stocks and flows are essential for effective migration management. In this report, we estimated the number of expatriates in 17 EU countries based on the number of Facebook Network users who are classified by Facebook as "expats". To this end, we proposed a method for correcting the over- or under-representativeness of Facebook Network users compared to countries' actual population. This method uses Facebook penetration rates by age group and gender in the country of previous residence and country of destination of a Facebook expat. The purpose of Facebook Network expat estimations is not to reproduce migration statistics, but rather to generate separate estimates of expatriates, since migration statistics and Facebook Network expats estimates do not measure the same quantities of interest. Estimates of social media application users who are classified as expats can be a timely, low-cost, and almost globally available source of information for estimating stocks of international migrants. Our methodology allowed for the timely capture of the increase of Venezuelan migrants in Spain. However, there are important methodological and data integrity issues with using social media data sources for studying migration-related phenomena. For example, our methodology led us to significantly overestimate the number of expats from Philippines in Spain and in Italy and there is no evidence that this overestimation may be valid. While research on the use of big data sources for migration is in its infancy, and the diffusion of internet technologies in less developed countries is still limited, the use of big data sources can unveil useful insights on quantitative and qualitative characteristics of migration.

1 Introduction

Migration is a top political priority for the European Union (EU). To address this, in May 2015 the European Commission (EC) introduced the European Agenda on Migration (European Commission 2018), highlighting explicitly the need for more and better use of information for several policy areas. Data on international migrant stocks and flows are essential for effective migration management, including the design, implementation and evaluation of policies. Improving data and their disaggregation by basic characteristics, among which migratory status, is also an overarching requirement of the 2030 Agenda for Sustainable Development, and part of the first objective of the Global Compact for Safe, Orderly and Regular Migration currently under negotiation.

Statistics on international migrant stocks are available from the United Nations' Department of Economic and Social Affairs (UNDESA), the World Bank, Eurostat (for EEA countries), and the Organization for Economic Co-operation and Development (OECD; for OECD countries). These statistics are characterized by a number of limitations and gaps, which reflect the limited availability of up-to-date and comprehensive statistics on international migrant stocks at the national level, particularly in low-income countries. Since international migrant stock data mainly derive from national population censuses, which are conducted infrequently in most countries, they hardly capture the age-sex distribution of international migrants in a country in a timely fashion. Second, these statistics are based on data provided by individual countries with separate collection systems and designs (Raymer et al. 2013). Third, they fail to describe new and transient forms of migration, such as transnationalism or circular migration, and often only give a picture of regular migration, since irregular migrants might not appear in censuses or official registers. For instance, as observed by Sinn, Kreienbrink and Von Loeffelholz (2005), in Germany even a thorough analysis of available data sources cannot provide reliable data on the size and composition of the irregular resident population. There are no official datasets on irregular migration and irregular migrants in the EU (Vespe, Natale and Pappalardo, 2017), although several existing datasets can be used as proxies to provide estimates. The lack of data on irregular migration is a matter of serious concern among scholars and the international community (Vono De Vilhena, 2018).

Innovative sources offer data that are timely, have a wide coverage, can be accessible at limited cost, and can potentially include information that may not always be provided by traditional migration data sources. As of April 2018, the Facebook Network (i.e. Facebook, Instagram, Messenger and the Audience Network) reported more than 2.15 billion monthly active users[5]. The digital traces that Internet users are actively or passively generating can potentially be exploited for studying migration-related phenomena. In 2014, the United Nations (UN) recognized the importance of big data for official statistics, and established a Global Working Group on this topic (UN 2018). The potential of new data sources (e.g. Facebook) to provide policy-relevant information on international migration is currently being explored by international bodies and initiatives such as the Big Data for Migration Alliance (BD4M), recently launched by the European Commission Knowledge Centre on Migration and Demography (KCMD) and IOM's Global Migration Data Analysis Centre (GMDAC), following a dedicated workshop on the topic (Rango and Vespe 2017).

However, social media users are not representative of the society at large as they are prone to a selection bias. People with different age, sex, socioeconomic and cultural backgrounds use social media applications to varying degrees (Smith and Anderson 2018). Apart from issues of representativeness, the use of personal data from social media applications raises concerns regarding the disclosure of personal information, as well as the integrity and the overall governance of the data by the entity that collects them. For example, the recent Facebook and Cambridge Analytica data breach scandal sparked significant public discussions about the lack of ethical and privacy standards in social media companies. The inherent bias of social media data described above, coupled with the risk of a lack of control on how the data are derived and processed cause uncertainties

[5]https://developers.facebook.com/docs/marketing-apis

regarding the possibility of effectively using such data for demographic research. This requires the development of appropriate methods for quantifying and mitigating such bias as well as the active collaboration with the social media companies themselves. Moreover, an in-depth evaluation of the potential of these data sources is needed to respond reliably to societal challenges and policy questions related to migration.

Several studies have used big data sources, such as social media and internet services, for analysing migration-related phenomena. Seminal work in this area was carried out by Zagheni, Weber and Gummadi (2017), who used data from Facebook's advertising platform to estimate the stock of international migrants in the US. Messias, Benevenuto, Weber and Zagheni, (2016) used Google+ data for studying location patterns of migrants who have lived in more than two countries. State, Ingmar and Zagheni (2013) and Zagheni and Weber (2012) estimated global flows of migrants and tourists using the IP geolocation of over 100 million anonymized users of Yahoo! Web services and of a large sample of Yahoo! Email messages sent between September 2009 and June 2011. Zagheni, Garimella, Weber and State (2014) and Hawelka et al. (2014) analysed trends in mobility and migration flows using geo-located 'tweets'. Ahas, Silm and Tiru (2017) used roaming data from mobile operators to map transnationalism from Estonia. Dubois, Zagheni, Garimella and Weber (2018) used data from Facebook's advertising platform to estimate the levels of assimilation of Arabic-speaking migrants in Germany. Herdagdelen, State, Adamic and Mason (2016), from the Facebook data science team, studied the composition of immigrants' social networks in the United States (US) using the structure of their friendship ties. Finally, State, Rodriguez, Helbing and Zagheni (2014) and Barslund and Busse (2016) investigated mobility patterns of highly skilled migrants using data from LinkedIn. A recent report prepared for the European Commission investigated the feasibility of using big data for studying migration issues (European Commission et al. 2016) and concluded that big data sources a) do not substitute traditional data sources but they can complement them; and b) can be used for estimating trends or changes in trends in migration flows in a timely manner.

Our research is based on the work of Zagheni, Weber and Gummadi (2017), and is innovative since it proposes a new method which takes into account the difference between the definition of "expat" as used by Facebook and the statistical definition of a foreign-born migrant as per the 1998 UN Recommendations on Statistics of International Migration[6]. The aim of the proposed method is to create independent estimates of Facebook Network "expats" instead of trying match the Facebook Network-provided expat estimations to existing official migration data in absolute terms. To this end, the proposed method calibrates the number of Facebook Network expats by using the penetration rate of Facebook Network in the country of destination and in the country of previous residence of migrants. To estimate the penetration rate of Facebook Network usage in a country we use population data. The proposed method uses migration data only to identify the degree to which a migrant assimilates to the Facebook Network usage patterns of the destination country.

The rest of the article is structured as follows. The next section describes the official population and migration data and the Facebook Network data used in this research. We then explain the methodology used to estimate the number of individuals who fulfil specific demographic criteria (e.g. "Italian expats in Germany"), based on Facebook Network statistics. The Results & Discussion section presents validated figures of the proposed model in Europe. Conclusions are outlined in the final section.

[6] https://unstats.un.org/unsd/publication/SeriesM/SeriesM_58rev1e.pdf

2 Data

In this study, we use both traditional data available from Eurostat and UNDESA, and data from an innovative source – the Facebook advertising platform. The following sections describe the data in more detail.

2.1 Facebook Network data

We use data from the Facebook advertising platform[7] to estimate stocks of "expats" in various countries (see below for a discussion on the definition of "expat"). The Facebook advertising platform allows advertisers to select the characteristics of their target audience, for instance, age and gender, and to obtain an estimate of the number of monthly active users of the Facebook Network (Facebook[8], Instagram[9], Messenger[10] and the Audience Network[11]) who meet the selected criteria and could be reached through an advertising campaign. According to Facebook, this estimation is a unique calculation based on factors such as user self-reported demographic characteristics, and is not intended to align with third-party calculations or population census data. The frequent discrepancy between Facebook Network estimations of the number of individuals with certain characteristics living in a country and census data on the same population groups opens up questions about the reliability of Facebook estimates on the one hand, and the possible gaps in traditional statistics, on the other.

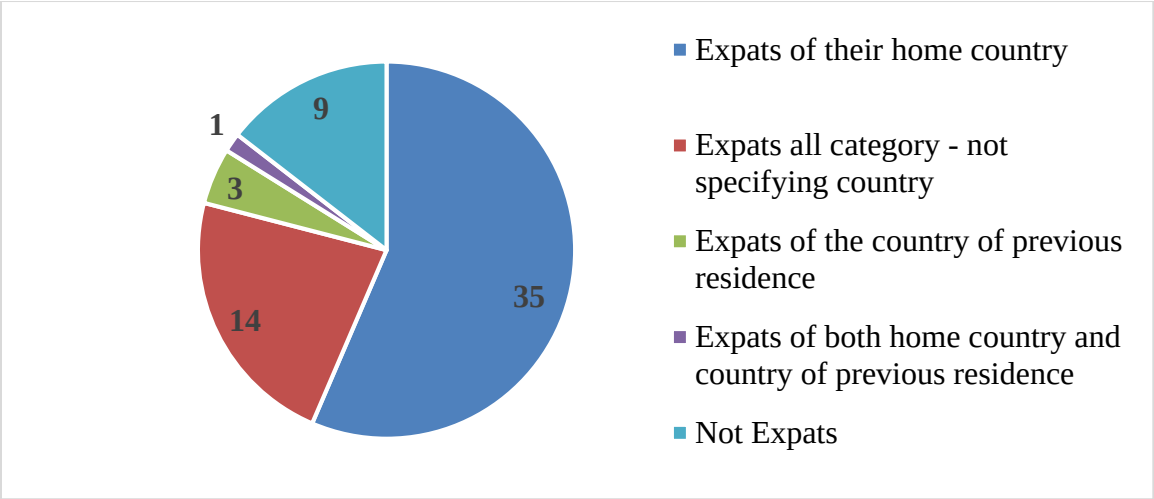
Through the Facebook's Marketing Application Programming Interface (API)[12], we collected data about the number of monthly active users of the Facebook Network based on the country of their current location, their age, gender and the country of their previous residence of which they are considered as expats. For example, we queried the number of Facebook Network users who now live in France, are male aged between 20 and 24 years old and are classified by Facebook as expats from Italy. As of February 2018, Facebook provided estimates about the expats of 89 countries. Facebook Network data for the EU case-study countries were collected between 30 January 2018 and 5 February 2018. Facebook does not provide the exact number of users that match specific criteria but gives a rounded number. As of 26 February 2018, the minimum response of Facebook marketing API to queries regarding its monthly active users was increased from 20 users to 1000. This means that one would not be able to obtain the number of monthly active users who match specific criteria unless the total number of users in this group is higher than 1000

The Facebook Network's definition of expats – "People who used to live in country X who now live abroad" – is quite generic. Facebook does not disclose details about the method used for classifying users as expats. A study by Facebook staff Herdagdelen et al. (2016) categorized Facebook users as expats based on their "hometown," as reported in their profiles. However, it is unclear whether the Herdagdelen et al (2016) approach is currently used by Facebook. We therefore conducted an online survey to understand how Facebook classifies its users as expats. The survey was limited to Facebook users and excluded users of other Facebook Network applications (Instagram, Messenger and Audience Network). As part of the survey, a) we requested participants to provide us with some personal information, i.e. country of origin/home country and current country of residence, b) to tell us what information they are reporting on Facebook, i.e. current city and hometown, and finally c) to access the "Ads preferences" Facebook webpage[13] and check whether Facebook classifies them as expats. A total of 114 Facebook users participated in this small survey, of whom 27 were not able to visualize any Facebook categories in the "Ads preferences" webpage. Of the remaining 87 participants, 62 were expatriates (living in a country other

[7] https://www.facebook.com/business/products/ads
[8] https://www.facebook.com/
[9] https://www.instagram.com/
[10] https://www.messenger.com/
[11] https://www.facebook.com/audiencenetwork
[12] https://developers.facebook.com/docs/marketing-apis
[13] https://www.facebook.com/ads/preferences

than their country of their hometown). **Figure 1** shows how Facebook classified these 62 participants.

Figure 1. Facebook classification of users who are expatriates (total: 62).



Of the 35 Facebook users who were classified by Facebook as expats of their home country (56% of the total), 14 stated that they did not report their hometown on their Facebook profile. Despite the small sample, this analysis suggests that Facebook uses additional attributes for estimating the country of an expat, in addition to user self-reported information on their home country. 14 participants were classified as expats but without a specification of the country ("Expats all category"); the 8 out of these 14 participants are originated from countries for which Facebook does not currently provide expat estimates, i.e. Bulgaria and Turkey. This simple survey was useful to understand how Facebook classifies expats but cannot be used for quantifying the accuracy of such a categorization, because of two main methodological limitations. First, our sample is not random, since most of the participants belong to the social and professional network of the authors of this study. Second, during a validation exercise that we conducted by contacting some participants, we realized that a proportion of those who are expatriates and declared that they are not classified by Facebook as expats, responded inaccurately. This is because they were not navigating correctly to the "Ads preferences" Facebook webpage where the categories (e.g. "Italian expat") were listed. To conclude, Facebook uses the "country of home town" and/or "country of previous residence" Facebook attributes for classifying its users as expats, among other attributes like geo-referenced information.

2.2 Official statistic

Official statistics on international migrant stocks disaggregated by age, sex, country of birth and destination were used to a) identify the degree to which a migrant assimilates to the Facebook Network usage patterns of the destination country and b) evaluate and compare the results of the model proposed. Migration statistics at this level of disaggregation are available from UNDESA (2008), the OECD in collaboration with the World Bank (2010) and Eurostat (2017a). We used Eurostat statistics since they were more updated. We additionally used updated population statistics from UNDESA (2017a) for calibrating the Facebook Network data.

2.2.1 Eurostat migration statistics

Statistics from Eurostat were used as a reference since they are more recent compared to statistics from the other two sources. Eurostat provides statistics disaggregated by country of birth and citizenship. The survey we performed suggested that Facebook mainly uses information on the country of the user's hometown and, secondarily, country of previous residence for defining a user as an expat of a country. Since the country of a user's hometown does not necessarily coincide to the country of birth or the country citizenship,

in this study we assume that the country where the hometown is located mostly refers to the country of birth. Thus, we selected the Eurostat dataset that provides disaggregation by country of birth, entitled "Population on 1 January by age group, sex and country of birth" (Eurostat 2017a), and it is hereafter called "Eurostat foreign-born migrant dataset". Eurostat adopts the UN definition of a (long-term) international migrant as a person who changes his or her place of usual residence for a period of at least 12 months (including people who arrive in a country with the intention of staying for at least 12 months)[14,15]. Eurostat provides figures of international migrant stocks for the reference year 2017, by age group, sex and country of birth for 18 EU countries.[16] It is worth mentioning that these countries do not include three major EU countries – Germany, France and the United Kingdom.

2.2.2 UNDESA population statistics

UNDESA population statistics are used in the study as an input in our model, to calibrate Facebook Network statistics. Using UNDESA population statistics, we estimated the penetration rates of the Facebook Network in each country, by age group and gender. UNDESA statistics are available for 200 countries globally, and are disaggregated by sex and 5-year age groups (UN/DESA 2017a).[17] We used population estimates (medium projection variant) for the year 2018, to match temporally the Facebook Network data.

[14] https://unstats.un.org/unsd/publication/SeriesM/SeriesM_58rev1e.pdf
[15] http://ec.europa.eu/eurostat/cache/metadata/en/demo_pop_esms.htm
[16] These countries are Austria, Belgium, Bulgaria, Czech Republic, Denmark, Estonia, Spain, Finland, Hungary, Italy, Lithuania, Luxembourg, Latvia, Netherlands, Romania, Sweden, Slovenia and Slovakia.
[17] The dataset used is "Population by 5-year age groups, annually from 1950 to 2100"

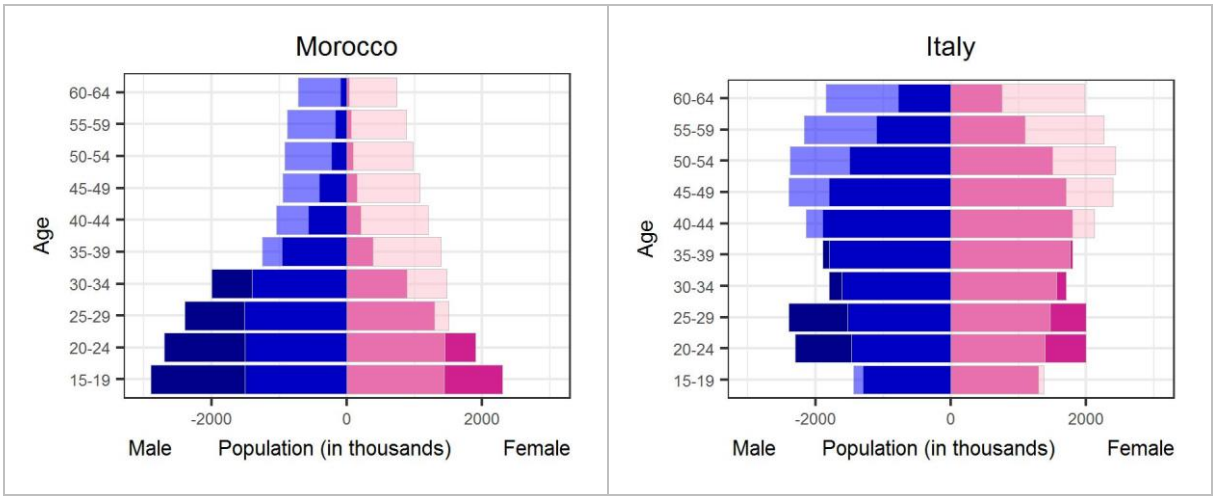
3 Methodology

We developed a methodology to estimate the number of individuals who fulfil specific demographic criteria based on non-representative Facebook Network statistics. For example, we want to estimate the number of individuals who are in a particular age group, are female or male, who used to live in, e.g. Germany, and who now live in another country, e.g. France, based on the number of Facebook Network users that meet those age and gender criteria and are classified by Facebook as German expats in France. In the following sections, we first analyse the Facebook Network data and its limitations, then we pre-assess and clean the Facebook Network statistics, and lastly, we present the model that we developed.

3.1 Understanding the data bias

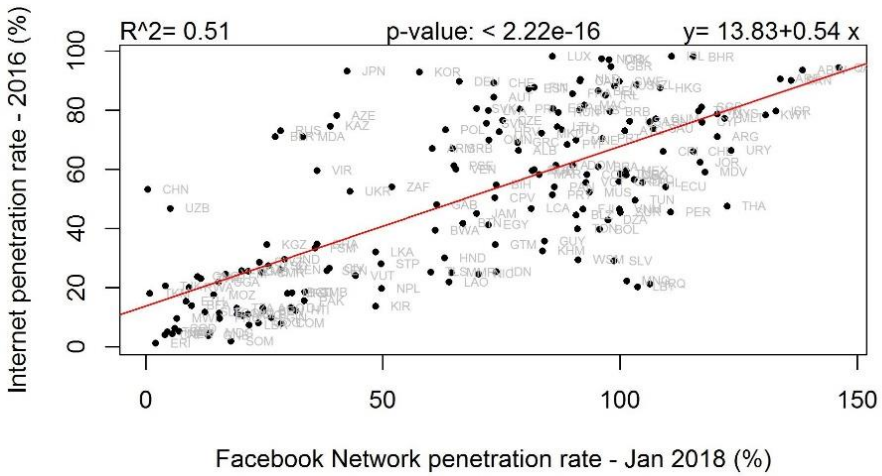
We analysed the characteristics of Facebook Network statistics to develop a robust model for correcting the bias given by the fact that Facebook Network users may over or under-represent a country's population at large. As shown in **Figure *2***, Facebook Network users' representativeness varies based on the country under consideration, as well as demographic characteristics of the population, namely gender and age. In Morocco, use of Facebook Network platforms is more widespread among males than females, while in Italy the differential in usage patterns based on gender is very small. When the number of Facebook Network users in a country and a given age group is higher than the actual number of residents in that age group (based on official statistics), it means that users have multiple unlinked Facebook Network accounts, for instance on Facebook, Instagram and Messenger. We assume that there are two main drivers of Facebook Network platforms' usage. The first is users' socio-psychological altitude towards Facebook Network platforms[18]. Second, there are technological and artificial constraints, for example, low internet penetration and restricted access to Facebook Network platforms in some countries. As shown in **Figure *3*** the percentage of individuals using the internet (International Telecommunication Union 2016a) explains 51% of the total variation in the Facebook Network penetration rates.

Figure 2. UNDESA population and Facebook Network (FN) users in Morocco and Italy by age and gender. Light shading=UNDESA population, mid shading=UNDESA & raw FN users, and dark shading = raw FN users



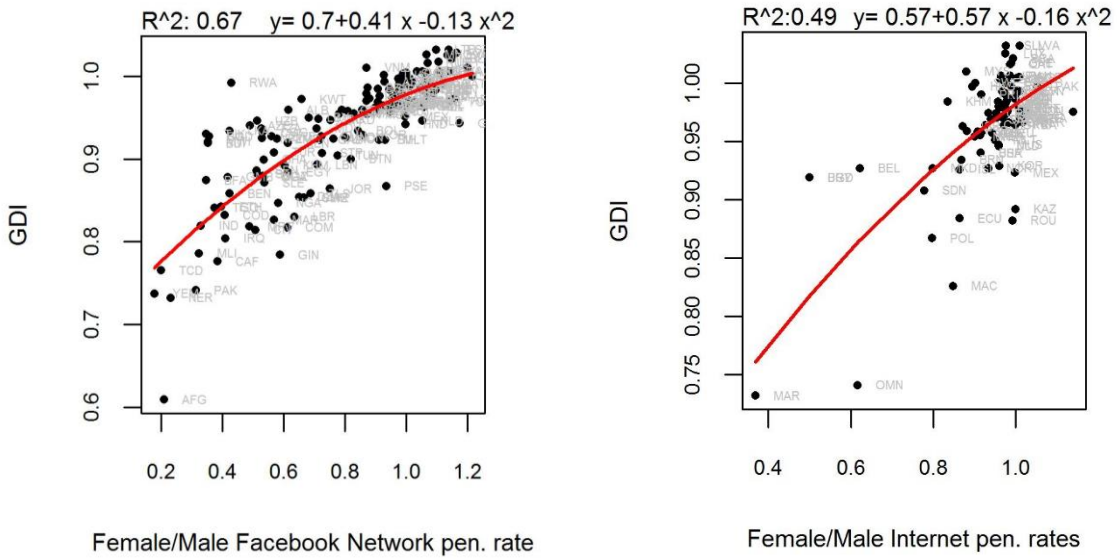
[18] For a detailed analysis of social media research theories we direct the interested reader to Ngai, Tao and Moon (2015).

Figure 3. Internet penetration rates Vs Facebook Network penetration rates across countries.



The popularity of different Facebook platforms also varies across countries due to the existence of alternative platforms -- for example, the most popular social networking site in Russia is not Facebook but VKontakte. Usage patterns of Facebook Network platforms by gender vary considerably across countries. A study by Fatehkia, Kashyap and Weber (2018) demonstrated the feasibility of using Facebook data for quantifying digital gender gaps. In our study, we demonstrate the correspondence between the Gender Development Index (GDI) and gender inequalities in the usage of Facebook Network applications. The GDI measures gender inequalities in three basic dimensions of human development: long and healthy life, education, and command over economic resources (United Nations Development Programme 2016). As shown on the left side of **Figure *4*** the GDI and gender inequalities in penetration rates of the Facebook Network across countries are correlated (R^2 =0.67, p<0.001). The GDI is also correlated with gender inequalities in internet access available from the International Telecommunication Union (2016b) (R^2 =0.49, p<0.001) (Figure 4). Interestingly, the GDI correlates to a higher degree with inequalities in penetration rates of Facebook Network platforms than with internet penetration rates. Financial, educational and cultural barriers in countries with conservative gender norms may prevent women from using social media (Fatehkia, Kashyap, and Weber 2018).

Figure 4. Correlation between the Gender Development Index (GDI) and female/male Facebook Network penetration rates (on the left side of the figure) and correlation between the GDI and female/male internet penetration rates across countries (on the right side of the figure).



3.2 Data preparation and cleaning

In the data preparation phase, we pre-assessed and cleaned the Facebook Network statistics. We evaluated Facebook Network data and we excluded from our analysis countries of previous residence and countries of destination for which either Facebook Network data were not reliable, or the use of Facebook was not permitted, or for which no third-party data were available. More specifically:

- *Expats from the US and Greece.* It appeared that the Facebook advertising platform is underestimating the number of Greek expats and overestimating the number of American expats. As of April 2018, the number of Facebook Network users, aged 15—64 years old, who are classified as Greek expats worldwide is 10,000, far lower than the official UNDESA (2017b) estimate of international migrants born in Greece and residing abroad in 2017, equal to 993,000. In the case of the US, there are 443,500 US expats Facebook Network users aged 15—64 residing in Italy, while US-born migrants residing in the country and within the same age group are only 40,073, based on Eurostat statistics.
- *Expats from China*. Access to Facebook is not possible in China[19], and since the proposed method requires knowledge of the Facebook Network penetration rate in the country of previous residence, we excluded Chinese expats from our analysis.
- *Expats from Cuba, Puerto Rico, Hong Kong and Monaco*. Expats from Cuba were excluded from the analysis since no Facebook Network statistics were available for this country. Expats from Puerto Rico and Hong Kong were excluded for lack of available Eurostat statistics, and expats from Monaco due to unavailability of UNDESA estimates.
- *Expats in Romania*. Facebook significantly overestimates expats residing in Romania, therefore we excluded Romania as a country of destination from our analysis. For example, there are 16,660 and 11,844 users in the 15—64 age group classified as Indian and Indonesian expats in Romania, respectively. The number of migrants born in India and Indonesia in that age group and residing in Romania, based on Eurostat statistics, are 360 and 48, respectively.

In total, of the 89 countries for which Facebook provides expat estimates, we used 82 countries of previous residence in our analysis (Annex 1). Since we excluded Romania as a country of destination, the countries of destination used in this study are 17 (Annex 2).

We also excluded combinations of countries of previous residence and destination for which the number of expats was very low. Since the minimum Facebook API response to each query was 20 users, and since we performed 20 queries for each previous residence-destination combination, one for each of the 10 age groups and for the 2 genders, the minimum aggregate number of expat users for each previous residence-destination was 400. To reduce the impact that this minimum responses might have in the model, we excluded from our analysis previous residence-destination combinations for which eight or more age-gender combinations had a (rounded) number of Facebook Network users equal to the minimum (20). As of 26 February 2018, the minimum number of Facebook Network monthly active users returned by the API increased from 20 to 1000. This increase makes difficult the repetition of this study, at the same level of disaggregation, since most of the queries will not super pass the new minimum API response (1000 users).

3.3 Model

As seen in Figure 2, Facebook Network penetration rates vary based on age, gender and country of residence. The representativeness of Facebook Network users by age group or gender changes depending on the country of residence. For instance females aged 30 to 34 years old in Italy have different usage patterns of the Facebook Network applications compared to females of the same age in Morocco. We therefore introduced a coefficient in our model for correcting the over- or under- representativeness of Facebook Network

[19] https://en.wikipedia.org/wiki/Censorship_of_Facebook

users. The coefficient is unique for each country of previous residence, country of destination, age group and gender. Our hypothesis is that the penetration rate of Facebook Network among e.g. males of age 20 to 24 years old, who used to live in Germany and now living in France is affected by the Facebook Network penetration rate among all males in the same age group living in Germany, as well as by the Facebook Network penetration rate of all males in the same age group residing in France.

We chose not to employ iterative proportional fitting procedure, or regression models that use the number of Facebook Network users for each age-gender-previous residence-destination as independent variables and known migration figures as depended variables, as methods for correcting the representativeness bias. Our aim is not to fit Facebook Network expat data to migration statistics in absolute terms, but rather to generate separate estimates, as migration statistics and estimates of Facebook Network expat users do not measure the same quantities of interest. However, as described in the next section, in order to select optimal $z_{a,g}$ and t parameters, which are used for the estimation of the Weight of the Destination country (*WD*), we assume that the age-gender distribution of Facebook Network expats users and of Eurostat foreign-born migrants will be somehow correlated.

The corrected number of Facebook Network users in age group *a*, of gender *g*, previous residence *o*, and who live in country *d,* denoted by $Fb\_cor_{a,g,o,d}$, corresponds to the estimated population of the this demographic group in the society. $Fb\_cor_{a,g,o,d}$ is estimated using equation (1) by dividing the raw number of Facebook Network expat users $Fb\_raw_{a,g,o,d}$ (the unprocessed number of users provided by Facebook) by a coefficient which is unique for each age-gender-previous residence-destination, and that is estimated using equation (2). The purpose of this coefficient is to correct Facebook Network's over or under-representativeness, considering the Facebook Network penetration rates in the country of previous residence and the country of the destination of a migrant. The $fpr_{a,g,c}$ in equation (2) is the penetration rate of the Facebook Network in country *c,* among users of gender *g* and age group *a,* and is estimated using equation (3). $Fb\_users_{a,g,c}$ in equation (3) is the number of raw Facebook Network users in country *c,* of age group *a* and gender *g, and* $UNDESA\_pop_{a,g,c}$ is the population in country *c,* of age group *a* and gender *g,* based on UNDESA population statistics.

$WD_{a,g,o,d}$ is the weight of the destination country and it reflects the extent to which migrants assimilate to the Facebook Network usage patterns in the destination country. The $WD_{a,g,o,d}$ is estimated using equation (4). The estimation of $z_{a,g}$ and t parameters is described in the next section. $WD_{a,g,o,d}$ can take any real number value between 0 and 1. This weight differs depending on gender, age group and the difference between the Facebook Network penetration rate in the country of destination and in the country of previous residence. The reason for considering the Facebook Network penetration rates in countries of previous residence and destination is that we assume that migrants who move from a country with higher penetration rates to a country with lower penetration rates will continue using the Facebook Network (if they are users), and will not fully adopt the digital habits of the destination country. Conversely, we expect that migrants moving from a country with lower Facebook Network penetration rates to a country with higher penetration rates, will be more likely to assimilate to the digital habits of the destination country and increase their use of the Facebook Network. Low Facebook Network penetration rates in a country may be due to low internet penetration rates, so migrants are more likely to adopt the digital habits of the destination countries once the technical constraints of internet access are removed. Finally, *WO* is the weight of the previous residence country and is estimated using equation (5).

$Fb\_cor_{a,g,o,d} = Fb\_raw_{a,g,o,d}/coefficient_{a,g,o,d}$ *where a = [15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64]* *g = [Male, Female]* *o ⊆ see Annex 1* *d = see Annex 2*	(1)

$coefficient_{a,g,o,d} = fpr_{a,g,d} * WD_{a,g,o,d} + fpr_{a,g,o} * WO_{a,g,o,d}$	(2)
$fpr_{a,g,c} = Fb\_users_{a,g,c} / UNDESA\_pop_{a,g,c}$ *where* $c = o \cup d$	(3)
$WD_{a,g,o,d} = z_{a,g} + t((fpr_{a,g,d} - fpr_{a,g,o}) / fpr_{a,g,d})$ *If* $WD_{a,g,o,d} > 1$ *then* $WD_{a,g,o,d} = 1$ *If* $WD_{a,g,o,d} < 0$ *then* $WD_{a,g,o,d} = 0$	(4)
$WO_{a,g,o,d} = 1 - WD_{a,g,o,d}$	(5)

3.3.1 Selection of the weight of the destination country *WD*

For the selection of the optimal values of the $z_{a,g}$ and t parameters, used for estimating $WD_{a,g,o,d}$ in equation (4), we assessed the correspondence between the corrected number of Facebook Network expat users and the number of Eurostat foreign-born migrants. We used natural logarithms of both variables to reduce the influence of observations with high numbers of migrants in the model. Also, we excluded age-gender-destination-previous residence groups in the Eurostat dataset ($Eurostat_{a,g,o,d}$) with less than 20 migrants to reduce the influence of non-significant migrant groups in the model. As mentioned in the Data section, Facebook Network classifies expats using an unknown methodology, so the definition of a Facebook Network expat may not be equivalent to that of a foreign-born migrant. For this reason, we decided not to select the optimal values of the parameters based on how well the corrected number of Facebook Network expat users approximate the number of migrants of X previous residence in country Y. Instead, we selected the optimal values based on how well the corrected number of Facebook Network expat users explains the age-gender distribution of Eurostat foreign-born migrants. We used the coefficient of determination, denoted R^2, as a statistical measure for evaluating how well the corrected Facebook Network data explain the age-gender distribution of foreign-born migrants in the EU, using different $WD_{a,g,o,d}$ values.

The optimal values of the $z_{a,g}$ and t parameters of the $WD_{a,g,o,d}$ are those for which the highest R^2 values were obtained when comparing the corrected Facebook Network expat estimations with the Eurostat foreign-born migrant statistics. We ran linear regressions (Table 1) between the log($Fb\_cor_{a,g,o,d}$) and the log($Eurostat_{a,g,o,d}$). We randomly selected 10 countries for the four training datasets and 7 for the four testing datasets. In each linear regression of the training or the testing datasets, we included the 10 or 7 destination countries, all possible countries of previous residence, 10 age groups and the 2 genders. We repeated each linear regression using all the possible parameter values t= [0,0.1,..,0.9,1] and $z_{a,g}$= [0,0.1,..,0.9,1] and we selected the combination of parameters yielding the highest R^2 when comparing log($Fb\_cor_{a,g,o,d}$) and log($Eurostat_{a,g,o,d}$). Due to computational limitations, we grouped the 10 age groups in the $z_{a,g}$ parameter to 4 age hyper-groups. We so reduced the number of required linear regressions for each training and testing dataset from 11^{21} to 11^{9}. We further reduced the number of the required iteration from 11^{9} to $\approx 11^{6}$ by splitting the process into two steps. In the first step, we considered 4 possible values $z_{a,g}$= [0.25.0.5,0.75,1] and all possible t parameter values t=[0,0.1,..,0.9,1]. In the second step, based on the optimal parameter for each gender and age hyper-group derived from step one – e.g. $z_{male,15\text{-}24}$= 0.75 – we targeted the potential values using 0.1 intervals, e.g. $z_{male,15\text{-}24}$= [0.6,0.7,0.8,0.9]. We finally speeded up the processing time by implementing 8-core parallel processing using the R software.

Table 1. Optimal values of the $z_{a,g}$ and t parameters for the 4 training and 4 testing datasets.

	g	a	Training					Testing				
			iteration				Median	iteration				Median
			1	2	3	4		1	2	3	4	
$z_{a,g}$ parameter	female	15-24	0.7	1	0.9	0.8	0.85	0.9	0.7	0.9	1	0.9
		25-39	0.5	0.5	0.5	0.5	0.5	0.4	0.4	0.3	0.2	0.35
		40-54	0.3	0.3	0.3	0.3	0.3	0.2	0.2	0.2	0	0.2
		55-64	0.3	0.3	0.3	0.4	0.3	0.3	0.3	0.4	0	0.3
	male	15-24	0.9	1	1	0.9	0.95	1	0.8	0.9	1	0.95
		25-39	0.8	0.8	0.8	0.9	0.8	0.8	0.8	0.9	0.6	0.8
		40-54	0.8	0.7	0.8	0.9	0.8	0.7	0.8	0.7	0.5	0.7
		55-64	0.8	0.7	0.9	1	0.85	0.9	0.8	0.8	0.6	0.8
t			0.3	0.1	0.3	0.2	0.25	0.3	0.4	0.2	0.6	0.35
R^2			0.75	0.82	0.84	0.83		0.85	0.81	0.78	0.78	
countries			SVK FIN SWE AUT HUN DNK BGR SVN BEL LTU	EST BGR SVN NLD SVK LTU ESP CZE LVA FIN	DNK SWE LUX BEL FIN EST ESP NLD AUT CZE	EST LTU FIN SVK ITA LVA HUN BEL SWE ESP		CZE ESP EST ITA LUX LVA NLD	AUT BEL DNK HUN ITA LUX SWE	BGR HUN ITA LTU LVA SVK SVN	AUT BGR CZE DNK LUX NLD SVN	

As shown in **Table *1***, the values of $z_{a,g}$ for females are lower than the values for males in all the iterations and on both training and testing datasets. In addition, the values of $z_{a,g}$ for younger age groups are higher than those for older age groups. This suggests that male migrants assimilate to a greater extent to the digital habits of the destination country compared to female migrants, and that young migrants assimilate to a higher degree compared to older migrants. These results are in line with those obtained by Dubois et al. (2018), where male and young migrants showed greater assimilation rates in terms of their Facebook interests compared to women and older people, respectively. We selected the median of the four training iterations shown in **Table *1*** as optimal values for parameters $z_{a,g}$ and t. The median optimal values of the training datasets are almost equivalent to the median optimal values of the testing datasets. The optimal values of parameters $z_{a,g}$ and t are presented in (6). **Figure *5***, shows the correlation between the log value of the corrected Facebook Network expats, estimated using the optimal parameter values, and the log value of the Eurostat foreign-born migrants for each previous residence-destination-age-gender combination. Finally, in **Figure *6*** we present how the R^2 values of the correlation between the log($Fb\_cor_{a,g,o,d}$) and the log($Eurostat_{a,g,o,d}$) change, based on different $z_{25\text{-}39,g}$ values for the estimation of $WD_{a,g,o,d}$, and consequently, of $Fb\_cor_{a,g,o,d}$.

$$t = 0.25$$

$$z_{15\text{-}20,female}=0.85,\ z_{20\text{-}24,female}=0.85,\ z_{25\text{-}29,female}=0.5,\ z_{30\text{-}34,female}=0.5,\ z_{35\text{-}39,female}=0.5,\ z_{40\text{-}44,female}=0.3,\ z_{45\text{-}49,female}=0.3,\ z_{50\text{-}54,female}=0.3,\ z_{55\text{-}59,female}=0.3,\ z_{60\text{-}64,female}=0.3$$

$$z_{15\text{-}20,male}=0.95,\ z_{20\text{-}24,male}=0.95,\ z_{25\text{-}29,male}=0.8,\ z_{30\text{-}34,male}=0.8,\ z_{35\text{-}39,male}=0.8,\ z_{40\text{-}44,male}=0.8,\ z_{45\text{-}49,male}=0.8,\ z_{50\text{-}54,male}=0.8,\ z_{55\text{-}59,male}=0.85,\ z_{60\text{-}64,male}=0.85 \quad (6)$$

Figure 5. Linear regressions between 11,725 log($Fb\_cor_{a,g,o,d}$) and log($Eurostat_{a,g,o,d}$) observations, using optimal $z_{a,g}$ and t parameters (see (6)) for the estimation of $WD_{a,g,o,d}$. For example, the red observation represents the Moroccan expats in Spain who are male and belong to the age group 45-49.

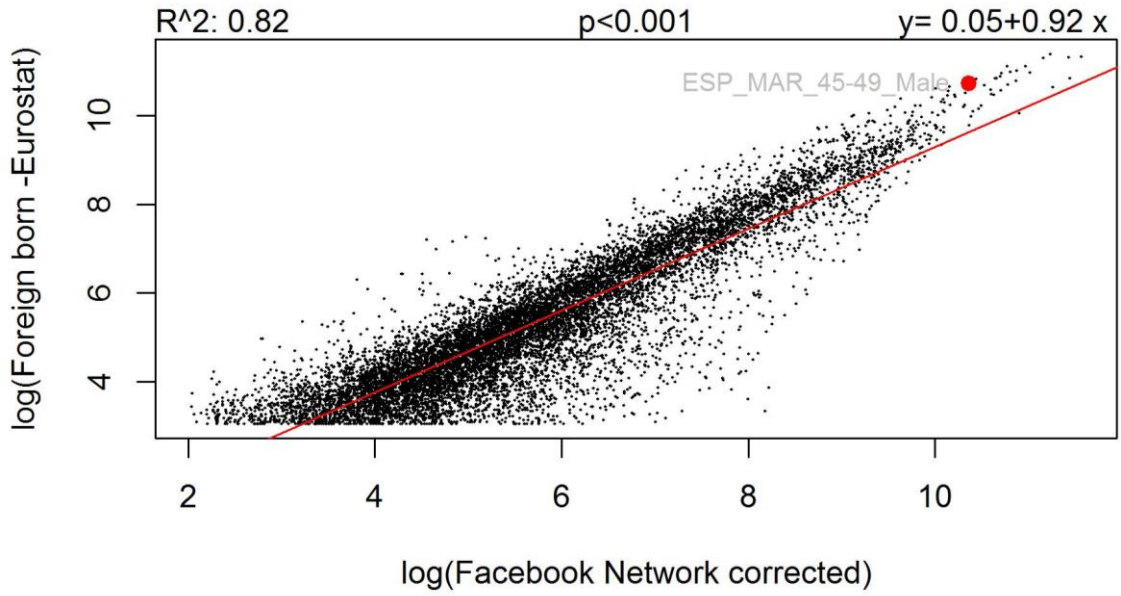
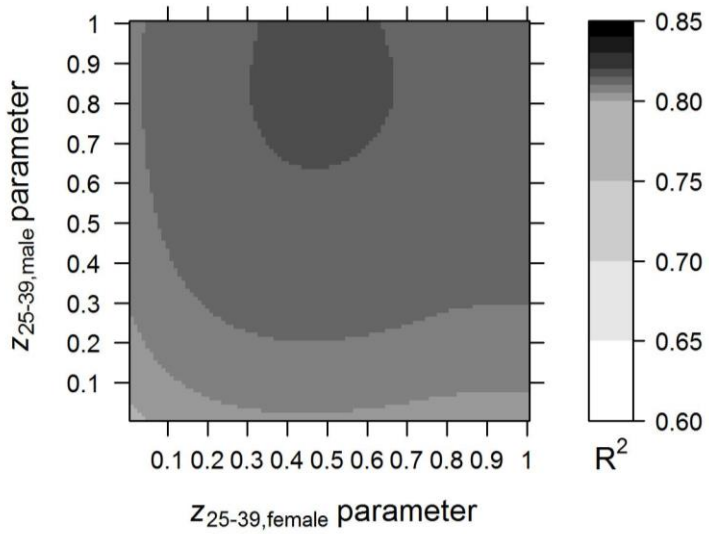


Figure 6. R^2 values of the correlation between the log of the number of Eurostat foreign-born migrants and the log of the number of corrected Facebook Network expats for different $z_{25\text{-}39,male}$ and $z_{25\text{-}39,female}$ parameter values. All the values of the remaining t and $z_{a,g}$ parameters, are the optimal.



3.3.2 Estimating lower and upper bounds

The Facebook Marketing API provides a rounded number of monthly active Facebook Network users that meet specific demographic criteria, not the precise figure. The applied rounding is proportional to the number of users, meaning that, for values between 0 and 1,000 the step is 10, for values between 1,000 and 10,000 the step is 100 and so forth. For example, if the real number of Facebook Network users is between 1,050 and 1,149 the Facebook API will return 1,100. Apart from the rounding, as already mentioned, Facebook uses a minimum response for the number of monthly active users; the minimum number was 20 when we collected the Facebook data. On 26 February 2018 Facebook increased the minimum response number to 1,000 users. To deal with the rounding

mechanism, we estimated the upper and the lower bounds of Facebook's API rounded responses. After identifying the lower and upper bound values for each rounded response, we estimated the lowest possible number and the highest possible number of the corrected Facebook Network expat users, based on equation (7).

$Fb\_cor^{i}_{a,g,o,d} = Fb\_raw^{i}_{a,g,o,d} / coefficient_{a,g,o,d}$ *where a = [15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64]* *g = [Male, Female]* *o ⊆ see Annex 1* *d = see Annex 2* *i = [rounded, lower bound, upper bound]*	(7)

where ^{i} indicates whether the number of Facebook Network users refers to the rounded numbers, to the lower bound or the upper bound.

4 Results & Discussion

In the following paragraphs, we present and discuss the results of the comparison between the Eurostat-estimated number of foreign-born migrants and the corrected Facebook Network-estimated number of expats. As shown in **Figure *7***, there is a higher correlation between the log of the corrected number of Facebook Network expats and the log of the number of Eurostat EU-born migrants compared to the log of the number of migrants born in a non-EU country. As shown on the left side of the **Figure *8*** Europeans aged between 50 and 64 are overestimated compared to other age groups.

Figure 7. Correlation between log(Eurostat foreign-born migrants) log(corrected Facebook Network expats) from EU countries (on the left side of the figure) and from non-EU (on the right side of the figure) in the 17 destination EU countries. Each observation represents a combination of previous residence and destination countries.

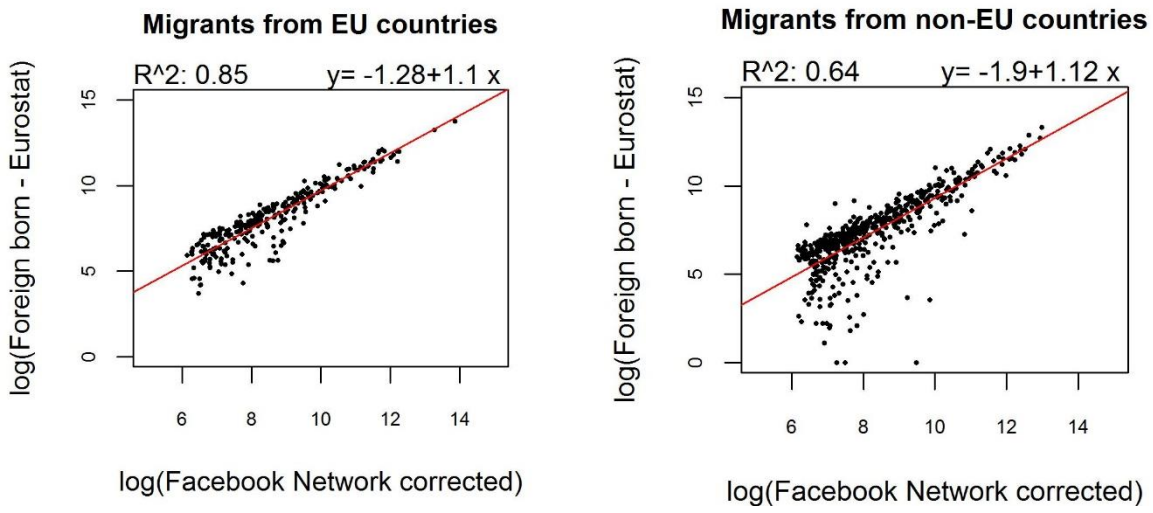
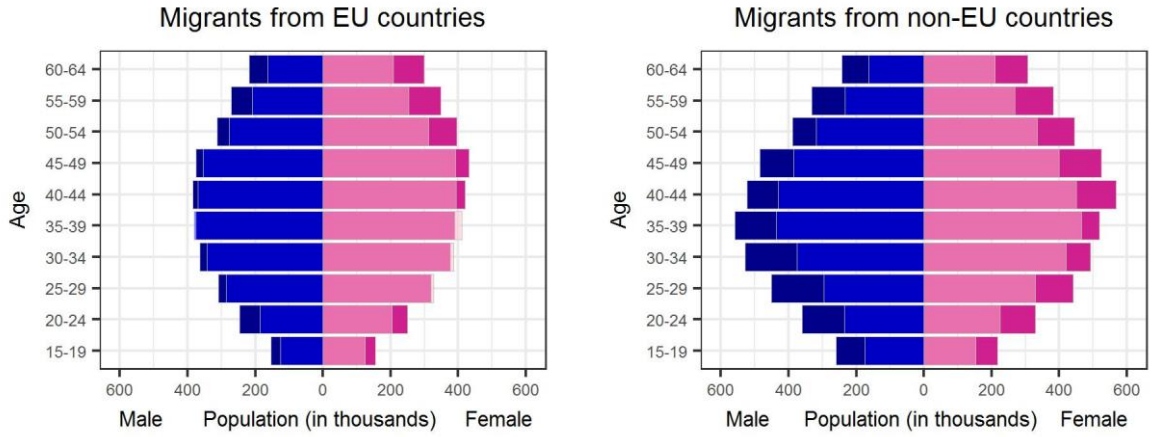


Figure 8. Demographic pyramids of Eurostat foreign-born migrants and of the corrected Facebook Network expat users from EU countries (on the left side of the figure) and from non-EU countries (on the right side of the figure) in the 17 destination EU countries. Light shading=Eurostat population, mid shading=Eurostat & corrected FN expat users, and dark shading = corrected FN expat users.



As shown in **Figure *9***, the number of Facebook Network expats and Eurostat migrants from Morocco in the 17 EU countries considered are well correlated (R^2 = 0.92, p<0.001). The slope of the fitted line is equal to 1.25 which means that the corrected number of Facebook

Network expats from Morocco are underestimated compared to the number of Moroccan-born migrants from Eurostat. **Table *2*** shows that Facebook Network expats from Brazil, Philippines and Serbia are those who are the most overestimated.

Figure 9. Correlation between the numbers of corrected Facebook Network expats from Morocco for each age group, gender and destination, and the number of migrants who were born in Morocco, according to Eurostat in the 17 EU destination countries. Points are labeled based on the county of destination.

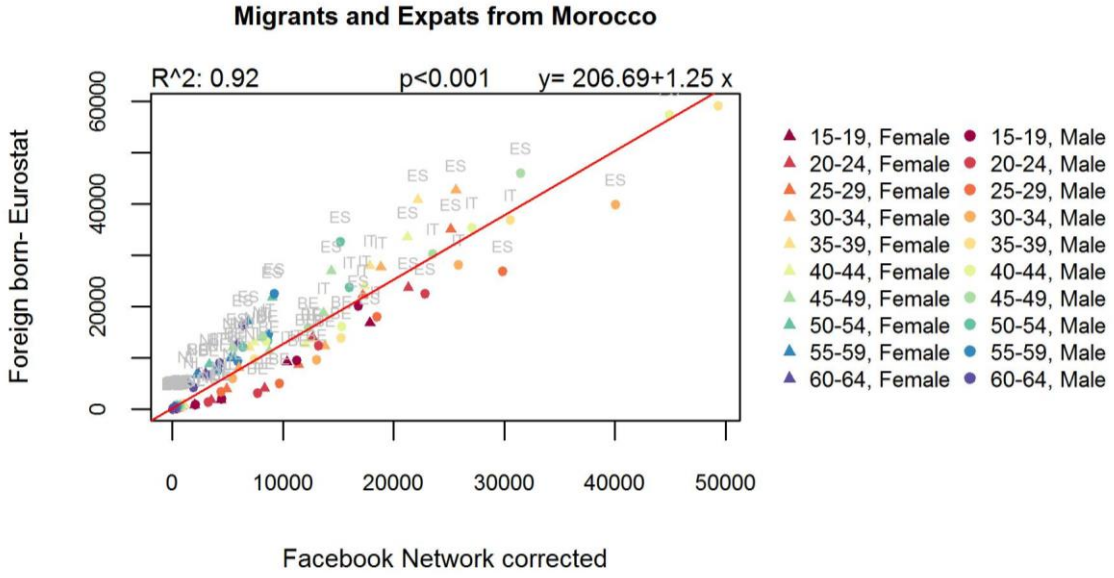


Table 2. Corrected numbers of Facebook Network expats aged 15 to 64 years old, by country of previous residence in the 17 EU countries listed in Annex 2. In the table, we present the 16 countries of previous residence with the highest number of corrected Facebook Network expats.

Previous residence	*Fb_cor rounded*	*Fb_cor lower*	*Fb_cor upper*	*Cob Eurostat*	*(Fb_cor_r-Cob)/Cob*
Romania	2,063,879	2,034,251	2,094,299	1,956,573	5%
Morocco	1,041,958	1,019,897	1,063,778	1,353,672	-23%
Germany	690,690	678,047	703,065	756,004	-9%
Philippines	560,417	546,959	574,569	233,805	140%
Poland	560,207	548,287	572,444	541,985	3%
Brazil	532,849	519,906	547,091	249,185	114%
France	518,918	509,710	527,784	476,993	9%
Colombia	518,714	506,261	531,783	402,007	29%
Serbia	488,596	479,512	497,693	211,744	131%
United Kingdom	480,928	473,876	489,997	333,615	44%
Russia	464,300	455,399	473,056	420,005	11%
Peru	413,156	404,688	423,599	289,275	43%
India	370,150	360,891	379,597	277,389	33%
Italy	364,144	357,536	370,890	249,014	46%
Venezuela	354,662	345,527	363,672	234,023	52%
Argentina	322,869	313,903	331,668	287,079	12%

Figure 10 shows that the corrected number of Facebook Network expats and the number of migrants in Spain according to Eurostat are well correlated (R^2 = 0.83, p<0.001). The slope of the fitted line is equal to 0.98, meaning that the corrected number of Facebook Network expats is almost equal to the number of foreign-born migrants in Spain as recorded by Eurostat. As shown in **Table *3***, in Slovakia, the Czech Republic and Bulgaria

the number of Facebook Network expats is much higher than the number of Eurostat foreign-born migrants. In the Czech Republic, the Facebook Network Slovakian expats aged 15 to 64 years old represent 40% of the total number of corrected Facebook Network expats. In Bulgaria the corrected Facebook Network Serbian expats in the same age group represents 30% of the total Facebook Network expats in the country.

Figure 10. Correlation between the corrected number of Facebook Network expats in Spain for each age group, gender and previous residence, and the number of foreign-born migrants who live in Spain according to Eurostat. Points are labelled according to migrants' country of previous residence.

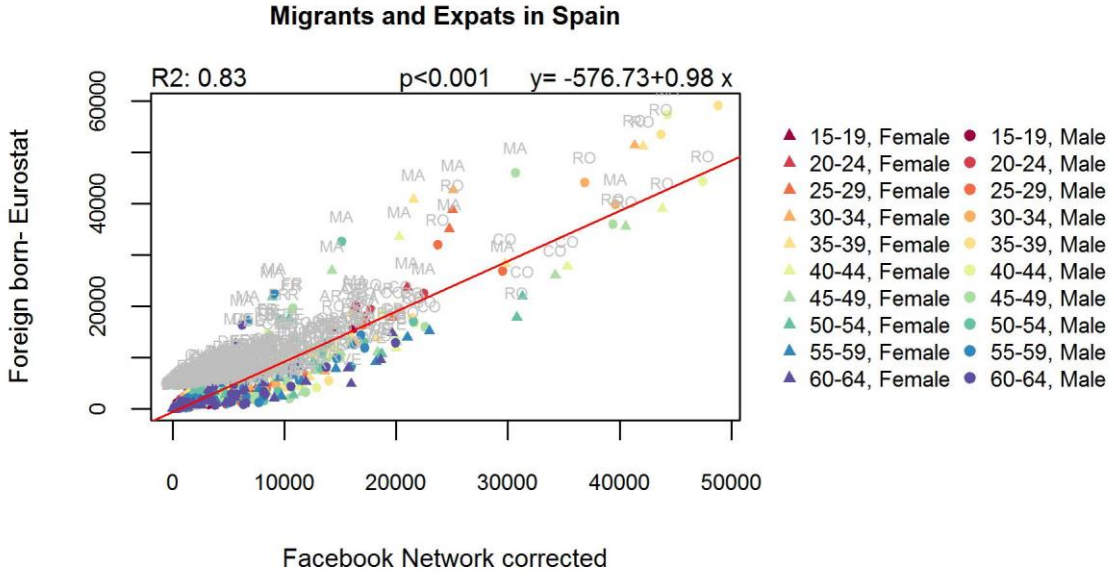


Table 3. The corrected numbers of Facebook Network expats of 82 countries of previous residence (see Annex 1) aged 15 to 64 years old, by country of destination.

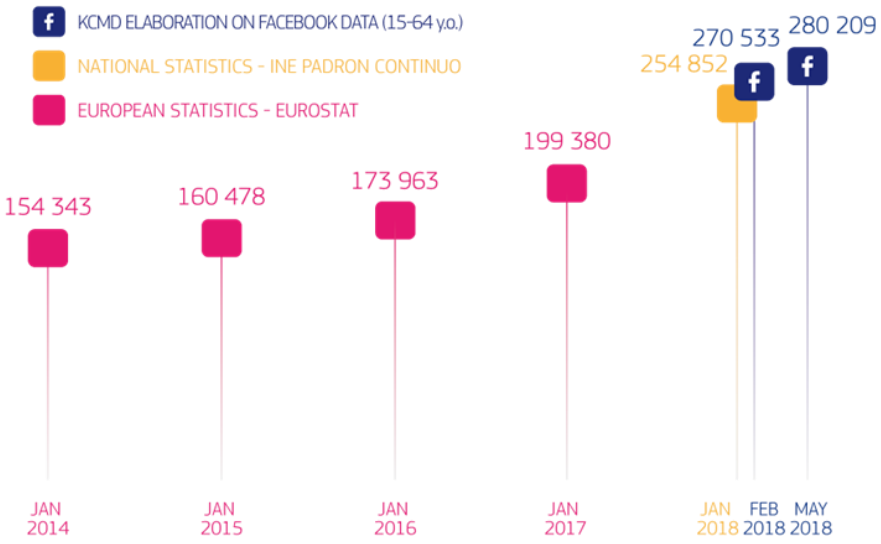
Destination	*Fb_cor rounded*	*Fb_cor lower*	*Fb_cor upper*	*Cob Eurostat*	*(Fb_cor_r-Cob)/Cob*
Spain	4,259,132	4,179,303	4,344,728	3,685,856	16%
Italy	4,179,378	4,104,143	4,257,452	3,497,140	20%
Belgium	1,454,117	1,422,116	1,484,748	1,046,252	39%
Netherlands	1,115,357	1,087,643	1,142,664	883,002	26%
Austria	1,027,823	1,001,621	1,053,292	732,773	40%
Sweden	667,449	647,850	686,512	651,148	3%
Czech Republic	486,512	466,226	507,118	244,268	99%
Denmark	304,558	292,951	315,180	315,857	-4%
Hungary	290,317	279,802	300,238	273,350	6%
Finland	232,444	218,621	245,436	165,080	41%
Luxembourg	181,487	173,715	188,785	170,422	6%
Slovakia	168,729	159,010	178,122	92,172	83%
Bulgaria	165,211	154,512	175,531	34,905	373%
Slovenia	63,649	57,452	69,529	32,529	96%
Lithuania	59,118	53,401	64,593	41,364	43%
Latvia	53,855	49,763	57,736	72,523	-26%
Estonia	35,079	32,119	37,868	74,678	-53%

To assess the corrected number Facebook Network expats, we present in Annex 3 the three most overestimated and the three most underestimated countries of previous residence for each country of destination. The overestimation or underestimation is determined by the difference between the corrected number of Facebook Network expats as of February

2018 and the number of foreign-born migrants according to Eurostat as of January 2017. In the next paragraphs we discuss the five most overestimated previous residence-destination combinations: Philippines-Spain (+120,546), Philippines-Italy (+112,925), Brazil-Spain (+106,385), Slovakia-Czech Republic (+105,409) and Venezuela-Spain (+91,809).

Still for the case of Spain, there were 270,533 corrected Venezuelan Facebook Network expats aged 15 to 64 in February 2018, and 178,724 Venezuelan-born migrants of the same age and 199,380 Venezuelan-born migrants of all ages, based on 2017 Eurostat data (**Figure *11***). On 1 January 2018, there were 223,463 Venezuelan-born migrants aged 15 to 64 years old in Spain, and 254,852 Venezuelan-born migrants of all ages, according to the National Statistical Office (Instituto Nacional del Estadística INE, 2018). In addition, there were 204,401 corrected Brazilian Facebook Network expats aged 15 to 64 years old in the country, and 98,016 and 119,137 Brazilian-born migrants based on 2017 Eurostat data and 2018 INE data, respectively. Brazilian male expats in Spain aged 40—64 are overestimated by 250% compared to the remaining 16 age-gender groups, which are overestimated by 85%. For Spain, the corrected number of Venezuelan Facebook Network expats is closer to the updated official data, while the corrected number of Brazilian Facebook Network expats is far above the updated official data.

Figure 11. Stocks of international migrants from Venezuela aged 15–64 in Spain, based on national and Eurostat statistics, vs. stocks of Venezuelan 'expats' based on authors' elaborations of Facebook data.



Again in the case of Spain, there are 160,163 corrected Facebook Network expats from Philippines in the 15—64 age group, and only 39,617and 42,437 Philippines-born migrants of the same age based on 2017 Eurostat data and 2018 INE data, respectively. Male expats from Philippines in Spain are overestimated by 413%, and female expat by 241%. In Italy, the corrected number of Facebook Network expats from Philippines in 2018 is 246,979 while, according to Eurostat, in 2017 there were only 134,054 Philippines-born migrants in the country. Male and female Facebook Network expats from Philippines aged 55 to 64 years old are overestimated by 138% compared to the remaining age groups, which are overestimated by 73%. According to Eurostat (2017b), only 105 and 35 Philippines nationals were found to be illegally present in 2017 in Italy and Spain, respectively. The difference between the Facebook Network and official data cannot be explained since there is no evidence of the presence of unregistered migrants from the Philippines in Italy or Spain.

Finally, in the Czech Republic there are 198,266 Slovakian corrected Facebook Network expats aged 15 to 64 years old, while according to Eurostat there are only 92,857 Slovak-

born migrants. This difference could be explained by the fact that the two countries were, until 1993, part of the Czechoslovakia, therefore the distinction between the two nationalities is prone to errors. The most overestimated demographic groups according to the corrected Facebook Network expat estimations are male and female expats aged 50 to 64 as well as young people aged 20 to 24 years old.

The authors would like to draw the readers' attention to a number of limitations in this study. First, we do not compare the same quantities of interest. The Eurostat foreign-born migrants dataset used for identifying the optimal parameters of our model refers to migrants who have their usual residence in an EU country (meaning they have lived in the country for a continuous period of at least 12 months, or intend to stay for at least one year) and are foreign-born.[20] Facebook seems to classify users as expats based mainly on the country of their self-reported hometown and/or the country of previous residence, while the duration of stay in the country of residence – a key criterion for distinguishing migrants from non-migrants, based on the UN definition – is unknown. Second, there is a one-year difference between the Eurostat foreign-born migrant statistics (January 2017) and the Facebook Network statistics (February 2018). Third, migrants may be more or less likely to use social media compared to non-migrants, while in our study we assume that both these groups use social media to the same degree. The proposed method may therefore overestimate or underestimate the number of expats. Fourth, the accuracy of Facebook's classification of expats is unknown. In addition some users might not declare their real personal information on Facebook platforms such as their "hometown". Fifth, official statistics refer to the sex of the individuals which is determined by their biological characteristics, while Facebook Network refers to the self-defined gender of the users. Lastly, we only estimated the number of Facebook Network expats aged 15 to 64 years old using 5 year age groups. This is because Facebook Network estimates are available for yearly age groups for users aged 13 years old up to 64 years old, and Eurostat statistics are available for 5-year age groups.

[20] http://ec.europa.eu/eurostat/cache/metadata/en/demo_pop_esms.htm

5 Conclusions

The study aimed to make a separate estimate of the number of expats in 17 EU countries based on the number of Facebook Network users who are classified by Facebook as expats. We proposed a method for correcting the over or under representativeness of Facebook Network users compared to the real population. To make Facebook Network user data more representative of the real population in the countries under analysis, we considered the penetration rates of the Facebook Network platforms (Facebook, Instagram, Messenger, and the Facebook Audience Network) for each age group and gender in the country of previous residence and the country of destination of a Facebook expat. We used official statistics from Eurostat to estimate how well migrants of different age and gender assimilate to the levels of Facebook Network usage in the destination country.

The paper shows that social media applications can be a timely, low-cost, and almost globally available data source for estimating stocks of international migrants, based on estimates of users who are classified as "expats." In the European context, it can be used for quantifying intra-EU mobility patterns of specific demographic groups, such as students, as well as for measuring migration movements that may not be captured in as timely a fashion by official statistics. For example, our methodology allowed us to now-cast an increase of Venezuelan migrants in Spain, validated by national official statistics, and not yet recorded in the Eurostat dataset. In addition to now-casting, social media data can provide insights into socio-economic indicators that are not collected yet by statistical offices. Examples, include personal interests, skills, educational attainment, and sector of employment disaggregated by country of previous residence, gender and age among other attributes.

There are, however, important limitations to using social media data sources for studying migration-related phenomena. There are well-founded concerns about the governance of the data, lack of transparency in processing users' personal information and insufficient information about the models that social media companies use for classifying users. For instance, we observed that Facebook Network numbers of expats for certain nationalities, e.g. Greek expats, were grossly underestimated. Such errors are maybe not related to non-representativeness of Facebook Network users but to probable errors in Facebook estimation models. In addition, as revealed by our study, Facebook network expats from Philippines in Spain and Italy are vastly overestimated and there is no evidence that such an overestimation may be valid. As stated in Zagheni et al. (2017), due to the lack of detailed documentation on how Facebook classifies expats, it is not possible to distinguish biases related to selection and non-representativeness of the users, or other noise and inconsistencies in the data.

The expected widespread diffusion of internet technologies in less developed countries where internet penetration rates are still low, presents a vast potential for studying not only quantitative characteristics of migration-related phenomena, such as migration numbers and trends, but also some qualitative characteristics, for instance migrants' interests and inclusion in society. The authors are planning to re-estimate the optimal parameters of their model discussed above following publication of the 2018 migration statistics by Eurostat.

References

Ahas, R., S. Silm, and M Tiru. 2017. *Tracking Transnationalism Originating in Estonia Through Mobile Roaming Data, Estonian Human Development Report 2017*. https://inimareng.ee/en/open-to-the-world/tracking-trans-nationalism-with-mobile-telephone-data/.

Barslund, Mikkel, and Matthias Busse. 2016. *How Mobile Is Tech Talent ? A Case Study of IT Professionals Based on Data from LinkedIn*. Brussels: CEPS.

Dubois, Antoine, Emilio Zagheni, Kiran Garimella, and Ingmar Weber. 2018. "Studying Migrant Assimilation Through Facebook Interests." *Under peer review*.

European Commission 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)*. Luxembourg.

———. 2018. "European Agenda on Migration." https://ec.europa.eu/home-affairs/what-we-do/policies/european-agenda-migration_en (March 20, 2018).

Eurostat. 2017a. "Population on 1 January by Age Group, Sex and Country of Birth." http://ec.europa.eu/eurostat/web/products-datasets/-/migr_pop3ctb (June 12, 2018).

———. 2017b. "Third Country Nationals Found to Be Illegally Present - Annual Data (Rounded)." http://ec.europa.eu/eurostat/web/products-datasets/-/migr_eipre (June 12, 2018).

Fatehkia, Masoomali, Ridhi Kashyap, and Ingmar Weber. 2018. "Using Facebook Ad Data to Track the Global Digital Gender Gap." *World Development* 107: 189–209.

Hawelka, Bartosz et al. 2014. "Geo-Located Twitter as Proxy for Global Mobility Patterns." *Cartography and Geographic Information Science* 41(3): 260–71.

Herdagdelen, Amaç, Bogdan State, Lada Adamic, and Winter Mason. 2016. "The Social Ties of Immigrant Communities in the United." In *Proceedings of the 8th ACM Conference on Web Science*, ACM, 78–84.

Instituto Nacional del Estadística. 2018. "Population by Country of Birth, Age (Five-Years Groups) and Sex, Advance of the Municipal Register at 1st January 2018. Provisional Results." http://www.ine.es/jaxiPx/Tabla.htm?path=/t20/e245/p04/provi/l0/&file=0ccaa005.px&L=1 (May 31, 2018).

International Telecommunication Union. 2016a. "Percentage of Individuals Using the Internet." https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx (March 20, 2018).

———. 2016b. "Percentage of Individuals Using the Internet by Gender." https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx (March 27, 2018).

Messias, Johnnatan, Fabricio Benevenuto, Ingmar Weber, and Emilio Zagheni. 2016. "From Migration Corridors to Clusters: The Value of Google+ Data for Migration Studies." In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*, 421–28.

OECD and World bank. 2010. "DIOC-E: Database on Immigrants in OECD and Non-OECD Countries." http://www.oecd.org/els/mig/dioc.htm (March 22, 2018).

Rango, Marzia, and Michele Vespe. 2017. *Big Data and Alternative Data Sources on Migration : From Case-Studies to Policy Support Summary Report*. Ispra, Italy.

Raymer, J. et al. 2013. "Integrated Modeling of European Migration." *Journal of the American Statistical Association* 108(503): 801–19.

Sinn, A, A Kreienbrink, and H D Von Loeffelholz. 2005. *Illegally Resident Third-Country*

Nationals in Germany Policy Approaches, Profile and Social Situation. Nürnberg Germany.

Smith, Aaron, and Monica Anderson. 2018. *Social Media Use in 2018*. http://www.pewinternet.org/2018/03/01/social-media-use-in-2018/.

State, Bogdan, Weber Ingmar, and Emilio Zagheni. 2013. "Studying Inter-National Mobility through IP Geolocation." In *WSDM'13*, Rome, Italy: ACM, 265–74.

State, Bogdan, Mario Rodriguez, Dirk Helbing, and Emilio Zagheni. 2014. "Migration of Professionals to the U.S. Evidence from LinkedIn Data." *Social Informatics* 8851: 531–43.

UN/DESA. 2008. "United Nations Global Migration Database (UNGMD)." https://esa.un.org/unmigration/.

———. 2017a. "Population by 5-Year Age Groups, Annually from 1950 to 2100: Medium Projection Variant." https://esa.un.org/unpd/wpp/Download/Standard/CSV/ (March 20, 2018).

———. 2017b. *Trends in International Migrant Stock: The 2017 Revision (United Nations Database, POP/DB/MIG/Stock/Rev.2017)*. http://www.un.org/en/development/desa/population/migration/data/estimates2/estimates17.shtml.

UN. 2018. "Big Data UN Global Working Group." https://unstats.un.org/bigdata/ (March 22, 2018).

United Nations Development Programme. 2016. United Nations Development Programme *Human Development Report 2016*.

Vespe, Michele, Fabrizio Natale, and Luca Pappalardo. 2017. "Measuring Irregular Migration: Innovative Data Practices." *Migration Policy Practice* VII(2): 26–33.

Vilhena, Daniela Vono De. 2018. "Knowing the Unknown Irregular Migration in Germany." *Population Europe Discussion Papers Series* (07).

Zagheni, Emilio, Kiran Venkata Rama Garimella, Ingmar Weber, and Bogdan State. 2014. "Inferring International and Internal Migration Patterns from Twitter Data." In *Proceedings of the 23rd International Conference on World Wide Web*, Seoul, Korea: ACM, 439–44.

Zagheni, Emilio, and Ingmar Weber. 2012. "You Are Where You E-Mail: Using e-Mail Data to Estimate International Migration Rates." *Proceedings of the 3rd Annual ACM Web Science …*: 1–10.

Zagheni, Emilio, Ingmar Weber, and Krishna Gummadi. 2017. "Leveraging Facebook's Advertising Platform to Monitor Stocks of Migrants." *Population and Development Review* 43(4): 721–34.

List of abbreviations and definitions

API	Application Programming Interface
BD4M	Big Data for Migration Alliance
EC	European Commission
EU	European Union
FN	Facebook Network
GDI	Gender Development Index
GMDAC	Global Migration Data Analysis Centre
OECD	Organization for Economic Co-operation and Development
UN	United Nations
UNDESA	United Nations' Department of Economic and Social Affairs
US	United States

List of figures

List of tables

Annexes

Annex 1. List of countries of previous residence taken into consideration

1	Algeria	29	Indonesia	57	Portugal
2	Argentina	30	Ireland	58	Qatar
3	Australia	31	Israel	59	Romania
4	Austria	32	Italy	60	Russia
5	Bangladesh	33	Ivory Coast	61	Rwanda
6	Belgium	34	Jamaica	62	Senegal
7	Brazil	35	Japan	63	Serbia
8	Cameroon	36	Jordan	64	Sierra Leone
9	Canada	37	Kenya	65	Singapore
10	Chile	38	KSA	66	Slovakia
11	Colombia	39	Kuwait	67	Slovenia
12	Congo DRC	40	Latvia	68	South Africa
13	Cyprus	41	Lebanon	69	South Korea
14	Czech Republic	42	Lithuania	70	Spain
15	Denmark	43	Luxembourg	71	Sri Lanka
16	Dominican Republic	44	Malaysia	72	Sweden
17	El Salvador	45	Malta	73	Switzerland
18	Estonia	46	Mexico	74	Tanzania
19	Ethiopia	47	Morocco	75	Thailand
20	Finland	48	Nepal	76	UAE
21	France	49	Netherlands	77	Uganda
22	Germany	50	New Zealand	78	United Kingdom
23	Ghana	51	Nicaragua	79	Venezuela
24	Guatemala	52	Nigeria	80	Vietnam
25	Haiti	53	Norway	81	Zambia
26	Honduras	54	Peru	82	Zimbabwe
27	Hungary	55	Philippines		
28	India	56	Poland		

Annex 2. List of countries of destination taken into consideration

1	Austria	10	Italy
2	Belgium	11	Lithuania
3	Bulgaria	12	Luxembourg
4	Czech Republic	13	Latvia
5	Denmark	14	Netherlands
6	Estonia	15	Sweden
7	Spain	16	Slovenia
8	Finland	17	Slovakia
9	Hungary		

Annex 3. Three most overestimated and underestimated countries of previous residence for each country of destination

Destination	Previous residence	Fb raw	Fb cor rounded	Fb cor lower	Fb cor upper	Cob Eurostat	Fb_cor_r - Cob	(Fb_cor_r-Cob) /Cob
Latvia	United Kingdom	2,570	2,715	2,582	2,826	217	2,498	1151%
Latvia	India	2,216	2,479	2,009	2,939	214	2,265	1058%
Latvia	Brazil	1,034	1,154	864	1,430	25	1,129	4516%
Latvia	Russia	15,630	34,970	34,630	35,260	60,922	-25,952	-43%
Latvia	Lithuania	2,450	3,746	3,571	3,888	7,309	-3,563	-49%
Latvia	Estonia	830	1,180	909	1,437	1,818	-638	-35%
Estonia	Congo DRC	800	1,170	829	1,487	8	1,162	14525%
Estonia	United Kingdom	1,334	1,331	1,108	1,544	477	854	179%
Estonia	India	952	1,179	762	1,591	444	735	166%
Estonia	Russia	10,870	22,165	21,968	22,324	62,644	-40,479	-65%
Estonia	Latvia	1,662	2,157	1,953	2,333	3,662	-1,505	-41%
Estonia	Germany	776	973	653	1,285	1,624	-651	-40%
Slovenia	Serbia	21,630	33,356	32,652	34,021	17,695	15,661	89%
Slovenia	Italy	3,520	4,541	4,392	4,659	2,354	2,187	93%
Slovenia	Russia	2,430	4,198	3,969	4,387	2,193	2,005	91%
Slovenia	Germany	3,020	4,182	3,990	4,334	5,590	-1,408	-25%
Slovenia	Austria	1,146	1,544	1,286	1,781	1,671	-127	-8%
Slovakia	Serbia	13,700	19,582	19,211	19,959	1,708	17,874	1046%
Slovakia	Czech Republic	39,850	63,277	61,742	64,975	54,679	8,598	16%
Slovakia	Hungary	11,350	13,833	13,572	14,059	6,193	7,640	123%
Slovakia	Romania	4,744	6,148	5,896	6,372	7,642	-1,494	-20%
Slovakia	Poland	3,320	4,974	4,756	5,152	4,999	-25	-1%
Sweden	Serbia	30,330	33,846	33,009	34,656	9,644	24,202	251%
Sweden	Philippines	21,450	23,730	23,224	24,282	11,475	12,255	107%
Sweden	Thailand	41,260	43,057	42,393	43,700	35,669	7,388	21%
Sweden	Finland	29,250	37,361	36,437	38,398	75,567	-38,206	-51%
Sweden	Poland	49,190	56,467	55,350	57,561	71,537	-15,070	-21%
Sweden	South Korea	2,338	2,299	2,104	2,473	9,521	-7,222	-76%
Netherlands	Poland	155,620	164,176	160,238	168,733	114,359	49,817	44%
Netherlands	Russia	19,410	24,317	23,738	24,953	2,593	21,724	838%
Netherlands	Philippines	28,680	33,367	32,534	34,182	11,810	21,557	183%
Netherlands	Morocco	79,560	94,607	93,502	95,679	144,094	-49,487	-34%
Netherlands	Germany	62,800	75,725	74,365	77,056	85,193	-9,468	-11%
Netherlands	Ethiopia	7,472	8,035	7,864	8,183	14,238	-6,203	-44%
Luxembourg	Portugal	67,920	74,820	73,763	75,854	61,551	13,269	22%
Luxembourg	Brazil	6,320	7,056	6,921	7,162	2,670	4,386	164%
Luxembourg	Serbia	2,664	3,035	2,836	3,215	1,243	1,792	144%
Luxembourg	France	23,310	26,364	25,760	26,937	31,868	-5,504	-17%
Luxembourg	Belgium	10,770	12,741	12,613	12,844	15,340	-2,599	-17%
Luxembourg	Poland	2,552	3,209	3,036	3,355	3,907	-698	-18%
Lithuania	United Kingdom	6,120	5,475	5,343	5,576	272	5,203	1913%
Lithuania	India	3,510	3,295	2,947	3,625	136	3,159	2323%
Lithuania	Brazil	3,308	3,001	2,780	3,205	15	2,986	19907%
Lithuania	Russia	14,970	24,591	24,283	24,927	33,473	-8,882	-27%
Lithuania	Latvia	2,240	3,089	2,918	3,226	4,527	-1,438	-32%

Italy	Philippines	197,800	246,979	239,926	254,630	134,054	112,925	84%
Italy	Brazil	163,700	176,204	172,425	180,941	100,733	75,471	75%
Italy	Romania	870,900	1,045,430	1,032,800	1,058,032	970,014	75,416	8%
Italy	Morocco	282,870	303,389	296,908	309,840	391,268	-87,879	-22%
Italy	Switzerland	72,790	102,161	100,982	103,313	179,049	-76,888	-43%
Italy	Germany	103,100	128,417	127,060	129,855	182,030	-53,613	-29%
Hungary	Serbia	35,240	44,299	43,265	45,306	31,677	12,622	40%
Hungary	Congo DRC	6,350	10,096	9,925	10,234	39	10,057	25787%
Hungary	Slovakia	13,630	16,212	15,888	16,581	9,793	6,419	66%
Hungary	Romania	87,100	116,432	114,975	117,861	166,539	-50,107	-30%
Hungary	Germany	11,830	14,895	14,683	15,112	19,455	-4,560	-23%
Hungary	Switzerland	970	1,235	1,091	1,354	1,386	-151	-11%
Finland	Russia	25,940	40,574	39,515	41,597	9,719	30,855	317%
Finland	Estonia	38,730	48,808	47,790	49,800	39,064	9,744	25%
Finland	Congo DRC	5,550	8,514	8,336	8,655	1,632	6,882	422%
Finland	Sweden	11,640	13,553	13,429	13,656	28,741	-15,188	-53%
Finland	Hungary	1,574	1,821	1,654	1,965	2,011	-190	-9%
Finland	Israel	550	544	339	743	695	-151	-22%
Spain	Philippines	131,300	160,163	158,778	161,517	39,617	120,546	304%
Spain	Brazil	185,900	204,401	199,611	209,742	98,016	106,385	109%
Spain	Venezuela	213,500	270,533	263,896	277,140	178,724	91,809	51%
Spain	Morocco	418,900	435,474	427,845	443,071	613,424	-177,950	-29%
Spain	France	114,700	137,032	135,687	138,349	169,510	-32,478	-19%
Spain	Germany	81,400	114,541	113,033	116,012	146,295	-31,754	-22%
Denmark	Philippines	16,790	17,504	17,232	17,756	10,366	7,138	69%
Denmark	Serbia	5,540	6,241	6,114	6,344	1,209	5,032	416%
Denmark	Thailand	13,770	15,162	14,779	15,520	11,697	3,465	30%
Denmark	Poland	25,120	27,326	26,754	27,873	34,391	-7,065	-21%
Denmark	South Korea	1,312	1,329	1,120	1,525	8,285	-6,956	-84%
Denmark	Germany	16,750	18,792	18,479	19,123	25,737	-6,945	-27%
Czech Republic	Slovakia	142,900	198,266	194,060	203,622	92,857	105,409	114%
Czech Republic	Vietnam	44,350	68,710	67,285	70,096	44,271	24,439	55%
Czech Republic	Romania	16,190	19,821	19,405	20,207	6,656	13,165	198%
Bulgaria	Serbia	37,450	49,991	49,016	50,931	1,436	48,555	3381%
Bulgaria	Russia	18,950	31,618	31,086	32,267	20,445	11,173	55%
Bulgaria	United Kingdom	11,930	13,142	12,922	13,339	2,456	10,686	435%
Belgium	France	171,400	178,461	175,205	181,692	131,967	46,494	35%
Belgium	Italy	92,100	100,199	98,980	101,394	61,298	38,901	63%
Belgium	Brazil	43,690	46,258	45,317	47,175	12,708	33,550	264%
Belgium	Germany	39,560	49,225	48,143	50,278	58,149	-8,924	-15%
Belgium	Morocco	164,280	178,847	174,386	183,280	184,202	-5,355	-3%
Belgium	Luxembourg	5,450	5,473	5,350	5,569	7,432	-1,959	-26%
Austria	Serbia	129,700	187,619	185,006	190,515	107,816	79,803	74%
Austria	Germany	145,300	209,380	205,141	213,582	164,284	45,096	27%
Austria	Romania	94,150	122,169	120,660	123,648	90,316	31,853	35%
Austria	Poland	36,700	57,760	56,547	58,969	61,441	-3,681	-6%
Austria	Slovenia	8,420	10,561	10,381	10,707	13,874	-3,313	-24%
Austria	Switzerland	6,870	9,703	9,536	9,838	11,235	-1,532	-14%

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: http://europea.eu/contact

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

- at the following standard number: +32 22999696, or

- by electronic mail via: http://europa.eu/contact

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: http://europa.eu

EU publications
You can download or order free and priced EU publications from EU Bookshop at: http://bookshop.europa.eu. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see http://europa.eu/contact).

KJ-NA-29273-EN-N

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

EU Science Hub
ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub - Joint Research Centre

Joint Research Centre

EU Science Hub

Publications Office

doi:10.2760/964282

ISBN 978-92-79-87989-0