

University of Southampton Research Repository

Copyright © and Moral Rights for this thesis and, where applicable, any accompanying data are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis and the accompanying data cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content of the thesis and accompanying research data (where applicable) must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holder/s.

When referring to this thesis and any accompanying data, full bibliographic details must be given, e.g.

Thesis: Francesco Rampazzo (2020) “Following a Trail of Breadcrumbs: a Study of Migration through Digital Traces”, University of Southampton, Social Statistics and Demography, PhD Thesis, pagination.

Data: Francesco Rampazzo (2020) Title. URI [dataset]

UNIVERSITY OF SOUTHAMPTON

Faculty of Social Sciences

School of Economic, Social and Political Sciences and

Max Planck Institute for Demographic Research

**Following a Trail of Breadcrumbs:
a Study of Migration through Digital Traces**

by

Francesco Rampazzo

Supervisors:

Jakub Bijak, Agnese Vitali,

Ingmar Weber, and Emilio Zagheni

ORCID: [0000-0002-5071-7048](https://orcid.org/0000-0002-5071-7048)

*A thesis for the degree of
Doctor of Philosophy*

January 2022

University of Southampton

Abstract

Faculty of Social Sciences

School of Economic, Social and Political Sciences and

Max Planck Institute for Demographic Research

Doctor of Philosophy

**Following a Trail of Breadcrumbs:
a Study of Migration through Digital Traces**

by Francesco Rampazzo

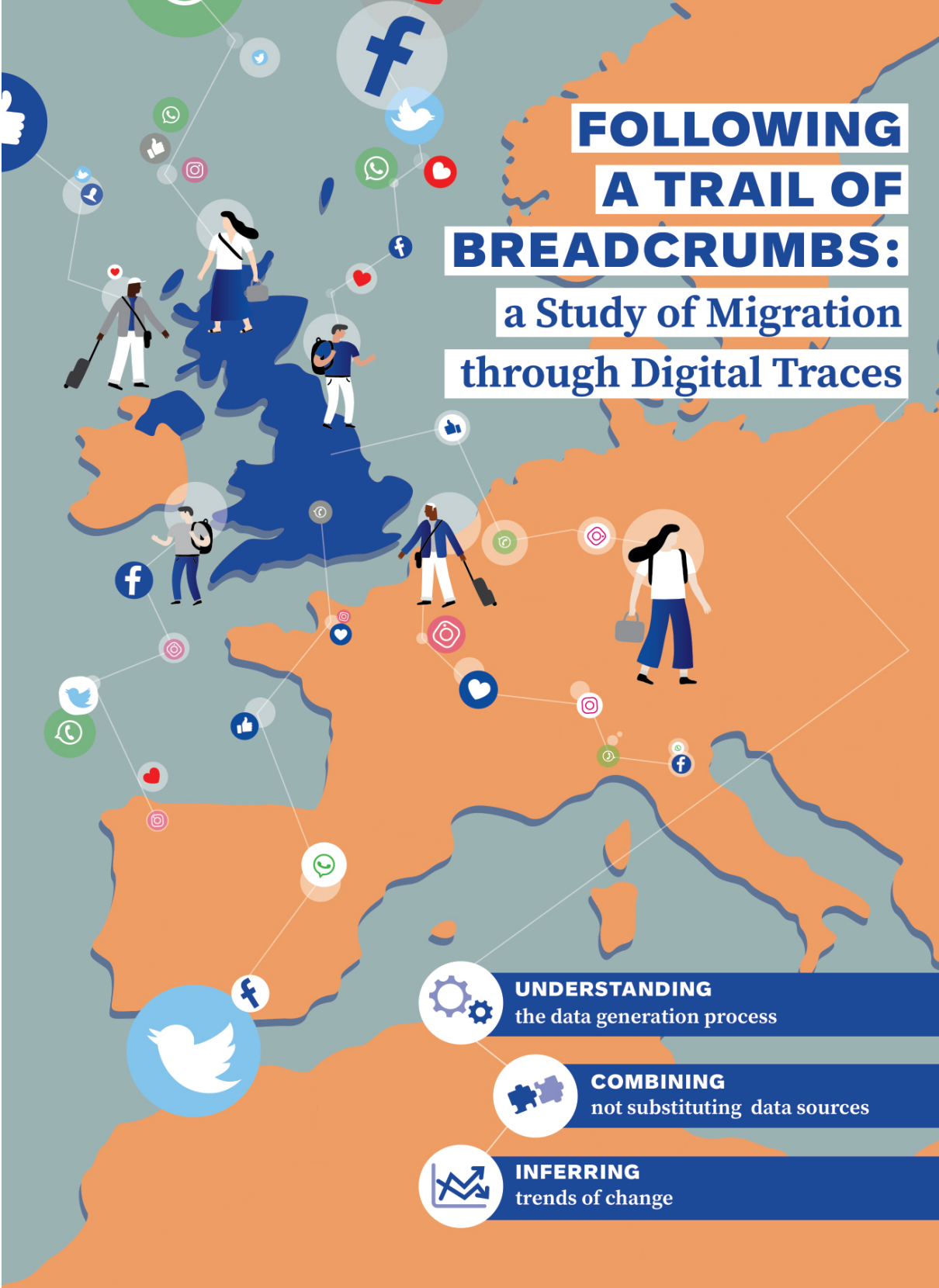
Supervisors:

Jakub Bijak, Agnese Vitali,

Ingmar Weber, and Emilio Zagheni

Abstract

An accurate estimation of international migration is hampered by a lack of timely and comprehensive data, with different definitions and measures of migration adopted by different countries. The aim of this thesis is to understand whether information from digital traces can help measure international migration. One of the approaches implemented in this thesis is to complement traditional data sources for the United Kingdom (UK) with digital traces data. The Bayesian framework proposed in the Integrated Model of European Migration (IMEM) is used to combine data from the Labour Force Survey (LFS) and the Facebook Advertising Platform in order to study the number of European migrants in the UK, aiming to produce more accurate estimates of European migrants. The thesis suggests an extension of the IMEM model to disaggregate the estimate by age and sex. Additionally, weekly time series from the Facebook Advertising Platform are analysed to infer trends of change in migration stocks over time. The quality of the data is reviewed paying particular attention to the biases of these sources. The results indicate visible yet uncertain differences between the model estimates using the Bayesian framework and individual sources. The advantages and limitations of this approach, which can be applied in other contexts, are also discussed. It seems that any individual source cannot be completely trusted, but combining sources through modelling can offer valuable insights. The main conclusions are that the data generation process should be examined, digital traces data should be combined with traditional data sources, and that digital traces data might be used to infer trends of change in migration stocks.



FOLLOWING A TRAIL OF BREADCRUMBS: a Study of Migration through Digital Traces

UNDERSTANDING
the data generation process

COMBINING
not substituting data sources

INFERRING
trends of change

Contents

List of Acronyms	xi
List of Figures	xiii
List of Tables	xv
Declaration of Authorship	xvii
Acknowledgements	xix
1 Introduction	1
1.1 Overview	2
1.2 Uncertainty	4
1.3 Contribution and Structure	5
1.4 Production of this thesis	6
2 Migration and Data: Literature Review	9
2.1 Context of Migration and the Digital Revolution	10
2.2 Migration to the United Kingdom	10
2.3 Measuring International Migration	12
2.3.1 Existing Models of Migrations	14
2.3.2 Migration Data	15
2.3.3 Survey-based Migration Data in the United Kingdom	16
2.4 Digital Revolution	18
2.4.1 Characteristics of Big Data and Digital Traces	19
2.5 Demography and Digital Traces	22
2.5.1 Mobility and Migration studies	23
2.5.2 Other Applications of Digital Traces in Demography	25
2.5.3 Potential of Digital Traces for Demographic Research	27
2.6 Facebook	27
2.6.1 Measurements in the Facebook Advertising Platform	30
3 A Framework for Estimating Migrant Stocks Using Digital Traces and Survey Data: An Application in the United Kingdom	33
3.1 Introduction	34
3.2 Data	36
3.2.1 Digital Traces Data and their Limitations	36
3.2.2 Comparison between LFS Data and Facebook Data	37

3.2.3	Additional Data Sources	38
3.3	Methodology	40
3.3.1	General Model Architecture	40
3.3.2	Measurement Error Models	43
3.3.2.1	Data Assessment of the Labour Force Survey	44
3.3.2.2	Data Assessment of the Facebook Advertising Platform	47
3.3.3	Theory-Based Model	49
3.4	Results	50
3.4.1	Model for Total Numbers	50
3.4.2	Model Disaggregated by Sex	52
3.4.3	Sensitivity Analysis	52
3.5	Discussion	55
3.6	Conclusions	57
4	Extending the Migration Estimation Model including Age and Sex Profiles of Migrants	59
4.1	Introduction	60
4.2	Background	61
4.3	Data	64
4.4	Methodology	64
4.4.1	Fitting the Rogers-Castro Age Schedule	65
4.4.2	The Multinomial-Dirichlet-Dirichlet Model for Age Schedules	69
4.5	Results	71
4.5.1	Model for the Western and Southern European Migrants	73
4.5.2	Model for the Republic of Ireland Migrants	73
4.5.3	Model for the Central and Eastern European Migrants	73
4.5.4	Sensitivity Analysis	73
4.6	Conclusions	75
5	A Brexitodus? Trends in the Numbers of European Migrants in the United Kingdom using Facebook Advertising Platform Data	79
5.1	Introduction	80
5.2	Background	81
5.3	Data	84
5.4	Methodology	86
5.5	Analysis	87
5.5.1	Descriptive	87
5.5.2	Results from the Model	88
5.6	Conclusions	91
6	Conclusions	95
6.1	Summary and Contributions	96
6.1.1	Limitations	98
6.2	Conclusions	98
6.3	Future of Demography	99
	Appendix A Supplementary Materials from Chapter 3	103

Appendix B	Supplementary Materials from the Chapter 4	113
Appendix C	Supplementary Materials from the Chapter 5	127
References		133

List of Acronyms

ACS American Community Survey

AIRE Anagrafe degli Italiani Residenti all'Estero

API Application Programming Interface

CDRs Call Detail Records

COVID-19 Corona Virus Disease 2019

DAUs Daily Active Users

EEC European Economic Community

EU European Union

GDP Gross Domestic Production

IMEM Integrated Model of European Migration

IoT Internet of Things

IPS International Passenger Survey

IQR Interquartile

IUSSP International Scientific Union of Population Studies

JAGS Just Another Gibbs Sampler

LFS Labour Force Survey

MAUs Monthly Active Users

MCMC Markov Chain Monte Carlo

MCMs Multiplicative Component Models

MEM Measurement Error Model

MIMOSA Migration Modelling for Statistical Analyses

NINo National Insurance Number

ONS Office for National Statistics

PERE Padrón Español Residente Extranjeros

PIN Personal Identification Number

SEC Securities and Exchange Commission

SNSs Social Network Sites

TBM Theory-Based Model

THESIM Towards Harmonised European Statistics on International Migration

UK United Kingdom

UN United Nations

USA United States of America

List of Figures

2.1	Net Migration estimates of long-term migrants since 1964 to 2018 (Personal elaboration with ONS data).	12
2.2	Screenshot of the Facebook Advertising Platform on 9 th January 2018. . .	28
2.3	Screenshot of the Facebook for developers web page, in which the definition of DAUs and MAUs are stated (26 th May 2019).	30
3.1	Facebook’s aggregated estimates for the expat and language variables and Labour Force Survey data of migrant stocks from 20 EU countries of origin in 2018 and 2019.	39
3.2	Change in the Facebook algorithm. Magnitude of the decline in Facebook’s estimates of EU migrant stocks in the UK in the middle of March 2019.	39
3.3	Diagram describing the steps (input, data assessment, model, and output) leading to configuring the model and obtain the estimates.	42
3.4	Graphical representation of the adapted IMEM (diagram inspired by Raymer et al. (2013, p. 804)). The hyperparameters are not shown for greater clarity of presentation. Indices: i , sending country; j , sex; t , time. Square nodes represent reported data (z_{ijt}^L, z_{ijt}^F) and covariates. Circle nodes represent parameters for the migration model (see Section 3.3.2) and the measurement model (see Section 3.3.3).	44
3.5	Comparison of Facebook, LFS, and model estimations of EU migrants aged 15+ for the years 2018 and 2019.	51
3.6	Comparison of Facebook, LFS, and model estimations of EU migrants aged 15+ by sex for the years 2018 and 2019.	53
3.7	Comparison between estimates from the first model and the sum of female and male migrants from the second model for 2018 and 2019. . . .	55
4.1	Diagram of the structure of the model.	65
4.2	Rogers-Castro age migration schedule.	66
4.3	Rogers-Castro estimates of the proportions of the EU migrant population living in the UK in 2018 and 2019 by the three country groups identified.	68
4.4	Diagram describing the hierarchical structure of the multinomial-Dirichlet-Dirichlet model. Indices: k , age groups, i , sending country; j , sex; t , time. Square nodes represent reported data ($T_{ijt}^L, T_{ijt}^F, N_{ijt}^L, N_{ijt}^F$). Circle nodes represent the parameters of the model ($a_k, \alpha_{ijt}^L, \alpha_{ijt}^F, \pi_{ijt}^L, \pi_{ijt}^F$, and pi_{ijt}).	70
4.5	Population pyramids from the multinomial-Dirichlet-Dirichlet estimates harmonising the Rogers-Castro estimates from Facebook and the LFS.	72
4.6	Population pyramids comparing the estimates from the model, the LFS and Facebook in the Western and Southern European group.	74

4.7	Population pyramids comparing the estimates from the model, the LFS and Facebook in the Republic of Ireland.	75
4.8	Population pyramids comparing the estimates from the model, the LFS and Facebook in the Central and Eastern European countries group. . .	76
4.9	Comparison between the true stock estimates by sex and the sum of the true proportion by age and sex for 2018 and 2019.	77
5.1	Some major events of Brexit and start of data collection.	82
5.2	Country time series with yearly data from the Labour Force Survey for 2017, 2018, and 2019, and with weekly data from the Facebook Advertising Platform from January 2018 to November 2020.	85
5.3	Country time series with weekly data from the Facebook Advertising Platform from January 2018 to July 2020 by age groups and countries. . .	87
5.4	Country time series with weekly data from the Facebook Advertising Platform from January 2018 to July 2020 by education levels and countries. . .	88
5.5	Values of c, b parameters estimated from the model for age.	89
5.6	Values of the c and b parameters estimated from the model for education.	90
5.7	Values of the c and b parameters estimated from the model for countries.	90
5.8	Values of c, b parameters estimated from the model for the interactions for Poland and Romania.	91
6.1	Updated drawing of a suggested structure of demography combining Kohler & Vaupel and Coleman's contributions in Pavlík (2000)	101
Appendix A.1	The number of Greek migrants in European countries based on Facebook Advertising Platform data and Eurostat data, and the number of Greek-speaking people on Facebook.	104
Appendix A.2	DHARMA of the model presented in Chapter 3	109
Appendix A.3	DHARMA of the Quasi-Poisson Model Specification	109
Appendix A.4	DHARMA of Negative Binomial Model Specification	109
Appendix A.5	DHARMA of the Negative Binomial Model Specification for the Facebook data and Quasi-Poisson for the LFS data	110
Appendix A.6	DHARMA of the "Model without Facebook data" in Table 3.4	110
Appendix A.7	DHARMA of the "Model with Facebook bias at 0%" in Table 3.4	110
Appendix A.8	DHARMA of the "Model with Facebook bias at 11%" in Table 3.4	111
Appendix A.9	DHARMA of the "Model with LFS bias at 4%" in Table 3.4	111
Appendix A.10	DHARMA of the "Model with LFS bias at 30%" in Table 3.4	111
Appendix A.11	DHARMA of the "Model with $\Gamma(1,1)$ " in Table 3.4	112
Appendix B.1	Population pyramids from the multinomial-Dirichlet-Dirichlet estimates harmonising the Rogers-Castro estimates from Facebook and the LFS with different values for the first Dirichlet.	125
Appendix B.2	DHARMA of the model presented in Chapter 4	126
Appendix C.1	Portrayal of the residuals from the first model without interaction terms.	129
Appendix C.2	Portrayal of the residuals from the model specified in Chapter 5, which includes interactions terms for Poland and Romania with age groups and education.	130

List of Tables

2.1	Tables with Facebook categories and variables.	29
3.1	Table summarising the parameters in the measurement error model for the Labour Force Survey and Facebook.	45
3.2	Aggregated estimates of the estimated number of EU migrants in England and Wales by the LFS and the census through which is computed the relative percentage change.	46
3.3	Undercount of the LFS estimates in comparison with the model estimates.	52
3.4	Undercount of the LFS estimates in three different models 1) the model specified only with the LFS data, 2) the model with the Facebook bias parameter set to 0%, 3) the model with the Facebook bias parameter set to 11%, 4) the model with the LFS bias parameter set to 4%, 5) the model with the LFS bias parameter set to 30%, and 6) the model with the $\text{Gamma}(1,1)$ distribution.	54
5.1	Distribution of the main effect of b and c	89
Appendix A.1	Posterior characteristics of the coefficients of the true stock estimates, y , in the first model for 2018 with \hat{R} and \hat{n}_{eff}	105
Appendix A.2	Posterior characteristics of the coefficients of the true stock estimates, y , in the first model for 2019 with \hat{R} and \hat{n}_{eff}	105
Appendix A.3	Posterior characteristics of the coefficients of the true stock estimates, y , in the second model for 2018 with \hat{R} and \hat{n}_{eff}	106
Appendix A.4	Posterior characteristics of the coefficients of the true stock estimates, y , in the second model for 2019 with \hat{R} and \hat{n}_{eff}	107
Appendix B.1	Posterior characteristics of the coefficients of the true stock estimates by age and sex for France by years with \hat{R} and \hat{n}_{eff}	114
Appendix B.2	Posterior characteristics of the coefficients of the true stock estimates by age and sex for Germany by years with \hat{R} and \hat{n}_{eff}	115
Appendix B.3	Posterior characteristics of the coefficients of the true stock estimates by age and sex for Ireland by years with \hat{R} and \hat{n}_{eff}	116
Appendix B.4	Posterior characteristics of the coefficients of the true stock estimates by age and sex for Italy by years with \hat{R} and \hat{n}_{eff}	117
Appendix B.5	Posterior characteristics of the coefficients of the true stock estimates by age and sex for Latvia by years with \hat{R} and \hat{n}_{eff}	118
Appendix B.6	Posterior characteristics of the coefficients of the true stock estimates by age and sex for Lithuania by years with \hat{R} and \hat{n}_{eff}	119
Appendix B.7	Posterior characteristics of the coefficients of the true stock estimates by age and sex for Poland by years with \hat{R} and \hat{n}_{eff}	120

Appendix B.8	Posterior characteristics of the coefficients of the true stock estimates by age and sex for Portugal by years with \hat{R} and \hat{n}_{eff}	121
Appendix B.9	Posterior characteristics of the coefficients of the true stock estimates by age and sex for Romania by years with \hat{R} and \hat{n}_{eff}	122
Appendix B.10	Posterior characteristics of the coefficients of the true stock estimates by age and sex for Spain by years with \hat{R} and \hat{n}_{eff}	123
Appendix C.1	Posterior characteristics of the coefficients of the intercept of the model, c , with \hat{R} and \hat{n}_{eff}	131
Appendix C.2	Posterior characteristics of the coefficients of the slope of the model, b , with \hat{R} and \hat{n}_{eff}	132

Declaration of Authorship

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:
 - Rampazzo, F., Bijak, J., Vitali, A., Weber, I., Zagheni, E. (2021). A framework for estimating migrant stocks using digital traces and survey data: an application in the United Kingdom. *Demography*. 58 (6): 2193–2218. doi: <https://doi.org/10.1215/00703370-9578562>
 - Rampazzo, F., Bijak, J., Vitali, A., Weber, I., Zagheni, E. (2021). Monitoring the Numbers of European Migrants in the United Kingdom using Facebook Data. Preface XIX 1 Plenary Sessions, 119.

Signed:.....

Date:.....

Acknowledgements

This thesis has been a journey. During these years as a student, I have received help from many people: friends, family and other researchers.

The first people I need to acknowledge for their help are my supervisors and mentors: Jakub Bijak, Agnese Vitali, Ingmar Weber, and Emilio Zagheni. They have all been incredible mentors during this process. I have learned a lot from working and discussing my ideas with them. I have been challenged and guided, and I am especially grateful for their support and humanity during this process. Even the COVID-19 pandemic has not stopped them from providing me with their feedback!

I am grateful to the Economic and Social Research Council (ESRC) for supporting my PhD at the University of Southampton through the South Coast Doctoral Training Partnership. Additionally, I have benefitted from the support of the Max Planck Institute for Demographic Research (MPIDR), who provided me with a scholarship to attend the European Doctoral School of Demography (EDSD). I have been fortunate to be part of the Digital and Computational Demography Laboratory and benefited from the feedback of its members. Being a member of the Centre of Population Change has also given me the opportunity to engage with several researchers at the University of Southampton and its network.

I have to acknowledge the help of several other researchers at the University of Southampton and the MPIDR. I am especially grateful to Jason Hilton for his help getting into JAGS (although I should probably be learning Stan now!) and his comments on my work. Emanuele Del Fava also helped me a lot in this last year of my PhD and provided very useful feedback. I have to thank all the PhD students with whom I shared an office at the University of Southampton, especially to those in office 2043 and the PhD students from the Social Statistics and Demography Department, as well as all the members of the Family Demography lunches.

During my years in Southampton I lived in a lively house with June and Tony Kelpie, Dimitra Amaxilati and even some squirrels. I have received a lot of support from my friends at the EDSD, with particular thanks to Alyce Raybould. Another special thank you goes to my adoptive family in Stockholm, who have continued to support me from a distance, and my friends in Padova who even provided me with their own Facebook accounts to collect data. Thank you to Mariapaola Ngaradoumbe Nanhorgue, Francesco Di Carlantonio, Sissi Mattiazzo, and Filippo Faccin.

Finally, I could not have completed this thesis without the support of my parents, Patrizia and Alberto, and my brother Pietro, without whom I would not have been able to manage many technical aspects of this thesis (including probably turning on the computer!).

*To my parents,
Patrizia and Alberto*

Chapter 1

Introduction

1.1 Overview

The Digital Revolution has brought opportunities to many different fields. [Mayer-Schönberger and Cukier \(2013\)](#) describes the Digital Revolution as a two-step process. First, a digitisation of our lives occurs with large amounts of data produced by the words we use online, our location, and our social interactions. Secondly, this information becomes accessible through indexes and searches. It is then that this new data becomes useful to researchers and profitable to advertisement markets ([Mayer-Schönberger and Cukier, 2013](#)). [Mayer-Schönberger and Cukier \(2013\)](#) also focused on the view that the “new oil” in capitalism is this digital data which is used to drive decision making in business ([Mayer-Schönberger and Ramge, 2018](#)).

Demographers are not ignoring this “datatification” ([Mayer-Schönberger and Cukier, 2013](#)). Indeed, [Billari and Zagheni \(2017\)](#) have called for attention to the Digital Revolution. The digitisation of our lives is creating new data sources that, in the context of this thesis, are called “digital traces” ([Latour, 2007](#); [Cesare et al., 2018](#)). Despite this attention to digital traces, there is not yet a clear definition. A digital trace is a footprint left by navigating a website, searching something online (e.g making a query online), or calling from a cellphone. These traces are not generated and collected by scientific research, but are created as a result of user interactions and experiences using digital equipment and platforms ([Karanasios et al., 2013](#)). Digital traces can include data derived from social media, financial transactions, as well as Call Detail Records (CDRs) ([Freelon, 2014](#)). This is a broad definition of the majority of data generated by digital mechanisms.

[Cesare et al. \(2018\)](#) addressed the challenges facing demographers in regard to using digital traces. One of the main challenges of this new data source is related to bias and non representativeness. A more thorough discussion of the characteristics of digital trace data can be found in Section 2.4. The main task for demographers is to understand how to measure the bias of these online non-representative sources in order to infer demographic trends for the population ([Zagheni and Weber, 2015](#)). After having understood the biases involved, one possible next step is to combine different data sources so as to extract more information and enhance the existing data. This is an ongoing process in which demographers have started to combine survey data with digital

traces, originally created for marketing, and repurposing them for scientific research (Zagheni et al., 2018, 2017; Gendronneau et al., 2019; Alexander et al., 2020). Many researchers have pointed out that the idea of repurposing data is not new to demography (Billari and Zagheni, 2017; Zagheni and Weber, 2015; Sutherland, 1963). For example, John Graunt's first Life Table (1662) was in fact a reworking of the public health data from the "*Bills of Mortality*" to infer the size of the population of London at the time (Sutherland, 1963).

In this thesis, the use of digital traces for migration research is investigated. In doing so, this work aims to contribute to the "*gradual and incremental process*" (Willekens, 2019, p. 240) of understanding how to combine traditional data sources and digital traces, while acknowledging that using digital traces as a source of migration data on its own right is challenging (Laczko and Rango, 2014). Therefore, the view taken in this thesis is that, for migration in particular, digital traces "*may complement but not replace traditional data sources*" (Willekens, 2019, p. 249). The focus of the thesis is on European migration to the United Kingdom (UK).

There are three main reasons why it is interesting to look at the migration system of the UK. First is the fact that the British Office of National Statistics (ONS) recently reclassified their estimates as experimental statistics in August 2019. This is because their estimates of international migration are based on surveys due to the lack of registers for migrants in the UK. Therefore, the ONS have admitted that their estimates might be inaccurate (ONS, 2019d). Moreover, the scientific literature suggests that the surveys used by the ONS are affected by multiple biases (Coleman, 1983; Kupiszewska and Nowok, 2008; Kupiszewska et al., 2010; Rendall et al., 2003). Among European countries, the UK is an example where there is no "gold standard" in migration data. The second reason is that the UK has experienced an increased positive net migration from European countries in the last two decades (Champion and Falkingham, 2016). Digital traces might provide important insights on trends in the age, sex, and country of origin of migrants by producing a more accurate estimate of European migrants in the UK. Thirdly, it is interesting to study the UK at the moment, and specifically migration to the UK, because of the country's current transition in leaving the European Union (EU), commonly known as Brexit. The uncertainty linked to Brexit might instigate changes

into the migration trends and decision-making process of European migrants currently living in the UK.

To model the bias in current migration estimates, accounting for the limitations of the data sources and the uncertainty of migration, Bayesian inference is adopted. Bayesian methods are gaining traction not only in statistics and demography (Bijak and Bryant, 2016), but also in many other areas of life (Ferrie, 2019). The Bayesian framework has been utilised in this thesis due to its coherent quantification of uncertainty produced by different sources (Bijak, 2010; Raftery, 1995).

This thesis seeks to contribute to the learning process in understanding how to use digital traces in demography by answering the following question: What can digital traces add to the existing ONS migration estimates, in a context where there is no “ground truth” data which model estimates can be validated against?

1.2 Uncertainty

The study of demographic components - fertility, mortality, and migration - is not deterministic. Although demographic change happens at a slow pace, policies, unpredictable events, and shocks might produce a change in population dynamics. Migration is less defined than other demographic events and might happen at a faster pace in comparison to fertility and mortality. Although it is acknowledged that there are many different reasons why people migrate, there is not yet a complete overarching theory of migration which considers all the situations and drivers that influence migration (Willekens, 1994; Bijak, 2010; Willekens, 2018). The United Nations (UN) have started in 2014 to produce projections based on the works of Raftery et al. (2012), Gerland et al. (2014), and Alkema et al. (2015), which take into account different degrees of uncertainty, using a Bayesian framework to forecast the world population. Migration is not yet probabilistic in the UN’s estimates due to problems in creating a global overarching model, however there is work in this direction (Azose and Raftery, 2019). Additionally, there are various other aspects that influence migration, including country’s policies and the politics of migration (Willekens, 1994; Bijak, 2010; Willekens, 2018).

The Bayesian inference attempts to resolve these issues by formally introducing uncertainty through probability distributions (Bijak, 2010; Raftery, 1995). In Bayesian statistics, using the Bayes theorem (Bayes and Price, 1763), the data is introduced in the model as a likelihood; the outcome of the analysis is a probabilistic distribution, called posterior distribution (Bijak and Bryant, 2016). Quantitative statements on the nature of the data, a prior probability density, can be introduced in the model to inform a parameter in a weaker or informative design (Bijak and Bryant, 2016). The prior distribution, however, is controversial because it is subjective to whomever is modelling the data (Raftery, 1995).

In this thesis, Bayesian modelling is used for three reasons. Firstly, it provides a coherent description of the different sources of uncertainty before linking the data together. This means that it is possible to provide a measure of the quality of the data to the model in terms of a probabilistic distribution. This is an advantage when data is missing or incomplete. Secondly, Bayesian modelling provides a formal mechanism for the inclusion of expert's judgements. It includes qualitative opinions of other experts through quantitative estimates of their beliefs and judgements on the data or future scenarios of the phenomenon under investigation (Bijak and Wiśniowski, 2010; Wiśniowski et al., 2013). Thirdly, it is a natural way to treat highly structured multi-level models (Bijak and Bryant, 2016). Complex population processes can be introduced in a hierarchical structure into the model, which might lead to an holistic representation of the process.

1.3 Contribution and Structure

Although digital traces are biased and non-representative, they also provide an opportunity for a more up-to-date picture of current demographic events. This thesis proposes the use of Bayesian methods to combine traditional and new data sources. Bayesian methods are a natural way of integrating different data sources and their inherent uncertainty. This thesis provides guidance on how to tackle the limitations of digital traces. The aim is to provide an example on the use of digital traces in migration research and to contribute to the ongoing learning in migration and data modelling.

The structure of the thesis is as follows. Firstly, Chapter 2 presents an overview of the current state of migration data and digital traces study. Chapter 3, Chapter 4, and

Chapter 5 outline three different examples of how to combine traditional data sources and social media data to estimate migration. Chapter 3 demonstrates how to complement LFS survey data with Facebook Advertising Platform data to estimate the total number of European migrants in the UK in 2018 and 2019. Chapter 4 builds on the results presented in Chapter 3, proposing a disaggregation by age and sex of the model estimates. Chapter 5 looks at European migration trends from March 2019 to March 2020; using weekly Facebook advertising data. Chapter 6 then comments on the contribution and the limitations of this thesis, describing the evolution of demography as a discipline in light of the Digital Revolution.

1.4 Production of this thesis

Several pieces of software have been used to work on this thesis, most of them open-source software. These softwares were used to write and manage the material of this thesis, then to download, store, and organise the data, as well as to model and analyse the data. This thesis was written in \LaTeX using the [University of Southampton template](#). [Overleaf](#), a collaborative cloud-based \LaTeX editor, was used to share this thesis with the supervisory team and to track changes to it. To compile and manage the bibliography, [Zotero](#), a free and open-source reference management software, was adopted.

The Facebook Advertising Platform data was downloaded through [PySocialWatcher](#), a data collector written in Python which accesses the Facebook Marketing API ([Araujo et al., 2017](#)). The *Ethics and Research Governance On-line* of the University of Southampton have approved this project (“Migrants in the UK through Facebook Advertising Platform”). The project ID is “31099.A2”. [cron job](#) was used to schedule and monitor the download of the Facebook Advertising data. Thanks to the scheduling of the download of the Facebook Advertising data, it was possible to continuously download data while receiving emails informing whether the download had completed or failed during the process. Following the University of Southampton protocols, the data was stored on [Microsoft OneDrive](#).

Data management, cleaning, and analysis was completed in [R](#). The data was manipulated with [Tidyverse](#), which contains several R packages for data science. The figures were then produced with [ggplot2](#) ([Grolemund and Wickham, 2016](#); [Healy, 2018](#)). [Just](#)

Another Gibbs Sampler (JAGS) was used to produce the estimate of the Bayesian models of this thesis. JAGS computes Bayesian hierarchical models using the Markov Chain Monte Carlo (MCMC). JAGS follows the syntax of its predecessors, WinBugs/OpenBugs, in writing the code of the models. For this reason, the manual for Winbugs was used to study the models (Ntzoufras, 2011). In order to use R as an interface of JAGS, the R package `rjags` written by Plummer et al. (2016) was adopted. The package `MCMCvis` was also used for the analysis of the MCMC chains.

Chapter 2

Migration and Data: Literature Review

2.1 Context of Migration and the Digital Revolution

In order to determine the importance of new sources of data and the Digital Revolution for migration research, it is important to first examine how migration is currently measured and the available traditional data sources used to estimate numbers of migrants. As the focus of this thesis is on the UK, Section 2.2 reviews the last four decades of migration to the UK, focusing on the latest developments in recent times due to Brexit. Section 2.3 provides a review of how migration is measured in the theories and models of migration, as well as traditional migration data sources. The data sources measuring the UK migration is described in detail in Subsection 2.3.3. The Digital Revolution is then explained in Section 2.4, exploring the link between demographic research and the Digital Revolution. Section 2.4 presents academic demographic studies and literature on the topic. Literature on the use of digital traces has been rapidly expanding in the last ten years (Edelmann et al., 2020) thanks to conferences such as the International Conference of Web and Social Media (ICWSM), the International Conference of Computational Social Science (IC²S²), and the Social Informatics (SocInfo) conference. While the literature is by no means uniform – as there is far more research on migration using digital traces than on fertility and mortality – Section 2.5 presents examples of all three components of demography that harness digital traces data. The chapter ends in Section 2.6 with an introduction to Facebook and the Facebook Advertising Platform.

2.2 Migration to the United Kingdom

In 1979, based on the International Passenger Survey, the UK recorded a positive net migration for the first time since records began, meaning that the number of immigrants exceeded the number of emigrants (Champion and Falkingham, 2016). From the 1980s onwards, but especially from the 1990s, the UK has changed from primarily a country of emigration to a country of immigration. The UK's story of immigration is linked mainly to two political organisations: The Commonwealth and the European Economic Community (EEC)/European Union (EU)¹. Since the 1960s, the UK received many migrants from its former Commonwealth colonies such as India and Pakistan after they

¹Since 1993, the European Economic Community changed its name to European Union (EU) (Hix and Hoyland, 2011).

become independent. The UK joined the EEC in 1973 (Hix and Høyland, 2011) and, in doing so, agreed to the Treaty of Rome (1957) which in Article 3 established freedom of movement for people in the member states. EEC membership has thus contributed to increases in migration to the UK from European countries since 1973 (Alfano et al., 2016).

Boswell and Geddes (2010) define “European mobility” as the migration of European people within Europe’s borders. Once Eastern European countries started to become part of the EU, the possibilities for European mobility increased. In 2004 the Eastern European countries of Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Slovenia and Slovakia, known as the EU8, became part of the EU alongside Cyprus and Malta. Between 2007 and 2015, the most common foreign nationality in the UK was Polish (before returning to Indian) (ONS, 2019c). In 2017, the Office for National Statistics (ONS) reported 9 million non-British born residents in the UK – of which 39.49% were born in a European country. The second most common European nationality in the UK in 2008 was Romanian. During 2017, the population of Romanian nationals in the UK vastly increased from 83,000 to 411,000. Figure 2.1 represents the overall trend of net migration to the UK since 1964, as well as migration from EU and non-EU countries since 1974. This data suggests that EU migration was increasing at the time of the “United Kingdom European Union membership referendum” - Brexit Referendum - (23rd June 2016), and that net migration may be back on a positive trend but has significantly declined since then.

When the UK has left the EU, the rights of the EU migrants coming to the UK are expected to change. Within the currently proposed withdrawal agreement the status and rights of EU migrants in the UK are expected to stay the same until 31st December 2020; after this date the UK will have a new migration policy for EU migrants. This process is still unclear, however, as there is uncertainty on when exactly the UK will completely leave the EU and on what terms. The initial political departure of the UK from the EU was on 31st January 2020. The bureaucratic burden for EU migrants in the UK has already started to change before Brexit is finalised. EU migrants are required to register for “settled status” in the UK, proving they have lived in the UK continuously for five years, and for “pre-settled status” if they have been living in the UK for less than 5 years continuously. In the absence of an official population register, EU migrants have

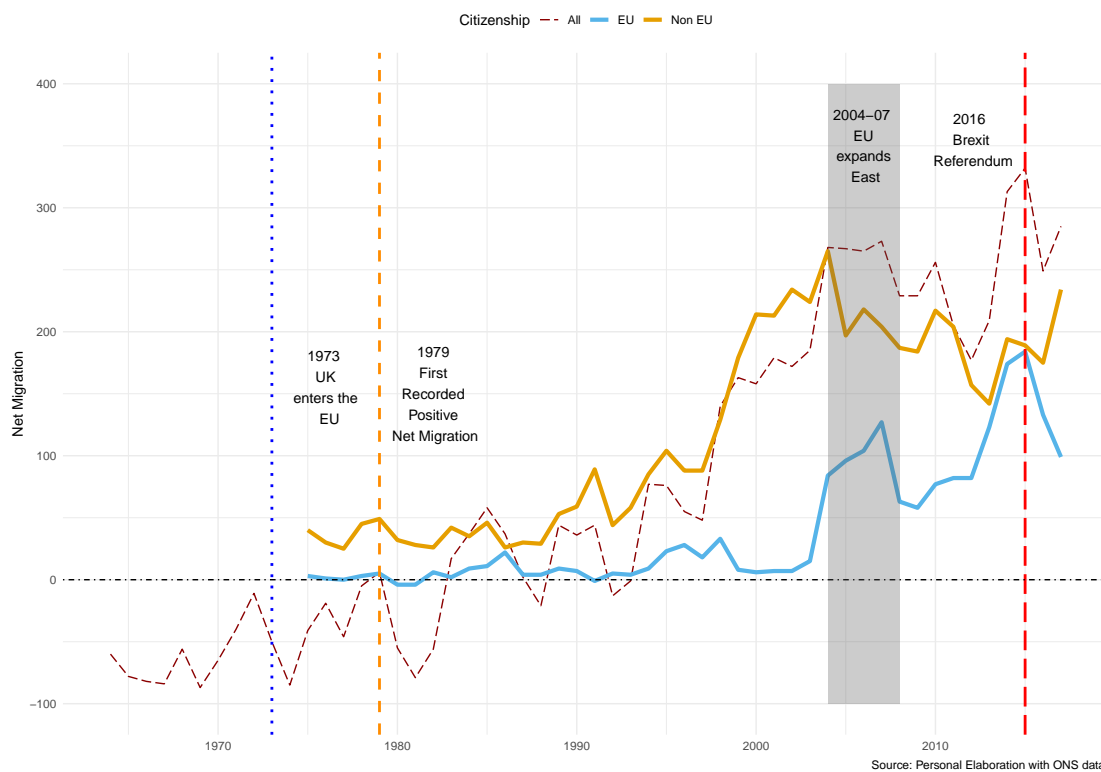


FIGURE 2.1: Net Migration estimates of long-term migrants since 1964 to 2018 (Personal elaboration with ONS data).

a burden on themselves to document their residence in the UK. EU migrants residing in the UK before 31st December 2020 have until 30th June 2021 to apply for the “settled status” or “pre-settled status” registration. Potential future EU migrants will have to apply for visas on a point-based migration system which will end the era of EU free movement to and from the UK.

2.3 Measuring International Migration

Measuring international migration is challenging (Bilsborrow et al., 1997). The lack of timely and comprehensive data about migrants, combined with the varying measures and definitions of migration used by different countries, are barriers to accurately estimating international migration (Bijak, 2010; Willekens, 1994, 2019). Despite the best efforts of many researchers and official statistics offices, it is a considerable problem that migration data sources lack quality in the comprehensiveness of their estimates (Kupiszewska and Nowok, 2008; Zlotnik, 1987; Poulain et al., 2006). Migration is a

topic widely discussed in several research fields including demography, sociology, political science, economics and politics. A lack of quality in data on migration can have a high social and political impact on many areas of life. Moreover, migration, and in particular international migration, has become of increasing importance in shaping population change, particularly in developed countries where fertility is decreasing (Bijak, 2010).

As a concept, international migration is characterised by five building blocks: legal nationality, residence, place of birth, time, and purpose of stay (Zlotnik, 1987). As these blocks are complexly entwined with each other, statistical systems use one or a combination of them to gather data on international migrants. The UN recommends a definition of international migration which explicitly focuses on residence and time (UN, 1998), defining a migrant as a “*person who moves from their country of usual residence for a period of at least 12 months*”. Migrants that stay between 3 and 12 months are considered to be short-term migrants. The intended purpose of the UN’s definition of international migrants is to harmonise data sources worldwide. The ONS takes place of birth and country of legal nationality to mean the “country of usual residence”, and follows the definition recommended by the UN for an international migrant (ONS, 2019a).

As described by Willekens (1994), there are many challenges related to measuring migration. Indeed, there are no adequate data sources to study migration as the existing data is incomplete or inaccurate. A gold standard for migration estimates does not yet exist. In fact, Swedish register data, long considered as the gold standard among demographic datasets, have been proved to overcount migrants (Monti et al., 2019). Given these limitations, Willekens suggested a possible solution; for a more coherent estimate of migration, as most of the data sources are inaccurate, multiple data sources should be combined. Both in 1994 and in 2019, Willekens (1994, 2019) called for the creation of a synthetic migration database, combining data from different sources. The purpose of this database would be to create “*the best possible estimates of the ‘true’ number of migrants*” (Willekens, 2019, p. 235). Managing and monitoring the database will contribute to building a “*learning process*” concerning the model and the data estimating the number of migrants.

2.3.1 Existing Models of Migrations

Migration is not only difficult to measure, but also to explain (Massey et al., 1993; Arango, 2000). In fact, there is no comprehensive migration theory. The drivers of migration have been described from the perspective of multiple isolated disciplines, such as economics, sociology, and geography, but not in a synergistic way. Theories of migration have been criticised as being difficult to operationalise, ambiguous in their applicability to the models and data available (Bijak and Wiśniowski, 2010; Bijak and Czaika, 2020). Overall, migration theories have had limited contribution to the estimation of migrants.

Modelling migration is necessary because of the lack of quality in migration data. The purpose of many existing migration models is to standardise estimates of migration. The oldest model of migration is the gravity model theorised by Ravenstein in the UK (1885). The gravity model explains migration with two factors: population size and distance. The assumption of the model is that the larger the population is and the closer two countries are, the higher the likelihood of migrating. The attractiveness of the two areas can also be included as a factor in the model. Although the gravity model has been widely used (Karemera et al., 2000; Ramos, 2016), it is a deterministic model that does not account for the decision-making process of the migrant themselves or for the uncertainty inherent in the data and estimates. Since the 1980s, other models started to gain popularity. The Poisson model became most commonly adopted in order to account for the discrete nature of the data using a log link function (Willekens, 1983). Other models used corrective factors to harmonise multiple data sources (Poulain, 1993, 1999). The latter models have been applied to migration data at a European level (Abel, 2010; de Beer et al., 2010).

Since the 1990s Europe has become a hub for the development of migration models. Several projects have been funded by the European Commission to create methods to estimate migration. One such project is “*Towards Harmonised European Statistics on International Migration*” (THESIM), which provides an inventory of all the European data sources estimating migrants (Poulain et al., 2006). This was followed by “*Migration Modelling for Statistical Analyses*” (MIMOSA), that proposed the use of Multiplicative Component Models (MCMs) to combine migration flows and data from sending and receiving countries (Raymer et al., 2011). MCMs are used to combine sources to study

elderly migration where data is sparse (Raymer, 2007) as well as to estimate migration in Asia where data on migration flows is considerably limited (Raymer et al., 2019). Additionally, Raymer and colleagues led another project, the “*Integrated Modelling of European Migration*” (IMEM), which proposed a Bayesian hierarchical model that combines multiple migration data sources considering the limitations of each data source (Raymer et al., 2013). The IMEM model is informed by a Poisson distribution, which forms one of the levels in the Bayesian hierarchical model. This thesis’ approach to combining traditional data sources and digital traces data is inspired by the IMEM.

2.3.2 Migration Data

The main data sources used for estimating migrants are censuses, administrative data sources, and surveys (hereafter referred to as ‘traditional data sources’). Traditional data sources have limitations related to the definition of migrants, coverage of the total migrant populations, and the accuracy of the estimates (Azose and Raftery, 2019; Willekens, 2019). Moreover, traditional data sources of migration are not promptly and regularly available. There might be a gap of many months or even years between when the data is collected and when it is released to the public.

While European-wide data sources follow the standard definition of an international migrant (European Parliament and Council of the European Union, 2007) (which is based on the recommended UN definition of international migrant), individual European countries use a variety of systems to track the number of international migrants living within their borders. While censuses are considered the best source of data for estimating migrant numbers, this data has at least three limitations (Willekens, 1994, 2019). First is that census data is collected every ten years, and so they do not provide a timely picture of migration. Secondly, the census records immigrants living in the country, but does not account for the emigrants that have left the country. And lastly, the census does not ask for important data such as the individual’s age at time of migration or return migration.

Administrative data sources, such as population registers, can also be used to estimate migrants. European countries like Italy, France, and Sweden use register permits or a Personal Identification Number (PIN) to register migrants (Poulain et al., 2006). These

register permits, however, are only given to migrants who register themselves in the country of new residence when they have lived there for more than twelve months. Moreover, there is an issue with the lack of de-registration which can affect this data source, as highlighted in Sweden (Monti et al., 2019). There are only two European countries, Italy and Spain, that keep a register of the emigration of their citizens, in the “*Anagrafe degli Italiani Residenti all’Estero*” (AIRE) and the “*Padrón Español Residente Extranjeros*” (PERE) respectively. Additionally, short-term migration is not tracked in these data sources.

Only a handful of countries use survey data to estimate migrants. The advantage of survey data collected from migrants is that they might provide additional information that is not included in the census or in administrative data sources. However, survey data may have issues related to the coverage of the migrant population. This can be related to the sampling framework as well as the questions asked in the survey. The next section describes the survey-based migration data system that is implemented in the UK.

2.3.3 Survey-based Migration Data in the United Kingdom

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

The IPS has been operating since its introduction in 1961, where it was originally intended to estimate overseas travel and tourism, and provides estimates of inflows and outflows of international migrants. The ONS, however, have suggested that the IPS “*has been stretched beyond its original purpose*” (ONS, 2019g) and should not be used as the only source to estimate international migration into the UK. Due to the COVID-19 pandemic, the survey has also been halted since March 2020 (ONS, 2020a). Furthermore, an additional issue with the IPS is that it measures the respondent’s stated intention to stay in the UK, but not their actual stay, which subsequently has to be adjusted for (Kupiszewska et al., 2010). This is clearly not the best measure to estimate migration

as an individual's intentions might change throughout their stay in the UK, or might not be honestly disclosed by the migrant in the first place. This might lead to either an underestimation or overestimation of migrant numbers. In light of this, the ONS tries to correct the raw IPS data when it is moved to the Long Term International Migrants (LTIM) estimates. The LTIM estimates use multiple data sources, including: the IPS, the LFS, Home Office data on asylum seekers, General Practitioner registration from the Northern Ireland Statistics and Research Agency, and adjustments from the Home Office and the Department for Work and Pensions (ONS, 2020b).

Prior to the discontinuation of data collection due to COVID-19, the IPS interviewed travellers into the UK 362 days a year (every day except Christmas Eve, Christmas Day, and Boxing Day) and covered 90% of passengers travelling to and from the UK (ONS, 2014). The interviews take place in 19 airports, 8 ports and at the entrance and exit of the Channel tunnel. The sample usually consists of 700,000-800,000 interviews, and of these only 4,000 interviews per year are of long-term migrants. The response rate is 80.4% (in 2013), and there is a delay of 11 months before the data is released (ONS, 2014).

The second main data source is the LFS, a Europe-wide quarterly household survey, which aims to estimate labour market conditions such as employment levels. The Annual Population Survey (APS) provides a sample boost in the LFS for the ONS to collect data on the stocks of foreign born and foreign citizens in the UK at a local level. The LFS interviews people that have lived at least six months in the same house, including short-term migrants, but excludes those in communal establishments (ONS, 2019e). Furthermore, the LFS does not ask about a person's intention to stay in the UK, just about the period already spent in the UK. The LFS interviews 41,000 UK households per quarter (ONS, 2018a) and combines the quarterly waves of the LFS, with a sample covering 360,000 individuals and 170,000 households per year, to estimate the number of migrants. The data is released 3 months after the end of the survey (ONS, 2018a).

The limitations of the sampling framework, the systematic bias, and the coverage of both the IPS and LFS have been examined by several researchers (Coleman, 1983; Rendall et al., 2003; Kupiszewska and Nowok, 2008; Kupiszewska et al., 2010). The ONS is aware of the limitations of its approach and is in response coordinating an ambitious plan that aims to use additional administrative data sources to complement the

IPS and LFS to obtain a comprehensive measure of migration, and thus better inform policy makers (ONS, 2018b, 2020d). This process is expected to be completed by 2023.

2.4 Digital Revolution

In present times we are experiencing a third industrial revolution; the sociologist Castells (1997) defined it as the Digital Revolution, and named the period from the start of the new millennium as the Information Age (Castells, 1997). It is not always easy to recognise a change when it is happening contemporaneously in our lives; it is usually easier to understand once it has happened. However, we can already recognise how our lives are altered by new digital technologies. Since the invention of the first personal computer in the 1970s, every decade has seen a new development in computing (Salganik, 2017; Waldrop, 2016). In the 1990s, the World Wide Web was launched by Tim Berners-Lee, providing access to the Internet (Berners-Lee and Fischetti, 2001), followed by the first social network sites (SNSs) in the 2000s. boyd and Ellison (2007) provide an overview of the history of SNSs, where they define an SNSs as web-based services where individuals can create their own profile and use it to connect and interact with other users (boyd and Ellison, 2007). Researchers in social psychology have defined the key elements of an SNS as the profile, the network, and the stream (Ellison and boyd, 2013). SNSs are also in continuous evolution; for example they have developed into social media platforms, which have been defined as SNSs that also have a messaging component in their platform (Bayer et al., 2020).

As a matter of fact, the Digital Revolution goes hand in hand with the invention of computers and the Internet. Microprocessors are getting progressively smaller (Waldrop, 2016) and computers can be anywhere: in our phones, pockets, watches, cars, houses, and more (Salganik, 2017). The Internet of Things (IoT)² has led computers and the Internet to be omnipresent through sensors and software introduced in everyday use items (Atzori et al., 2010). It is even possible to control our house or domestic appliances through our smartphones. The Digital Revolution is not only having an impact on our lives, but also on research. Salganik (2017) describes the paradigm shift from analogue to digital research as an opportunity to change from custom-made datasets, created

²Industrial devices with chips that connect them to the Internet.

ad-hoc for research, to readymade datasets, waiting to be repurposed by researchers. In this way repurposing complements readymade datasets, created by companies or governments, with custom-made datasets. Social science research, however, should not only be motivated by the opportunity to explore new data sources, but also by the ideal to expand understanding of the dynamics of the present society. This approach has started to be developed under the umbrella of *computational social science*, in which social science disciplines are integrated with computer science approaches (Edelmann et al., 2020).

2.4.1 Characteristics of Big Data and Digital Traces

It is a challenge to define big data. De Mauro et al. (2016) tried to address this problem by reviewing the fast growing literature in Data Science and suggesting that big data is characterised by four attributes: high volume, velocity, variety, and analytical methodologies. In the book *“Bit by Bit”* (Salganik, 2017), big data is described as being *“incomplete, inaccessible, non representative, drifting, algorithmically confounded, dirty, sensitive”*, as well as *“big, always on, and nonreactive”*. The latter characteristics listed are helpful to researchers, while the former just create more complexity.

Throughout this thesis, the term *“digital traces”* is used in juxtaposition to big data. Digital traces are described as new forms of data sources, produced by phone records, money transfers, geotagged posts on social media, and more (Latour, 2007; Cesare et al., 2018). The main characteristic of digital traces is that they are fast-moving, meaning that the time window available to collect them is much shorter than for surveys or censuses. However, in contrast to traditional data sources digital traces are non representative; in fact, digital traces sets do not include the entire population. In the introduction of this thesis, a digital trace has been defined as a footprint left by navigating a website, querying a platform, or calling from a smartphone. These traces are not generated and collected within the scope of scientific research, but are created as a product of user interactions and experiences using digital instruments (Karanasios et al., 2013). Digital traces can include data derived from social media, financial transactions, as well as Call Detail Records (CDRs) (Freelon, 2014).

In some ways, digital traces still share some features with the more traditional data sources. This is because most of the existing migration data is a by-product of collecting information for other purposes - be it administrative sources such as registers or surveys like the LFS and IPS - and can only be repurposed by adding migration variables in the form of country of birth or country of nationality. In this thesis, the characteristics of digital traces are defined as:

1. originated by web-based platforms and phones;
2. biased, due to the age and sex structure of its users (of web-based platforms and phones);
3. fragile, because of the environment in which they are created, and the unclear definition of what they are measuring; and
4. not always accessible to researchers.

The first characteristic highlights the fact that digital traces are mostly produced by web-based platforms, like social media websites, that collect the information we share online and aggregate them to be used for advertising purposes. Additionally, the second characteristic suggests that these data sources are non-representative of the entire population (Couper, 2013; Baker, 2017; Amaya et al., 2020). Couper (2013) describes two types of biases in digital traces: selection bias and measurement bias. The selection bias causes issues in making the analysis made from digital traces generalisable to the entire population, while the measurement bias is associated with how users portray themselves on the digital platforms.

Hargittai (2018) analyses the potential bias of different platforms in the USA: Facebook, LinkedIn, Twitter, Tumblr and Reddit. She found that Facebook was the most representative social media platform across education levels and Internet skills, while the other social media platforms were used by smaller and more specific groups of the population. The work of Hargittai builds on Lazer et al.'s (2014) critique of the assumption that we can substitute traditional data sources with big data, as it is problematic without considering the bias of these new data sources. The authors also highlight the issue with algorithm dynamics, the fragile characteristic of digital traces, as companies constantly modify their algorithms and therefore are in full control of the information

researchers ultimately receive. Moreover, especially for young adults, we should consider fake accounts. These can be used to portray the life of pets or fictional characters, or be a “*finsta*”, meaning fake Instagram account, showing the individual’s everyday life to their closest friends (Kang and Wei, 2019). These “fake” accounts are more honest and closer to the reality than their main or “real” accounts which present a built personality made to impress.

Furthermore, digital traces are not always available to academics. Some new companies like LinkedIn, for example, are providing access to their users’ data through the “*Economic Graph Challenge*” competition, in which researchers can submit a proposal and, if selected, use LinkedIn’s data to research computational social science topics. Following this approach, “*Social Science One*”³ tries to create partnerships between academic researchers and businesses. At the moment, it has an active partnership with Facebook, established in April 2018. The initiative is led by Gary King (Harvard University) and Nathaniel Persily (Stanford University). The goal is to give researchers access to Facebook’s micro-level data after having submitted a research proposal. There are significant privacy concerns from this, however, which has created delays in the process. On 13th February 2020 the first Facebook URLs Dataset was made available; “*The dataset itself contains a total of more than 10 trillion numbers that summarize information about 38 million URLs shared more than 100 times publicly on Facebook (between 1/1/2017 and 7/31/2019)*” (Gary and Persily, 2020). A research proposal is needed to apply for access to data like this; this is the first step in analysing large micro-level datasets from private companies. Companies also often control the analysis produced on their data. Researchers using companies’ data have to follow strict contracts on its use and seek approval on the results before publication. The Social Science One initiative is interesting in this regard as it comes with pre-approval from Facebook.

Another interesting new initiative intended to expand the collaboration and exchange of data is the [Data Collaboratives](#) organised by the GovLab and the New York University. This has been very active during the COVID-19 pandemic as a repository of COVID-19 projects and data. For example, Cuebiq, one of the largest owners of mobility data worldwide, provided anonymised and aggregated estimates of their data at small granularity to study the effect of COVID-19 and mobility patterns.⁴ This data

³<https://socialscience.one>

⁴<https://list.data4covid19.org/projects/covid19-mobility-data-collaborative>

was used, for example, by the Northeastern Mobs Lab COVID-19 Mobility to understand how people's behaviours adapted in response to the pandemic.⁵

2.5 Demography and Digital Traces

Being a data-driven discipline, demography is one of the fields of research which can benefit the most from the abundance of digital data. Digital traces are an opportunity to collect data from sources that just a short time ago were not yet available (Alburez-Gutierrez et al., 2019). As mentioned before, there is a problem with the representativeness of the data, but by employing calibration and difference in differences (e.g. a quasi-experimental design) this bias can be taken into account and reduced (Zagheni and Weber, 2015). An example of accurate estimates produced from biased data is provided by Wang and colleagues who used Xbox data to forecast the electoral result of the USA presidential election in 2012 (Wang et al., 2015). For 45 days before the election, less than five questions were asked every day to Xbox users. Among these questions some demographic characteristics were measured, which were used to poststratify the sample. A multilevel regression and post-stratification in a Bayesian framework were used to produce a reliable forecast of the electoral result. Billari and Zagheni (2017) expressed their hope that the Digital Revolution currently in progress will lead to studies at smaller granularities and on topics not yet explored by demography. They stress how important it will be to use digital traces in formal demography using modelling techniques from computational disciplines. In this way, different online platforms might be used to investigate different populations.

Indeed, researchers (Gil-Clavel and Zagheni, 2019; Hargittai, 2018) have shown that different sub-groups of the population in age, sex, and education are represented differently on social media. For example, if we are interested in a highly skilled section of the population we should focus on LinkedIn, while if we want to research a broader and more heterogeneous audience we should consider Facebook (Hargittai, 2018). However, as the website or social media platform used could disappear in the next couple of years, we should not base our models on just that one source; an example of this is the use of Google+ to study migration, which is no longer used widely now, but was

⁵<https://list.data4covid19.org/projects/covid19-mobility-data-collaborative>

ten years ago (Messias et al., 2016). This thesis aims to provide an example of how to combine traditional data sources and digital traces, as well as start the discussion around the development of these methods.

2.5.1 Mobility and Migration studies

New data sources are a gold mine for migration studies because they can address the lack of information which hinders this field of research. Digital traces are quick to collect using Twitter or the Facebook Application Programming Interface (API)⁶, and it is possible to know in real time how many of the users are in a specific location contributing to nowcasting migration (e.g. predicting the present).

A popular, but ultimately unsuccessful, example of an attempt to predict the present was the use of queries from Google through Google Trend to estimate the spread of influenza (Ginsberg et al., 2009). Google Trend has also been used to create economic indicators (Choi and Varian, 2012). This kind of digital data has many benefits; for example, as digital data is geolocated, email location has been used to estimate international migration rates (Zagheni and Weber, 2012). This data is also cheap as we do not need to create new data collection infrastructure when repurposing datasets originally intended for advertising. Furthermore, new data sources can also add insights to expand the working definition of an international migrant.

Current definitions of migrants vary between countries. While they all depend on the time of stay outside of the country of usual residence, definitions are by no means harmonized worldwide (Kupiszewska and Nowok, 2008; Willekens, 1994). Fiorio et al. (2017) highlight the potential of using geotagged Twitter data to investigate short-term mobility and long-term migration. They suggest that digital traces can help refine migration theory and modelling. In addition, this data can be augmented through online surveys; it is possible to survey hard to reach sections of the population that would

⁶Companies can provide data through interrogations to their servers. An API is required for this. An API is the link between us, the client, to the server where the data is stored in a database (Cooksey, 2014; Sloan and Quan-Haase, 2017). The job of the API is to connect to and interrogate a server several times until it creates a file with the requested data. To be able to connect to an API a key authentication is usually needed. This is a long series of letters and numbers that identifies the account querying the API (Cooksey, 2014).

be too difficult and expensive to conduct with a traditional sampling framework. Indeed, Polish migrants have been interviewed on Facebook in Austria, Ireland, Switzerland, and the UK, with the intention of supplementing existing cross-national surveys (Pötzschke and Braun, 2017). In a period of four weeks, 1,100 Polish migrants were interviewed for the small budget of 500 euros. Not only is conducting online surveys cheaper than traditional surveys, they often produce more timely results. This has certainly been proved to be the case also during the COVID-19 pandemic (Perrotta et al., 2020).

Nevertheless, these sources also have limitations to consider. As mentioned, researchers do not have direct access to these new datasets, and need to create partnerships with private companies to gain access. Once access is granted, useful research can be conducted. For example, Blumenstock (2012), in partnership with Rwanda's primary telecommunications operator, was able to obtain the mobile phone records from 2005 to 2008 of 1.5 million mobile subscribers in order to study internal migration within the country. Another example is a study published by the University of Berkeley and Facebook researchers (Chi et al., 2020), which uses microlevel Facebook data to analyse the network of friendship between migrant and non-migrant profiles on the platform. They showed that the Facebook network is strengthened and enmeshed by migrants. Additionally, LinkedIn data can give insights on trends of highly skilled migrants moving to the USA (State et al., 2014), and Web of Science data can be used to follow trends and patterns of scholars' migration across the world (Aref et al., 2019).

The most prominent contribution to demography by Facebook data so far has been in the work of Zagheni et al. (2017). They proposed that migration to the USA can be estimated by combining Facebook's Advertising Platform data with data from the high-quality American Community Survey (ACS). The new data sources were compared with traditional data sources in order to understand the differences between the repurposed dataset (the Facebook Advertising Platform) and the official statistics. Building on this research, Alexander et al. (2020) have tried to combine Facebook data with the ACS' data to nowcast migration in the USA. They use a Bayesian approach to combine the two data sources. The bias adjustment to the Facebook Advertising data is done by recalibration with the ACS data, which provides a representative sample of the migrant population. Facebook might be used to offer a timelier picture of the state

of affairs whereas the ACS takes longer to produce. A similar approach in a difference in differences design has been used by [Alexander et al. \(2020\)](#) to study the effect of Hurricane Maria on the number of people leaving from Puerto Rico. By combining Facebook Advertising data with the ACS data they demonstrated the benefits of using digital traces to study the effect of a natural disaster, as the Facebook Advertising Platform data was much faster than the official statistics in capturing the migrants' outflows from Puerto Rico to the USA. This kind of analysis has also proven to be successful with Twitter data ([Martín et al., 2020](#)) and studying other topics such as out migration from Venezuela using Facebook data ([Palotti et al., 2020](#)) and Twitter data ([Mazzoli et al., 2020](#)) respectively. Many studies using digital data that focus on demographic topics such as forced migration ([Singh et al., 2019](#)) are not yet published in demographic outlets, but instead are in computer science proceedings or journals. Facebook data has also been used in studying the integration of migrants in Germany and the USA ([Dubois et al., 2018](#); [Stewart et al., 2019](#)).

Additionally, official statistics producers are investigating the use of digital traces. The European Commission have published three pieces of research that investigate the effectiveness of inferring migration from a combination of traditional and digital data sources such as mobile phones, social media, and other big data. In the first report the authors described the different data sources and the ways to reduce bias through calibration techniques ([Hughes et al., 2016](#)). The second report reviewed the possibility of using Facebook data to study migration and inform policy ([Spyratos et al., 2019](#)). In the third, a similar model to the one proposed in Chapter 3 of this thesis was applied to estimate European migration stocks across European countries ([Gendronneau et al., 2019](#)). It would seem it is still too early to overarchingly judge the usefulness of digital traces. Continued research is needed to understand how academia can benefit from digital traces, as the methods cannot yet achieve the same high standards as official statistics.

2.5.2 Other Applications of Digital Traces in Demography

Digital traces have started to be used in demography both for explaining fertility behaviours and for estimating fertility events. The Internet has been shown to have three effects on fertility: it provides easy access to information on contraception, a wider

pool to find a partner, and, finally, it allows people to combine family time with work. The latter influence has been demonstrated by [Billari et al. \(2019\)](#) in Germany. Internet data has been used to investigate dating dynamics as well as norms of age at marriage spread in relation to Internet use ([Hitsch et al., 2010](#); [Bellou, 2015](#)). Google Trends and Facebook Advertising data have been used to estimate fertility in the USA and to estimate male fertility around the world respectively ([Billari et al., 2016](#); [Rampazzo et al., 2018](#)). [Ojala et al. \(2017\)](#) contributed to the literature of socio-economic differences in childbearing by combining data from an old version of Google Trends (Google Correlates) and the ACS. [Markey and Markey \(2013\)](#) used Google keywords to investigate the seasonality of mating-related searches and consequential fertility, while [Reis and Brownstein \(2010\)](#) looked at web searches related to abortion to estimate abortion rates. Twitter data has been used to explore the area of fertility desires and intentions ([Adair et al., 2014](#)), as well as parents' attitudes towards the birth and future of their children ([Mencarini et al., 2019](#)).

Moreover, [Fatehikia et al. \(2018\)](#) have shown using Facebook Advertising data that there are differences between genders in the use of social media around the world; the authors maintain a dashboard⁷ that constantly updates the results on global gender gaps in Internet and mobile access. Mobile access seems to be particularly important for access to information about contraception and therefore reduction of fertility in sub-Saharan Africa, as well as a measure of gender equality when women have access to phones ([Billari et al., 2020](#); [Rotondi et al., 2020](#)). Digital traces have also been used to show the geographical patterns of homosexuality online ([Gilroy and Kashyap, 2018](#)).

All these studies are still at the macrolevel as individual level data is not available for academic research. However, this data can be augmented through surveys on advertising platforms, as in the case of the Global South in the work of [Rosenzweig et al. \(2020\)](#). The area of mortality research is also employing digital traces, including using family tree websites to study life expectancy ([Fire and Elovici, 2015](#); [Kaplanis et al., 2018](#)) and other areas of the Internet to study the causes of death and the health of populations ([Gittelman et al., 2015](#)).

⁷<https://www.digitalgendergaps.org>

2.5.3 Potential of Digital Traces for Demographic Research

Certainly, demographic literature and interest from the academic audience on digital traces is growing. Demography has always been situated between multiple disciplines, and is now approaching the computational social science field. Demographers' desire for representativeness is an opportunity to expand these studies using digital traces to the entire population. As we have seen, the area of research in Demography that has seen the majority of the study with digital traces data is migration, but there are also examples in fertility and mortality.

2.6 Facebook

Facebook is a social media platform that was founded by Mark Zuckerberg in the USA in 2004, accounting for 2.74 billion users worldwide as of 30th September, 2020 (Facebook Inc., 2020c). Facebook's main source of business is online advertising. By the spring of 2004, Facebook had started to show adverts on its website; *"this option was typically used by businesses and educational institutions that were of potential interest to students"* (Brügger, 2015). In order to create a personal account on Facebook, a user needs a valid email address and, since 2006, must be older than 13 years old⁸. The content on Facebook is created by the user, who fills out information related to themselves (e.g. age, gender, education, relationship status, political views, music and films tastes etc.) and then connects with friends and acquaintances online.

In 2007 Facebook launched an initial version of the Facebook Advertising service (Facebook Inc., 2020a). This service then became Facebook Adverts Manager; this website provides freely accessible aggregated Facebook data. It is a targeted advertising platform designed to allow online advertisers (Facebook's main source of revenue) to target a particular audience of people with their advertisements. Facebook suggests that currently *"more than 140 million businesses use our apps every month"* (Facebook Inc., 2020a). Audience targeting is performed based on a range of both self-declared and inferred attributes assigned to Facebook users.

⁸When creating an account on Facebook, the user must provide their date of birth to prove they are older than 13 years old, however, this information is not verified.

To facilitate budget planning, Facebook provides ‘user estimates’ of the potential reach of targeted adverts. The data provides an instantaneous picture of the population at the time of the query, but it is not possible to get a historical picture. Figure 2.2 shows a screenshot of the Facebook Adverts Manager in use. In this example, an anonymised aggregated group of individuals living in the UK between 18 and 30 years old have been selected. In the orange box on the right-hand side of the screenshot, it is possible to see the “potential reach” of the advertisement and the “estimated daily results reach”. The “potential reach” is the audience size that is used in this thesis.

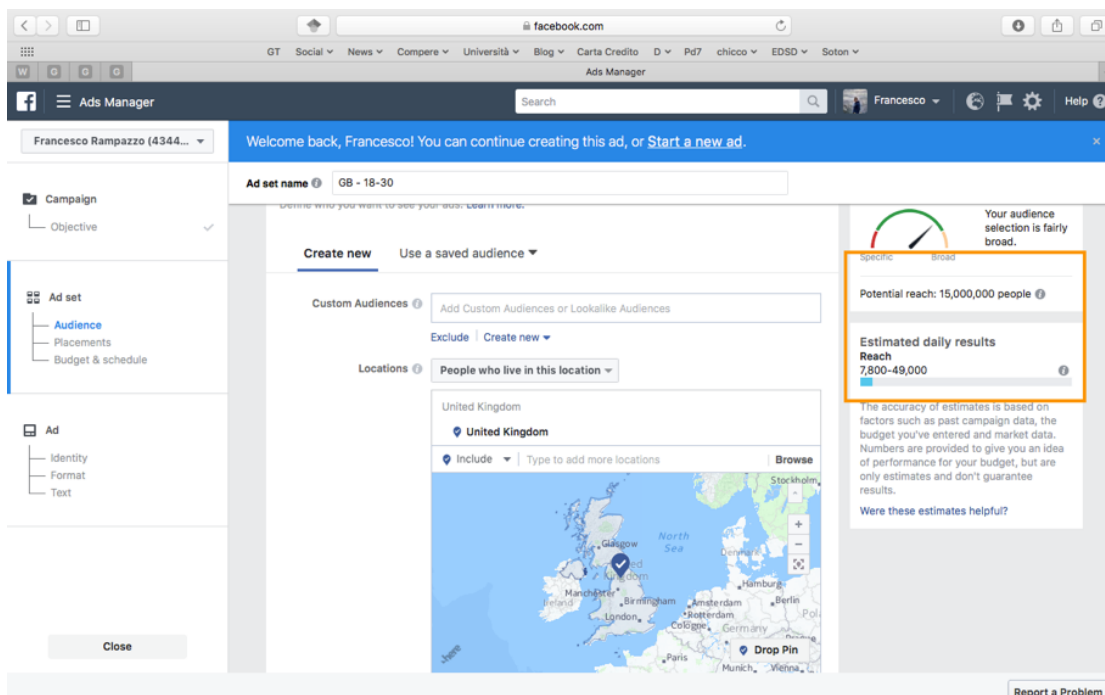


FIGURE 2.2: Screenshot of the Facebook Advertising Platform on 9th January 2018.

In this thesis, the category “*Ex-pats(*)*” is selected on Facebook Adverts Manager in order to receive an anonymous aggregated estimate of the immigrant population within a country. This category is used to estimate the number of European immigrants in the UK which will then be stratified by a selection of variables, namely age, gender, and education.

The Facebook Advertising Platform is rich with information. The variables are categorised as Location, Demographics, Interests, and Behaviours⁹. Table 2.1 shows that these four macro categories have multiple sub-categories that provide more detailed information with further supplemental sub-categories within these.

⁹This list follows the categories found on the Facebook Advertising Manager on the 12th February 2020.

TABLE 2.1: Tables with Facebook categories and variables.

Location	Demographics	Interests	Behaviours
People living in or recently in this location,	Education,	Business Industry,	Anniversary,
People living in this location	Financial,	Entertainment,	Consumer classification,
People recently in this location,	Life Events,	Family and relationships,	Digital activities,
People travelling in this location	Parents	Fitness and wellness,	Ex-pats,
	Relationship	Food and drink,	Mobile Device Users,
	Work	Hobbies and activities,	Mobile Device Users/device use time,
		Shopping and fashion,	More Categories,
		Sports and outdoors,	Multicultural affinity,
		Technology	Politics (US), Purchase behaviour, Ramadan (Month), Soccer, Travel

To download the data from the Facebook API a Python package, [pySocialWatcher](#), was used (Araujo et al., 2017). To download Facebook data, a Facebook account is required, as well as a Facebook advertising account and its ID in order to create an app. It is also

possible to download the data with R through `r-estimates-fb-ads` Gilroy and Kashyap (2018), or with `Using-Facebook-API` by Gil-Clavel and Zagheni (2019). All the packages are available with documentation on GitHub.

2.6.1 Measurements in the Facebook Advertising Platform

Facebook Marketing API provides two usage metrics: Daily Active Users (DAUs), and Monthly Active Users (MAUs). On Facebook for developers DAUs is defined as the “estimated number of people that have been active on your selected platforms and satisfy your targeting spec in the past day”, while the MAUs is the “estimated number of people that have been active on your selected platforms and satisfy your targeting spec in the past month”. Figure 2.3 shows a screenshot of the Facebook for developers web page.

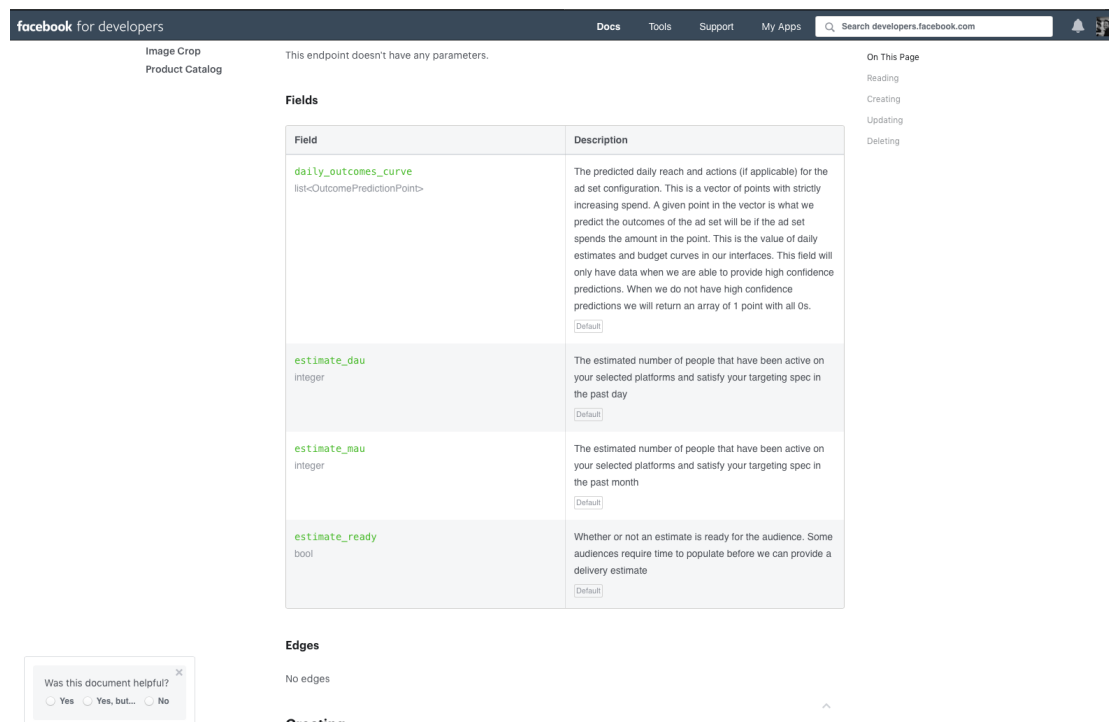


FIGURE 2.3: Screenshot of the Facebook for developers web page, in which the definition of DAUs and MAUs are stated (26th May 2019).

Until March 2018 the Facebook Marketing API only provided the MAUs estimates. Since then, estimates of the DAUs are also provided. Additionally, there has been a change in the granularity of the estimates provided; previously the smallest integer in the MAUs was 20 profiles, whereas now it is 1000. This aggregation is implemented to avoid reidentification of individual users. The DAUs can reach smaller granularities,

including a count of zero. In this thesis, the MAUs estimates are used in Chapter 3, Chapter 4, and Chapter 5.

Chapter 3

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

3.1 Introduction

In recent years, scholars have started using Bayesian methods to combine different sources of migration data in order to provide better estimates of the migrant stock; the total number of migrants present in a country at a certain date (Azose and Raftery, 2019). In this chapter, the aim is to improve estimates by complementing survey data with social media data. This is important as, when designing migration policies, it is crucial to have access to valid sources of data on international migration. This chapter proposes the use of a Bayesian data assessment model that combines data from the Labour Force Survey (LFS) and the Facebook Advertising Platform to assess the number of European migrants in the United Kingdom (UK). The aim is to demonstrate how such a model can produce a more accurate estimate of European migration. The UK is used as an example in this study, as it is a Western country for which the migration data is of poor quality.

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

At the same time, a new strand of research has emerged recently that has been repurposing digital data to complement traditional demographic data sources, and to improve their coverage and timeliness of production. Since digital traces data are often geolocated, migration has received particular attention in this literature. As Cesare et al. (2018) have suggested, using digital traces data sources has advantages, such as

the speed and low cost of data collection, but also limitations, with issues in the lack of accessibility, transparency, and representativeness. Drawing on data from the Facebook Advertising Platform and the LFS, this chapter investigates whether the digital traces that individuals leave on Facebook can be used to estimate stocks of migrants in the UK.

This is by no means the first study that has tried to combine digital traces with survey data (Zagheni et al., 2018; Alexander et al., 2019, 2020). However, in this chapter, an overarching framework is proposed, for the first time, including both a theoretical model that takes into account push and pull factors related to migration theories as well as a data assessment model that aims to reduce the bias of the data used in the model. This framework provides a more context-specific model for examining migration to the UK from several sending countries. Moreover, this chapter provides important insights into the complex reality of international migration to the UK by shedding light on the demographics of migrants by country of origin, which are hard to obtain using currently available official statistics. The attention is limited to migrants from European countries because, in the UK context, these migrant stocks are the hardest to estimate due to the EU's "freedom of movement". At least until December 2020, there is no requirement for EU migrants to register their residence in the UK. Thus, up to now, survey data has been used to estimate the stock of migrants from the EU. The aim of this thesis is to complement these existing, but incomplete, official estimates of migrant stocks by analysing digital traces. As a result, an estimate of the total number of EU migrants in the UK for 2018 and 2019 is produced.

There are two additional reasons why it is interesting to look at the migration system of the UK. First is that the UK Office of National Statistics (ONS) bases its estimates of international migration on surveys. In August 2019, the ONS reclassified their estimates as experimental statistics, emphasising that the estimates might be inaccurate (ONS, 2019d). Furthermore, the scientific literature has suggested that these surveys are affected by different sources of bias (Coleman, 1983; Kupiszewska and Nowok, 2008; Kupiszewska et al., 2010; Rendall et al., 2003). In Europe, the UK is an example of a country in which there is only a "bronze standard", meaning that the UK migration data sources are inferior to the "gold standard" but are of "sufficient quality for validation" (Azose and Raftery, 2019). Secondly, although the UK has experienced a

net positive increase in migration from European countries over the past two decades (Champion and Falkingham, 2016), the ONS reported an undercount of 16% for the net migration estimates for the EU8 countries (Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Slovakia, and Slovenia) in 2016, suggesting that the relevant migration statistics are of insufficient quality (ONS, 2019f). Using digital traces might provide insights into UK migration trends in sex and country of origin by enabling researchers to produce estimates of stocks of European migrants in the UK.

3.2 Data

3.2.1 Digital Traces Data and their Limitations

The variable that is used to estimate international migrants is defined by Facebook as “People that used to live in country x and now live in country y ”. This variable was first used by Zagheni et al. (2017) and was compared to data from the American Community Survey (ACS). Facebook’s definition of a migrant has changed over time, as until December 2018 the variable was called “Expatriate from country x ”. Facebook’s documentation, however, does not provide information on which individual characteristics have been used to create the variable, or on whether the algorithm identifying a user as a migrant was changed along with the re-wording of the definition in 2018. Two studies have investigated how Facebook processes this category. In the first, researchers at Facebook suggested that Facebook users are considered “expats” based on the location of their hometown and the structure of their friendship networks (Herdağdelen et al., 2016). In the second study, Spyratos et al. (2019) ran a survey in which 114 Facebook users were asked whether the Facebook Advertising Platform identified them as an “expat”. They concluded that Facebook uses other types of information that are not specified in the users’ profiles to categorise them, such as geolocation outputs. The final clue can be found in Facebook’s form “10-K”, which is a “USA Securities and Exchange Commission” (USA SEC) documents that provides a summary of Facebook Inc.’s financial performance on the stock market. In these documents, Facebook wrote that “the geographic location of our users is estimated based on a number of factors, such as user’s IP address and self-disclosed location” (USA SEC, 2018, 2019). In this chapter, the additional variable of “language” from the Facebook Advertising Platform is included.

Facebook reported that it is possible to “*target people with language other than common language for a location*”¹.

The same USA SEC documents (USA SEC, 2018, 2019) reported estimates of the bias of these MAUs statistics, finding that in 2018 and 2019 11% of Facebook accounts were duplicated and another 5% were fake accounts. Most of these anomalies were detected in south east Asia. This chapter uses the MAUs estimates because Facebook’s documentation makes clear that this measure is more stable than the DAUs metric. The MAUs metric does not report numbers under 1000 to prevent the targeting of very small groups of individuals. Through the Facebook Marketing API, this thesis includes all Facebook users in an aggregated and anonymised format.

It would seem, however, that Facebook’s coverage of the general population varies by age and gender. Pew Research (2018) reported that while Facebook is used across all age groups, the numbers of younger users on Facebook has been declining. Nevertheless, Facebook has noted that some younger users register on Facebook with an inaccurate age (USA SEC, 2018, 2019). In addition to the age composition of Facebook users, the coverage differences between men and women should be considered. Fatehkia et al. (2018), and Garcia et al. (2018) explored patterns in the use of Facebook to describe the digital gender gap. While the gap is growing smaller, there are still more men than women using Facebook (Fatehkia et al., 2018).

3.2.2 Comparison between LFS Data and Facebook Data

In this chapter, the two main data sources used are the LFS and the Facebook Advertising Platform. Twenty countries are included in the study: Austria, Belgium, the Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, the Netherlands, Poland, Portugal, Romania, Slovakia, Spain, and Sweden. Malta and Luxembourg have been excluded because of their small size which breaches the confidentiality limit both in the LFS and Facebook data regulations; Bulgaria and Croatia have been excluded because Facebook does not provide estimates of expat numbers for them; and Estonia and Slovenia have been excluded as they have missing values in the covariate variables’ data. The aggregated estimates of European

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

migrants from the Facebook Advertising Platform were collected in the third week of July 2018 and 2019. *pySocialWatcher*, a Python package, was used to download the data (DAUs and MAUs) from the Facebook API (Araujo et al., 2017). The data from the LFS was provided by the ONS for the period of June to July in both 2018 and 2019.

Figure 3.1 shows a comparison between these two data sources for the two years included in the analysis. Three variables from the data sources are shown: the migrant variable and the language variable from Facebook, and estimates of migrant stocks by country of birth from the LFS. Looking at the figure, a correlation can be seen between the Facebook migrant variable and the Facebook language variable for many countries. The correlation between the Facebook expat variable and the Facebook language variable is 0.92 for both years, while the correlation between the Facebook migrant variable and the LFS estimates is 0.91 in 2018 and 0.88 in 2019. However, there are exceptions for:

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

Figure 3.1 shows a visible drop in the Facebook migrant variable estimates between 2018 and 2019. This is not due to out-migration from the UK, but rather because of an algorithm change that affected the Facebook estimates. Figure 3.2 highlights the shift that occurred in the middle of March 2019, which led to an average change in the estimates of 48%. The impact of the change was both country-, age-, and sex-specific.

3.2.3 Additional Data Sources

In this analysis, additional sources of migration data are used as covariates that can help us estimate migrant stocks. Data from the IPS on the inflows and outflows of migrants for 2017 and 2018 are considered. Furthermore, information on the populations of the countries of origin from the projections produced by Eurostat, together with the

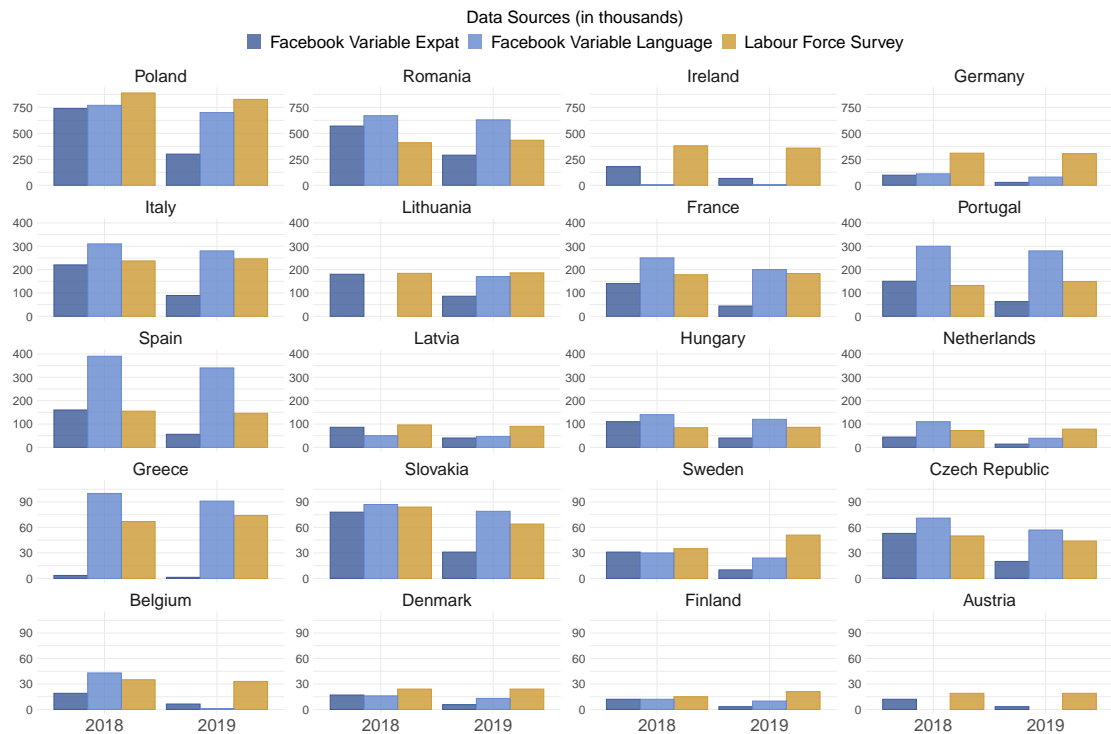


FIGURE 3.1: Facebook’s aggregated estimates for the expat and language variables and Labour Force Survey data of migrant stocks from 20 EU countries of origin in 2018 and 2019.

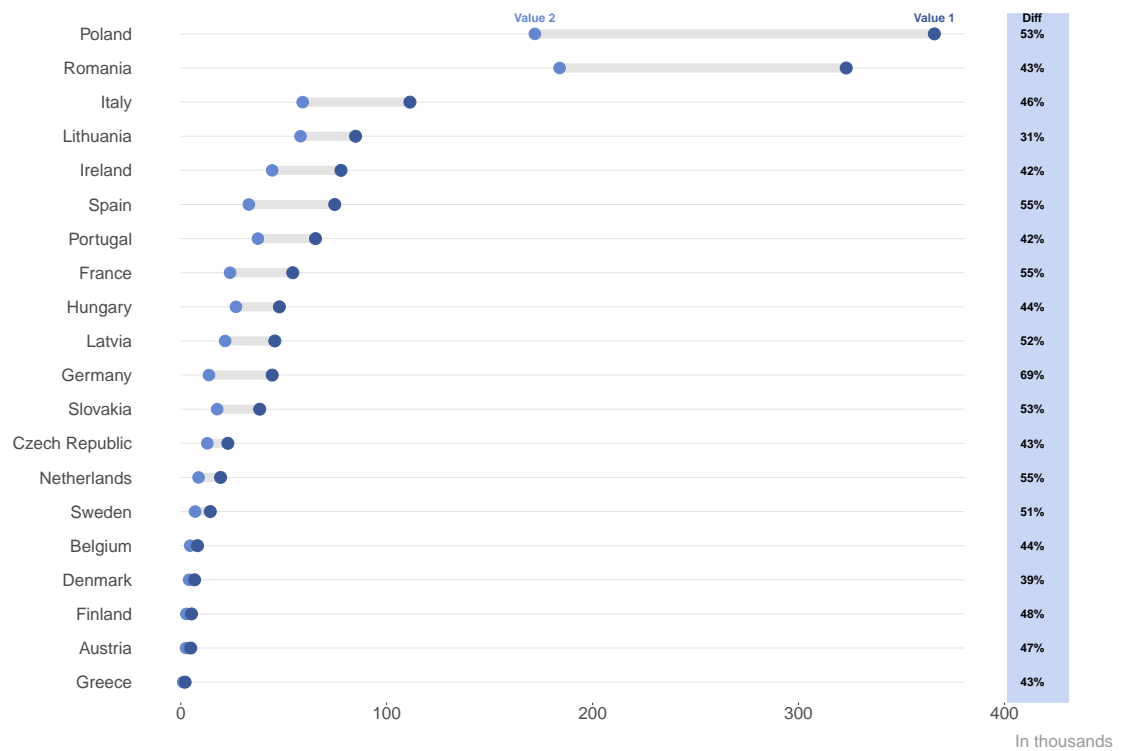


FIGURE 3.2: Change in the Facebook algorithm. Magnitude of the decline in Facebook’s estimates of EU migrant stocks in the UK in the middle of March 2019.

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

Data from the UK settled and pre-settled status scheme is added to make an additional comparison. This scheme allows European migrants already residing in the UK to apply for pre-settled status if they have been living in the UK for less than five years, and for settled status if they have been living in the UK for five or more years. The measure of applications to the scheme provides an indication of the number of Europeans who want to continue to have the right to remain in the UK after Brexit has been finalised. The data represents an estimate for the total number of applications, and includes data from 28th August 2018 to 31st December 2019.

3.3 Methodology

3.3.1 General Model Architecture

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

The model is divided into two parts: The Measurement Error Model (MEM) and the Theory-Based Model (TBM). In the MEM, the Facebook Advertising Platform and LFS

data are combined; while in the TBM, other variables are also considered in the estimation of the true stock. In this framework, the IMEM quantifies the limitations of the data sources and provides the appropriate prior distribution in order to reduce the bias.

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

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

In Figure 3.3, the model is explained using a diagram that is divided into four parts: input, data assessment, model, and output. In the input column, the data sources are presented as being survey data, digital traces, and migration theory covariates for the TBM. The data assessment is followed by a summary of the limitations of the data in terms of definition, bias, and coverage. In the model box, the true stock at the centre of the figure is estimated by the TBM and the MEM, which combine the stock estimates from the LFS with those from the Facebook Advertising Platform, while incorporating considerations related to definition, bias, and accuracy. Finally, in the output, the diagnostics and results are shown.

The model is constructed as follows. The number of European migrants (stocks), z_{ijt}^k , from a certain country, i , in the UK with a certain characteristic, j , is observed. In this case the characteristic selected is sex. This is done using data from Facebook, F , and from the LFS, L , and the value k is then used to represent either L or F depending on which data is used to measure the European migrants stock (z^k). The year, t , in this case is 2018 and 2019.

The datasets used can thus be described in the form of matrices Z^F for Facebook, and Z^L for the LFS. The model borrows strength across the two years.

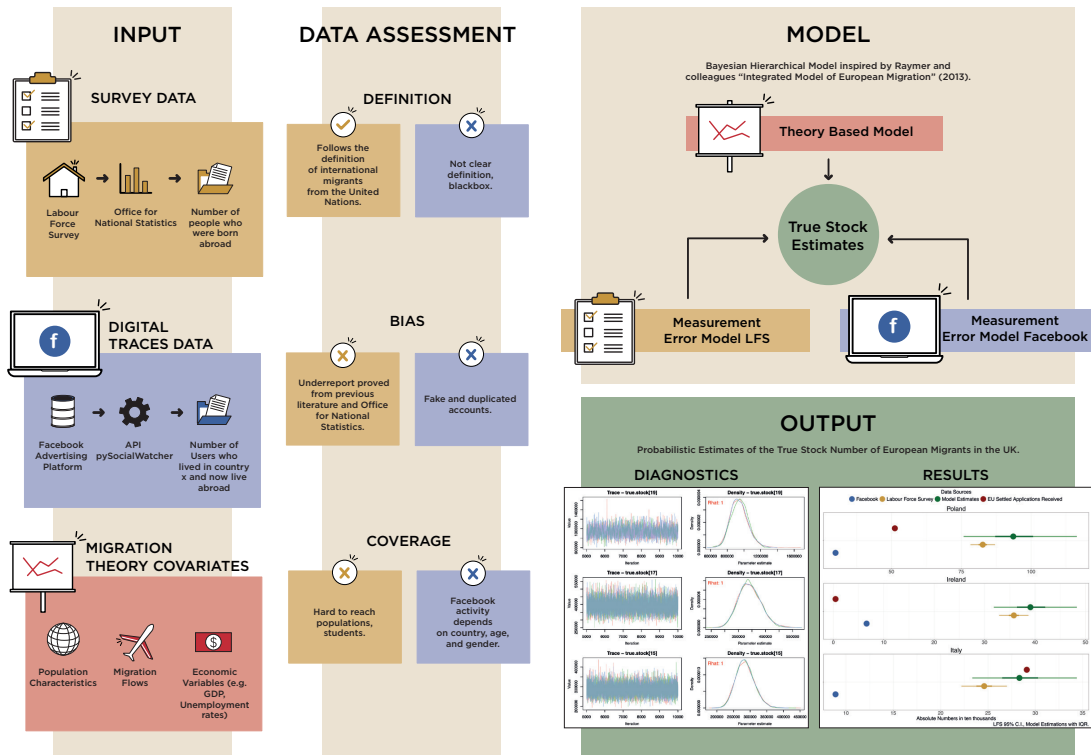


FIGURE 3.3: Diagram describing the steps (input, data assessment, model, and output) leading to configuring the model and obtain the estimates.

$$Z^F = \begin{pmatrix} z_{11t}^F & z_{12t}^F & \cdots & z_{1Jt}^F \\ z_{21t}^F & z_{22t}^F & \cdots & z_{2Jt}^F \\ \vdots & \vdots & \ddots & \vdots \\ z_{i1t}^F & z_{i2t}^F & \cdots & z_{iJt}^F \end{pmatrix} \quad (3.1)$$

$$Z^L = \begin{pmatrix} z_{11t}^L & z_{12t}^L & \cdots & z_{1Jt}^L \\ z_{21t}^L & z_{22t}^L & \cdots & z_{2Jt}^L \\ \vdots & \vdots & \ddots & \vdots \\ z_{i1t}^L & z_{i2t}^L & \cdots & z_{iJt}^L \end{pmatrix} \quad (3.2)$$

The value Y_{ij} is the random variable estimate of the true stock. It is a matrix with dimension $I \times J$.

$$Y = \begin{pmatrix} y_{11t} & y_{12t} & \dots & y_{1Jt} \\ y_{21t} & y_{22t} & \dots & y_{2Jt} \\ \vdots & \vdots & \ddots & \vdots \\ y_{I1t} & y_{I2t} & \dots & y_{IJt} \end{pmatrix} \quad (3.3)$$

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

$$z_{ijt}^k \sim Po(\mu_{ijt}^k). \quad (3.4)$$

Figure 3.4 intended to graphically illustrate the hierarchical structure of the model. In the next section, the model is explained in detail. The model is estimated using JAGS in R (Plummer et al., 2016). In JAGS, the normal distributions are defined in terms of the mean, μ , and precision (i.e. one over the variance), τ . The JAGS notation is used.

3.3.2 Measurement Error Models

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

$$\log \mu_{ijt}^k = \log y_{ijt} + \delta^k + \beta^k + \chi_{ij}^k + \zeta_{ijt}^k + \lambda_{ijt}^k + \epsilon_{ijt}^k \quad (3.5)$$

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

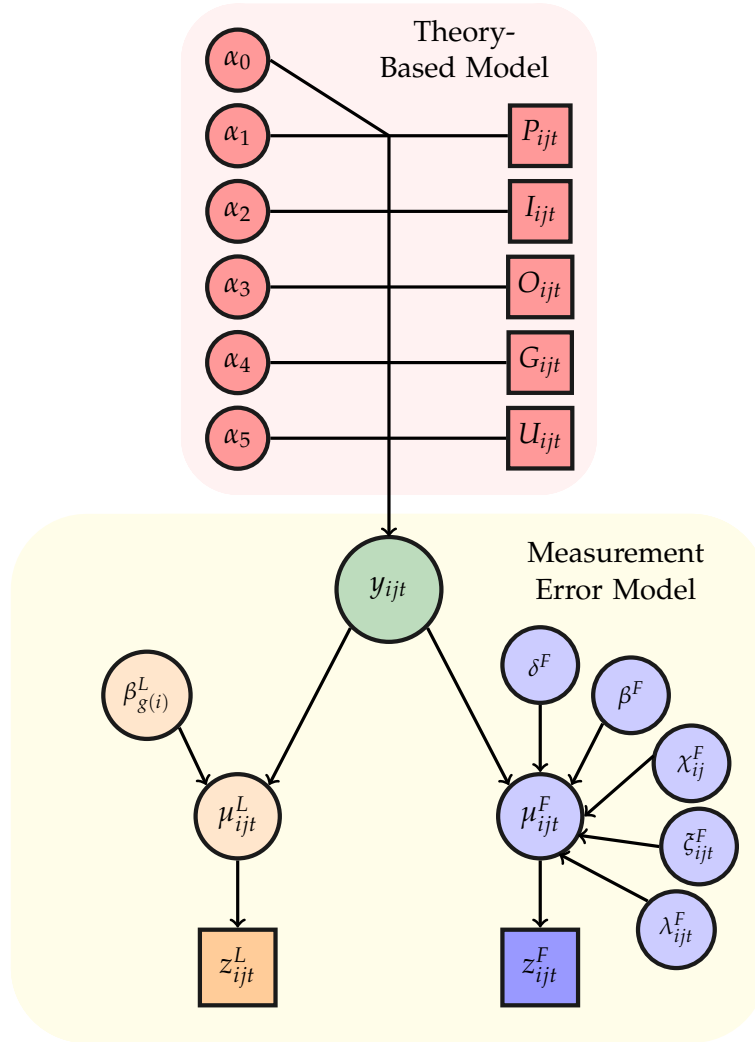


FIGURE 3.4: Graphical representation of the adapted IMEM (diagram inspired by Raymer et al. (2013, p. 804)). The hyperparameters are not shown for greater clarity of presentation. Indices: i , sending country; j , sex; t , time. Square nodes represent reported data (z_{ijt}^L, z_{ijt}^F) and covariates. Circle nodes represent parameters for the migration model (see Section 3.3.2) and the measurement model (see Section 3.3.3).

certain language. The term ϵ_{ijt}^k is the error term with normal distribution $N(0, \tau_{ijt})$, the precision τ_{ijt} has Gamma distribution $G(100, 1)$, (where 100 is the shape parameter and one is the rate parameter) which has a mean equal to 100 and precision equal to 1 (e.g variance equal to 100). Table 3.1 summarises the parametrisation of the model and the direction of the prior distributions.

3.3.2.1 Data Assessment of the Labour Force Survey

The LFS defines a long-term international migrant in the same way as the UN (ONS, 2018a), and provides data on each migrant's country of birth and citizenship. For the

TABLE 3.1: Table summarising the parameters in the measurement error model for the Labour Force Survey and Facebook.

Measurement Error Model			
Parameter	Interpretation	Labour Force Survey	Facebook
δ	Definition		Unknown definition, but with some variation
β	Bias	Inflation of the estimates $\left\{ \begin{array}{l} 4\% \text{ undercount low} \\ 12\% \text{ undercount medium} \\ 30\% \text{ undercount high} \end{array} \right.$ $+$	Deflation of the estimates $- 4\% \text{ fake, duplicates}$
χ	Coverage		\pm coverage by sex in the home country
ξ	Algorithm Change		\sim effect of an algorithm change in 2019
λ	Language Parameter		\sim Greek language dummy parameter

purposes of this paper, the country of birth criterion is used because it captures individuals with a migrant background, including those who acquired citizenship through naturalisation. Since the LFS is used to estimate the stock of migrants in the UK, many researchers have investigated the quality of the survey's estimates and have found that they underestimate migrants. [Rendall et al. \(2003\)](#) for example, reported that the 2001 LFS under-reported international migrants by 26% compared to the 2001 census. Other research has shown that the bias in the LFS might be as high as 30% for nationalities with smaller stocks, such as Greeks and Lithuanians ([Kupiszewska et al., 2010](#)), and that the survey has a non-response rate of over 15% ([Martí and Ródenas, 2007](#)). Furthermore, the sampling framework of the LFS does not cover the entire target population ([Kupiszewska et al., 2010](#)) as students and more mobile migrants might not fully appear in the sample.

Table 3.2 compares data from the LFS collected between January and December 2011 with the British census that occurred on 27th March 2011. The data is aggregated for England and Wales only. It reveals the relative percentage change between the LFS and the Census. The relative percentage change gives a sense of the bias between the LFS and the census. It has to be stressed that the ONS has already attempted to recalibrate the LFS estimates with the results of the census. Despite this, there is still a problem with both undercounting and overcounting. The range of the bias is between -21% and 15%. This issue suggests the LFS Measurement Error Equation to be:

$$\log \mu_{ijt}^L = \log y_{ijt} + \beta_{g(i)}^L + \epsilon_{ijt}^L \quad (3.6)$$

TABLE 3.2: Aggregated estimates of the estimated number of EU migrants in England and Wales by the LFS and the census through which is computed the relative percentage change.

	LFS January - March 2011	Census March 2011	Relative Percentage Change
Austria	19000	19087	-0,46
Belgium	28000	25472	9,03
Czech Republic	37000	37150	-0,41
Denmark	18000	21445	-19,14
Finland	10000	12149	-21,49
France	134000	129804	3,13
Germany	279000	273564	1,95
Greece	33000	34389	-4,21
Hungary	44000	48308	-9,79
Ireland	353000	407357	-15,40
Italy	121000	134619	-11,26
Latvia	57000	54669	4,09
Lithuania	115000	97083	15,58
Netherlands	52000	59081	-13,62
Poland	572000	579121	-1,24
Portugal	83000	88161	-6,22
Romania	94000	79687	15,23
Slovakia	52000	57824	-11,20
Spain	69000	79184	-14,76
Sweden	30000	30694	-2,31

Note: The relative percentage change is computed from the LFS data from January to December 2011 and the census in 2011. The LFS data available for January to March 2011 is already recalibrated through 2011 census data.

In this chapter, the LFS measurement error equation is assumed to be:

$$\log \mu_{ijt}^L = \log y_{ijt} + \beta_{g(i)}^L + \epsilon_{ijt}^L \quad (3.7)$$

As for this assessment, the LFS data is deflated only by one parameter, β^L , which considers both the bias and the coverage of the data. A separate parameter, such as δ^L , is redundant as the definition of international migrant in the LFS follows the UN standard. The literature (Rendall et al., 2003; Kupiszewska et al., 2010; Martí and Ródenas, 2007) suggests that for countries with small migrant populations in the UK, LFS migrant estimates may be around 30% lower than the true numbers. This percentage is

reduced, at around 15%, for those nationalities with large populations in the UK. Table 3.2 provides a measure of the bias at a country level. The ONS reports that the quality of the LFS estimates decreases over time when distanced from the census year (ONS, 2020d). The classification relies on the literature, the data from Table 3.2 as well as assessment from the ONS and our (my and supervisors) own expertise. The LFS bias is anchored to the relative percentage change between the LFS and the census, and an increase of bias over time is also considered. As a matter of fact, the countries are divided into three groups:

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

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

$$g(i) = \begin{cases} 1, & \text{if the undercount is assumed to be low;} \\ 2, & \text{if the undercount is assumed to be medium;} \\ 3, & \text{if the undercount is assumed to be high.} \end{cases} \quad (3.8)$$

The prior distribution is set to:

$$\beta_i^L \sim \begin{cases} N(-0.04, 100), & \text{if the undercount is assumed to be low;} \\ N(-0.13, 100), & \text{if the undercount is assumed to be medium;} \\ N(-0.35, 100), & \text{if the undercount is assumed to be high} \end{cases} \quad (3.9)$$

The term ϵ_{ijt}^k is the error term with normal distribution $N(0, \tau_{ijt})$, and the precision τ_{ijt} has Gamma distribution $G(100, 1)$, as previously described.

3.3.2.2 Data Assessment of the Facebook Advertising Platform

Given the description of the Facebook data in the data section, a parameter was created for both the definition, bias, and coverage of the Facebook data. The Facebook δ^F is a

priori assumed to be normally distributed with $N(0, 100)$, while β^F has a normal distribution $N(0.04, 100)$. The mean of β^F is set at 4% to deflate the Facebook estimates in order to account for fake and duplicate accounts. This value is lower than the 11% suggested by Facebook themselves, because it is assumed that the percentage of fake and duplicated accounts labelled as belonging to migrants is lower in Europe. The mean of the coverage parameter χ_{ijt}^F is the rate of non-Facebook users in the country of origin of the European migrants, since the aim is to correct by this adjustment. It is computed as:

$$\chi_{ijt} = \log \left(1 - \frac{\text{Number of Facebook Users}_{ijt}}{\text{Eurostat Population Size}_{ijt}} \right) \quad (3.10)$$

Additionally, the digital traces data is described as unstable. Indeed, it seems that Facebook reviewed its algorithm on expats in the middle of March 2020, and there was a drop in the migrant estimates after this time. The change is country- and sex-specific. For this reason, a parameter was introduced for the rate algorithm ζ_{ij}^F , which aims to adjust the Facebook data for this bias caused by the change in the algorithm.

$$\zeta_{ij} = \log \left(\frac{\text{Estimates before}_{ij} - \text{Estimates after}_{ij}}{\text{Estimates before}_{ij}} \right) \quad (3.11)$$

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

$$\lambda_{ij} = \log \left(\frac{\text{FB Language}_{ij}}{\text{FB Migrant}_{ij}} \times \frac{\text{LFS Greece Migrant}_{ij}}{\text{LFS Greece Migrant}_{ij} + \text{LFS Cyprus Migrant}_{ij}} \right) \quad (3.12)$$

After this assessment, the Facebook measurement error equation is:

$$\log \mu_{ij}^F = \log y_{ij} + \delta^F + \beta^F + \chi_{ij}^F + \zeta_{ij}^F + \lambda_{ij}^F + \epsilon_{ij}^F \quad (3.13)$$

3.3.3 Theory-Based Model

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

$$\log y_{ij} = \alpha_0 + \alpha_1 P_{ij} + \alpha_2 I_{ij} + \alpha_3 O_{ij} + \alpha_4 \log G_{ij} + \alpha_5 \log U_{ij} + \epsilon_{ij} \quad (3.14)$$

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

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

- **P**: a normalised measure of population size in the country of origin, divided by the mean of the population in the same countries considered in the model. The data is from the latest estimates by Eurostat (2019, 2020);
- **I**: a normalised measure of the inflows from European countries to the UK, divided by the mean of the inflows of migrants from the countries considered in the model. The data is from the IPS in 2017 and 2018;
- **O**: a normalised measure of the outflows to the European countries from the UK, divided by the mean of the outflows of emigrants from the countries considered in the model. The data is from the IPS in 2017 and 2018;
- **G**: ratio of GDP growth rate in the European country of origin in 2017 and 2018, divided by the GDP growth rate in the UK. The data is from Eurostat;
- **U**: ratio of the unemployment rate in the European country of origin in 2017 and 2018, divided by the unemployment rate in the UK. The data is from Eurostat.

The normalised measure of the population size is a predictor of the possible number of migrants informed by a gravity model; i.e. the larger the population, the larger the number of possible migrants. The normalised measures of inflows and outflows from the IPS provide an indication of the levels of fluctuation in terms of arrivals and departures for every nationality, and thus help to capture fluctuations in the stocks. The ratio of the GDP growth rate to the unemployment rate provides information on how the economy of the country of origin compares to that of the UK, and therefore is a form of economic gravity indicator.

3.4 Results

The models were estimated in R using JAGS (Plummer et al., 2016). The models were run with three chains, 100,000 interactions, 1,001 burn-in and 10-fold thinning (i.e. every 10th value of the chains is kept and all other values are discarded to avoid autocorrelation). Two sets of models are presented. The first is for the total number of European migrants in the UK, and the second disaggregate the estimates by sex. The two sets of models are run simultaneously by year (2018 and 2019) to borrow strength across the years. In the first model, the aim is to explain the magnitude of the undercount of the LFS data relative to the estimates produced by the model for the two years. Finally, all the estimates of the models converge. Detailed results and some diagnostic statistics are included in Appendix A.

3.4.1 Model for Total Numbers

Figure 3.5 shows data from three datasets and our estimates: the Facebook Advertising data is in blue, the LFS data is in yellow, the settled status application data is in red, and the model estimates are in green. The settled status data is used for comparison, and is not used in the analysis. LFS data is shown with a 95% confidence interval (CI), while model estimates are shown with the quartiles. The data for the two years is identified by a circle for 2018, and by a square for 2019.

The settled status data is used for comparison, and is not used in the analysis. LFS data is shown with a 95% confidence interval, while model estimates are shown with the

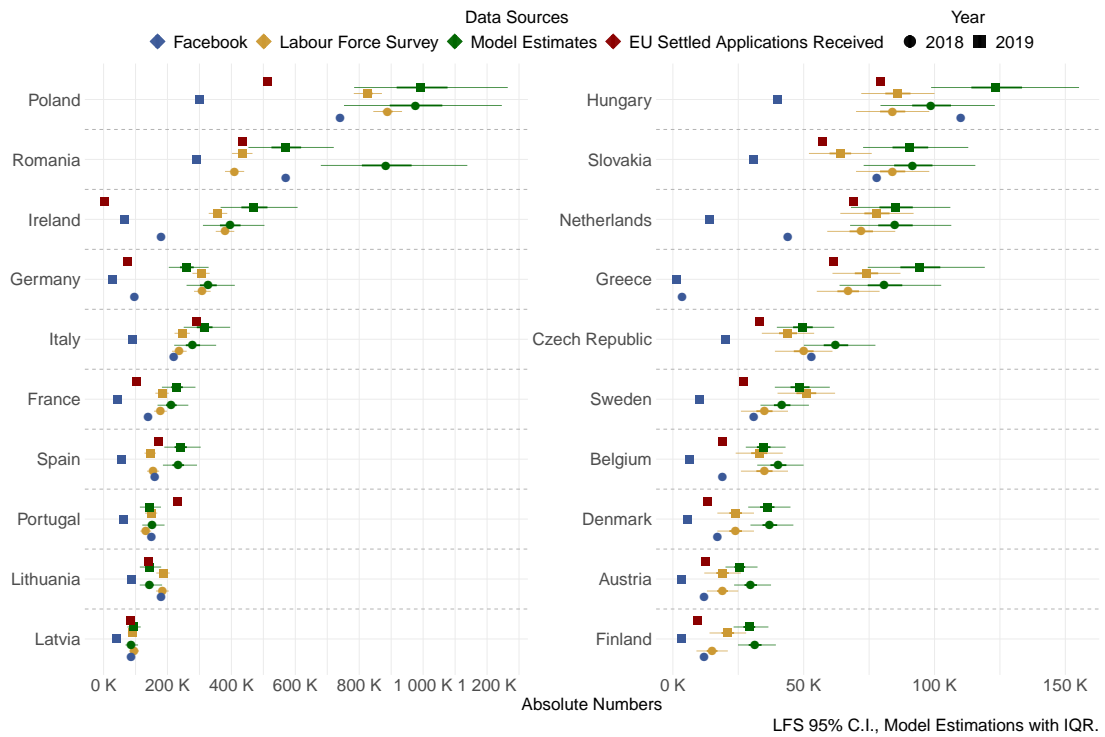


FIGURE 3.5: Comparison of Facebook, LFS, and model estimations of EU migrants aged 15+ for the years 2018 and 2019.

quartiles. The data for the two years are identified by a circle for 2018, and by a square for 2019.

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

In Figure 3.5 the estimates for the second group of countries are also shown. The parameter on Greece seems to be effective in bringing the estimates closer to the LFS values. In Appendix A, the posterior characteristics of the true stock estimates for all of the models and the \hat{R} are reported, a measure that helps determine whether chains have converged depending on whether it is close to one (Gelman et al., 2013). All of the chains have converged when \hat{R} is strictly equal to one (except for Romania in 2018 and

Poland in 2019, where \hat{R} is 1.01 as shown in Appendix A). The algorithm for estimating all of the other parameters has converged as well.

In Table 3.3, a comparison of the undercounted LFS estimates with the model estimates is presented. While the ONS has estimated an undercount of 16%, the model estimates an undercount of 24.47% for 2018 and 19.84% for 2019.

TABLE 3.3: Undercount of the LFS estimates in comparison with the model estimates.

	2.5%	25%	50%	75%	97.5%
2018	13.11 %	20.55 %	24.47 %	28.47 %	36.69 %
2019	9.64 %	16.23 %	19.84 %	23.53 %	30.88 %

The undercount for 2018 has larger intervals, likely due to the prior on the algorithm change. Additionally, the model for 2019 estimates a higher number of migrants of certain nationalities (e.g., Polish, Italian, and Hungarian), and a lower number of migrants of other nationalities (e.g., Romanian, German and Czech). The interquartile range of these distributions is large, highlighting the uncertainty in the estimates. However, the models for the two years indicate that the undercount and the uncertainty are in the same direction.

3.4.2 Model Disaggregated by Sex

In this part of the model the estimates are disaggregated by sex. It is important to study the age and sex differences of migrants. The model proposed works for sex disaggregation, and Figure 3.6 shows the estimates. In this case, the comparison with migrants who have applied for the settled status scheme is not available because the data from the Home Office is not disaggregated by sex. There are no large differences between the two sexes in the number of migrants within countries and across years.

3.4.3 Sensitivity Analysis

Some sensitivity checks of the model are provided. First, the model was run while only including the LFS data. For the model specified in this paper, the undercount is estimated at 25% in 2018 and at 20% in 2019. In Table Table 3.4, the undercount of this new specification of the model is reported, estimated at a median level of 8% in 2018

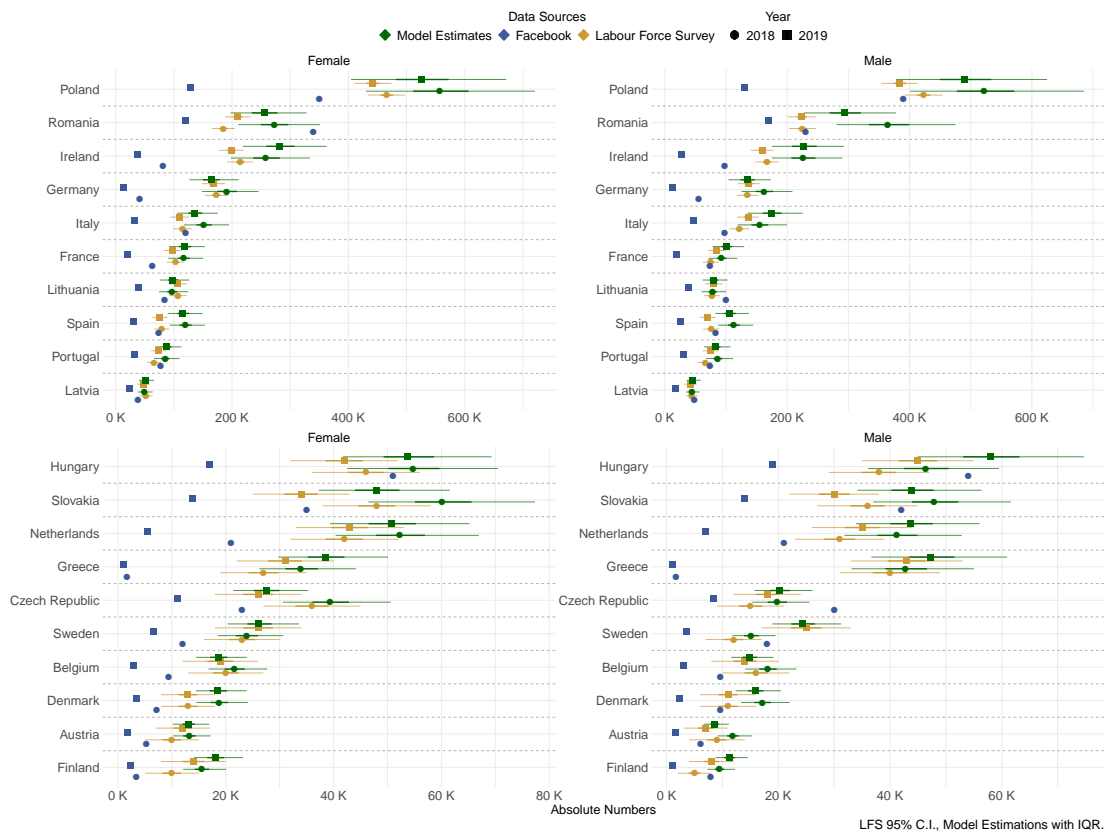


FIGURE 3.6: Comparison of Facebook, LFS, and model estimations of EU migrants aged 15+ by sex for the years 2018 and 2019.

and 22% in 2019. These two median levels are not close to those produced by the model that combines Facebook and LFS data, with a smaller undercount in 2018 and a larger one in 2019. Overall, the uncertainty of the undercount estimate is greater when using only LFS data. The second sensitivity check was to modify the parameters from the Facebook and LFS Measurement Error Models. In the models included in this paper, the parameters are informed by previous research and calculations on the data, except the β^F , which is the bias parameter for Facebook. It is assumed the value is lower than the percentage of fake and duplicate accounts worldwide. In the sensitivity analysis the Facebook bias parameter was first modified to 0%, indicating no bias in the Facebook estimates, and then to 11%.

In Table 3.4 the undercount value of the new specifications of the model is reported. The undercount with no bias attributed to the Facebook estimates is 22% for 2018 and 19% for 2019, which is slightly lower than that specified in the suggested model. The undercount with a higher β^F is 25% for 2018 and 20% for 2019. The undercount with a β^F at 4% and at 11% are very similar.

TABLE 3.4: Undercount of the LFS estimates in three different models 1) the model specified only with the LFS data, 2) the model with the Facebook bias parameter set to 0%, 3) the model with the Facebook bias parameter set to 11%, 4) the model with the LFS bias parameter set to 4%, 5) the model with the LFS bias parameter set to 30%, and 6) the model with the $Gamma(1,1)$ distribution.

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

The model is sensitive to the choice of the assumed bias of the LFS parameter. In Table 3.4 we modified the bias of the LFS to 4% (the minimum level assumed) and to 30% (the maximum level assumed) for all the countries. With the low minimum bias level assumed, the undercount reaches negative median values, while it is larger when the maximum bias level assumed. We also tried different specifications of the precision distribution term, which is assumed to follow a $Gamma(100,1)$ in the presented model. In Table 3.4, the model was specified with a $Gamma(1,1)$, which is less informative than $Gamma(100,1)$. The gradient of the median of the undercount is similar to the one in the presented model, though the uncertainty is larger. There is some impact of the prior selection on the uncertainty of the estimates.

Additionally, in Figure 3.7 the estimates from the model on the total estimates (model 1) are compared to the sum of the estimates from the sex disaggregation model. While the estimates are close to each other, there are cases in which the sum from the sex disaggregation model is not completely aligned with the distribution from model 1. This is due to inconsistencies in the Facebook and LFS data disaggregated by sex. While the estimates from our models seem to be stable to different prior distributions, the precision of those prior distributions had to be carefully chosen to ensure model convergence, while exploring reasonable areas of the parameter space with respect to the precision parameters.

Finally, I perform additional goodness-of-fit tests comparing the data with simulated data from the models. These checks are contained and described in Appendix A. Overall, the result is that the model specified could be improved considering alternative distributions for the data.

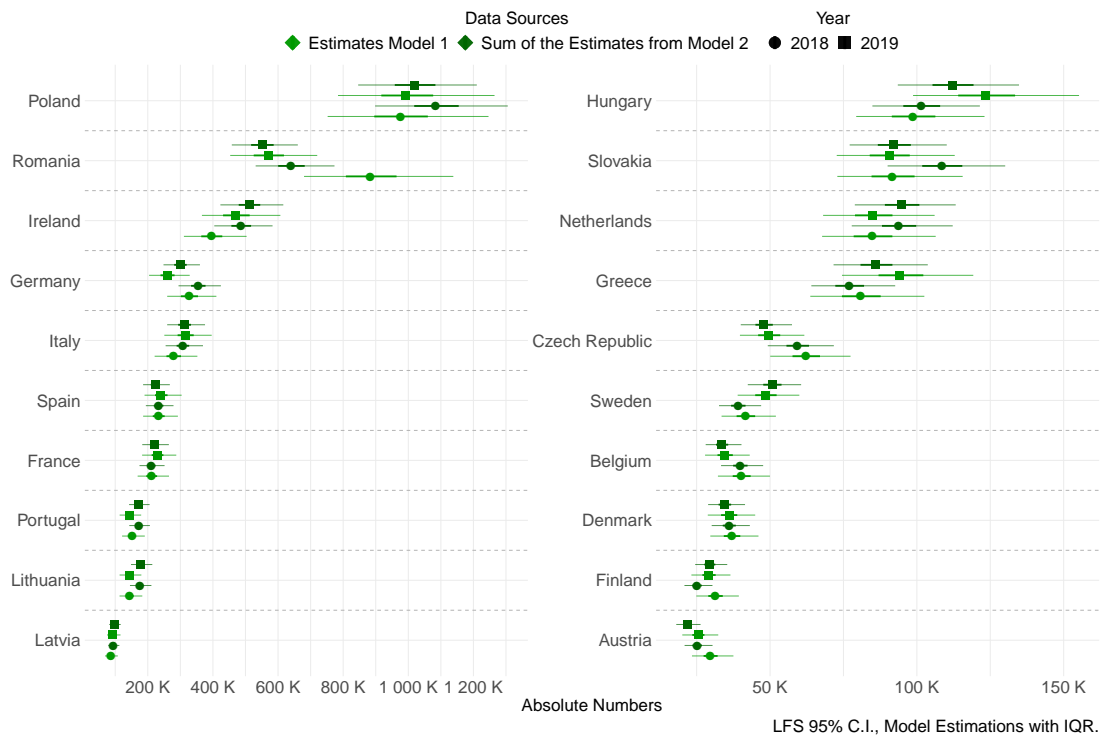


FIGURE 3.7: Comparison between estimates from the first model and the sum of female and male migrants from the second model for 2018 and 2019.

3.5 Discussion

The model estimated the migrant stocks for 2018 and 2019. In the 2018 model, a prior distribution was used to account for an algorithm change that Facebook implemented in March 2019 which led to a decrease in the estimate of European migrant numbers. This algorithm change was not uniform, however, as it varied by country and sex of the migrants. This finding highlights the importance of monitoring digital traces, and that using digital traces alone is not sufficient to generate better estimates of stocks of migrants. The parameters associated with the algorithm change and the Greek factor (e.g. the factor that Greeks are underrepresented in the Facebook migrant variable) were shown to be effective in bringing the model estimates in line with the LFS estimates.

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

An estimate of the total number of European migrants by sex is also provided. The sum of the estimates from this second model were equal to the total from the first. There was uncertainty in our estimates, greatest for the countries of origin with the highest number of migrants in the UK: Poland, and Romania. This might suggest that for nationalities where the level of uncertainty is higher, the sample of households and migrants interviewed should be increased. A possible solution to reduce the uncertainty would be to include a prior distribution driven by expert opinion in the model, as well more informative priors on the Facebook and LFS data once they become available.

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

The model proposed is similar to the one suggested by [Gendronneau et al. \(2019\)](#); however, it focuses only on one country, the UK. On the other hand, the model in [Gendronneau et al. \(2019\)](#) focuses on three years, ranging from 2016 to 2018, and uses both MAU and DAU (for 2018) from the Facebook Advertising Platform. Moreover, it considers not only the LFS data, but also Census and Eurostat data. The model proposed in this thesis could also be extended to combine multiple datasets, as suggested by [Gendronneau et al. \(2019\)](#). Given that the Facebook data are rounded to the closest thousand, it is interesting the use of a tobit regression to correct for any biases coming from censoring. The model proposed in this thesis, which contains additional priors to correct for biases and coverage of the Facebook data, could be extended to several EU countries and analytically compared with [Gendronneau et al. \(2019\)](#).

3.6 Conclusions

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

This analysis has made three contributions to digital and computational demography. First, it has proposed to apply a framework that is already in use in migration research to digital traces. The proposed model is a flexible framework, in which it is possible to include new information as soon as it becomes available, including additional digital trace data, such as from other advertising platforms like Instagram, Snapchat, and LinkedIn, as well as from other administrative sources. Second, it has addressed the biases of both traditional and digital trace data. The use of a prior distribution has been shown to fix these issues in a probabilistic fashion. Third, it has produced an estimate of the undercount of migration levels. Overall, the model estimated an undercount of

24.47% for 2018 and 19.84% for 2019 based on the LFS data. For migrants to the UK from the EU8 countries, the ONS had estimated an undercount of 16% for March 2016. It would be possible to compute this measure based on data from both the LFS and Facebook at the time of the next census (which in the UK is scheduled for 2021). In this way, the model could be used to help nowcast migration in a timely manner, comparing the estimates to those of the census.

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

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

Chapter 4

Extending the Migration Estimation Model including Age and Sex Profiles of Migrants

4.1 Introduction

Migration is a crucial process that shapes not only the population size, but also the age and sex structure of countries and regions across the world (Rogers and Castro, 1981; Rogers and Watkins, 1987). Estimates of migrant population stocks by age and sex are important for making population projections, as well as understanding migrant populations' types of and motives for migration (Bernard et al., 2014; Wiśniowski et al., 2016; Bijak et al., 2007, 2008). For example, it has been shown that the majority of migration movements are concentrated during young adulthood (Rogers and Castro, 1981; Wilson, 2010). Measurement of the number of migrants by age and sex is sometimes difficult, however, as traditional data sources often lack the necessary timeliness and coverage. Especially in the case of surveys, when the estimates are disaggregated by age and sex, the uncertainty increases due to the small sample sizes examined. Furthermore, traditional data sources are designed to measure stable populations and might be lacking in their representation of migrants. Indeed, the ONS has stated that the LFS is not adequately designed to study migrants (ONS, 2018a, 2019e).

Digital traces, however, are timely and, although not entirely representative, have a higher coverage of younger adults and individuals with higher digital skills (Hargittai, 2018). Previous research has looked at the potential of the Facebook Advertising Platform's data on the US population disaggregated by age and sex (Alexander et al., 2020; Zagheni et al., 2017). Such an approach may be especially useful when applied to estimate the size of the migrant population in a particular country. The aim of this chapter is to produce migration estimates disaggregated by age and sex, extending the model from Chapter 3. In the previous chapter, a Bayesian hierarchical model was used to combine survey data with digital traces data. Subsequently, the model redistributed the estimates by age and sex following a harmonised migration schedule derived from the data sources. This general framework can be applied in the context of limited and inconsistent data.

In this chapter, data from the LFS, a Europe-wide household survey, and the Facebook Advertising Platform are used, both disaggregated by age and sex. The model proposes an extension of the IMEM by age and sex (Raymer et al., 2013) and an application of digital traces data. A simplified Rogers-Castro model (Rogers and Castro, 1981; Rogers

and Watkins, 1987) computes the age profiles of the migrants from the LFS and the Facebook Advertising Platform data. A multinomial-Dirichlet-Dirichlet model is then used to harmonise the age profile of migrants. As an illustration, the model is applied to estimate the age and sex distribution of the ten most numerous European migrant groups in the UK. The analysis was limited to these ten migrant groups, with the corresponding coverage reduced from twenty to ten countries, to ensure model reliability. Due to privacy concerns, both the LFS and Facebook Advertising Platform estimates are rounded by Facebook when the number of migrants is low.

4.2 Background

In 2016 the ONS reported 9 million non-British born residents in the UK, 38% of which were born in a European country (ONS, 2018c). The main reasons driving European migration to the UK were reported to be work and study (Kierans, 2020). In the academic year 2018/19, the Higher Education Statistics Agency (HESA) reported that 30% of international students in the UK were EU citizens, with the most numerous nationalities being Italian, French, and German (HESA, 2020). Moreover, the Department for Work and Pensions (DWP) has recently provided estimates from the National Insurance Number (NINo) system, a register number that foreign citizens are required to obtain in order to work in the UK (ONS, 2020c). These estimates highlight that EU migrants to the UK are generally reasonably young and in working age. The ONS is currently in the process of transforming their migration statistics by combining multiple administrative data sources (ONS, 2019g); this process might help in reducing the uncertainty of migrant estimates disaggregated by age and sex by not drawing solely on the LFS, but by adding information from the HESA and DWP to the LFS (Disney, 2015). However, there is still the possibility to add digital traces data to these estimates calculations.

Given that the LFS is a household survey that targets households through post codes, the sampling framework might underrepresent young migrant adults, who might be in more unstable jobs and therefore be in more unstable renting conditions. Indeed, the ONS has suggested that the LFS is not designed to study long-term migrants and does not provide complete coverage (ONS, 2018a, 2019e). Moreover, the LFS only interviews

individuals who have lived at the same address for at least six months, including short-term migrants but excluding communal establishments (ONS, 2019e).

As mentioned, traditional data sources lack timeliness and comprehensive coverage of migrants in terms of age and sex. Notwithstanding the problems of Facebook data, it seems digital traces might be better suited than the LFS to capture younger migrants who do not live in standard households, such as university campuses or other types of accommodation not easily captured by the LFS. In this chapter, age migration patterns discernible from Facebook and the LFS are combined in order to add coverage to the traditional data sources.

The literature on the relationship between age and migration has mainly focused on flows. Rogers and Castro (1981), for example, described the age-specific regularities of migration flows with mathematical expressions. Courgeau (1985) made an explicit connection to the link between the life course and migration. Individuals seemed to be less inclined to migrate after marriage, divorce or widowhood, but an overall increase in migration is supported by the leave of children from the parental home (Courgeau, 1985). Furthermore, migration rates have been shown to be positively affected by education and type of employment (Courgeau, 1985). Rogers and Castro (1981) used a multi-exponential model to capture the different dependencies between age and migration. Empirical data on migration flows from developed countries were used to understand the age regularities. The curve, $M(x)$, described by the Rogers-Castro model displays high levels of migration in the first years of life (the “pre-labour force”), which then drops rapidly before increasing again in the late teenage years. The peak in migration was shown to be in the early twenties age group. Additionally, there is another hump in migration in the years after retirement. Therefore, the model is constituted by four components: the labour force, the pre-labour force, the post-labour force, and a constant component, included to improve the model fit. Four main model specifications were presented: the “standard model” with 7 parameters, the “elderly post-retirement migration model” with 9 parameters, the “elderly retirement peak model” with 11 parameters (Rogers and Castro, 1981), and the “elderly retirement peak plus post-retirement” with 13 parameters (Rogers and Watkins, 1987). The model is flexible, given that it is possible to choose how many parameters are included given the shape of the migration

pattern. In the following equations, the Rogers-Castro model is presented in $M(x)$ with 13 parameters.

$$M(x) = a_1 \times \exp(a_1 x) + \text{Pre-Labour Curve (4.1)}$$

$$a_2 \times \exp(-a_2(x - \mu_2)) - \exp[-\lambda_2(x - \mu_2)] + \text{Working-Age Curve (4.2)}$$

$$a_3 \times \exp(-a_3(x - \mu_3)) - \exp[-\lambda_3(x - \mu_3)] + \text{Post-Retirement Curve (4.3)}$$

$$a_4 \times \exp(\lambda_4 x) + \text{Retirement Peak (4.4)}$$

$$c_1 \text{ Constant Component (4.5)}$$

Wilson (2010) suggested a further extension of the model to include a student component for internal migration. This curve is informed by a spike in the intensity of migration in the late teenage years, which coincides with the transition from high school to university. The suggested parametrisation of the student curve is similar to the working age curve. This additional curve might be necessary when the data used is disaggregated by year, as looking at 5 years' worth of data would suppress these intensities. The student curve should be considered specific to each country.

Previous attempts to model age and sex structure were based on multiple imperfect data sources. An extension of the IMEM with age and sex has already been suggested by Wiśniowski et al. (2016). Wiśniowski et al. (2016) computed the flows by origin (O), destination (D), age (A), and sex (S). First, the "true flow" was estimated by origin and destination using the Measurement Error Model and the Migration Theory Model. The structure of these models follows the Integrated Model of European Migration (Raymer et al., 2013), which has been described in Chapter 3. In Wiśniowski et al. (2016), the analysis considered the data from 31 European countries from 2002 to 2008. Disaggregation by age and sex was then performed by harmonising the proportions of the age profiles from sending and receiving countries through a multiplicative model; a multinomial Poisson model with over-dispersion. Finally, the "true flow" by origin and destination were redistributed by the harmonised migration schedule. Several limitations of the age and sex flows were reported. In this chapter, a similar method structure

is used. Thus, firstly the “*true stock*” by sex is calculated, as presented in Chapter 3, followed by the use of a multinomial-Dirichlet-Dirichlet model to redistribute the true stock for female and male migrants using a harmonised age profile.

4.3 Data

The data used here is the same as in Chapter 3 of this thesis (see Section 3.2), but with the data from the LFS and Facebook Advertising Platform now disaggregated by age and sex. The LFS disaggregate their estimates into 5 year age groups from 0 to 84 years old, with an open-ended group for those aged 85+ years. The LFS data is not provided when there were only 0 to 3 interviews/contacts in that specific age, sex, and country of origin group¹. Facebook has a minimum age of 13 for its users, so the Facebook Marketing API provides data for ages 13-65+. Data can be collected for individual ages, but there are limitations when the group falls under the 1000 minimum threshold of the Monthly Active Users (MAUs) data on Facebook. An additional limitation of the Facebook Advertising Platform data is that it does not provide further disaggregation of the age group 65+. Moreover, Facebook has suggested that “*a disproportionate number of our younger users register with an inaccurate age*” (USA SEC, 2018, 2019). Therefore, even if it is possible to obtain single year age groups, it is advisable to aggregate the data into age groups of five years in order to reduce the uncertainty of the Facebook data. Consequently, this chapter uses the following age structure for the two data sources: 15-19 to 60-64, with an open-ended age group for the population aged 65+.

4.4 Methodology

The model is divided into three parts, which is graphically represented in Figure 4.1:

1. the Measurement Error Model (see Section 3.3.2) and Theory-Based Model (see Section 3.3.3) is used to estimate the true stock by sex;
2. the multinomial-Dirichlet-Dirichlet model is used to estimate the *true proportion by age and sex*, and

¹The reference to this statement is in the notes of the data file. The special values represent the following: . for “no contact”, c for “not available due to disclosure control”, and 0 for “rounded to zero”.

3. where the two parts are combined to disaggregate the *true stock by sex* into the *true proportion by age and sex* to obtain the *true stock by age and sex*.

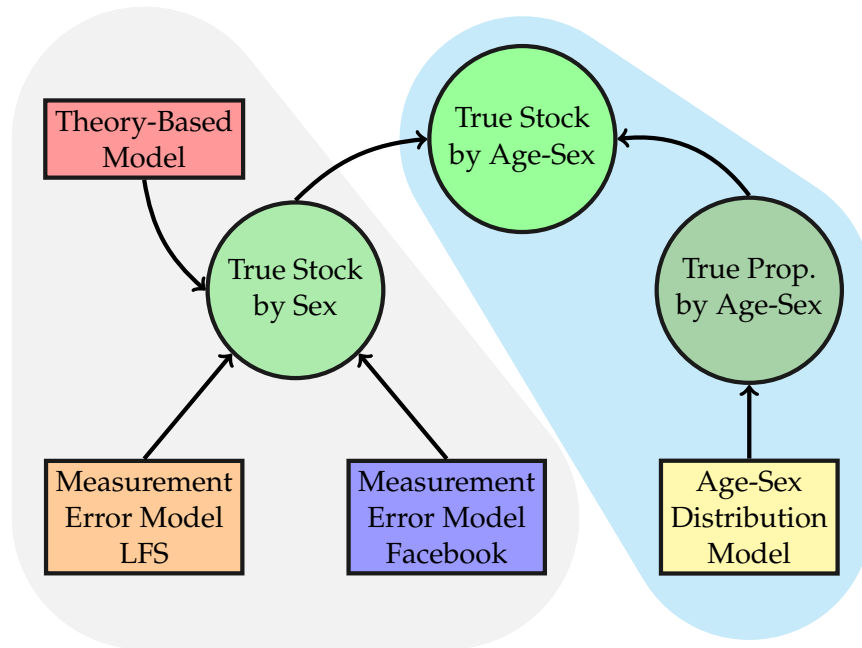


FIGURE 4.1: Diagram of the structure of the model.

The first part of the model has already been presented in Chapter 3; this is highlighted in grey in Figure 4.1. The second part of the model, highlighted in light blue in Figure 4.1, is the focus of this chapter. First, a Rogers-Castro age schedule model is fitted to the data from the LFS and Facebook Advertising Platform. Then the estimated proportions obtained from the Rogers-Castro model are used to redistribute the sex-specific migrant stocks into the different age groups through a multinomial-Dirichlet-Dirichlet model. The goal is to obtain a harmonised migrant population age structure from the two data sources. In the next section, the rationale of using the Rogers-Castro model is explained.

4.4.1 Fitting the Rogers-Castro Age Schedule

Migration flows have previously been modelled using the Rogers-Castro model. Whilst the model was not explicitly designed to estimate migrant stocks, as we aim to do here, the model is flexible to being adapted for this purpose: it can be assumed that the accumulated flow of individuals over time (as captured in the Rogers-Castro model) is equal to migrant stocks. Further, it can be assumed that constant immigration in low

fertility contexts may result in a stationary population, meeting the assumptions of the Roger-Castro model. It is therefore deemed that the model is appropriate for estimating a broad range of possible stock outcomes (Espenshade et al., 1982). The Rogers-Castro model can be fitted with 7 to 13 parameters. Given that our data covers the age groups 15 to 65+, a reduced formulation of the model is used as not all the parameters are needed, so the Rogers-Castro model was thus fitted with just 8 parameters. A *working age curve*, a *post-retirement curve* and the constant were included. These components are described as:

1. **working age curve**: this is a left-skewed unimodal curve for the age pattern of migration of people of working age;
2. **post-retirement curve**: while normally used to describe post-retirement migration, in this case it is used to describe older aged migrants as well as families that brought their older parents with them;
3. **constant** component: constant term which represents “background” migration.

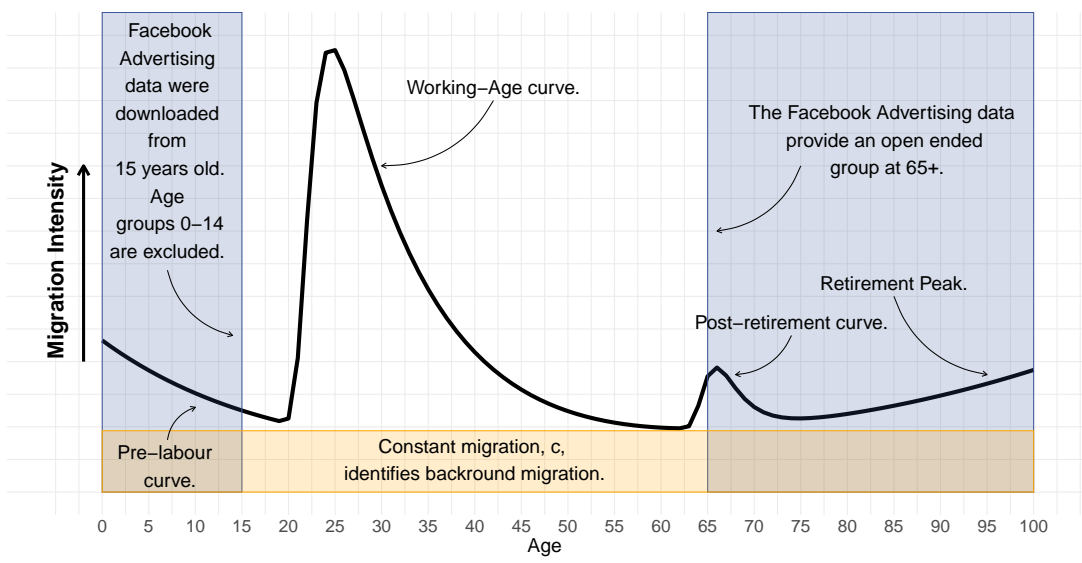


FIGURE 4.2: Rogers-Castro age migration schedule.

Figure 4.2 shows the rationale of the choice of the parameters given the age structure of the data used. The model was fitted on the Facebook and LFS data simultaneously. Three groups were identified of countries with similar shapes. To fit the Rogers-Castro

Model the R code on Github² from Ruiz-Santacruz (2019) was modified. Finally, the equation of the specified Rogers-Castro model is:

$$M(x) = a_2 \times \exp(-\alpha_2(x - \mu_2)) - \exp[-\lambda_2(x - \mu_2)] + \text{Working Age Curve (4.6)}$$

$$a_4 \times \exp(\lambda_4 x) + \text{Retirement Peak (4.7)}$$

$$c_1 \text{ Constant Component (4.8)}$$

The three different groups of countries identified were:

1. **Group 1:** “Western and Southern European” countries, including France, Germany, Italy, Portugal, and Spain;
2. **Group 2:** the Republic of Ireland;
3. **Group 3:** “Central and Eastern European” countries, including Latvia, Lithuania, Poland, and Romania.

Figure 4.3 represents the proportions, $p_{1s} \dots p_{sk}$ (where k identifies the data sources), from the LFS and Facebook Advertising Platform data with circles and the Rogers-Castro estimates with lines. The LFS data is shown in yellow, while the Facebook Advertising Platform’s data is in blue.

The three groups of countries identified above display different age profiles of migrants. In the first group, Western and Southern European countries show a younger proportion of migrants on Facebook and an older proportion on the LFS. The second country group includes only the Republic of Ireland given the particular age distribution from the LFS data, done to account for the long history of Irish migration. Migration from the Republic of Ireland is inextricably connected to the evolution of the UK, highlighted by the high proportion of 65 year olds born in the Republic of Ireland and now living in the UK. Finally, the third group includes Central and Eastern European countries. The migration from the Central and Eastern European block is recent, starting in force after they began to join the EU from 2004. The proportions of age groups

²<https://github.com/elflacosebas/migraR>

shown by the LFS and Facebook are almost identical for these countries, with a slightly higher proportion of the younger age groups in the Facebook data. The Facebook proportion between group 1 and 3 follow a similar shape, however, the LFS data provide different curves.

The peak age group is 25-29 years old for the Western and Southern European group, and is 30-34 years old for the Eastern European group. It would seem that the Rogers-Castro model fits better for those countries with an earlier migration schedule, such as and Southern European groups. This means the approach has limitations when applied to the Republic of Ireland, where the migrant stocks are mostly driven by the 65+ years old migrants who have already been living in the UK for a long time. As Facebook mostly captures younger migrants, the 65+ years old migrants are likely underestimated.

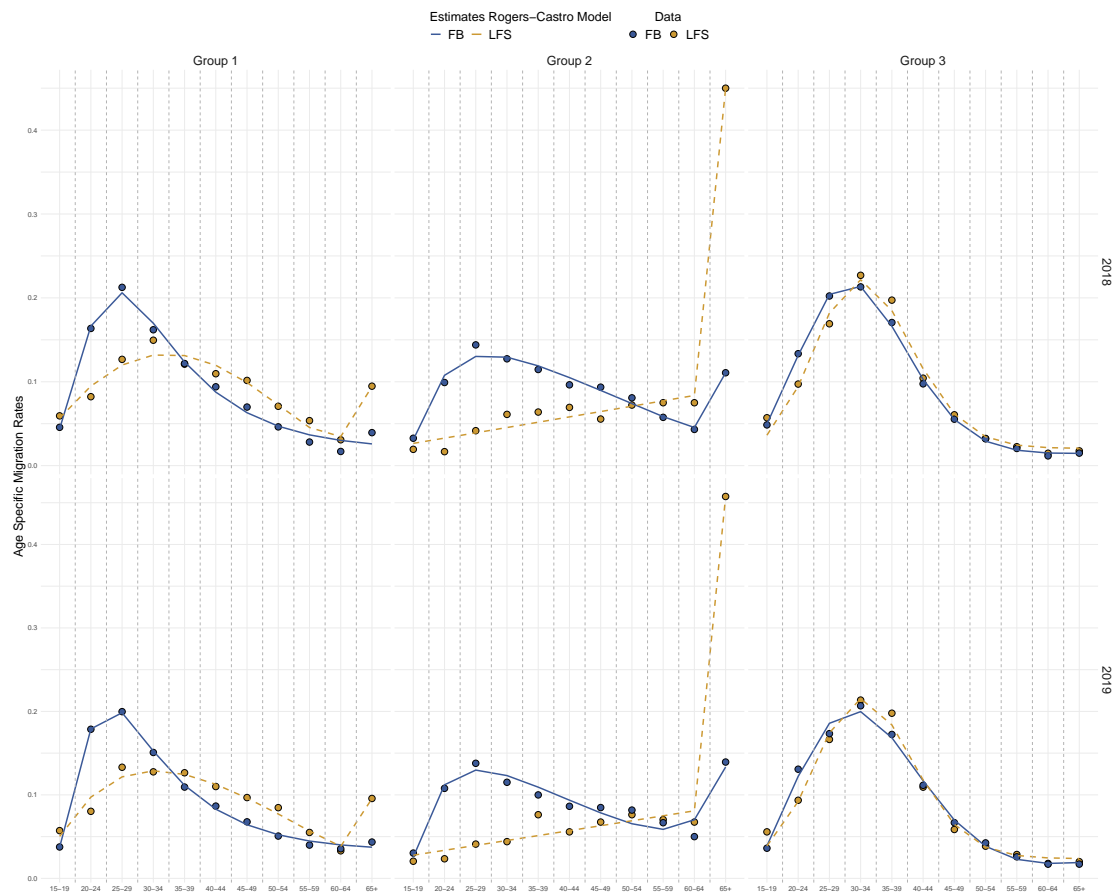


FIGURE 4.3: Rogers-Castro estimates of the proportions of the EU migrant population living in the UK in 2018 and 2019 by the three country groups identified.

4.4.2 The Multinomial-Dirichlet-Dirichlet Model for Age Schedules

Given the structure of the data, there is the need for a statistical model for proportions; a natural choice is the Dirichlet model, which has the following density distribution:

$$\text{Dir}(\theta|\alpha) = \frac{1}{\text{Beta}(\alpha)} \prod_{i=1}^K \theta_i^{\alpha_i-1}, \text{ where } \text{Beta}(\alpha) = \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^K \alpha_i)} \text{ and } \alpha = (\alpha_1, \dots, \alpha_K) \quad (4.9)$$

In the context of demography, the use of the Dirichlet distribution was suggested by [Caussinus and Courgeau \(2010\)](#) in studies of paleo-demography: they showed an application which estimates the age structure of past populations with limited or no data. The Dirichlet distribution allows the redistribution of the population components into categories around a total. In this vein, the approach has been used for example to study the demographic features and dynamics of a Scythian population from the Black Sea region ([Łukasik et al., 2017](#)). In the Bayesian context, an additional advantage is that the Dirichlet distribution is the conjugate for the multinomial distribution. In this study, a multinomial-Dirichlet-Dirichlet model is used to combine the migration age profile from the LFS data with the Facebook Advertising data. The model is estimated including all the variables simultaneously to borrow strength between the parameters for country, age, sex, and year.

Figure 4.4 graphically describes the hierarchical structure of the multinomial-Dirichlet-Dirichlet model. The indicators parameter are k for age groups, i for country, j for sex, t for year. The top-level Dirichlet model has parameters a_k equal to 1 - this carries little information *a priori* and does not favour any population group. It is used as a hyper-prior for the multinomial distribution that harmonises the Rogers-Castro estimates from the LFS and the Facebook Advertising data. The Rogers-Castro estimates of age-specific counts enter into the model via the $\alpha_{kijt}^{L,F}$. The top-level Dirichlet model is applied to the LFS and Facebook age structures that are assumed to follow a common Rogers-Castro distribution.

$$a_1, \dots, a_K \sim \text{Dir}(1, \dots, 1) \quad (4.10)$$

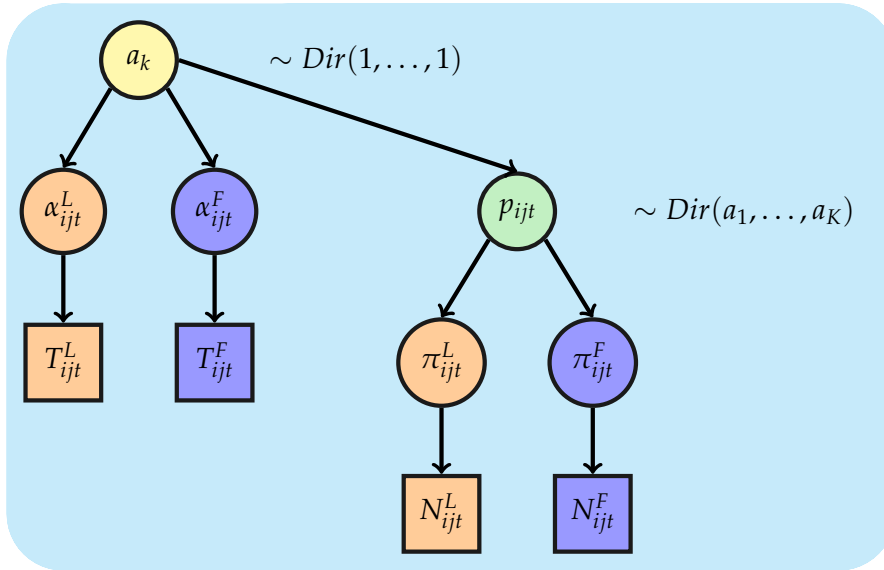


FIGURE 4.4: Diagram describing the hierarchical structure of the multinomial-Dirichlet-Dirichlet model. Indices: k , age groups, i , sending country; j , sex; t , time. Square nodes represent reported data (T_{ijt}^L , T_{ijt}^F , N_{ijt}^L , N_{ijt}^F). Circle nodes represent the parameters of the model (a_k , α_{ijt}^L , α_{ijt}^F , π_{ijt}^L , π_{ijt}^F , and p_{ijt}).

The pseudo-counts are obtained from a multinomial-Dirichlet model which applies the vector a_k to the sum of counts from the Rogers-Castro model $T_{ijt}^{L,F}$. From this top-level Dirichlet distribution harmonised migration schedule is obtained from the LFS and Facebook Advertising Platform Rogers-Castro estimates:

$$\alpha_{1ijt}^L, \dots, \alpha_{Kijt}^L \sim \text{Mult}(a_1, \dots, a_K, T_{ijt}^L) \quad (4.11a)$$

$$\alpha_{1ijt}^F, \dots, \alpha_{Kijt}^F \sim \text{Mult}(a_1, \dots, a_K, T_{ijt}^F) \quad (4.11b)$$

The second Dirichlet applies the a_k obtained from the Rogers-Castro estimates applied to the top-level Dirichlet model to the total estimates from the LFS and Facebook Advertising data. The k indicator specifies the age groups, with $k = (1, \dots, 11)$. This second Dirichlet model is necessary in order to apply the harmonised Rogers-Castro schedule to the true stock data through a combined harmonised migration schedule, which has the following prior:

$$p_{1ijt}, \dots, p_{Kijt} \sim \text{Dir}(\alpha_1, \dots, \alpha_K) \quad (4.12)$$

from Equation 4.12, $\pi_{1ijt}, \dots, \pi_{Kijt}$ is obtained; the proportions of migrants by age and sex for each country. Here, an additional indicator is j for sex, since the proportions are now also disaggregated over sex. The likelihood of the model is multinomial, defined by $\pi_{1ijt}^{L,F}, \dots, \pi_{Kijt}^{L,F}$ which are the proportions by age and sex for each country and $N_{ijt}^{L,F}$, which are the totals by country, sex, and year to be distributed:

$$\pi_{1ijt}^L, \dots, \pi_{Kijt}^L \sim \text{Mult}(p_{1ijt}, \dots, p_{Kijt}, N_{ijt}^L) \quad (4.13a)$$

$$\pi_{1ijt}^F, \dots, \pi_{Kijt}^F \sim \text{Mult}(p_{1ijt}, \dots, p_{Kijt}, N_{ijt}^F) \quad (4.13b)$$

Figure 4.5 highlights the three groups of countries, the Rogers-Castro estimates from Facebook in blue, and the LFS in yellow, in grey the harmonised age profile by country. These are used to redistribute the true stock estimates by sex through multiplication obtaining a true stock estimate disaggregated by age and sex.

4.5 Results

In this section, the results for the models for the three groups of countries and for the two years are presented. As the models in Chapter 3, the models were estimated in R using JAGS (Plummer et al., 2016). The models were run with three chains, 100,000 interactions, 1,001 burn-in and 10-fold thinning. First, the results for the Western and Southern European migrants are discussed, then for the Republic of Ireland, and finally for Central and Eastern European migrants. The model was run simultaneously for each year (2018 and 2019) and country. The model was applied to the ten most numerous migrant groups in the UK. In Figure 4.6, Figure 4.7, and Figure 4.8, the results are shown with green box-plots showing the IQR. The data from the LFS is represented by yellow box-plots with CI at 50% and 95%. The Facebook Advertising Platform data is represented by blue points. The results are shown as counts of migrants by age and sex. \hat{R} is strictly equal to 1, meaning that the chains converged (Gelman et al., 2013); the statistics are fully reported in the Appendix B.

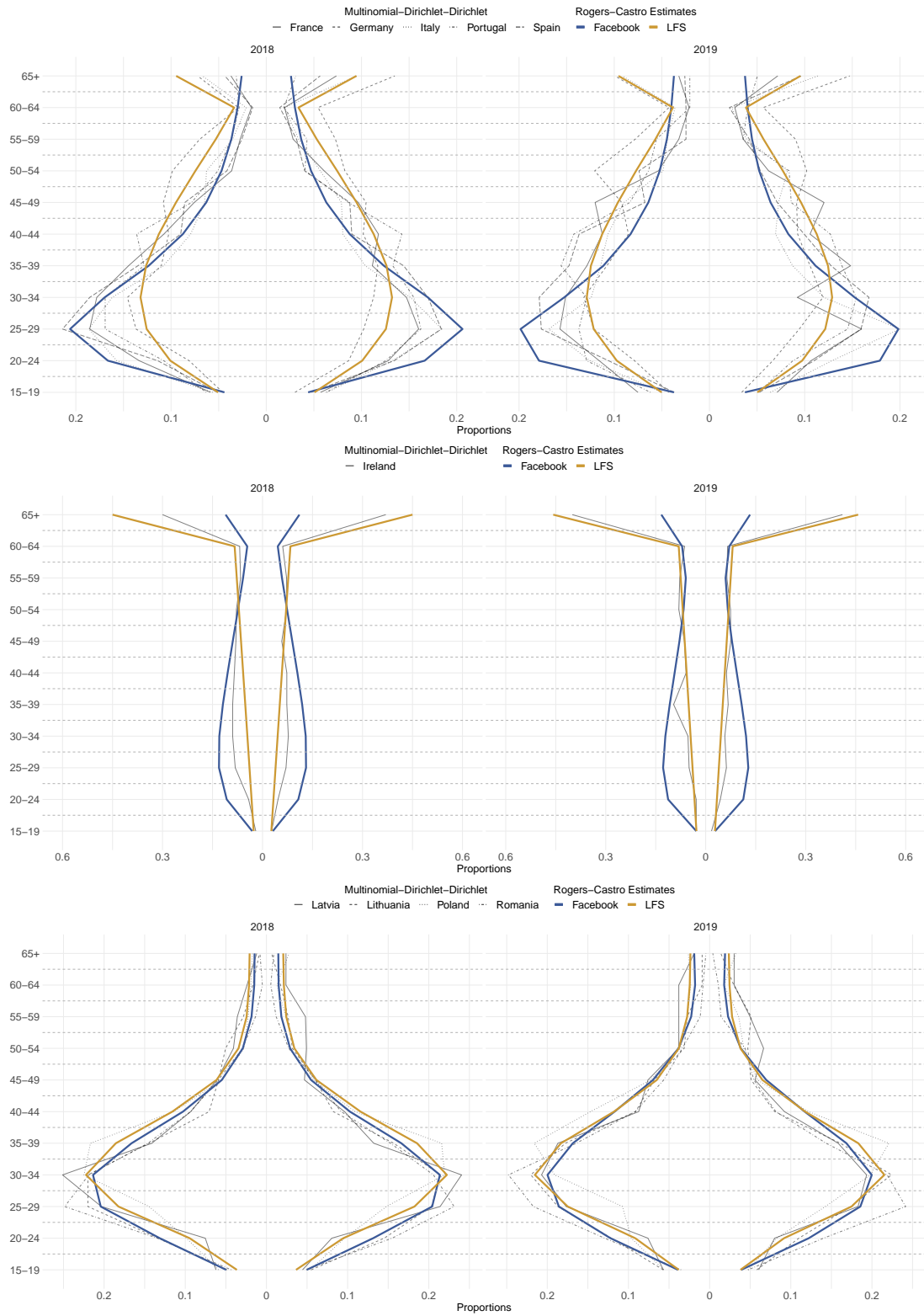


FIGURE 4.5: Population pyramids from the multinomial-Dirichlet-Dirichlet estimates harmonising the Rogers-Castro estimates from Facebook and the LFS.

4.5.1 Model for the Western and Southern European Migrants

The model estimates for the Western and Southern European migrants are shown in Figure 4.6. A reduction of the number of migrants between 2018 and 2019 is discernible. For the countries in this group, the largest number of migrants is in the 20-29 age group. There is, however, higher uncertainty in those age groups with larger numbers of migrants. The model estimates show more migrants across the age group (except for Germany in 2019) than the LFS.

4.5.2 Model for the Republic of Ireland Migrants

The population pyramids for the Irish migrants in the UK are presented in Figure 4.7. The estimates of the number of migrants are highest for the age group 65+ years old. It is clear from the 2019 results that the estimates are sensitive to sudden changes in LFS data collection, given the bump between 35-39 and 40-44 years old at this time. There is higher uncertainty in the estimates at older ages, and generally the model estimates more total migrants than the LFS.

4.5.3 Model for the Central and Eastern European Migrants

As shown in Figure 4.8, it is evident that Central and Eastern European migrants follow a similar pattern to each other. Compared to the previous two groups, the largest group of migrants are older than the Western and Southern European migrants, being between 25-29 and 30-34 years old. Combining the Facebook and LFS data helps to fill in gaps for older age groups from Latvia, Lithuania, and Romania, where the data is not reported by the LFS.

4.5.4 Sensitivity Analysis

Figure 4.9 shows a sensitivity analysis in the form of a quality of the estimates. Within the model previously shown the true stock by sex and the sum of true proportion by age and sex are computed. Figure 4.9 show that the sum of the true proportion by age

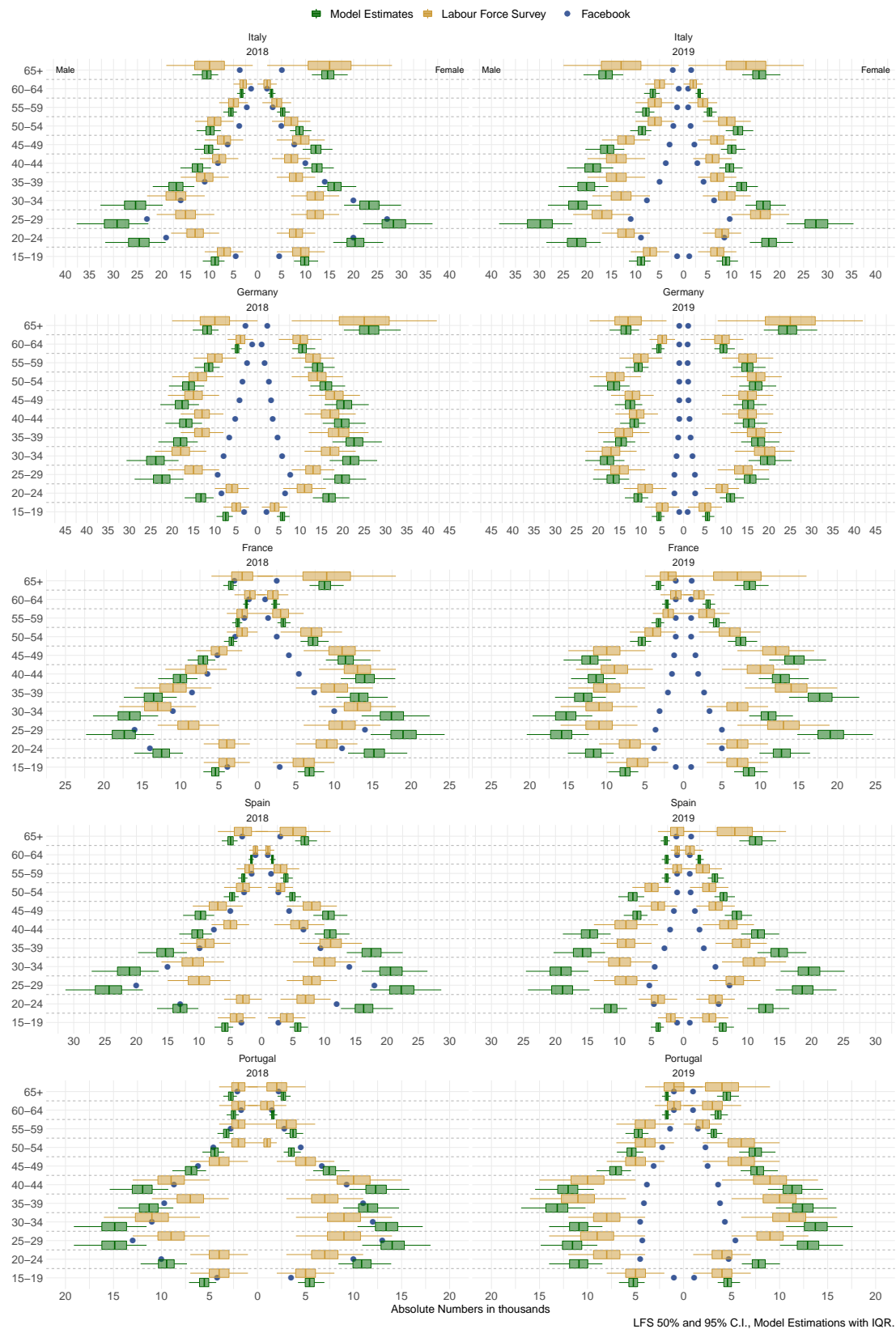


FIGURE 4.6: Population pyramids comparing the estimates from the model, the LFS and Facebook in the Western and Southern European group.

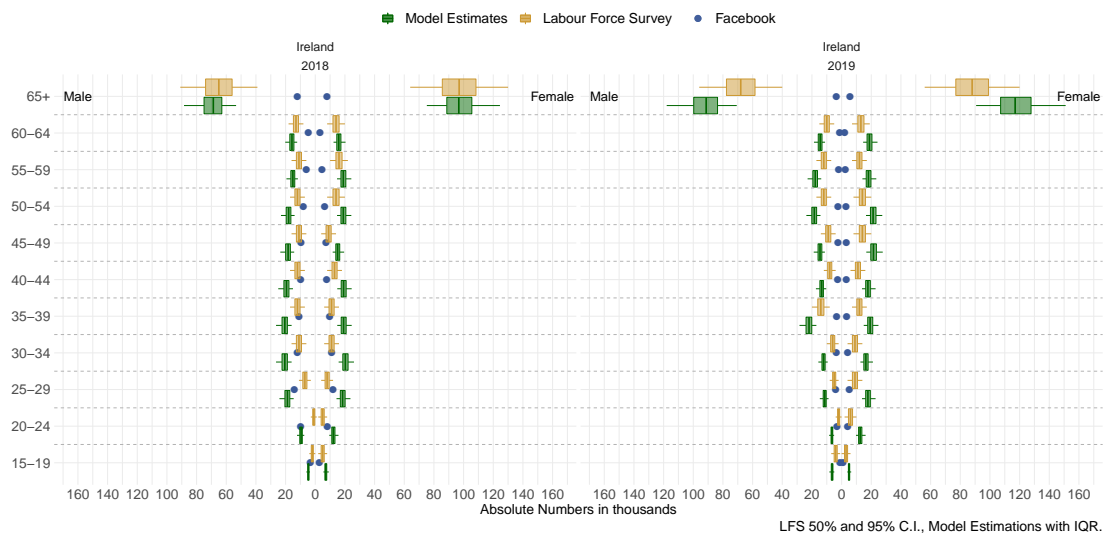


FIGURE 4.7: Population pyramids comparing the estimates from the model, the LFS and Facebook in the Republic of Ireland.

and sex is equal to the true stock by sex, which means that the multinomial-Dirichlet-Dirichlet works in redistributing the totals in age and sex categories. Indeed, the median and IQR have the same distributions across the two measures.

In Appendix B, Figure B.1 shows that the model is robust to changes of the value of the first Dirichlet. The model was specified with α equal to 0.1, 10, and 100. The model is robust to these changes. Moreover, it is presented an analysis of the residuals as a posterior predictive check of the model (Figure B.2). As suggested in Chapter 3, the model specified could be improve considering alternative distributions for the data.

4.6 Conclusions

In this chapter, the estimates from the Chapter 3 were extended by including an age and sex disaggregation. The use of a hierarchical multinomial-Dirichlet-Dirichlet model was proposed, utilising a Rogers-Castro model at one level of the hierarchy to borrow strength between the data sources. The model is flexible and the stocks data seems to show regularities similar to the flows data. First, three groups with different age and sex profiles were identified. The Western and Southern European countries have the largest number of young migrants, but the Eastern European countries migrants are

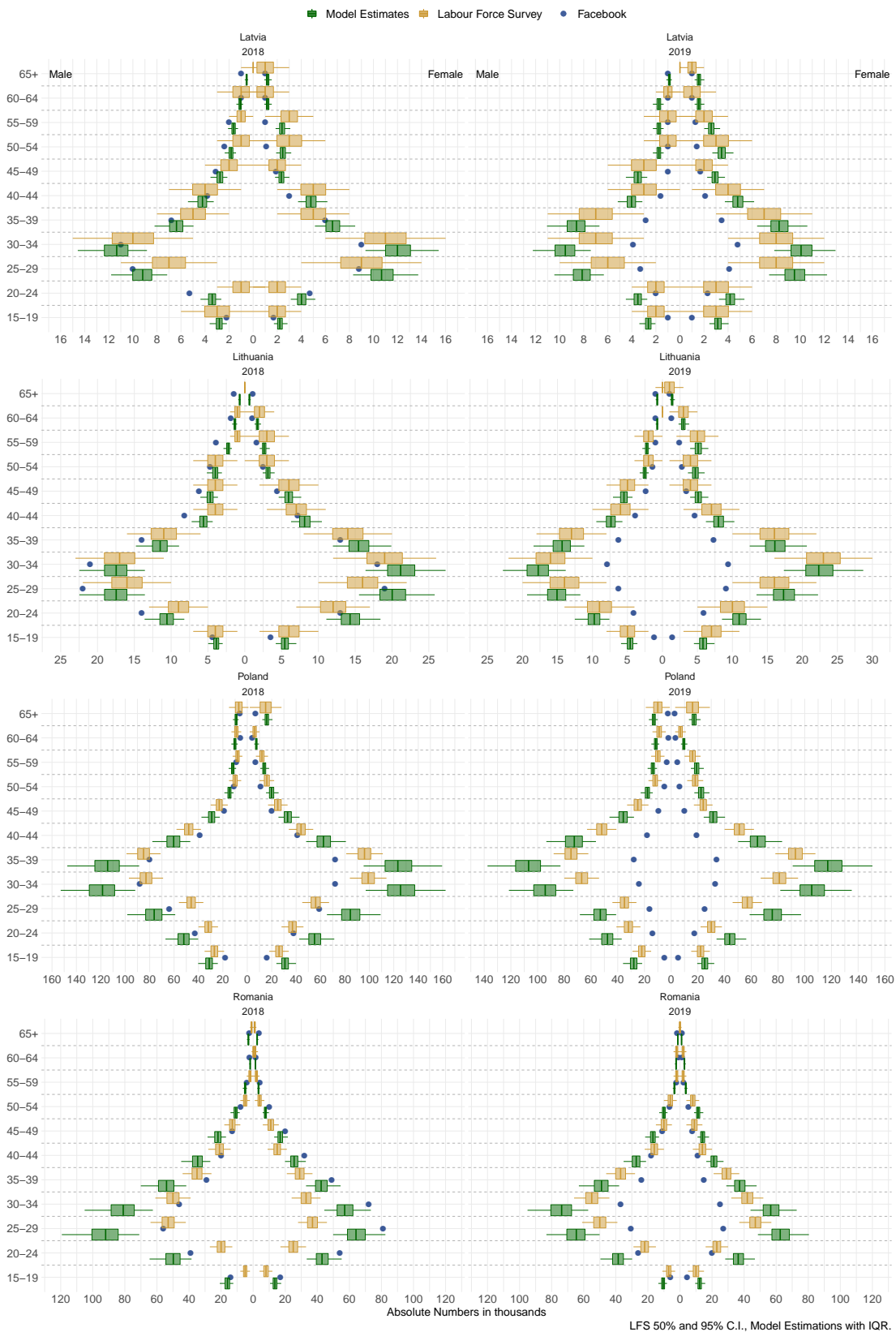


FIGURE 4.8: Population pyramids comparing the estimates from the model, the LFS and Facebook in the Central and Eastern European countries group.

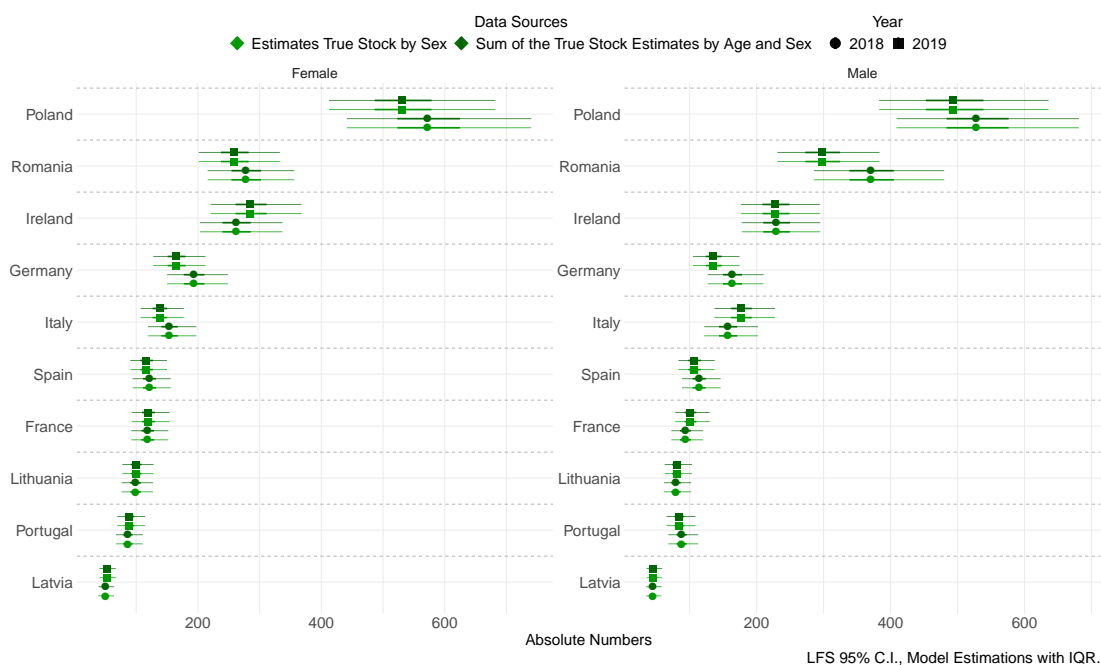


FIGURE 4.9: Comparison between the true stock estimates by sex and the sum of the true proportion by age and sex for 2018 and 2019.

only slightly older, with a lower total of older people. The Republic of Ireland has an overall older migrant population.

Secondly, Facebook has been shown to add coverage of younger migrants to the age structure. Chapter 3 showed how the estimates from the model report higher number of migrants than the LFS. This chapter, therefore, shows that European migrants are generally young, except in the case of those from the Republic of Ireland. The Rogers-Castro model application to stock data shows its limitations in this way for migrants from Ireland, which has a long history of migration and therefore an accumulation of old adults as migrants.

Harmonising the age and sex profiles from two (or more) data sources might add young migrants that are difficult to capture from traditional types of household surveys into the picture. Differing profiles of migrants might highlight different reasons to migrate as well as their length of stay. From this research, for example, it seems that Western and Southern Europeans might come to the UK for a period of their life, while Eastern Europeans might move to settle in the UK, as shown by their overall older age.

Thirdly, this chapter presents a way to harmonise age and sex profiles from two data sources. In doing so, we illustrate that it is possible to apply new profiles and fill gaps

in migration estimates from traditional data sources. The model is shown to be sensitive to changes in the age profile from the two data sources. The model was run simultaneously by year given the algorithm changes in 2019.

In this chapter, the framework previously presented was extended and a multinomial-Dirichlet-Dirichlet model was incorporated into the IMEM model to provide disaggregation by age and sex. It is crucial for policy makers to understand the age and sex structure of migrants in order to better plan policy and services for the population living in a country. The digital traces data combined with traditional data sources might help to increase the coverage of younger migrants in estimates. It further helps to answer to this thesis' research question of "What can Facebook Advertising data contribute to ONS migration estimates in a context where there is no "groundtruth" data against which model estimates can be validated?". By proposing a way to combine migration age profiles across two data sources, and to disaggregate the estimates of migrants by age and sex through a multinomial-Dirichlet-Dirichlet model, this chapter demonstrates the benefit of combining traditional and digital traces data sources to improve the coverage of particular groups of migrants in estimates. Furthermore, the model used is simple to apply and requires data that should be easily accessible from statistical offices.

The model presented in this chapter by age and sex and [Wiśniowski et al. \(2016\)](#) are different. On the one hand, the model presented is less complicated than [Wiśniowski et al. \(2016\)](#) and requires less parametrisation. On the other hand, in [Wiśniowski et al. \(2016\)](#) a measure of accuracy of the age and sex disaggregation is included. This part is missing in the suggested model, but could be included. A potential benefit of the model is that it runs in R using JAGS and could be reproduced in a short period of time without any licensed software.

The analysis was done on the yearly data for 2018 and 2019, which already show a change in the estimates between the two years. It would be interesting to understand how the Brexit transition might impact not only the number of European migrants living in the UK, but also their age profile. In the next chapter, this aspect is investigated using weekly data from the Facebook Advertising Platform.

Chapter 5

A Brexodus? Trends in the Numbers of European Migrants in the United Kingdom using Facebook Advertising Platform Data

5.1 Introduction

The two previous chapters of this thesis, Chapter 3 and Chapter 4, look at the use of a single yearly estimate of the number of migrants from the Facebook Advertising Platform. One of the highly discussed benefits of digital traces data is its timeliness: this type of data is often described as “*always on*” or “*timely*” or “[*having*] *velocity*” (Salganik, 2017; De Mauro et al., 2016). Previous research has used digital traces data to study migration (Zagheni et al., 2017; Gendronneau et al., 2019; Alexander et al., 2019, 2020), but has not yet explored the capacity of this data to study migration change at a detailed timely granularity. For example, Alexander et al. (2019) collected Facebook advertising data “*every two to three months*” to study the impact of Hurricane Maria on out-migration from Puerto Rico. This chapter verifies whether Facebook Advertising Platform data can be used without additional data sources to study the weekly trends in the number of European migrants in the UK. To do this, a simple Bayesian trend model with indicator variables for age, education, and country is used.

In this chapter, the aim is to analyse the effect of the uncertainty and threat related to the departure of the UK from the EU (known as Brexit) on the stocks of European migrants by age, education, and country over time. Brexit is an ongoing, long-lasting transition bringing economic and political change to the UK and the rest of the world. The end of the Brexit transition period and the reciprocal changes in EU and UK migration policies are still, at time of writing, uncertain. The new UK migration system aims to focus on skills through a point-based system (Home Office Government, 2020). Since 2016 the ONS has reported a positive but declining net migration of EU nationals to the UK (ONS, 2017). Paraphrasing the motto of Theresa May (former British Prime Minister), “*Brexit means Brexit*”, the interest in this chapter is in investigating whether “*Brexit means Brexodus*” of Europeans from the UK.

To evaluate this, weekly time series data of EU migrants in the UK was collected from the Facebook Advertising Platform starting in January 2018. However, in this chapter, only the period from March 2019 to March 2020 is analysed with weekly estimates of European migrants. The focus is on this time period for three reasons. First, an algorithm change in March 2019 affected the estimates of migrants worldwide (Palotti et al., 2020). In Chapter 3, the effect of the algorithm change in the Facebook data was

described; a drop in the expat estimates that was addressed with a parameter informed by a prior distribution. The second reason is that this algorithm change coincided with the beginning of the transition period for the UK's exit from the EU following the referendum. This makes it reasonable to focus the analysis on the time period after the algorithm change to avoid causal inference effects. And thirdly, the data after March 2020 is affected by the COVID-19 pandemic and lockdown restrictions on movement. It is interesting to study this period of time (March 2019 to March 2020) because the trend in stocks of EU migrants might give us insights into migrants' decision-making process about whether to remain or leave the UK. Four years have passed since the EU Referendum (known as the Brexit Referendum), allowing for a lag in the decision-making process of migrants regarding the decision to stay or leave the UK. Furthermore, disaggregation by age, education, and country might inform us of the differences in change of trends across these groups.

5.2 Background

On 22nd February 2016, David Cameron (then British Prime Minister), announced the date of the Brexit Referendum when UK citizens would be asked whether they wanted to continue to be part of the EU. The referendum on the UK's EU membership was held on 23rd June 2016; 51.9% of the voters chose to leave the EU. Since then the UK has continued to debate how exactly to leave the EU, while there have been three successive prime ministers (David Cameron, Theresa May, and Boris Johnson), two general elections (June 2017 and December 2019), and three Brexit extensions (March 2019, April 2019, and October 2019). On 23rd January 2020, the Withdrawal Agreement became law and the UK left the EU politically, but not economically, on 31st January 2020. This also marked the start of the transition period that ended on 31st December 2020. All the events mentioned are summarised in a briefing from the House of Commons ([Walker, 2020](#)); Figure 5.1 represents a timeline of the main events of Brexit in the last four years.

The entire Brexit process has been characterised by uncertainty; the political uncertainty has certainly had major repercussions on the lives of British citizens both in the UK and the EU, as well as EU migrants in the UK. At the time of writing, the UK is expected to change their EU migration policy at the end of the transition period. On

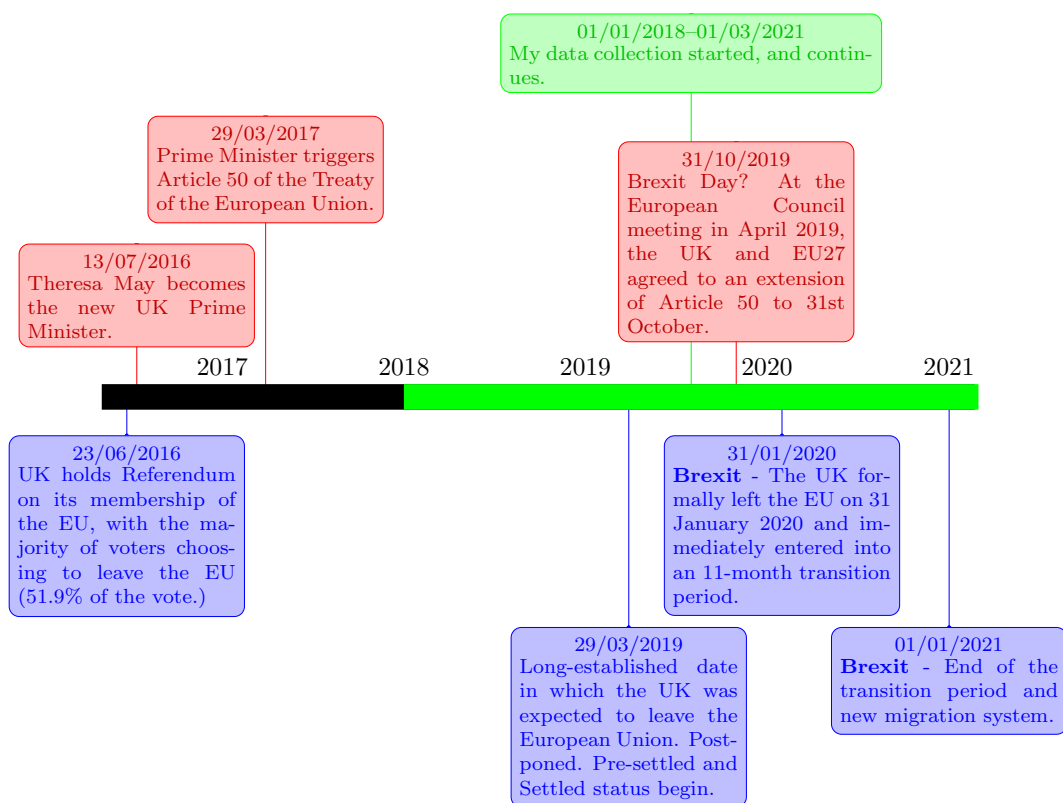


FIGURE 5.1: Some major events of Brexit and start of data collection.

1st January 2021, the UK will introduce a points-based migration system¹. This new system will end the era of the free movement of EU citizens to the UK and vice versa. EU migrants not living in the UK before the end of the transition period will thereafter have to apply for a visa to study and work in the UK. The EU migrants already living in the UK are qualified to apply for right of residence in the UK until 30th June 2021 under the EU Settlement Scheme (Home Office Government, 2020). The points-based migration system imposes visa costs on the applying migrant as well as the migrant’s employer in the UK, English language requirements, and a general salary threshold of £25,600 (Home Office Government, 2020). The system is intended to only give highly skilled migrants with higher levels of education the right to live in the UK. This change in policy might have an effect on the migration flows to and from the UK given that it drastically changes the rights of EU migrants to live and work in the UK.

Until the Brexit Referendum, the UK was an attractive destination within the EU migration system, with high levels of net migration (also shown in Figure 2.1). However,

¹<https://www.gov.uk/guidance/new-immigration-system-what-you-need-to-know>

several Quarterly Reports from the ONS (ONS, 2017, 2019b, 2020c) report that net migration from EU countries to the UK is still positive, but has been falling since 2016. The ONS also suggests that there is a decreasing number of EU migrants moving for work reasons to the UK. Working and studying are the two most popular reasons for EU migrants to move to the UK (Kierans, 2020). The latest NINo estimates show a decreasing trend in the number of EU migrants registrations compared to previous years, with 57% of the registration in the year ending March 2020 made by Romanian, Italian, and Polish nationals (DWP, 2020). The ONS is not the only source that is suggesting a loss in attractiveness for EU migrants to the UK labour market: indeed, LinkedIn data also suggests that professional migration from EU countries has decreased by 30% since 2016, and that recruiters have started to consider mostly individuals already in the UK ².

Drivers of migration are complex to study, and in fact there is not yet an overarching theory of migration (Bijak and Czaika, 2020; Willekens, 1983, 2019). The decision to migrate is linked to economic, social, and political factors related to both the area of origin and destination; i.e. push and pull factors that make one area more attractive than another (Lee, 1966). There are certain contexts that facilitate migration (Massey et al., 1993, 1999) because their environment is perceived as suitable for new opportunities (Bijak and Czaika, 2020). However, as suggested by Bijak and Czaika (2020), “migration drivers are generally not static but change dynamically”. The Brexit Referendum and transition process are currently functioning as a “shock in the migration system” (Bijak and Czaika, 2020), and will end with a drastic change in the migration system in place. This uncertainty might be seen as an obstacle for migrants, influencing their decision-making process. It must be noted, however, that some migrants clearly still see opportunities in moving to the UK, as although migration from the EU has visibly decreased, it has not stopped entirely.

In this chapter, Facebook Advertising Platform data denoting age, education, and country of origin is used to investigate whether there is a decreasing trend in the stock of EU migrant MAUs in the UK and to compare groups by their trends. The LFS might not be suitable as a traditional data source to investigate this change, because it cannot be disaggregated at such small time granularities.

²<https://economicgraph.linkedin.com/blog/linkedin-workforce-insights--how-has-brexit-affected-the-uk-labo>

5.3 Data

The data was downloaded from the Facebook Advertising Platform every week since January 2018. pySocialWatcher was used to query the Facebook Marketing API (Araujo et al., 2017). The interest was in downloading the number of migrants disaggregated by age, education, and country of origin. The analysis did not include the sex disaggregation as Chapter 3 and Chapter 4 underline that there does not seem to be strong divergence across sex by country. Therefore, further dividing the estimates by sex in this analysis might lead to a loss of power in the Facebook estimates that are already disaggregated by age, education, and country of origin. As a consequence, the data is separated into categories of:

- **age groups:** 15-19, 20-29, 30-39, 40-49, and 50+ years old;
- **education levels:** Secondary (No Degree, In High School, High School), Tertiary (In College, In Grad School, Graduated), and Unspecified;
- **countries:** France, Germany, Ireland, Italy, Latvia, Portugal, Poland, Romania, and Spain.

These data series offer an opportunity for researchers; despite their limitations, they bring timeliness and a wider coverage of the migrant population. Digital traces data is, however, not representative of the entire population and is based on self-reported and algorithmically confounded variables (Cesare et al., 2018). The education variable is an example of a self-reported variable, in which the “Unspecified” category makes it difficult to fully interpret the distribution in terms of education of the migrant Facebook users. In Figure 5.2, the data from the LFS and the Facebook Advertising Platform are compared through a logarithmic transformation. The data from the LFS represents weekly estimates for the years 2017, 2018, and 2019. It shows a stable trend in the stock estimates of migrants across the nine countries studied, with only a slightly decreasing trend for Latvia and a slightly increasing trend for Romania. Figure 5.2 also shows the total number of estimated migrants from each country of origin studied, using data from the Facebook Advertising Platform from the start of the download in January 2018 until July 2020. It is evident that there were two break points, in March 2019 and March

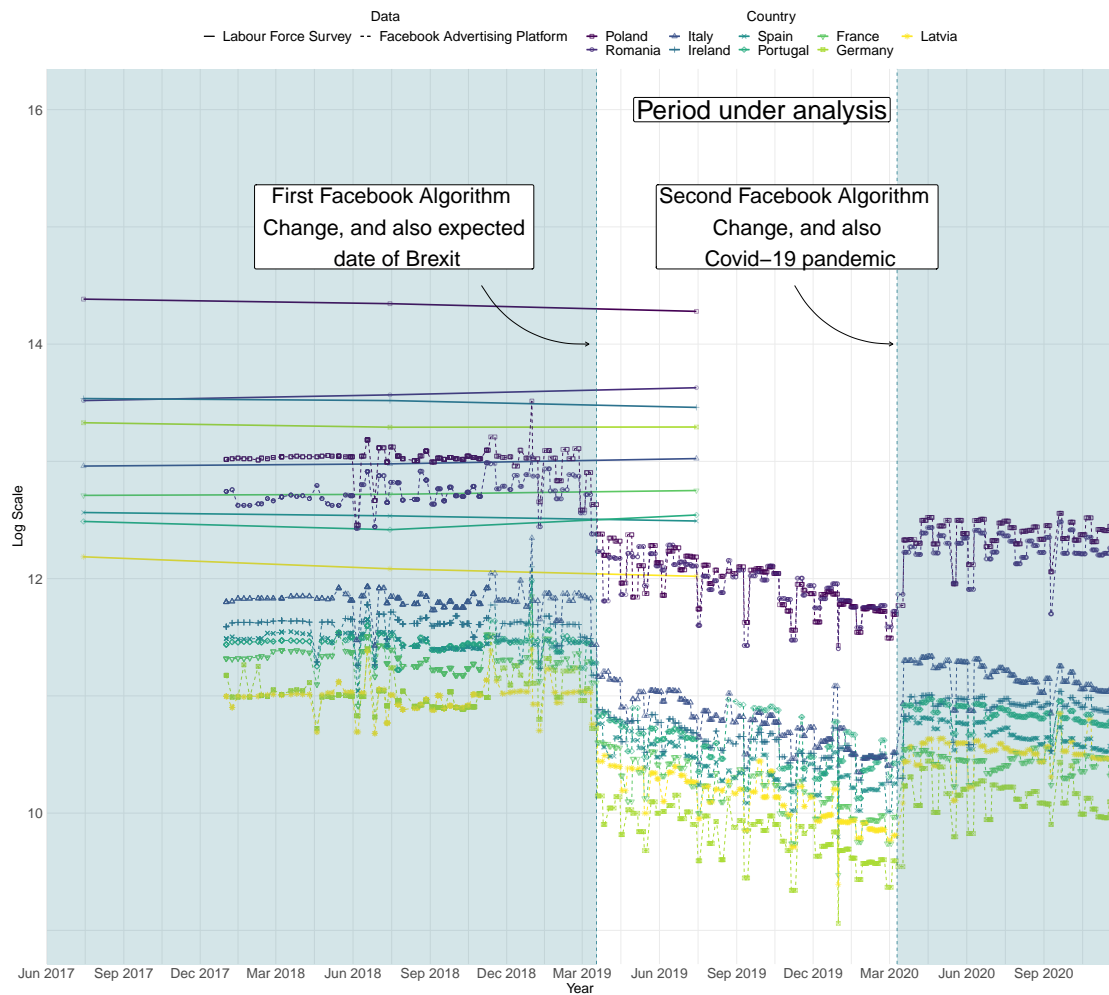


FIGURE 5.2: Country time series with yearly data from the Labour Force Survey for 2017, 2018, and 2019, and with weekly data from the Facebook Advertising Platform from January 2018 to November 2020.

2020 respectively, in which the estimates provided by the Facebook Advertising Platform were affected by an algorithm change. Since the first algorithm change occurred in March 2019 when Brexit was supposed to take place, and the second algorithm change occurred in mid-March during the COVID-19 pandemic, this chapter focuses solely on the data in between these two time periods in order to study the change in trend. From Figure 5.2 it seems that digital traces data, despite their limitations, in comparison to the LFS data, might be able to capture changes in migration stocks in a timelier manner. The time series have been cleaned by removing 0, 1000, and counts that were a standard deviation out from a weekly mean computed by age, education, and country.

5.4 Methodology

A simple Bayesian trend model with indicator variables for age, education, and country was used to analyse the changes in the number of migrants. The trend equation m_{aeit} is the mean of y_{aeit} , a log-normal distribution with precision parameter t . The precision parameter τ is *a priori* distributed as a Gamma with both shape and rate parameters equal to 0.01.

$$y_{aeit} \sim \text{Log-Normal}(m_{aeit}, \tau) \quad (5.1)$$

$$m_{aeit} = c + c_1^T d_a + c_2^T d_e + c_3^T d_i + c_{13}^T (d_a \times d_i) + c_{23}^T (d_e \times d_i) + (b + b_1^T d_a + b_2^T d_e + b_3^T d_i + b_{13}^T (d_a \times d_i) + b_{23}^T (d_e \times d_i)) \times t \quad (5.2)$$

$$\tau \sim \text{Gamma}(0.01, 0.01) \quad (5.3)$$

The parameters $\mathbf{c} = (c, \dots, c_3)$ and $\mathbf{b} = (b, \dots, b_3)$ are assumed to be normally distributed $N(0, 0.0001)$, with mean 0 and precision 0, 0.0001.

The trend m_{aeit} is divided into two parts: the c component, which is the intercept of the trend that describes the initial magnitude, and the b component, the slope of the model that describes the gradient of the decline. The parameters c and b are the overall effects, c_1^T and b_1^T are vectors of the parameters for age, c_2^T and b_2^T are for education, and c_3^T and b_3^T are for country of origin. The vectors d_a , d_e , and d_i contain the dummy indicator for the variables of age, education, and country. The reference category group for age is 15-19 years old, for education it's Secondary education, and for country it's Italy. Interactions between age and country, c_{13}^T and b_{13}^T , and education and country, c_{23}^T and b_{23}^T , are included in the model. The choice of the interactions in the model was driven by an initial descriptive analysis of the residuals from a model with all the main effects included. In Appendix C, the residuals are reported for the model without interactions (Figure C.1) and with interactions (Figure C.2). After analysing the residuals, the following interactions were included: Unspecified Education, Tertiary Education, and age groups 20-29, 30-39 and 40-49 for Romania and Poland. Moreover, the two models are compared using the Deviance Information Criterion, a generalisation of the Akaike Information Criterion (AIC), which is used to compare hierarchical Bayesian models

(Gelman et al., 2013). In the model without interactions the mean deviance is 217041 (penalized deviance 217072, penalty 30.97), while for the model with interactions the mean deviance is 205705 (penalized deviance 205756, penalty 51.19. Therefore, the interactions helps, even if just slightly, to reduce the mean deviance.

5.5 Analysis

5.5.1 Descriptive

Figure 5.3 and Figure 5.4 present the raw estimates from the Facebook Advertising Platform data in the period under analysis. In Figure 5.3, the Facebook weekly time series data is disaggregated by age group and country. An exponential decline is evident, especially in the age groups 20-29 and 30-39, but less pronounced in older age groups. In Figure 5.4, the weekly estimates are presented disaggregated by education level and country. The trends are negative across all the educational levels. Given the large estimated numbers in the unspecified education level, it is evident that Facebook does not provide educational information for a large group of profiles.

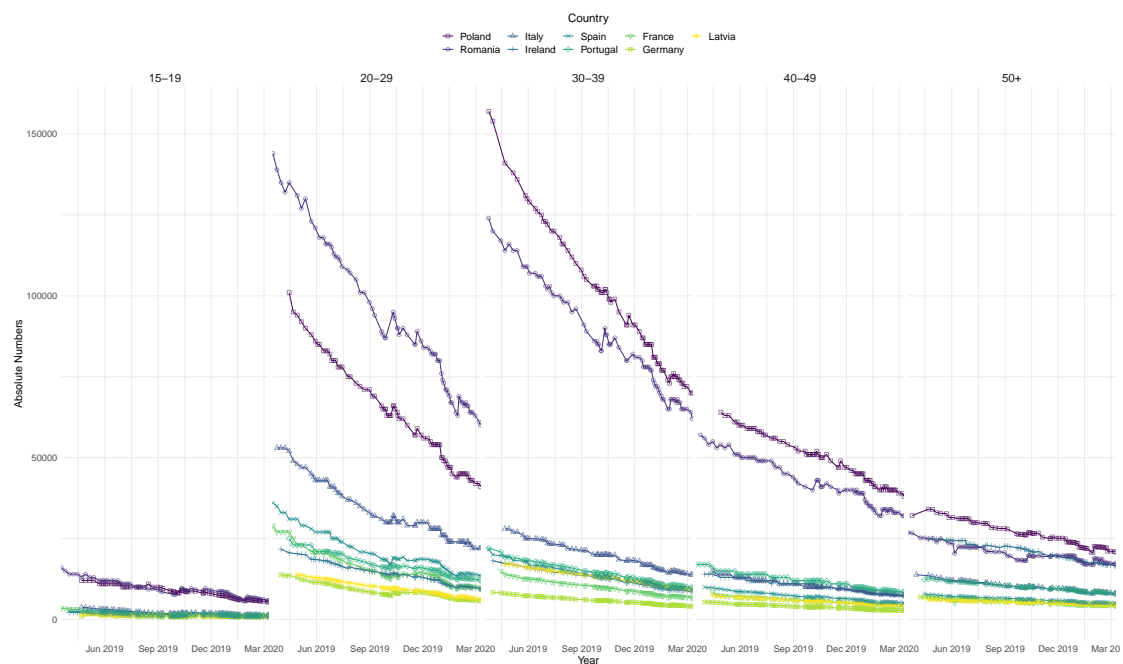


FIGURE 5.3: Country time series with weekly data from the Facebook Advertising Platform from January 2018 to July 2020 by age groups and countries.

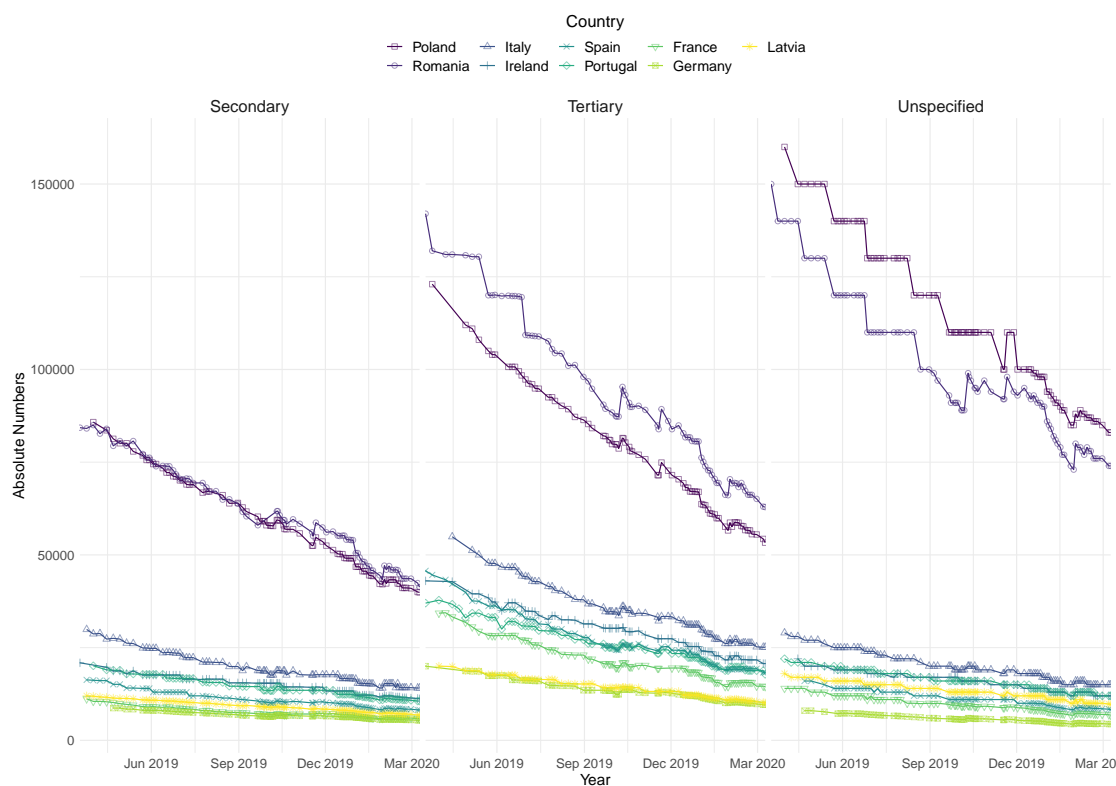


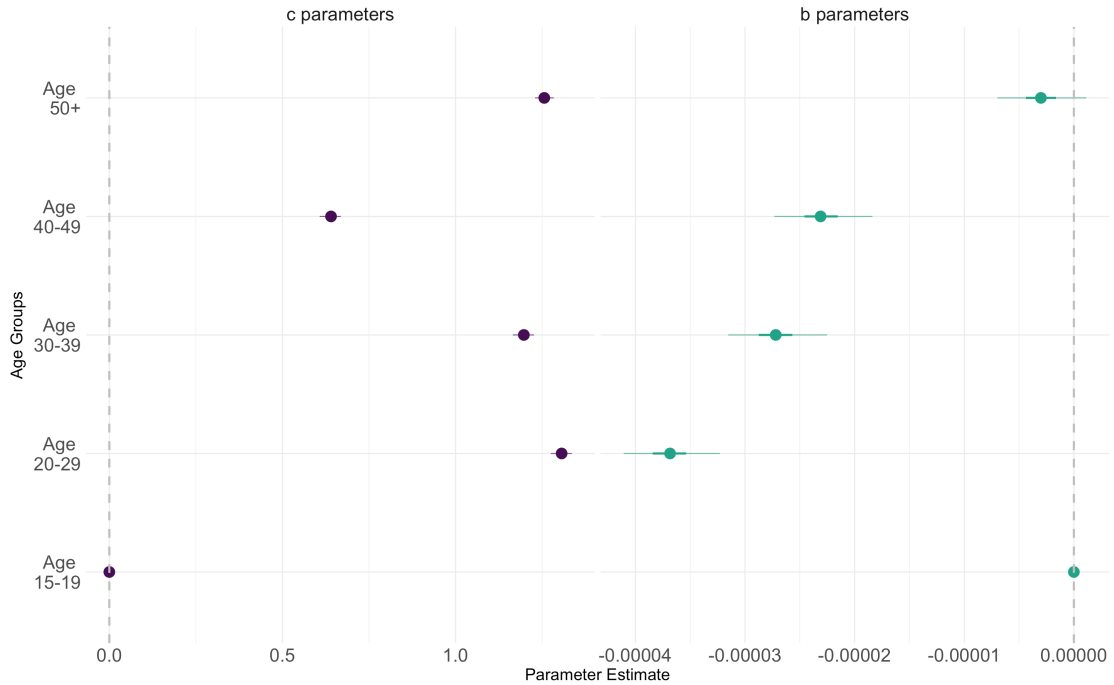
FIGURE 5.4: Country time series with weekly data from the Facebook Advertising Platform from January 2018 to July 2020 by education levels and countries.

5.5.2 Results from the Model

The model was estimated in R using JAGS (Plummer et al., 2016). The model contains 11,880 observations of the number MAUs by country, age, and education. The model is run with three chains, 10,000 interactions, 1,001 burn-in and 10-fold thinning (i.e. every 10th value of the chains is kept and all other values are discarded to avoid autocorrelation). All the chains and parameters of the model converge. In Appendix C, Table C.1 and Table C.2 report the values of each of the parameters, \hat{R} and \hat{n}_{eff} . Table 5.1 reports the main effect of b and c . The parameter c indicates the initial magnitude level of migrants (i.e. the model intercept) in comparison to the reference category, while the parameter b indicates the direction of the regression slope in comparison to the reference category. The estimated median of b is negative, indicating a decreasing slope over time of the log-transformed stocks of migrants. In Table 5.1, the two main effects of b and c are presented. The interesting value here is b , as it indicates that the slope has an overall negative trend across all categories.

TABLE 5.1: Distribution of the main effect of b and c .

	2.5%	25%	50%	75%	97.5%	\hat{R}	\hat{n}_{eff}
b	-5.45×10^{-6}	-1.92×10^{-6}	-1.32×10^{-6}	1.82×10^{-6}	5.80×10^{-6}	1.02	107
c	8.00	8.03	8.04	8.06	8.08	1.01	97

FIGURE 5.5: Values of c , b parameters estimated from the model for age.

In Figure 5.5, the respective estimated distributions of the different age groups are shown in comparison to the reference category of 15-19 years old. Looking at the b effect, the size of the age group that is decreasing fastest in comparison to the reference group is the 20-29 age group, followed by the 30-39 age group. The two groups decreasing fastest are also the two groups with a higher initial level in comparison to the 15-19 age group. The 50+ age group has a similar slope to the 15-19 age group, but a higher initial level.

Figure 5.6 shows an interesting part of the analysis in terms of education. The trend is decreasing fastest for the unspecified category, followed by the tertiary education and then secondary education level. Figure 5.6 shows that the category with the highest number of migrants is Tertiary Education, followed by Secondary and Unspecified.

In Figure 5.7, the effects of various individual countries of origin are shown. The reference category is Italy, the third largest country of origin after Poland and Romania

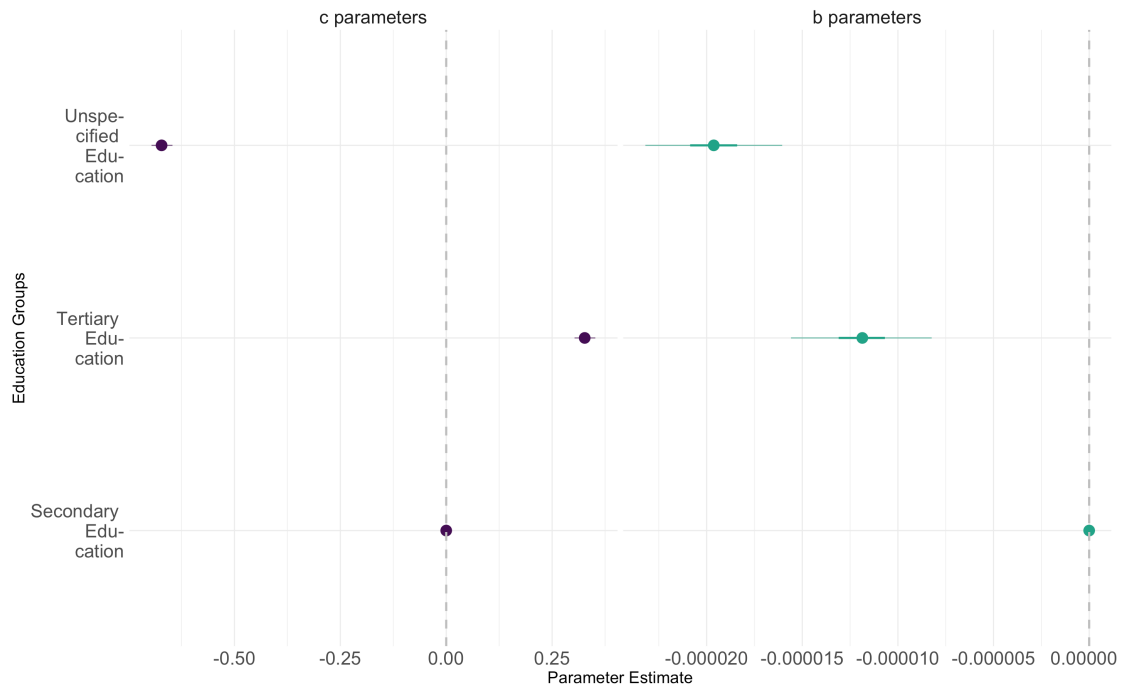


FIGURE 5.6: Values of the c and b parameters estimated from the model for education.

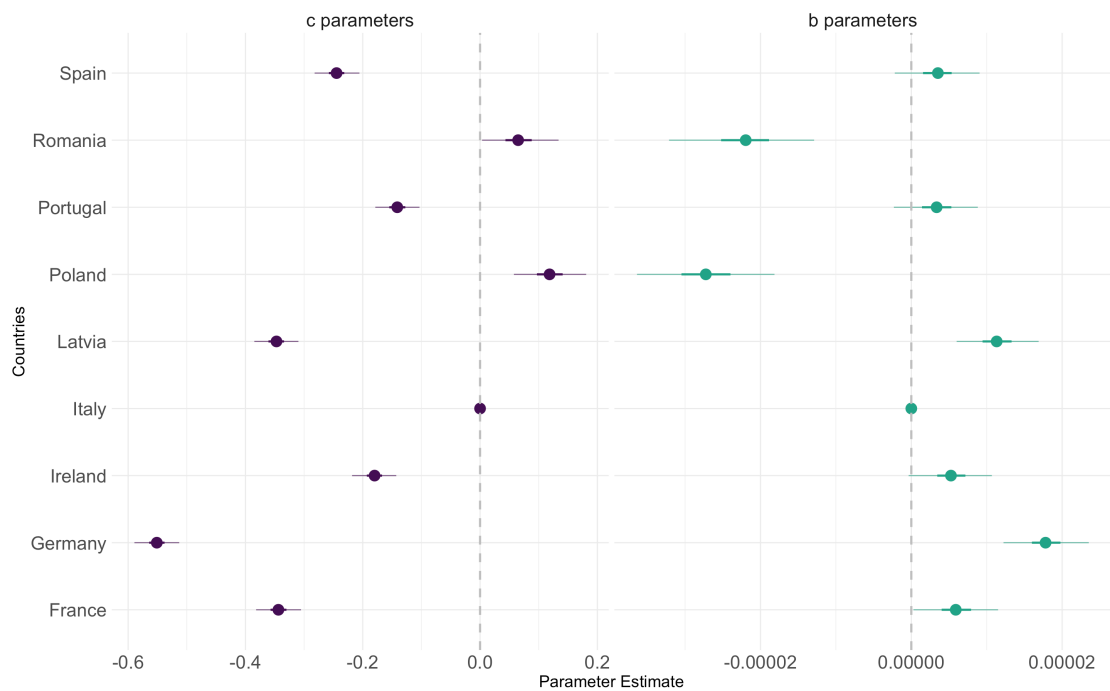


FIGURE 5.7: Values of the c and b parameters estimated from the model for countries.

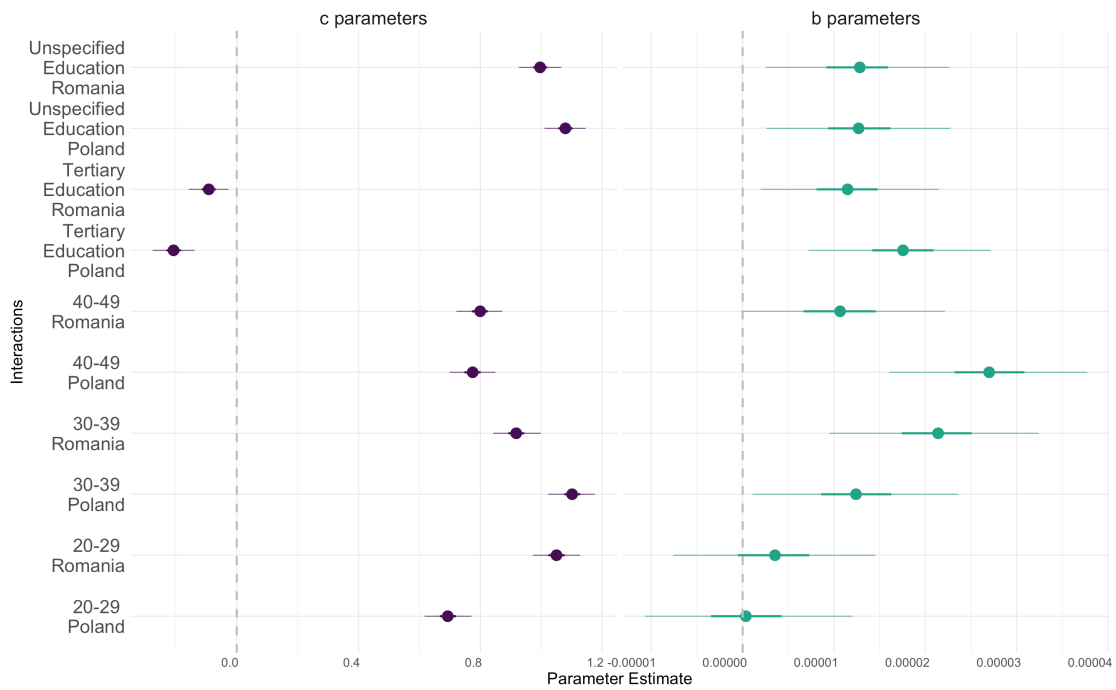


FIGURE 5.8: Values of c , b parameters estimated from the model for the interactions for Poland and Romania.

which have a faster decrease in comparison. The other European countries have positive outputs, which means that are not decreasing as fast as Italy. As introduced in Section 5.4, interactions for Poland and Romania are included in the model, the outputs of which are shown in Figure 5.8. The interactions parameters have positive outputs, meaning that for Poland and Romania the decline is not as fast as for the reference category and the overall effect of education, age, and country.

5.6 Conclusions

The analysis of Facebook data has shown that the stocks of migrants in the UK are decreasing. These changes are reported in relations to the chosen reference categories. Overall, the age groups that are decreasing fastest are 20-29 and 30-39 in comparison to the 15-19 age group. These two age groups are also the two age groups with the highest number of migrants. Additionally, the migrant stocks are decreasing for all education levels. Although it is difficult to assess the quality of the education variable provided by the Facebook Advertising Platform, the Tertiary Education group appears to be decreasing faster than Secondary Education. Given that the UK's new migration system

aims to attract highly-skilled and educated migrants, this is not a good sign for the current attractiveness of the UK for this targeted group. In terms of countries of origin, the numbers of migrants from the largest European countries, Poland and Romania, are decreasing faster than Italy and the other European countries in the analysis. The interactions effects on Poland and Romania slightly slow down the decrease for both age groups 30-39 and 40-49, as well as for Unspecified and Tertiary Education groups.

While digital traces data is timely, it is unstable, and therefore we cannot completely substitute traditional data sources with them. We can, however, obtain useful indicators from digital traces data, in this case the trend of numbers of migrants in the UK. In this chapter, weekly estimates from the Facebook Advertising Platform were used. The main result is that, for the nine European countries included in the analysis, the stock of migrants living in the UK was found to be decreasing between March 2019 and March 2020. Figure 5.2 shows that the decreasing trend appears to continue in the period after March 2020, which is not included in the primary analysis. The current analysis is exploratory in nature. It might be interesting to explore different specifications of the model. For example, it could be interesting to use random effects, and different smoothing techniques such as time series or exponential smoothing.

So, does Brexit mean Brexodus? It does seem that there is a declining trend in migrants coming to and living in the UK. The decline started after the expected Brexit date in March 2019 (Figure 5.2), however as this coincided with an algorithm change in how the Facebook estimates were produced, it is difficult to establish cause. The UK is clearly losing its attractiveness for the migrants living in the UK as well as new migrants coming to the UK, and this might be linked to the ongoing uncertainty surrounding Brexit. Although the LFS data shows a decline in their estimates, it is not as pronounced as the decline shown by the digital traces data. This might be linked to the intrinsic timely nature of digital traces data.

The useful result of this chapter lies in showing the trend of change, rather than providing an exact percentage change or estimate of the change over time. The analysis crucially shows that digital traces data might be used to monitor change over time in the stocks of migrants present in a country. This aspect is especially important in current times, given that the IPS has been paused since the beginning of March 2020, NINo are not provided to migrants who are not on a visa, and that the LFS might have more

issues with representation as many European migrants may have decided to leave the UK during the COVID-19 pandemic. Working from home in the UK during 2020 has been at times enforced and other times strongly encouraged in the UK, meaning migrants may have decided to move back to their home countries or might even have been stuck in their home countries due to travel restrictions during the pandemic.

From the trends presented above, it seems that the UK is losing young migrants from European countries. In addition, the Tertiary Education group is decreasing at a faster pace in comparison to the Secondary Education group. Facebook is potentially not, however, the best social media platform to use to investigate the level of education of migrants; LinkedIn might be a more appropriate data source to study skills of migrants and investigate further the importance of digital traces data in assessing the effect of Brexit on the number of migrants in the UK. Nevertheless, it will be interesting to continue following the change in migrants coming from the different education categories over time, as the UK looks to change their migration system on 1st January 2021 to the new point-based system outlined.

Although the COVID-19 pandemic may continue to have an effect on migration behaviour in 2021 and it might take some time to understand the effect of the UK's migration policy change, some signs of a loss in attraction to the UK are already apparent. This chapter's analysis could be additionally enhanced by analysing a longer time series of data once the data from the two Facebook algorithm changes is corrected, if the nature of these changes becomes known.

Chapter 6

Conclusions

6.1 Summary and Contributions

This thesis has demonstrated that digital traces can contribute to the study of stocks of European migration to the UK. With three pieces of analysis, an illustration of how the Facebook Advertising Platform might be used to complement traditional data sources even where there is no “ground truth” is presented. This illustration might be expanded to a broader range of digital traces data sources beyond Facebook. This thesis contributes to the *learning process* advocated by Willekens (1994, 2019) by analysing the use of digital traces data for migration research with attention paid to the varying definitions and biases of the different data sources. Moreover, through suggesting a way to combine digital and traditional data sources, trends of change are inferred from these timely data sources.

The model suggested in Chapter 3 builds on the Integrated Model of European Migration, a Bayesian model combining data on flows at the European level suggested by Raymer et al. (2013). In this thesis, the IMEM was repurposed to estimate the true stock of migrants in the UK from 20 different European nationalities, combining digital traces data with survey data. The UK was selected as the destination country in this thesis, as the traditional migration data sources have been particularly criticised for their quality (Coleman, 1983; Kupiszewska and Nowok, 2008; Kupiszewska et al., 2010; Rendall et al., 2003). Digital traces can therefore contribute to improving the quality of these estimates more significantly than in countries with more extensive migration records. Given that Facebook usage is high across Europe and that the LFS is a Europe-wide survey, the model suggested could be expanded to the rest of Europe, as already proposed by Gendronneau et al. (2019). Most of the traditional migration data sources are biased in similar ways to digital traces data. This points to the need for an accurate data assessment in terms of definitions, bias, and coverage of all the data sources that enter the model. Although the estimates produced are uncertain, the model is flexible enough to be updated with new available information once it becomes available to ensure it is as comprehensive as possible.

Data on migrants disaggregated by age and sex is not only important for policy makers, but also for forecasting future populations size. In Chapter 4, an extension to the

model from Chapter 3 was proposed, including an age and sex disaggregation. An extension of the IMEM by age and sex has already been proposed by [Wiśniowski et al. \(2016\)](#); in this thesis, an alternative approach is considered. The Rogers-Castro model ([Rogers and Castro, 1981](#); [Rogers and Watkins, 1987](#)) was repurposed to create a harmonised migration stock schedule, and then a multinomial-Dirichlet-Dirichlet model was used to combine and distribute the harmonised schedule of migration by age and sex. This model is inspired by the [Caussinus and Courgeau \(2010\)](#) use of the approach in paleodemography.

As digital traces and traditional data sources have different levels of coverage of the migrant population, harmonising the migration schedule from each might provide a larger coverage in migration estimates. The different data sources complement each other, as while digital traces data might omit the older section of the migrant population, traditional data sources like surveys may miss the young and more mobile section of the migrant population due to their sampling framework. The estimates are more uncertain for the age groups in which there are more migrants.

Velocity and timeliness are highly appreciated characteristics of digital traces data. In Chapter 5, weekly time series data from the Facebook Advertising Platform was analysed to infer whether the trends produced might capture migration changes faster than those from traditional data sources. This paper stresses the importance of using digital traces data to capture fast changes in trends that traditional data sources are not systematically constructed to grasp. This advantageous aspect of digital traces data has already been used to analyse migration following a natural disaster ([Alexander et al., 2020](#); [Martín et al., 2020](#)) or political crisis ([Palotti et al., 2020](#)), and now can be used to monitor changing stocks of migrants in relation to uncertain international events such as Brexit.

Chapter 3 and Chapter 4 answer the overarching research question of this thesis by producing an estimate of the number of migrant stocks at the yearly level, while the Chapter 5 expands the time granularity of the digital traces. This thesis contributes both to methodological refinement and our knowledge of how digital traces can be used.

6.1.1 Limitations

Despite the contributions of the three research chapters to advancing knowledge on the use of digital traces data in demographic research, there are limitations related to this form of data and the methods used. Digital traces data often lacks a clear definition of what is being measured. In demographic research it is common to use strict standards to define measures of interests. As digital traces data is obtained from private companies, it is somewhat blurry what the algorithms behind these estimates are using, for example, to define an Expat or how long it takes to consider a user as an Expat. A clearer understanding of the construction of these measures would allow these data sources to be included in models with more precision. Another aspect that is limiting our understanding of the Facebook estimates, is that the process behind the algorithm providing the MAUs estimates is not clear. Additionally, social media companies rarely transparently inform the public of changes to their algorithm infrastructure. For example, throughout this thesis there is mention of Facebook's algorithm change in March 2019 related to the Expat variable; as a result, if the models in Chapter 3 and Chapter 4 were run again for 2020 the prior distributions for the algorithm change would need to be reconsidered. A further limitation of this form of data is issues around its privacy. Although the data used in this research was aggregated and anonymised, there is still the need for rigorous safeguarding of privacy to maintain high ethical standards of research when using digital sources. Therefore, the main limitation of digital traces can be summarised as a lack of transparency in terms of its definitions, algorithms, and ethics.

6.2 Conclusions

Focusing on the UK, this thesis provides a deeper understanding of the use of digital traces data in the context of migration research. The contribution of this thesis can be summarised into three main points. Firstly, it highlights that it is necessary to understand the data generation process through investigation. Secondly, that digital traces data should be combined with, and not substitute, traditional data sources. As described, digital and traditional data both have pros and cons, and combining them

might lead to a clearer picture. And thirdly, that digital traces is more timely than traditional data sources, and therefore, despite being biased, might be used to infer trends of change and nowcast short-term changes.

The COVID-19 pandemic has made evident the importance of obtaining estimates of migration from multiple data sources. The IPS and some other administrative data sources, for example, have not been running since March 2020. This makes clear that administrative data, surveys, and digital traces data should be combined to form a clearer picture of migration. The models presented in this thesis might be further expanded to include more data sources and countries, and there is an opportunity for new research to focus on understanding how to use digital traces for the study of flows of migrants as well as stocks.

6.3 Future of Demography

The International Scientific Union of Population Studies (IUSPP) reports several definitions of *demography*; these converge to suggest that demography is the “*scientific or statistical study of the human population*” (IUSPP, 2014). Pavlík (2000) published a collection titled “*Position of Demography Among Other Disciplines*”, in which fifteen demographers discussed their views on demography as a discipline. Kohler & Vaupel’s contribution (Pavlík, 2000) defined methods and data as the two foundations of demography; they position demography at the centre of a rectangle with connections to the four other related disciplines: mathematics and statistics, bio-sciences, social sciences, and public policy. They stressed that demographic studies have an important role in shaping policies to influence or shape population change. In the chapter by Coleman (Pavlík, 2000), a similar but more vague view of demography is described; the methods, models and data of formal demography are described as the core of the discipline, which is also shaped by other fields. Overall, the collection concludes demography is an interdisciplinary discipline across statistics, social and biological sciences.

In the first decade of the 2000s, demographers started to use new tools provided by a wider access to computers, and began to take advantage of microsimulation and

Agent-based Modelling techniques (Billari et al., 2006). In more recent years, demographers have become aware of developments in data science in the shape of unstructured datasets not designed for scientific research. Acknowledging computer science and data science is twofold process in terms of adopting the methodology from computer science (e.g. simulations and network science), and also managing new data sources made available by the expansion of the digital world.

In *“What is Digital Sociology?”*, Selwyn (2019) examines the concept of digital sociology, describing it as a sub-discipline of sociology that enhances traditional sociological research. The recent advancements in the use of digital traces data in demography has fallen under the umbrella term of digital demography, however, given that demography has always been interdisciplinary, this type of new research should just be considered a natural development and evolution of the discipline. Demography can harness new types of data from the digital world and use its established academic rigour to study them. Figure 6.1 aims to update and combine Kohler & Vaupel’s and Coleman’s attempts to describe demography (Pavlík, 2000). The discipline of demography is at the centre of a circle surrounded by the disciplines that shape it. Data science and computer science have been added to the list of disciplines having an influence on demography. Furthermore, it is stressed that the two pillars of demography are its methods and data.

The expansion of the digital world brings new opportunities for demographic research in terms of its implications and methodologies. Certainly, more and more people are using digital technologies and family formation dynamics have been shown to be changing due to online dating (Hitsch et al., 2010; Potârcă and Mills, 2015; Danielsbacka et al., 2020). Additionally, other demographic events such as planning to have a child might be informed by digital mechanisms such as fertility tracking apps that monitor ovulation (Earle et al., 2020; Zwingerman et al., 2020). New sources of data might lead to new data-driven discoveries that can update or create new theories (Billari and Zagheni, 2017). Indeed, the Digital Revolution has also an enabling effect on migration and intentions to migrate, given that the Internet supports potential migrants providing information on potential destinations (Pesando et al., 2021). Moreover, digital traces data might contribute in expanding definitions of international migration thanks to its time granularity (Fiorio et al., 2017). It would be possible to use traditional demographic methodologies to study this data, repurposing them for unstructured data sources.

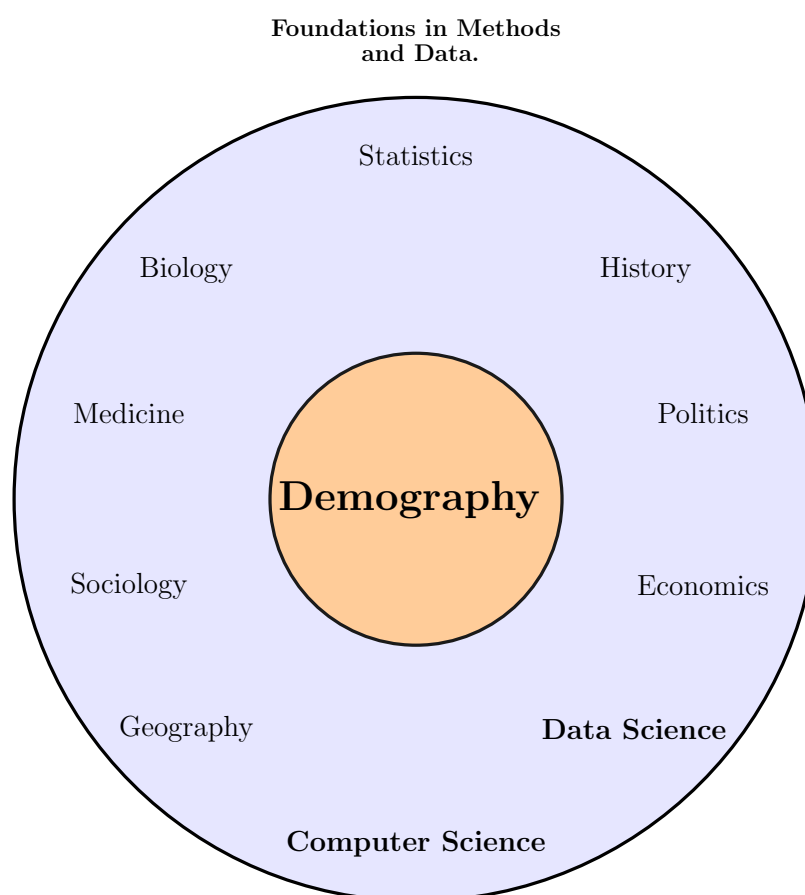


FIGURE 6.1: Updated drawing of a suggested structure of demography combining Kohler & Vaupel and Coleman's contributions in Pavlík (2000).

Demography is still largely based on traditional data sources that have encountered problems during the COVID-19 pandemic, such as censuses, administrative sources, and surveys. In fact, some censuses have been postponed (as in India, Scotland, the Republic of Ireland, and other countries), while many administrative data sources¹ and surveys have been suspended. This might be taken as incentive to innovate and understand how to utilise new data infrastructures when population change seems to occur quickly in events such as a pandemic, natural disasters, or refugee crises². Traditional data sources might not be able to record these changes. In a fast-changing world, it is important to understand who is using digital technologies in order to be able to generalise the research findings to the entire population. Furthermore, it is important to

¹In the UK, administrative data sources have been affected by the COVID-19 pandemic. The Home Office visa application centres were closed since March 2020, and the NINo allocation process was suspended since March 2020 (ONS, 2020d).

²Francesco Billari's presentation at UNECE: https://www.unece.org/fileadmin/DAM/pau/age/Icpd/ICPD-25/Presentations/Session-1/1_-_1st-thematic-session-Francesco-Billari.pdf

understand the impact of new technologies on people's lives as access to technology and digital skills may continue to become increasingly important in the future, just as education has been an important control variable for many aspects of our lives.

There are important issues to consider on the ethics, privacy and ownership of digital traces data. Digital traces data may give us more insights, but of course "just because we can do something does not mean we should". The United Nation Global Pulse has created a *Level of Risks, Harms and Benefits Assessment* (UN Global Pulse, 2017) to assess the various risks from analysis such as in this thesis. The assessment is divided into two steps with checklists to better understand harms and risks of the research. These considerations should not be seen as an obstacle, however, but as a challenge to create ethical infrastructures and committees to provide access to this kind of data and evaluate research questions. Moreover, companies such as Facebook, which now have data on some of the world's largest human networks, have started to share useful aggregated estimates for research in the initiative *Data for Good* (Facebook Inc., 2020b). For example, the available datasets contain information on COVID-19, natural disasters maps, population density maps, gender equality at home, and more.

Demography can assist in providing safe access to digital traces data, as well as create better data by combining traditional data sources with newly available ones. These steps will produce relevant research that can create positive change. It seems that the use of digital traces data should simply be seen as a challenge to demography and a natural evolution of the discipline in the Digital Revolution.

Appendix A

Supplementary Materials from Chapter 3

Figure A.1 shows the number of Greek migrants across European countries according to Eurostat data from 2018. We compare the Eurostat data with Facebook Advertising Platform data from 2020 estimating the number of Greek migrants, defined as “People that used to live in Greece and now live abroad”, and with the number of people speaking Greek on Facebook. The latter variable seems to better approximate the number of Greek migrants living abroad, in line with Eurostat’s estimates. For the majority of the countries; (except the UK, the Netherlands, France, Germany, Portugal, and Spain), the variable ‘Greek migrants’ from Facebook does not account for any actual Greek migrants.

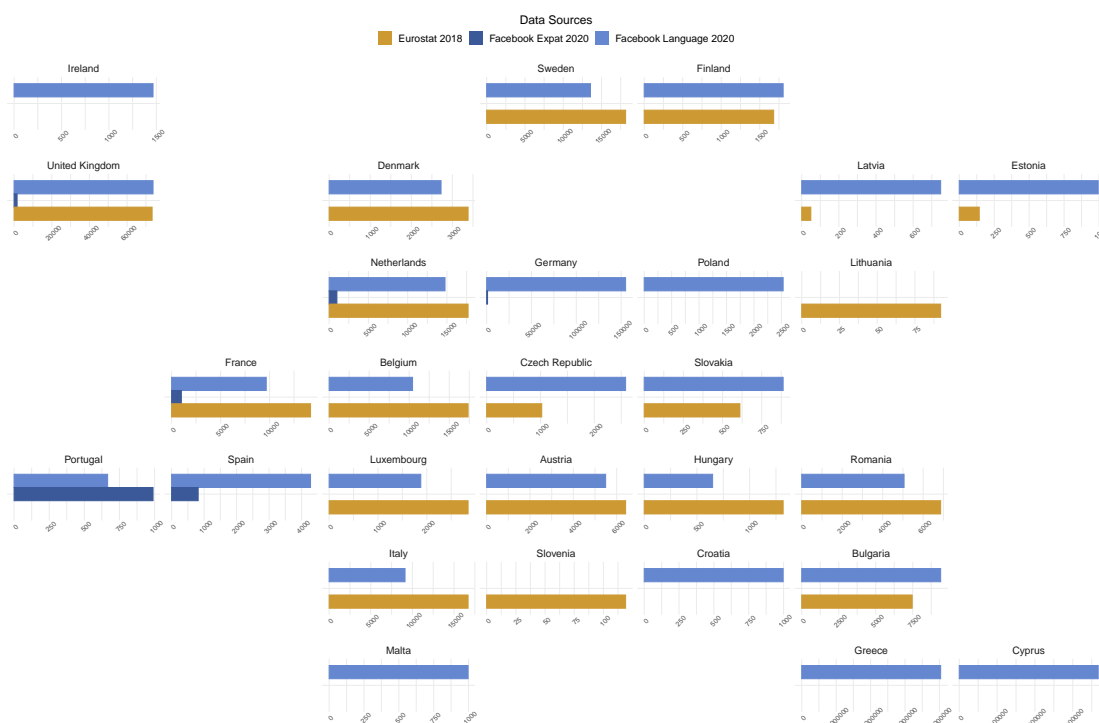


FIGURE A.1: The number of Greek migrants in European countries based on Facebook Advertising Platform data and Eurostat data, and the number of Greek-speaking people on Facebook.

Tables A.1, A.2, A.3, A.4 report the posterior characteristics of the coefficients of the true stock estimates, y , for models 1 and 2 respectively for the two years of analysis. The tables report \hat{R} and \hat{n}_{eff} , which is the effective number of simulation draws (Gelman et al., 2013); it is reported as an additional measure to show the series converge.

TABLE A.1: Posterior characteristics of the coefficients of the true stock estimates, y , in the first model for 2018 with \hat{R} and \hat{n}_{eff} .

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

TABLE A.2: Posterior characteristics of the coefficients of the true stock estimates, y , in the first model for 2019 with \hat{R} and \hat{n}_{eff} .

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

TABLE A.3: Posterior characteristics of the coefficients of the true stock estimates, y , in the second model for 2018 with \hat{R} and \hat{n}_{eff} .

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

TABLE A.4: Posterior characteristics of the coefficients of the true stock estimates, y , in the second model for 2019 with \hat{R} and \hat{n}_{eff} .

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

Sensitivity tests are performed on the models presented in Chapter 3. The R package DHARMA was used, which provides residual diagnostics for hierarchical regression models (Hartig and Lohse, 2021). The package uses a simulation-based approach to produce posterior predictive simulation from JAGS; the idea behind the approach is to simulate replicated data under the model specified and to compare them to the observed data (Gelman and Hill, 2006). The residuals are standardised between 0 and 1; the observed residuals are plotted against the expected. We would expect that a model correctly specified can simulate correctly the data. Analysing the residuals, it is possible to look at systematic discrepancies between real and replicated data (Gelman et al., 2013).

For each simulation are presented two plots: a scatterplot which shows two sets of quantiles against one another, and a residuals plot, which shows the residuals on the vertical axis and the independent variable on the horizontal axis. From the first plot, we should expect an almost straight line from the quantiles, while from the second plot, we would like to see the residuals randomly distributed on the horizontal axis.

Figure A.2 shows the plots produced through simulation and the DHARMA package. The figure on the left shows the Facebook side of the model, while the one on the right the LFS side. Considering the two scatterplots, it seems that the model does a better job at simulating the LFS data rather than the Facebook ones. The residuals plots shows that the residuals are not randomly distributed. This analysis stresses that the model specification could be improved; it seems that one of the problems might be over-dispersion. As a matter of fact, Figure A.3 and Figure A.4 explore the use of a Quasi-Poisson and a Negative Binomial distributions respectively. The two alternative specifications do not seem to improve the scatterplots. However, the Quasi-Poisson seems to improve the residual plot in the case of the LFS (Figure A.3), while the Negative Binomial seems to produce more randomly distributed residuals in the case of the Facebook data (Figure A.4). Therefore, it was attempted to specify a model in which the Facebook data follows a Negative Binomial, while the LFS data a Quasi-Poisson. Figure A.5 shows the results from this attempt, which seems to improve the random distribution of the residuals for the Facebook data, but not significantly improving the LFS side of the model. In terms of the estimates from the models, these are robust across the different specifications of the models, though with larger uncertainty intervals.

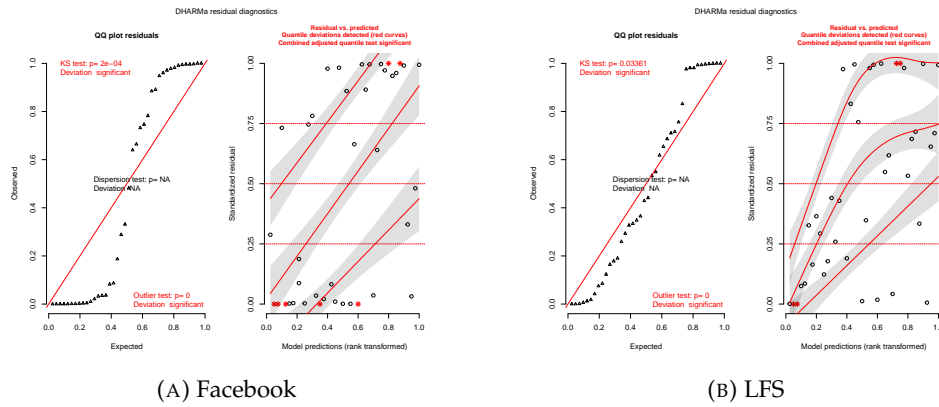


FIGURE A.2: DHARMA of the model presented in Chapter 3

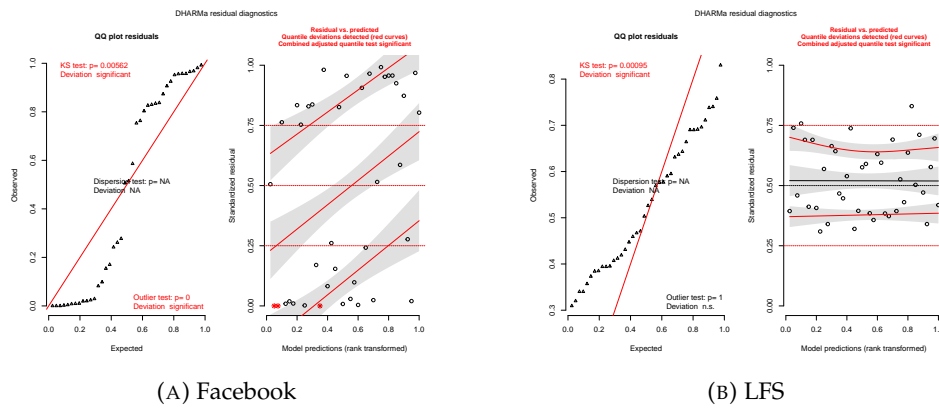


FIGURE A.3: DHARMA of the Quasi-Poisson Model Specification

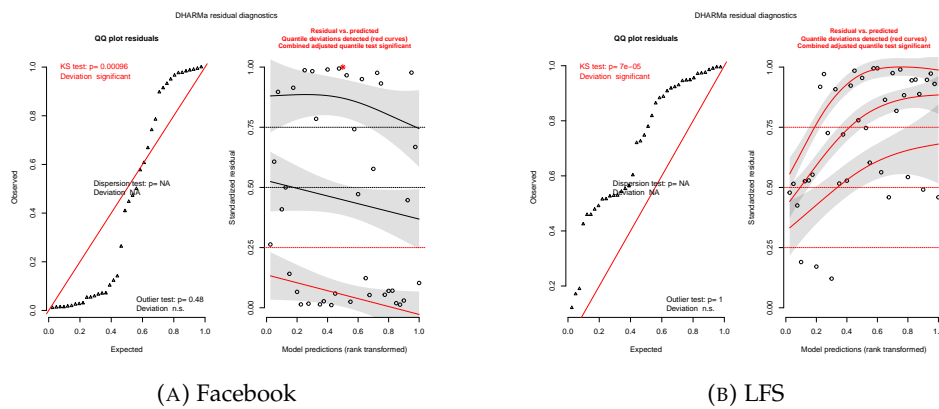


FIGURE A.4: DHARMA of Negative Binomial Model Specification

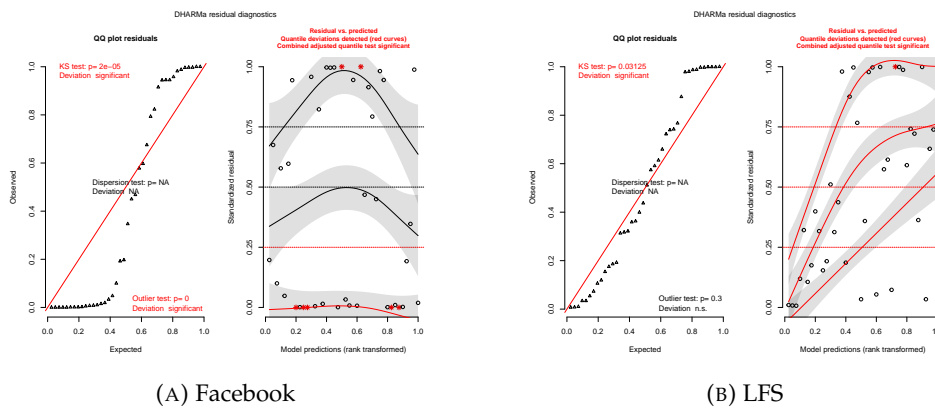


FIGURE A.5: DHARMA of the Negative Binomial Model Specification for the Facebook data and Quasi-Poisson for the LFS data

The simulation are run also for all the models presented in Table 3.4. The figures do not vary much from Figure A.2. However, the model with $\text{Gamma}(1, 1)$ (Figure A.11) shows slightly better results in comparisons to the previous models.

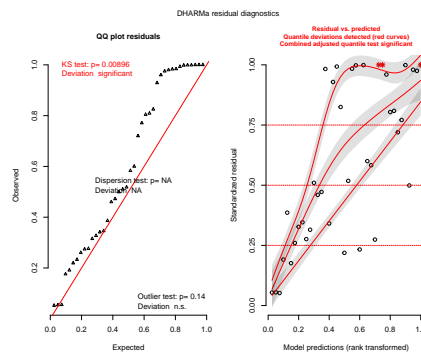


FIGURE A.6: DHARMA of the “Model without Facebook data” in Table 3.4

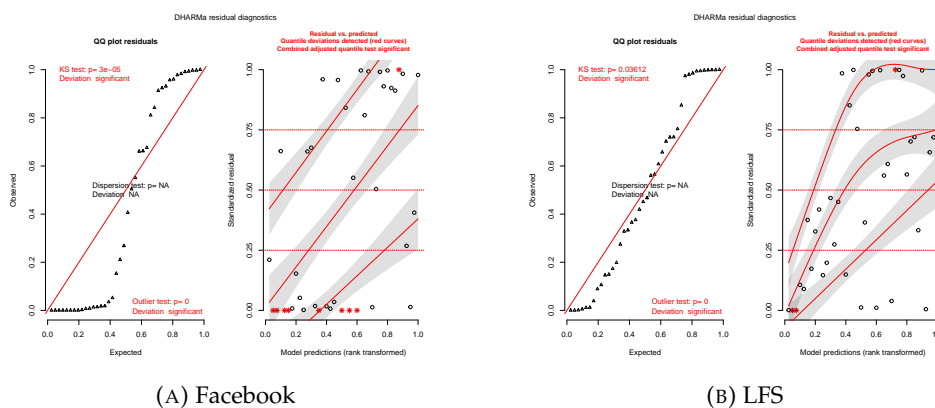


FIGURE A.7: DHARMA of the “Model with Facebook bias at 0%” in Table 3.4

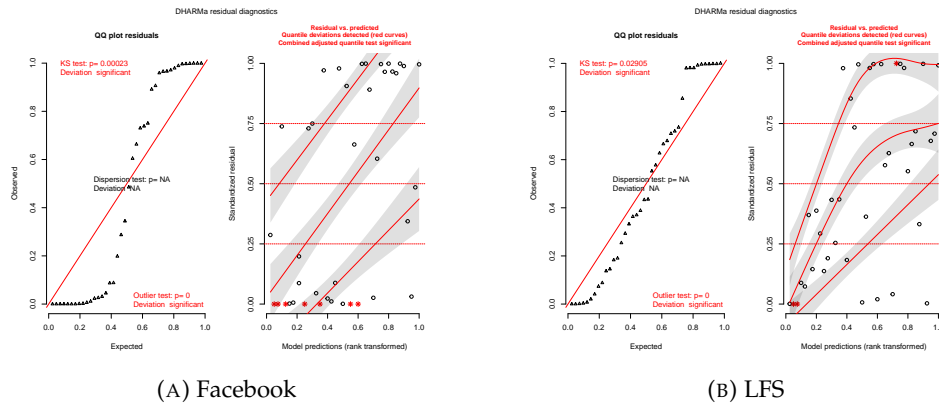


FIGURE A.8: DHARMA of the “Model with Facebook bias at 11%” in Table 3.4

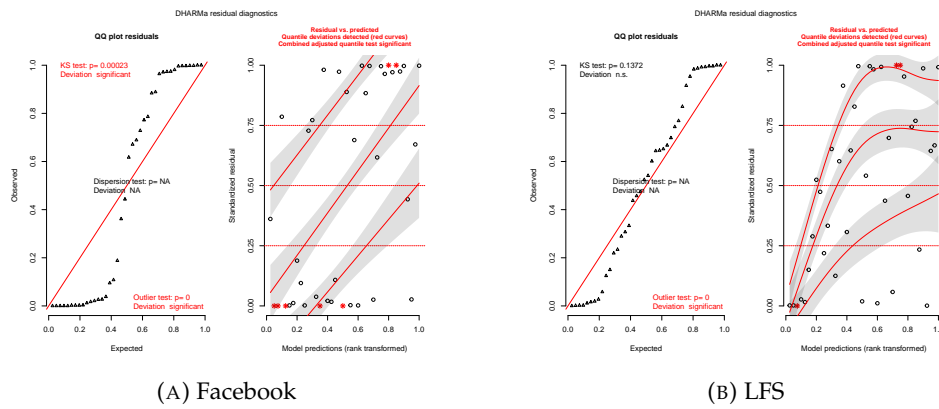


FIGURE A.9: DHARMA of the “Model with LFS bias at 4%” in Table 3.4

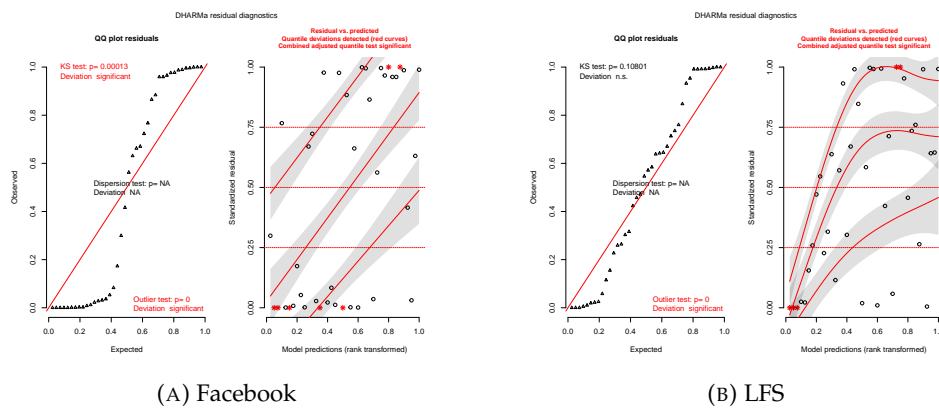


FIGURE A.10: DHARMA of the “Model with LFS bias at 30%” in Table 3.4

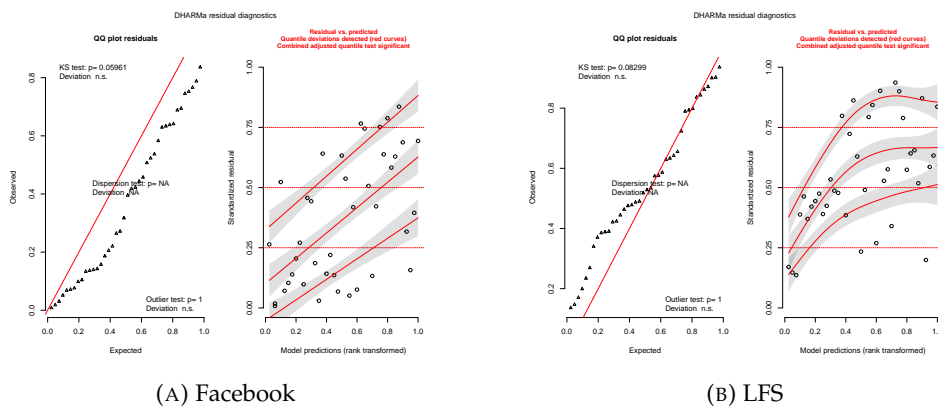


FIGURE A.11: DHARMA of the “Model with $\text{Gamma}(1,1)$ ” in Table 3.4

Appendix B

Supplementary Materials from the Chapter 4

Table B.1, Table B.2, Table B.3, Table B.4, Table B.5, Table B.6, Table B.7, Table B.8, Table B.9, and Table B.10 report the posterior characteristics of the coefficients of the true stock estimates by age and sex for the two years of analysis. The tables report \hat{R} and \hat{n}_{eff} , which is the effective number of simulation draws (Gelman et al., 2013); it is reported as an additional measure to show the series converge.

TABLE B.1: Posterior characteristics of the coefficients of the true stock estimates by age and sex for France by years with \hat{R} and \hat{n}_{eff} .

Year	Age	Male					\hat{R}	\hat{n}_{eff}	Female					\hat{R}	\hat{n}_{eff}
		2.5%	25%	50%	75%	97.5%			2.5%	25%	50%	75%	97.5%		
2018	15-19	4259	5024	5481	5977	7050	1.00	293262	5243	6181	6742	7351	8674	1.00	289677
	20-24	9711	11451	12488	13617	16044	1.00	293165	11786	13893	15150	16516	19490	1.00	289290
	25-29	13493	15904	17346	18913	22296	1.00	293081	14736	17367	18936	20643	24360	1.00	289577
	30-34	12949	15269	16652	18155	21396	1.00	293245	13556	15977	17424	18991	22409	1.00	289568
	35-39	10517	12405	13528	14750	17386	1.00	293306	10253	12086	13181	14367	16958	1.00	289492
	40-44	7821	9224	10059	10969	12933	1.00	293239	10841	12782	13939	15195	17929	1.00	289427
	45-49	5500	6488	7076	7718	9099	1.00	293261	8897	10489	11436	12469	14712	1.00	289443
	50-54	2641	3116	3399	3708	4375	1.00	293244	5598	6598	7196	7846	9263	1.00	289347
	55-59	1993	2352	2566	2800	3305	1.00	293158	2589	3055	3333	3634	4292	1.00	289436
	60-64	1130	1334	1457	1590	1879	1.00	293179	1764	2082	2272	2479	2928	1.00	289764
65+	2695	3179	3469	3783	4463	1.00	293117	6775	7987	8711	9497	11205	1.00	289607	
2019	15-19	5895	6957	7588	8275	9771	1.00	290077	6572	7772	8484	9258	10940	1.00	290808
	20-24	9098	10735	11706	12767	15069	1.00	289857	9874	11680	12748	13909	16422	1.00	287273
	25-29	12338	14555	15872	17304	20421	1.00	289890	14796	17490	19087	20824	24602	1.00	290415
	30-34	11883	14015	15284	16665	19668	1.00	289743	8544	10102	11030	12035	14218	1.00	290590
	35-39	10109	11931	13009	14184	16742	1.00	289906	13724	16226	17711	19323	22824	1.00	290549
	40-44	8845	10438	11382	12412	14647	1.00	289654	9779	11563	12621	13768	16269	1.00	290469
	45-49	9437	11136	12141	13239	15628	1.00	289579	11144	13176	14381	15690	18537	1.00	290355
	50-54	4210	4969	5420	5912	6980	1.00	289809	5749	6800	7424	8101	9572	1.00	290468
	55-59	2522	2980	3252	3548	4192	1.00	289498	3284	3884	4242	4630	5473	1.00	290541
	60-64	1679	1985	2167	2366	2797	1.00	290077	2461	2913	3182	3473	4108	1.00	290103
65+	2521	2980	3251	3548	4191	1.00	289677	6640	7849	8570	9351	11050	1.00	290343	

TABLE B.2: Posterior characteristics of the coefficients of the true stock estimates by age and sex for Germany by years with \hat{R} and \hat{n}_{eff} .

Year	Age	Male					\hat{R}	\hat{n}_{eff}	Female					\hat{R}	\hat{n}_{eff}
		2.5%	25%	50%	75%	97.5%			2.5%	25%	50%	75%	97.5%		
2018	15-19	5790	6836	7459	8137	9609	1.00	285720	4514	5337	5820	6350	7493	1.00	290702
	20-24	10300	12154	13262	14462	17074	1.00	285669	12963	15317	16697	18216	21490	1.00	291052
	25-29	17383	20513	22381	24401	28800	1.00	285387	15330	18116	19754	21541	25407	1.00	290188
	30-34	18526	21862	23857	26009	30698	1.00	285472	16892	19955	21758	23730	27996	1.00	290152
	35-39	14022	16544	18052	19684	23229	1.00	291378	17553	20743	22616	24666	29095	1.00	291016
	40-44	13013	15362	16761	18276	21564	1.00	285138	15259	18030	19656	21439	25301	1.00	290971
	45-49	13734	16207	17682	19280	22761	1.00	285297	15703	18554	20228	22062	26032	1.00	291033
	50-54	12516	14771	16120	17576	20746	1.00	285557	12369	14617	15937	17382	20504	1.00	291137
	55-59	8867	10467	11420	12453	14691	1.00	291491	10885	12865	14026	15299	18055	1.00	291153
	60-64	3717	4387	4789	5224	6170	1.00	284832	8147	9626	10497	11448	13507	1.00	290253
65+	9152	10806	11791	12857	15173	1.00	284760	20223	23896	26048	28416	33531	1.00	291036	
2019	15-19	4463	5281	5760	6287	7436	1.00	288956	4324	5115	5587	6098	7209	1.00	289781
	20-24	8266	9772	10660	11630	13744	1.00	288895	8495	10046	10966	11972	14138	1.00	288950
	25-29	12749	15072	16439	17934	21188	1.00	288523	12045	14246	15549	16969	20044	1.00	289686
	30-34	13852	16374	17864	19484	23016	1.00	288746	15222	17999	19646	21441	25324	1.00	289567
	35-39	11321	13382	14597	15922	18815	1.00	288827	13489	15950	17411	19002	22443	1.00	289064
	40-44	8935	10564	11524	12573	14846	1.00	288619	11826	13987	15269	16664	19683	1.00	289245
	45-49	9677	11445	12486	13621	16097	1.00	288594	11682	13819	15084	16463	19438	1.00	288775
	50-54	12660	14968	16325	17811	21045	1.00	288625	13039	15423	16836	18371	21703	1.00	289021
	55-59	8190	9684	10564	11524	13621	1.00	288450	11539	13648	14897	16259	19208	1.00	290085
	60-64	4463	5280	5763	6288	7433	1.00	288628	7210	8529	9310	10163	12005	1.00	289117
65+	10424	12324	13443	14669	17330	1.00	288691	18811	22249	24283	26500	31303	1.00	289885	

TABLE B.3: Posterior characteristics of the coefficients of the true stock estimates by age and sex for Ireland by years with \hat{R} and \hat{n}_{eff} .

Age	Year	Male					\hat{R}	\hat{n}_{eff}	Female					\hat{R}	\hat{n}_{eff}
		2.5%	25%	50%	75%	97.5%			2.5%	25%	50%	75%	97.5%		
2018	15-19	3596	4251	4639	5065	5983	1.00	291034	5499	6503	7097	7741	9141	1.00	289902
	20-24	7402	8751	9547	10420	12300	1.00	290753	9431	11150	12165	13271	15657	1.00	289385
	25-29	14533	17172	18738	20448	24141	1.00	292213	14291	16895	18433	20105	23727	1.00	289944
	30-34	15919	18812	20525	22397	26434	1.00	290905	15721	18585	20276	22115	26087	1.00	290233
	35-39	15919	18809	20523	22395	26429	1.00	290913	14791	17488	19077	20809	24551	1.00	289822
	40-44	15019	17746	19363	21130	24937	1.00	290758	14790	17487	19077	20808	24555	1.00	290118
	45-49	14191	16767	18289	19963	23567	1.00	290790	11722	13853	15115	16485	19454	1.00	290018
	50-54	13914	16436	17936	19569	23103	1.00	290562	14646	17319	18894	20607	24318	1.00	290082
	55-59	11695	13821	15080	16456	19425	1.00	292387	14649	17318	18894	20607	24322	1.00	289999
	60-64	12180	14394	15705	17139	20232	1.00	292268	12291	14529	15852	17292	20406	1.00	290031
65+	53315	62970	68710	74973	88467	1.00	290700	75043	88715	96768	105537	124505	1.00	290029	
2019	15-19	4940	5845	6385	6973	8262	1.00	270605	3862	4572	4992	5452	6470	1.00	276718
	20-24	5041	5962	6513	7114	8425	1.00	271542	9684	11460	12505	13657	16184	1.00	274917
	25-29	8802	10407	11366	12411	14696	1.00	270180	13732	16238	17725	19352	22932	1.00	276038
	30-34	9395	11108	12132	13243	15680	1.00	270947	12665	14983	16345	17854	21159	1.00	275566
	35-39	17014	20117	21967	23980	28391	1.00	270629	14897	17613	19218	20988	24875	1.00	276075
	40-44	10387	12279	13408	14642	17340	1.00	271072	13729	16242	17722	19355	22931	1.00	276303
	45-49	11277	13332	14559	15893	18822	1.00	270987	16635	19674	21469	23442	27765	1.00	276077
	50-54	14244	16841	18393	20077	23769	1.00	270829	16442	19445	21218	23167	27450	1.00	275204
	55-59	13753	16254	17752	19381	22945	1.00	271097	14021	16581	18096	19762	23410	1.00	275099
	60-64	11179	13214	14430	15756	18649	1.00	270266	14501	17158	18718	20443	24220	1.00	275197
65+	70758	83630	91303	99676	117956	1.00	270954	90623	107187	116946	127676	151240	1.00	275085	

TABLE B.4: Posterior characteristics of the coefficients of the true stock estimates by age and sex for Italy by years with \hat{R} and \hat{n}_{eff} .

Year	Age	Male					\hat{R}	\hat{n}_{eff}	Female					\hat{R}	\hat{n}_{eff}
		2.5%	25%	50%	75%	97.5%			2.5%	25%	50%	75%	97.5%		
2018	15-19	6876	8114	8854	9657	11402	1.00	287952	7607	8977	9789	10679	12600	1.00	285600
	20-24	19138	22580	24638	26869	31709	1.00	287930	15788	18621	20301	22145	26124	1.00	285544
	25-29	22736	26816	29257	31909	37659	1.00	287880	21991	25936	28279	30844	36374	1.00	285405
	30-34	19739	23287	25409	27712	32701	1.00	287877	18040	21284	23204	25311	29849	1.00	285392
	35-39	13158	15525	16942	18475	21813	1.00	287807	12402	14630	15953	17403	20524	1.00	288760
	40-44	9686	11431	12472	13604	16065	1.00	287972	9581	11305	12327	13448	15860	1.00	288514
	45-49	7892	9314	10163	11085	13088	1.00	287759	9412	11106	12109	13209	15582	1.00	285537
	50-54	7655	9031	9854	10749	12691	1.00	288049	6763	7979	8702	9493	11199	1.00	288543
	55-59	4302	5080	5544	6046	7141	1.00	287799	4056	4787	5221	5696	6722	1.00	283832
	60-64	2567	3033	3310	3611	4268	1.00	288086	2250	2658	2900	3165	3737	1.00	289781
65+	8193	9665	10549	11504	13584	1.00	287805	11332	13367	14574	15899	18753	1.00	285492	
2019	15-19	6855	8103	8842	9648	11409	1.00	289047	6849	8087	8819	9614	11349	1.00	290858
	20-24	17231	20352	22202	24223	28618	1.00	288252	13791	16278	17747	19344	22820	1.00	290742
	25-29	23149	27345	29831	32543	38445	1.00	288672	21411	25272	27554	30036	35443	1.00	290571
	30-34	17029	20115	21948	23941	28284	1.00	288555	12867	15192	16563	18058	21304	1.00	290742
	35-39	15677	18515	20197	22040	26043	1.00	288581	9351	11048	12045	13133	15495	1.00	290852
	40-44	14632	17282	18855	20572	24309	1.00	288673	7435	8779	9573	10434	12318	1.00	290757
	45-49	12315	14549	15872	17319	20465	1.00	288481	7767	9172	10004	10905	12874	1.00	290837
	50-54	6693	7907	8629	9416	11127	1.00	288936	8770	10358	11294	12312	14523	1.00	290804
	55-59	6064	7166	7819	8532	10088	1.00	289069	4173	4931	5377	5864	6925	1.00	292003
	60-64	4953	5858	6390	6975	8246	1.00	288842	2502	2958	3226	3520	4157	1.00	290662
65+	12566	14843	16191	17666	20878	1.00	288390	12199	14404	15704	17118	20191	1.00	290762	

TABLE B.5: Posterior characteristics of the coefficients of the true stock estimates by age and sex for Latvia by years with \hat{R} and \hat{n}_{eff} .

Year	Age	Male					\hat{R}	\hat{n}_{eff}	Female					\hat{R}	\hat{n}_{eff}
		2.5%	25%	50%	75%	97.5%			2.5%	25%	50%	75%	97.5%		
2018	15-19	2184	2575	2808	3062	3613	1.00	290731	1724	2035	2219	2419	2855	1.00	288392
	20-24	2648	3120	3402	3709	4377	1.00	290680	3124	3687	4018	4381	5168	1.00	289189
	25-29	7149	8422	9181	10007	11801	1.00	290641	8308	9798	10674	11633	13717	1.00	289258
	30-34	8832	10403	11340	12359	14578	1.00	290709	9338	11010	11996	13070	15414	1.00	288591
	35-39	4963	5845	6371	6946	8194	1.00	290807	5134	6055	6597	7191	8481	1.00	288671
	40-44	3279	3864	4211	4592	5418	1.00	290704	3730	4402	4797	5230	6169	1.00	288355
	45-49	2143	2525	2754	3003	3543	1.00	290979	1818	2145	2338	2550	3011	1.00	288933
	50-54	1427	1683	1836	2002	2364	1.00	290885	1910	2256	2458	2681	3164	1.00	289121
	55-59	1259	1485	1620	1767	2087	1.00	290775	1865	2200	2399	2615	3086	1.00	289599
	60-64	838	989	1080	1178	1393	1.00	290636	930	1100	1199	1309	1546	1.00	288881
65+	417	494	540	590	698	1.00	290441	930	1099	1199	1309	1546	1.00	289200	
2019	15-19	2037	2405	2624	2863	3375	1.00	291111	2448	2888	3147	3430	4042	1.00	293935
	20-24	2717	3208	3498	3817	4501	1.00	291691	3244	3825	4170	4545	5353	1.00	293401
	25-29	6325	7462	8134	8869	10448	1.00	291300	7414	8738	9520	10375	12210	1.00	293243
	30-34	7411	8748	9533	10396	12249	1.00	291313	7841	9243	10069	10975	12916	1.00	293507
	35-39	6692	7895	8606	9386	11056	1.00	291339	6421	7568	8245	8985	10573	1.00	293533
	40-44	3126	3689	4023	4388	5173	1.00	291508	3736	4405	4798	5232	6158	1.00	293499
	45-49	2716	3207	3499	3816	4500	1.00	291231	2263	2671	2911	3173	3740	1.00	293584
	50-54	1356	1603	1749	1909	2253	1.00	291311	2692	3177	3462	3774	4443	1.00	293589
	55-59	1356	1603	1748	1909	2253	1.00	291526	2018	2381	2596	2831	3335	1.00	293478
	60-64	1356	1603	1749	1909	2254	1.00	292576	1221	1443	1573	1716	2024	1.00	293308
65+	676	800	874	956	1131	1.00	291278	1222	1443	1573	1716	2023	1.00	293400	

TABLE B.6: Posterior characteristics of the coefficients of the true stock estimates by age and sex for Lithuania by years with \hat{R} and \hat{n}_{eff} .

Year	Age	Male					\hat{R}	\hat{n}_{eff}	Female					\hat{R}	\hat{n}_{eff}
		2.5%	25%	50%	75%	97.5%			2.5%	25%	50%	75%	97.5%		
2018	15-19	2994	3537	3859	4213	4981	1.00	292133	4207	4981	5437	5936	7011	1.00	289002
	20-24	8204	9688	10567	11534	13629	1.00	292205	11077	13109	14307	15619	18444	1.00	288797
	25-29	13558	16007	17462	19053	22512	1.00	290493	15507	18353	20030	21869	25815	1.00	288822
	30-34	13555	16006	17461	19055	22514	1.00	292240	16397	19400	21175	23115	27302	1.00	288752
	35-39	8915	10530	11487	12535	14812	1.00	292134	11963	14156	15452	16868	19924	1.00	288862
	40-44	4350	5139	5605	6117	7229	1.00	292086	6290	7445	8126	8872	10481	1.00	288699
	45-49	3635	4295	4687	5115	6046	1.00	292103	4607	5452	5952	6498	7678	1.00	288857
	50-54	3101	3664	3997	4363	5157	1.00	291896	2435	2883	3147	3438	4061	1.00	288856
	55-59	1745	2063	2251	2458	2905	1.00	291053	2035	2411	2632	2875	3398	1.00	289129
	60-64	1032	1221	1332	1455	1722	1.00	291002	1326	1572	1717	1875	2218	1.00	289176
65+	532	631	689	753	893	1.00	292475	484	575	629	688	817	1.00	288838	
2019	15-19	3584	4226	4608	5021	5918	1.00	291889	4493	5309	5792	6317	7449	1.00	290859
	20-24	7590	8945	9751	10625	12519	1.00	291779	8512	10050	10965	11954	14099	1.00	291054
	25-29	11737	13827	15073	16423	19348	1.00	291789	13436	15865	17310	18868	22257	1.00	291106
	30-34	13833	16298	17763	19354	22808	1.00	291789	17347	20482	22343	24359	28717	1.00	291123
	35-39	11167	13160	14343	15632	18414	1.00	290808	12472	14730	16066	17518	20660	1.00	291046
	40-44	5727	6749	7358	8020	9449	1.00	291756	6208	7332	7999	8723	10289	1.00	291025
	45-49	4280	5044	5499	5994	7063	1.00	291828	3959	4676	5103	5565	6565	1.00	291144
	50-54	1963	2317	2527	2755	3251	1.00	291874	3639	4298	4689	5114	6035	1.00	291203
	55-59	1732	2044	2229	2431	2868	1.00	292878	3957	4676	5102	5563	6564	1.00	290879
	60-64	574	680	743	811	961	1.00	292594	2298	2717	2965	3233	3817	1.00	291045
65+	574	680	743	812	960	1.00	292456	1067	1263	1379	1505	1780	1.00	290837	

TABLE B.7: Posterior characteristics of the coefficients of the true stock estimates by age and sex for Poland by years with \hat{R} and \hat{n}_{eff} .

Year	Age	Male					\hat{R}	\hat{n}_{eff}	Female					\hat{R}	\hat{n}_{eff}
		2.5%	25%	50%	75%	97.5%			2.5%	25%	50%	75%	97.5%		
2018	15-19	24191	28598	31219	34068	40239	1.00	292844	23848	28235	30855	33703	39928	1.00	284365
	20-24	40320	47661	52024	56788	67065	1.00	288910	42571	50430	55097	60184	71295	1.00	284483
	25-29	59144	69902	76306	83285	98362	1.00	288909	65290	77320	84481	92285	109328	1.00	284431
	30-34	91936	108671	118632	129467	152896	1.00	292754	97077	114973	125622	137231	162520	1.00	284358
	35-39	88720	104860	114467	124916	147476	1.00	292855	95363	112955	123418	134818	159709	1.00	284424
	40-44	46779	55287	60357	65868	77806	1.00	292815	48256	57155	62439	68218	80800	1.00	284424
	45-49	22579	26690	29135	31801	37549	1.00	288876	25542	30251	33056	36115	42779	1.00	284424
	50-54	11290	13345	14569	15901	18781	1.00	288936	15326	18153	19837	21668	25667	1.00	284414
	55-59	9190	10864	11862	12948	15294	1.00	288814	10668	12638	13810	15090	17876	1.00	284366
	60-64	7850	9276	10128	11057	13058	1.00	292930	5673	6722	7346	8026	9511	1.00	284396
65+	7094	8387	9157	9995	11811	1.00	288870	12314	14588	15941	17415	20633	1.00	284349	
2019	15-19	21705	25635	27973	30521	36044	1.00	291465	19467	23006	25101	27368	32290	1.00	289603
	20-24	36988	43676	47658	52001	61417	1.00	293037	33933	40085	43743	47689	56257	1.00	289701
	25-29	41175	48619	53043	57881	68336	1.00	293121	58704	69349	75676	82500	97339	1.00	289754
	30-34	73164	86411	94269	102875	121472	1.00	293074	81599	96422	105205	114697	135308	1.00	289695
	35-39	82818	97797	106718	116435	137464	1.00	292891	90918	107419	117193	127775	150740	1.00	289709
	40-44	56272	66466	72525	79130	93433	1.00	292990	50109	59202	64598	70427	83074	1.00	289659
	45-49	27822	32851	35845	39108	46178	1.00	293024	24327	28754	31377	34209	40348	1.00	289687
	50-54	13740	16236	17715	19330	22837	1.00	292773	17321	20466	22331	24351	28735	1.00	289702
	55-59	10610	12530	13675	14924	17627	1.00	293243	14744	17420	19012	20728	24451	1.00	289813
	60-64	8918	10538	11501	12550	14825	1.00	291498	7311	8641	9431	10285	12138	1.00	289612
65+	10046	11867	12950	14132	16694	1.00	293044	13284	15697	17128	18676	22043	1.00	289823	

TABLE B.8: Posterior characteristics of the coefficients of the true stock estimates by age and sex for Portugal by years with \hat{R} and \hat{n}_{eff} .

Year	Age	Male					\hat{R}	\hat{n}_{eff}	Female					\hat{R}	\hat{n}_{eff}
		2.5%	25%	50%	75%	97.5%			2.5%	25%	50%	75%	97.5%		
2018	15-19	4310	5081	5541	6042	7129	1.00	290346	4209	4969	5416	5903	6963	1.00	296818
	20-24	7361	8675	9460	10315	12172	1.00	290627	8421	9938	10830	11805	13924	1.00	296380
	25-29	11571	13634	14865	16207	19122	1.00	290320	10898	12862	14018	15276	18013	1.00	296865
	30-34	11567	13633	14866	16207	19125	1.00	290415	10404	12278	13380	14582	17194	1.00	296490
	35-39	8782	10350	11284	12305	14519	1.00	290580	8918	10523	11469	12500	14743	1.00	296518
	40-44	9307	10970	11959	13041	15388	1.00	290585	9564	11283	12296	13401	15801	1.00	296494
	45-49	5360	6320	6891	7516	8869	1.00	290362	5795	6840	7455	8125	9583	1.00	296652
	50-54	3468	4090	4460	4863	5740	1.00	290445	2722	3214	3505	3821	4508	1.00	290318
	55-59	2521	2973	3243	3538	4177	1.00	291251	2870	3390	3695	4029	4754	1.00	290610
	60-64	1942	2291	2500	2727	3221	1.00	290771	1234	1460	1593	1737	2053	1.00	290475
65+	2152	2540	2770	3021	3567	1.00	290926	2078	2454	2676	2918	3443	1.00	290610	
2019	15-19	4053	4787	5221	5698	6729	1.00	290067	3547	4192	4572	4987	5882	1.00	290170
	20-24	8451	9976	10877	11871	14010	1.00	290351	6054	7151	7798	8506	10035	1.00	289278
	25-29	8976	10599	11558	12612	14886	1.00	290133	10027	11837	12904	14074	16596	1.00	289356
	30-34	8448	9976	10877	11872	14011	1.00	290336	10653	12579	13714	14954	17637	1.00	290477
	35-39	10210	12052	13141	14339	16920	1.00	290144	9611	11345	12370	13489	15906	1.00	288926
	40-44	9331	11014	12009	13106	15462	1.00	290403	8771	10359	11293	12317	14528	1.00	290232
	45-49	5473	6463	7049	7694	9084	1.00	290442	5929	7003	7635	8329	9823	1.00	290267
	50-54	4187	4946	5395	5889	6955	1.00	290630	5774	6822	7439	8114	9570	1.00	289100
	55-59	3646	4308	4699	5129	6056	1.00	290506	2432	2875	3137	3422	4041	1.00	289990
	60-64	1347	1594	1740	1901	2249	1.00	290779	2781	3286	3585	3912	4617	1.00	290574
65+	1347	1594	1740	1901	2248	1.00	290326	3477	4109	4481	4888	5770	1.00	289101	

TABLE B.9: Posterior characteristics of the coefficients of the true stock estimates by age and sex for Romania by years with \hat{R} and \hat{n}_{eff} .

Year	Age	Male						\hat{R}	\hat{n}_{eff}	Female						\hat{R}	\hat{n}_{eff}
		2.5%	25%	50%	75%	97.5%	2.5%			25%	50%	75%	97.5%				
2018	15-19	12388	14678	16047	17537	20785	1.00	284488	10578	12467	13586	14814	17454	1.00	292986		
	20-24	38479	45576	49830	54453	64535	1.00	284708	33445	39401	42936	46815	55144	1.00	293066		
	25-29	71082	84207	92060	100608	119187	1.00	284792	49943	58853	64132	69923	82377	1.00	293114		
	30-34	62598	74167	81081	88609	104971	1.00	284686	44445	52368	57071	62221	73296	1.00	293140		
	35-39	41735	49437	54053	59071	69982	1.00	284829	33015	38903	42392	46218	54460	1.00	293054		
	40-44	26730	31674	34626	37838	44835	1.00	284955	19888	23437	25545	27852	32804	1.00	293078		
	45-49	16949	20084	21957	23997	28446	1.00	284816	13117	15461	16849	18370	21644	1.00	293117		
	50-54	8405	9964	10895	11909	14113	1.00	284736	5923	6981	7609	8299	9778	1.00	292071		
	55-59	3842	4556	4983	5448	6457	1.00	285751	2579	3041	3316	3616	4263	1.00	293533		
	60-64	1429	1697	1858	2033	2413	1.00	285160	1140	1345	1467	1601	1890	1.00	292461		
65+	2277	2701	2956	3232	3835	1.00	286278	1944	2292	2500	2727	3216	1.00	292867			
2019	15-19	8106	9584	10457	11410	13482	1.00	288999	9521	11236	12254	13359	15754	1.00	292772		
	20-24	29847	35289	38495	41994	49609	1.00	289172	28240	33320	36338	39617	46697	1.00	292858		
	25-29	50126	59253	64643	70513	83308	1.00	288974	48619	57343	62533	68172	80355	1.00	292773		
	30-34	57217	67635	73780	80485	95083	1.00	289068	44008	51920	56620	61729	72756	1.00	292849		
	35-39	37927	44839	48919	53362	63038	1.00	289115	28898	34094	37183	40531	47780	1.00	292879		
	40-44	21137	24993	27268	29748	35147	1.00	289165	16416	19370	21125	23033	27144	1.00	292879		
	45-49	13053	15438	16841	18373	21710	1.00	290680	10964	12940	14112	15386	18136	1.00	293842		
	50-54	7767	9187	10025	10937	12924	1.00	289222	8800	10382	11323	12347	14561	1.00	292803		
	55-59	2669	3159	3448	3765	4452	1.00	291117	2819	3330	3633	3963	4678	1.00	293822		
	60-64	1861	2203	2405	2626	3109	1.00	291011	2163	2555	2788	3042	3590	1.00	293820		
65+	1053	1247	1364	1489	1764	1.00	290914	783	928	1014	1108	1312	1.00	294378			

TABLE B.10: Posterior characteristics of the coefficients of the true stock estimates by age and sex for Spain by years with \hat{R} and \hat{n}_{eff} .

Year	Age	Male					\hat{R}	\hat{n}_{eff}	Female					\hat{R}	\hat{n}_{eff}
		2.5%	25%	50%	75%	97.5%			2.5%	25%	50%	75%	97.5%		
2018	15-19	4549	5364	5852	6382	7520	1.00	292298	4457	5262	5740	6262	7401	1.00	290221
	20-24	10118	11925	13002	14181	16702	1.00	292937	12650	14927	16282	17753	20966	1.00	290178
	25-29	18973	22360	24386	26587	31295	1.00	292624	17315	20424	22280	24293	28705	1.00	290199
	30-34	16442	19383	21131	23042	27133	1.00	292834	15979	18853	20563	22424	26488	1.00	290116
	35-39	11953	14088	15361	16749	19723	1.00	292202	13588	16027	17479	19061	22507	1.00	290186
	40-44	7966	9390	10241	11165	13150	1.00	292182	8455	9975	10882	11869	14018	1.00	290268
	45-49	7585	8943	9753	10635	12524	1.00	292736	8254	9740	10625	11587	13686	1.00	290253
	50-54	3665	4321	4714	5141	6060	1.00	292253	3792	4476	4884	5328	6296	1.00	290372
	55-59	2272	2681	2926	3192	3763	1.00	292403	2993	3533	3855	4206	4974	1.00	290308
	60-64	1260	1488	1625	1774	2095	1.00	292535	1328	1569	1713	1871	2214	1.00	289028
65+	3855	4545	4957	5407	6373	1.00	292559	5324	6284	6854	7476	8831	1.00	290259	
2019	15-19	3072	3628	3960	4322	5106	1.00	291832	4724	5585	6096	6655	7868	1.00	287953
	20-24	8817	10407	11353	12384	14612	1.00	291431	9925	11732	12800	13973	16505	1.00	287441
	25-29	14689	17329	18904	20616	24330	1.00	291316	14380	16985	18532	20228	23891	1.00	287677
	30-34	14882	17548	19140	20872	24636	1.00	291458	15133	17878	19509	21289	25140	1.00	288548
	35-39	12270	14474	15789	17218	20318	1.00	291452	11539	13633	14873	16235	19178	1.00	288951
	40-44	11386	13433	14653	15982	18864	1.00	291458	8983	10614	11583	12642	14939	1.00	287864
	45-49	5637	6654	7260	7921	9347	1.00	291508	6427	7597	8292	9051	10692	1.00	288836
	50-54	6149	7259	7920	8640	10197	1.00	291368	4818	5697	6217	6788	8021	1.00	288561
	55-59	2044	2417	2640	2882	3407	1.00	291406	3776	4467	4876	5325	6295	1.00	288233
	60-64	2045	2417	2640	2882	3409	1.00	291586	1885	2233	2438	2663	3152	1.00	287986
65+	2148	2538	2771	3026	3577	1.00	291744	8698	10278	11217	12243	14466	1.00	288741	

Figure B.1 shows the population pyramids from the multinomial-Dirichlet-Dirichlet with different α values for the first Dirichlet. The values for the α are 0.1, 1, 10, and 100. Changing the values of the α does not affect the harmonisation between Facebook and LFS.

Figure B.2 shows the plots produced through simulation and the DHARMA package for the model presented in Chapter 4. As discussed, the model specified could be improved.

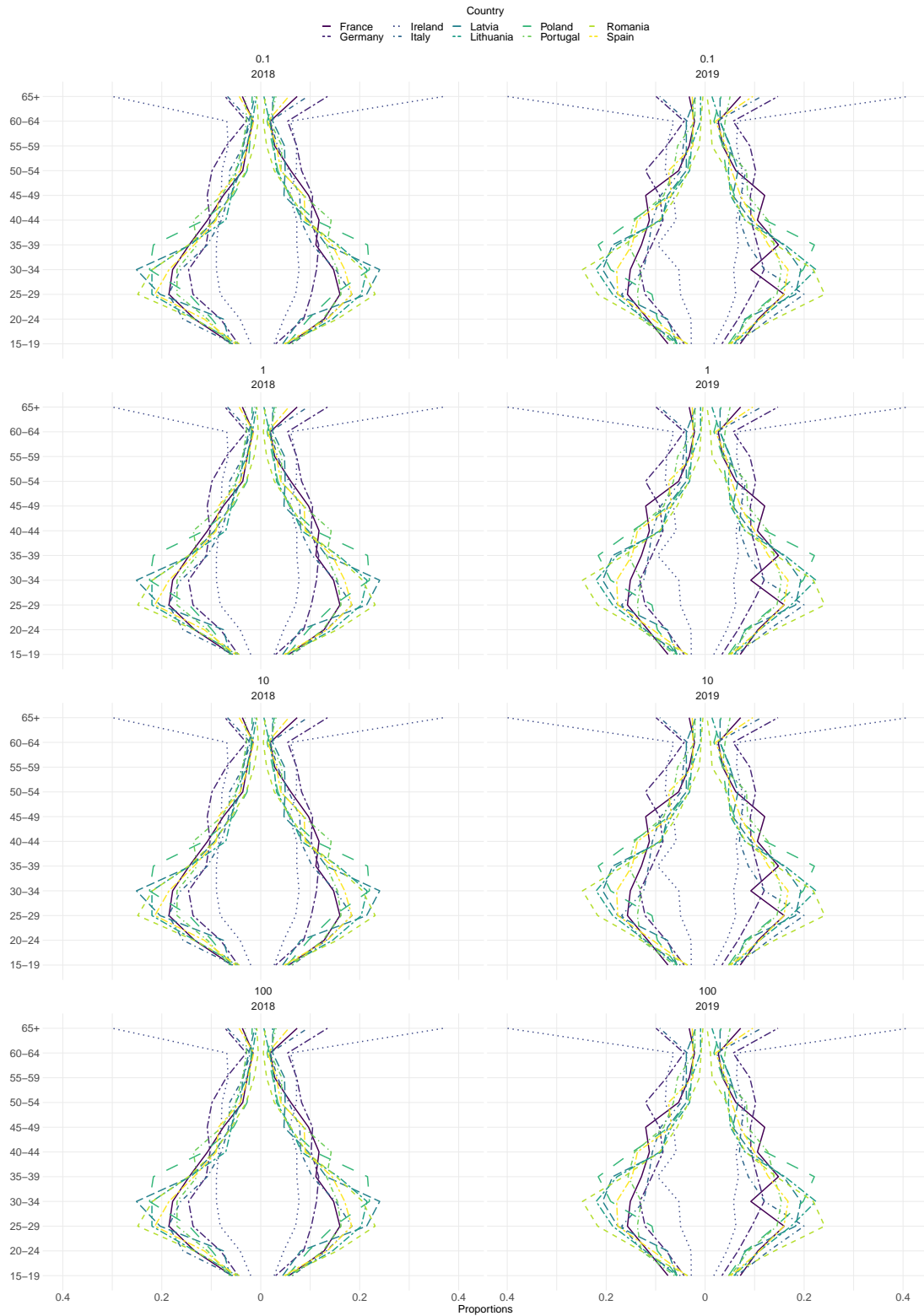


FIGURE B.1: Population pyramids from the multinomial-Dirichlet-Dirichlet estimates harmonising the Rogers-Castro estimates from Facebook and the LFS with different values for the first Dirichlet.

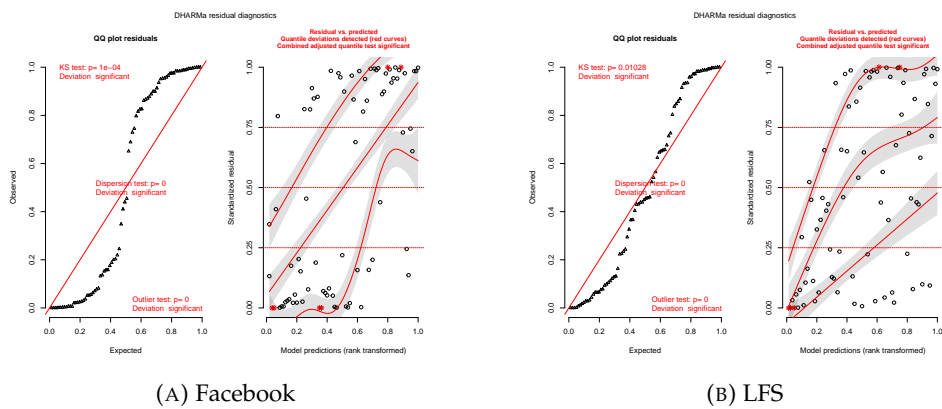


FIGURE B.2: DHARMa of the model presented in Chapter 4

Appendix C

Supplementary Materials from the Chapter 5

Figure C.1 shows the residuals from the model specified without interactions terms:

$$m_{aeit} = c + c_1^T d_a + c_2^T d_e + c_3^T d_i + (b + b_1^T d_a + b_2^T d_e + b_3^T d_i) \times t \quad (\text{C.1})$$

The lighter and darker colours identify combinations in which the residuals are not close to 0. It is evident that the model is not able to account for variability in the estimates of Poland and Romania: especially, in the combination of the countries factors with education and age groups.

Figure C.2 shows the residuals from the model specified in Chapter 5, which includes the interactions terms (see Section 5.4). The interactions seems to have an effect in reducing the residuals, meaning they reduce the difference between the data observed and the expected values.

Table C.1 and Table C.2 report the posterior characteristics of the coefficients b and c respectively, for the model specified in Chapter 5. The tables report \hat{R} and \hat{n}_{eff} , which is the effective number of simulation draws (Gelman et al., 2013); it is reported as an additional measure to show the series converge.

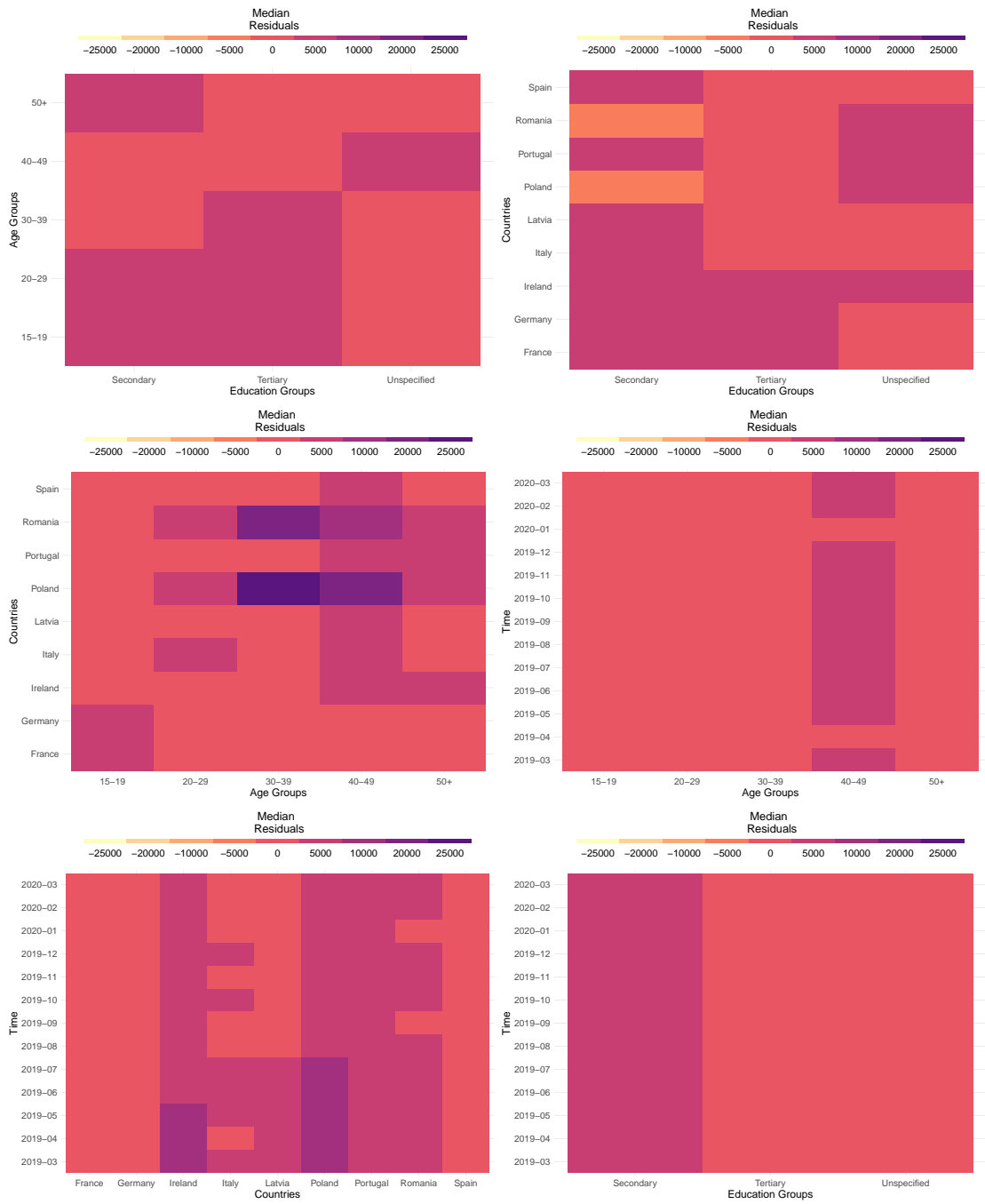


FIGURE C.1: Portrayal of the residuals from the first model without interaction terms.

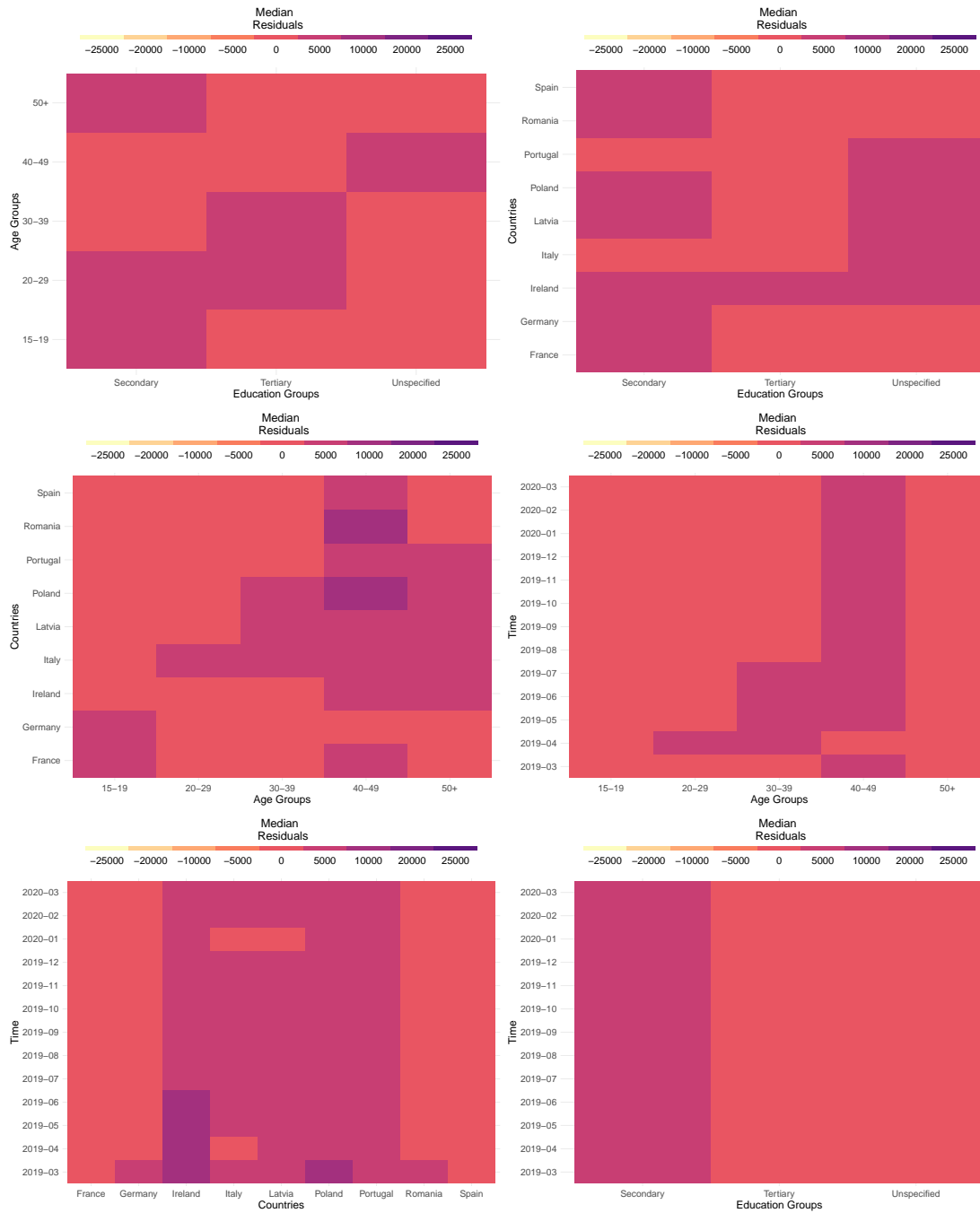


FIGURE C.2: Portrayal of the residuals from the model specified in Chapter 5, which includes interactions terms for Poland and Romania with age groups and education.

TABLE C.1: Posterior characteristics of the coefficients of the intercept of the model, c , with \hat{R} and \hat{n}_{eff} .

	2.5%	25%	50%	75%	97.5%	\hat{R}	\hat{n}_{eff}
c	8.00	8.03	8.04	8.06	8.08	1.01	97
$c_{Edu\ Unspecified}$	-0.69	-0.68	-0.67	-0.66	-0.65	1.02	513
$c_{Edu\ Tertiary}$	0.30	0.32	0.33	0.33	0.35	1.03	452
$c_{Age\ 20-29}$	1.27	1.29	1.31	1.32	1.33	1.02	556
$c_{Age\ 30-39}$	1.16	1.19	1.20	1.21	1.23	1.01	424
$c_{Age\ 40-49}$	0.61	0.63	0.64	0.65	0.67	1.03	483
$c_{Age\ 50+}$	1.23	1.25	1.26	1.26	1.28	1.00	521
c_{France}	-0.38	-0.36	-0.34	-0.33	-0.30	1.00	259
$c_{Germany}$	-0.59	-0.56	-0.55	-0.54	-0.51	1.00	267
$c_{Ireland}$	-0.22	-0.19	-0.18	-0.17	-0.14	1.00	282
c_{Poland}	0.06	0.10	0.12	0.14	0.18	1.01	268
c_{Latvia}	-0.38	-0.36	-0.35	-0.33	-0.31	1.00	288
$c_{Portugal}$	-0.18	-0.15	-0.14	-0.13	-0.10	1.01	216
$c_{Romania}$	0.00	0.04	0.06	0.09	0.13	1.01	234
c_{Spain}	-0.28	-0.26	-0.24	-0.23	-0.20	1.00	249
$c_{Edu\ Unspecified} \times Romania$	0.93	0.97	1.00	1.02	1.06	1.02	589
$c_{Edu\ Tertiary} \times Romania$	-0.16	-0.11	-0.09	-0.07	-0.03	1.02	601
$c_{Edu\ Unspecified} \times Poland$	1.01	1.066	1.08	1.10	1.14	1.00	650
$c_{Edu\ Tertiary} \times Poland$	-0.28	-0.23	-0.21	-0.18	-0.14	1.01	604
$c_{Age\ 20-29} \times Romania$	0.97	1.02	1.05	1.08	1.123	1.01	928
$c_{Age\ 30-39} \times Romania$	0.84	0.89	0.92	0.94	1.00	1.00	943
$c_{Age\ 40-49} \times Romania$	0.72	0.77	0.80	0.82	0.87	1.00	847
$c_{Age\ 20-29} \times Poland$	0.62	0.67	0.69	0.72	0.77	1.01	655
$c_{Age\ 30-39} \times Poland$	1.02	1.08	1.10	1.13	1.18	1.00	758
$c_{Age\ 30-39} \times Poland$	0.70	0.75	0.77	0.80	0.85	1.01	951

TABLE C.2: Posterior characteristics of the coefficients of the slope of the model, b , with \hat{R} and \hat{n}_{eff} .

	2.5%	25%	50%	75%	97.5%	\hat{R}	\hat{n}_{eff}
b	-5.45×10^{-6}	-1.92×10^{-6}	-1.32×10^{-6}	1.82×10^{-6}	5.80×10^{-6}	1.02	107
$b_{Edu\ Unspecified}$	-2.32×10^{-5}	-2.09×10^{-5}	-1.96×10^{-5}	-1.84×10^{-5}	-1.60×10^{-5}	1.02	521
$b_{Edu\ Tertiary}$	-1.56×10^{-5}	-1.31×10^{-5}	-1.19×10^{-5}	-1.07×10^{-5}	-8.23×10^{-6}	1.03	387
$b_{Age\ 20-29}$	-4.11×10^{-5}	-3.84×10^{-5}	-3.68×10^{-5}	-3.53×10^{-5}	-3.23×10^{-5}	1.02	587
$b_{Age\ 30-39}$	-3.15×10^{-5}	-2.87×10^{-5}	-2.72×10^{-5}	-2.57×10^{-5}	-2.25×10^{-5}	1.01	436
$b_{Age\ 40-49}$	-2.73×10^{-5}	-2.46×10^{-5}	-2.31×10^{-5}	-2.15×10^{-5}	-1.83×10^{-5}	1.02	496
$b_{Age\ 50+}$	-6.97×10^{-6}	-4.37×10^{-6}	-3.01×10^{-6}	-1.63×10^{-6}	1.13×10^{-6}	1.00	528
b_{France}	2.85×10^{-6}	4.02×10^{-6}	5.89×10^{-6}	7.92×10^{-6}	1.15×10^{-6}	1.00	268
$b_{Germany}$	1.22×10^{-5}	1.60×10^{-5}	1.78×10^{-5}	1.98×10^{-5}	2.35×10^{-5}	1.01	271
$b_{Ireland}$	-3.58×10^{-6}	3.44×10^{-5}	5.26×10^{-5}	7.17×10^{-6}	1.07×10^{-5}	1.00	291
b_{Poland}	-3.64×10^{-5}	-3.05×10^{-5}	-2.72×10^{-5}	-2.39×10^{-5}	-1.81×10^{-5}	1.01	260
b_{Latvia}	6.01×10^{-6}	9.45×10^{-5}	1.13×10^{-5}	1.33×10^{-5}	1.68×10^{-5}	1.00	289
$b_{Portugal}$	-2.31×10^{-6}	1.43×10^{-5}	3.36×10^{-6}	5.31×10^{-6}	8.81×10^{-6}	1.01	238
$b_{Romania}$	-3.21×10^{-5}	-2.52×10^{-5}	-2.19×10^{-5}	-1.89×10^{-5}	-1.29×10^{-5}	1.01	243
b_{Spain}	-2.19×10^{-6}	1.56×10^{-6}	3.52×10^{-6}	5.35×10^{-6}	9.05×10^{-6}	1.01	283
$b_{Edu\ Unspecified} \times Romania$	2.57×10^{-6}	9.17×10^{-6}	1.28×10^{-5}	1.59×10^{-5}	2.26×10^{-5}	1.02	575
$b_{Edu\ Tertiary} \times Romania$	1.98×10^{-6}	8.11×10^{-6}	1.15×10^{-5}	1.48×10^{-5}	2.15×10^{-5}	1.02	549
$b_{Edu\ Unspecified} \times Poland$	2.61×10^{-6}	9.37×10^{-6}	1.27×10^{-5}	1.62×10^{-5}	2.27×10^{-5}	1.00	624
$b_{Edu\ Tertiary} \times Poland$	7.22×10^{-6}	1.42×10^{-5}	1.75×10^{-5}	2.11×10^{-5}	2.71×10^{-5}	1.00	591
$b_{Age\ 20-29} \times Romania$	-7.58×10^{-6}	-4.91×10^{-6}	3.54×10^{-6}	7.31×10^{-6}	1.45×10^{-5}	1.00	865
$b_{Age\ 30-39} \times Romania$	9.51×10^{-6}	1.74×10^{-5}	2.14×10^{-5}	2.51×10^{-5}	3.24×10^{-5}	1.00	955
$b_{Age\ 40-49} \times Romania$	-1.44×10^{-7}	6.66×10^{-6}	1.07×10^{-5}	1.46×10^{-5}	2.21×10^{-5}	1.00	922
$b_{Age\ 20-29} \times Poland$	-1.07×10^{-5}	-3.46×10^{-6}	3.54×10^{-7}	4.27×10^{-6}	1.20×10^{-5}	1.01	798
$b_{Age\ 30-39} \times Poland$	1.08×10^{-6}	8.60×10^{-6}	1.24×10^{-5}	1.62×10^{-5}	2.36×10^{-5}	1.01	802
$b_{Age\ 40-49} \times Poland$	1.60×10^{-5}	2.32×10^{-5}	2.70×10^{-5}	3.08×10^{-5}	3.77×10^{-5}	1.00	925

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.
- Adair, L. E., Brase, G. L., Akao, K., and Jantsch, M. (2014). #babyfever: Social and media influences on fertility desires. *Personality and Individual Differences*, 71:135–139.
- Alburez-Gutierrez, D., Zagheni, E., Aref, S., Gil-Clavel, S., Grow, A., and Negraia, D. V. (2019). Demography in the Digital Era: New Data Sources for Population Research. Preprint, SocArXiv.
- Alexander, M., Polimis, K., and Zagheni, E. (2019). The Impact of Hurricane Maria on Out-migration from Puerto Rico: Evidence from Facebook Data. *Population and Development Review*, 45(3):617–630.
- Alexander, M., Polimis, K., and Zagheni, E. (2020). Combining Social Media and Survey Data to Nowcast Migrant Stocks in the United States. *Population Research and Policy Review*.
- Alfano, M., Dustmann, C., and Frattini, T. (2016). Immigration and the UK: Reflections after Brexit. SSRN Scholarly Paper ID 2900373, Social Science Research Network, Rochester, NY.
- Alkema, L., Gerland, P., Raftery, A. E., and Wilmoth, J. (2015). The United Nations Probabilistic Population Projections: An Introduction to Demographic Forecasting with Uncertainty. *Foresight (Colchester, Vt.)*, 2015(37):19–24.
- Amaya, A., Biemer, P. P., and Kinyon, D. (2020). Total Error in a Big Data World: Adapting the TSE Framework to Big Data. *Journal of Survey Statistics and Methodology*, 8(1):89–119.

- Arango, J. (2000). Explaining Migration: A Critical View. *International Social Science Journal*, 52(165):283–296.
- Araujo, M., Mejova, Y., Weber, I., and Benevenuto, F. (2017). Using Facebook Ads Audiences for Global Lifestyle Disease Surveillance: Promises and Limitations. In *Proceedings of the 2017 ACM on Web Science Conference, WebSci '17*, pages 253–257, New York, NY, USA. ACM.
- Aref, S., Zagheni, E., and West, J. (2019). The Demography of the Peripatetic Researcher: Evidence on Highly Mobile Scholars from the Web of Science. In Weber, I., Darwish, K. M., Wagner, C., Zagheni, E., Nelson, L., Aref, S., and Flöck, F., editors, *Social Informatics*, Lecture Notes in Computer Science, pages 50–65, Cham. Springer International Publishing.
- Atzori, L., Iera, A., and Morabito, G. (2010). The Internet of Things: A survey. *Computer Networks*, 54(15):2787–2805.
- Azose, J. J. and Raftery, A. E. (2019). Estimation of emigration, return migration, and transit migration between all pairs of countries. *Proceedings of the National Academy of Sciences*, 116(1):116–122.
- Baker, R. (2017). Big Data. In *Total Survey Error in Practice*, chapter 3, pages 47–69. John Wiley & Sons, Ltd.
- Bayer, J. B., Triêu, P., and Ellison, N. B. (2020). Social Media Elements, Ecologies, and Effects. *Annual Review of Psychology*, 71(1):471–497.
- Bayes, T. and Price, n. (1763). LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society of London*, 53:370–418.
- Bellou, A. (2015). The impact of Internet diffusion on marriage rates: Evidence from the broadband market. *Journal of Population Economics*, 28(2):265–297.
- Bernard, A., Bell, M., and Charles-Edwards, E. (2014). Life-Course Transitions and the Age Profile of Internal Migration. *Population and Development Review*, 40(2):213–239.
- Berners-Lee, T. and Fischetti, M. (2001). *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. DIANE Publishing Company.

- Bijak, J. (2010). *Forecasting International Migration in Europe: A Bayesian View*. Springer Science & Business Media.
- Bijak, J. and Bryant, J. (2016). Bayesian demography 250 years after Bayes. *Population Studies*, 70(1):1–19.
- Bijak, J. and Czaika, M. (2020). Assessing Uncertain Migration Futures: A Typology of the Unknown. Technical Report QuantMig Project Deliverable D1.1., Southampton / Krems: University of Southampton and Danube University Krems.
- Bijak, J., Kupiszewska, D., and Kupiszewski, M. (2008). Replacement Migration Revisited: Simulations of the Effects of Selected Population and Labor Market Strategies for the Aging Europe, 2002–2052. *Population Research and Policy Review*, 27(3):321–342.
- Bijak, J., Kupiszewska, D., Kupiszewski, M., Saczuk, K., and Kicing, A. (2007). Population and labour force projections for 27 European countries, 2002–052: Impact of international migration on population ageing: Projections de population et de population active pour 27 pays européens 2002–052: Impact de la migration internationale sur le vieillissement de la population. *European Journal of Population / Revue européenne de Démographie*, 23(1):1–31.
- Bijak, J. and Wiśniowski, A. (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.
- Billari, F., D’Amuri, F., and Marcucci, J. (2016). Forecasting Births Using Google. *CARMA 2016: 1st International Conference on Advanced Research Methods in Analytics*, pages 119–119.
- Billari, F. C., Fent, T., Prskawetz, A., and Scheffran, J. (2006). *Agent-Based Computational Modelling: Applications in Demography, Social, Economic and Environmental Sciences*. Taylor & Francis.
- Billari, F. C., Giuntella, O., and Stella, L. (2019). Does broadband Internet affect fertility? *Population Studies*, 0(0):1–20.
- Billari, F. C., Rotondi, V., and Trinitapoli, J. (2020). Mobile phones, digital inequality, and fertility: Longitudinal evidence from Malawi. *Demographic Research*, 42:1057–1096.

- Billari, F. C. and Zagheni, E. (2017). Big data and population processes: A revolution. *Statistics and Data Science: New Challenges, New Generations*, pages 167–178.
- Bilsborrow, R. E., Hugo, G., Zlotnik, H., and Oberai, A. S. (1997). *International Migration Statistics: Guidelines for Improving Data Collection Systems*. International Labour Organization.
- Blumenstock, J. E. (2012). Inferring patterns of internal migration from mobile phone call records: Evidence from Rwanda. *Information Technology for Development*, 18(2):107–125.
- Boswell, C. and Geddes, A. (2010). *Migration and Mobility in the European Union*. Macmillan International Higher Education.
- boyd, D. M. and Ellison, N. B. (2007). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230.
- Brügger, N. (2015). A brief history of Facebook as a media text: The development of an empty structure. *First Monday*, 20(5).
- Castells, M. (1997). *End of Millennium*. Blackwell Publishers, Inc. Cambridge, MA, USA.
- Caussinus, H. and Courgeau, D. (2010). Estimating Age without Measuring it: A New Method in Paleodemography. *Population (English Edition, 2002)*, 65(1):117–144.
- Cesare, N., Lee, H., McCormick, T., Spiro, E., and Zagheni, E. (2018). Promises and Pitfalls of Using Digital Traces for Demographic Research. *Demography*, 55(5):1979–1999.
- Champion, T. and Falkingham, J. (2016). *Population Change in the United Kingdom*. Rowman & Littlefield.
- Chi, G., State, B., Blumenstock, J. E., and Adamic, L. (2020). Who Ties the World Together? Evidence from a Large Online Social Network. In Cherifi, H., Gaito, S., Mendes, J. F., Moro, E., and Rocha, L. M., editors, *Complex Networks and Their Applications VIII*, volume 882, pages 451–465. Springer International Publishing, Cham.
- Choi, H. and Varian, H. (2012). Predicting the Present with Google Trends. *Economic Record*, 88(s1):2–9.

- Coleman, D. (1983). Some problems of data for the demographic study of immigration and of immigrant and minority populations in Britain. *Ethnic and Racial Studies*, 6(1):103–110.
- Cooksey, B. (2014). An Introduction to APIs. <https://zapier.com/learn/apis/>.
- Couper, M. P. (2013). Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys. *Survey Research Methods*, 7(3):145–156.
- Courgeau, D. (1985). Interaction between spatial mobility, family and career life-cycle: A French survey. *European Sociological Review*, 1(2):139–162.
- Danielsbacka, M., Tanskanen, A. O., and Billari, F. C. (2020). Meeting online and family-related outcomes: Evidence from three German cohorts. *Journal of Family Studies*, 0(0):1–26.
- 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.
- De Mauro, A., Greco, M., and Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. *Library Review*, 65(3):122–135.
- Del Fava, E., Wiśniowski, A., and Zagheni, E. (2019). Modelling International Migration Flows by Integrating Multiple Data Sources. Preprint, SocArXiv.
- Disney, G. (2015). *Model-Based Estimates of UK Immigration*. PhD thesis, University of Southampton.
- Dubois, A., Zagheni, E., Garimella, K., and Weber, I. (2018). Studying Migrant Assimilation Through Facebook Interests. *arXiv:1801.09430 [cs]*.
- DWP (2020). National Insurance numbers allocated to adult overseas nationals to March 2020 - GOV.UK. <https://www.gov.uk/government/statistics/national-insurance-numbers-allocated-to-adult-overseas-nationals-to-march-2020>.
- Earle, S., Marston, H. R., Hadley, R., and Banks, D. (2020). Use of menstruation and fertility app trackers: A scoping review of the evidence. *BMJ Sexual & Reproductive Health*.

- Edelmann, A., Wolff, T., Montagne, D., and Bail, C. A. (2020). Computational social science and sociology. *Annual Review of Sociology*, 46(1):null.
- Ellison, N. B. and boyd, D. M. (2013). Sociality Through Social Network Sites. In Dutton, W. H., editor, *The Oxford Handbook of Internet Studies*, volume 1. Oxford University Press.
- Espenshade, T. J., Bouvier, L. F., and Arthur, W. B. (1982). Immigration and the Stable Population Model. *Demography*, 19(1):125–133.
- European Parliament and Council of the European Union (2007). Regulation (EC) No 862/2007 of the European Parliament and of the Council of 11 July 2007 on Community statistics on migration and international protection and repealing Council Regulation (EEC) No 311/76 on the compilation of statistics on foreign workers (Text with EEA relevance).
- Facebook Inc. (2020a). Company Info. <https://about.fb.com/company-info/>.
- Facebook Inc. (2020b). Facebook Data for Good. <https://dataforgood.fb.com/>.
- Facebook Inc. (2020c). Facebook Reports Third Quarter 2020 Results. <https://investor.fb.com/investor-news/press-release-details/2020/Facebook-Reports-Third-Quarter-2020-Results/default.aspx>.
- Fatehkia, M., Kashyap, R., and Weber, I. (2018). Using Facebook ad data to track the global digital gender gap. *World Development*, 107:189–209.
- Ferrie, C. (2019). *Bayesian Probability for Babies*. Sourcebooks Jabberwocky, Naperville, illustrated edition edition.
- Fiorio, L., Abel, G., Cai, J., Zagheni, E., Weber, I., and Vinué, G. (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 - WebSci '17*, pages 103–110, Troy, New York, USA. ACM Press.
- Fire, M. and Elovici, Y. (2015). Data Mining of Online Genealogy Datasets for Revealing Lifespan Patterns in Human Population. *ACM Trans. Intell. Syst. Technol.*, 6(2):28:1–28:22.

- Freelon, D. (2014). On the Interpretation of Digital Trace Data in Communication and Social Computing Research. *Journal of Broadcasting & Electronic Media*, 58(1):59–75.
- Garcia, D., Kassa, Y. M., Cuevas, A., Cebrian, M., Moro, E., Rahwan, I., and Cuevas, R. (2018). Analyzing gender inequality through large-scale Facebook advertising data. *Proceedings of the National Academy of Sciences*, 115(27):6958–6963.
- Gary, K. and Persily, N. (2020). Unprecedented Facebook URLs Dataset now Available for Academic Research through Social Science One. <https://socialscience.one/blog/unprecedented-facebook-urls-dataset-now-available-research-through-social-science-one>.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition*. CRC Press.
- Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press.
- Gendronneau, C., Wiśniowski, A., Yildiz, D., Zagheni, E., Fiorio, L., Hsiao, Y., Stepanek, M., Weber, I., Abel, G., and Hoorens, S. (2019). *Measuring Labour Mobility and Migration Using Big Data: Exploring the Potential of Social-Media Data for Measuring EU Mobility Flows and Stocks of EU Movers*. Publications Office of the European Union.
- Gerland, P., Raftery, A. E., Sevčiková, H., Li, N., Gu, D., Spoorenberg, T., Alkema, L., Fosdick, B. K., Chunn, J., Lalic, N., Bay, G., Buettner, T., Heilig, G. K., and Wilmoth, J. (2014). World population stabilization unlikely this century. *Science (New York, N.Y.)*, 346(6206):234–237.
- Gil-Clavel, S. and Zagheni, E. (2019). Demographic Differentials in Facebook Usage Around the World. *arXiv:1905.09105 [cs]*.
- Gilroy, C. and Kashyap, R. (2018). Extending the Demography of Sexuality with Digital Trace Data. In *Population Association of America*.
- Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014.

- Gittelman, S., Lange, V., Crawford, C. A. G., Okoro, C. A., Lieb, E., Dhingra, S. S., and Trimarchi, E. (2015). A New Source of Data for Public Health Surveillance: Facebook Likes. *Journal of Medical Internet Research*, 17(4):e98.
- Grolemund, G. and Wickham, H. (2016). *R for Data Science*. O'Reilly.
- Hargittai, E. (2018). Potential Biases in Big Data: Omitted Voices on Social Media. *Social Science Computer Review*, page 089443931878832.
- Hartig, F. and Lohse, L. (2021). DHARMA: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models.
- Healy, K. (2018). *Data Visualization*. Princeton University Press.
- Herdağdelen, A., State, B., Adamic, L., and Mason, W. (2016). The social ties of immigrant communities in the United States. In *Proceedings of the 8th ACM Conference on Web Science, WebSci '16*, pages 78–84, New York, NY, USA. Association for Computing Machinery.
- HESA (2020). Where do HE students come from? — HESA. <https://www.hesa.ac.uk/data-and-analysis/students/where-from>.
- Hitsch, G. J., Hortaçsu, A., and Ariely, D. (2010). Matching and Sorting in Online Dating. *American Economic Review*, 100(1):130–163.
- Hix, S. and Høyland, B. (2011). *The Political System of the European Union*. Macmillan International Higher Education.
- Home Office Government (2020). The UK's Points-Based Immigration System – Further Details.
- Hughes, C., Zagheni, E., Abel, G. J., Wi'sniowski, A., Sorichetta, A., Weber, I., and Tatem, A. J. (2016). *Inferring Migrations, Traditional Methods and New Approaches Based on Mobile Phone, Social Media, and Other Big Data: Feasibility Study on Inferring (Labour) Mobility and Migration in the European Union from Big Data and Social Media Data*. Publications Office, Luxembourg.
- IUSSP (2014). What is demography? <https://iussp.org/en/about/what-is-demography>.

- Kang, J. and Wei, L. (2019). Let me be at my funniest: Instagram users' motivations for using Finsta (a.k.a., fake Instagram). *The Social Science Journal*.
- Kaplanis, J., Gordon, A., Shor, T., Weissbrod, O., Geiger, D., Wahl, M., Gershovits, M., Markus, B., Sheikh, M., Gymrek, M., Bhatia, G., MacArthur, D. G., Price, A. L., and Erlich, Y. (2018). Quantitative analysis of population-scale family trees with millions of relatives. *Science*, 360(6385):171–175.
- Karanasios, S., Thakker, D., Lau, L., Allen, D., Dimitrova, V., and Norman, A. (2013). Making sense of digital traces: An activity theory driven ontological approach. *Journal of the American Society for Information Science and Technology*, 64(12):2452–2467.
- Karemera, D., Oguledo, V. I., and Davis, B. (2000). A gravity model analysis of international migration to North America. *Applied Economics*, 32(13):1745–1755.
- Kierans, D. (2020). Who migrates to the UK and why? <https://migrationobservatory.ox.ac.uk/resources/briefings/who-migrates-to-the-uk-and-why/>.
- Kupiszewska, D., Kupiszewski, M., Martí, M., and Ródenas, C. (2010). Possibilities and limitations of comparative quantitative research on international migration flows. Technical Report Promoting Comparative Quantitative Research in the Field of Migration and Integration in Europe (PROMINSTAT), Project funded by the European Commission, DG Research Sixth Framework Programme, Priority 8.
- Kupiszewska, D. and Nowok, B. (2008). *Comparability of Statistics on International Migration Flows in the European Union*, pages 41–71. John Wiley & Sons.
- Laczko, F. and Rango, M. (2014). "Can Big Data Help Us Achieve a 'Migration Data Revolution'?" <https://publications.iom.int/books/migration-policy-practice-volume-iv-number-2-april-june-2014>.
- Latour, B. (2007). Beware, your imagination leaves digital traces. *Times Higher Literary Supplement*, page 3.
- Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science*, 343(6176):1203–1205.
- Lee, E. S. (1966). A Theory of Migration. *Demography*, 3(1):47–57.

- Łukasik, S., Bijak, J., Krenz-Niedbała, M., Liczbińska, G., Sinika, V., and Piontek, J. (2017). Warriors Die Young: Increased Mortality in Early Adulthood of Scythians from Glinoe, Moldova, Fourth through Second Centuries BC. *Journal of Anthropological Research*, 73(4):584–616.
- Markey, P. M. and Markey, C. N. (2013). Seasonal Variation in Internet Keyword Searches: A Proxy Assessment of Sex Mating Behaviors. *Archives of Sexual Behavior*, 42(4):515–521.
- Martí, M. and Ródenas, C. (2007). Migration Estimation Based on the Labour Force Survey: An EU-15 Perspective. *The International Migration Review*, 41(1):101–126.
- Martín, Y., Cutter, S. L., Li, Z., Emrich, C. T., and Mitchell, J. T. (2020). Using geotagged tweets to track population movements to and from Puerto Rico after Hurricane Maria. *Population and Environment*.
- Massey, D. S., Arango, J., Hugo, G., Kouaouci, A., Pellegrino, A., and Taylor, J. E. (1993). Theories of International Migration: A Review and Appraisal. *Population and Development Review*, 19(3):431–466.
- Massey, D. S., Arango, J., Hugo, G., Kouaouci, A., Pellegrino, A., and Taylor, J. E. (1999). *Worlds in Motion: Understanding International Migration at the End of the Millennium*. OUP Catalogue, Oxford University Press.
- Mayer-Schönberger, V. and Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt.
- Mayer-Schönberger, V. and Ramge, T. (2018). *Reinventing Capitalism in the Age of Big Data*. Basic Books, Inc., New York, NY, USA.
- Mazzoli, M., Diechtiareff, B., Tugores, A., Wives, W., Adler, N., Colet, P., and Ramasco, J. J. (2020). Migrant mobility flows characterized with digital data. *PLOS ONE*, 15(3):e0230264.
- Mencarini, L., Hernández Farías, D. I., Lai, M., Patti, V., Sulis, E., and Vignoli, D. (2019). Happy parents' tweets: An exploration of Italian Twitter data using sentiment analysis. *Demographic Research*, 40(25):693–724.

- Messias, J., Benevenuto, F., Weber, I., and Zagheni, E. (2016). From migration corridors to clusters: The value of Google+ data for migration studies. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 421–428.
- Monti, A., Drefahl, S., Mussino, E., and Härkönen, J. (2019). Over-coverage in population registers leads to bias in demographic estimates. *Population Studies*, 0(0):1–19.
- Ntzoufras, I. (2011). *Bayesian Modeling Using WinBUGS*, volume 698. John Wiley & Sons.
- Ojala, J., Zagheni, E., Billari, F., and Weber, I. (2017). Fertility and Its Meaning: Evidence from Search Behavior. In *Eleventh International AAAI Conference on Web and Social Media*.
- ONS (2014). International passenger survey (IPS) methodology. Text User Guide (Volume 1), Office for National Statistics.
- ONS (2017). Migration Statistics Quarterly Report - Office for National Statistics. <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/bulletins/migrationstatisticsquarterlyreport/november2017>.
- ONS (2018a). Labour Force Survey – user guidance - Office for National Statistics. Technical report, Office for National Statistics.
- ONS (2018b). Migration statistics transformation update - Office for National Statistics. <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/migrationstatisticstransformationupdate/2018-05-24>.
- ONS (2018c). Population of the UK by country of birth and nationality - Office for National Statistics. <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/bulletins/ukpopulationbycountryofbirthandnationality/2016>.
- ONS (2019a). International migration – terms, definitions and frequently asked questions - Office for National Statistics. <https://www.ons.gov.uk/peoplepopulationandcommunity/>

populationandmigration/internationalmigration/methodologies/
longterminternationalmigrationfrequentlyaskedquestionsandbackgroundnotes.

ONS (2019b). Migration Statistics Quarterly Report - Office for National Statistics. <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/bulletins/migrationstatisticsquarterlyreport/august2019>.

ONS (2019c). Population of the UK by country of birth and nationality - Office for National Statistics. <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/bulletins/ukpopulationbycountryofbirthandnationality/july2018tojune2019>.

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

ONS (2019e). Understanding different migration data sources - Office for National Statistics. <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/understandingdifferentmigrationdatasources/juneprogessreport>.

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

ONS (2019g). Update on our population and migration statistics transformation journey: A research engagement report - Office for National Statistics. <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/updateonourpopulationandmigrationstatisticstransformationjourneyaresearchengagementreport/2019-01-30>.

ONS (2020a). International migration and mobility: What's changed since the coronavirus pandemic - Office for National Statistics. <https://www.ons.gov.uk/peoplepopulationandcommunity/>

- populationandmigration/internationalmigration/articles/
internationalmigrationandmobilitywhatschangedsincethecoronaviruspandemic/
2020-11-26.
- ONS (2020b). Long-Term International Migration estimates methodology - Office for National Statistics. <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/methodologies/longterminternationalmigrationestimatesmethodology>.
- ONS (2020c). Migration Statistics Quarterly Report - Office for National Statistics. <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/bulletins/migrationstatisticsquarterlyreport/may2020>.
- ONS (2020d). Population and migration statistics system transformation – overview - Office for National Statistics. <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/transformationofthepopulationandmigrationstatisticssystemoverview/>
2019-06-21.
- Palotti, J., Adler, N., Morales-Guzman, A., Villaveces, J., Sekara, V., Herranz, M. G., Al-Asad, M., and Weber, I. (2020). Monitoring of the Venezuelan exodus through Facebook’s advertising platform. *PLOS ONE*, 15(2):e0229175.
- Pavlík, Z. (2000). *Position of Demography among Other Disciplines*. Department of Demography and Geodemography : Charles University in Prague, Faculty of Science, Prague.
- Perrotta, D., Grow, A., Rampazzo, F., Cimentada, J., Fava, E. D., Gil-Clavel, S., and Zagheni, E. (2020). Behaviors and attitudes in response to the COVID-19 pandemic: Insights from a cross-national Facebook survey. *medRxiv*, page 2020.05.09.20096388.
- Pesando, L. M., Rotondi, V., Stranges, M., Kashyap, R., and Billari, F. C. (2021). The internetization of international migration. *Population and Development Review*, 47(1):79–111.

- Pew Research (2018). Social Media Use 2018: Demographics and Statistics — Pew Research Center. <http://www.pewinternet.org/2018/03/01/social-media-use-in-2018/>.
- Plummer, M., Stukalov, A., Denwood, M., and Plummer, M. M. (2016). Package ‘rjags’. <https://cran.r-project.org/web/packages/rjags/index.html>.
- Potârca, G. and Mills, M. (2015). Racial Preferences in Online Dating across European Countries. *European Sociological Review*, 31(3):326–341.
- Pöttschke, S. and Braun, M. (2017). Migrant Sampling Using Facebook Advertisements: A Case Study of Polish Migrants in Four European Countries. *Social Science Computer Review*, 35(5):633–653.
- Poulain, M. (1993). Confrontation des Statistiques de migrations intra-européennes: Vers plus d’harmonisation? *European Journal of Population / Revue européenne de Démographie*, 9(4):353–381.
- Poulain, M. (1999). Confrontation des statistiques de migration intra-européennes: Vers une matrice complète? - (3/1999/E/N 5). <https://ec.europa.eu/eurostat/de/web/products-statistical-working-papers/-/KS-AP-01-016>.
- Poulain, M., Perrin, N., and Singleton, A. (2006). *THESIM: Towards Harmonised European Statistics on International Migration*. Presses univ. de Louvain.
- Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology*, 25:111–163.
- Raftery, A. E., Li, N., Sevcikova, H., Gerland, P., and Heilig, G. K. (2012). Bayesian probabilistic population projections for all countries. *Proceedings of the National Academy of Sciences*, 109(35):13915–13921.
- Ramos, R. (2016). Gravity models: A tool for migration analysis. *IZA World of Labor*.
- Rampazzo, F., Zagheni, E., Weber, I., Testa, M. R., and Billari, F. (2018). Mater Certa Est, Pater Numquam: What Can Facebook Advertising Data Tell Us about Male Fertility Rates? In *Twelfth International AAAI Conference on Web and Social Media*.
- Ravenstein, E. G. (1885). The Laws of Migration. *Journal of the Statistical Society of London*, 48(2):167–235.

- Raymer, J. (2007). The Estimation of International Migration Flows: A General Technique Focused on the Origin-Destination Association Structure. *Environment and Planning A: Economy and Space*, 39(4):985–995.
- Raymer, J., de Beer, J., and van der Erf, R. (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., Guan, Q., and Ha, J. T. (2019). Overcoming data limitations to obtain migration flows for ASEAN countries. *Asian and Pacific Migration Journal*, 28(4):385–414.
- Raymer, J., Wiśniowski, A., Forster, J. J., Smith, P. W. F., and Bijak, J. (2013). Integrated Modeling of European Migration. *Journal of the American Statistical Association*, 108(503):801–819.
- Reis, B. Y. and Brownstein, J. S. (2010). Measuring the impact of health policies using Internet search patterns: The case of abortion. *BMC Public Health*, 10:514.
- Rendall, M. S., Tomassini, C., and Elliot, D. J. (2003). Estimation of annual international migration from the Labour Force Surveys of the United Kingdom and the continental European Union. *Statistical Journal of the United Nations Economic Commission for Europe*, 20(3,4):219–234.
- Rogers, A. and Castro, L. J. (1981). Model Migration Schedules. <http://pure.iiasa.ac.at/id/eprint/1543/>.
- Rogers, A. and Watkins, J. (1987). General Versus Elderly Interstate Migration and Population Redistribution in the United States. *Research on Aging*, 9(4):483–529.
- Rosenzweig, L., Bergquist, P., Pham, K. H., Rampazzo, F., and Mildemberger, M. (2020). Survey sampling in the Global South using Facebook advertisements. Technical report, SocArXiv.
- Rotondi, V., Kashyap, R., Pesando, L. M., Spinelli, S., and Billari, F. C. (2020). Leveraging mobile phones to attain sustainable development. *Proceedings of the National Academy of Sciences*, 117(24):13413–13420.

- Ruiz-Santacruz, J. S. (2019). Estimación de calendarios migratorios mediante la simulación de los valores iniciales en las optimizaciones de parámetros de los modelos de migración multi-exponenciales : una aplicación a la migración internacional intra-latinoamericana. Technical report, Papers de Demografia, 463: 1-69. Bellaterra: Centre d'Estudis Demogràfics.
- Salganik, M. J. (2017). *Bit by Bit: Social Research in the Digital Age*. Princeton University Press.
- Selwyn, N. (2019). *What Is Digital Sociology?* John Wiley & Sons.
- Singh, L., Wahedi, L., Wang, Y., Wei, Y., Kirov, C., Martin, S., Donato, K., Liu, Y., and Kawintiranon, K. (2019). Blending Noisy Social Media Signals with Traditional Movement Variables to Predict Forced Migration. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19*, pages 1975–1983, Anchorage, AK, USA. Association for Computing Machinery.
- Sloan, L. and Quan-Haase, A. (2017). *The SAGE Handbook of Social Media Research Methods*. SAGE.
- Spyratos, S., Vespe, M., Natale, F., Weber, I., Zagheni, E., and Rango, M. (2019). Quantifying international human mobility patterns using Facebook Network data. *PLOS ONE*, 14(10):e0224134.
- State, B., Rodriguez, M., Helbing, D., and Zagheni, E. (2014). Migration of Professionals to the U.S. In Aiello, L. M. and McFarland, D., editors, *Social Informatics*, volume 8851, pages 531–543. Springer International Publishing, Cham.
- Stewart, I., Flores, R., Riffe, T., Weber, I., and Zagheni, E. (2019). Rock, Rap, or Reggaeton?: Assessing Mexican Immigrants' Cultural Assimilation Using Facebook Data. *arXiv:1902.09453 [cs]*.
- Sutherland, I. (1963). John Graunt: A Tercentenary Tribute. *Journal of the Royal Statistical Society: Series A (General)*, 126(4):537–556.
- UN (1998). *Recommendations on Statistics of International Migration*. Number no. 58, rev. 1 in Statistical Papers. Series M. United Nations, New York.
- UN Global Pulse (2017). Risks, Harms and Benefits Assessment. <https://www.unglobalpulse.org/policy/risk-assessment/>.

- USA SEC (2018). Facebook Inc. 2018 Annual Report 10-K. <https://www.sec.gov/Archives/edgar/data/1326801/000132680119000009/fb-12312018x10k.htm>.
- USA SEC (2019). Facebook Inc. 2019 Annual Report 10-K. <https://sec.report/Document/0001326801-20-000013/fb-12312019x10k.htm>.
- Waldrop, M. M. (2016). The chips are down for Moore's law. *Nature News*, 530(7589):144.
- Walker, N. (2020). Brexit timeline: Events leading to the UK's exit from the European Union. *Commons Briefing papers CBP-7960*.
- Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3):980–991.
- Willekens, F. (1994). Monitoring international migration flows in Europe: Towards a statistical data base combining data from different sources. *European Journal of Population*, 10(1):1–42.
- Willekens, F. (2018). Towards causal forecasting of international migration. *Vienna Yearbook of Population Research*, Vienna Yearbook of Population Research:20.
- Willekens, F. (2019). Evidence-Based Monitoring of International Migration Flows in Europe. *Journal of Official Statistics*, 35(1):231–277.
- Willekens, Frans, G. J. (1983). Twenty-Second European Congress of the Regional Science Association. *The Professional Geographer*, 35(1):99–100.
- Wilson, T. (2010). Model migration schedules incorporating student migration peaks. *Demographic Research*, 23:191–222.
- Wiśniowski, A. (2017). Combining Labour Force Survey data to estimate migration flows: The case of migration from Poland to the UK. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(1):185–202.
- Wiśniowski, A., Bijak, J., Christiansen, S., Forster, J. J., Keilman, N., Raymer, J., and Smith, P. W. (2013). Utilising Expert Opinion to Improve the Measurement of International Migration in Europe. *Journal of Official Statistics*, 29(4):583–607.

- Wiśniowski, A., Forster, J. J., Smith, P. W. F., Bijak, J., and Raymer, J. (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.
- Zagheni, E., Polimis, K., Alexander, M., Weber, I., and Billari, F. C. (2018). Combining Social Media Data and Traditional Surveys to Nowcast Migration Stocks. In *Annual Meeting of the Population Association of America*.
- Zagheni, E. and Weber, I. (2012). You Are Where You e-Mail: Using e-Mail Data to Estimate International Migration Rates. In *Proceedings of the 4th Annual ACM Web Science Conference, WebSci '12*, pages 348–351, New York, NY, USA. ACM.
- Zagheni, E. and Weber, I. (2015). Demographic research with non-representative internet data. *International Journal of Manpower*, 36(1):13–25.
- Zagheni, E., Weber, I., and Gummadi, K. (2017). Leveraging Facebook’s Advertising Platform to Monitor Stocks of Migrants. *Population and Development Review*, 43(4):721–734.
- Zlotnik, H. (1987). The Concept of International Migration as Reflected in Data Collection Systems. *The International Migration Review*, 21(4):925–946.
- Zwingerman, R., Chaikof, M., and Jones, C. (2020). A Critical Appraisal of Fertility and Menstrual Tracking Apps for the iPhone. *Journal of Obstetrics and Gynaecology Canada*, 42(5):583–590.