European Commission

Measuring Labour Mobility and Migration Using Big Data

Exploring the potential of social-media data for measuring EU mobility flows and stocks of EU movers

Cloé Gendronneau, Arkadiusz Wiśniowski, Dilek Yildiz, Emilio Zagheni, Lee Fiorio, Yuan Hsiao, Martin Stepanek, Ingmar Weber, Guy Abel, Stijn Hoorens

Social Europe

EUROPEAN COMMISSION

Directorate-General for Employment, Social Affairs and Inclusion
Directorate A — Employment and Social Governance
Unit A.4 — Thematic Analysis
Contact: Simone Rosini and Stefano Filauro
E-mail: Simone.ROSINI@ec.europa.eu or Stefano.FILAURO@ec.europa.eu *European Commission*
B-1049 Brussels

Measuring Labour Mobility and Migration Using Big Data

European Commission
Directorate-General for Employment Social Affairs and Inclusion
Directorate A

2019

Luxembourg: Publications Office of the European Union, 2019

PDF ISBN 978-92-76-08749-6 doi:10.2767/474282 KE-01-19-556-EN-N

PREFACE

Internal freedom of movement is one of the European Union's four fundamental freedoms and is necessary for the EU single market to function. Yet official statistics on the migration of workers are constrained. They are limited in their ability to distinguish population subgroups, come with a considerable time lag of a year or more and are fully reliant on individual member states' measurements. Current data sources also tend to underestimate the overall extent of mobility by not covering short-term moves and not capturing the most recent movers.

Given the importance of freedom of movement, it is crucial for European institutions to have robust, rich and up-to-date data to monitor it. Big data sources from social media, such as Twitter and Facebook, offer opportunities to bridge the gap between official statistics and recent migration trends.

The European Commission's Directorate-General for Employment Social Affairs and Inclusion commissioned RAND Europe to investigate social media data's potential use for measuring EU mobility. Researchers collaborated with experts from the Vienna Institute for Demography, the University of Manchester, Washington University, Max Planck Institute for Demographic Research and the Qatar Computing Research Institute. This report discusses the activities, results and findings of this study and presents recommendations for future work in this area. It is aimed at a specialist audience of academics and policy-makers with a specific interest in measuring and monitoring migration flows.

RAND Europe is an independent, not-for-profit policy research organisation that helps to improve policy and decision-making through research and analysis. This report has been peer-reviewed in accordance with RAND's quality assurance standards**.**

For more information about RAND Europe or this document, please contact Stijn Hoorens (hoorens@rand.org).

RAND Europe
Rue de la Loi 82, Bte 3
1040 Brussels
Belgium
Tel: +32 (2) 669 2400

RAND Europe
Westbrook Centre, Milton Road
Cambridge CB4 1YG
United Kingdom
Tel: +44 1223 353 329

TABLE OF CONTENTS

TABLE OF TABLES

TABLE OF FIGURES

ABBREVIATIONS

API	Application Programming Interface
CES	Conference of European Statisticians
DAU	Daily Active Users
EFTA	European Free Trade Area
EU	European Union
IIASA	International Institute for Applied Systems Analysis
IMEM	Integrated Modelling of European Migration
IP	Internet Protocol
JAGS	Just Another Gibbs Sampler
LFS	Labour Force Survey
MAU	Monthly Active Users
MCMC	Markov Chain Monte Carlo
MIMOSA	Migration Modelling for Statistical Analyses
OECD	Organisation for Economic Cooperation and Development
QA	Quality Assurance
RMSE	Root Mean Squared Error
SEC	Securities and Exchange Commission
UN	United Nations

EXECUTIVE SUMMARY

The free movement of people and workers, introduced by the Maastricht Treaty, ensures that all EU citizens and their family members have the right to seek work, become self-employed, be a pensioner or student anywhere across the EU. Latest official available data from Eurostat report that 17.6 million EU citizens are currently living and working abroad and 4 per cent of the EU population of working age lives in another EU country.

Having up-to-date information about the nature and extent of such mobility is important for policy making, such as labour market policy or social services. However, timely and reliable statistics on the number of EU citizens residing in or moving across other Member states are difficult to obtain. Official statistics on EU movers are developed by national offices of statistics and published by Eurostat, but they come with a considerable time lag of about two years.

With the rise of the Internet, new data sources potentially offer opportunities to complement traditional sources for EU mobility statistics. In particular, the availability of high quantities of individual geo-tagged data from social media (i.e. metadata that contains information linked to the geographical location of the content) has opened new opportunities. In this report, we study these opportunities in detail, exploring the possibilities of using these new "big data" sources – focusing in particular on social-media data, such as those from Facebook and Twitter – to develop a method to provide more timely and potentially more accurate EU mobility statistics. As such, we **investigate the potential of geo-referenced social-media data to facilitate "nowcasting", providing nearly real-time estimates that will serve as early warnings about changes in EU mobility.** In order to tailor the study's scope to the policy area of European Commission's Employment, Social Affairs and Equal Opportunities DG, we focus on intra-EU migration – or in other words, EU mobility.

Against this background, we have collected a number of different datasets to **develop methodologies to provide recent estimates of stocks of EU movers and EU mobility flows using social-media data, complemented with traditional data sources**.

Stocks and flows are concepts regularly used in migration research, stemming from the field of system dynamics. A **stock** is a measure of a quantity at one specific time that may have accumulated in the past – in this case, EU movers residing in a different member state. A **flow** variable is roughly equivalent to a rate or a speed measured over an interval of time. Mobility flows are expressed as the total number of EU citizens per time unit (e.g. year) moving from one member state to the other.

This document reports the results of this attempt to develop such an approach. The report takes stock of the advantages and disadvantages of new and traditional data sources, and what is known about new methodologies using social-media data. It subsequently describes the data collected, the proposed models for estimating stocks and flows, and the results of the application of these models using real-world data. Furthermore, the report offers direction for the European Commission to potentially use this approach in the policy process.

A. Taking stock of existing approaches

Whilst EU (labour) mobility and mobile workers or movers are the terms used in the policy context of the European Commission's Employment, Social Affairs and Inclusion DG, the methodologies to estimate these phenomena are no different than those used for migration and migrants in general. Hence, the academic literature refers to migrants as those who have established permanent residence in a new country. Consistent with the Commission's terminology, **we will use the terms mobility or EU movers** when referring to intra-EU migration, unless the terms migration or migrants are explicitly used in the variable definitions of data sources or in the literature.

When reviewing the available literature reporting on methods and data sources used to measure migration, the main conclusion is that no single source of data is of sufficient quality to provide accurate, timely and unbiased information on migration:

- **Census data** are reliable as they cover the entire population. However, they are only available once every 5 or 10 years.
- **Aggregated statistics from national offices** (e.g by Eurostat) are available every year and their coverage is also good. However, aggregating these national statistics causes a time lag of typically one or two years, and statistics based on registries tend to underestimate migration.
- **Data from standardised household surveys** (e.g. Labour Force Surveys) are often timely as they are set up as continuous data-collection efforts. However, because migration is a relatively rare event, the sample sizes are relatively small.
- **Social-media data** are available almost in real-time, can represent relatively good coverage of the population and can offer granularity on location and demographics. However, samples are biased and not representative of the population.

This applies to both traditional and officially used data sources and new forms of data, including those from social media. Both traditional and new forms of data have their strengths and both suffer from different shortcomings. The main advantage of social-media data is their volume and their timeliness, and hence their potential to facilitate "nowcasting".

Based on the lessons from existing research, **we take a two-pronged approach to develop models to estimate stocks of EU movers and EU mobility flows**. First, we focus on estimating stocks of EU movers, which facilitates comparisons with other methods, as stocks data are more easily measured by national statistical agencies and/or relevant national administrations, and there is more consistency across EU Member states in the definitions for stocks. Moreover, distinguishing EU mobility from other types of movement, such as travel, is difficult in social-media data sources. In contrast, proxies for estimates of migrant stocks can be obtained directly from a platform such as Facebook. We first present the methodology for estimating stocks of EU movers, the data sources it requires, and how it can be used by policy makers. Subsequently, we present a first proof of concept methodology for estimating flows, and a set of recommendations for further research.

B. Approach and data to measure stocks of EU movers

The approach presented in this report for measuring stocks of EU movers uses geo-referenced big data from Facebook, complemented with traditional migration statistics from Eurostat, data from the EU Labour Force Survey (EU-LFS) and Population and Housing Census data. **Facebook** is the largest social networking platform with 2.3 billion users who log on at least once a month worldwide, and about 262 million of such monthly active users (MAU) in the EU. Facebook coverage varies in the EU between 56 per cent of the working-age population in Germany to 92 per cent in Denmark. Users share personal information with the platform and provide and consume media content that is either publicly accessible or only from or for their friend connections. Facebook provides access to a large amount of aggregate and anonymous data on the characteristics of its users through its marketing application programming interface (API). While Facebook network users may not be a perfectly representative sample of the general population, the data available through the marketing API provides information on nearly three quarters of the EU population between 15 and 64 years old. We have used the Facebook Marketing API to collect data about the approximate number of users within a Member state that fit the selected criterion of "People who used to live in [country] who now live abroad".

In order to facilitate "nowcasting" of stocks of EU movers, this study proposes Bayesian inference methods to estimate model parameters, which takes into account the limitations of the various datasets (see Box below). This approach permits explicit inclusion of qualitative information on each data source in terms of prior probability distributions of the dataset's coverage and undercounting (i.e. the number of EU movers missing in the data due to an under-reporting of the total arrivals).

Box 1: Modelling approach to estimate stocks of EU movers

A Bayesian inferential framework offers a powerful mechanism to combine data sources and provide measures of uncertainty. Models previously developed have primarily combined traditional migration data sources. Nevertheless, in this study, we adapt the basic methodologies of these former models to combine data from both traditional and new, social-media sources. The framework requires specifying a probability distribution that represents uncertainty about those "unknowns" (so called "priors") by combining the observed data, represented by a statistical model. These priors contain information about the characteristics and quality of the data source, indicating the errors which are known or are believed to exist.

The methodology estimates the unobserved stocks of EU movers for each EU member state by EU country of birth, by broad age groups and gender for years between 2011 and 2018. First, we use Bayesian inference to estimate the true stocks of EU movers for each year. Then, as a second step, we further disaggregate the migrant stocks by broad age group and gender.

The level of migrant stocks in reported data tends to be systematically biased compared with estimates based on census data. In order to obtain an estimate of the true migration quantity, our model adjusts the reported stocks of EU movers using meta-data on the bias and the accuracy of the reported migrant stocks via five measurement models – one for each data source: Facebook Monthly Active Users (MAU), Facebook Daily Active Users (DAU), Eurostat statistics, Labour Force Survey (LFS) and Census data.

The hierarchy in our Bayesian model is overlaid with a migration module that estimates true stocks of EU movers. The migration model takes a non-theoretical perspective that allows forecasting and dealing with missing values across time and countries. In doing so, the autoregressive "nowcasting" model (in which the current stocks of EU movers are estimated using past observations) creates a "bridge" between the official statistics and traditional data sources on the one hand, and the newly collected social-media data on the other.

C. Estimating stocks of EU movers

For each year between 2011 and 2018, we estimate the number of EU movers for each combination of origin and destination within the EU. In total, our model estimates just over 15 million EU citizens of working age (between 15 and 64) living in another EU member state than their country of birth in 2018, a slight increase compared to 2016 and 2017. Compared to previous years, the estimates are more uncertain (i.e. predictive intervals are wider) in 2018, the year for which we only have Facebook data available. Consistent with the Eurostat migrant stocks, our model estimates higher numbers of female migrants than male migrants. Male and female migrant stocks follow the same trend over time at different levels.

Additionally, the predictive intervals for countries such as Germany, the UK and France are wider than the predictive intervals of other countries, because their estimates are based on less information. This is due to missing values in some data sources, and the size of migrant stock in these countries being higher than the migrant stock in other countries. Within the countries with high immigration, the figure shows that the increase in the numbers of migrants in the United Kingdom and Germany are slowing, while there is a decreasing trend in France, Italy and Spain.

The model estimates stocks of male and female EU movers in three age groups (15–24, 25–54 and 55–64). As expected, the numbers of male migrants aged 25 to 54 are highest in the UK and in Germany compared to other member states. The same estimates show a decreasing trend in Spain, Italy and France. With the exception of Austria, Belgium and Netherlands, the other countries are estimated to have fewer than 250,000 male migrants in this age interval.

When comparing the results with official statistics, we observe that for most of the countries the estimates are comparable to those reported by Eurostat, taking into account the missing data for some countries in the official statistics. We observe no pattern of under- or over-counting over time. However, for a handful of countries (mainly Italy and Spain) our estimates are higher than the reported number of migrants in Eurostat, suggesting that these countries may have missing observations.

D. Approach and data to measure EU mobility flows

In addition to estimating stocks of EU movers, we also made a first attempt to develop a method for "nowcasting" mobility flows. Whilst the Facebook Marketing API makes data available on total number of users and the number of users who live in a country and used to live somewhere else, Twitter offers real-time information on the location (of a small share) of their users. We explored if these data can be used to estimate flows. In the approach explained in this report, we use the data from Twitter and migration statistics available from Eurostat.

Twitter is a popular social networking platform that enables users to post and interact with messages known as "tweets". Twitter had 335 million monthly active users in 2018, sending more than 200 billion tweets per year, and its coverage in the EU is much lower than Facebook, varying from 32 per cent of the population in Ireland to 2 per cent in Romania. A key feature of the Twitter data is that a small proportion of tweets provides information about the user's location. The main advantages of Twitter data are their good accessibility, coverage across countries and population subgroups, and relative simplicity of the available information. While the population of Twitter users may not be representative of the EU population overall, the spatial and temporal detail of these data provide a unique opportunity to study population mobility.

As a first proof of concept methodology to estimate mobility flows, we used a basic estimation framework to combine the Twitter data with the Eurostat statistics. From the model, we aimed to estimate EU mobility flows for each member state (i.e., the proportion of EU movers compared to the population size).

The initial results of this approach indicate that the model running with only Eurostat data outperforms the application that includes Twitter data. The prediction accuracy is much lower for the joint model. This is the case for every country in the model. There are multiple reasons that may have caused the discrepancy in the model performances, which require further investigation and testing beyond the scope of this study.

E. The potential of social-media data to facilitate "nowcasting" EU mobility

The aim of this study was to investigate the potential of geo-referenced social-media data to facilitate "nowcasting" stocks of EU movers and mobility flows, providing more recent estimates than official statistics to serve as early warning signs for the European Commission.

The first **results of the application of the stocks model are experimental, but they are promising**. Complementary research and data would be needed to improve the robustness of the new estimates. In case the European Commission wishes to continue the development of this approach, we have formulated some steps to replicate the methodology and update the estimates with more recent data. The first step is to collect and prepare the data sources, and the second step is to update the prior distributions used in the Bayesian model.

The main limitation of the current application is the lack of overlapping time-series between the official data on the one hand and Facebook data on the other. Furthermore, any future changes in the representativeness of these datasets, for example due to declining popularity of Facebook, will affect the reliability of the model.

The approach taken to estimate **EU mobility flows has not yet offered any plausible results.** We therefore do not recommend applying this approach, in its current form, to estimate EU mobility flows. Further research would be required to develop a robust and reliable sample of Twitter data. Notwithstanding the required improvements, it is important to note that Twitter imposes restrictions to developers with regards to sharing data with third parties. Therefore, if the European Commission were to consider further pursuing an approach to estimate flows based on Twitter data, we would recommend

starting to build a dataset in accordance with the approach specified in this report to build its own sample that could be used for further research in this area.

F. Improving the approach and estimates

We discuss several recommendations to improve the proposed method for estimating stocks of EU movers in the future.

Investigate different migration models. While we employed a non-theoretical perspective that permits forecasting and dealing with missing values across time and countries, further work would be required to assess various theory-based models, such as a gravity model that considers the drivers (i.e. pull and push factors) of migration.

Longer time-series LFS data. The Labour Force Surveys (LFS) have been conducted for many years. However, in this project we have only been able to use data collected in 2016 and 2017. Incorporating longer time series of LFS data in the model, for all countries, allows benchmarking against census and official Eurostat data, which can help in estimating the completely missing cells, bearing in mind the typical caveats of using survey data related to sampling and non-sampling errors.

Longer time series from Facebook. The Bayesian modelling framework aims to harmonise traditional data sources with social-media data. Going forward, longer time series of data in which Facebook data and official data overlap, will become available for most of the countries. This overlap will mean that the posterior distributions of estimated true migrant stocks will shrink with more information. Given the data protection provisions in the user agreement of the Facebook API, we recommend that the European Commission should start building its own dataset following the steps outlined in this report.

If the strengths and weaknesses of the data sources used for the proposed approach remain relatively stable, over time the model can be expected to perform better in "nowcasting" EU mobility.

ACKNOWLEDGEMENTS

This project could not have been conducted without the funding provided by the European Commission. At the European Commission's Directorate-General for Employment, Social Affairs and Equal Opportunities (DG-EMPL), we are particularly grateful to Simone Rosini (DG-EMPL A4) and Stefano Filauro (DG-EMPL A4) who acted consecutively as project officers for this study. We thank them for their active engagement, constructive feedback and facilitation of activities.

We are thankful to the members of the Steering Committee assembled by DG-EMPL, who offered their expertise throughout the study, providing feedback on the methodology and research design, and commenting on draft versions of this report. The Steering Committee comprised Filip Tanay (DG-EMPL A1), Lambert Kleinmann (DG-EMPL D1), Benoît Paul (DG EMPL-D1), Mantas Sekmokas (DG-EMPL E3), Michele Vespe (JRC E6) and Martin Ulbrich (DG-CNECT F4).

We are also grateful to Prof Frans Willekens (University of Groningen and Netherlands Interdisciplinary Demographic Institute, NIDI) and Charlene Rohr (RAND Europe) for the input, guidance and constructive criticism on earlier versions of this report, which they provided in their role as peer reviewers in the context of RAND Europe's Quality Assurance system.

Finally, we are grateful for the excellent research support and contributions from a number of colleagues including William Phillips, Tor Richardson-Golinski, Clément Fays (RAND Europe), Julian Glenesk, Harry McNeill, Carlijn Straathof (former RAND Europe) and Beatriz Sofia Gil (Max Planck Institute for Demographic Research).

1. INTRODUCTION

The free movement of people and workers is one of the founding pillars of the Treaty on the Functioning of the European Union (TFEU). While the Treaty rules on free movement of persons, as laid out in Article 45 – which initially only applied to economically active persons (i.e. employed persons and jobseekers) – the Maastricht Treaty ensured all EU citizens and their family members have the right to seek work, become self-employed or be a pensioner or student across the EU. Latest official available Eurostat data from 2018 state that 17.6 million EU citizens are currently living and working in another member state, and 4 per cent of the EU working-age population lives in another EU country (Eurostat 2019a).

In light of this, the European Commission's Directorate-General for Employment, Social Affairs and Inclusion aims to make it easy for citizens to work in another EU country and protect their social security rights when moving within Europe. Having access to up-to-date information about EU mobility is essential for policy making, such as labour-market, social or health policy, and for broader research purposes. However, the right for citizens and workers to move freely across the borders of EU countries has made it difficult to measure migration flows within the EU, as well as the number of EU migrants living in another EU country.

Official migration statistics are developed by Member States' national offices of statistics and collated and published by Eurostat, the EU's statistical agency. While these statistics are based on rigorous internationally harmonised principles, they come with a considerable time lag.

In 2005, the European Commission made a proposal for the development of harmonized European statistics on migration. It stated that, due to the development of Community policies and legislation on migration and asylum, the need for comprehensive and comparable European statistics on a range of migration-related issues had become a priority. The European Parliament adopted the proposal in 2007 (Regulation No. 862/2007 of the European Parliament and of the Council on Community statistics on migration and international protection). This regulation provides clear definitions for important terms, including usual residence, emigration (i.e. moving out of a country) and immigration (i.e. moving into a country). In addition, it describes which data Member States have to provide to Eurostat. However, the regulation leaves it to the Member States to decide how they collect the required data (e.g. using population registries, or conducting a sample survey), which means that the challenge of obtaining robust statistics on migration remains.

1.1. Developing a method to measure migration

In the light of these issues, it is important to continue innovating in ways to measure and analyse mobility and migration patterns in a faster, more precise and smarter manner. With the rise of information technologies and advanced computing, new data sources potentially offer opportunities to complement traditional sources for migration statistics. In particular, the availability of high quantities of individual geo-tagged data from social media has opened new opportunities. The metadata of these geo-tagged data contain information linked to the geographical location of the content. In this report, we will study these opportunities in detail, exploring the possibilities of using these new data sources – focusing in particular on social-media data, such as those from Facebook and Twitter – to develop a method to provide more timely and potentially more accurate statistics on EU mobility.

In light of on the one hand, rigorously calculated yet limited official statistics, and on the other hand, prospective new data with enormous potential and wide coverage, the pressing need to measure labour mobility and migration in a precise and timely manner necessitates the overarching objective of this study:

To investigate the potential of geo-referenced social-media data to facilitate "nowcasting", and develop methodologies that can provide nearly real-time estimates to serve as early warnings about the changes in EU mobility.

In order to tailor the study's scope to the policy area of the European Commission's Employment, Social Affairs and Inclusion DG, we focus on intra-EU migration, or in other words, EU mobility.

Ultimately, this study aims to be a first step in the development and application of a robust and sustainable method that may be applied, repeated and updated by the European Commission. It is intended that the new indicators will be disaggregated by country of origin and destination, and potentially by other socio-demographic characteristics, such as age group or gender. Similar to the official statistics, we aim to cover all EU countries and allow for a comparison with official statistics to assess the reliability of these new data and methods.

In order to assess the potential of these social-media data, we carried out the following activities:

- Collection of "big" social-media data, in particular from Facebook and Twitter, that can be used for measuring EU mobility;
- Specification and development of a method of measurement, combining these data with traditional data sources;
- Production of estimates for stocks of EU movers and EU mobility by EU Member State; and
- Comparison of estimates obtained from this research with official statistics and assessment of the reliability of the method.

1.2. Definitions of key concepts

When developing an approach to measure EU mobility, it is important to establish common definitions of key terms. Not only will it allow consistency in analysis and estimates over time, but it is important to acknowledge potential inconsistencies between estimates from social-media data sources and different official statistics. This section discusses the different definitions proposed for the purpose of this study, their differences and their application to social-media data.

Datasets carrying information from online social media are often described as "**big data**" sources. Laney (2001) argued that big data can be described by the 3 V's: "data sets characterized by huge amounts (**volume**) of frequently updated data (**velocity**) in various formats, such as numeric, textual, or images/videos (**variety**)" (Chen et al. 2012).

The concept of EU mobility refers to the movement of EU nationals within the EU, whether within a Member State or between Member States, as mobile workers (Eurofound 2019). Mobility includes the phenomena of posted workers and cross-border commuters. But for the purpose of this study, we are interested in those **EU citizens who have established their usual residence in another EU country rather than where they were born**. According to EU Regulation 862/2007 on migration statistics, establishing "usual residence" requires that a citizen has lived in a country for a continuous period of at least 12 months, or is expected to be living there for at least 12 months.

In the international literature, and indeed in the EU Regulation (862/2007), this minimum period of 12 months is used to distinguish migration from other types of mobility. Despite this international convention, the European Commission's Employment, Social Affairs and Inclusion DG does not use the term migration when referring to mobility within the EU. Regardless of establishing usual residence, moving to reside in a different Member State is referred to as **mobility, not migration**. And EU citizens moving to another Member State are not labelled migrants, but **EU movers**. Migration is only used to refer to movements between the EU and non-EU countries.

Consistent with the Commission's terminology, **we will use the term EU mobility** for intra-EU migration and **EU movers** for EU migrants when referring to the target group. In some cases, however we use the terms migration or migrants when they are explicitly used in the variable definitions of data sources or in the literature. In this section, we briefly reflect on the different definitions of mobility and migration.

1.2.1. Migration and migrant

As discussed by The Migration Observatory (2017), the definition of "migrant" differs across different data sources and between datasets and law. Numbers of migrant estimates vary significantly depending on the definition used, and so does the analysis of the drivers and impacts of migration. Not only does the definition differ between countries and institutions, but it also varies across European sources as described in Table 1 below.

Table 1: Definitions of immigration and migrant

	Term	Definition	Source
1	Immigration	"The action by which a person establishes his or her usual residence in the territory of a Member State for a period that is, or is expected to be, at least 12 months, having previously been usually resident in another Member State or a third country."	Eurostat[1]
2	Immigration	"In EU context, the action by which a person from a non-EU country establishes his or her usual residence in the territory of an EU country for a period that is, or is expected to be, at least 12 months."	European Commission[2]
3	Migrant	"A broader term of an immigrant and emigrant that refers to a person who leaves from one country or region to settle in another, often in search of a better life."	European Commission[3]
4	Migrant	"A person who established their usual residence in another country rather than where they were born, for a period that is – or is expected to be – at least 12 months."	EU Labour Force Survey[4]

There are a number of determinants that play a role in these definitions, for example:

- **EU and Third Countries** – Definitions 1, 3 and 4 do not differentiate between EU nationals and third countries' nationals, while according to definition 2, migration applies to movements across the EU borders only.
- **Country of Birth** – Definition 4 implies that a migrant is a foreign-born person in the country of his or her usual residence while definitions 1, 2 and 3 do not specify country of birth. In definitions 1 and 3, a native-born person returning after having lived in another country would be considered as an immigrant.
- **Length of stay** – Definitions 1, 2 and 4 state that a person is considered a migrant if they have established, or are expected to establish, their usual residence in the country for a period of at least 12 months.

[1] Eurostat 2019c
[2] European Commission 2019a
[3] European Commission 2019a
[4] Eurostat 2014

1.2.2. Migrant stocks, migration flows and corridors

Stocks and flows are concepts stemming from the field of system dynamics. A stock is a measure of a quantity at one specific time that may have accumulated in the past. A flow variable is roughly equivalent to a rate or a speed measured over an interval of time. Therefore, flows are always expressed as a quantity per unit of time, for instance, a year.

In the context of migration, international migrant stocks reflect "the total number of international migrants present in a given country at a particular point in time" (UN 2017). Data on migrant stocks are typically measured as the share of a country's population that is born abroad, or on those holding a foreign citizenship. Migrant flows also reflect a total number of migrants, but they measure the number of migrants entering or leaving during a specified time period, typically a calendar year (UN 2017).

When reporting on stocks of EU movers and mobility flows in this report, we also refer to origin-destination corridors. With a "mobility or migration corridor" we imply the combination of an origin and destination member state within the EU. An origin member state in a corridor is the country of birth for traditional sources and home country for Facebook. Consequently, there are 756 migration corridors in the EU (= 28 member states * (28 – 1) member states, excluding corridors within a member state).

1.2.3. Labour mobility and mobile workers

Two further concepts – "labour migration" and "labour mobility" – are concepts that represent a sub-set of mobility. There are a number of different but related terms in this context, such as mobile workers, migrant workers, posted worker and cross-border workers. Although definitions are more homogenous than in the case of "migration", subtle differences exist. For the purposes of this study, determining those individuals that qualify as "mobile workers" is an essential step towards the development of migrant statistics. Table 2 below presents the central definitions around the theme of labour mobility.

Table 2: Definitions around the theme of Labour Mobility

#	Term	Definition	Source
1	Intra-EU mobility	"The movement of EU nationals within the EU, whether within a Member State or between Member States, as mobile workers. In cases where this move is between Member States and at least semi-permanent, this constitutes internal migration. Shorter term movement includes the phenomena of posted workers and cross-border commuters."	Eurofound[5]
2	EU-28/EFTA movers	"EU-28 or EFTA citizens who reside in an EU-28 or EFTA country other than their country of citizenship (definition created for the purposes of the study)."	2017 Annual Report on intra-EU Labour Mobility
3	Mobile worker	"Mobile workers are defined as economically active EU-28 citizens who reside in a Member State or EFTA country other than their country of citizenship."	2017 Annual Report on intra-EU Labour Mobility
4	Migrant worker	"A person who is to be engaged, is engaged or has been engaged in a	European Commission[6]

[5] Eurofound 2019
[6] European Commission 2019b

#	Term	Definition	Source
		remunerated activity in a state of which they are not nationals."	
5	Posted Worker	"A worker who, for a limited period, carries out his/her work in the territory of a Member State other than the State in which he/she normally works.[7] The posted worker has a regular employment relationship in the usual country of work and maintains this employment relationship during the period of posting.[8]"	2017 Annual Report on intra-EU Labour Mobility
6	Cross-border worker	"For the purposes of the report, cross-border workers are defined as EU citizens who live in one EU or EFTA country and work in another, regardless of their precise citizenship (provided they are EU-28/EFTA citizens). Cross-border workers therefore move across borders regularly.[9] They can be EU-28/EFTA movers – meaning they live in a different Member State than their country of citizenship – and cross-border workers at the same time (for example, where a British person lives in Belgium and works in Luxembourg).[10] Cross-border workers are employed or self-employed in a country other than their country of residence."	2017 Annual Report on intra-EU Labour Mobility

The key factors that help distinguishing between these different but related concepts include:

- **Duration of stay** – Definition 1 differentiates between internal migration and shorter term movements using the notion of duration of stay, but does not state a time threshold. Definition 3 also mentions that the move should be made "on a long-term or permanent basis".
- **Active population** – Definitions 1, 3, 4 and 5 seem to agree that the mobile worker status applies to the whole active population. Definitions of employment and unemployment are complex and the Labour Force Survey uses a specific set of questions and a derivation chart to determine whether the respondent can be considered as part of the active population.[11]

Although we do not differentiate between economically active and inactive movers in this study, we do focus on the working-age population (15 to 64 years old).

1.3. Scope of the study

As explained in previous sections, in this study we aim to explore the potential of geo-referenced social-media data to facilitate "nowcasting" stocks of EU movers and EU mobility flows. EU mobility is the terminology used by the European Commission's Employment, Social Affairs and Inclusion DG to describe intra-EU migration and other forms of mobility within the EU. We are interested in those working-age EU citizens who

[7] Article 2(1), Directive 96/71/EC of the European Parliament and of the Council of 16 December 1996 concerning the posting of workers in the framework of provision of services.
[8] Article 1(3)(a-c), Directive 96/71/EC of the European Parliament and of the Council of 16 December 1996 concerning the posting of workers in the framework of provision of services.
[9] The frequency of commuting cannot be identified in the EU-LFS, which is the data source for the estimation of numbers of cross-border workers.
[10] For a more detailed definition, see European Commission, 2011, Mobility in Europe, p. 86.
[11] Eurostat 2019d

have established their usual residence in another EU country than where they were born. In accordance with EU regulation, "usual residence" requires a minimum (intended) duration of stay of 12 months. This implies that, while we refer to EU "mobility", the scope of this study is consistent with what is referred to as migration between EU member states by statistical agencies and in the academic literature.

In summary:

- **Working-age population**: population between 15 and 64 years of age.
- **EU movers**: EU citizens who have established their usual residence (at least 12 months) in another EU country than where they were born.
- **EU mobility**: the action by which an EU citizen establishes usual residence in an EU member state for a period that is, or is expected to be, at least 12 months, having previously been usually resident in another member state.
- **Stocks of EU movers**: the total number of EU movers present in a given country at a particular point in time.
- **Flows of EU mobility**: the total number of EU movers from an origin member state and to a destination member state within the EU during a specified time period.

With this in mind, we take a two-pronged approach to developing models to estimate EU stocks of EU movers and flows. The emphasis in this report is on estimating stocks of EU movers. This approach is motivated by a number of factors. First, migrant stocks are more easily measured by national statistical agencies, and hence migrant stocks by country of birth are more widely available in the EU (revealing the bilateral connections between countries). Second, official migrant stock data is far more uniformly defined in comparison to migration flow estimates. Third, non-traditional data sources on migrant stocks are more easily obtained than flow estimates, and possess fewer biases. Social-media platforms such as Facebook make aggregate data of specific target groups available through their advertising platform.

1.4. The structure of this report

This final report describes the methodology of the study and discusses its findings. Chapter 2 summarises the findings from existing literature – focussing on previous studies that have used geo-tagged data sources to measure migration – with a view to developing a new method.

Subsequent chapters describe in detail the data sources, methodology and results of the approach applied to estimate stocks of EU movers (Chapters 3, 4 and 5) and EU mobility flows (Chapters 6, 7 and 8). Specifically, Chapter 3 describes the data sources used for the stocks methodology, as well as their main benefits and limitations. Chapter 4 describes the applied methodology in detail. Chapter 5 presents the results obtained using this approach and discusses the total estimates of stocks of EU movers, their limitations and caveats. In a similar vein, Chapters 6, 7 and 8 describe the data sources, methodology and results of the approach applied to estimate EU mobility flows.

Finally, Chapter 9 concludes the report with a series of steps the European Commission can undertake to apply the proposed approach in the policy process, and gives an overview of areas for further research.

2. A REVIEW OF THE AVAILABLE LITERATURE

This chapter presents a summary of the findings from existing literature on the subject of measuring migration and labour mobility using geo-tagged data. The review starts with a discussion of potential sources of geo-tagged data that can be used to measure migration and includes some examples of their applications. Subsequently we explain how a Bayesian framework can be used to combine various data sources, and review a number of previous applications using this approach in the context of migration. Finally, we summarise some of the main lessons that can be drawn from the literature for the purpose of this study.[12]

2.1. Sources and applications of geo-referenced data in measurement of labour mobility and migration[13]

The study of human migration and mobility is not confined to a single discipline. Several lines of literature, which often cross disciplinary borders, have emerged. Historically, censuses have been used to gather estimates on migration, defined as changes in residence over a defined period of time, such as one or five years. Elsewhere, migration patterns might be monitored through population registers, administrative data, travel history surveys or national representative surveys (see e.g. Bilsborrow et al. 1997).

The increasing availability of geo-tagged digital records has led to a growing trend of interaction and exchange between scholars with different backgrounds. Estimating flows of migrants is important to understand migration processes, to assess the success of policy interventions and to forecast future trends. Abel (2013) has developed statistical techniques to estimate flows of migrants from census data, in order to generate a historical time series of migration flows. Among others, Abel's work is intended to inform the population projections of the International Institute for Applied Systems Analysis (IIASA). The Population Division of the United Nations (UN) has recently moved towards offering probabilistic population projections (Raftery et al. 2012). Forecasting migration remains one of the most difficult tasks for the UN. Currently, there is a continuing collaboration between the UN and the University of Washington to develop statistical models to forecast net migration rates for all countries (Azose & Raftery 2013).

Statistical approaches to the study of human migration and mobility have been integrated within the framework of models used in physics. For instance, the most widely known model for migration flows is the "gravity model"[14] (Zipf 1946; Cohen et al. 2008), which has been applied to, for instance, estimating the missing information on migration flows between countries in the European Union (De Beer et al. 2010; Abel 2010; Raymer et al. 2011; Raymer et al. 2013). More recently, the "radiation" model, an approach that addresses some of the limitations of gravity-type models, has been suggested (Simini et al. 2012).

The increasing availability of geo-tagged "big data" from online sources has opened new opportunities to identify migrants and to follow them, in an anonymous way, over time (Hui et al. 2012; Cesare et al. 2018). Various types of data sources have been used to estimate human mobility. Cell-phone data have been used mainly to evaluate mobility patterns, particularly in terms of regular patterns, within a country (Bayir et al. 2009; Gonzalez et al. 2008; Candia et al. 2008; Blumenstock 2012), but also between countries (in terms of international calls) (Blumenstock 2012). Travel itineraries for tourists have been inferred using geo-tagged pictures in Flickr (Choudhury et al. 2010) and recommendations posted on Couchsurfing (Pultar & Raubal 2009). Localized mobility,

[12] Since this chapter summarises the findings from previous research, we refer to concepts of migration and migrants, which are more common in the academic literature than EU mobility or EU movers.
[13] This section is based on the review of existing work to infer migration and international mobility using alternative data sources in Zagheni et al. (2014).
[14] Gravity models are based on Newton's universal law of gravitation (which measures the attraction between two objects based on their mass and distance). The gravity model of migration is used to predict the degree of migration interaction between two places (Rodrigue et al. 2009).

often within a city, has been measured using data from Twitter (Ferrari et al. 2011), Google Latitude (Ferrari & Mamei 2011), Foursquare (Noulas et al. 2011) and public transport fare collection sensors (Lathia et al. 2012; Smith et al. 2013). IP addresses have been used to evaluate internal mobility (Pitsillidis et al. 2010). Sequences of geo-tagged tweets have been used to infer movements of individuals over time (Hawelka et al. 2014; Lenormand et al. 2014; Zagheni et al. 2014) and investigate temporal patterns of migration (Fiorio et al. 2017). A study on Facebook users' hometown and current locations has been used, amongst others, to indicate coordinated migration[15] (Hofleitner et al. 2013), measure stocks of migrants (Zagheni et al. 2017), visualise geo-demographic data (Araujo et al. 2018) and improve predictions of crime (Fatehkia et al. 2019). Recent trends in international flows have been estimated by tracking the locations of users who repeatedly login into Yahoo! Services, inferred from their IP addresses (State et al. 2013; Zagheni & Weber 2012).

2.2. Bayesian modelling to estimate migration

The availability of geo-tagged data from online sources has opened up new opportunities to track recent trends in population movements. As discussed in Hughes et al. (2016; see also Bijak & Bryant 2016; Raymer et al. 2013; Willekens 2016; Wiśniowski 2017), Bayesian methods may be particularly useful to combine different migration data in a consistent way. Within the Bayesian inferential framework, the interest is in quantities of interest which are "unknowns" and are usually the parameters of the statistical model or the forecasts of the variable of interest. The aim is to produce a probability distribution that represents uncertainty about those "unknowns" by combining the observed data, represented by a statistical model, with the prior beliefs of the researcher on the quantities of interest which are represented by prior probability distributions. By using Bayes Theorem, a so-called posterior distribution for the "unknowns", conditional on the observed data and the assumed statistical model, is produced (Bijak & Bryant 2016).

None of the relevant datasets can measure migration or mobility with absolute precision – estimates using geo-tagged data may not uniformly cover the entire population, censuses are done with insufficient frequency, and surveys may underrepresent migrant stocks for a host of reasons. While a reasonable estimate of true migrant stocks may potentially be obtained using a simple average measure, Bayesian methods take this approach a step further and allow explicit inclusion of prior beliefs in the form of, for example expert opinion or official statistics, and allow statistical inferences based on a limited number of observations.

Clearly, assumptions – or caveats – need to be made and explicitly stated when using numbers of data sources. Social-media users are hardly representative of the general population as they tend to be younger and are more likely to be from urbanised areas (e.g. Mislove et al. 2011; Perrin 2015; Sloan et al. 2015 – for US and Yildiz et al. 2017 – for UK). While appropriate weighting may help to diminish or entirely correct the bias inherent in the data, one must be cautious when using these data. The main reason is that it may not be possible to produce the exact weights for the entire population, but only relative weights for comparisons over time or amongst geographical units such as countries. For instance, Zagheni et al. (2014) clearly state that migration rates estimated through analysis of Twitter users cannot be considered representative of a broader population without making additional adjustments, specifically looking at the relative changes in outcomes over time. Further they note that those living in the US with "Lived In Mexico" status (formerly expat status) on Facebook (e.g. Mexicans in US) are not necessarily representative of all Mexicans living in the US, and investigate the bias in migrant stock estimates derived from Facebook advertising platform marketing data.

As mentioned before, the definitions of migration may also differ from one data source to another. This leads to a second caveat: while the outputs should be generally comparable

[15] Defined by the authors as a large proportion of a population, which migrates as a group from city A to city B.

in relative terms, the raw migrant stock estimates will likely not be comparable in absolute terms due to definitional differences across datasets.

Modelling frameworks for harmonising migration flows have been developed in two projects: Migration Modelling for Statistical Analyses (MIMOSA, see De Beer et al. 2010 and Raymer et al. 2011) and Integrated Modelling of European Migration (IMEM, see Raymer et al. 2013; Wisniowski et al. 2013 and 2016). The former approach used optimisation procedures to harmonise migration flows collected from sending and receiving countries, benchmarking them to Sweden, which was deemed to be the best country in terms of accurately collecting this information, and estimating completely missing flows (Abel 2010). The IMEM approach applied a statistical model that reconciled the differences between the data observed in sending and receiving countries by correcting for their main limitations: coverage, undercount, accuracy and a difference in the duration-of-stay definition, applied across countries. It also relied on informative prior distributions, especially on undercount and accuracy, elicited from experts on migration data in Europe (Wisniowski et al. 2013). The hierarchical model contained an extended theory-based gravity model that utilised the observed data for some of the countries (so-called borrowing strength[16] from the observed data) and estimated missing flows.

The statistical model, described in Chapter 4, is based on the idea developed in Raymer et al. (2013). The model integrates three main sources of data – (i) officially reported data, (ii) survey data and (iii) social-media data – and accounts for differences in the definitions applied therein as well as their other characteristics – e.g. coverage of subpopulations, undercount and accuracy – to produce a harmonized, timely and up-to-date set of migrant stocks. Further, by using the officially reported Eurostat data on stocks,[17] disaggregated by age groups and gender, and calibrating the resulting estimates to them, a much more timely system of "early warnings" regarding the changes in migrant stocks could be implemented (see for example, Disney et al. 2015).

2.3. Lessons from the literature for the purpose of this study

The main conclusion stemming from the literature in the area of estimating migration is that no single source of data is of sufficient quality to provide accurate, timely and unbiased information on migration. This applies to both traditional and officially used data sources and new forms of data. Both have their strengths and both suffer from different shortcomings, such as biases resulting from under- or over-counting migrants, coverage issues when the data are collected only on a part of the entire population and low accuracy in the case of using survey data (Disney et al. 2015; Willekens et al. 2016). Official data also require extensive processing, which delays their official publication. The main advantage of the new forms of data, such as social-media data (described in Section 2 of this report), is their timeliness, though they still suffer from the lack of representativeness (e.g. if the use of particular social media is selective by age or educational attainment). Nevertheless, such data have the potential to provide estimates that can serve as early warnings about changes in population dynamics and thus in EU mobility.

In the next section, we propose a method to overcome those limitations by combining the various datasets within a complex statistical model. We recommend using Bayesian inference methods (Bijak & Bryant 2016; Willekens 2016) to estimate model parameters and forecasts, as this approach permits explicit inclusion of qualitative information on each data source in terms of prior probability distributions (e.g. Wisniowski et al. 2013). This is especially relevant in the context where: (i) none of the sources are sufficiently reliable to alone provide an exhaustive picture of labour mobility within the EU in recent

[16] The concept of "borrowing of strength" was first used by John W. Tukey (Brillinger 2002). Borrowing of strength refers to assuming a distribution over parameters of interest so that information on one parameter contributes to determine information on others.
[17] Eurostat 2019e

years and (ii) we have knowledge on the mechanisms of data collection and their comparative qualities. The results yielded by the statistical model can be used to create a synthetic data base for migration and forecasts accompanied by measures of uncertainty, which subsequently can inform planners and policymakers (Raymer et al. 2013; Willekens 2016).

A Bayesian inferential framework offers a powerful mechanism to combine data sources and provide measures of uncertainty. Models previously developed have primarily combined traditional migration data sources (see for example, Bijak and Wiśniowski. 2010; Raymer et al. 2013; Wiśniowski et al. 2013; Wiśniowski 2017; Wiśniowski et al. 2016). In this study, we adapt the basic methodologies of these former models to combine migration data from both traditional and new data sources derived from social media.

As explained in the Introduction, we take a two-pronged approach to developing models to estimate stocks of EU movers and EU mobility flows. First, we focus on estimating stocks of EU movers and subsequently, the stocks module is complemented by a proposed methodology estimating EU mobility flows (discussed in Chapters 6, 7 and 8). This approach is motivated by three main reasons:

1. Migrant stocks are more easily measured by national statistical agencies, and hence migrant stocks by country of birth are more widely available in the EU (revealing the bilateral connections between countries).
2. Migrant stock data is far more uniformly defined in comparison to migration flow estimates.
3. Non-traditional data sources on migrant stocks are more easily obtained than flow estimates and with fewer biases (Nowok, Kupiszewska & Poulain 2006; Abel & Sander 2014). For example, proxies for estimates of migrant stocks can be obtained directly from the Facebook advertising platform.

Therefore, we start by focusing on estimating stocks of EU movers in the following three chapters (Chapters 3, 4 and 5), which describe in detail the data sources, the methodology and the results of the approach applied to estimate stocks. Subsequently, Chapters 6, 7 and 8 use a similar structure to describe the methodology for estimating EU mobility flows.

3. DATA SOURCES TO ESTIMATE STOCKS OF EU MOVERS

In this chapter we outline the data sources used for the estimation of stocks of EU movers. In the following sections, we provide a brief description of each dataset, their characteristics and limitations. We explore the potential of geo-tagged social-media data from Facebook to facilitate "nowcasting" of EU mobility. Although other social-media sources could have been used in combination to the Facebook data, we have focused on Facebook as it has the highest coverage and provides information on people who moved a long time ago. Section 3.1 describes the available data from Facebook, their definitions, the data-collection process and important limitations. Subsequent sections outline the traditional data sources that complement these geo-tagged social-media data: migration statistics from Eurostat (Section 3.2), data from the EU Labour Force Survey (EU-LFS) (Section 3.3) and Population and Housing Census data (Section 3.4). Section 3.5 summarises the strengths and weaknesses of each data source. And finally, Section 3.6 discusses the similarities and differences in the definitions of mobility and migration in these data sources and their consequences for estimating stocks of EU movers.

3.1. Facebook data

Facebook, the online social-networking platform, was initially launched as TheFacebook on 4 February 2004 with membership limited to students from a number of universities in the United States. Since September 2006, the network has expanded eligibility to anyone above 13 years old who possesses a valid email address. To access the platform, users must create a profile where they are asked to provide their gender, date of birth, education, employment status or history and a large number of other characteristics. Users can then interact by, for example, expanding their network of friends, posting information, writing on friends' "walls" or updating their status. As of the fourth quarter of 2018, Facebook counts 2.32 billion monthly active users (MAU), defined as users who have logged in to Facebook during the last 30 days. Among them, 1.52 billion people log on average daily to the platform and are considered daily active users (DAU). While Facebook network users may not be a representative sample of the general population (for more information, see Section 3.1.5), Facebook has about 262 million[18] MAUs in the EU. This is 59 per cent of the population aged 13 years and older, according to the last estimate of Eurostat from 2017 (about 442 million). Focusing on the population aged 15 to 64 years old, this number increases to 73 per cent. Facebook provides access to a large amount of aggregate and anonymous data on the characteristics of its users through its marketing API.

This section describes the Facebook Marketing interface, the data available and the general mechanism to download the data.

3.1.1. Facebook Marketing API and Ads Manager interface

The Facebook Marketing API and Ads Manager interface are online tools designed to guide advertisers on the type of Facebook users to whom their ad should be shown. These tools allow registered advertisers to assess the approximate number of users that fit selected criteria. Once the user has specified the characteristics of the targeted audience, the interface will display the number of monthly and daily active users who fit this particular profile. These services are publicly available online and have been used in previous studies either as a way to reach a specific population to whom a survey should be distributed (Pötzschke and Braun 2016) or, as is the case here, to estimate stocks of EU movers (Zagheni et al. 2017).

The Facebook Ads Manager interface is accessible to any registered Facebook user. From a personal account, it is possible to reach the Ads Manager platform by creating a page and then clicking on "create an ad" in the parameters. After entering details about their

[18] According to Facebook API, as of 4 February 2019.

marketing campaign, the users are redirected towards the Audience section of the ad set. Users can then choose the profile of the targeted audience by changing the parameters. There are four types of parameters that can be used to define the targeted profile:

1. Demographics;
2. Location;
3. Interests; and
4. Behaviours.

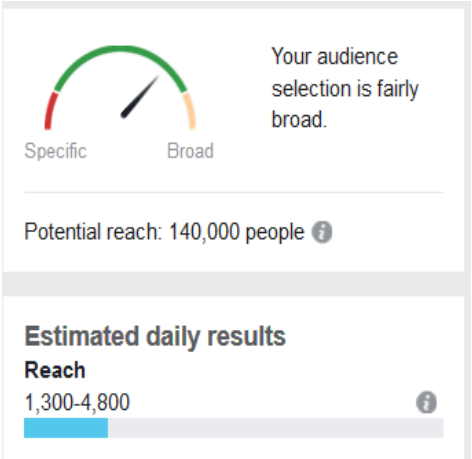
Locations of interest can be defined as a group of countries, of cities, of post codes or by a latitude-longitude point and a radius. The interface proposes to reach Facebook Network users:

- "who live in this location";
- "who are travelling in this location";
- "who recently were in this location"; or
- "everyone in this location".

The demographic criteria allow the user to filter by gender, age, education level, relationship status or workplace. Specific interests are available to choose from – such as "Entertainment" or "Sports and outdoors" – and advertisers can also target people with specific behaviours such as whether they use mobile devices and whether they "Lived in [country]", where [country] can be chosen from a list of 89 countries (as of February 2018; Spyratos et al. 2018).[19] In addition to the above, the users can also choose the platform on which they wish to advertise (Facebook, Instagram, Audience Network and Messenger). For the purpose of this study we chose Facebook only, to avoid the potential double-counting issue described in Spyratos et al. (2018).

Once the target criteria have been defined, the interface returns the estimated number of users who will be shown the ad. As displayed in the example in Figure 1 below, two estimates are shown. The first one is called the "potential reach" and is an assessment of the number of Facebook users corresponding to the given profile.[20] The second number refers to the "Estimated daily results – reach" and is an estimate of how many people the ad will reach per day if the full budget is spent.[21]

Figure 1: Facebook Audience estimates screenshot



Source: Facebook Ads Manager interface

[19] Until 2018, this information category was labelled "Expats-[country]"
[20] Facebook 2019a
[21] Facebook 2019b

3.1.2. Definitions and methodology to estimate Facebook monthly and daily active users

To compute the "potential reach" and the "estimated daily results", the interface appears to be calling two estimates from a larger Facebook database based on the profile selected in the interface called monthly and daily active users estimates. The monthly active users (MAU) and daily active users (DAU) are the estimated number of users who satisfy the selected profile and have been active on Facebook during the past month or day (respectively).[22]

Since 26 February 2018, the MAU estimate is rounded to two significant digits for numbers above 1,000[23] (Spyratos et al. 2018). One cannot assess the MAU estimate for categories with fewer than 1,000 users. On another hand, the DAU does not seem to be rounded and can bring useful precision to the dataset.

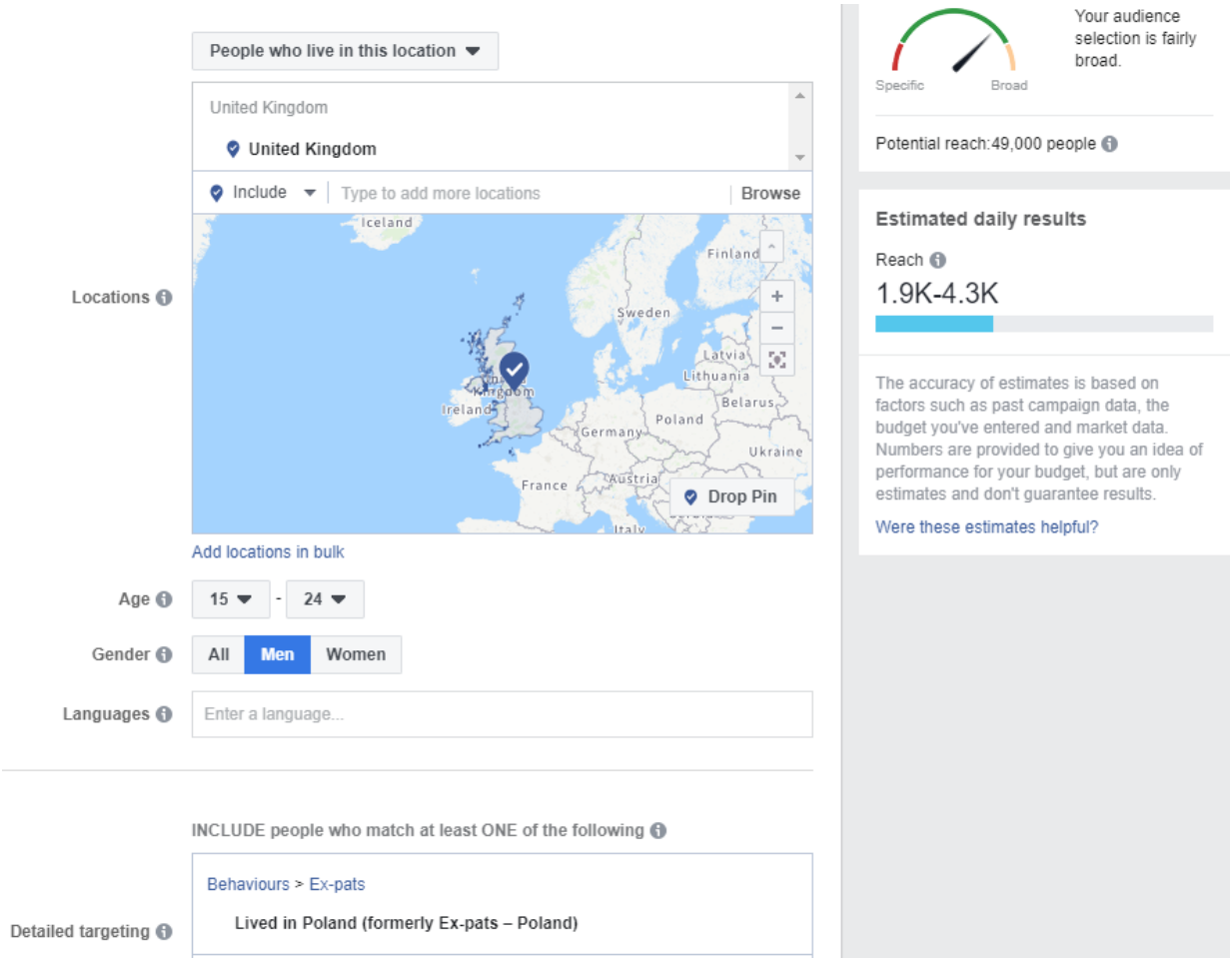
As explained above, there is a pre-defined list of parameters which can be used to define the population of interest. Some of the parameters are self-declared by Facebook users while others are determined by a set of algorithms developed by the service providers. To build our Facebook dataset, we have targeted each combination of the characteristics detailed in Table 3. Figure 2 shows an example of a query for Facebook male monthly active users between 15 and 24 years old who are currently living in the UK, and who used to live in Poland. The result for this example is 49,000.

Table 3: Facebook dataset variable descriptions

Characteristic	Value
Gender	All; Male; Female
Age	15 to 24; 25 to 54; 55 to 64; 15 to 64 years old
Location	Each of the 28 Member States
Lived in (formerly Expat status)	Lived in each of 26 Member States **(excluding Bulgaria and Croatia)**

[22] Facebook for Developers 2019
[23] Numbers from 1,000 to 10,000 would be rounded to the hundred (e.g. 1,200, 1,300); numbers from 10,000 to 100,000 would be rounded to the thousand (e.g. 12,000, 13,000), etc.

Figure 2: Example of a Facebook Audience in the Facebook Ads Manager interface



Source: Screenshot of Facebook Ads Manager Interface

Gender, date of birth and education level are self-reported. Age is computed from the self-reported date of birth. We have aligned the age ranges in the analysis to match the ones used in the data source with less granularity (Labour Force Survey tables, see Section 3.3). The Facebook dataset compiled for the purpose of this study is targeted at "people who live in this location". This is determined by the stated city on Facebook users' profiles and is validated based on device and connection information according to the help section of the Ads Manager (see Section 3.1.5 for more information).

As of 21 November 2018, Facebook describes "Lived In country X (formerly Expats – Country X)" as "People who used to live in [country] who now live abroad", although it does not provide more details on the determination process of "Lived in" status. According to Herdagdelen et al. (2016), produced by researchers working internally at Facebook, the "Lived in" status is determined using the self-reported "current city" mentioned above and "hometown" populated in the "places you have lived" of users' Facebook Network profiles. It is then validated using the structure of the friendships network of the users in each country. Spyratos et al. (2018) recently conducted an online survey to understand how Facebook assigns "Lived In" status and concluded that both "country of home town" and "country previous residence" might be determinants, as well as other attributes such as geo-tagged information. There is therefore uncertainty around the method used by Facebook to determine the "Lived In" status. Moreover, it does not seem that Facebook includes a time component to its definition of "Lived In" unlike the official definition of "migrant" cited in Table 1 or the subsequent definition of EU movers. Facebook data will therefore give an overview of mobility without an explicit threshold value for the length of stay, as does the official definition of intra-EU mobility of Eurofound cited in Table 2. The dataset may therefore include stocks of both short- and long-term mobility. However, to have a "Lived In" status, the user must have changed its

status on the platform, which we assume, would not be done for short stays (e.g. holidays or commuting). Similarly to the survey conducted by Spyratos et al. (2018), a survey of Facebook members could help understand how long elapses, on average, before they change their status.

3.1.3. Data collection process

The MAU and DAU estimates can be accessed both manually and through a computer-automated process. Specifically, it is possible to create a list of characteristics as described in Table 3 and to use a Python code to send requests to the Facebook Marketing API. The program will then populate a spreadsheet with the estimated values delivered by the website. In this study, we have used the Python code developed by pySocialWatcher (Araujo et al, 2017), available on GitHub, a repository for source code.[24] We have downloaded six datasets starting in April 2018.We have also conducted similar requests without filtering for "Lived in" status to have an estimate of Facebook penetration for each category (by age and gender) in each member state. It provides us the estimated number of monthly and daily active users in each member state for each category and allows for a comparison with the size of the whole population in each member state.

3.1.4. Missing values

The data from Facebook contain a number of missing values. There are several ways to treat these missing values:

- **Historical data from Facebook.** It is only possible to download current estimates from Facebook API. Therefore, we were not able to download Facebook estimates for the period 2013 to 2016. Although we have been in contact with some representatives at Facebook to discuss a potential data sharing agreement, after July 2018, the Facebook representatives no longer responded to our requests. Therefore, in addition to the data downloaded since April 2018, we used Facebook data for the period 2016–2017, shared with us by project members. The 2016 data include each of the 28 member states as country of residence. However it only covers 11 member states as origin countries, namely Austria, France, Germany, Greece, Hungary, Ireland, Italy, Poland, Portugal, Romania and Spain. The 2017 data include 9 member states as country of residence (Belgium, Denmark, Finland, France, Germany, Italy, the Netherlands, Spain and the United Kingdom) and 6 as countries of origin (France, Germany, Greece, the Netherlands, Poland and Portugal). The data were not specifically targeted at the economically active population, and no age constraint was applied. In addition to allowing to cover a larger period of time for a few migration corridors, the data were not rounded to the hundred as is the case for the 2018 data, but rather to the tenth digit, and estimates go as low as 20.

- **Bulgaria and Croatia as countries of origin are missing.** As of August 2018, Facebook does not allow "Lived in Croatia" and "Lived in Bulgaria" as "Lived in" status. In the future, one way to estimate the number of users with the status "Lived in Bulgaria" and "Lived in Croatia" could be to look at Bulgarian-speakers and Croatian-speakers in each member state as users speaking one of these two languages are highly likely to be from these countries.

- **Categories with less than 1,000 monthly active users.** As mentioned above, Facebook does not provide estimates for categories with less than 1,000 MAUs (e.g. users aged between 55 and 64 years old who used to live in Luxembourg, but who now live in Poland). Targeting larger groups (e.g. by ignoring the filter on age: users who lived in Luxembourg who now live in Poland) can help mitigate

[24] GitHub, Inc. 2019

against this problem, but it is not always sufficient. We have therefore applied the methodology outlined in Box 2 below to try to estimate the missing values.

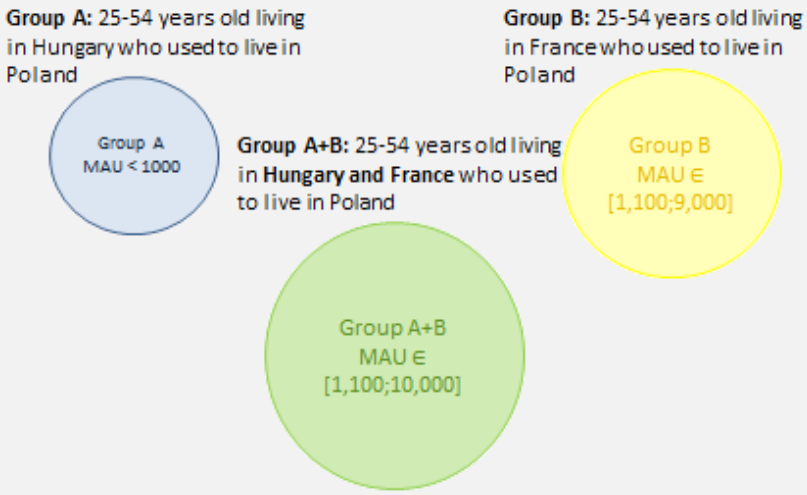
Box 2: Method to deal with missing values

Dealing with missing values in the Facebook data (< 1,000 monthly active users)
Since February 2018, Facebook has increased the minimum response of its marketing API to queries of its monthly active users from 20 to 1,000. As a result, estimates for categories with less than 1,000 monthly active users are missing from our datasets. To overcome this problem, we have used the following methodology:

1. Let's call group A the group for which the estimate is missing. Find a group B for which the estimate is not missing and such that group B does not overlap with group A.
2. Get the number of monthly active users for group B.
3. Get the number of monthly active users for the joint group A+B.
4. By subtracting the estimate for group B to the joint estimate for group A+B, we obtain an estimate for group A.

As we are looking for a number below 1,000, it is important that group A+B is between 1,100 and 10,000 (and so, that group B is between 1,100 and 9,000) as any number above 10,000 will be rounded to the thousand as explained above.

The country of residence ("people who live here") is unique to each user and it is possible to make joint queries on this field (e.g. to query the number of monthly active users aged 15 to 64 living in France and in Germany). Given the high number of groups with missing values to cover, it was not feasible to select a tailored "group B" for each of them, and it was therefore necessary to find a "group B" that could help to eradicate the most missing values. In other words, it was necessary to identify the member states with the less missing values in their immigration corridors, but also with number of immigrants below 9,000. Selecting several such countries would guarantee to cover as many missing values as possible, but as each download takes some time, we had to limit our "group B" countries to three: United Kingdom, France and Italy. This helped to reduce the number of missing values by 75 per cent. When several combinations of countries allowed us to get an estimate, we took the average of all estimates.
Example: measuring the number aged 25–54 years old living in Hungary who used to live in Poland



We assume that the number aged 25–54 years old living in Hungary who used to live in Poland is missing. We then get the number aged 25–54 years old living in Hungary and France who used to live in Poland which is not missing and the number aged 25–54 years old living in France who used to live in Poland which is not missing either. We then subtract this number for France to the number for France and Hungary and we obtain a measure of the number aged 25–54 years old who live in Hungary who used to live in Poland. We repeat the same process replacing France first by the UK and then by Italy. Our estimate for Hungary will then be the average of the measure we successfully computed using this method.

3.1.5. Caveats

As discussed in the previous section, Facebook data as a source for migration estimates, or in this case stocks of EU movers, has its limitations and caveats. Bayesian modelling requires specifying the prior probability distributions (so called "priors") that explicitly take these into account in a statistical model. Priors contain information about the characteristics and quality of the data source indicating the errors which are known (e.g. exclusion of specific subpopulations from the survey which can be found in the survey's documentation) or are believed to exist (e.g. a relative difference in social-media penetration by age based on aggregate data and researcher's judgement). Depending on the source of information used to construct a given prior distribution, the priors may be subjective representations of limitations in the data sources under study. This section summarises the quality issues related to the data collected from Facebook Marketing API, namely the issues related to the generalisability of the Facebook data to the entire population, the lack of transparency with regards to the definitions and methodology used to compute the data and other potential issues related to Facebook memberships.

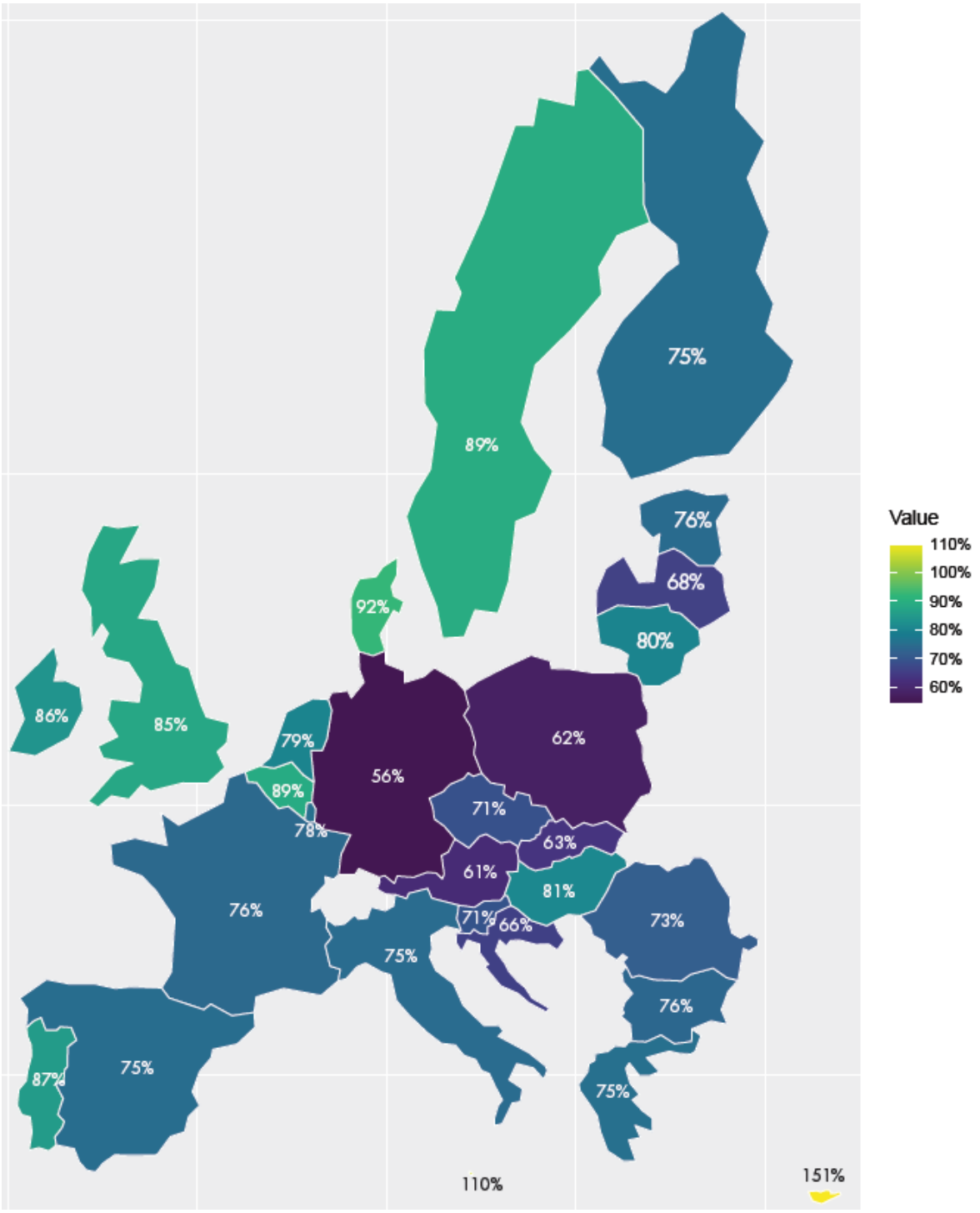
Facebook penetration rate
One of the main drawbacks of using social media to estimate statistics for the general population is that the penetration of social media varies between the different categories of the population, making social-media users an imperfect sample of the population. In the case of low penetration by specific subpopulations, this may lead to truncation of the data at the lowest reported figures (e.g. 1,000) which effectively leads to a missing data problem.

Comparing Facebook penetration between countries and by age and gender supports understanding some of the bias of the data. Spyratos et al. (2018) analysed the bias in the data and suggested a methodology (different from the one we propose in Chapter 4) to adjust Facebook estimates using the penetration rate of Facebook in different population groups. We compare the penetration rates across countries in Figure 3. One issue that we observe is that Malta and Cyprus have penetration rates[25] higher than 100 per cent for those aged 15 to 64 years old (respectively 110 per cent and 151 per cent). The large difference in the case of Cyprus can mainly be explained by the fact that the Facebook estimate includes users living on the whole island of Cyprus, while the Eurostat reference area is the government-controlled area of the Republic of Cyprus.[26] Arguably, the remaining difference could be explained by three other reasons:

- the estimates of Facebook penetration rates were downloaded in November 2018 while the Eurostat data dates back to 2017, and the population size is likely to have increased since.
- It is likely that children under 13 years old misreport their age to be able to obtain a Facebook membership (see section paragraph on "Definitions and Methodology" below).[27]
- As revealed by Facebook in 2017, fake Facebook profiles might artificially inflate the size of the audience reported by its Marketing API (for more information, see below).

[25] Estimate of MAUs in a country divided by the total population of this country.
[26] Eurostat 2016b, section 15.2
[27] An online survey of 1,001 10–12 years olds conducted by ComRes on behalf of BBC Newsround early 2017 finds that 49 per cent of respondents have used Facebook.

Figure 3: Facebook penetration rates for people aged 15–64 years old in EU member states



Source: Facebook monthly active users aged 15–64 per member states as of November 2018 and Eurostat 2017 15–64 year-old population estimate (demo_pop).

Comparing penetration rates by age group, we observe that the number of monthly active users decreases with age, as expected. Although the penetration rate seems to be slightly higher for males than females in the EU member states, the difference is quite small and is in the opposite direction in some countries.

Definitions and methodology caveats

As explained in Section 3.1.2, the definitions and methodology used by Facebook to compute some variables are unclear and the information available online about these computations is scarce. The methodology used by Facebook to estimate the "Lived in" status of users, for example, or their current home location is not publicly available.

Although some studies give indications of the principles used, it is difficult to estimate the reliability and the quality of the data. Despite our attempts to open the discussion with Facebook employees, we have not been able to obtain information around these.

Moreover in the annual report released on 31 January 2019, Facebook admits that while user-provided data suggests a lower penetration rate among the younger age groups, this age data might be unreliable since "a disproportionate number of [their] younger users register with an inaccurate age". In the same report, Facebook discloses that the geographic location of users is estimated using a number of factors, such as the user's IP address and self-reported location. These two methods can be subject to measurement errors, for example if a user is using a proxy server from a different location to access Facebook.

Facebook Membership

Related to the above, there is also some uncertainty pertaining to the ability of the company to identify and close fake or duplicate accounts. In 2017, Facebook revealed that about 2–3 per cent of Facebook accounts were believed to be fake, while 6–10 per cent of accounts were believed to be duplicates. This number is subsequently updated in their quarterly and annual reports. In their US Securities and Exchange Commission (SEC) filing[28] released on 31 January 2019[29] Facebook defines false accounts as falling into one of two categories:

(1) User-misclassified accounts: this happens when a user mistakenly creates a personal profile instead of a page for a non-human entity such as a business, an organisation or a pet. Under Facebook Terms and Conditions such entities are permitted on Facebook using a Page rather than a personal profile.
(2) Undesirable accounts: described as user profiles believed to have been created with the intention to violate Facebook's Terms and Conditions, for purposes such as spamming.

A duplicate account, on the other hand, is defined as an account that an owner maintains in addition to his or her principal account. It is important to note that for a duplicate account to be included in the MAU estimate, the owner must sign into Facebook at least once in the last 30 days. In the last quarter of 2018, Facebook estimated that 11 per cent of worldwide MAUs were duplicates while 5 per cent of accounts were estimated to be false (Facebook SEC filing from 31 January 2019). The document reports that a meaningful share of the duplicate accounts that have been located is from developing markets, such as the Philippines.

To estimate the number of fake or duplicate accounts, Facebook conducts "an internal review of a limited sample of accounts". To identify duplicate accounts, Facebook says it compares IP addresses or user names to identify accounts that are similar. To identify false accounts, it looks for names which "appear to be fake or other behaviour that appears inauthentic to the reviewers". In the document, Facebook recognises that their estimates might not be representative of reality and might be influenced by the methodology they use to evaluate them. The number of false accounts is also subject to episodic spikes in the creation of such accounts, which are reported to often originate from specific countries, such as Indonesia and Vietnam.

[28] US Securities and Exchange Commission 2017
[29] US Securities and Exchange Commission 2019

Box 3: Data Protection

Data Protection
As the retrieved from the Facebook Marketing API only contains aggregated estimates for groups Facebook users, it does not include any personal data. Therefore GDPR does not apply. We have complied with the user agreement for the Facebook Marketing API.

Other social-media data sources
In addition to Facebook data as a source for estimating stocks of EU movers, we have explored the possibility of data from LinkedIn, a professional networking platform that operates via websites and mobile apps. Launched in 2003, the platform includes employers posting jobs and job seekers sharing their education and employment history. Such data would be helpful to compute meaningful migration estimates for different education levels and job categories. However, it rapidly became clear that we would not have access to these data. The data available through LinkedIn API allows filtering by age, gender, education level and sectors for example, but the platform does not allow filtering based on country of origin. It is possible to filter by university attended but it does not seem to be a reliable proxy for the country of origin, especially in the era of the Erasmus Programme, in which student mobility has become common. We have been in contact with representatives of LinkedIn. However, our request for collaboration was rejected and hence, we were unable to collect any relevant LinkedIn data.

3.2. Eurostat population statistics

Eurostat's online data portal provides access to a large number of public datasets classified by themes.[30] For the purpose of this study, we are interested in extracting data regarding the total population for each EU country to benchmark the stocks of EU movers by country and to analyse Facebook penetration by country and by gender and age group, as well as data on population by country of birth in EU countries.

3.2.1. Definitions and data-collection methodology

Population statistics from Eurostat are gathered individually by each member state. Most member states measure the size of the population as of 31 December in the year of reference, and transmit it to Eurostat to be published on 1 January in the following year. Although Eurostat prescribes a set of common definitions to guide member states, countries use different methods to compute their statistics. This section provides an overview of these data and their metadata derived from the most recent report on this matter, "Demographic statistics: A review of definitions and methods of collection in 44 European countries" (Eurostat 2015a).

3.2.1.1. Definition

The definition recommended by the Conference of European Statisticians (CES) (Lanzieri 2014) to measure the population of a country is based on the place of "usual residence", which "means the place where a person normally spends the daily period of rest, regardless of temporary absences for purposes of recreation, holidays, visits to friends and relatives, business, medical treatment or religious pilgrimage."[31] It considers persons who have been living in the Member State for at least 12 months or have arrived during the 12 months before the reference time with the intention to stay for at least a year. However, the definition actually used by member states is much more varied. As shown in Table 4, 23 of 28 member states use the concept of "usually resident population", and 20 use the 12-month time criteria.

[30] Eurostat 2019f
[31] Regulation (EU) No 1260/2013

Table 4: Eurostat population definition by member state

	Registered population	Legal population	Usually resident population	Time criteria
Austria	X			90 days (main residence)
Belgium	X	x	X	None
Bulgaria			X	12 months
Croatia			X	At least 12 months
Cyprus			X	12 months
Czech Republic	X			None
Denmark	X			3 months
Estonia			X	At least 12 months
Finland		x	X	12 months
France		x	X	At least 12 months
Germany	X		X	None
Greece			X	12 months
Hungary			X	12 months
Ireland			X	12 months
Italy	X		X	None
Latvia			X	12 months
Lithuania			X	12 months
Luxembourg			X	12 months
Malta			X	At least 12 months
Netherlands	X			None
Poland	X	x	X	12 months
Portugal			X	12 months
Romania		x	X	At least 12 months
Slovakia	X	x		None
Slovenia			X	At least 12 months
Spain	X		X	12 months
Sweden	X			12 months
United Kingdom			X	12 months

Source: Eurostat (2015a)

3.2.1.2. Methods of collection

Member states also use different approaches to estimate their population size. The majority use their most recent census and adjust it by the variation in population size since the census (12 Member States). Others base the estimate on population registers (8 Member States) and some on both (8 Member States). According to Eurostat (2015a), 26 Member States produce population estimates at reference dates 31 December or 1 January (the difference between these two dates being negligible at national level except where legislation affecting the population count enters into force on 1 January). The two exceptions are Ireland and the UK, where the reference dates used are 1 April and 30 June. Table 5 below summarises the characteristics of estimation methods in each country.

Table 5: Eurostat population estimation methods

	Source for estimating population			Reference date of population estimates			
	Population register(s)	Census - based	Other	1 January	31 December	Mid-year	Other
Austria	x			x			
Belgium	x			x			
Bulgaria		X			X		
Croatia		X	Survey		X	x	
Cyprus		X			X		
Czech		X		x	X	x	Quarterly

Republic							
Denmark	x			x			
Estonia		X		x			
Finland	x				x		
France	x	X		x			
Germany	x	X	Local register and others		x		Average
Greece		X		x			
Hungary	x	X		x			
Ireland		X	Survey				Mid-April
Italy	x		Survey	x	x		
Latvia	x	X	Mathematical methods; several registers	x			
Lithuania	x	X	Foreigners' Register	x			
Luxembourg		X		x			
Malta	x	X	Survey		x		
Netherlands	x			x			
Poland		X			x		
Portugal		X			x		
Romania	x	X	Survey; Econometric models	x		x	
Slovakia		X			x		
Slovenia	x			x		x	1 April, 1 October
Spain	x	X		x		x	
Sweden	x				x		
United Kingdom		X				x	

Source: Eurostat (2015a)

3.2.2. Eurostat datasets used in this study

For the purpose of this study, two datasets were collected from the Eurostat database:

- *demo_pjangroup*: the population on 1 January by age group and gender; and
- *migr_pop3ctb*: the population on 1 January by age group, gender and country of birth.

The first dataset is assumed to be representative of the whole population for each country. We are interested in the information for each of the 28 member states, and we redefine age groups to match the ones described in Table 3. It is useful to benchmark the stocks of migrants by country as well as to analyse Facebook penetration by country and by gender and age group. We used data for years 2011 to 2017.

The second information extracted from Eurostat's online portal breaks down the population of each member state by country of birth. We are particularly interested in EU migrants, i.e. the population born in a different EU member state, by age and gender. We used data for years 2011 to 2017.

Box 4: Data Protection

Data Protection
As Eurostat datasets used in this section only contain aggregated estimates by age group and gender, it does not include any personal data. Therefore, GDPR does not apply.

3.3. EU Labour Force Survey

The third source of data on migrant stocks we use in our model comes from the EU Labour Force Survey (LFS) for years 2016 and 2017. This is a household survey conducted in all EU member states plus Iceland, Norway and Switzerland. While Eurostat's data portal makes publicly available a database of statistics computed from the Labour Force Survey,[32] it does not offer information on the country of birth of respondents, which is the variable of interest for this study. It is possible to apply for access to LFS microdata.[33] For the purpose of this project, Eurostat provided us with statistics for each quarter of 2016 and 2017 disaggregated by gender, age groups (see Table 3), member state and country of birth. This section describes the LFS, the variables relevant to this study and potential caveats of use.

3.3.1. Data collection methodology

All participating countries of the LFS are expected to submit their survey results to Eurostat, which is in charge of processing the data and releasing the main indicators data on a quarterly calendar (except for monthly unemployment indicators, which are realised according to a different calendar).[34] National offices of statistics are responsible for questionnaire design, fieldwork and interviews, based on a common coding scheme. Their questionnaire is split into three categories. The "main indicators" are collected on a quarterly basis and updated four times per year. Annual statistics are also published, which are an average of quarterly statistics. These include indicators such as persons employed part-time and total unemployment rate.

All countries conduct the LFS as a continuous survey, with interviews spread uniformly over all weeks of each quarter. A quarter is defined as beginning on the Monday of the week that contains the first Thursday of the quarter (Eurostat 2017). Various Commission Regulations govern the content and conduct of the survey. Regulation 577/98, for example, specifies in some detail the data that must be provided. Survey designs, characteristics, methods and decision-making processes are also regulated.[35] Eurostat is charged with monitoring compliance with these regulations. Once collected, the data is processed by Eurostat.

3.3.2. Caveats

There are important differences in the organisation of the LFS in member states. In a majority of countries, participation is voluntary (16 member states) while in others it is compulsory (12) (Eurostat 2016a). The creation of the sampling frame also varies. Some countries use census data while others rely exclusively on registers of residents as the basis of their sampling frame. In general, the sampling frame excludes homeless people and those with no registered residences, such as caravans. People living in collective dwellings are excluded from the sampling frames for many countries. Likewise, the basis of stratification and variables for weighting differ between states. Confidence limits for LFS estimates therefore vary. Germany's confidence interval for employment rate as a percentage of population aged between 20 and 64 is 0.2 percentage points. For Lithuania the equivalent figure is 1.3 percentage points. Finally, only 20 of the 28 member states undertook the questionnaire in other languages in 2016 (Eurostat 2018a). Migrants who do not speak the local language (or proposed languages) might refuse to participate in the survey, causing potential under-reporting, or might answer inaccurately, triggering measurement errors. A detailed description of under/over reporting per country is available in Appendix 10 of the EU-LFS quality report (Eurostat 2018a).

[32] Eurostat 2019g
[33] Eurostat 2019h
[34] Eurostat 2019i
[35] Eurostat 2019j

3.3.3. Variables of interest

The LFS contains information on an individual's country of birth (the "CountryB" variable) in order to analyse the labour-market participation of migrants by country of origin, and to be able to identify naturalisations. There is clear guidance from Eurostat on how the CountryB variable is to be determined. It must be collected quarterly, and ISO country classifications used for coding.

The most recent report on the quality of the LFS is from 2015. It indicates some problems with collecting data for 'CountryB'. For example, it was not filled in in Germany for legal reasons.

Nationality information is also collected in the LFS (the "National" variable) in order to analyse participation in the labour market by nationality, according to three groups: national citizens, non-national EU citizens and non-EU citizens. As in the Eurostat data, nationality is clearly defined as the legal concept of citizenship rather than any ethnic concept of nationality. Clear implementation rules are provided in cases where individuals hold multiple citizenships.

Box 5: Data Protection

Data Protection
As EU Labour Force Survey data used in this section only contains aggregated estimates by age group and gender, it does not include any personal data. Therefore, GDPR does not apply.

3.4. Population and Housing Census

The Population and Housing Census is a rich dataset collected every ten years by the EU. It contains information on the total population and housing stock of a country, as well as other demographic and socio-economic characteristics (Eurostat 2011). EU member states retain a deal of autonomy over the exact methods and sources used for collecting their national data, however legislation is in place to harmonise these different datasets. For the purpose of this study, we are interested in extracting data about individuals who were born in an EU country but have since moved to a different EU country. Specifically, we are collecting data on their country of birth and country of usual residence (providing these are different) and the corresponding age range of these people.

3.4.1. Definitions and data-collection methodology

Since each member state collects its own raw data via its designated National Statistical Institute (NSI), in order for the data to be comparable, there are a set of technical specifications for every member state to follow. These are laid out in Commission Regulation (EC) No 1201/2009, which details the statistical units that must be enumerated and under which class they must be assigned to in various questions.[36] For example, age can be measured at different levels of detail and broken down into different levels. Age can be measured at the lowest level of detail – which is measured by broad groups, e.g. "under 15 years" and "15 to 29 years" – or at the highest level of detail (which is every single age). This is one example where the legislation helps ensure comparability and completeness. The legislation largely corresponds to the specifications given in the CES Recommendations (UNECE 2006).

Advantages
The census is a rich and comprehensive source of data since it allows information to be collected at detailed levels, such as individual municipalities and regions. This gives census data a distinct advantage as it goes into a fine level of detail that is unrivalled by

[36] Commission Regulation (EC) No 1201/2009 of 30 November 2009: Implementing Regulation (EC) No 763/2008 of the European Parliament and of the Council on population and housing censuses as regards the technical specifications of the topics and of their breakdowns.

many other data sources. Hence, census data provides the "most reliable and geographically detailed count of the population" (Eurostat 2011, 9). There are several other key advantages of census data (Eurostat 2015b, 10): the characteristics of individual people are recorded separately; the information obtained is in reference to a specific time period, providing a reliable picture of a certain snapshot in time; and it offers good coverage, since much census data covers all individuals within a region.

Disadvantages
However, census data also has some shortcomings, an obvious one being the data is collected only once every ten years (Eurostat 2011, 3). This means the obtained information can go out of date very quickly; for instance, the most recent collection point occurred in 2011. Using it to estimate stock levels for the 2013–2016 period does mean that it is not as accurate as it could be. Reasons for this infrequent data collection are simply a matter of practicality, as collecting data at such a fine level can be very cost-intensive (Eurostat 2011, 9). Furthermore, the EU legislation is output focused – meaning that it makes sure the results are comparable and interpretable, but the methodologies and technologies used to collect the data in the first place are left up to each individual country (Eurostat 2011, 15). This could prove troublesome as there are no assurances of uniformity; one country's data-collection methods could prove to be more sophisticated and more reliable than others.

Definition
As defined by the European Parliament and Council Regulation No 763/2008 on population and housing censuses, the place of usual residence is taken to mean the place that a person "normally spends the daily period of rest, regardless of temporary absences for purposes of recreation, holidays, visits to friends and relatives, business, medical treatment or religious pilgrimage" (Eurostat 2011, 58).
Similarly, information on the country of birth is defined as "the place of usual residence of the mother at the time of the birth, or, if not available, the place in which the birth took place" (Eurostat 2011, 78).

Methods of collection
Methods of collection for the raw data are left up to the individual member states. For instance, the United Kingdom maintained a legal requirement for everyone who had lived or intended to live in the country for three months or more to complete a questionnaire, whereas Belgium did not even approach its population directly, deciding instead to draw solely upon pre-existing register databases (Eurostat 2018b). Eurostat's primary role is as a compiler, with data collection, validation, processing and formatting all ultimately the responsibility of the NSIs. The data-collection methods used by the NSIs can broadly be broken down into the following categories (UNECE 2014):

- **Traditional:** Information is collected on all individuals (full field enumeration), using a combination of methods such as: a paper census form; face-to-face and telephone interviews, or via online means.
- **Rolling census:** Data is gathered from different samples of the population each year, enabling an aggregated census to be created annually if required.
- **Register-based:** Data is drawn from registers of administrative data, which means there is no requirement for new collection of data.
- **Combined:** As the name suggests, involves a mixed-methods approach using a combination of register-based information, as well as some form of "traditional" or sample-survey approach.

Table 6: Population censuses in the UNEC region, 2010 round

Country	Reference date	Type of census
Austria	31 October 2011	Register-based
Belgium	1 January 2011	Register-based (+ data from surveys)
Bulgaria	1 February 2011	Traditional
Croatia	31 March 2011	Traditional
Cyprus	1 October 2011	Traditional

Czech Republic	26 March 2011	Traditional
Denmark	31 December 2010	Register-based
Estonia	31 December 2011	Combined (registers + enumeration)
Finland	31 December 2010	Register-based
France	1 January 2011	Rolling census
Germany	9 May 2011	Combined (registers + enumeration + survey)
Greece	16 March 2011	Traditional
Hungary	1 October 2011	Traditional
Ireland	10 April 2011	Traditional
Italy	23 October 2011	Traditional
Latvia	1 March 2011	Combined (registers + enumeration)
Lithuania	1 March 2011	Combined (registers + enumeration)
Luxembourg	1 February 2011	Traditional
Malta	20 November 2011	Traditional
Netherlands	1 January 2011	Register-based (+ data from surveys)
Poland	31 March 2011	Combined (registers + survey)
Portugal	21 March 2011	Traditional
Romania	22 October 2011	Traditional
Slovakia	21 May 2011	Traditional
Slovenia	1 January 2011	Register-based
Spain	1 November 2011	Combined (registers + survey)
Sweden	31 December 2011	Register-based
United Kingdom	27 March 2011	Traditional

Source: UNECE (2014)

3.4.2. Variables of interest

In this study, information on the following three variables was collected from the 2011 census database (Eurostat 2018):

- C_BIRTH: The place of birth, according to international boundaries in place as of the 1st January 2011, recorded at the country level.
- AGE: Age in years at the time of the census being carried out, broken down into five year bands from age 15 to 64 (i.e. 15–19, 20–24, …60–64).
- GEO: The place of usual residence, measured at the country level.

Given the focus on EU mobility (i.e. migration within the EU), we were only interested in data points where individuals recorded two **different** countries for place of birth and place of usual residence (for example, an individual who recorded the same country for usual residence and place of birth was of no interest to this dataset).

Box 6: Data Protection

Data Protection
As the Population and Housing Census data used in this section only contains aggregated estimates by age group and gender, it does not include any personal data. Therefore, GDPR does not apply.

3.5. Summary of strengths and weaknesses of data sources used in the Bayesian model

As explained in previous sections, each data source has characteristics that can be classified as advantages or disadvantages when they are used for policy-making. Our objective is to use Bayesian modelling to combine all these data sources to take advantage of each source's strengths to overcome other sources' weaknesses. This

section summarises the strengths and weaknesses of each data source in the context of intra-EU mobility policy-making.

To correctly inform policy-making, there is a need for timely, unbiased and accurate estimates of intra-EU mobility. Unfortunately, at the moment none of the four sources cited in the previous section offers all of these characteristics. Arguably, the most reliable source of data with regards to bias and accuracy are censuses, as they cover the entire population and are therefore deemed to be perfectly representative of the population. However, censuses are only conducted every five or ten years depending on the country, so they are rapidly outdated. Although member states share demographic data every year with Eurostat, this data is often based on registries – which are likely to underestimate mobility – and is also rapidly outdated. The labour force survey (LFS) continuously run by member states has the advantages to provide more timely estimates and to be fairly representative of the population by design. However, mobility is a somewhat rare event that is hard to capture through the relatively small samples used for labour force surveys, hence the variance of the estimates may be big as rare events require large samples (Hughes et al. 2016). Facebook marketing API offers both the opportunity to obtain timely estimates and to have a relatively good coverage compared to LFS (the survey was conducted on about 1.5 million individuals quarterly in 2017,[37] while there are more than 242 million Facebook MAU between 15 and 64 years old, as of February 2019). Moreover, Facebook, compared to other social media like Twitter, offers a high level of disaggregation both across space and demographics.[38] Although the bias due to selection might be substantial, it can be monitored using penetration-rate data (Zagheni & Weber 2012; Spyratos et al. 2018). These methods help diminish the selection bias, but might not entirely correct it as the penetration of social media is likely to be driven by unobservable factors. Table 7 offers an overview of the discussion above.

Table 7: Strengths and weaknesses of data sources used in the Bayesian model

Sources	Strengths	Weaknesses
Census	Complete coverage	Only available every 5 or 10 years
Eurostat	Available every year; Coverage is also good	Rapidly outdated; Registries tend to underestimate mobility
LFS	Timely availability as continuous survey; Good representativeness thanks to methods used	Migration is a somewhat rare event and the samples being quite small, variance of mobility estimates is likely to be high across periods
Facebook	Timely availability; Fair coverage compared to LFS; Offers granularity on location and demographics	Biased (can be informed by using penetration rates)

The idea of combining all these sources to produce timely, accurate and unbiased estimates is therefore appealing, and Bayesian modelling offers a method to do so in a consistent way by specifying parameters that will inform bias and accuracy for each data source.

3.6. Comparison of definitions used in the data

As mentioned in the previous sections, since each data source serves its own purpose, the definitions used can vary significantly from one dataset to another. More precisely, the definitions of importance are the concepts of mobility and migration measured by each data source. As each source has its own advantages (see Section 3.5) and these

[37] Eurostat 2019k
[38] Doha Demographics 2017

multiple sources are combined in our model, it is important to reflect on the different definitions of the variables used. These definitions will determine the definition of the output variable: stocks of EU movers. The objective of this section is to offer an overview of the definitions used in each data source and to define the population described by the new model estimates.

3.6.1. Census

From each national census, we use the "usual residence" and the "country of birth" to compute the stocks of EU movers. As explained in Section 3.4, censuses carried out in each member state must comply with Regulation No 763/2008 of the European Parliament and of the Council. As set out by Regulation No 768/2008, the "usual residence" is defined as:

> *"the place where a person normally spends the daily period of rest, regardless of temporary absences for purposes of recreation, holidays, visits to friends and relatives, business, medical treatment or religious pilgrimage."*

The following persons alone shall be considered to be usual residents of the geographical area in question:

- "those who have lived in their place of usual residence for a continuous period of at least 12 months before the reference date; or
- those who arrived in their place of usual residence during the 12 months before the reference date with the intention of staying there for at least one year.
- Where the circumstances described in point (i) or (ii) cannot be established, "usual residence" shall mean the place of legal or registered residence."

In addition, the place/country of birth is "the place of usual residence of the mother at the time of the birth, or, if not available, the place in which the birth took place". For each origin-destination pair we use the usual residence to estimate the destination, and the country of birth to estimate the origin.[39] Due to the definitions presented above, these data points will then comply with the definition of migrant used by Eurostat and the ad-hoc module of LFS 2014 presented in Table 1: "A person who established their usual residence in another country rather than where they were born, for a period that is – or is expected to be – at least 12 months."

3.6.2. Eurostat

Similarly, we also use the "usual residence" variable as the destination country and the country of birth as the origin country in the Eurostat data. The metadata of the dataset (Eurostat 2016a) define:

> *"**Usual residence** means the place where a person normally spends the daily period of rest, regardless of temporary absences for purposes of recreation, holidays, visits to friends and relatives, business, medical treatment or religious pilgrimage."*

The following persons alone are considered to be usually residents of the geographical area in question:

- "those who have lived in their place of usual residence for a continuous period of at least 12 months before the reference time; or
- those who arrived in their place of usual residence during the 12 months before the reference time with the intention of staying there for at least one year."

[39] Origin and destination terms are often used to denote country of previous or next residence, but we do not entertain this concept in the report

"**Country of birth** is the country of residence (in its current borders, if the information is available) of the mother at the time of the birth or, in default, the country (in its current borders, if the information is available) in which the birth took place."

Therefore the Eurostat population data comply with the same definition of migrant described above.

3.6.3. LFS

The definitions used in the LFS main survey are less clear than the ones used in the censuses and by Eurostat. The metadata for the online dataset (Eurostat 2019b) states that the statistical population covers "the total population usually residing in Member States, except for persons living in collective or institutional households". In section 15.3 of the same document, it is explained that "sometimes the rules for defining the usual resident population differ in the LFS from the rule in population statistics". When looking at the definition of the statistical population used by each country (Eurostat 2018c), it appears that each country applies slightly different criteria to define the statistical population and that the 12-month minimum length of (intended) stay is rarely mentioned. We use the country of residence as the destination country and the country of birth as the country of origin. Therefore it is possible that the definition of a migrant in the LFS dataset is less stringent for some countries than the definition used in the census and Eurostat population data, and might include shorter-term migration.

3.6.4. Facebook

The definitions used to compute the Facebook data differ significantly from the three other sources, as Facebook does not comply with any regulation. Moreover, the purpose of the data computed by Facebook is very different from the official data: while official data are computed mainly to inform policy-making, the estimates provided by Facebook aim to advise marketers on the potential audience of their advertisement on the platform. We use the query argument "who live in this location" as a proxy for country of destination and the "Lived in" as a proxy for the country of origin in the Facebook dataset. As explained in Section 3.1, information on the definitions and methods used to compute these two fields is scarce. According to the help section of the Facebook Ads Manager, the "who live in this location" field is determined using the stated city on Facebook users' profiles and is validated based on device and connection information.[40] As of 21 November 2018, Facebook describes "Lived In country X (formerly Expats – Country X)" as "People who used to live in [country] who now live abroad. According to Herdagdelen et al. (2016), the "Lived in" status is determined using the self-reported "current city" mentioned above and "hometown" populated in the "places you have lived" of users' Facebook Network profiles. It is then validated using the structure of the friendships network of the users in each country. Spyratos et al. (2018) recently conducted an online survey to understand how Facebook assigns "Lived In" status and concluded that both "country of home town" and "country of previous residence" might be determinants, as well as other attributes such as geo-tagged information.

Therefore the target group included in the Facebook dataset does not have any condition on the duration of stay in the destination country. We currently lack any information about the distribution of the actual duration of stay within this group. More research would be necessary to determine after how long, or under which conditions Facebook users decide to change their current city, but we believe it is reasonable to assume that users would not change it for very short moves, such as holidays or if they intend to stay for a short period of time (e.g. posted workers).

Moreover, while we use "country of birth" as the country of origin in three datasets from official sources, the definition of a country of origin in the Facebook dataset is "home

[40] Facebook for Developers 2019

country" (derived from "hometown"). This is a self-reported field and there is no clear definition of it. It is left to the interpretation of the user and might be influenced by a feeling of emotional attachment to a town or location, rather than the actual place of birth. Therefore, estimates based on Facebook may not take account of early migrations for such users.

3.6.5. Model-based estimates

The model, presented in Section 4, utilises these four datasets to compute new estimates. As the datasets do not use a single consistent definition of EU movers, it is essential to reflect on the implication this will have for the model-based estimates delivered in this report. The ambiguity around the definitions used by Facebook implies ambiguity about the definitions of the estimates of the stocks of EU movers. Since the duration of residency is more ambiguous in the Facebook dataset, the model-based estimates are also more inclusive in their definition of EU movers. The outputs also include movers who have been residing in a new member state for a shorter period of time than 12 months. Moreover, the estimates are also ambiguous regarding the definition of the country of origin, since they amalgamate the country of birth used in the official data and the "hometown" used in Facebook data.

To account for these inconsistencies with regards to the definition of migration used by Eurostat and avoid confusion, we will refer to the new model estimates as "stocks of EU movers".

4. METHODOLOGY TO ESTIMATE STOCKS OF EU MOVERS

4.1. Modelling Framework

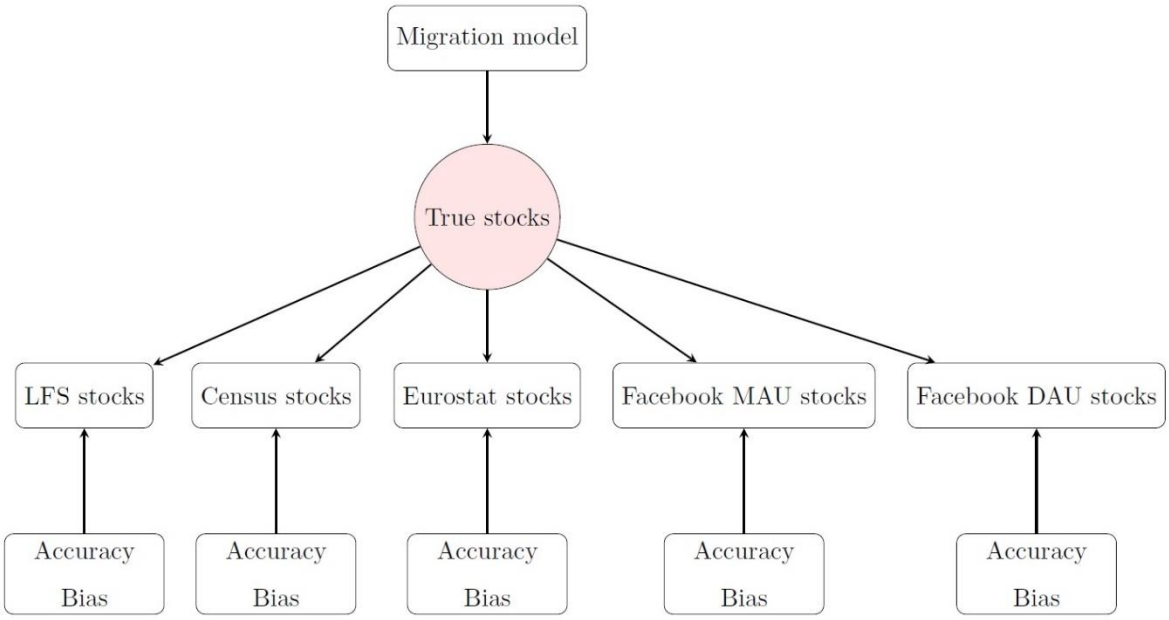
This chapter presents the methodology to estimate unobserved stocks of EU movers by broad age groups and gender for years 2011 to 2018. The methodology consists of two steps. First, we use Bayesian inference to estimate the true stock of EU movers for each year. Then, as a second step, we further disaggregate the stocks of EU movers by age group and gender.

As indicated in earlier chapters, Bayesian inferential framework offers a powerful mechanism to combine data sources and provide measures of uncertainty. Previously, models have been developed for combining traditional migration data sources by members of the research team (see e.g. Bijak et al. 2010; Raymer et al. 2013; Wisniowski et al. 2013; Wisniowski 2017; Wisniowski et al. 2016). We adapt the basic methodologies of these former models to combine population by country-of-birth data from traditional sources and new data derived from social media as explained in Section 3.1.

Our modelling framework to combine traditional data sources with Facebook data is shown in Figure 4. Each layer of the Figure illustrates a hierarchy of the Bayesian model. On top of the figure is the "migration model", which, in our case, relies on a simple autoregressive process where the stocks of EU movers in a current year depend on the stocks of EU movers in the previous year.[41] The hierarchical structure of the parameterisation of that model allows "borrowing of information" from countries that have data available to countries that do not report data on stocks. Below the recursive migration model are the true stocks of EU movers, which are informed by the migration model and fed into the measurement error models. Measurement error models are based on the reported data from Eurostat, 2011 Census, LFS and Facebook. Below each data source are factors that drive the level and variation of the reported data through systematic and random errors respectively.

[41] This assumption relies on the fact that stocks of EU movers in year t depend on the stocks of EU movers observed in year t-1 and on the net migration flows between t-1 and t. Moreover, stocks tend to vary less than flows over time.

Figure 4: Conceptual framework for Bayesian hierarchical model for stocks of EU movers



Note: MAU = monthly active users; DAU = daily active users.

The level of stocks of EU movers (population by country of birth) in reported data tends to be systematically biased compared with the true level of stocks of EU movers. In order to obtain an estimate of the true migration quantity, our model adjusts the reported population by country of birth using metadata on the bias and the accuracy of the reported numbers via five measurement models – one for each data source. This approach is described below.

Our measure of bias refers to the fraction of the unknown true stock of EU movers that is captured in each data source. It is a composite measure that captures the coverage of population (the population targeted to be measured by statistical offices, who provide the population by country-of-birth data to Eurostat), undercount (the number of EU movers missing in the data due to an under-reporting of the total arrivals), or overcount (the number of migrants over-reported due to, for instance, omission of de-registration or differences in definitions).

The variation in reported population by country of birth can be driven by a number of random errors. The size of the noise of the measure of reported population by country-of-birth quantities in each data source can vary according to a number of factors. For example, traditional migration data (reported data from census or administrative data-bases) tend to be more accurate (with lower margins of error) than reported data based on surveys. For Facebook "migrant" measures, the accuracy levels vary much more in reported-data based on Daily Active Users (DAU) than for reported-data based on Monthly Active Users (MAU). We also assume that Facebook data are less precise than the traditional administrative sources, i.e. population by country of birth reported by Eurostat and found in censuses. In all data sources, the variation in the reported data is related to the size of the underlying EU-movers quantity – where, for example, larger stocks have greater associated margins of error.

Total stocks from all data sources are lower than our estimates because of missing data (none of the data sources provides information about population by country-of-birth for each EU member state). However, both Eurostat and LFS show increasing patterns similar to our estimates. The stocks of EU movers for 2017 Facebook MAU estimates are lower than in 2016 and 2018, because data were collected for fewer countries of residence than in the latter two years.

In the Bayesian inferential framework, all parameters require prior distributions. This is the case in our model, in which we construct priors for bias and accuracy, i.e. the measurement model parameters. In our model we have assumed that all parameters in the measurement model have prior distributions based either on meta-information of the data sources or on expert opinion – that may vary by country of reporting data and year – to provide robust estimates of true EU movers quantities. When testing for the impact of these assumptions, we find that the results are sensitive to prior distributions due to short time series of the Facebook data.

4.2. Measurement models

The reported data from different sources are harmonised via measurement models (explained in Section 2.2 as the IMEM approach). These models take both the bias and the accuracy of each data source into account. Measurement model for source *k* is as follows:

$$z_{ijt}^{k} = y_{ijt} \times \gamma_{t}^{k} \times e_{ijt}^{k}.$$

z_{ijt}^{k} represents the reported stocks of EU movers in data source k from origin i residing in destination j at time t. How the country of origin is defined depends on the particular data source at hand; in Eurostat, Census and LFS it is country of birth, whereas in Facebook data, country of origin is defined by "hometown" and derived from their "home country" (see Section 3.1.2). We treat any differences resulting from this inconsistency as negligible and captured in the accuracy e_{ijt}^{k} of the data source described below. This assumption would be of lesser importance when longer series of Facebook data are available, together with meta-information on how the category "Lived In" is constructed by the Facebook Marketing API. Nevertheless, the underlying idea is that the stock data derived from Facebook are benchmarked against the officially reported data from Eurostat and in censuses, the definition of which is known. When the benchmarking is done on sufficiently long series with the implicit assumption that Facebook usage is stable over time and amongst countries, then the difference between traditional data sources and Facebook will become stable over time and the true stocks will reflect the stocks as defined in the official and census data, which are more accurate than Facebook data.

The y_{ijt} component represents the true stocks of EU movers from origin i and residing in country (destination) j in year t. The autoregressive migration model provides a prior distribution for the true stocks. The γ_{t}^{k} parameter reflects the bias from data source k as a fraction of the true stocks. For different data sources, bias parameters are grouped differently (by year, country of residence etc.). The e_{ijt}^{k} reflects the accuracy of data source k. In particular, reported stocks from census data have a much higher level of accuracy than data drawn from a social-media data source. Prior distributions for each data source measurement parameters (γ_{t}^{k} and e_{ijt}^{k}) are explained in detail in Section 4.3.

Finally, to account for the censoring in Facebook data (rounding values lower than 1,000; see Section 3.1.2), in the measurement model we utilize a so-called tobit regression that corrects any potential bias that may occur due to censoring (Tobin 1958). The tobit model also provides a more robust assessment of uncertainty in the resulting posterior distribution for the true stocks of EU movers, especially for the corridors where data are affected by censoring.

4.3. Prior distributions

As mentioned previously, biases in data sources result from a combination of coverage issues and over- or under-counting (referred to as "*undercount"* henceforth to avoid confusion). We assume that the bias in source k is sum of coverage and undercount as follows:

$$\gamma_t^k = \kappa^k + u^k,$$

where coverage and undercount of source k are denoted by κ^k and u^k,respectively. However, without additional information on the quality of sources, especially their coverage of EU movers, it is difficult to estimate these parameters separately as the same level of total bias can be due to pure effect of undercounting with perfect coverage, insufficient coverage with no undercount, or a combination of the two (i.e. the exact source of bias cannot be identified). Such information is usually available through recapture surveys after censuses and used to estimate the missing, hard-to-count population size via capture-recapture methodology (Brown 2000; Chen & Tang 2011; ONS 2012). However, such information is not available for all data sources (especially for social-media data). Therefore, to avoid issues with identifying exact source of bias in the model resulting from the lack of detailed information on both components, and to be consistent in measurement models, we decided to model the sum of these two parameters as the total bias.

When there is no bias the parameter is equal to one. A bias parameter smaller than one suggests that the reported stock of EU movers (population born in another EU member state by country of birth) in the data source is lower than the true stock of EU movers, and a bias parameter higher than one suggests that the data source is subject to overcount. That is, a bias parameter equal to 0.80 means that the data source is reporting the 80 per cent of the true stocks of EU movers.

The 2011 Census series aim at capturing the whole resident population of each country. We assume a high coverage, which does not vary significantly between countries. Therefore, we use a normal prior with a relatively high precision:

$$\gamma^{Census} \sim N\left(b^{Census}, \zeta^{Census}\right),$$

where b^{Census} is the mean and ζ^{Census} is the precision (inverse of variance). This prior distribution is constructed in a way that represents our beliefs on the coverage of population in the censuses (cf. Section 2.2).

To construct prior distributions for bias parameter in Eurostat data, we grouped countries into low (undercount group 1) and high undercount (undercount group 2) countries following Raymer et al. (2013, p. 803).[42] As the same bias parameter is used in a given group, such a grouping allows borrowing of information from countries with available data to inform about the bias in the countries that do not provide data to Eurostat. The same prior distribution $\gamma_{u,t}^{Eurostat} \sim N\left(b_{u,t}^{Eurostat}, \zeta_{u,t}^{Eurostat}\right)$ is used for both groups (u = low, high) and each year (t = 2011, 2012, 2013, 2014, 2015, 2016, 2017). Table 8 shows the undercount groups of countries. The same group is allocated to a country in each year.

Table 8: Groups for Eurostat bias prior distributions. 1 = Low undercount, 2 = High undercount.

Country of residence	Undercount group	Country of residence	Undercount group
Austria	1	**Latvia**	2
Belgium	1	**Lithuania**	2
Bulgaria	2	**Luxembourg**	1
Cyprus	1	**Netherlands**	1
Czech Republic	2	**Poland**	2
Denmark	1	**Romania**	2

[42] The caveat here is that the original grouping in the reference related to migration flows rather than migrant stocks.

Country of residence	Undercount group	Country of residence	Undercount group
Estonia	2	**Slovakia**	2
Finland	1	**Slovenia**	2
Germany	1	**Spain**	1
Hungary	1	**Sweden**	1
Ireland	1	**United Kingdom**	1
Italy	1		

Source: Raymer et al. (2013, p. 803) and authors' own assessment.

It is mentioned in Section 3.3.2 that LFS sampling frame differs for each country. Therefore, country of residence and year-specific bias parameters ($\gamma_{d,t}^{LFS}$) are utilised for LFS measurement error model. The $\gamma_{d,t}^{LFS}$ parameter is normally distributed, $\gamma_{d,t}^{LFS} \sim N(b_{d,t}^{LFS}, \zeta^{LFS})$, where d = Austria, Belgium, …, United Kingdom; t = 2016, 2017.

Facebook MAU and DAU prior distributions for bias parameters are grouped according to the country of residence's estimated coverage of the population and year to reflect the change in the number of Facebook users. The ratios between the estimated number of Facebook users in each country and the population of that country are calculated, and countries of residence are grouped according to these ratios.

Countries of residence in Facebook MAU are allocated to four groups in 2016 and 2017, and five groups in 2018, and countries of residence in Facebook DAU are allocated to five groups in 2018. In each year, both Malta and Cyprus had their own groups with a higher mean value than the rest of the countries. Similar to other sources, the bias parameters are normally distributed with their associated mean and variance, $\gamma_{c,t}^{Facebook\ MAU} \sim N(b_{c,t}^{Facebook\ MAU}, \zeta^{Facebook\ MAU})$, where c in the country of residence group (c = 1, 2, 3, 4 for 2016 and 2017 and c = 1, 2, 3, 4, 5 for 2018) and t = 2016, 2017, 2018. The prior distributions for bias parameters for each group and year are presented in Annex 1.

Table 9: Groups for Facebook bias prior distributions for each country by year and data source

Country of residence	MAU 2016	MAU 2017	MAU 2018	DAU 2018
Austria	1	1	2	2
Belgium	2	3	3	3
Bulgaria	2	2	2	2
Croatia	2	2	2	2
Cyprus	4	4	5	5
Czech Republic	1	2	2	2
Denmark	2	3	3	3
Estonia	2	2	2	2
Finland	2	2	2	2
France	2	2	2	2
Germany	1	1	1	1
Greece	2	2	2	2
Hungary	2	2	3	2
Ireland	2	2	3	3
Italy	2	2	2	2
Latvia	1	1	2	2
Lithuania	2	2	2	2

Country of residence	MAU 2016	MAU 2017	MAU 2018	DAU 2018
Luxembourg	2	2	2	2
Malta	3	3	4	4
Netherlands	2	2	2	2
Poland	1	1	1	2
Portugal	2	2	3	2
Romania	1	2	2	2
Slovakia	2	2	2	2
Slovenia	1	2	2	2
Spain	2	2	2	2
Sweden	2	3	3	3
United Kingdom	2	3	3	3

Source: Authors' own calculations.

The sampling error in each data source is denoted by a precision parameter, e_{ijt}^{k}, which is the inverse of the variance. A high precision means that the data source has high accuracy. For each source except LFS one precision parameter (neither year nor country of residence specific) is used, e^{k}, where k is the data source (k = 2011 Census, Eurostat, Facebook MAU, and Facebook DAU). We assume that the census has the highest precision, then Eurostat and LFS, followed by Facebook MAU and Facebook DAU.

Since LFS sampling design and coverage could differ for each country and year, we estimate the accuracy within the Bayesian hierarchical model assuming that the accuracy is related to the size of the stock of EU movers in each corridor and year. The following model is used to estimate the accuracy for each iteration in our Bayesian model:

$$e_{ijt}^{LFS} = \mathrm{A} + \mathrm{B}\,\mathrm{n}_{ijt}^{LFS},$$

where A is the intercept, B is the slope and n_{ijt}^{LFS} is LFS reported stock of EU movers for origin (country of birth) i, destination (country of residence) j, and year t.

4.4. Migration model

The true stocks of EU movers are driven by a migration model that sits at the top of the hierarchy in our Bayesian model (see conceptual framework in Figure 4). In the current application, rather than relying on the migration model that reflects migration theories on migrant stocks, we take a non-theoretical perspective that allows forecasting and borrowing of strength across time and countries. Forecasting (or now-casting) is crucial as the data provided by Eurostat and LFS are reported with a significant delay (e.g. the most recent Eurostat data available are from 2017; see Section 3), compared to Facebook data which are practically updated in real-time. The Facebook data have been collected over the course of 2018 and provided by project members for 2016–17. The autoregressive forecasting model allows creating a "bridge" between the officially reported data and the newly collected social-media data.

In our model we assume a migration model in which the level of stocks of EU movers can vary. A time series model is used to estimate stocks of EU movers in year *t* based on the stocks of EU movers for the same corridor in year *t-1*. Further, we assume that the stocks of EU movers by country of birth follow a stationary process, that is, we do not expect oscillatory or explosive behaviours of stocks in any particular corridor, but rather stabilisation in the long term (which can be much longer than the period under consideration). However, this specification of the model still allows capturing any sudden changes in the stocks of EU movers, should they appear in any specific corridor. To borrow information from corridors with many observations available to those with scarce or missing data, we create a hierarchical prior distribution for the autoregressive models in which all corridors have their own parameter capturing the corridor-specific mean and

variability over time with all parameters converging to the grand mean autoregressive and intercept parameters.

4.5. Stocks of EU movers by age and gender

Previous sections explained the methodology to estimate the true stocks of EU movers in each of the 28 EU member states between 2011 and 2018, based on the available data. In this section we explain how stocks of EU movers are disaggregated to three age groups (15–24, 25–54 and 55–64) for males and females.

Each data source in our Bayesian model has a different age and gender distribution because of the issues explained in Chapter 3. In addition, many stocks of EU movers by country of birth are missing in the sources, and there is no source which provides age-group and gender distribution for stocks of EU movers in each EU member state by country of birth and for each year. Therefore, age-group gender proportions averaged over origin, destination and year are used to estimate stocks of EU movers by age groups and gender.

The first idea to disaggregate the stocks is using information from every data source. We want to give more weight to data sources which have higher precision, hence we trust the proportions more. To achieve this, we first calculated the normalised weights for each data source according to their posterior accuracy, as follows:

$$w^k = \frac{\mathrm{T}^k}{\sum \mathrm{T}^k},$$

where T^k is the median precision of data source k. Then, stocks of EU movers by origin, destination, age group, gender and year ($\hat{n}_{ijast}$) are estimated using a weighted average. This step is repeated 3,000 times for each of the three chains, with 1,000 iterations saved as the result of the Bayesian model in order to estimate the uncertainty around the estimates. To simplify, we show calculation for one iteration in one chain below:

$$\hat{n}_{asijt} = \hat{n}_{ijt} \sum w^k \, p_{as}^k$$

In the above equation, $\hat{n}_{ijt}$ is the estimated stocks of EU movers for origin i, destination j, and time t; $w^k$is the weight for source k; and p_{as}^k is the proportion of age group a, gender s, in data source k.

The posterior precision of Eurostat is significantly higher than all other data sources' precisions, with a weight of $w^{Eurostat} \simeq 0.99$. This is likely due to the fact that most of the data used in the model come from that source. Therefore, the final estimates are disaggregated using the population by country-of-birth data by age group and gender proportions from Eurostat:

$$\hat{n}_{asijt} = \hat{n}_{ijt} \, p_{as}^{Eurostat}.$$

The median and credible intervals are calculated using samples from posterior distributions obtained from the origin-destination only model as described in Sections 4.2 and 4.3. Those measures of uncertainty do not include variability in age-gender profiles.

4.6. Software and computational details

The full modelling framework has been implemented in open-source Bayesian software, JAGS, operated from within the "R" environment (Plummer 2003). JAGS splits each part of the model into its hierarchical components and underlying sub-model components. The migration model parameters, that are used to derive the estimates of true stocks of EU movers, as well as the measurement parameters, that are highly influenced by the prior distributions based on metadata, are simultaneously estimated using Markov chain Monte Carlo (MCMC) methods operationalised by JAGS.

5. ESTIMATES OF TOTAL STOCKS OF EU MOVERS

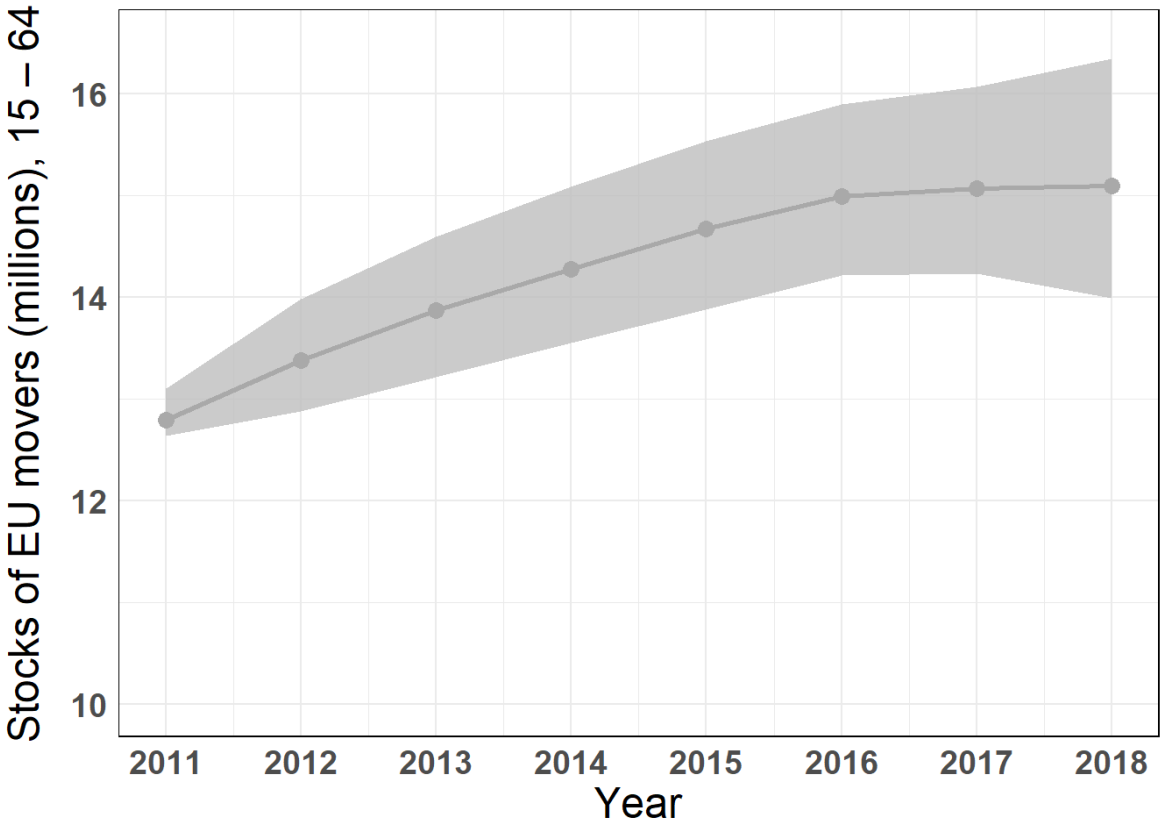
The previous chapter provided a detailed explanation of the methodology to estimate stocks of EU movers in the 28 EU member states. Applying this Bayesian model to a combination of official statistics, household surveys, census data and Facebook data, we can estimate the total number of EU in-movers and out-movers for each individual member state. These estimates can subsequently be disaggregated by age group and gender. The migration model, measurement-error models and prior distributions for the bias and accuracy parameters are presented in Annex 1. The following sections discuss these results in more detail.

5.1. Overall stocks of EU movers

In our model we use a migration model that allows variation in the level of EU movers for each origin-destination pair, and each year, to vary alongside prior distributions for the measurement parameters: the bias and accuracy.

Figure 5 plots the overall stocks of EU movers of working age (between 15 and 64) residing in another EU country. We used reported data from 2011 Census, Eurostat (beginning in 2011, ending in 2017), LFS (2016 and 2017), Facebook MAU (2016, 2017 and 2018) and Facebook DAU (2018) data. The model specification can be found in Annex 1. Our stocks model estimates slightly more than 15 million EU movers living in another EU member state in 2018, an increase from 2016 and 2017.

Figure 5: Estimate of total stocks of EU movers (15–64) living in EU countries with 80% prediction interval



5.2. Immigrants

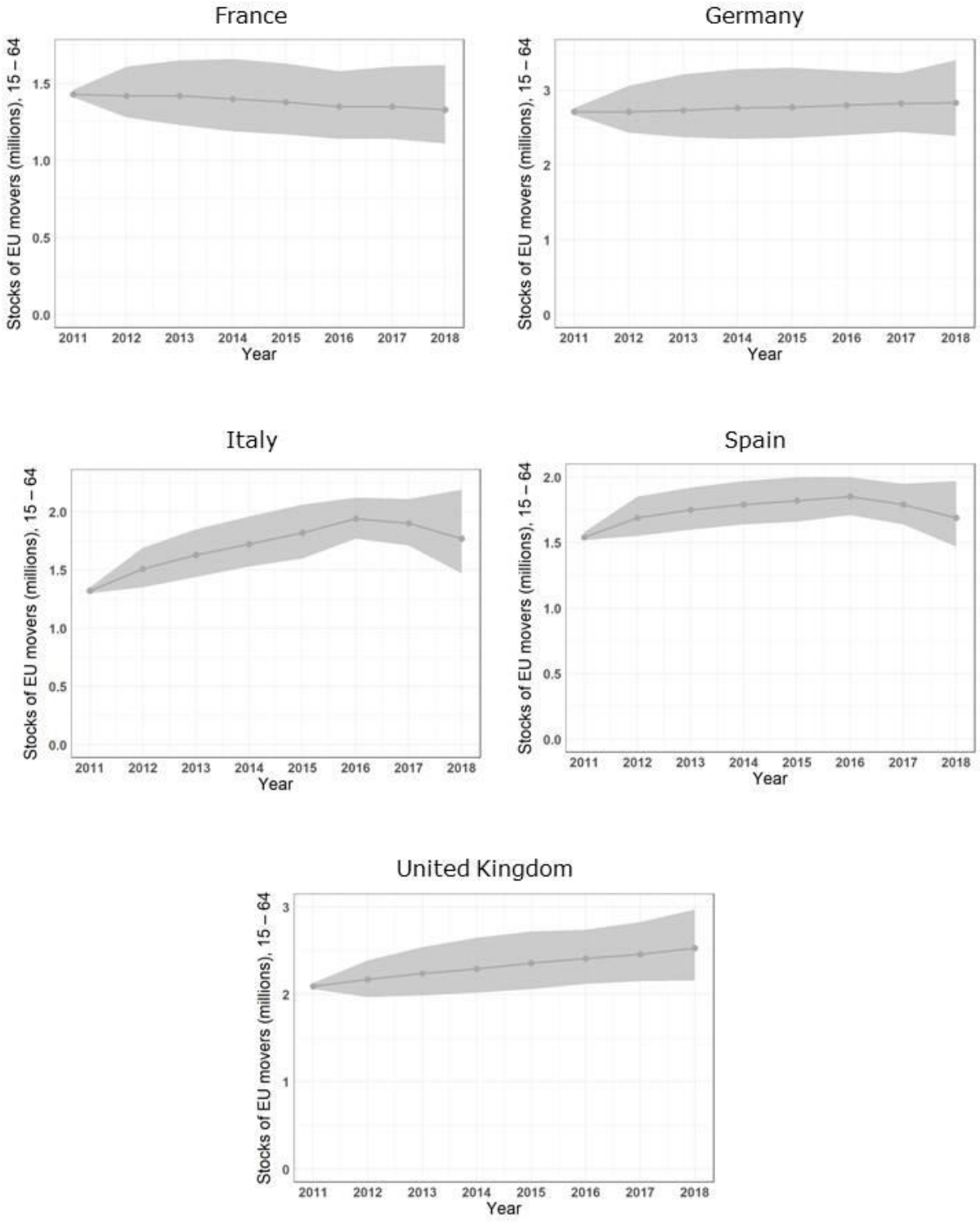
Decomposing the EU-wide stock, we can observe the estimated trend in the stock of EU movers by country of destination.

In countries such as Germany, where reported Eurostat data on the number of foreign-born EU movers (by country of birth) are not available, the number of immigrants in earlier years is estimated between the levels of LFS and Facebook data. Consequently, the predictive intervals for Germany (and similarly for France) are wider than the predictive intervals for other countries – i.e. a greater uncertainty in the estimated levels of EU movers – because their estimates are based on less information, and the size of stock of EU movers in these countries is higher than the stock of EU movers in other countries

Figure 6 shows the total stock of EU movers of working age (15-64) in Germany, France, Italy, Spain and the United Kingdom respectively as estimated by our model. Notably, all five countries show either near constant (Germany and the United Kingdom) or decreasing (France, Italy and Spain) stocks of EU movers.

In countries such as Germany, where reported Eurostat data on the number of foreign-born EU movers (by country of birth) are not available, the number of immigrants in earlier years is estimated between the levels of LFS and Facebook data. Consequently, the predictive intervals for Germany (and similarly for France) are wider than the predictive intervals for other countries – i.e. a greater uncertainty in the estimated levels of EU movers – because their estimates are based on less information, and the size of stock of EU movers in these countries is higher than the stock of EU movers in other countries

Figure 6: Stocks of EU movers in millions (15-64) living in major destination countries



We continue with illustrating the results of our model by showing stocks of EU movers in two exemplary countries in more detail: Netherlands and Poland. These countries are chosen according to the difference in their data availabilities. While Eurostat reports migrant stocks for Netherlands, such data are unavailable for Poland.

Figure 7 shows that Eurostat reports population in Netherlands born in other EU member states. In contrast, Figure 8 shows that no Eurostat-reported population by country-of-birth data are available for Poland. The EU movers living in Poland are estimated based

on 2011 Census, which is missing for many countries of birth, Facebook DAU and Facebook MAU data sources. Whereas, the EU immigrants in Netherlands are estimated based on Eurostat, LFS, Facebook DAU and Facebook MAU data sources. Therefore, the estimated stocks of EU movers in Netherlands have significantly narrower 80 per cent predictive intervals than in Poland.

The number of EU movers from the UK, Germany and Poland living in Netherlands are estimated with a relatively wider predictive interval than the stocks of EU movers from other countries. This is due to the larger size of the stocks of EU movers. In these two figures, the German-born stock of EU movers in Poland has the largest predictive interval. For 2018, the 80 per cent predictive interval for numbers of Germany-born movers living in Poland is between less than 13,000 and 38,000 people.

Figure 7: Total EU movers (15–64) in Netherlands by country of origin. Displays arranged by broad geographic location of origin country

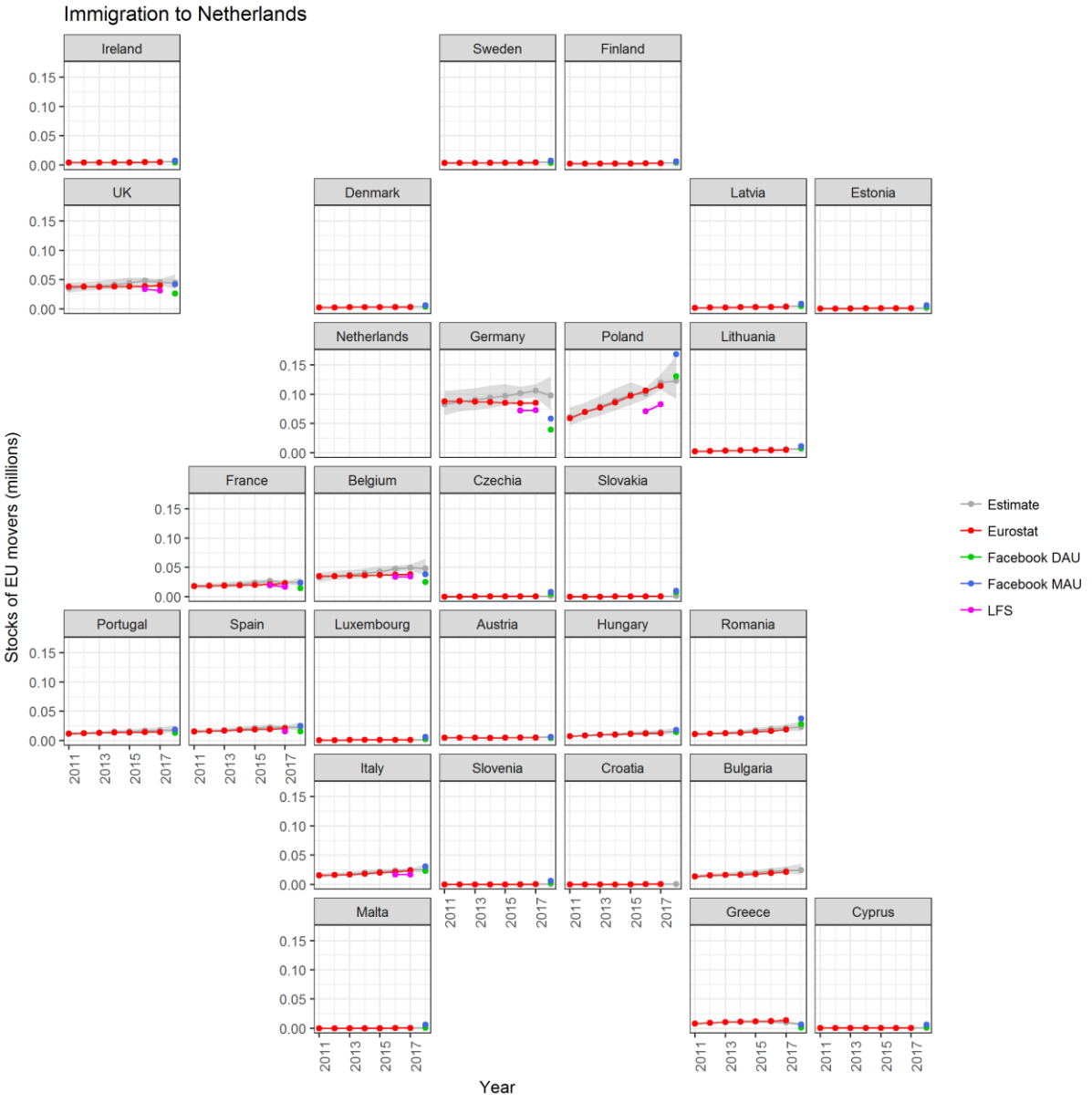
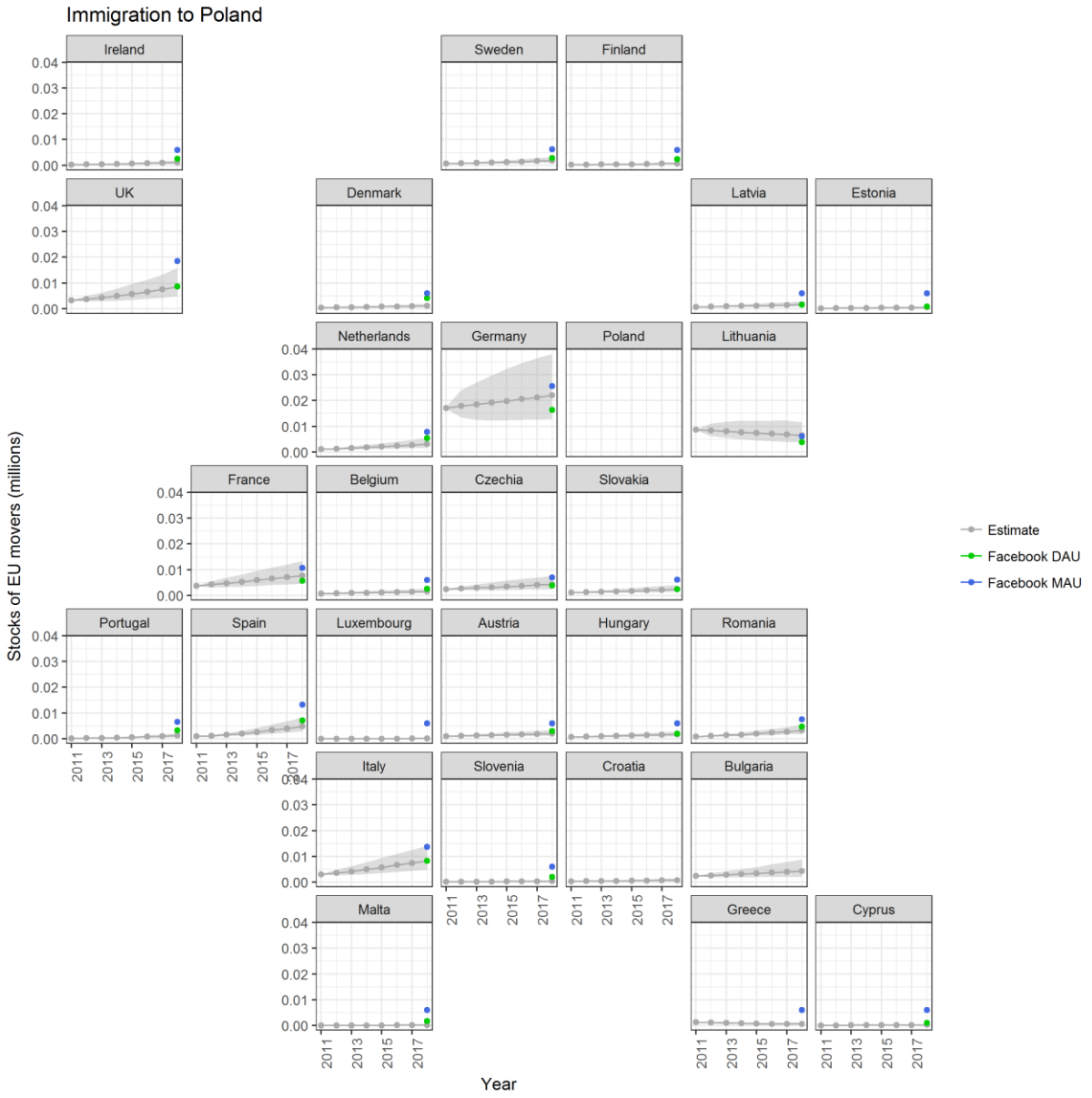


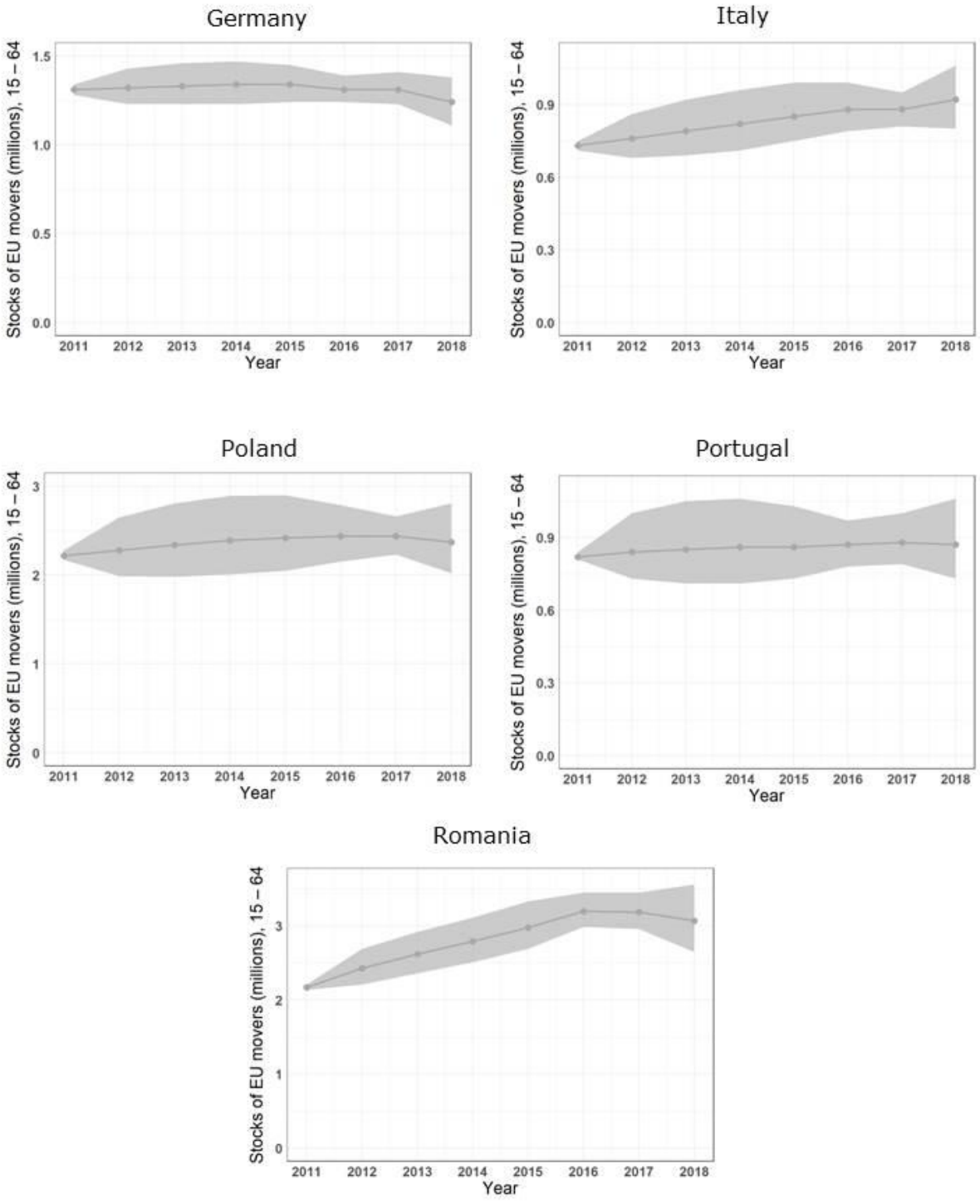
Figure 8: Total EU movers (15–64) in Poland by country of origin. Displays arranged by broad geographic location of origin country



5.3. Emigrants by country of destination

Figure 9 shows the total stock of EU movers from Germany, Italy, Poland, Portugal and Romania, the five first countries of origin. As Eurostat data are collected in the country of residence, not birth, the totals for the foreign-born populations are incomplete in all countries, as there is at least one country in each year that does not collect data. Our model takes advantage of the completeness in availability of Facebook MAU data and the estimates of the true stocks of EU movers are based on information over all corridors. Similar to stocks of EU movers by country of residence presented in Figure 6, the estimates of EU movers abroad from these EU member states are relatively constant over time (Poland, Portugal), slightly decreasing in the last years (Germany, Romania) or slightly increasing in 2018 (Italy).

Figure 9: Stocks of EU movers in million (15 - 64) by major countries of origin



Similar to Section 0, here we present the stocks of EU movers from Netherlands (Figure 10) and Poland (Figure 11) living in another EU member state.

Figure 10: Total EU movers (15–64) from Netherlands living in an EU country. The EU destination countries are arranged by broad geographic location

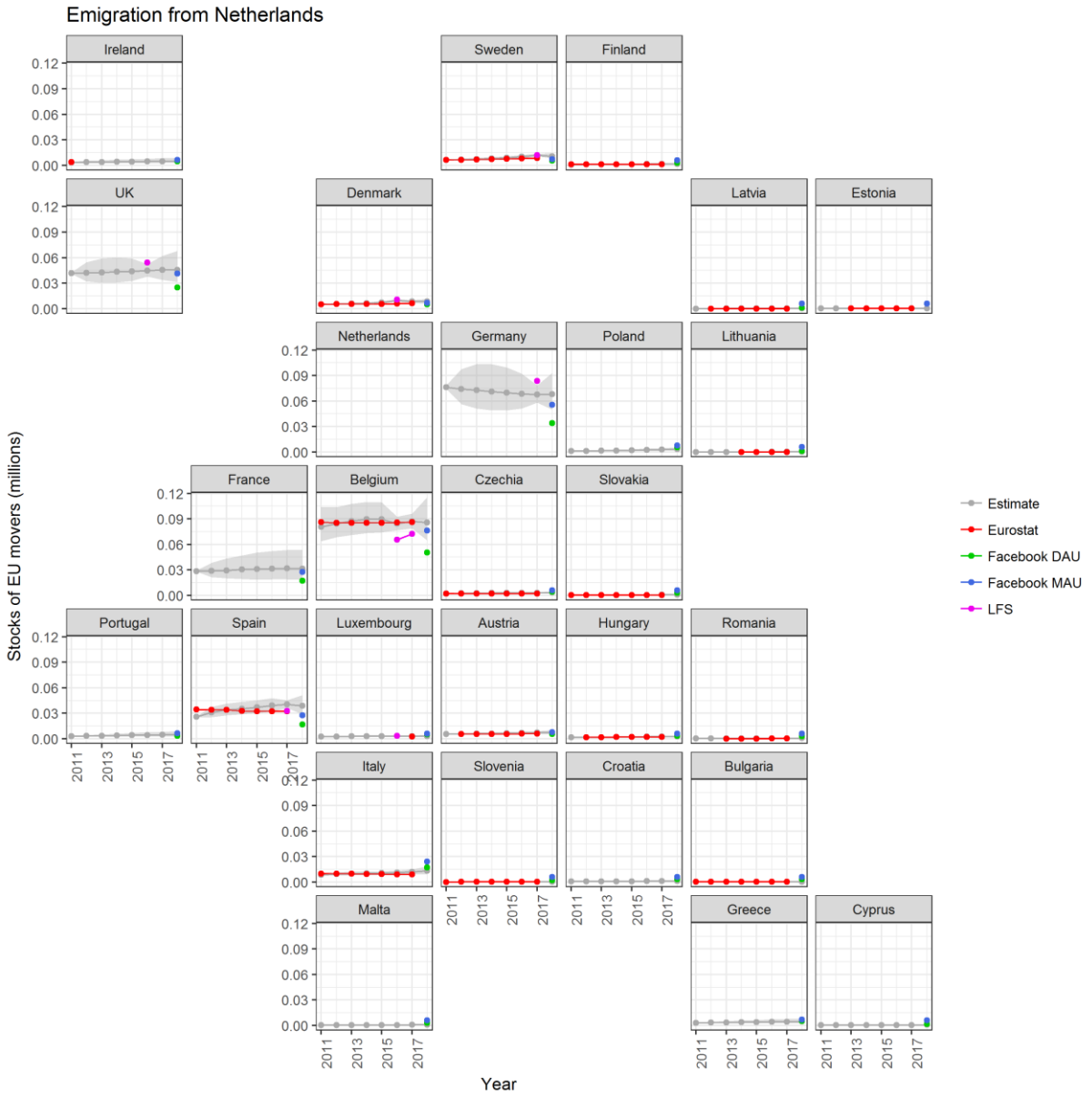
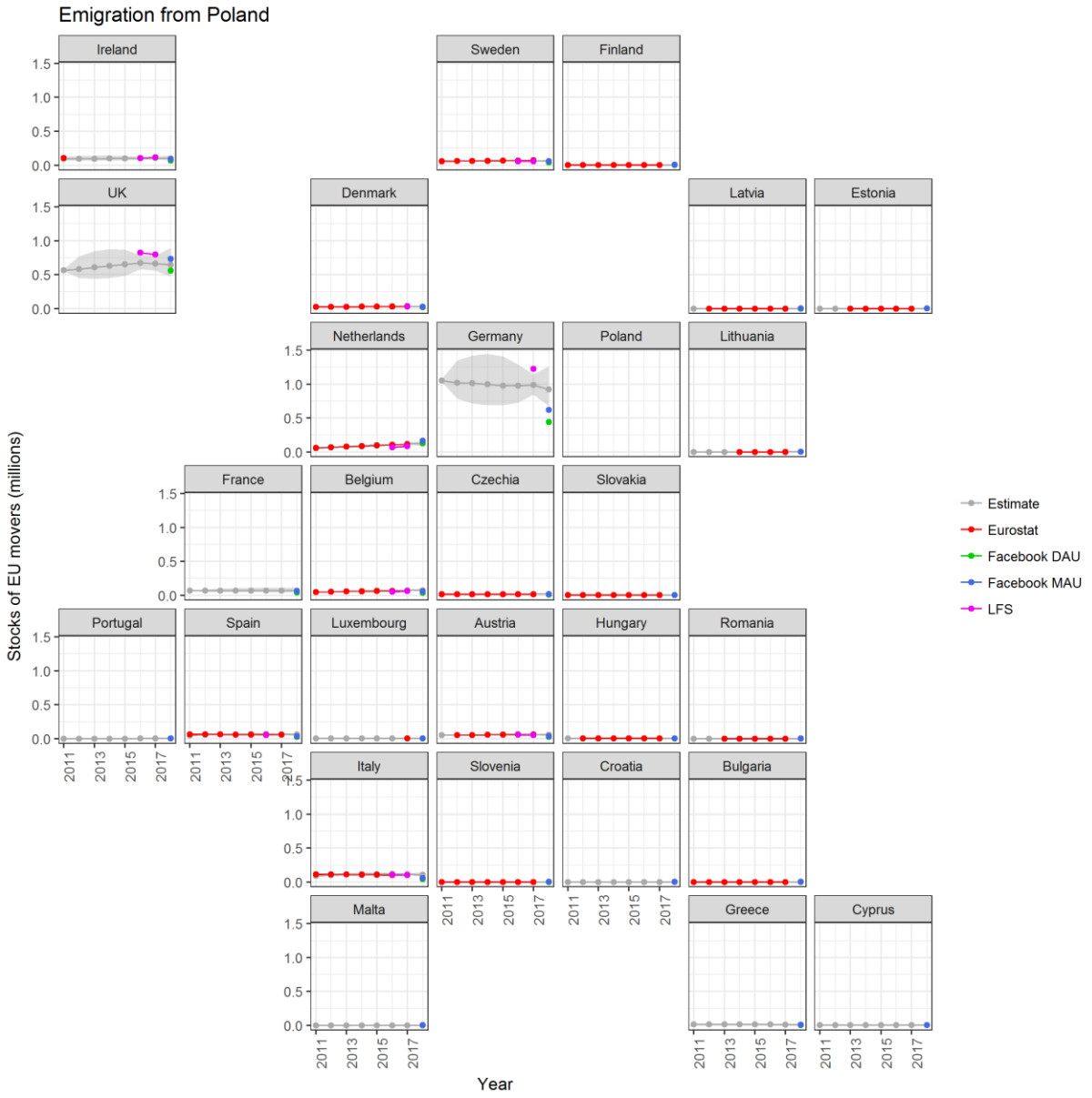


Figure 11: Total movers (15–64) from Poland living in an EU country. The EU destination countries are arranged by broad geographic location

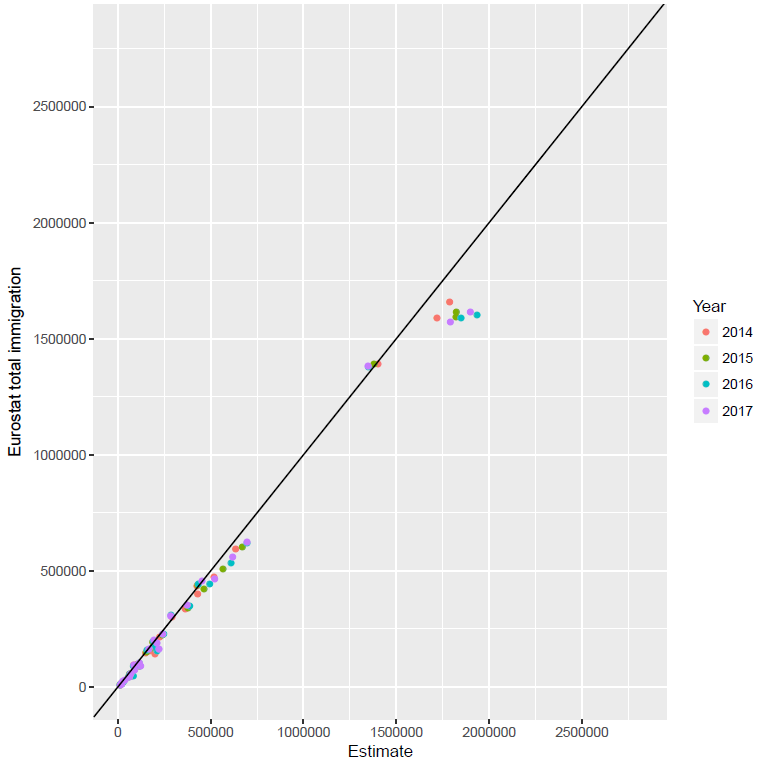


5.4. Comparison with official statistics

To conclude this chapter, in this section we discuss how the estimates compare with the official statistics.

Figure 12 compares the total EU movers living in another EU country (*migr_pop3ctb)* and the medians of the posterior distributions for true stocks of EU movers. These Eurostat estimates are aggregate numbers (total number EU movers in each member state, published by Eurostat) (note that these aggregate data without detailed breakdown by country of birth are not used in the model). These data include "total stocks of EU movers" living in all other EU countries. Sometimes, countries have missing data on the origin of an EU mover and hence do not have origin-destination corridor estimates, but they have an estimate about how many total EU citizens are living in that country. Each EU country of residence is denoted by a dot. The points under the line are the countries in which our estimates are higher than reported number of EU movers in Eurostat. It shows that for most of the countries our estimates are similar to those from Eurostat. We observe no pattern of under- or over-counting over time. However, for a handful of countries (mainly Italy and Spain) our estimates are higher than the reported number of EU movers in Eurostat, suggesting that these countries may have missing observations.

Figure 12: Total stocks of EU movers (15–64) living in another EU country reported in Eurostat vs estimates



Facebook 2018 MAU and DAU include the highest number of origin-destination corridors. However, they do not contain data on EU movers residing in Bulgaria and Croatia. Moreover, as mentioned in Chapter 3, Facebook neither represents everyone in a country nor provides information on how the "Lived In" status is estimated. Facebook MAU data for 2018 includes information for more corridors than that for 2016 and 2017. In addition, the age and gender disaggregation is not available for the earlier two years. Both Eurostat and Census report stocks of movers for most of the corridors.

Among all data sources included in our Bayesian model, LFS has the highest share of missing values. As mentioned before we only included LFS population by country-of-birth data in the model if we have information for both genders in all three age groups. Unfortunately, among 1,512 corridors (28 destination x 27 origin x 2 years), only 177 corridors have complete information.

6. DATA SOURCES TO ESTIMATE EU MOBILITY FLOWS

In this chapter, we discuss the data sources and provide a description of the data used in the estimation of EU mobility flows. The methodology for measuring mobility flows makes use of geo-tagged Twitter data and migration statistics from Eurostat. In this section, we describe both datasets and their limitations.

6.1. Twitter data

Twitter is a social-networking platform founded in 2006. It has shown constant growth over the past decade, reaching 335 million monthly active users in Q2 2018, with more than 200 billion tweets sent per year.[43] Much of the activity on Twitter comprises of public posts or "tweets" that are available – with some significant caveats – for data collection. Some studies document that approximately 0.85 per cent of tweets provide exact information about the location of the tweeting users, who need to opt in to provide such information (Sloan et al. 2013; Sloan & Morgan 2015). However, the sheer number of data points ensures that millions of geo-tagged tweets are still sent daily. The main advantages of Twitter data are good accessibility, coverage across countries and population subgroups, and relative simplicity of access to the available information. While the population of Twitter users may not be representative of the working-age EU population overall, the spatial and temporal detail of these data provide a unique opportunity to study population mobility and migration. This section describes the structure of Twitter data and how to access the data (Section 6.1.1), the strategy for using Twitter data to measure mobility flows (Section 6.1.2), the Twitter dataset used in this study (Section 6.1.3) and the constraints of using Twitter data to estimate EU mobility (Section 6.1.4).

6.1.1. Structure and access to the Twitter data

Twitter users are required to create an account representing themselves on the online platform, with all their tweets being associated with that account. Users can follow or be followed by other users and can make a variety of associated personal data – including their username and other optional information, such as language and geolocation – available to others. Since Twitter does not require users to share information such as gender or age in their profile, information about an individual's characteristics is relatively poor compared to information that Facebook collects from its users.

Tweets themselves contain text produced by users as well as a variety of metadata associated with their production, including location, time (stamp) and semantic information relative to other tweets and users, such as hashtags, user references and retweets. Information from newly created tweets can be downloaded through the **Twitter Streaming API**[44], which gives users access to Twitter's stream of tweets following a request through a programmatic interface. Twitter allows programmers to download up to 1 per cent of the tweet feed. To meet the temporal requirement of this project to cover the period 2013 to 2016, a three-step strategy was set up.

1. Newly created tweets were collected through the Twitter Streaming API. The selection of Tweets can be based on a set of keywords, users or location. Using this approach and selecting tweets based on location, the program downloads approx. 700,000 random tweets written from any European country every day.[45] Upon receiving the data, the program automatically stores the tweet and all its metadata in a secure local database so that it can be easily accessed at a later time.

[43] Statista 2019
[44] Twitter, Inc. 2019a
[45] Quantity observed between December 2017 and January 2018.

2. While the streaming service automatically gives priority to geo-tagged tweets, only about 10–20 per cent of all downloaded tweets have the precise geographical coordinates available.[46] Hence, we filter and save only the geo-tagged tweets for future analysis. A separate automated algorithm then goes through each such geo-tagged tweet and extracts its author – a Twitter user who is known to have tweeted at least one geo-tagged tweet in the past – and adds the user to a database of such users if he/she is not in it already. This way we have been progressively building a database of users who tweeted at least once from any European country and whose location or movement may be analysed for a specific period of time.
3. Unfortunately, the Twitter streaming service does not provide information about users' past tweets, which are necessary to assess whether the user changed location. To obtain this information we utilise the **Twitter Search API**, particularly the **Twitter GET user_timeline** REST API, which returns a collection of the most recent tweets (up to 3,200 most recent tweets) posted by the user indicated in the request parameters. For each user in the database of users with at least one geo-tagged tweet a third automated script downloads all recent tweets of the user and saves the data in a database. The number of requests that can be made to the Twitter GET user_timeline API per period of time is limited, and it is thus a bottleneck of the Twitter data-collection process.

According to the Twitter Developer Policy paragraph F.2 (see Annex 4), sharing of Twitter Content accessed through the API must be limited to Tweet IDs, direct message IDs and user IDs and sharing and retention beyond 30 days of a large amount of such information is only allowed in the context of academic research (for more information see Annex 4).[47] For the purposes of this research, we used an existing collection of tweet IDs covering the period running from March 2016 to February 2018.

Using Twitter API, we "rehydrated" the tweet IDs (i.e. requested the full Tweet content through Twitter API using the tweet ID) to obtain information about these tweets, such as the user ID, the geo-location data and the tweet content. Some Twitter users had deleted some of their tweets or had unsubscribed since the tweet IDs had been collected, therefore we were only able to rehydrate about 68 per cent of the whole dataset.

From this dataset, we identified over 450,000 unique users (as described in Step 2) who appeared regularly and frequently over the 2016–2018 period[48]. We passed their user ID to the user_timeline REST API (as described in Step 3). As the process of gathering users' timeline data is time-consuming, we applied some inclusion criteria to filter user IDs to be included in Step 2 and 3. We have data from nine quarters (the first and last quarters are incomplete): the first quarter is considered to be from 1 to 31 March 2016,[49] and quarter 9 is from 1 January to 28 February 2018.[50] The criteria for users to be included in Step 2 and 3 were then as follow:

1. The user must have at least 50 tweets over the period (therefore 2 tweets per month on average).
2. The user must appear in at least 7 of the 9 quarters.
3. The user must appear in at least 2016 Q1 or 2016 Q2.
4. The user must have 2,000 tweets or less over the period.

Criteria 1 and 2 allow filtering users who tweeted "enough" over the period to allow us to estimate their movement (as explained in Section 6.1.2). Criterion 3 is in place to exclude users who likely started to tweet from mid-2016 and are therefore irrelevant to the project. Since geo-tagged tweets are much less frequent than other tweets and the

[46] Proportion observed in our sample during the same time period.
[47] Twitter, Inc. 2017
[48] See section 6.1.3 for more information.
[49] This is because we do not have the data for January and February 2016 at this stage.
[50] Similarly the data for March 2018 was not yet available.

user timeline will contain all 3,200 last tweets (not necessarily just geo-tagged tweets),[51] criterion 4 is in place to limit the number of users for whom we will not retrieve new geo-tagged tweets. Table 10 below shows the number of user IDs included in Step 3 per country, i.e. the users who met criteria 1 to 4. Of the more than 550 million tweets included in the collection of user IDs, only 9 were geotagged in Cyprus. This might be due to a misspecification of the geolocation of the EU, and the location of Cyprus within the EU, in the script that was used to collect the data. Unfortunately, due to this lack of coverage, no flows into or out of Cyprus are observed in the Twitter data.

Table 10: Number of Twitter user IDs passed through the Twitter GET user_timeline per member state

Member State	Number of user IDs
Austria	1,557
Belgium	6,554
Bulgaria	499
Croatia	213
Czech Republic	1,603
Denmark	2,162
Estonia	267
Finland	3,204
France	33,789
Germany	15,746
Greece	2,537
Hungary	878
Ireland	10,540
Italy	28,660
Latvia	1,068
Lithuania	180
Luxembourg	247
Malta	4
Netherlands	17,094
Poland	3,545
Portugal	5,250
Romania	665
Slovakia	245
Slovenia	441
Spain	65,171
Sweden	7,327
United Kingdom	163,308

6.1.2. Estimating mobility flows from geo-tagged Twitter data

Unlike the Facebook Marketing data used to estimate stocks (see Sections 3, 4 and 5), the Twitter data are comprised of raw data documenting user activity on the platform. Whereas the analysis of Facebook data begins with estimates of migrant stocks *provided*

[51] As explained above, the Twitter Search API returns up to 3,200 most recent Tweets.

by Facebook, the analysis of Twitter data begins with raw Twitter data that we must convert into estimates of flows.

When a Twitter user posts to Twitter, they generate a record containing the content of the post (i.e. a combination of text, image, video and links to websites or other posts) as well as metadata corresponding to the post, like the user handle (user ID) and the timestamp. The user can also choose to make public the location associated with the specific post by adding a location tag. Though the popularity of the location tag feature and the precision of locational information captured have varied over time (for a more detailed discussion see Section 6.1.4), roughly 1 per cent of tweets contain locational information.

A geo-tagged Twitter post, however, is not a flow. It is simply a record (a receipt or a trace, perhaps) of a Twitter user in a specific location at a specific time. To estimate flows, we must analyse the posts of each user and infer whether the user transitioned from living in one country to another over a given period. The overall number of users who transitioned is an estimate of the flow.

Adding to the complexity of this task are the definitional issues in measuring EU mobility discussed in Section 1.2. Movement is the norm, and everything from holiday travel to permanent migration is present in the raw Twitter data. The issue is how to parse one kind of movement (i.e. travel, mobility, migration) from the others. While on the one hand we are limited by the overall quality of the Twitter data (see Section 6.1.4), on the other, we are afforded flexibility in how we choose to operationalise these concepts. Below is an outline of the strategy we used for distinguishing mobility from the raw Twitter data. We describe this strategy in more detail in Chapter 7, and it could be improved upon in future iterations of this analysis.

- First, to improve the quality of the flow estimates, we make exclusions based on frequency. Users who are not observed at least three times per calendar year are dropped from the analysis. This means that when estimating flows that occurred between 2016 and 2017, for example, a user must have at least three geo-tagged tweets in both 2016 and 2017.

- Second, to distinguish EU mobility from migration into or out of the EU, we make exclusions based on geography. We focus on flows between EU member states only. For example, a flow from Germany to France is included, while a flow from Algeria to France is not.

- Third, to distinguish travel from mobility, we make decisions based on regularity. We assume that the majority of a user's posts will come from their country of primary residence. Thus, a user contributes to a flow when they are regularly observed in one country in one year (e.g. 2016) and regularly observed in a different country the next year (e.g. 2017). However, those users for whom there is not a clear regular location in either or both years are dropped from the analysis. Following Zagheni et al. (2014), we determine a user as having a "regular location" when their most frequently observed location is at least three times as frequent as their next most frequently observed location. For example, a person who split their time equally between Germany and France in 2016 would be excluded from the 2016 to 2017 flow, regardless of whether they were observed in just one country in 2017. At the same time, a person who is regularly observed in France in 2016 but who took a month-long holiday in Germany would be included in the 2016 to 2017 flow. Thus, by summarizing over a long period (i.e. a year) we attempt to remove information corresponding to temporary travel.

6.1.3. Description of the Twitter dataset

By combining the 2016–2018 collection of rehydrated tweets with the complementary data collected from the user-history API, we created a collection of Tweets spanning 2012 to 2017. There are some caveats associated with this approach that should be noted here. First, because the user-history API caps access to a user's timeline to 3,200

tweets, this approach biases against high-volume Twitter users. If someone tweeted more than 3,200 times in 2016 and 2017, then we cannot access their earlier tweets. Second, the original tweet dataset uses a selection criteria based on geography (i.e. all tweets occurring in Europe) while the user-history API data use a selection criteria based on user (i.e. all tweets from a specified set of users). This causes a slight incongruity in coverage. Whereas the original data cover Europe by definition, the user-history consists of a sample of users who appeared to be residing in the EU in 2016 and 2017.

Table 11 contains counts of frequently and regularly observed users (see Section 6.1.2) by year and by country for the 2012 to 2016 period.[52] For each year, the count is conditional on being observed the following year. The original data, which span roughly 2016 to 2017, contain a large number of users. The data obtained for earlier years is based on the users observed in the original data. Using the Twitter User History API, we retrieved tweets from these users that occurred prior to 2016. The limitations of this approach, discussed in Section 6.1.1, mean that the number of users observed each year declines as we go backward in time.

Table 11: Number of Frequently and Regularly Observed Users by Year and EU Member State

EU member State	2012	2013	2014	2015	2016
Austria	101	209	318	589	3,983
Belgium	517	997	1,659	2,910	16,005
Bulgaria	0	51	106	197	1,379
Croatia	0	56	72	88	307
Cyprus	0	0	0	0	0
Czech Republic	77	170	340	749	4,034
Denmark	201	361	648	1,050	5,509
Estonia	0	28	62	120	635
Finland	163	369	750	1,512	8,475
France	1,247	2,741	5,362	12,040	108,894
Germany	1,237	2,181	3,596	6,727	40,037
Greece	252	422	680	1,235	7,730
Hungary	96	167	255	414	2,199
Ireland	525	1,273	2,384	4,846	27,231
Italy	2,851	5,106	8,230	14,900	76,913
Latvia	99	236	416	677	2,940
Lithuania	16	22	42	75	442
Luxembourg	0	16	28	74	430
Malta	2	0	13	12	7
Netherlands	1,895	3,434	5,702	9,690	46,484
Poland	166	381	713	1,459	10,123
Portugal	159	301	526	1,085	18,541
Romania	0	80	141	287	1,999
Slovakia	0	0	54	111	638
Slovenia	28	52	0	176	902
Spain	3,459	7,883	14,972	28,139	193,686
Sweden	753	1,528	2,488	4,024	18,776
United Kingdom	9,545	23,061	41,030	81,152	452,754

[52] 2017 and 2018 data were not used as we did not have overlap with the Eurostat data for these years.

6.1.4. Caveats

There are several further observations that we must make about the quality of the Twitter data for the purpose of estimating labour mobility, before we conduct our analysis. In this subsection, we enumerate some caveats.

The number of total users varies by year. A flow is most commonly thought of as a *count*. The number of Twitter users, however is not stable. This is true both for our sample (see Table 11) and for Twitter overall. This poses a problem for thinking of flows as counts of people who move from one country to another. As the number of users in our sample increases with time, so should our estimated flows in absolute terms. To get around this issue, we convert our flow estimates to proportions (we call them rates) by dividing them by the total number of users observed in either the origin country or the destination country. We define in-movers to country X as persons who were observed changing their country of residence to country X and out-movers from country X as persons who were observed changing their country of residence X to another country. We compute the rates among observed Twitter users using these formulae:

$$In-mover\ rate\ for\ year\ t = \frac{number\ of\ in-movers\ in\ year\ t}{population\ in\ destination\ country\ in\ year\ t}$$

$$out-mover\ rate\ for\ year\ t = \frac{number\ of\ out-movers\ in\ year\ t}{population\ in\ origin\ country\ in\ year\ t}$$

In the Twitter dataset, "population in country X in year t" is the number of observed users living in country X in year t.

For each year and country, we measure the proportion of users observed in a different EU member state the following year. We call this the out-mover rate. Similarly, for each year and country, we measure the proportion of users observed in a different EU member state the previous year. We call this the in-mover rate. Table 12 shows the mean yearly out-mover rate for all years 2012 to 2016 (i.e. 2012–2013 to 2016–2017).

Table 12: Mean yearly EU member out-mover rate and in-mover rate by EU member states 2012 to 2016

Member State	Average yearly out-mover rate	Average yearly in-mover rate
Austria	0.038	0.035
Belgium	0.011	0.015
Bulgaria	0.031	0.034
Croatia	0.053	0.059
Cyprus	0	0
Czech Republic	0.032	0.049
Denmark	0.019	0.024
Estonia	0.034	0.032
Finland	0.014	0.014
France	0.014	0.012
Germany	0.015	0.017
Greece	0.029	0.016
Hungary	0.036	0.02
Ireland	0.018	0.014
Italy	0.011	0.008
Latvia	0.015	0.008

Lithuania	0.077	0.054
Luxembourg	0.043	0.079
Malta	0.365	0.385
Netherlands	0.007	0.008
Poland	0.024	0.018
Portugal	0.029	0.02
Romania	0.028	0.026
Slovakia	0.022	0.045
Slovenia	0.029	0.03
Spain	0.01	0.008
Sweden	0.009	0.008
United Kingdom	0.004	0.006

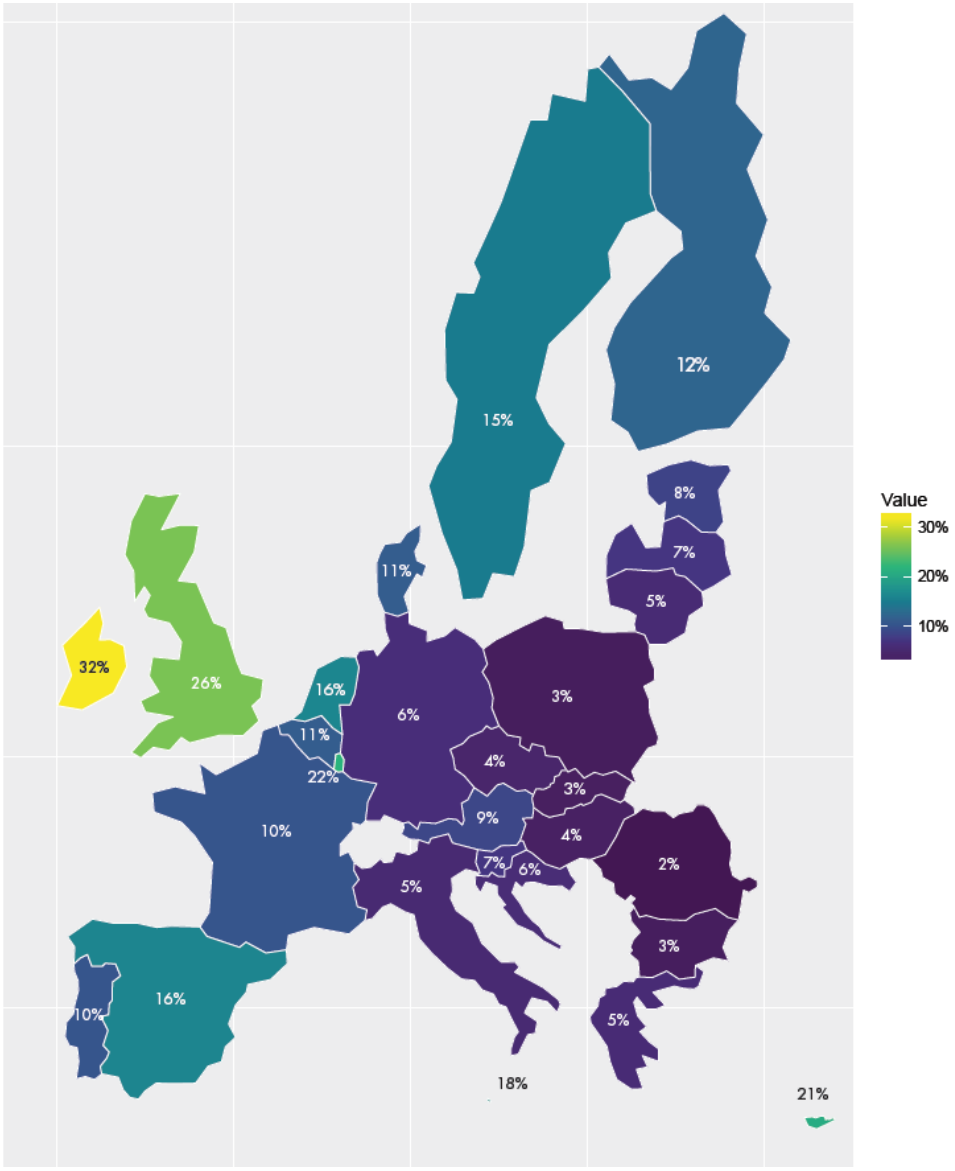
The penetration rate of Twitter varies by country. Twitter is not uniformly popular across EU member states. This means that the distribution of users by EU member state is not the same as the distribution of real-world population by EU member state. While some of the potential bias this may cause will be mitigated by converting our flow estimates into rates, there remains the issue that certain flows may be more visible in the Twitter data than others. In particular, given the hegemony of the English language on the platform, it may be that flows into or out of English-speaking countries are more easily observed. Figure 13 below shows an estimation of the penetration of Twitter in the general population of EU member states. The penetration was computed using the "reach estimate" provided by Twitter to marketers on their Ads Manager[53] (similar process to Facebook),[54] and dividing it by the Eurostat[55] population estimate for 2018 (tps00001).

[53] Twitter, Inc. 2019b

[54] Similar to Facebook, the user needs to "launch a Twitter Ads campaign", enter the specifications of their campaign and access the "Find your audience" page in Targeting, where they can estimate the number of users corresponding to certain targeting criteria who are likely to see the ad (e.g. 13–49 year-old males in Croatia). In Twitter however, the marketer will have to enter a list of "interests" that their audience must have. For the purpose here, we have entered a long list of broad interests until each additional interest only marginally changed the estimate. We have selected each member state as location by turn and all genders and all ages. The estimate comes as an interval with a minimum and a maximum, and we have used the upper bound.

[55] Eurostat 2019f

Figure 13: Twitter penetration rate in the total population, per EU member state



Source: Twitter Ads Manager (as of 14/02/2019) and Eurostat population estimate for 2018 (tps00001 as of 14/02/2019).

The way Twitter captures locational information changed over the period in question. In the early development of Twitter, the default was to capture the precise latitude and longitude corresponding to each tweet (Tasse et al. 2017). Users had to opt-out of posting their location. In spring of 2015, Twitter substantially changed their strategy by making it so users had to consciously choose to post their location. In addition, the level of geographic detail was reduced. Instead of capturing precise latitude and longitude, Tweets with geo-information now contain the latitude and longitude of the place-tag users opted to include. For our purposes, the level of detail included in a place tag is sufficient for enumerating users by country; however, the behaviour that generates locational information is not consistent. Prior to 2015 locational information was captured somewhat passively. Since 2015, it has been captured when users decide to do so. The post-2015 strategy favours certain kinds of mobility behaviour (holiday travel, for example) over others.

Multiple Accounts. Twitter allows users to have several accounts that they can choose to link together or not. This means that tweets coming from two different user IDs in our dataset could actually belong to the same physical person, but we would treat them as two persons in our dataset. There is no known way at the moment to deal with this. As we estimate rate of migrants rather than absolute numbers, this could potentially be an issue if individuals holding several accounts were more or less likely to move than others.

This would artificially increase or decrease the EU mobility rate creating a bias. Unfortunately, we cannot test that this is the case, but we think this is a low risk.

On a related note, some fake accounts are created to artificially increase the number of followers of some accounts and/or to artificially increase the number of re-tweets to spread fake news, for example. The issue is well-known and Twitter claims that they are challenging and closing a large number of such accounts every month (more than 9.9 million accounts were identified and challenged in May 2018, according to Twitter).[56] Most of these fake accounts are Twitter bots, which means that they are controlled via the Twitter API and autonomously perform actions such as tweeting, re-tweeting or following. We think it is unlikely that these Twitter bots activate the geo-location on their tweets and therefore the risk that they impact our EU mobility rates would be low.

Twitter Personal and Corporate accounts. Unlike Facebook, which encourages the creation of pages rather than membership for organisations, Twitter allows organisations to create their own account. Ideally, it would be best to filter corporate accounts out of our list of user IDs. Unfortunately, there is no clear indicator that a Twitter account is a personal account or a corporate account in the information shared by Twitter for each tweet. However, among this information, some could be used as indicators that an account belongs to a person or an organisation. Examples of such indicators are suggested online,[57] such as comparing the numbers of followers against following and whether the account has been verified or not, but more research would be necessary in order to test them.

Box 7: Data Protection

Data Protection
Pursuant to the Developer Agreement (see Annex 3), we have the right to use the Twitter Content for analysis pursuant to Twitter's terms.

The terms include reference to the terms of access by data subjects including those on privacy. The Twitter Privacy Policy (see Annex 4) makes clear that the information provided on Twitter is intended to "broadly and instantly disseminate information [subjects] share publicly through [Twitter's] services." Twitter users are therefore notified that their information publicly posted will be widely used and disseminated.

As third-party analysers of this data, we rely on Article 9(2) (j) of GDPR to process the user's data for scientific research purposes and in doing so we "safeguard the fundamental rights and interest of the data subject(s)" through only using non-identifiable data in our analysis, in the form of aggregated EU mobility flows. No use of user-identifiable data is made, nor any decisions about users' data. To the extent that potentially identifiable user names are stored with "Twitter Content" (as defined in Twitter's terms), these are stored in a secure manner and will be deleted on conclusion of the research.

6.2. Eurostat migration statistics

Our model uses Eurostat migration statistics in combination with the data from Twitter to estimate migration flows. Eurostat's online data portal provides access to a large number of public datasets classified by themes. For this section, we are interested in extracting data regarding the emigration flows per country. This section presents the definitions used by Eurostat (Section 6.2.1), the methods used for the collection of emigration data (Section 6.2.2) and the dataset used in our model and its limitations (Section 6.2.3).

[56] Twitter, Inc. 2018b
[57] Stackoverflow 2019

6.2.1. Definitions

Regarding the definition of emigration, Eurostat's statistics rely on the Regulation (EC) No.862/2007 of the European Parliament and of the Council of 11 July 2007, stating that emigration is "the action by which a person, having previously been usually resident in the territory of a Member State, ceases to have his or her usual residence in that Member State for a period that is, or is expected to be, of at least 12 months". The time criteria used for the definition of emigration slightly varies among the members state, the detail of which is described in Table 13 below.

Table 13: Countries' definition of emigration

	Actual 12-month	Intended 12-month	Actual & intended 12-month	6 months criterion
Austria	X			
Belgium	X			
Bulgaria			X	
Croatia			X	
Cyprus			X	
Czech Republic	X*	X		
Denmark	X			
Estonia	X			
Finland				X
France			X	
Germany	X			
Greece			X	
Hungary		X		
Ireland			X	
Italy		X		
Latvia				
Lithuania			X	
Luxembourg			X	
Malta			X	
Netherlands	X			
Poland			X	
Portugal			X	
Romania	X			
Slovakia			X	
Slovenia			X	
Spain			X	
Sweden				X
United Kingdom		X		

*Only Czech nationals. Source : Eurostat (2018).

Member states assigned to the first column "Actual 12-month" count individuals as emigrants after they have relocated their usual residence outside of the country for at least 12 months. The second column ("Intended 12-month"), on the other hand, contains member states considering individuals as emigrants if they have relocated their usual residence to another country with the intention to leave for at least 12 months. Consequently the third column ("Actual or intended 1-month") shows countries using either definition. Finland and Sweden fall into the last column ("6-month criterion") as they use a buffer zone of 6 months instead of 12. Member states included under column 3 and 4 therefore apply looser definitions of emigration and will likely include more individuals in their emigration statistics. It is also worth noting that Czech Republic only reports emigrants having Czech nationality.

The reference period is the calendar year, hence an emigration will be accounted for the administrative year during which the relocation occurred.

6.2.2. Data collection methodology

The data are primarily gathered by national statistical institutes. Each member state can freely use any relevant data sources, as availability and practices can vary from one member state to another. However, each of them shall comply with the harmonised definitions of migration statistics given by the Regulation (EC) No.862/2007.

The most common sources of data used to evaluate the size of emigration flow are administrative sources (e.g. social insurance data, offices of foreigner registrations), sample surveys (e.g. International Passengers Surveys, National Statistical Surveys), census data, mirror data (usage of other countries' national statistics on immigration to estimate their own statistics), mathematical methods (such as regression analysis and other econometric modelling), or a combination of these sources. The detail of methodologies used in each Member State is displayed in Table 14.

Table 14: Countries' data collection methodologies

	Administrative	Sample survey	Census data	Mirror data	Mathematical methods
Austria			X		
Belgium			X		
Bulgaria			X		
Croatia	X				
Cyprus	X	X			
Czech Republic	X				
Denmark	X		X		
Estonia	X		X		
Finland			X		
France			X		X
Germany					
Greece	X				X
Hungary	X		X		
Ireland	X				
Italy			X		
Latvia	X		X		X
Lithuania	X		X		
Luxembourg			X		
Malta	X	X			X
Netherlands			X		
Poland	X	X	X	X	
Portugal	X	X			
Romania	X			X	
Slovakia		X			
Slovenia			X		
Spain			X		
Sweden			X		
United Kingdom	X	X			

Source: Eurostat (2018).

6.2.3. Eurostat dataset used in our model and limitations

Two datasets were used in our model to estimate migration flows; they were retrieved from the Eurostat database. The first dataset is:

- *migr_emi3nxt* : emigration by age, gender and country of next usual residence.

This dataset contains yearly time-series information on the number of individuals – by age, by gender and by country of next-usual-residence – that emigrated from each member state country from 1990 to 2016 (upon data availability). We emphasise that this study only focuses on the 2012–2016 time period for comparability purposes with the Twitter dataset.

The data suffers from a number of limitations. The first one is the absence of obligation for the member states to break down their number of EU foreigners by individual citizenship. While some of the member states produce such figures, they are doing so on a voluntary basis. This is the case for 20 member states, while 8 member states (Greece, France, Croatia, Cyprus, Luxemburg, Malta, Austria and Poland) only publish the total number of EU foreigners living on their soil, making the estimation of emigration flows difficult, as it prevents the use of mirror statistics.

Another limitation comes from the fact that many member states rely on administrative data to estimate the number of foreigners living on their territories. As a share of migrants might not register with the authorities, the figures are likely to underestimate the true number of emigrants. Even when they actually register, there might be a delay between their arrival and their accounting, leading to a potential bias in the dynamic analysis.

The second data set is:

- *Migr_imm5prv: immigration by age, gender and country of previous usual residence.*

Similarly to the emigration data, this dataset is a yearly time series of the number of individuals by age and gender and by country of previous usual residence, who immigrated to each member state country from 1990 to 2016. While we are interested in the 2012 to 2016 period, data on immigration from other EU countries are also only available from 2013 to 2016 in this dataset.

We also rely on:

- *demo_pjangroup*: the population on 1 January by age group and gender.

This dataset was used to estimate the share of emigrants in a member state's population for each year between 2012 and 2016. More information about this dataset can be found in Section 3.2.

Similar to the rates computed from the Twitter data, we use these three datasets to compute emigration and immigration rates in the following way:

$$Immigration\ rate\ in\ year\ t = \frac{number\ of\ immigrants\ in\ year\ t}{population\ in\ country\ of\ destination\ in\ year\ t}$$

$$Emigration\ rate\ in\ year\ t = \frac{number\ of\ emigrants\ in\ year\ t}{population\ in\ country\ of\ origin\ in\ year\ t}$$

Box 8: Data Protection

Data Protection
As this dataset only contains aggregated estimates by age group and gender, it does not include any personal data, therefore GDPR does not apply.

7. METHODOLOGY TO ESTIMATE MIGRATION FLOWS

The goal of the estimation framework is to combine the two sources of data – Twitter data and Eurostat data. To capture these sources, we use a *joint-modelling approach* to understand the data-generating mechanisms for both Eurostat and Twitter data. From the model, we estimate emigration rates for each country (i.e. the proportion of emigrants compared to the population size). Thus, we aggregate the destination countries into a single-rate measure.

The assumption of the model is that there is a common process for the "true EU mobility rates" in the population and we have two sources of data that measure this process. The Eurostat data estimates this process with random error, while the Twitter data estimates this process with both random error and bias. We assume Eurostat draws from sources of representative samples, but the estimates are affected by random sampling error. On the other hand, because Twitter users are not representative of the general population, estimates from Twitter may be inherently biased as well as being affected by random sampling error. In other words, the Eurostat and the Twitter estimates for a particular country-year would be bivariate normal with a partially shared mean component (the part without the bias). Over the countries and years, the Eurostat and Twitter estimates would be multivariate normal with a partially shared mean component.

Since emigration rates can be interpreted as probabilities[58] that lie between zero and one, we model the logit[59] of the emigration probabilities to allow for numeric space to cover continuously from negative infinity to infinity. Formally, denote:

Y_{ES-ct} as the logit of the emigration estimates from Eurostat for country *c*, year *t*
Y_{TW-ct} as the logit of the emigration estimates from Twitter for country *c*, year *t*

Then, following the method by Mercer et al. (2015):

$$Y_{ES-ct} \sim N(\mu_{ct}\ ,\ V_{ES-ct})$$
$$Y_{TW-ct} \sim N(\mu_{ct} + B_{ct}\ ,\ V_{TW-ct})$$

We emphasise the common mean component μ_{ct} in both Eurostat and Twitter estimates. Also note B_{ct}, which represents the bias component for Twitter estimates (we assume that Eurostat estimates are unbiased[60]).

Because of the common mean component, we model the two processes jointly. That is:

$$\begin{pmatrix} Y_{ES-ct} \\ Y_{TW-ct} \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_{ct} \\ \mu_{ct} + B_{ct} \end{pmatrix}, \begin{pmatrix} V_{ES-ct} & 0 \\ 0 & V_{TW-ct} \end{pmatrix}\right]$$

Here we assume that the covariance terms are zero as the measurement errors of Eurostat and Twitter are independent. We model both Y_{ES-ct} and Y_{TW-ct} as processes of space-time interactions. That is:

$$\mu_{ct} = \mu + \theta_c + \varphi_c + \alpha_t + \gamma_t + \delta_{ct}$$

Where μ is an overall mean, θ_c is a spatial intrinsic conditional autoregressive process (ICAR) defined over the adjacency matrix of European countries (i.e. a 28x28 matrix of 1s and 0s for country-pairs that are adjacent), φ_c is a random IID intercept for each country, α_t is a random walk of order 2 process (RW2), γ_t is a random IID intercept for

[58] If we randomly pick an individual in year t in a population, the emigration rate for year t in this population can be interpreted as the probability that this individual will emigrate in year t.

[59] Logarithm of the odds ratio: $\log(\frac{p}{1-p})$

[60] As explained in section 6.2, Eurostat data have a number of non-negligible limitations (e.g. difference in definitions across countries, missing values, underestimations). However, for the purpose of this research, and as a first step to investigating the bias of Twitter data, we first make this assumption.

each year, δ_{ct} is a structured interaction between the ICAR process and the RW2 process. Similarly, the bias is also modeled as:

$$B_{ct} = \mu + \theta_c + \varphi_c + \alpha_t + \gamma_t + \delta_{st}$$

with the same spatial and temporal components. We fit the model using the *INLA* package in the statistical software *R*.

To evaluate the performance of the model, we adopt a cross-validation approach and try to predict the last year (i.e. year 2016) of the Eurostat estimates. The intuition is to see if the model can offer reliable estimates of EU mobility flows for recent periods when official statistics are not yet available, which would provide more timely estimates using recent geo-tagged data from social media. We compare two models:

1. The joint model that uses Eurostat data from years 2012–2015 and Twitter data from years 2012–2016.
2. A "Eurostat only" model that uses Eurostat data from years 2012–2015.

Ideally, the joint model should have higher prediction accuracy. Prediction accuracy is assessed with the squared error for each country and the summary statistics of root mean squared error (RMSE). Formally, let $\hat{Y}_{ES-c,2016}$ be the Eurostat estimate for country *c*, year 2016; also let $\hat{Y}_{Joint-c,2016}$ be the estimate from the joint model and $\hat{Y}_{ES\_only-c,2016}$ be the estimate from the "Eurostat only" model, then:

$$RMSE_{Joint} = \sqrt{\sum (\hat{Y}_{Joint-c,2016} - \hat{Y}_{ES-c,2016})^2}$$

and

$$RMSE_{ES\_only} = \sqrt{\sum (\hat{Y}_{ES\_only-c,2016} - \hat{Y}_{ES-c,2016})^2}$$

Intuitively, $RMSE_{Joint}$ should be smaller than $RMSE_{ES\_only}$

8. ESTIMATES OF TOTAL EU MOBILITY FLOWS

8.1. Initial results

We first present the raw estimates of emigration computed from the Twitter data, and compare them with the Eurostat estimates in Figure 14. As examples, we also present estimates for Belgium and Germany in Figure 15. The horizontal axis is the year, while the vertical axis is the emigration rate bounded by 0 and 1. The red line represents the Eurostat estimates of emigration rates, while the blue line represents the Twitter estimates. We can observe a few points:

1. In general, the Twitter estimates are larger than the Eurostat estimates. If we treat the Eurostat estimates as unbiased, then the Twitter estimates are upward-biased.
2. The variance of the Twitter estimates is much larger than the Eurostat estimates.
3. There are certain country-years where the Twitter estimates are much larger (e.g. Croatia 2014). This is most likely due to the small sample sizes of Twitter users in particular country-years.
4. There does not appear to be a consistent relationship between the Eurostat estimates and the Twitter estimates.
5. There appears to be an increase of Twitter estimates in year 2016 even when the migration rates from Eurostat do not increase. This could be due to our sampling strategy, where we select Twitter users observable in 2016 and collect their Twitter history.

Figure 14: Comparison of Twitter and Eurostat emigration estimates by country

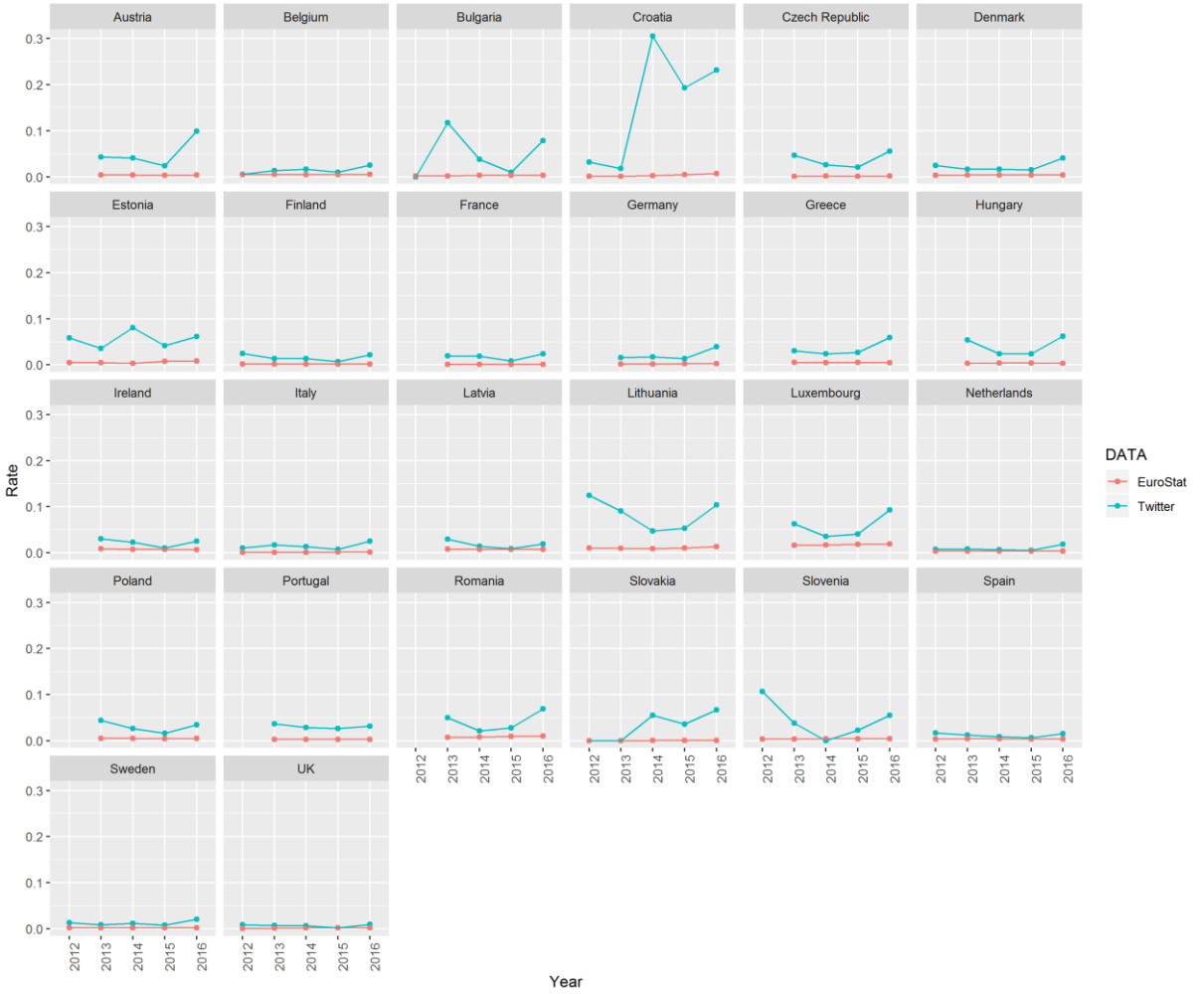
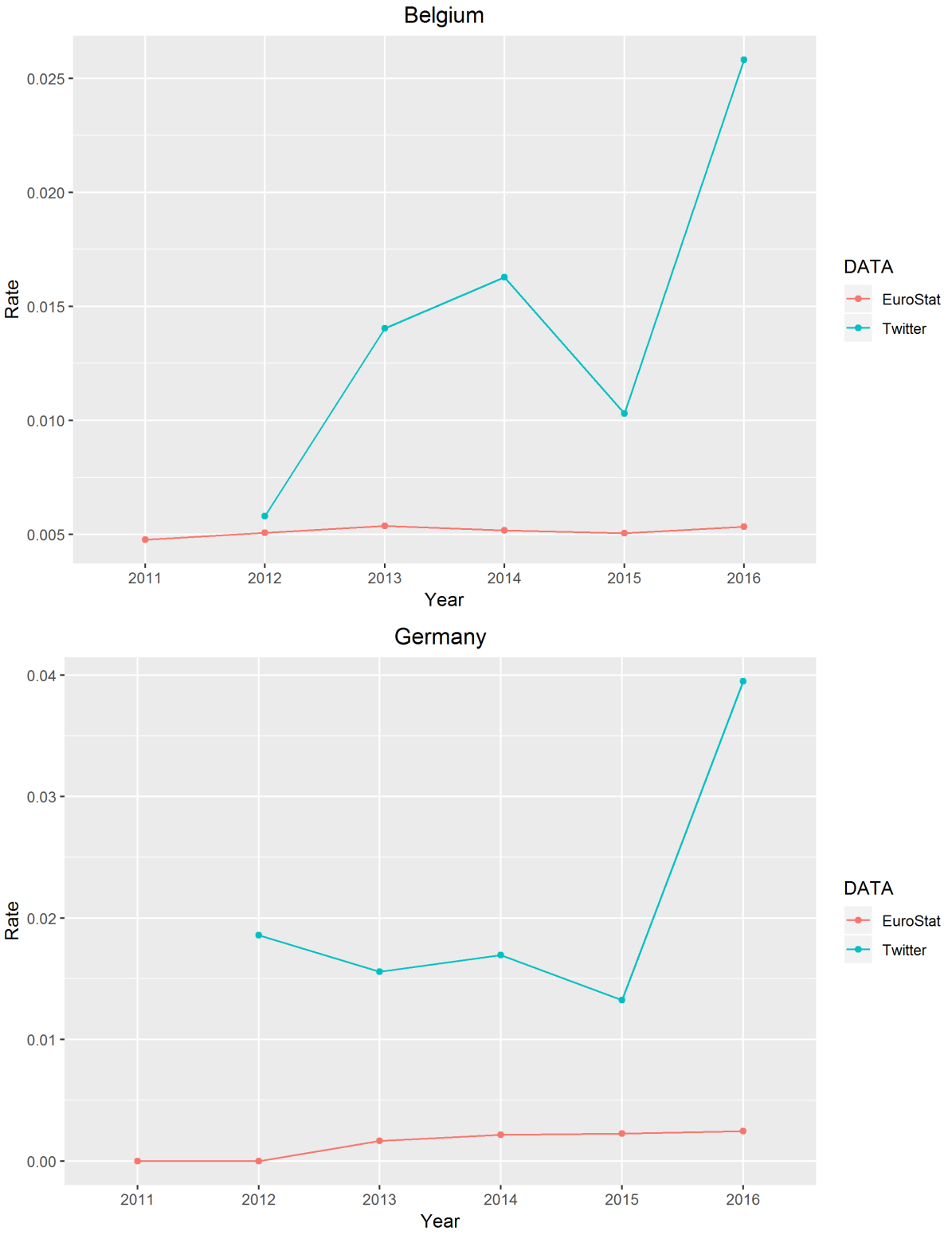


Figure 15: Comparison of Twitter and Eurostat emigration estimates in Belgium and Germany



We then present the raw estimates of *immigration* computed from the Twitter data, and compare them with the Eurostat estimates in Figure 16. As examples, we also present estimates for Belgium and Germany in Figure 17. We can observe a few points:

1. Similar to the results for emigration, the Twitter estimates are larger than the Eurostat estimates. If we treat the Eurostat estimates as unbiased, then the Twitter estimates are upward-biased.

2. The plots suggest that the variance of the Twitter estimates is again much larger than the Eurostat estimates.
3. Similar to the results on emigration rates, there are certain country-years where the Twitter estimates are much larger (e.g. Croatia 2014). This is most likely due to the small sample sizes of Twitter users in particular country-years.
4. There does not appear to be a consistent relationship between the Eurostat estimates and the Twitter estimates.
5. There appears to be an increase of Twitter estimates in year 2017 (i.e. the last year of observation from Twitter), even when the migration rates from Eurostat do not increase. Thus, it seems that for both emigration and immigration rates the Twitter estimates for the last year are particularly high.

Figure 16: Comparison of Twitter and Eurostat immigration estimates by country

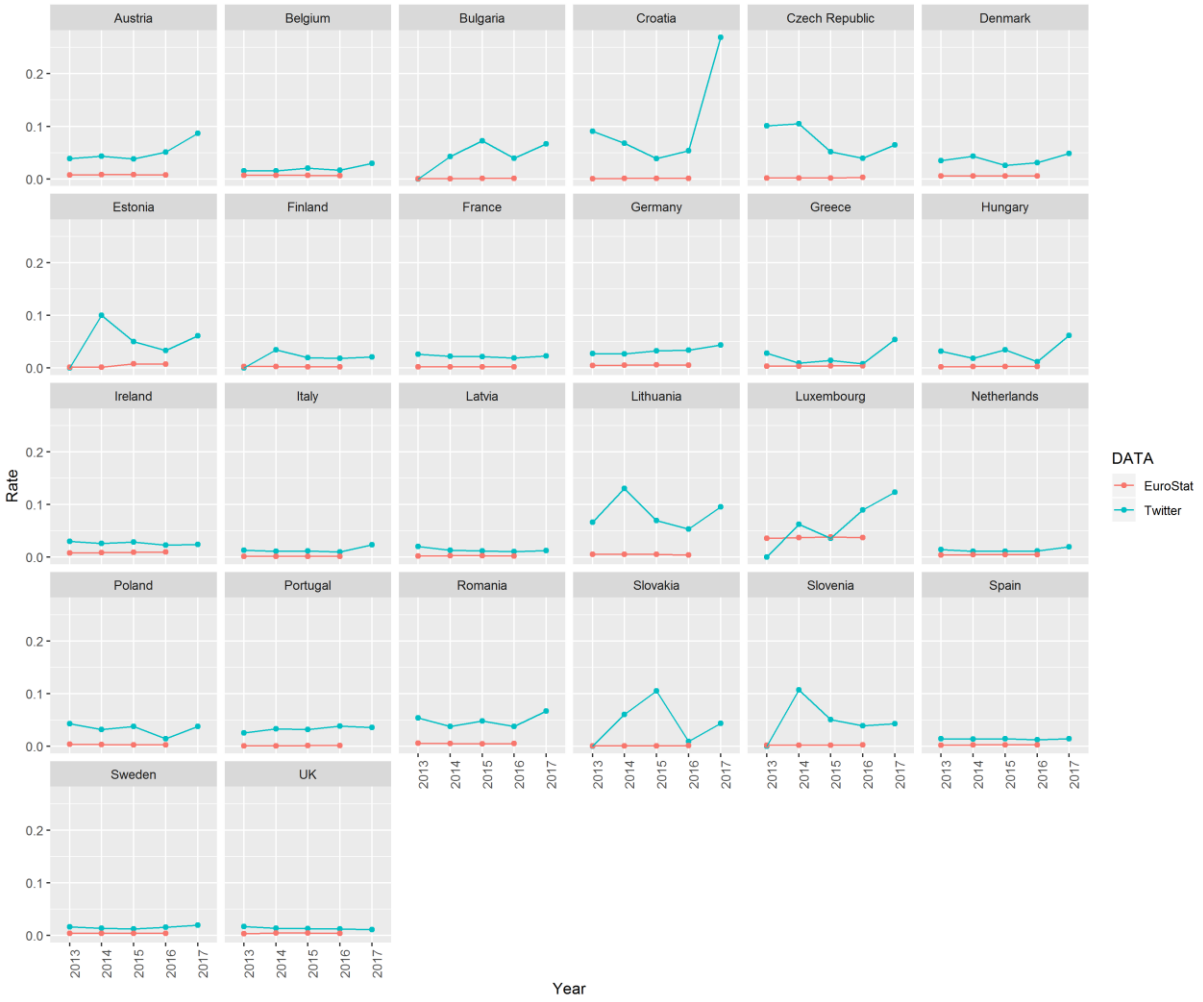
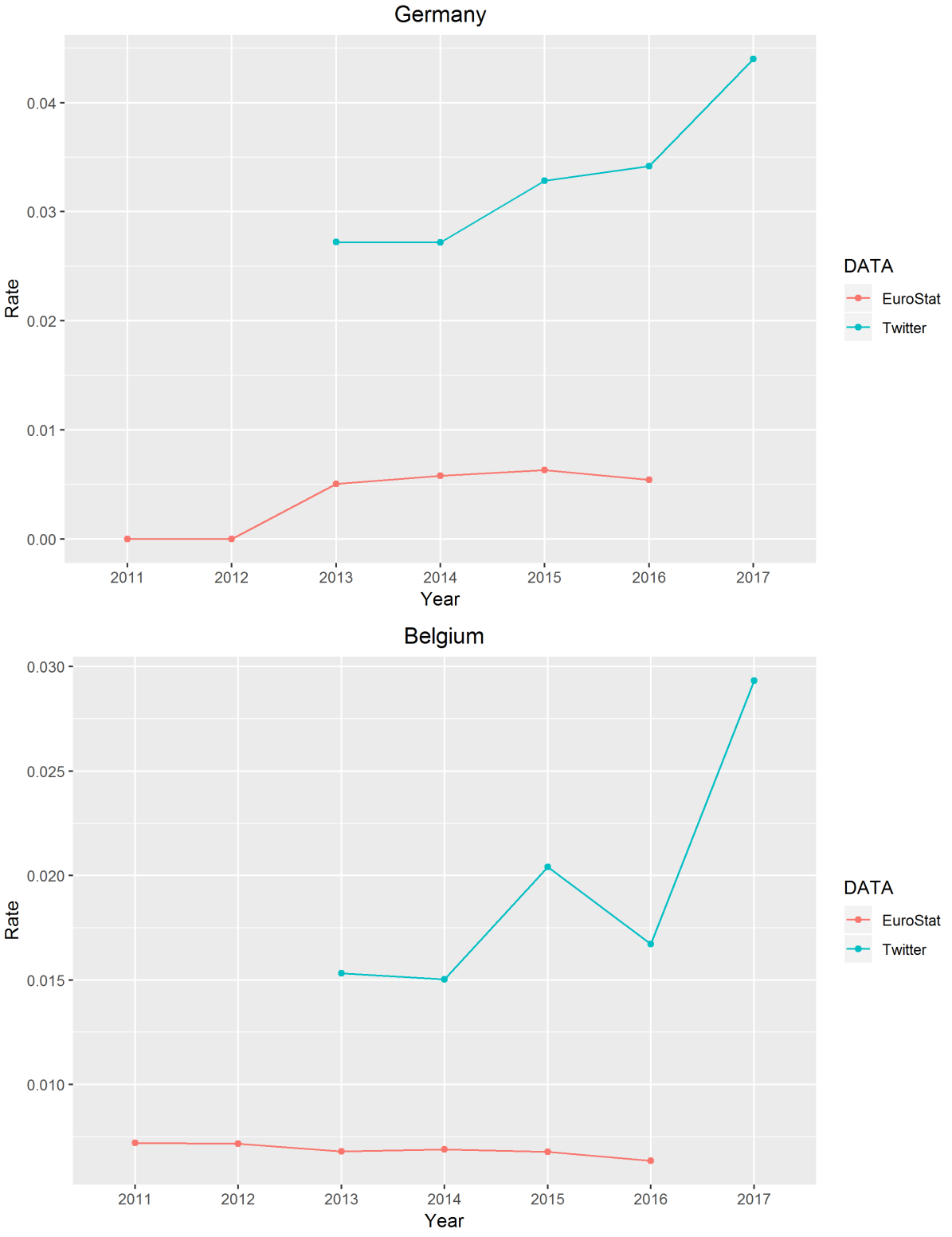


Figure 17: Comparison of Twitter and Eurostat immigration estimates in Belgium and Germany



8.2. Discussion of model results

We first test whether the joint model predicts emigration rates for year 2016 better than the "Eurostat only" model. Results indicate that the "Eurostat only" model outperforms the joint model. The RMSE for the "Eurostat only" model is 0.0012, while the RMSE for the joint model is 0.0133. The prediction accuracy is much lower for the joint model. Further inspection indicates that the "Eurostat only" model performs better for every country in the model, as shown in Figure 18. In Figure 18, the horizontal axis is the country abbreviation, while the vertical axis is the absolute value of the error, defined as

the absolute value of the discrepancy between the Eurostat estimates in the year 2016 and the values predicted by the model. There are multiple reasons that may have caused the discrepancy in the model performances, which require future work to investigate and test.

The results for immigration rates are better. The RMSE for the "Eurostat only" model is 0.0052, while the RMSE for the joint model is also 0.0020. However, as seen in the comparison of prediction errors in Figure 19, almost all of the better model performance occurs in Estonia, and otherwise the joint model does not appear to perform significantly better.

In short, the joint model that adds information from Twitter data does not appear to outperform a model that only utilises data from official statistics. There are a number of reasons why this may be the case, which require future work to investigate and test:

(1) As shown in Figure 14 and Figure 16, the variation of the Twitter estimates is very large compared to the Eurostat estimates to allow for precise estimates. The Eurostat estimate for each country remains quite stable over the years, while the Twitter estimates fluctuate a lot. Part of the reason could be that for certain country-years, the number of Twitter users is too small, resulting in substantial random errors. In such cases, it would be difficult to model the bias, as the large random errors overshadow the possibility to detect bias. To examine the latter concern, one could re-fit the model while deleting country-years that have very few Twitter users (for example, if $N < 30$). For future work, one might consider adding more Twitter data to increase the sample size and reduce random error. It would also be useful to use a measure of relative error (e.g. coefficients of variations) rather than a measure of absolute error to determine whether a similar finding is observed.

(2) As shown in Figure 14 and Figure 16, there appears to be an increase of Twitter estimates for the year 2016. Perhaps Twitter data in 2016 has some additional bias that is different from previous years. Recall that our sampling strategy was to sample from users active in Europe in the year 2016 and recollect their Twitter history. Perhaps this sampling strategy causes the representativeness of migration estimates to be different in 2016 compared to the earlier years. To test this possibility, one might re-validate the model in other years (e.g the year 2015) rather than the year 2016. For future work, if we continue to track the current pool of users, we might be able to compare Twitter estimates with Eurostat estimates for years 2017 or 2018 when the Eurostat estimates for the respective years are available. This would allow us to see whether the uptick is due to our sampling strategy.

(3) It could be that adjacency matrices do not capture the spatial relationships, and hence the spatial structure of the bias. Currently, the bias structure is set up so that adjacent countries may share a correlation in the bias, whereas in reality there may either be no such correlation or perhaps there are alternative ways to capture spatial dependency. To test this possibility, one could revalidate the model without a spatial component, and for future work explore whether there are alternative forms of spatial dependency that are compatible with the model (e.g. distance to country centroids).

(4) The model currently does not incorporate country-level covariates, which might reduce the residuals and better parse out the bias component. For future work, we might consider innovative ways to incorporate country-level covariates in the model, or try models with fixed country-level intercepts rather than random country-level intercepts. The latter strategy may effectively incorporate some country-level covariates that are constant across the years.

Figure 18: Comparison of absolute model errors on 2016 emigration rates

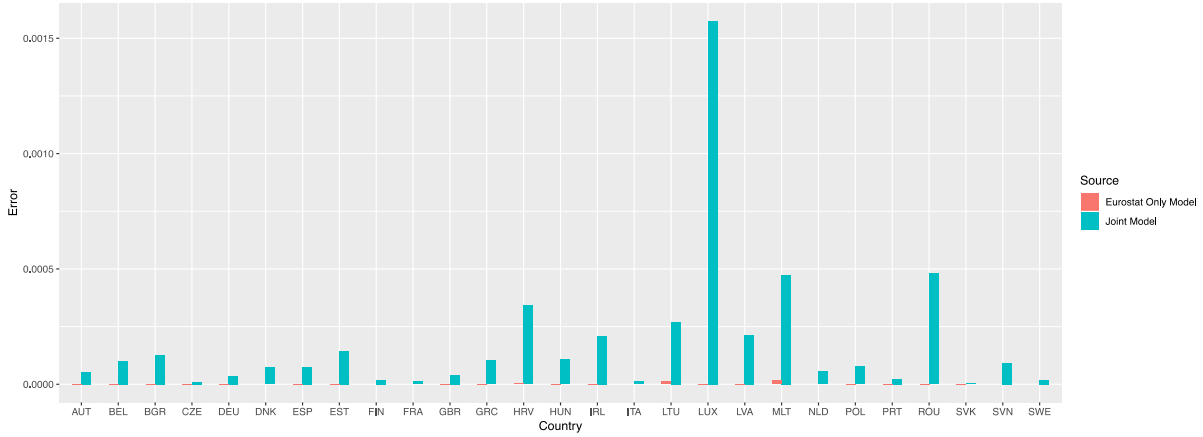
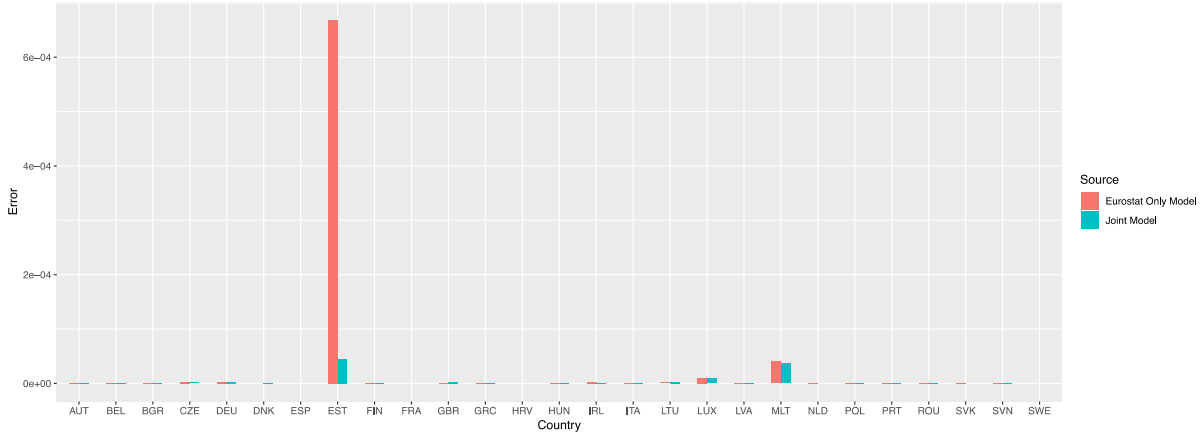


Figure 19: Comparison of absolute model errors on 2016 immigration rates



9. CONCLUDING REMARKS

The aim of this study was to investigate the potential of geo-referenced social-media data to facilitate "nowcasting" stocks of EU movers and mobility flows, providing more recent estimates than official statistics to serve as early warning signs for the European Commission. Our first results are experimental, and complementary research and data are needed to improve the robustness of the new estimates.

The first **results of the application of the stocks model are experimental, but they are promising**. Complementary research and data would be needed to improve the robustness of the new estimates. In case the European Commission wishes to continue the development of this approach, we have formulated some steps to replicate the methodology and update the estimates with more recent data.

The approach taken to estimate **EU mobility flows has not yet offered any plausible results.** We therefore do not recommend applying this approach to estimate EU mobility flows in its current form. Further research would be required to develop a robust and reliable sample of Twitter data.

This chapter offers a summary of the steps which will need to be taken in order to use the stocks model (Section 9.1), a summary of the current caveats and limitations of our approach (Section 9.2) and recommendation of areas for future research (Section 9.3).

9.1. Instruction manual on how to use the model

This section presents the steps to replicate the methodology and update the estimates with more recent data. The first step is to collect and prepare the data sources, and the second step is to update the prior distributions used in the Bayesian model.

9.1.1. Data collection

The Bayesian model harmonizes different data sources to provide more timely estimates (see Section 2.2). These data sources can be either traditional sources such as censuses, registers and surveys or social-media sources such as Facebook advertisement platform and Twitter API. The novelty of our approach is to create new estimates that incorporate strengths from each data source to overcome their weaknesses (for more information, see Section 3.5). Therefore, in addition to origin, destination, year and stock data in the form presented in Annex 1, meta-information such as survey sampling frame, sampling errors and number of active users of the social-media platform needs to be collected. This information is incorporated in the model as the parameters capturing bias and the accuracy of the data source. Next, we explain which meta-data are required for each data source for prior elicitation.

Census
Traditional national censuses aim to collect information from all usual residents of a country. Therefore, they have high coverage and accuracy and low bias. Hence, unless traditional censuses are systematically excluding a part of the population, there is no need to collect additional meta-information. For this report, we used 2011 Census data-sets downloaded from the Eurostat website.

Eurostat migrant stocks
Eurostat statistics are gathered individually from each EU member country. As mentioned in Chapter 3, despite using common definitions, various methods of classifying subpopulations are employed by different countries. Although these official statistics have high accuracy, it is reasonable to assume that different countries have different levels of coverage, meaning the migration statistics are subject to undercount or overcount when

people fail to deregister (i.e. to inform the authorities that they emigrating).[61] For this research, we divided countries into two groups: low-bias and high-bias countries. The allocation of the countries to groups has been done according to the measurement aspects table presented in Raymer et al. (2013, p. 803). However, we recommend updating the allocation into groups whenever more information is available. For example, if it is known that a country initiated a new, more reliable mechanism for collecting data on migrants, it is advisable to move that country from a high undercount to low undercount group.

LFS
LFS sampling frame, survey participation, population interviewed and the response rate differ for each country. For example, the participation is compulsory in some countries while it is voluntary in others; population in institutional households are typically excluded from the sampling frame; some countries only include students whereas others also include servicemen. Additionally, these survey designs can change over time. Therefore, it is important to collect meta-information about surveys when they are included in the model to construct the prior distributions for bias parameter. For the accuracy prior distribution, we recommend collecting sampling error of the LFS estimates for each origin, destination and year and using it as prior input, although it was not available at the time of this research.

Facebook advertisement data
The number of Facebook users and their characteristics for each country change constantly. Therefore, as a first step, it is important to collect data on number of Facebook users for each country, which will later be used in prior distributions for bias parameter.

9.1.2. Prior distributions

The second step is to update the prior distributions according to the new data collected in the first step. The prior distributions and how the countries are grouped are explained in Chapter 4. In addition, the details on prior distributions used in the model can be found in the Annex 1. In order to update the estimates with more recent Facebook data, we recommend calculating Facebook users' proportions for each country in each year. We used a ratio of the number of total Facebook users in a country to the Eurostat population estimate for that country in the corresponding age group. Then, we grouped them according to low and high penetration. In the short term, we do not expect that significant changes for prior distributions will be required. One exception for this could be to include sampling errors for LFS data – instead of estimating them with a regression model – as another level within the hierarchical Bayesian model.

9.1.3. Additional information required in the model

In addition to the data illustrated in Annex 1 and the prior information, the only information required to run the model are the indices reflecting available (i.e. not NA) data for each source and for each prior group when applicable. This information needs to be saved as a vector using an R command. For example, for Eurostat, rows in which the origin, destination, year and migrant stocks are available need to be saved in this vector. The allocation of countries into bias groups to assign prior distributions is also required.

The last step to produce the estimates is to update the model using additional data and the prior distributions outlined in Section 9.1.2. As mentioned before, we do not envisage significant changes in migrant stock patterns in the short term. Practically, the only relevant input required is to update the prior distribution for parameters capturing bias in Facebook data.

[61] In the model we simplify the notation by denoting under-, over-count and coverage together as "bias".

9.2. Caveats and limitations

Bayesian inferential framework combines prior beliefs with the observed data. The prior beliefs are included in the model as prior distributions. Estimates from the model are then weighted depending on the amount of data and how strong the prior beliefs incorporated in prior distributions are. Hence, in the cases with large portions of data missing, the estimates are more strongly affected by the prior beliefs; and vice versa, if the lengthy time-series of data are available for most of the countries, the role of the priors will diminish. In our research, the crucial point to produce good estimates is to have a long time-series dataset in which Facebook data and official data overlap. More overlap between Facebook data and official data would decrease the impact of prior distributions on the estimated true stock of EU movers, and therefore increase the robustness of our results. We believe that the lack of direct comparison through overlapping series of the official Eurostat and Facebook data is the main limitation of the current results. This may result in sensitivity to the particular assumptions about the prior distributions about bias and relative accuracy of the data sources. However, we envisage that this limitation can be tackled by incorporating Facebook data collected over a longer time period and, in general, more data. Furthermore, we addressed this issue by introducing to the model the data from the Labour Force Survey, which provides harmonised measures of migrant stocks for more recent years and all countries under study. Finally, any future changes in the representativeness of these datasets, for example due to declining popularity of Facebook, will affect the reliability of the model.

9.3. Recommendations for improving the approach and estimates of stocks of EU movers

In this section we present recommendations for future work to improve the proposed method and estimates of stocks of EU movers. The recommendations relate to the amount of data available as well as the method itself.

Investigate different migration models. As mentioned before, the migration model that aims at estimating true flows where the data are not available is an important aspect of our modelling framework. For this research, we employed a non-theoretical perspective that permits forecasting and borrowing strength across time and countries. Further work would be required to assess specific theoretical models, such as a gravity model that relies on sending and receiving countries' populations and distance between them, or other so-called pull-and-push factors (cf. Raymer et al. 2013).

Longer time series from Facebook. The Bayesian modelling framework aims at harmonising traditional data sources with social-media data. The estimates would benefit from a longer time-series of Facebook advertisement platform data that overlaps with traditional data sources. This way, the model will lead to more shrinkage in the posterior distributions for bias and accuracy parameters and, thus, lead to improved estimates of stocks of EU movers. Therefore, if the European Commission were to consider pursuing further research based on Facebook data, we would recommend the continuation of the regular data collection from the Facebook Marketing API in accordance with the approach specified in Chapter 3.

Longer and more time-series LFS data. The Labour Force Surveys have been conducted for many years. However, in this project we have only been able to use data for 2016 and 2017. Further disaggregation by age and gender was not feasible due to small sample sizes and statistical disclosure issues. We argue that using the LFS data can provide a better picture of the patterns in changing stocks of EU movers, especially when the official administrative data are not available through Eurostat. Incorporating the LFS data in the model for all countries allows benchmarking against census and official Eurostat data, which can help in estimating the completely missing cells, bearing in mind the typical caveats of using survey data related to sampling and non-sampling errors. Further, we believe that the current approach would benefit from using the LFS sampling errors to inform prior distributions on accuracy parameters in the model. Unfortunately, such data were not available for this study.

9.4. Recommendations for further developing the approach to estimate mobility flows

The approach taken to estimate **EU mobility flows has not yet offered any plausible results.** The added value of Twitter data for the purpose of measuring flows is limited. We therefore do not recommend applying this approach to estimate EU mobility flows in its current form. When Twitter data are combined with Eurostat data, the flows model does outperform a model based on Eurostat data exclusively. Further research would be required to develop a robust and reliable sample of Twitter data, and to understand the reasons behind these implausible results.

In particular, we identify the following avenues for future research:

- The relatively low coverage of the Twitter data, particularly when compared with Facebook, may explain the implausible results. For certain countries and years, our Twitter dataset contains too few Twitter users, making the Twitter estimate of flows unreliable. A larger Twitter dataset, for example by continuing the approach described in Chapter 7, would likely reduce the random error.
- As our Twitter data for years 2012 to 2015 and year 2016 result from two different sampling strategies, this might have caused the representativeness of migration estimates to be different in the two samples, which is likely to have affected our results. We would advise to keep downloading data going forward and to add Eurostat estimates for years 2017 and 2018 once they are available, to test this hypothesis.
- We observe a high RMSE for the joint model, perhaps due to the small number of Twitter users observed. It would be useful to try a measure of relative error (e.g. coefficients of variations) rather than a measure of absolute error to determine whether a similar finding is observed.
- Our model currently uses adjacency matrices to capture the spatial relationships and the spatial structure of the bias. It is possible that such matrices are not accurately representing the spatial dependencies, and we would recommend testing alternative forms of spatial dependencies to determine whether some would perform better than others.
- Some adjustments to the model could also return better results. For example, one might want to try a model with fixed country-level intercepts rather than random-effect intercepts, or to incorporate country-level covariates.
- Further, while we have used Eurostat data, the model could also be tested using a different source of data, such as data from the Labour Force Survey, using the question about the country of residence of the respondent in the year prior to the survey.

Notwithstanding the required improvements in the data set, it is important to note that Twitter imposes restrictions to developers with regards to sharing data with Third Parties (see section F.2 of the Developers Policy in Annex 4). Only tweet IDs, direct messages IDs and user IDs can be shared, in the context of academic research, and respecting a number of rules with regards to the quantity of such objects that can be shared in a given time period. In addition, the rehydration of the dataset will require some delay itself. Therefore, if the European Commission were to consider pursuing further research based on Twitter data, we would recommend starting to build a data set in accordance with the approach specified in this report to build its own sample.

REFERENCES

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

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

Abel, G.J., & N. Sander. (2014). "Quantifying global international migration flows." *Science*, 343(6178), 1520–1522.

Andrew, A.H., K. Eustice, & A. Hickl. (2013). "Using location lifelogs to make meaning of food and physical activity behaviors." In *Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, 2013, 7th International Conference on, 408–411. IEEE.

Araujo, M., Mejova, Y., Weber, I., & Benevenuto, F. (2017). "Using Facebook ads audiences for global lifestyle disease surveillance: Promises and limitations." In Proceedings of the 2017 ACM on Web Science Conference (pp. 253-257). ACM.

Araujo, M., Y. Mejova, M. Aupetit, & I. Weber. (2018) "Visualizing Geo-Demographic Urban Data." Companion of the ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW), 2018. As of 15 April 2019: https://dl.acm.org/citation.cfm?id=3273001

Azose, J.J., & A.E. Raftery. (2013). "Bayesian probabilistic projection of international migration rates." arXiv preprint arXiv:1310.7148.

Barslund, M., & M. Busse. 2016. "Labour Mobility in the EU: Addressing challenges and ensuring 'fair mobility'". CEPS Special report 139, July 2016. As of 15 April 2019: https://www.ceps.eu/system/files/SR139%20MB%20and%20MB%20LabourMobility.pdf

Bayir, M.A., M. Demirbas, & N. Eagle. (2009) "Discovering spatiotemporal mobility profiles of cellphone users." World of Wireless, Mobile and Multimedia Networks & Workshops, 2009. WoWMoM 2009. IEEE International Symposium on a. 1–9. IEEE.

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

Bijak, J., & A. Wisniowski. (2010). "Bayesian forecasting of immigration to selected European countries by using expert knowledge." *Journal of the Royal Statistical Society*: Series A (Statistics in Society), 173(4), 775–796.

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

Blanford, J.I., Z. Huang, A. Savelyev, & A.M. MacEachren. (2015). "Geo-located tweets. Enhancing mobility maps and capturing cross-border movement." *PloS one*, 10(6), e0129202.

Blumenstock, J.E. (2012). "Using mobile phone data to measure the ties between nations." In *Proceedings of the 2011 iConference*, iConference '11, 195–202, New York, NY, USA, 2011. ACM.

Brillinger, D.R. "John W. Tukey: his life and professional contributions." *The Annals of Statistics* 30.6 (2002): 1535–1575.

Brown J. (2001). "Design of a census coverage survey and its use in the estimation and adjustment of census underenumeration: a contribution towards creating a one-number census in the UK in 2001". PhD Thesis, University of Southampton

Brown, D.M., & A. Soto-Corominas. (2017). "Overview–The Social Media Data Processing Pipeline." *The SAGE Handbook of Social Media Research Methods*, 125.

Buchel, O., & D. Pennington. (2017). "Geospatial analysis." *The SAGE Handbook of Social Media Research Methods*, 285.

Candia J., M.C. González, P. Wang, T. Schoenharl, G. Madey, & A.-L. Barabási. (2008). "Uncovering individual and collective human dynamics from mobile phone records." *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015.

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

Chen, S. and Tang, C. (2011). "Properties of census dual system population size estimators". International Statistical Review, 79(3):336-361.

Chen, H., R.H.L. Chiang & V.C. Storey (2012) "Business intelligence and analytics: From big data to big impact". *MIS Quarterly*, 36 (4): 1165-1188

Chen, M., S. Mao, Y. Zhang, & V.C. Leung. (2014). "Big data analysis". In *Big Data*, 51–58). Springer International Publishing.

Choudhury, M. De, M. Feldman, S. Amer-Yahia, N. Golbandi, R. Lempel, and C. Yu. (2010). "Automatic construction of travel itineraries using social breadcrumbs." In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, 35–44. ACM.

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

ComRes on behalf of BBC Newsround. (2017). "A survey of 10–12 year olds who use social media on behalf of BBC Newsround." As of 6 February 2019: https://www.comresglobal.com/polls/bbc-newsround-safer-internet-day-2017/

De Beer, J., Raymer, J., Van der Erf, R., and Van Wissen, L. (2010). "Overcoming the problems of inconsistent international migration data: A new method applied to flows in europe." European Journal of Population/Revue européenne de Démographie, 26(4):459–481.

Disney, G., A. Wiśniowski, J.J. Forster, P.W. Smith, & J. Bijak. (2015). "Evaluation of existing migration forecasting methods and models." Report for the Migration Advisory Committee: Commissioned research. ESRC Centre for Population Change, University of Southampton.

Doha Demographics. (2017). "Demographic Distribution in Doha." As of 17 April 2019: http://fb-doha.qcri.org/

Eurofound. (2019). "Migration and mobility." As of 22 February 2019: https://www.eurofound.europa.eu/topic/migration-and-mobility

European Commission. (2019a). "EU Immigration Portal: Glossary". As of 17 April 2019: https://ec.europa.eu/immigration/content/glossary_en

European Commission. (2019b). "Migration and Home Affairs: migrant worker". As of 17 April 2019: https://ec.europa.eu/home-affairs/content/migrant-worker-0_en

Eurostat. (2011). "EU legislation on the 2011 Population and Housing Censuses: Explanatory Notes." As of 10 December 2018: https://ec.europa.eu/eurostat/documents/3859598/5916677/KS-RA-11-006-EN.PDF/5bec0655-4a55-466d-9a00-fabe83d54649?version=1.0

Eurostat. (2014). "2014. Migration and labour market (lfso_14)" As of 17 April 2019: http://ec.europa.eu/eurostat/cache/metadata/en/lfso_14_esms.htm

Eurostat. (2015a). "Demographic statistics: A review of definitions and methods of collection in 44 European countries 2015 edition."

Eurostat. (2015b). "People in the EU: Who are we and how do we live?" As of 10 December 2018: https://ec.europa.eu/eurostat/documents/3217494/7089681/KS-04-15-567-EN-N.pdf/8b2459fe-0e4e-4bb7-bca7-7522999c3bfd

Eurostat. (2016)a. "Labour Force Survey in the EU, candidate and EFTA countries. Main characteristics of national surveys, 2015 – 2016 Edition." As of 15 April 2019: http://ec.europa.eu/eurostat/documents/3888793/7751652/KS-TC-16-021-EN-N.pdf/8475c2e2-c037-4ba2-9029-93db1ade41fe

Eurostat. (2016)b. "Population (demo_pop): Reference Metadata in Euro SDMX Metadata Structure (ESMS)." As of 28 February 2019: https://ec.europa.eu/eurostat/cache/metadata/en/demo_pop_esms.htm

Eurostat. (2017). "Quality report of the European Union Labour Force Survey 2015." ec.europa.eu. As of 12 July 2018: http://ec.europa.eu/eurostat/documents/7870049/7887033/KS-FT-17-003-EN-N.pdf/22ed8f4e-9eb3-455c-924a-8df102620f89

Eurostat. (2018a). "Quality report of the European Union Labour Force Survey 2016" ec.europa.eu. As of 28 November 2018: https://ec.europa.eu/eurostat/documents/7870049/9350257/KS-FT-18-008-EN-N.pdf/d547620d-33fc-426b-8946-30b5a634fbda

Eurostat. (2018b). "2011 Census Hub." As of 11 December 2018: https://ec.europa.eu/eurostat/web/population-and-housing-census/census-data/2011-census

Eurostat. (2018c). "Labour Force Survey in the EU, candidate and EFTA countries. Main characteristics of national surveys, 2016 – 2018 edition." As of 15 April 2019: https://ec.europa.eu/eurostat/documents/7870049/8699580/KS-TF-18-002-EN-N.pdf/ce2e7a97-6b8c-44b8-8603-3a4606e5b335

Eurostat. (2019a). "Migration and migrant population statistics - Statistics Explained." As of 7 March 2019: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Main_Page

Eurostat. (2019b). "Employment and unemployment (Labour force survey) (employ): Reference Metadata in Euro SDMX Metadata Structure (ESMS)." As of 28 February 2019: https://ec.europa.eu/eurostat/cache/metadata/en/employ_esms.htm

Eurostat. (2019c). "International Migration Statistics." As of 17 April 2019: http://ec.europa.eu/eurostat/cache/metadata/en/migr_immi_esms.htm

Eurostat. (2019d). "EU labour force survey – methodology". As of 17 April 2019: http://ec.europa.eu/eurostat/statistics-explained/index.php/EU_labour_force_survey_-_methodology

Eurostat. (2019e). "Population (Demography, Migration and Projections), Table migr_imm4ctb". As of 17 April 2019: http://ec.europa.eu/eurostat/web/population-demography-migration-projections/migration-and-citizenship-data/database

Eurostat. (2019f). "Database". As of 17 April 2019: https://ec.europa.eu/eurostat/data/database

Eurostat. (2019g). "Employment and Unemployment (LFS): Database." As of 17 April 2019: https://ec.europa.eu/eurostat/web/lfs/data/database

Eurostat. (2019h). "Access to Microdata: Overview." As of 17 April 2019: https://ec.europa.eu/eurostat/web/microdata/overview

Eurostat (2019i). "Employment and unemployment (Labour force survey) (employ)." As of 17 April 2019: https://ec.europa.eu/eurostat/cache/metadata/en/employ_esms.htm

Eurostat. (2019j). "EU labour force survey – main features and legal basis." As of 17 April 2019: http://ec.europa.eu/eurostat/statistics-explained/index.php?title=EU_labour_force_survey_%E2%80%93_main_features_and_legal_basis#Legal_basis

Eurostat. (2019k). "European Union labour force survey (EU LFS)." As of 17 April 2019: https://ec.europa.eu/eurostat/web/microdata/european-union-labour-force-survey

Facebook Ads Manager. "Advertiser help – About location targeting". Facebook.com. As of 15 February 2018: https://www.facebook.com/ads/manager/creation/creation/

Facebook Business. (2019a) "About potential reach". Facebook.com. As of 17 April: https://www.facebook.com/business/help/1665333080167380

Facebook Business. (2019b). "About estimated daily results". Facebook.com. As of 17 April: https://www.facebook.com/business/help/1438142206453359

Facebook for Developers. (2019). "Ad Campaign Delivery Estimate". Facebook.com. As of 17 April: https://developers.facebook.com/docs/marketing-api/reference/ad-campaign-delivery-estimate/

Faris, D. M. (2013). *Dissent and revolution in a digital age: Social media, blogging and activism in Egypt.* London/New York: IB Tauris.

Fatehkia, M., D. O'Brien, & I. Weber. (2019). "Correlated impulses: Using Facebook interests to improve predictions of crime rates in urban areas." *PloS one*, 14(2), e0211350.

Ferrari L., & M. Mamei. (2011) "Discovering daily routines from google latitude with topic models." In *Pervasive Computing and Communications Workshops (PERCOM Workshops)*, 2011, IEEE International Conference on, 432–437. IEEE, 2011.

Ferrari L., A. Rosi, M. Mamei, & F. Zambonelli. (2011) "Extracting urban patterns from location-based social networks." In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, 9–16. ACM, 2011.

Fiorio, L., G. Abel, J. Cai, E. Zagheni, I. Weber, & G. Vinué. (2017). "Using Twitter data to estimate the relationship between short-term mobility and long-term migration." In *Proceedings of the 2017 ACM on Web Science Conference*, 103–110. ACM.

Fries-Tersch, E., T. Tugran, & H. Bradley. (2017). "2016 Annual report on intra-EU Labour Mobility". The European commission. Second edition May 2017.

Fries-Tersch, E., T. Tugran, L. Rossi, & H. Bradley. (2018) "2017 Annual Report on intra-EU Labour Mobility." Prepared for European Commission, DG EMP. Second edition September 2018. As of 21 January 2019: http://publications.europa.eu/publication/catalogue_number/KE-BQ-18-101-EN-N

Gelman, A., J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, & D. B. Rubin. (2014). "Bayesian data analysis (Vol. 2)." Boca Raton, FL: CRC press.

GitHub, Inc. (2019). "Social Watcher on Facebook Marketing API." As of 17 April: https://github.com/maraujo/pySocialWatcher

Gonzalez, M.C., C.A. Hidalgo, & A.-L. Barabasi. (2008). "Understanding individual human mobility patterns." *Nature*, 453(7196): 779–782, 2008.

Grimmelmann, J. (2015). "The law and ethics of experiments on social media users." 13 Colo. Tech. L.J. 219, 2015; U of Maryland Legal Studies Research Paper NO.2015-15. As of 15 April 2019: https://ssrn.com/abstract=2604168

Grinberg, N., M. Naaman, B. Shaw, & G. Lotan. (2013). Extracting diurnal patterns of real world activity from social media. In *ICWSM*.

Haustein, S., V. Larivière, M. Thelwall, D. Amyot, & I. Peters. (2014). "Tweets vs. Mendeley readers: How do these two social media metrics differ?" *IT-Information Technology*, 56(5), 207–215.

Hawelka, B., I. Sitko, E. Beinat, S. Sobolevsky, & P.K.C. Ratti. (2014). "Geo-located twitter as proxy for global mobility patterns," *Cartography and Geographic Information Science* 41: 260–271.

Herdagdelen, A., L. Adamic, W. Mason, et al. (2016). "The social ties of immigrant communities in the United States". Proceedings of the 8th ACM Conference on Web Sciences, 78–84. ACM.

Hofleitner, A., T.V. Chiraphadhanakul, & B. State. (2013). "Coordinated migration." Facebook Data Science Team.

Holland, D., T. Fic, A. Rincon-Aznar, L. Stokes, & P. Paluchowski. (2011). "Labour mobility within the EU – The impact of enlargement and the functioning of the transitional arrangements." Study commissioned by the Employment, Social Affairs and Inclusion Directorate General of the European Commission.

Hughes, C., E. Zagheni, G.J. Abel, A. Wisniowski, A. Sorichetta, I. Weber, & A.J. Tatem. (2016). "Inferring Migrations: Traditional Methods and New Approaches based on Mobile Phone, Social Media, and other Big Data." European Commission project #VT/2014/093, February 2016.

Hui, P., R. Mortier, M. Piorkowski, T. Henderson, & J. Crowcroft. (2012). "Planet-scale human mobility measurement." doi:10.1145/1834616.1834618.

Jiang, B., & Y. Miao. (2015). "The evolution of natural cities from the perspective of location-based social media." *The Professional Geographer*, 67(2), 295–306.

Kingdom of Belgium Foreign Affairs, Foreign Trade and Development Cooperation. 2018. 'Nationality.' As of 18 January 2019: https://diplomatie.belgium.be/en/services/services_abroad/nationality

Laney, D. (2001). "3D data management: Controlling data volume, velocity, and variety." META Group. As of 21 May 2019: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Lanzieri, G. (2014). "Population bases - Draft text for the Conference of European Statisticians Recommendations for the 2020 census round."

Lathia, N., D. Quercia, & J. Crowcroft. (2012). "The hidden image of the city: sensing community well-being from urban mobility." In *Pervasive Computing*, 91–98.

Lenormand, M., A. Tugores, P. Colet, J.J. Ramasco. (2014). "Tweets on the road." *PloS one* 9.8 (2014): e105407.

Masse, M. (2011). "REST API Design Rulebook: Designing Consistent RESTful Web Service Interfaces". O'Reilly Media, Inc.

Mercer, L., Wakefield, J., Pantazis, A., Lutambi, A., Masanja, H., and Clark, S. (2015). "Small area estimation of child mortality in the absence of vital registration." The Annals of Applied Statistics 9(4): 1889-1905.

Ministry of the Interior of the Czech Republic. (2017). "Citizenship of the Czech Republic." As of 18 January 2019: https://www.mvcr.cz/mvcren/article/citizenship-of-the-czech-republic.aspx

Mislove, A., S. Lehmann, Y.Y. Ahn, J.P. Onnela, & J.N. Rosenquist. (2011). "Understanding the Demographics of Twitter Users." *Proceedings of the Fifth International AAAI Conference on Web and Social Media*: 554–557.

Neubauer, G., H. Huber, A. Vogl, B. Jager, A. Preinerstorfer, S. Schirnhofer, G. Schimak, & D. Havlik. (2015). „On the volume of geo-referenced tweets and their relationship to events relevant for migration tracking." In *Environmental Software Systems. Infrastructures, Services and Applications*, 520–530. Springer.

Noulas, A., S. Scellato, C. Mascolo, & M. Pontil. (2011). "An empirical study of geographic user activity patterns in foursquare." *ICWSM*, 11:70–573.

Nowok, B., Kupiszewska, D., & Poulain, M. (2006). "Statistics on international migration flows." THESIM: Towards harmonised European statistics on international migration, 203-232.

ONS (2012). "Beyond 2011: Exploring the Challenges of Using Administrative Data available". https://www.ons.gov.uk/census/censustransformationprogramme/beyond2011censustransformationprogramme/reportsandpublications

Ortega, F., & G. Peri. (2013). "The effect of income and immigration policies on international migration." *Migration Studies*, 1(1), 47–74.

Perrin, A. (2015). "Social Networking Usage: 2005–2015." Pew Research Center. As of 15 April 2019: http://www.pewinternet.org/2015/10/08/2015/Social-Networking-Usage-2005-2015/

Pitsillidis A., Y. Xie, F. Yu, M. Abadi, G.M. Voelker, & S. Savage. (2010). "How to tell an airport from a home: Techniques and applications." In *Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks*, p13. ACM.

Plummer, M. (2003). "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling." In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, p125. Vienna, Austria.

Pötzschke, S., & M. Braun. (2016). "Migrant sampling using Facebook advertisements: A case study of Polish migrants in four European countries." *Social Science Computer Review* 35(5): 633–653.

Pultar, E. & M. Raubal. (2009). "A case for space: physical and virtual location requirements in the couchsurfing social network." In *Proceedings of the 2009 International Workshop on Location Based Social Networks*, 88–91. ACM.

Raftery, A.E., N. Li, H. Ševčíková, P. Gerland, & G.K. Heilig. (2012). "Bayesian probabilistic population projections for all countries." *Proceedings of the National Academy of Sciences*, 109(35), 13915–13921.

Raymer, J., J. de Beer, & R. van der Erf. (2011). "Putting the pieces of the puzzle together: Age and sex-specific estimates of migration amongst countries in the EU/EFTA, 2002–2007." *European Journal of Population/Revue européenne de Démographie*, 27(2), 185–215.

Raymer, J., A. Wiśniowski, J.J. Forster, P.W., Smith, & J. Bijak. (2013). "Integrated modeling of European migration." *Journal of the American Statistical Association*, 108(503), 801–819.

Regulation (EC) No.862/2007 of the European Parliament and of the Council of 11 July 2007 on Community statistics on migration and international protection, and repealing Council Regulation (EEC) No 311/76 on the compilation of statistics on foreign workers.

Regulation (EU) No.1260/2013 of the European Parliament and of the Council of 20 November 2013 on European Demographic Statistics. As of 17 April 2019: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32013R1260&from=EN

Rodrigue, J.-P., C. Comtois, & B. Slack. (2009). *The Geography of Transport Systems*. London/New York: Routledge.

Simini, F., M.C. González, A. Maritan, & A.-L. Barabási. (2012). "A universal model for mobility and migration patterns." *Nature*, 484(7392):96–100.

Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P., & Rana, O. (2013). "Knowing the Tweeters: Deriving sociologically relevant demographics from Twitter." Sociological research online, 18(3), 1-11.

Sloan, L., & Morgan, J. (2015). "Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter." PloS one, 10(11), e0142209.

Sloan, L., J. Morgan, P. Burnap, & M. Williams. (2015). "Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data." *PloS one*, 10(3): e0115545.

Smith, C., D. Quercia, & L. Capra. (2013) "Finger on the pulse: identifying deprivation using transit flow analysis." In *Proceedings of the 2013 conference on Computer supported cooperative work*, 683–692. ACM.

Spyratos, S., M. Vespe, F. Natale, I. Weber, E. Zagnei, & M. Rango. (2018). "Migration Data using Social Media: a European Perspective." EUR 29273 EN, Publications Office of the European Union, Luxembourg. ISBN 978-92-79-87989-0, doi:10.2760/964282, JRC112310.

Stackoverflow. (2019). "How to differentiate Personal vs Corporate accounts in Twitter API?" As of 17 April 2019: https://stackoverflow.com/questions/29761227/how-to-differentiate-personal-vs-corporate-accounts-in-twitter-api#

State, B., I. Weber, & E. Zagheni. (2013). "Studying inter-national mobility through IP geolocation." In *WSDM*, 265–274.

Statista. (2019). "Number of monthly active Twitter users worldwide from 1st quarter 2010 to 4th quarter 2018 (in millions)." As of 17 April 2019: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

Stefanidis, A., A. Cotnoir, A. Croitoru, A. Crooks, M. Rice, & J. Radzikowski. (2013). "Demarcating new boundaries: mapping virtual polycentric communities through social media content." *Cartography and Geographic Information Science*, 40(2), 116–129.

Tasse, D., Z. Liu, A. Sciuto, & J.I. Hong. (2017). "State of the Geotags: Motivations and Recent Changes." In *ICWSM*, 250-259.

The Migration Observatory. (11 January 2017). "Who counts as a migrant? Definitions and their consequences". As of 15 February 2018: http://www.migrationobservatory.ox.ac.uk/resources/briefings/who-counts-as-a-migrant-definitions-and-their-consequences/

Thorson, K., K. Driscoll, B. Ekdale, S.L. Edgerly, L.G. Thompson, A. Schrock, L. Swartz, E.K. Vraga, & C. Wells. (2013). "YouTube, Twitter and the Occupy movement: Connecting content and circulation practices." *Information, Communication & Society*, 16(3), 421–451.

Tobin, J. (1958). "Estimation of relationships for limited dependent variables." *Econometrica* 26, no.1: 24–36. doi:10.2307/1907382.

Twitter, Inc. (3 November 2017). "Developer Policy." As of 17 April 2019: https://developer.twitter.com/en/developer-terms/policy

Twitter, Inc. (2018a). "Developer Agreement." As of 17 April 2019: https://developer.twitter.com/en/developer-terms/agreement

Twitter, Inc. (2018b). "How Twitter is fighting spam and malicious automation." As of 17 April 2019: https://blog.twitter.com/en_us/topics/company/2018/how-twitter-is-fighting-spam-and-malicious-automation.html

Twitter, Inc. (2019a). "Docs." As of 17 April 2019: https://dev.twitter.com/streaming/overview

Twitter, Inc. (2019b). "Business." As of 17 April 2019: https://business.twitter.com/en/a/get-started.html

UNDESA. (2017). "Trends in International Migrant Stock: The 2017 Revision." As of 17 January 2019: http://www.un.org/en/development/desa/population/migration/data/estimates2/docs/MigrationStockDocumentation_2017.pdf

UNECE. (2014). "Measuring Population and Housing: Practices of UNECE countries in the 2010 round of censuses." New York/Geneva. As of 13 December 2018: https://www.unece.org/fileadmin/DAM/stats/publications/2013/Measuring_population_and_housing_2010.pdf

UNECE. (2006). "Conference of European Statisticians Recommendations for the 2010 Censuses of Population and Housing' As of 11 December 2018: https://www.unece.org/fileadmin/DAM/stats/publications/CES_2010_Census_Recommendations_English.pdf

UN (2017) "Handbook on Measuring International Migration through Population Censuses." Prepared by the Secretariat, Statistical Commission Background document. (1 March 2017). As of 21 December 2018: https://unstats.un.org/unsd/demographic-social/Standards-and-Methods/files/Handbooks/international-migration/2017-draft-E.pdf

US Securities and Exchange Commission. (2017). "Filings & Forms." As of 17 April 2019: https://www.sec.gov/edgar.shtml

US Securities and Exchange Commission. (2019). "Annual Report, Facebook, Inc." As of 17 April 2019: http://d18rn0p25nwr6d.cloudfront.net/CIK-0001326801/a109a501-ed16-4962-a3af-9cd16521806a.pdf

Willekens, F. (2019). "Evidence-Based Monitoring of International Migration Flows in Europe." Journal of Official Statistics, 35(1), 231–277. doi: https://doi.org/10.2478/jos-2019-0011

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

Wisniowski, A. (2017). "Combining Labour Force Survey data to estimate migration flows: the case of migration from Poland to the UK." *Journal of the Royal Statistical Society*: Series A (Statistics in Society), 180(1), 185–202.

Wisniowski, A., J. Bijak, S. Christiansen, J.J. Forster, N. Keilman, J. Raymer, & P.W. Smith. (2013). "Utilising expert opinion to improve the measurement of international migration in Europe." *Journal of Official Statistics*, 29(4), 583–607.

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

Yildiz, D., J. Munson, A. Vitali, R. Tinati, & J. Holland. (2017) "Using Twitter data for demographic research." *Demographic Research* (37):1477–1514.

Zagheni, E., & I. Weber. (2012) "You are where you e-mail: using e-mail data to estimate international migration rates." In *WebSci*, 348–351.

Zagheni, E., & I. Weber. (2015). "Demographic research with non-representative internet data." *International Journal of Manpower*, 36(1), 13–25.

Zagheni, E., V.R.K. Garimella, I. Weber, & B. State. (2014). "Inferring international and internal migration patterns from twitter data." In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, 439–444. International World Wide Web Conferences Steering Committee.

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

Zipf, G.K. (1946). "The p1 P2/D hypothesis: On the intercity movement of persons." *American Sociological Review*, 11(6):677–686.

ANNEXES

Annex 1: Stock Model details

The corridor and year index, i, is structured as follows:

i	Origin	Destination	Year
1	Austria	Belgium	2011
2	Austria	Bulgaria	2011
28	Belgium	Austria	2011
29	Belgium	Bulgaria	2011
756	United Kingdom	Sweden	2011
757	Austria	Belgium	2012
758	Austria	Bulgaria	2012
6048	United Kingdom	Sweden	2018

n: number of corridors, n = 756 = 28 x 27

y: Logarithm of true stocks i.e. migrant stock estimate

z_i^k: The reported migrant stocks in data source k in logarithmic scale

Precision: Inverse of variance $(1/\sigma)$

Migration model relies on a simple autoregressive process where the migrant stocks in a in a current year depend on the migrant stocks in the previous year. Therefore, we model the first year, 2011, separately than the rest of the years.
For 2011, we assume that the logarithm of the migrant stock for each corridor, y1[i], is distributed normally mean $\beta[1,i]$ and precision τ_{y1} as follows:
For i = 1,2,…, 756:

$$y1[i] \sim \text{Normal}\left(\beta[1,i], \tau_{y1}\right).$$

For years between 2012 and 2018, we similarly assume a normal distribution with mean $(\beta[1,i] + \beta[2,i] \text{ x } y1[i-n])$ and the same precision parameter, τ_{y1}. . The difference between 2011 model and this one is that the mean of migrant stock at year t +1 depends on the migrant stock estimate at year t.

For i = 757, 758,…, 6048 and n = 756:

$$y1[i] \sim \text{Normal}\left(\beta[1,i] + \beta[2,i] \text{ x } y1[i-n], \tau_{y1}\right).$$

Measurement models are used to harmonise each data source by taking their bias (γ) and accuracy (e) into account.

- <u>2011 Census:</u>

$$z_i^{Census} \sim \text{Normal}\left(y_i + \log\left(\gamma^{Census}\right), e^{Census}\right)$$

γ^{Census}: Bias parameter for 2011 Census, not corridor specific.

We assume all the census migrant stocks distributed normally with mean one (assuming no bias on average) and precision as follows:

$$\gamma^{Census} \sim N(1, 100) \text{ and}$$
$$e^{Census} \sim \text{Normal}(5000, 0.01).$$

- Eurostat:

$$z_i^{Eurostat} \sim \text{Normal}\ (y_i + \log(\gamma_{u,t}^{Eurostat}), e^{Eurostat})$$

U: undercount group, u= 1,2

t = 2011, 2012, …, 2017

The prior distributions for bias parameter of Eurostat data are grouped into two categories as low undercount and high undercount destination countries. Please see Table 8 for the categories. We allowed for variation in posterior distributions by undercount group and year. However, the same prior distribution is used for each year and undercount group. Similar to 2011 Census we assume no bias on average for Eurostat data (mean = 1) as follows:

$$\gamma_{u,t}^{Eurostat} \sim N(1{,}100).$$

However, we assume that the migrant stocks from Eurostat will be slightly less precise than those from census. Hence, the accuracy parameter is assumed to be distributed normally as follows:

$$e^{Eurostat} \sim \text{Normal}(1111, 0.01).$$

- LFS:

$$z_i^{LFS} \sim \text{Normal}\ (y_i + \log(\gamma_{d,t}^{LFS}), e_i^{LFS})$$
$$\gamma_{d,t}^{LFS} \sim N(b_{d,t}^{LFS}, \zeta^{LFS})$$

d: country of residence i.e. where the survey took place

t: 2016 and 2017

As mentioned in Chapter 4, LFS sampling frame differs for each country. Therefore, country of residence and year specific bias parameters ($\gamma_{d,t}^{LFS}$) are utilised for LFS measurement error model. In accordance with this, to account for abovementioned differences, the accuracy of LFS is assumed to be related to the size of the migrant stock in each corridor and year. The following model is used to estimate the accuracy for each iteration in our Bayesian model:

$$e_i^{LFS} = \text{A} + \text{B}\ \text{n}_i^{LFS},$$

where A is the intercept, B is the slope and $\text{n}_\text{i}^{\text{LFS}}$is LFS reported migrant stock for each origin-destination corridor and year.

$$A \sim \text{Normal}\ (0.5, 1)$$
$$B \sim \text{Normal}\ (0.00275, 10000)$$

- Facebook:

$$z_i^{Facebook\ MAU} \sim \text{Normal}\ (y_i + \log(\gamma_{c,t}^{Facebook\ MAU}), \tau^{Facebook\ MAU})$$
$$z_i^{Facebook\ DAU} \sim \text{Normal}\ (y_i + \log(\gamma_{c,t}^{Facebook\ DAU}), \tau^{Facebook\ DAU})$$

As mentioned in Chapter 4, destination countries in Facebook MAU are allocated to four groups in 2016 and 2017, and five groups in 2018, and destination countries in Facebook

DAU are allocated to five groups in 2018. In each year, both Malta and Cyprus had their own groups with a higher mean value than rest of the countries. The group allocation of the destination countries can be found inTable 9.

The prior distributions for Facebook Mau and Facebook DAU bias parameters for each group and year are as follows:

$$\gamma_{1,2016}^{Facebook\ MAU} \sim N(0.35,100)$$
$$\gamma_{2,2016}^{Facebook\ MAU} \sim N(0.50,100)$$
$$\gamma_{3,2016}^{Facebook\ MAU} \sim N(0.70,100)$$
$$\gamma_{4,2016}^{Facebook\ MAU} \sim N(0.85,100)$$
$$\gamma_{1,2017}^{Facebook\ MAU} \sim N(0.45,100)$$
$$\gamma_{2,2017}^{Facebook\ MAU} \sim N(0.60,100)$$
$$\gamma_{3,2017}^{Facebook\ MAU} \sim N(0.80,100)$$
$$\gamma_{4,2017}^{Facebook\ MAU} \sim N(1.15,100)$$
$$\gamma_{1,2018}^{Facebook\ MAU} \sim N(0.55,100)$$
$$\gamma_{2,2018}^{Facebook\ MAU} \sim N(0.70,100)$$
$$\gamma_{3,2018}^{Facebook\ MAU} \sim N(0.90,100)$$
$$\gamma_{4,2018}^{Facebook\ MAU} \sim N(1.05,100)$$
$$\gamma_{5,2018}^{Facebook\ MAU} \sim N(1.50,100)$$
$$\gamma_{1,2018}^{Facebook\ DAU} \sim N(0.35,100)$$
$$\gamma_{2,2018}^{Facebook\ DAU} \sim N(0.50,100)$$
$$\gamma_{3,2018}^{Facebook\ DAU} \sim N(0.70,100)$$
$$\gamma_{4,2018}^{Facebook\ DAU} \sim N(0.90,100)$$
$$\gamma_{5,2018}^{Facebook\ DAU} \sim N(1.00,100)$$

The accuracy of Facebook MAU and DAU are as follows:

$$e^{Facebook\ MAU} \sim \text{Normal}(50, 0.1)$$
$$e^{Facebook\ DAU} \sim \text{Normal}(100, 0.01)$$

Annex 2: Input data structure

The required data for the JAGS jags.parfit() function is structured as a list. Almost all of the items (except indices used in the measurement error models to inform a specific data source is available for which row i) in this list are saved in a data frame in R environment. This data frame consists of six columns (origin, destination, year, corridor, source and stock) with destination changing the fastest, then in the following order; origin, year and source. Corridor is the index number starting from one to 756 (28 x 27) and replicated for each year and source in our model.

Origin	Destination	Year	Corridor	Source	Stock
Austria	Belgium	2011	1	Eurostat	
Austria	Bulgaria	2011	2	Eurostat	
Belgium	Austria	2011	28	Eurostat	
Belgium	Bulgaria	2011	29	Eurostat	
United Kingdom	Sweden	2011	756	Eurostat	
Austria	Belgium	2012	1	Eurostat	
United Kingdom	Sweden	2018	756	Eurostat	
Austria	Belgium	2011	1	Census	
United Kingdom	Sweden	2018	756	Facebook DAU	

Annex 3: Twitter Developer Agreement[62]

Developer Agreement
Effective: May 25, 2018.

This Twitter Developer Agreement ("Agreement") is made between you (either an individual or an entity, referred to herein as "you") and Twitter, Inc. and Twitter International Company (collectively, "Twitter") and governs your access to and use of the Licensed Material (as defined below). Your use of Twitter's websites, SMS, APIs, email notifications, applications, buttons, embeds, ads, and our other covered services is governed by our general Terms of Service and Privacy Policy.

PLEASE READ THE TERMS AND CONDITIONS OF THIS AGREEMENT CAREFULLY, INCLUDING WITHOUT LIMITATION ANY LINKED TERMS AND CONDITIONS APPEARING OR REFERENCED BELOW, WHICH ARE HEREBY MADE PART OF THIS LICENSE AGREEMENT. BY USING THE LICENSED MATERIAL, YOU ARE AGREEING THAT YOU HAVE READ, AND THAT YOU AGREE TO COMPLY WITH AND TO BE BOUND BY THE TERMS AND CONDITIONS OF THIS AGREEMENT AND ALL APPLICABLE LAWS AND REGULATIONS IN THEIR ENTIRETY WITHOUT LIMITATION OR QUALIFICATION. IF YOU DO NOT AGREE TO BE BOUND BY THIS AGREEMENT, THEN YOU MAY NOT ACCESS OR OTHERWISE USE THE LICENSED MATERIAL. THIS AGREEMENT IS EFFECTIVE AS OF THE FIRST DATE THAT YOU USE THE LICENSED MATERIAL ("EFFECTIVE DATE").

IF YOU ARE AN INDIVIDUAL REPRESENTING AN ENTITY, YOU ACKNOWLEDGE THAT YOU HAVE THE APPROPRIATE AUTHORITY TO ACCEPT THIS AGREEMENT ON BEHALF OF SUCH ENTITY. YOU MAY NOT USE THE LICENSED MATERIAL AND MAY NOT ACCEPT THIS AGREEMENT IF YOU ARE NOT OF LEGAL AGE TO FORM A BINDING CONTRACT WITH TWITTER, OR YOU ARE BARRED FROM USING OR RECEIVING THE LICENSED MATERIAL UNDER APPLICABLE LAW.

I. Twitter API and Twitter Content

A. Definitions

1. **Twitter Content** – Tweets, Tweet IDs, Twitter end user profile information, Periscope Broadcasts, Broadcast IDs and any other data and information made available to you through the Twitter API or by any other means authorized by Twitter, and any copies and derivative works thereof.
2. **Broadcast ID** - A unique identification number generated for each Periscope Broadcast.
3. **Developer Site** – Twitter's developer site located at https://developer.twitter.com.
4. **End Users** – Users of your Services.
5. **Licensed Material** – A collective term for the Twitter API and Twitter Content.
6. **Periscope Broadcast** - A live or on-demand video stream that is publicly displayed on Twitter Services and is generated by a user via Twitter's Periscope Producer feature (as set forth at https://help.periscope.tv/customer/en/portal/articles/2600293).
7. **Services** – Your websites, applications and other offerings that display Twitter Content or otherwise use the Licensed Material as approved by Twitter through any onboarding process.
8. **Tweet ID** – A unique identification number generated for each Tweet.
9. **Tweet** – a short-form text and/or multimedia-based posting made on Twitter Services.
10. **Direct Message** - A text and/or multimedia-based posting that is privately sent on Twitter Services by one end user to one or more specific end user(s).
11. **Twitter API** – The Twitter Application Programming Interface ("**API**"), Software Development Kit ("**SDK**") and/or the related documentation, data, code, and other materials provided by Twitter with the API, as updated from time to time, including without limitation through the Developer Site.
12. **Twitter Marks** – The Twitter name, trademarks, or logos that Twitter makes available to you, including via the Developer Site.
13. **Twitter Services** – Twitter's offerings and platforms, including without limitation, those offered via https://twitter.com and Twitter's mobile applications.

[62] Twitter, Inc. 2018a

B. License from Twitter. Subject to the terms and conditions in this Agreement (as a condition to the grant below), Twitter hereby grants you and you accept a non-exclusive, royalty free, non-transferable, non-sublicensable, revocable license solely to:

1. Use the Twitter API to integrate Twitter Content into your Services or conduct analysis of such Twitter Content;
2. Copy a reasonable amount of and display the Twitter Content on and through your Services to End Users, as permitted by this Agreement;
3. Modify Twitter Content only to format it for display on your Services; and
4. Use and display Twitter Marks, solely to attribute Twitter's offerings as the source of the Twitter Content, as set forth herein.

C. License to Twitter You hereby grant Twitter and Twitter accepts a non-exclusive, royalty free, non-transferable, non-sublicensable revocable license to access, index, and cache by any means, including web spiders and/or crawlers, any webpage on which you display Twitter Content using embedded Tweets or embedded timelines.

D. Incorporated Terms. Your use of the Licensed Material is further subject to and governed by the following terms and conditions:

1. the Twitter Developer Policy located at https://developer.twitter.com/en/developer-terms/policy ("**Developer Policy**");
2. as it relates to your display of any of the Twitter Content, the Display Requirements located at https://developer.twitter.com/en/developer-terms/display-requirements ("**Display Requirements**");
3. as it relates to your use and display of the Twitter Marks, the Twitter Brand Assets and Guidelines located at https://twitter.com/logo and https://www.periscope.tv/trademarkpolicy ("**Brand Guidelines**"); and
4. as it relates to taking automated actions on your account, the Automation Rules located at https://support.twitter.com/articles/76915 ("**Automation Rules**").

The Developer Policy, Display Requirements, Brand Guidelines, and Automation Rules are collectively referred to herein as the "**Developer Terms**". You agree to the Developer Terms, which are hereby incorporated by reference and are available in hardcopy upon request to Twitter. In the event of a conflict between the Developer Terms and this Agreement, this Agreement shall control. None of the Developer Terms expand or extend the license to the Twitter API, Twitter Content or Twitter Marks granted in this Agreement.

II. Restrictions on Use of Licensed Materials

A. **Reverse Engineering and other Limitations.** You will not or attempt to (and will not allow others to) **1)** reverse engineer, decompile, disassemble or translate the Twitter API, or otherwise attempt to derive source code, trade secrets or know-how in or underlying any Twitter API or any portion thereof; **2)** interfere with, modify, disrupt or disable features or functionality of the Twitter API, including without limitation any such mechanism used to restrict or control the functionality, or defeat, avoid, bypass, remove, deactivate or otherwise circumvent any software protection or monitoring mechanisms of the Twitter API; **3)** sell, rent, lease, sublicense, distribute, redistribute, syndicate, create derivative works of, assign or otherwise transfer or provide access to, in whole or in part, the Licensed Material to any third party except as expressly permitted herein; **4)** provide use of the Twitter API on a service bureau, rental or managed services basis or permit other individuals or entities to create links to the Twitter API or "frame" or "mirror" the Twitter API on any other server, or wireless or Internet-based device, or otherwise make available to a third party, any token, key, password or other login credentials to the Twitter API; or **5)**use the Licensed Material for any illegal, unauthorized or other improper purposes.

B. **Rate Limits.** You will not attempt to exceed or circumvent limitations on access, calls and use of the Twitter API ("**Rate Limits**"), or otherwise use the Twitter API in a manner that exceeds reasonable request volume, constitutes excessive or abusive usage, or otherwise fails to comply or is inconsistent with any part of this Agreement. If you exceed or Twitter reasonably believes that you have attempted to circumvent Rate Limits, controls to limit use of the Twitter APIs or the terms and conditions of this Agreement, then your ability to use the Licensed Materials may be temporarily suspended or permanently blocked. Twitter may monitor your use of the Twitter API to improve the Twitter Services and to ensure your compliance with this Agreement and the Developer Terms.

C. **Geographic Data.** Your license to use Twitter Content in this Agreement does not allow you to (and you will not allow others to) aggregate, cache, or store location data and other geographic information contained in the Twitter Content, except in conjunction with the Twitter Content to which it is attached. Your license only allows you to use such location data and geographic information to identify the location tagged by the Twitter Content. Any use of location data or geographic information on a standalone basis or beyond the license granted herein is a breach of this Agreement.
D. **Use of Twitter Marks.** The Twitter Marks may not be included in or as part of your registered corporate name, any of your logos, or any of your service or product names. Moreover, you may not create any derivative works of the Twitter Marks or use the Twitter Marks in a manner that creates or reasonably implies an inaccurate sense of endorsement, sponsorship, or association with Twitter. You will not otherwise use business names and/or logos in a manner that can mislead, confuse, or deceive users of your Services. All use of the Twitter Marks and all goodwill arising out of such use, will inure to Twitter's benefit. You shall not use the Twitter Marks except as expressly authorized herein without Twitter's prior consent. You will not remove or alter any proprietary notices or Twitter Marks on the Licensed Material.
E. **Security.** You will maintain the security of the Twitter API and will not make available to a third party, any token, key, password or other login credentials to the Twitter API. You will use industry standard security measures to prevent unauthorized access or use of any of the features and functionality of the Twitter API, including access by viruses, worms, or any other harmful code or material. Additionally, you will keep Twitter Content (including, where applicable, personal data) confidential and secure from unauthorized access by using industry-standard organizational and technical safeguards for such data, and with no less care than it uses in connection with securing similar data you store. You will immediately notify Twitter consult and cooperate with investigations, assist with any required notices, and provide any information reasonably requested by Twitter if you know of or suspects any breach of security or potential vulnerability related to the Licensed Material and will promptly remedy such breach or potential vulnerability resulting from Your access to the Licensed Material.

III. Updates

You acknowledge that Twitter may update or modify the Twitter APIs from time to time, and at its sole discretion (in each instance, an "**Update**"). You are required to implement and use the most current version of the Twitter API and to make any changes to your Services that are required as a result of such Update, at your sole cost and expense. Updates may adversely affect the manner in which your Services access or communicate with the Twitter API or display Twitter Content. Your continued access or use of the Twitter APIs following an update or modification will constitute binding acceptance of the Update.

IV. Ownership and Feedback

A. **Ownership.** The Licensed Material is licensed, not sold, and Twitter retains and reserves all rights not expressly granted in this Agreement. You expressly acknowledge that Twitter, its licensors and its end users retain all worldwide right, title and interest in and to the Licensed Material, including all rights in patents, trademarks, trade names, copyrights, trade secrets, know-how, data (including all applications therefor), and all proprietary rights under the laws of the United States, any other jurisdiction or any treaty ("**IP Rights**"). You agree not to do anything inconsistent with such ownership, including without limitation, challenging Twitter's ownership of the Twitter Marks, challenging the validity of the licenses granted herein, or otherwise copying or exploiting the Twitter Marks during or after the termination of this Agreement, except as specifically authorized herein. If you acquire any rights in the Twitter Marks or any confusingly similar marks, by operation of law or otherwise, you will, at no expense to Twitter, immediately assign such rights to Twitter.
B. **Feedback.** You may provide Twitter with comments concerning the Licensed Material, Twitter Services or your evaluation and use thereof (collectively, "**Feedback**"). You hereby grant Twitter all rights, title and ownership of such Feedback (including all intellectual property rights therein), and Twitter may use

the Feedback for any and all commercial and non-commercial purposes with no obligation of any kind to you.

V. Termination

Twitter may immediately terminate or suspend this Agreement, any rights granted herein, and/or your license to the Licensed Materials, at its sole discretion at any time, for any reason by providing notice to you. You may terminate this Agreement at any time by ceasing your access to the Twitter API and use of all Twitter Content. Upon termination of this Agreement, (a) all licenses granted herein immediately expire and you must cease use of all Licensed Material; and (b) you shall permanently delete all Licensed Material and Twitter Marks in all forms and types of media, and copies thereof, in your possession. The parties to this Agreement will not be liable to each other for any damages resulting solely from termination of this Agreement as permitted under this Agreement. Sections II, IV, V, VI and VII of this Agreement will survive the termination of this Agreement.

VI. Confidentiality

You may be given access to certain non-public information, software, and specifications relating to the Licensed Material ("**Confidential Information**"), which is confidential and proprietary to Twitter. You may use this Confidential Information only as necessary in exercising your rights granted in this Agreement. You may not disclose any of this Confidential Information to any third party without Twitter's prior written consent. You agree that you will protect this Confidential Information from unauthorized use, access, or disclosure in the same manner that you would use to protect your own confidential and proprietary information of a similar nature and in no event with less than a reasonable degree of care.

VII. Other Important Terms

A. **User Protection.** Twitter Content, and information derived from Twitter Content, may not be used by, or knowingly displayed, distributed, or otherwise made available to:
    1. any public sector entity (or any entities providing services to such entities) for surveillance purposes, including but not limited to:
        a. investigating or tracking Twitter's users or their Twitter Content; and,
        b. tracking, alerting, or other monitoring of sensitive events (including but not limited to protests, rallies, or community organizing meetings);
    2. any public sector entity (or any entities providing services to such entities) whose primary function or mission includes conducting surveillance or gathering intelligence;
    3. any entity for the purposes of conducting or providing surveillance, analyses or research that isolates a group of individuals or any single individual for any unlawful or discriminatory purpose or in a manner that would be inconsistent with our users' reasonable expectations of privacy;
    4. any entity to target, segment, or profile individuals based on health (including pregnancy), negative financial status or condition, political affiliation or beliefs, racial or ethnic origin, religious or philosophical affiliation or beliefs, sex life or sexual orientation, trade union membership, data relating to any alleged or actual commission of a crime, or any other sensitive categories of personal information prohibited by law;
    5. any entity that you reasonably believe will use such data to violate the Universal Declaration of Human Rights (located at http://www.un.org/en/documents/udhr/), including without limitation Articles 12, 18, or 19.

    If law enforcement personnel request information from you about Twitter or its users for the purposes of an ongoing investigation, you must refer them to Twitter's Guidelines for Law Enforcement located at https://t.co/le.
B. **Additional Terms for Permitted Government Use.** The Twitter API and Twitter Content are "commercial items" as that term is defined at 48 C.F.R. 2.101, consisting of "commercial computer software" and "commercial computer software documentation" as such terms are used in 48 C.F.R. 12.212. Any use, modification, derivative, reproduction, release, performance, display, disclosure

or distribution of the Twitter API or Twitter Content by any government entity is prohibited, except as expressly permitted by the terms of this Agreement. Additionally, any use by U.S. government entities must be in accordance with 48 C.F.R. 12.212 and 48 C.F.R. 227.7202-1 through 227.7202-4. If you use the Twitter API or Twitter Content in your official capacity as an employee or representative of a U.S., state or local government entity and you are legally unable to accept the indemnity, jurisdiction, venue or other clauses herein, then those clauses do not apply to such entity, but only to the extent as required by applicable law. For the purpose of this provision, contractor/manufacturer is Twitter, Inc., 1355 Market Street, Suite 900, San Francisco, California 94103.

C. **Data Protection.** Twitter International Company ("**TIC**"), an Irish registered company, controls some of the Twitter Content, as set forth in the Twitter Privacy Policy (https://www.twitter.com/privacy), and has authorized Twitter to license such Twitter Content under this Agreement (such data is "TIC Data"). To the extent that you are relying upon the EU Commission's implementing Decision 2016/1250 pursuant to Directive 95/46/EC on the adequacy of the protection provided by the EU-U.S. Privacy Shield (the "**Privacy Shield**") and is certified under Privacy Shield to receive categories of data which include the TIC Data, you represent and warrant it will comply with the Privacy Shield principles. Without limiting the foregoing, if for any reason you are unable to comply with such principles or your Privacy Shield certification should end, you will immediately notify Twitter and take reasonable and appropriate steps to remedy any non-compliance, or cease access to the Twitter API and use of any and all TIC Data. If a transfer of TIC Data by you is not covered by Privacy Shield, and then only if you are located or transfer such TIC Data out of (a) the European Economic Area, or (b) a jurisdiction where a European Commission positive adequacy decision under Article 25(6) of Directive 95/46/EC is in force and covers such transfer, then use of such TIC Data is subject to the model contractual clauses annexed to Commission Decision 2004/915/EC (the "**Clauses**"), which are hereby incorporated into this Agreement. In such cases, TIC is the 'data exporter' and you are the 'data importer' as defined in the Clauses, and you select option (iii) of Clause II(h) and agree to the data processing principles of Annex A to the Clauses. For the purposes of Annex B to the Clauses, the following shall apply: (i) 'Data subjects' are the users of the Twitter Services or individuals whose personal data is in the TIC Data; (ii) the 'Purpose of the transfer(s)' is the performance of this Agreement and the provision of services by you to End Users; (iii) the 'Categories of data' are TIC Data as defined herein; (iv) the 'Recipients' are End Users and you; (v) 'Sensitive data' is personal data regarding an individual's racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, health or sex life, criminal convictions or alleged commission of an offense; and (vi) the 'contact points for data protection enquiries' are the representatives of TIC and you with responsibility for data privacy.

D. **Compliance Audit.** Twitter, or a third party agent subject to obligations of confidentiality, shall be entitled to inspect and audit any records or activity related to your access to the Licensed Material for the purpose of verifying compliance with this Agreement. Twitter may exercise its audit right at anytime upon notice. You will provide your full cooperation and assistance with such audit and provide access to all Licensed Material in your possession or control, applicable agreements and records. Without limiting the generality of the foregoing, as part of the audit, Twitter may request, and you agree to provide, a written report, signed by an authorized representative, listing your then-current deployment of the Licensed Material and Twitter Content. The rights and requirements of this section will survive for one (1) year following the termination of this Agreement.

E. **Compliance with Laws; Export and Import.** Each party will comply with all applicable foreign, federal, state, and local laws, rules and regulations, including without limitation, all applicable laws relating to bribery and/or corruption. The Licensed Material is subject to U.S. export laws and may be subject to import and use laws of the country where it is delivered or used. You agree to abide by these laws. Under these laws, the Licensed Material may not be sold, leased, downloaded, moved, exported, re-exported, or transferred across borders without a license, or approval from the relevant government authority, to any country or to any foreign national restricted by these laws, including countries embargoed by the U.S. Government (currently Cuba, Iran, North Korea, Northern Sudan and Syria); or to any restricted or denied end-user including, but not limited to, any person or entity prohibited by the U.S. Office of Foreign

Assets Control; or for any restricted end-use. You will maintain all rights and licenses that are required with respect to your Services.

F. **Warranty Disclaimer.** THE LICENSED MATERIAL IS PROVIDED TO YOU "AS IS", "WHERE IS", WITH ALL FAULTS AND EACH PARTY DISCLAIMS ALL WARRANTIES, WHETHER EXPRESS, IMPLIED, STATUTORY, OR OTHERWISE, INCLUDING WITHOUT LIMITATION WARRANTIES OF MERCHANTABILITY, NONINFRINGEMENT, FITNESS FOR A PARTICULAR PURPOSE, AND ANY WARRANTIES OR CONDITIONS ARISING OUT OF THIS AGREEMENT, COURSE OF DEALING OR USAGE OF TRADE. TWITTER DOES NOT WARRANT THAT THE LICENSED MATERIAL OR ANY OTHER TWITTER PRODUCT OR SERVICE PROVIDED HEREUNDER WILL MEET ANY OF YOUR REQUIREMENTS OR THAT USE OF SUCH LICENSED MATERIAL OR OTHER PRODUCTS OR SERVICES WILL BE ERROR-FREE, UNINTERRUPTED, VIRUS-FREE OR SECURE. THIS DISCLAIMER OF WARRANTY MAY NOT BE VALID IN SOME JURISDICTIONS AND YOU MAY HAVE WARRANTY RIGHTS UNDER LAW WHICH MAY NOT BE WAIVED OR DISCLAIMED. ANY SUCH WARRANTY EXTENDS ONLY FOR THIRTY (30) DAYS FROM THE EFFECTIVE DATE OF THIS AGREEMENT (UNLESS SUCH LAW PROVIDES OTHERWISE).

G. **Indemnification.** You shall defend Twitter against any and all actions, demands, claims and suits (including without limitation product liability claims), and indemnify and hold Twitter harmless from any and all liabilities, damages and costs (including without limitation reasonable attorneys' fees) to the extent arising out of: (i) your use of the Licensed Material in any manner that is inconsistent with this Agreement; or (ii) the performance, promotion, sale or distribution of your Services. In the event Twitter seeks indemnification or defense from you under this provision, Twitter will promptly notify you in writing of the claim(s) brought against Twitter for which it seeks indemnification or defense. Twitter reserves the right, at its option and sole discretion, to assume full control of the defense of claims with legal counsel of its choice. You may not enter into any third party agreement, which would, in any manner whatsoever, affect the rights of Twitter, constitute an admission of fault by Twitter or bind Twitter in any manner, without the prior written consent of Twitter. In the event Twitter assumes control of the defense of such claim, Twitter shall not settle any such claim requiring payment from you without your prior written approval.

H. **Limitation of Liability.** IN NO EVENT WILL TWITTER BE LIABLE TO YOU OR ANY END USERS FOR ANY INDIRECT, SPECIAL, INCIDENTAL, EXEMPLARY, PUNITIVE OR CONSEQUENTIAL DAMAGES OR ANY LOSS OF OR DAMAGE TO USE, DATA, BUSINESS, GOODWILL OR PROFITS ARISING OUT OF OR IN CONNECTION WITH THIS AGREEMENT. IN ANY CASE, TWITTER'S AGGREGATE LIABILITY FOR ANY AND ALL CLAIMS UNDER THIS AGREEMENT WILL NOT EXCEED $50.00 USD. THE FOREGOING LIMITATIONS, EXCLUSIONS AND DISCLAIMERS SHALL APPLY REGARDLESS OF WHETHER SUCH LIABILITY ARISES FROM ANY CLAIM BASED UPON CONTRACT, WARRANTY, TORT (INCLUDING NEGLIGENCE), STRICT LIABILITY OR OTHERWISE, AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH LOSS OR DAMAGE. INSOFAR AS APPLICABLE LAW PROHIBITS ANY LIMITATION ON LIABILITY HEREIN, THE PARTIES AGREE THAT SUCH LIMITATION WILL BE AUTOMATICALLY MODIFIED, BUT ONLY TO THE EXTENT SO AS TO MAKE THE LIMITATION COMPLIANT WITH APPLICABLE LAW. THE PARTIES AGREE THAT THE LIMITATIONS ON LIABILITIES SET FORTH HEREIN ARE AGREED ALLOCATIONS OF RISK AND SUCH LIMITATIONS WILL APPLY NOTWITHSTANDING THE FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY.

I. **Updates.** Twitter may update or modify this Agreement, Developer Terms, and other terms and conditions, from time to time at its sole discretion by posting the changes on this site or by otherwise notifying you (such notice may be via email). You acknowledge that these updates and modifications may adversely affect how your Service accesses or communicates with the Twitter API. If any change is unacceptable to you, your only recourse is to cease all use of the Licensed Material. Your continued access or use of the Licensed Material will constitute binding acceptance of the such updates and modifications.

J. **Miscellaneous.** This Agreement constitutes the entire agreement among the parties with respect to the subject matter and supersedes and merges all prior proposals, understandings and contemporaneous communications. Any modification to this Agreement must be in a writing signed by both you and Twitter, Inc. You may not assign any of the rights or obligations granted hereunder, in whole or in part, whether voluntarily or by operation of law,

contract, merger (whether you are the surviving or disappearing entity), stock or asset sale, consolidation, dissolution, through government action or otherwise, except with the prior written consent of Twitter, Inc. Twitter, Inc. is authorized to sign modifications and consents on behalf of Twitter International Company, an Irish company responsible for the information of Twitter users who live outside the United States. Any attempted assignment in violation of this paragraph is null and void, and Twitter may terminate this Agreement. This Agreement does not create or imply any partnership, agency or joint venture. This Agreement will be governed by and construed in accordance with the laws of the State of California, without regard to or application of conflicts of law rules or principles. Any dispute, claim or controversy arising out of or relating to this Agreement or the breach, termination, enforcement, interpretation or validity thereof, including the determination of the scope or applicability of this Agreement to arbitrate, shall be determined by arbitration in San Francisco, CA before a single arbitrator. The arbitration shall be administered by JAMS pursuant to its Comprehensive Arbitration Rules and Procedures. Judgment on the Award may be entered in any court having jurisdiction. You and Twitter hereby expressly waive trial by jury. As an alternative, you may bring your claim in your local "small claims" court, if permitted by that small claims court's rules. You may bring claims only on your own behalf, and unless Twitter agrees, the arbitrator may not consolidate more than one person's claims. Despite the foregoing, you agree that money damages would be an inadequate remedy for Twitter in the event of a breach or threatened breach of a provision of this Agreement protecting Twitter's intellectual property or Confidential Information, and that in the event of such a breach or threat, Twitter, in addition to any other remedies to which it is entitled, is entitled to such preliminary or injunctive relief (including an order prohibiting Company from taking actions in breach of such provisions), without the need for posting bond, and specific performance as may be appropriate. The parties agree that neither the United Nations Convention on Contracts for the International Sale of Goods, nor the Uniform Computer Information Transaction Act (UCITA) shall apply to this Agreement, regardless of the states in which the parties do business or are incorporated. No waiver by Twitter of any covenant or right under this Agreement will be effective unless memorialized in a writing duly authorized by Twitter. If any part of this Agreement is determined to be invalid or unenforceable by a court of competent jurisdiction, that provision will be enforced to the maximum extent permissible and the remaining provisions of this Agreement will remain in full force and effect.

Annex 4: Twitter Developer Policy[63]

Developer Policy
Effective: November 3, 2017.

In addition to the Developer Agreement, this Developer Policy ("Policy") provides rules and guidelines for developers who interact with Twitter's ecosystem of applications, services, website, web pages and content including any content that we may make available through our other covered services set forth at https://support.twitter.com/articles/20172501("Twitter Services"). Policy violations are also considered violations of the Developer Agreement. Take a look at the Definitions for the meaning of capitalized words used in this Policy. These policies may be changed from time to time without notice. Please check here for any updates.

I. Guiding Principles

A. A Few Key Points

1. Keep any API keys or other access credentials private and use only as permitted.
2. Respect our requirements on how to display and interact with users' content.
3. If your application will need more than 1 million user tokens, you must contact us about your Twitter API access, as you may be subject to additional terms.

---

[63] Twitter, Inc. 2017

4. Twitter may monitor your use of the Twitter API to improve the Twitter Services, examine commercial use and ensure your compliance with this Policy.

5. Remember, Twitter may suspend or revoke access to the Twitter API if we believe you are in violation of this Policy. Do not apply for or register additional API tokens if Twitter has suspended your account. Instead, contact us.

B. Maintain the Integrity of Twitter's Products

1. Follow the Display Requirements, Twitter Rules and Periscope Community Guidelines. If your Service facilitates or induces users to violate the Twitter Rules or Periscope Community Guidelines, you must figure out how to prevent the abuse or Twitter may suspend or terminate your access to the Twitter API. We've provided guidance in our Abuse Prevention and Security help page.

2. If your Service submits content to Twitter that includes a Twitter username, submit the correct Twitter username ("@username").

3. Do not modify, translate or delete a portion of the Twitter Content.

4. Maintain the features and functionality of Twitter Content and Twitter API. Do not interfere with, intercept, disrupt, filter, or disable any features of Twitter or the Twitter API, including the Twitter Content of embedded Tweets and embedded timelines.

5. Do not exceed or circumvent limitations on access, calls, sharing, privacy settings, or use permitted in this Policy, or as otherwise set forth on the Developer Site, or communicated to you by Twitter.

6. Do not remove or alter any proprietary notices or marks on Twitter Content or the Twitter API.

7. Do not (and do not allow others to) aggregate, cache, or store location data and other geographic information contained in the Twitter Content, except as part of a Tweet or Periscope Broadcast. Any use of location data or geographic information on a standalone basis is prohibited.

C. Respect Users' Control and Privacy

1. Get the user's express consent before you do any of the following:

    a. Take any actions on a user's behalf, including posting Twitter Content, following/unfollowing other users, modifying profile information, starting a Periscope Broadcast or adding hashtags or other data to the user's Tweets. A user authenticating through your Service does not constitute user consent.

    b. Republish Twitter Content accessed by means other than via the Twitter API or other Twitter tools.

    c. Use a user's Twitter Content to promote a commercial product or service, either on a commercial durable good or as part of an advertisement.

    d. Store non-public Twitter Content such as Direct Messages or other private or confidential information.

    e. Share or publish protected Twitter Content, private or confidential information.

    f. Configure media to be sent in a Direct Message as "shared" (i.e. reusable across multiple Direct Messages). You must also provide the user with clear notice that "shared" media sent in a Direct Message will be viewable by anyone with the media's URL.

2. Do not (and do not permit others to) associate the Twitter Content with any person, household, device, browser, or other individual identifier, unless you or the entity on whose sole behalf you make such an association do so (a) with the express opt-in consent of the applicable individual; or (b) based solely on publicly available data and/or data provided directly by the applicable individual that the individual would reasonably expect to be used for that purpose.

3. If Twitter Content is deleted, gains protected status, or is otherwise suspended, withheld, modified, or removed from the Twitter Service (including removal of location information), you will make all reasonable efforts to delete or modify such Twitter Content (as applicable) as soon as reasonably possible, and in any case within 24 hours after a request to do so by Twitter or by a Twitter user with regard to their Twitter Content, unless otherwise prohibited by applicable law or regulation, and with the express written permission of Twitter.

4. If your Service will display Twitter Content to the public or to end users of your Service, and you do not use Twitter Kit or Twitter for Websites to do so, then you must use the Twitter API to retrieve the most current version of the Twitter Content for such display. If Twitter Content ceases to be available through the Twitter API, you may not display such Twitter Content and must remove it from non-display portions of your Service as soon as reasonably possible.

5. If your Service allows users to post Twitter Content to Twitter, then, before publishing, show the user exactly what will be published, including whether any geotags will be added to the Twitter Content. If you will send read receipt events for Direct Messages, you should inform users they will be sent as part of a conversation, such as by directly providing this notice to users in your application or by displaying read receipts from other participants in a conversation.

6. If your Service allows users to post Twitter Content to your Service and Twitter, then, before publishing to the Service:

    a. Explain how you will use the Twitter Content;

    b. Obtain proper permission to use the Twitter Content; and

    c. Continue to use such Twitter Content in accordance with this Policy in connection with the Twitter Content.

7. Display your Service's privacy policy to users before download, installation or sign up of your application. Your privacy policy must be consistent with all applicable laws, and be no less protective of end users than Twitter's Privacy Policy located at https://twitter.com/privacy including any relevant incorporated policies such as the supplemental terms located at https://support.twitter.com/articles/20172501. You must comply with your privacy policy, which must clearly disclose the information you collect from users, how you use and share that information (including with Twitter), and how users can contact you with inquiries and requests regarding their information. If for any reason you are unable to comply with your privacy policy or any privacy requirement of the Developer Agreement or Policy, you must promptly inform Twitter and take reasonable and appropriate steps to remedy any non-compliance, or cease your access to the Twitter API and use of all Twitter Content.

8. If your Service uses cookies, disclose in your privacy policy:

    a. Whether third parties collect user information on your Service and across other websites or online services;

    b. Information about user options for cookie management and whether you honor the Do Not Track setting in supporting web browsers.

9. If your Service adds location information to users' Tweets or Periscope Broadcasts:

a. Disclose when you add location information, whether as a geotag or annotations data, and whether you add a place or specific coordinates.

b. Comply with Geo Developers Guidelines if your application allows users to Tweet with their location.

10. Do not store Twitter passwords.

D. Clearly Identify Your Service

1. Make sure users understand your identity and the source and purpose of your Service. For example:

    a. Don’t use a name or logo that falsely implies you or your company is related to another business or person.

    b. Don’t use a shortened URL for your Service that attempts to mask the destination site.

    c. Don’t use a URL for your Service that directs users to

        i. a site that is unrelated to your Service

        ii. a site that encourages users to violate the Twitter Rules or the Periscope Community Guidelines.

        iii. a spam or malware site.

2. Do not replicate, frame, or mirror the Twitter website or its design.

E. Keep Twitter Spam Free

1. Follow the Abuse and Spam rules here.

2. Comply with the automation rules if your Service performs automatic actions.

3. Do not do any of the following:

    a. Mass-register applications.

    b. Create tokens/applications to sell names, prevent others from using names, or other commercial use.

    c. Use third-party content feeds to update and maintain accounts under those third parties’ names.

    d. Name squat by submitting multiple applications with the same function under different names.

    e. Publish links to malicious content.

    f. Publish pornographic or obscene images to user profile images and background images.

F. Be a Good Partner to Twitter

1. Follow the guidelines for using Tweets in broadcast if you display Tweets offline and the guidelines for using Periscope Broadcasts in a broadcast if you display Periscope Broadcasts offline.

2. If you provide Twitter Content to third parties, including downloadable datasets of Twitter Content or an API that returns Twitter Content, you will only distribute or allow download of Tweet IDs, Direct Message IDs, and/or User IDs.

a. You may, however, provide export via non-automated means (e.g., download of spreadsheets or PDF files, or use of a "save as" button) of up to 50,000 public Tweet Objects and/or User Objects per user of your Service, per day.

b. Any Twitter Content provided to third parties remains subject to this Policy, and those third parties must agree to the Twitter Terms of Service, Privacy Policy, Developer Agreement, and Developer Policy before receiving such downloads.

    i. You may not distribute more than 1,500,000 Tweet IDs to any entity (inclusive of multiple individual users associated with a single entity) within any given 30 day period, unless you are doing so on behalf of an academic institution and for the sole purpose of non-commercial research or you have received the express written permission of Twitter.

    ii. You may not distribute Tweet IDs for the purposes of (a) enabling any entity to store and analyze Tweets for a period exceeding 30 days unless you are doing so on behalf of an academic institution and for the sole purpose of non-commercial research or you have received the express written permission of Twitter, or (b) enabling any entity to circumvent any other limitations or restrictions on the distribution of Twitter Content as contained in this Policy, the Twitter Developer Agreement, or any other agreement with Twitter.

3. Use and display Twitter Marks solely to identify Twitter as the source of Twitter Content.

4. Comply with Twitter Brand Assets and Guidelines as well as the Periscope Trademark guidelines.

5. Do not do any of the following:

    a. Use a single application API key for multiple use cases or multiple application API keys for the same use case.

    b. Charge a premium above your Service's standard data and usage rates for access to Twitter Content via SMS or USSD.

    c. Sell or receive monetary or virtual compensation for Tweet actions, Periscope Broadcasts actions or the placement of Tweet actions on your Service, such as, but not limited to follow, retweet, like, heart, comment and reply.

    d. Do not use, access or analyze the Twitter API to monitor or measure the availability, performance, functionality, usage statistics or results of Twitter Services or for any other benchmarking or competitive purposes, including without limitation, monitoring or measuring:

        i. the responsiveness of Twitter Services; or

        ii. aggregate Twitter user metrics such as total number of active users, accounts, total number of Periscope Broadcast views, user engagements or account engagements.

    e. Use Twitter Content, by itself or bundled with third party data, to target users with advertising outside of the Twitter platform, including without limitation on other advertising networks, via data brokers, or through any other advertising or monetization services.

    f. Use Twitter Marks, or Twitter Certified Products Program badges, or similar marks or names in a manner that creates a false sense of endorsement, sponsorship, or association with Twitter.

g. Use the Twitter Verified Account badge, Verified Account status, or any other enhanced user categorization on Twitter Content other than that reported to you by Twitter through the API.

G. Avoid Replicating the Core Twitter Experience

1. Twitter discourages online services from replicating Twitter Service's core user experience or features.

2. The following rules apply solely to Services or applications that attempt to replicate Twitter's core user experience:

    a. You must obtain our permission to have more than 100,000 user tokens, and you may be subject to additional terms.

    b. Use the Twitter API as provided by Twitter for functionalities in your Service that are substantially similar to a Twitter Service feature and present this to your users as the default option.

    c. Display a prominent link or button in your Service that directs new users to Twitter's sign-up functionality.

    d. Do not do the following:

        i. Pay, or offer to pay, third parties for distribution. This includes offering compensation for downloads (other than transactional fees) or other mechanisms of traffic acquisition.

        ii. Arrange for your Service to be pre-installed on any other device, promoted as a "zero-rated" service, or marketed as part of a specialized data plan.

        iii. Use Twitter Content or other data collected from users to create or maintain a separate status update, social network, private messaging or live broadcasting database or service.

H. Engage in Appropriate Commercial Use

1. Advertising Around Twitter Content

    a. You may advertise around and on sites that display Tweets and Periscope Broadcasts, but you may not place any advertisements within the Twitter timeline or on or within Periscope Broadcasts on your Service other than Twitter Ads or advertisements made available through the official Twitter Kit integration with MoPub. Access to MoPub ads through Twitter Kit requires a MoPub supply account and is subject to MoPub terms of service & policies.

    b. Your advertisements cannot resemble or reasonably be confused by users as a Tweet or Periscope Broadcast.

    c. You may advertise in close proximity to the Twitter timeline or a Periscope Broadcast (e.g., banner ads above or below timeline), but there must be a clear separation between Twitter Content and your advertisements.

2. Twitter reserves the right to serve advertising via Twitter APIs ("Twitter Ads"). If you decide to serve Twitter Ads once we start delivering them, we will share a portion of advertising revenue with you in accordance with the relevant terms and conditions.

II. Rules for Specific Twitter Services or Features

A. Twitter Login

1. Present users with easy to find options to log into and out of Twitter, for example, via the OAuth protocol or Twitter Kit.

2. Provide users without a Twitter account the opportunity to create a new Twitter account.

3. Display the "Sign in with Twitter" option at least as prominently as the most prominent of any other third party social networking sign-up or sign-in marks and branding appearing on your Service.

B. Social Updates

1. If you allow users to create social updates from your own social service or a third party social networking, micro-blogging, or status update provider integrated into your Service ("Update"), you must display a prominent option to publish that content to Twitter.

2. If Updates are longer than 140 characters or not text, you must display a prominent link to publish that content to Twitter and:

    a. URLs must direct users to the page where that content is displayed. You may require users to sign in to access that page, but the content must not otherwise be restricted from being viewed.

    b. URLs must not direct users to interstitial or intermediate pages.

C. Twitter Identity

1. Once a user has authenticated via "Sign in with Twitter" via your Service, you must clearly display the user's Twitter identity via your Service. Twitter identity includes visible display of the user's avatar, Twitter user name and the Twitter bird mark.

2. Displays of the user's followers on your Service must clearly show that the relationship is associated with the Twitter Service.

D. Twitter Cards

1. Develop your Card to have the same quality experience across all platforms where Cards are displayed.

2. If your Service provides a logged-in experience, the experience prior to a user’s login must be of equivalent quality and user value.

3. Mark your Card as ‘true’ for sensitive media if such media can be displayed.

4. Use HTTPS for hosting all assets within your Card.

5. For video and audio content:

    a. Default to ‘sound off’ for videos that automatically play content.

    b. Include stop or pause controls.

6. Do not do any of the following:

    a. Exceed or circumvent Twitter’s limitations placed on any Cards, including the Card’s intended use.

    b. Attach the App Card to a user’s Tweet, unless the user is explicitly promoting or referring to the app in the Tweet.

    c. Place third-party sponsored content within Cards without Twitter’s prior approval.

d. Include content or actions within your Card that are not contextually relevant to the user's Tweet text and Tweet entities, such as URLs and media.

e. Generate active mixed content browser warnings.

f. Attach monetary incentives or transactions (including virtual currency) to activities that occur within the Card or on Twitter from your Card.

g. Apply for Cards access for domains you do not manage to prevent others from registering or utilizing Cards on those domains.

E. Twitter for Websites

1. If you expect your embedded Tweets and embedded timelines to exceed 10 million daily impressions, you must contact us about your Twitter API access, as you may be subject to additional terms.

2. If you use Twitter for Websites widgets, you must ensure that an end user is provided with clear and comprehensive information about, and consents to, the storing and accessing of cookies or other information on the end user's device as described in Twitter's cookie use where providing such information and obtaining such consent is required by law.

3. If you use embedded Tweets or embedded timelines, you must provide users legally sufficient notice that fully discloses Twitter's collection and use of data about users' browsing activities on your website, including for interest-based advertising and personalization. You must also obtain legally sufficient consent from users for such collection and use, and provide legally sufficient instructions on how users can opt out of Twitter's interest-based advertising and personalization as described here.

4. If you operate a Service targeted to children under 13, you must opt out of tailoring Twitter in any embedded Tweets or embedded timelines on your Service by setting the opt-out parameter to be true as described here.

F. Periscope Producer

1. You must provide a reasonable user-agent, as described in the Periscope Producer technical documentation, for your Service when accessing the Periscope API.

2. If you expect the number of broadcasts created by your hardware will exceed (10 million) daily broadcasts, you must contact us about your Twitter API access, as you may be subject to additional terms.

3. You must honor user requests to log out of their Periscope account on your Service.

4. You may not provide tools in your service to allow users to circumvent technological protection measures.

G. Definitions

1. **Twitter Content** - Tweets, Tweet IDs, Direct Messages, Direct Message IDs, Twitter end user profile information, User IDs, Periscope Broadcasts, Periscope Broadcast IDs and any other data and information made available to you through the Twitter API or by any other means authorized by Twitter, and any copies and derivative works thereof.

2. **Developer Site** – Twitter's developer site located at https://developer.twitter.com.

3. **Periscope Broadcast** - A user generated live video stream that is available live or on-demand, that is publicly displayed on Twitter Services.

4. **Broadcast ID** - A unique identification number generated for each Periscope Broadcast.

5. **Tweet** - A short-form text and/or multimedia-based posting made on Twitter Services.

6. **Tweet ID** - A unique identification number generated for each Tweet.
7. **Direct Message** - A text and/or multimedia-based posting that is privately sent on the Twitter Service by one end user to one or more specific end user(s).
8. **Direct Message ID** - A unique identification number generated for each Direct Message.
9. **Twitter API** - The Twitter Application Programming Interface ("API"), Software Development Kit ("SDK") and/or the related documentation, data, code, and other materials provided by Twitter, as updated from time to time, including without limitation through the Developer Site.
10. **Twitter Marks** - The Twitter name, or logos that Twitter makes available to you, including via the Developer Site.
11. **Service** - Your websites, applications, hardware and other offerings that display or otherwise use Twitter Content.
12. **User ID** - Unique identification numbers generated for each User that do not contain any personally identifiable information such as Twitter usernames or users' names.

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://europa.eu/contact

On the phone or by e-mail

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)

EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex at: http://eur-lex.europa.eu

Open data from the EU

The EU Open Data Portal (http://data.europa.eu/euodp/en/data) provides access to datasets from the EU. Data can be downloaded and reused for free, both for commercial and non-commercial purposes.

Publications Office
of the European Union