Journal of the Royal Statistical Society Series A: Statistics in Society, 2026, 00, 1–25
https://doi.org/10.1093/jrsssa/qnag090
Published: 29 July 2026
Original article

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society **Series A** Statistics in Society
OXFORD

Migrant age profiles reconciling digital trace and survey data: an example of the United Kingdom in 2018 and 2019

Francesco Rampazzo[1], Jakub Bijak[2], Agnese Vitali[3], Ingmar Weber[4], Emilio Zagheni[5]

[1]Department of Social Statistics, Cathie Marsh Institute, University of Manchester, Manchester, UK
[2]Nuffield Department of Population Health, Reuben College, Leverhulme Centre for Demographic Science, University of Oxford, Oxford, UK
[3]Department of Sociology and Social Research, University of Trento, Trento, Italy
[4]Saarland Informatics Campus, Saarland University, Saarbrücken, Germany
[5]Department of Digital and Computational Demography, Max Planck Institute for Demographic Research, Rostock, Germany
Address for correspondence: Francesco Rampazzo, Department of Social Statistics, Cathie Marsh Institute, University of Manchester, Humanities Bridgeford Street, Manchester M15 6AD, UK. Email: francesco.rampazzo@manchester.ac.uk

Abstract

Accurate estimates of migrant population stocks by age and sex are essential for demographic analysis and policy. This study develops a hierarchical Bayesian model that integrates survey data from the Labour Force Survey with digital trace data from the Facebook Advertising Platform to improve estimation of migrant age–sex distributions. The approach applies reduced Rogers–Castro age schedules as first-stage smoothing and harmonizes profiles within a multinomial-Dirichlet–Dirichlet framework. We illustrate the model using data on the 10 largest European migrant groups in the United Kingdom in 2018–2019. Results reveal 3 broad patterns: younger age structures among Western and Southern European migrants, slightly older working-age profiles for Central and Eastern Europeans, and a predominantly older Irish migrant population. External validation against Census 2021 country-of-birth data for England and Wales shows that the proposed model more closely matches benchmark age–sex distributions than a specification inspired by the area model of Wiśniowski et al. (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. https://doi.org/10.1111/rssa.12177). Facebook data improve representation of younger migrants, who are often underrepresented in surveys, while the Labour Force Survey provides broader demographic structure. Overall, the findings demonstrate the value of combining traditional and digital data to address gaps in migration statistics and produce timely, detailed population estimates.

Keywords bayesian modelling, data integration, international migration statistics, migrant stocks

Received: November 26, 2025. **Revised:** July 13, 2026. **Accepted:** July 14, 2026

1 Introduction

Migration is a crucial demographic process that shapes not only population size but also the age and sex structures of countries and regions worldwide. Estimates of migrant populations (migrant stocks) disaggregated by age and sex are essential for making accurate population projections and for understanding migrants' characteristics and motives for migration (Rogers & Castro, 1981; Rogers & Watkins, 1987). Additionally, challenges such as undercounting in migration statistics, highlighted more recently by Dańko et al. (2024) and systematically addressed years earlier by Poulain et al. (2006), underscore the need to improve data quality to support accurate demographic analysis. At the same time, there are visible regularities in the migration patterns by age, which can help with estimation. Migration movements, particularly those driven by job opportunities, are predominantly concentrated in young adulthood, especially during the 20s and early 30s (Rogers & Castro, 1981; Wilson, 2010).

However, accurately measuring the number of migrants by age and sex presents multiple challenges. Traditional data sources often lack timeliness and sufficient coverage of all relevant population groups. Survey-based data, such as those from the Labour Force Surveys (LFS), may introduce greater uncertainty when disaggregated by age and sex due to small sample sizes (Bernard et al., 2014; Wiśniowski et al., 2016). Moreover, surveys like the LFS are designed to measure stable populations and may not adequately represent migrants (Office for National Statistics, 2018a, 2019c). This limitation persists despite ongoing methodological improvements.

Digital trace data offer a complementary set of sources for migration measurement. Rather than forming a single homogeneous stream, they constitute a heterogeneous ecosystem in which different data streams capture different migration-related constructs, including identity and stated affiliation, day-to-day movement, professional trajectories, spatial presence, migration-related information seeking, and platform-internal mobility flows. Social media advertising audiences, particularly from the Facebook Advertising Platform, have been used to measure and estimate migrant stocks disaggregated by age and sex in the United States (Zagheni et al., 2017; Alexander et al., 2022; Sanlitürk et al., 2024), the United Kingdom (Rampazzo et al., 2021, 2025), Venezuela-related displacement (Palotti et al., 2020), and internal displacement during humanitarian crises (Leasure et al., 2023). Exploratory work on Africa has also shown that Meta/Facebook migrant-related advertising variables were available beyond high-income settings, although their interpretation depends on context-specific patterns of platform use (Rampazzo & Weber, 2020). More recent collaboration with Meta has also enabled estimates of monthly country-to-country migration flows from privacy-protected Facebook records across 181 countries (Chi et al., 2025).

Other digital traces capture different aspects of mobility and migration. Mobile-phone records have been used to study internal mobility in low- and middle-income contexts (Blumenstock, 2012), while georeferenced traces such as Twitter, check-in, and mobile-phone data raise important definitional questions about residence and observation windows (Fiorio et al., 2021). Internet-search activity provides a further window onto migration intentions and information-seeking behaviour, with Google Trends data used to now-cast Romanian migration flows to the UK (Avramescu & Wiśniowski, 2021). Bibliometric and online-platform data, including LinkedIn-derived career histories, have also been used to study high-skilled and gendered professional mobility (Aref et al., 2019; Jacobs et al., 2025; Sanlitürk et al., 2024; State et al., 2014). The reach and accuracy of these sources depend on patterns of digital access and use (Hargittai, 2018).

The present paper focuses on Facebook advertising-audience data because they have historically provided aggregate, demographically disaggregated audience estimates across countries, at low cost, and through repeatable programmatic access via the platform's marketing Application Programming Interface (API) and tools such as *pySocialWatcher* (Araujo et al., 2017). However, these measures remain vulnerable to platform coverage bias, changing migrant classifications, licensing restrictions, and discontinuities associated with algorithmic or interface changes (Palotti et al., 2020; Rampazzo et al., 2025). Building on Rampazzo et al. (2021), this study combines Facebook advertising-audience data with LFS data in a Bayesian hierarchical model to estimate the age and sex distribution of European migrant populations in the UK. By harmonizing estimates from two imperfect data sources, the model provides a framework for addressing inconsistencies in data quality, coverage, and availability.

The proposed approach employs an extension of the Integrated Model of European Migration (IMEM) by age and sex (Raymer et al., 2013). A simplified Rogers–Castro model (Rogers & Castro, 1981; Rogers & Watkins, 1987) is used as a parsimonious prior smoother for age profiles derived from the LFS and Facebook data. Since the Rogers–Castro schedule was developed for migration flows rather than migrant stocks, we do not interpret it here as a substantive generative model of stock formation. Instead, it provides a structured first-stage prior representation of noisy age shares; wherever source data are available, this prior shape is modified by the multinomial-Dirichlet–Dirichlet (MDD) hierarchy. To illustrate the model's utility, it is applied to estimate the age and sex distribution of the 10 largest European migrant groups in the UK for 2018 and 2019. This analysis focuses on these groups and years to ensure the model's reliability, with coverage reduced from 20 to 10 countries compared to Rampazzo et al. (2021).

This study addresses critical challenges in migration data by reconciling discrepancies between sending and receiving countries, incorporating missing data, and improving age and sex disaggregation, building on the area model of Wiśniowski et al. (2016).

This paper makes three contributions. First, it provides a Bayesian framework for integrating survey and digital trace data to estimate migrant age–sex composition. Second, it introduces a MDD harmonization approach for combining source-specific age profiles. Third, it applies and externally validates the framework using UK migrant-stock data and Census 2021 country-of-birth profiles.

2 Background

In 2016, the Office for National Statistics (ONS) reported that there were 9 million non-British-born residents in the UK, 38% of whom were born in a European country (Office for National Statistics, 2018b). These figures differ slightly from the 2021 population census, which showed that one in six usual residents of England and Wales were born outside the UK, an increase of 2.5 million since 2011, from 7.5 million (13.4%) to 10 million (16.8%) in 2021 (Office for National Statistics, 2022). Work and study were identified as the main drivers of European migration to the UK during the 2010s (Kierans, 2020). Similarly, estimates from the Department for Work and Pensions, based on National Insurance Number registrations, indicate that most EU migrants engaging with the UK labour market are of working age, a pattern expected given that National Insurance Number registration is primarily required for employment purposes (Office for National Statistics, 2020). For example, in the academic year 2018/2019, the Higher Education Statistics Agency reported that 30% of international students in the UK were EU citizens, with the largest groups being Italian, French, and German nationals (HESA, 2020).

In August 2019, the ONS reclassified its long-term migration statistics as experimental, reflecting increased uncertainty in the figures (Office for National Statistics, 2019a). These statistics were previously based on the International Passenger Survey and the LFS. Since then, the ONS has been transforming its migration statistics by integrating multiple administrative data sources (Office for National Statistics, 2019a). The resulting methodology, known as the Registration and Population Interaction Database, draws on data from the Department for Work and Pensions and HM Revenue and Customs (Office for National Statistics, 2021). However, this approach remains provisional, and the resulting estimates are currently classified as official statistics in development (Office for National Statistics, 2021, 2025).

The LFS is another key source of migration data in many countries across Europe. In the UK, the LFS sampling framework targets private households using the Postcode Address File, an official database of addresses. However, this sampling framework may underrepresent young migrant adults, who often experience unstable jobs and housing conditions. The ONS acknowledges that the LFS is not designed to study long-term migrants and does not provide complete coverage of this group (Office for National Statistics, 2018a, 2019b). Furthermore, the UK LFS only interviews individuals who have lived at the same address for at least six months, excluding many short-term migrants and those in communal establishments such as student halls of residence, which introduces additional biases (Office for National Statistics, 2018a, 2019b).

In addition, traditional data sources like the LFS lack timeliness and comprehensive coverage of migrants, especially in terms of age and sex. Digital trace data, despite its limitations (e.g. Rampazzo et al., 2021), may better capture younger migrants not living in standard households, such as university students. For that reason, in this article, we set out to combine age patterns from Facebook data and the LFS, in order to enhance the coverage of traditional data sources for these groups. This approach aims to reduce the uncertainty in migrant estimates disaggregated by age and sex by incorporating data from multiple administrative sources and surveys (Disney, 2015), with digital trace data providing further improvements.

As hinted before, studies of migration patterns by age have often focused on flows rather than stocks. For instance, Rogers and Castro (1981) described age-specific regularities in migration flow propensities using mathematical models, while Courgeau (1985) linked these patterns to the life course. Migration tendencies vary with life events such as marriage, divorce, or children leaving the parental home (Courgeau, 1985).

In their seminal work, Rogers and Castro (1981) proposed a multi-exponential model to capture the relationships between age and migration. Based on empirical data from developed countries, the Rogers–Castro model describes a pattern of migration rates, $M(x)$, across all age groups x. This pattern includes a high rate of migration in early childhood (children migrating with parents), a sharp drop during teenage years, a peak in the early 20s (associated with education or early career moves), and a gradual decline throughout the economically active years. Additional smaller peaks are observed postretirement and at the end of life, as individuals move closer to family or into care institutions.

The full Rogers–Castro model is represented by the following equation:

$$\begin{aligned} M(x) = \;& a_1 \times \exp(\alpha_1 x) + && \text{(Pre-labour force curve)},\\ & a_2 \times \exp(-\alpha_2(x-\mu_2)) - \exp[-\lambda_2(x-\mu_2)] + && \text{(Labour force peak)},\\ & a_3 \times \exp(-\alpha_3(x-\mu_3)) - \exp[-\lambda_3(x-\mu_3)] + && \text{(Retirement peak)},\\ & a_4 \times \exp(\lambda_4 x) + && \text{(Post-retirement increase)},\\ & c_1 && \text{(Constant)}, \end{aligned}$$

where

- a_1, α_1: prelabour force trend
- $a_2, \alpha_2, \mu_2, \lambda_2$: labour force peak
- $a_3, \alpha_3, \mu_3, \lambda_3$: retirement peak
- a_4, λ_4: postretirement increase
- c_1: constant term

The model (1) comprises four additive components and a constant term, which improves the fit to observed data. These components correspond to different stages in the life course: the prelabour force period, the labour force peak, retirement, and postretirement phases. The model's flexibility allows for adaptation to varying migration patterns, depending on data availability and the specific context (Rogers & Castro, 1981; Rogers & Watkins, 1987). In addition, Wilson (2010) proposed an extension to the model to account for student migration, introducing a curve that reflects the sharp increase in migration propensity during late teenage years. This spike, relevant especially for internal migration, coincides with transitions from secondary school to university, particularly in countries like the UK with high student mobility. The additional student curve is crucial when analysing disaggregated data, as aggregated datasets may suppress these specific patterns.

Previous attempts to model migration by age and sex relied on multiple imperfect data sources. For instance, the area model of Wiśniowski et al. (2016) extended the IMEM model to estimate 'true flows' of migration between European countries by origin, destination, age, and sex. They reconciled discrepancies between data sources using a measurement error model and harmonized age profiles through a multiplicative framework. However, limitations in estimating age and sex flows were noted. These included inconsistencies between sending and receiving country reports, divergent national definitions of migration (e.g. 'permanent' versus 'no time limit'), and missing or incomplete disaggregation by age and sex in many cases. The approach also relied heavily on expert judgment to fill data gaps and adjust for undercounting, which, while necessary, introduced subjectivity.

In this article, a similar logic and structure is applied, with key innovations to address prior limitations in estimating age- and sex-disaggregated migration data. First, the 'true stock' of migrants by sex is estimated, following the approach outlined in Rampazzo et al. (2021). Then, a MDD model is used to redistribute this stock across age groups for male and female migrants, using harmonized age profiles derived from both survey and digital trace data. Unlike the area model of Wiśniowski et al. (2016), which often faced gaps in age–sex breakdowns and inconsistencies between data sources, this method leverages Facebook data alongside the LFS to enhance coverage, particularly for younger migrants and those underrepresented in traditional surveys. The integration of traditional and digital sources, combined with a flexible Bayesian framework, enables more reliable, transparent, and detailed estimates of migrant-stock distributions by age and sex.

3 Data

In this work, as in Rampazzo et al. (2021), we use data from the LFS and the Facebook Advertising Platform. The LFS disaggregates its estimates into five-year age groups from 0–4 to 80–84 years old, with an open-ended group for those aged 85+. The LFS data are not provided when there were between zero and three interviews/contacts in that specific age, sex, and country of origin group.[1]

Facebook requires users to be at least 13 years old, so the Facebook Marketing API provides data for ages 13–65+. While data can be collected for individual ages, the platform applies a minimum-audience threshold (commonly described as around 1,000 monthly active users, though the exact threshold is not always specified) below which audience size is reported as zero rather than the underlying count. According to Facebook for developers, monthly active users are defined as the 'estimated number of people that have been active on your selected platforms and satisfy your targeting spec in the past month'. An additional limitation of the Facebook Advertising Platform is the lack of further disaggregation within the 65+ age group. Moreover, Facebook has acknowledged that 'a disproportionate number of our younger users register with an inaccurate age' [USA SEC (U.S. Securities and Exchange Commission) 2018, 2019]. Nevertheless, the platform achieves high overall accuracy in classifying user demographics: sex classification matches user-reported data in 98%–99% of cases, and age classification shows similar reliability across major age groups. As such, when appropriately validated, Facebook data can serve as a dependable tool for demographic research (Grow et al., 2022).

Therefore, even if it is possible to obtain single-year age groups, it is advisable to aggregate the data into five-year age groups to reduce errors in the Facebook data. Consequently, this work 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+. The Facebook data were downloaded through *pySocialWatcher* (Araujo et al., 2017).

4 Methodology

The model architecture consists of two components, as illustrated in Figure 1. The first-component estimates total migrant stocks by sex; the second component redistributes those sex-specific totals across age groups.

4.1 Brief overview of the Rampazzo et al. (2021) framework

The first component reproduces the framework introduced in Rampazzo et al. (2021). That framework treats observed migrant counts from the LFS and from the Facebook Advertising Platform as

[1] The special values in the data file represent the following: . for 'no contact', c for 'not available due to disclosure control', and 0 for 'rounded to zero'.

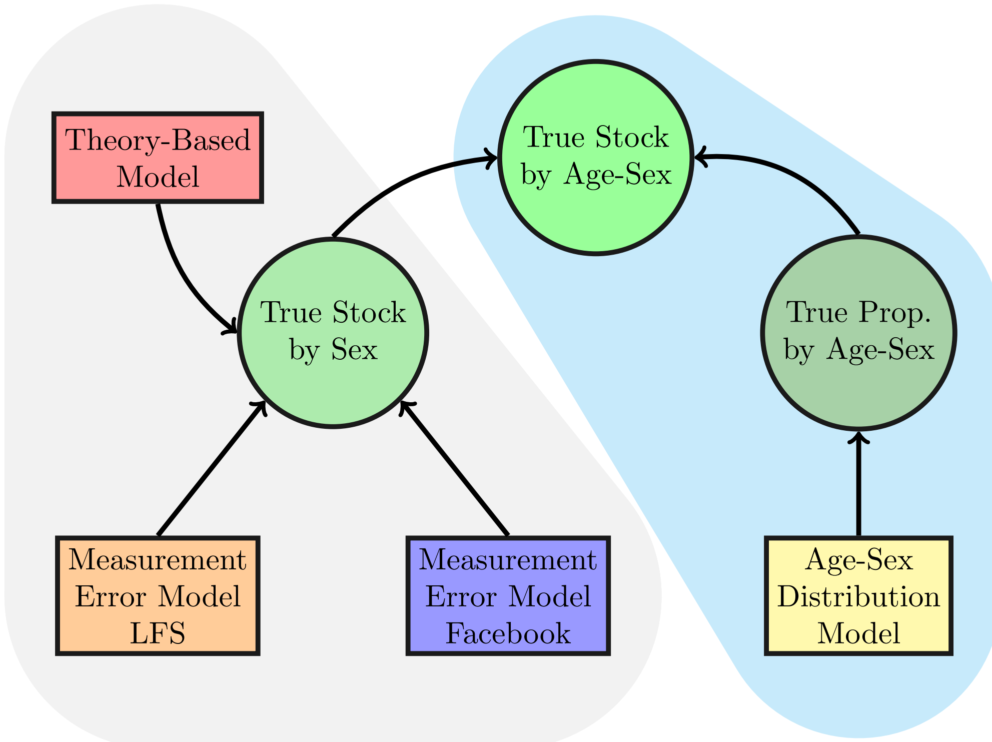

Figure 1. Overview of the model structure. The diagram highlights the two main components: (1) estimation of the true stock by sex using a measurement error and theory-based model and (2) disaggregation of the true stock by age and sex through an age–sex distribution model. Grey indicates components previously developed in Rampazzo et al. (2021), while light blue highlights the novel contributions of this study.

imperfect, source-specific realizations of an unobserved latent migrant stock by origin country, sex, and year. Two submodels combine to identify this latent stock: a *measurement error model* that links each observed source to the latent stock through additive log-scale adjustments for definitional, coverage, algorithmic, and survey-bias effects; and a *theory-based model* that places a regression on the latent stock, with covariates motivated by classical demographic and economic determinants of migration. The novelty of the present article lies in the second component, which redistributes the resulting sex-specific stock totals across age groups.

1. *Measurement error model and theory-based model*: This component estimates the latent true migrant stock by origin country, sex, and year using the IMEM framework proposed by Raymer et al. (2013) and adapted in Rampazzo et al. (2021).

Let i be the index origin country, j be the sex (male, female), t be the year (2018, 2019), and k be the data source, where $k = L$ denotes the LFS and $k = F$ denotes Facebook. The observed source-specific migrant count is denoted by z^k_{ijt}, the corresponding latent reporting mean by μ^k_{ijt}, and the latent true migrant stock by y_{ijt}. The measurement error model is

$$z^k_{ijt} \sim \text{Poisson}(\mu^k_{ijt}),$$

$$\log \mu^k_{ijt} = \log y_{ijt} + \delta^k + \beta^k_{ijt} + \chi^k_{ijt} + \xi^k_{ijt} + \lambda^k_{ijt} + \epsilon^k_{ijt}.$$

The source-specific terms adjust for definitional differences, source bias, coverage, Facebook algorithm changes, language effects, and residual error. Not all terms apply to both sources: for example, the Facebook component includes coverage and algorithm-change adjustments, whereas the LFS component uses survey-bias and residual-error terms. The full prior specification follows Rampazzo et al. (2021) and is summarized in Tables S1 and S2, online supplementary material.

The latent stock y_{ijt} is linked to demographic and economic covariates through the theory-based model

$$\log y_{ijt} = \gamma_0 + \gamma_1 P_{it} + \gamma_2 I_{it} + \gamma_3 O_{it} + \gamma_4 \log G_{it} + \gamma_5 \log U_{it} + \varepsilon_{ijt}.$$

Here, P_{it} is the origin-country population, I_{it} and O_{it} are inflow and outflow measures, and G_{it} and U_{it} are gross domestic product and unemployment ratios relative to the UK. We use γ for these regression coefficients to avoid confusing them with the Dirichlet parameters used in the age-composition model below. We use ε_{ijt} for the theory-level residual here to distinguish it from the source-specific residual ϵ_{ijt}^{k} in the measurement error equation above. The full prior specification, including the priors on γ and the source-specific bias terms, follows Rampazzo et al. (2021) and is reproduced for completeness in the online supplementary material.

4.2 Treatment of unobserved and threshold-governed source-specific cells

The two data sources differ in how they handle small or low-coverage cells. LFS age–sex–country cells with very few interviews are suppressed by ONS and returned as non-numeric flags (. for no contact, c for disclosure control, or rounded zero); we treat these *suppressed* cells as missing, so the latent stock y_{ijt} in the first-component model is identified through the theory-based regression and the available Facebook data, rather than through observed zero counts. The Facebook Advertising Platform is governed by platform-specific minimum-audience thresholds rather than statistical disclosure control, and the platform documentation does not always specify the exact threshold. In the country–age groups considered here, no Facebook audience estimates take the minimum threshold value, so this issue does not arise empirically in our sample. Accordingly, Facebook observations are treated as reported values rather than as suppressed or missing, and are interpreted with caution where coverage may be limited. This treatment is consistent with the differing data-generating processes of the two sources and is implemented directly within the second-component multinomial likelihood, without introducing additional missing-data or measurement adjustments.

2. *Age–sex distribution and integration*: A MDD model is used to estimate the *true proportion by age and sex*. These proportions are then applied to the *true stock by sex* to produce the *true stock by age and sex*. Posterior uncertainty in the sex-specific stock totals y_{ijt} from the first component is propagated forward into the age–sex estimates: the second-component MDD model is run on the same Markov chain Monte Carlo (MCMC) machinery and the age-specific stocks $Y_{sa} = y_s \pi_{sa}$ are computed at each posterior draw, so the credible intervals reported in Section 7 combine uncertainty from both stages rather than treating the first-stage stocks as fixed point estimates.

The first component, highlighted in grey in Figure 1, has already been described in detail in Rampazzo et al. (2021), with equations and parameters provided in the online supplementary material and summarized in this methodological section. The focus of this article is on the second component, the novel component is shown on the right hand-side of Figure 1.

The age-distribution component has two steps. First, reduced Rogers–Castro schedules are fitted to the LFS and Facebook age profiles to obtain smoothed age-specific proportions. These smoothed profiles are then converted into source-specific pseudo-counts that act as a *prior summary* of the demographic shape of the age distribution; this is the only point at which the Rogers–Castro schedule enters the Bayesian hierarchy. Second, the MDD model harmonizes these pseudo-counts and the observed LFS and Facebook age counts within a single hierarchy, producing a posterior age distribution for each country-sex-year stratum. Thus, the Rogers–Castro component supplies a structured prior input, but the final age profiles are updated by the observed source-specific counts wherever those data are available. These age distributions are then multiplied by the posterior sex-specific stock totals y_{ijt} from the first component to obtain migrant stocks by country, sex, year, and age, with full propagation of posterior uncertainty from the first component.

5 Fitting the Rogers–Castro age schedule

The model by Rogers and Castro (1981) was originally developed to describe the intensity of migration flows by age, but its mathematical structure—a sum of one or more exponential-family components plus a constant—is flexible enough to be used more broadly as a parametric smoother for demographically shaped age profiles, including age profiles of migrant *stocks*. In this application, the reduced Rogers–Castro schedule is used in exactly that pragmatic sense: it is a flexible smoothing device for the noisy age-specific stock shares observed in the LFS and Facebook data, not a structural model of how stocks form. It enters the Bayesian model only through pseudo-count vectors constructed before the MDD step is run; its parameters are not estimated jointly with the rest of the hierarchy, and any other parsimonious smoother of age-specific shares could substitute for it. We assess robustness of the substantive age–sex conclusions to changes in the smoothing step in the validation and sensitivity checks below.

As shown in Equation (1), the Rogers–Castro model can be fitted with up to 13 parameters. Since the data used in this study cover the age range 15–65+, a reduced version of the model is applied here, using only the parameters relevant to this age range. To avoid notational clashes with the working-age and postretirement components of Equation (1), we use distinct symbols (A_w, λ_w, A_o, λ_o, μ_o, ν_o, c) here (using ν_o rather than η_o to avoid clashing with the Dirichlet concentration parameters η_a introduced later). The reduced schedule has seven parameters: two for a working-age component, four for an older-age component, and one constant. The postretirement component of Equation (1) is therefore subsumed into the older-age double-exponential of Equation (5), which captures both retirement-age concentration and any postretirement increase compatible with the available age groups. For age midpoint x, the fitted schedule is

$$S(x) = A_w \exp\{-\lambda_w |x - 25|\} + A_o \exp\{-\lambda_o (x - \mu_o) - \exp[-\nu_o (x - \mu_o)]\} + c,$$

which is then normalized to sum to one across the 11 age groups. The working-age component captures the young-adult concentration in migrant stocks, the older-age component allows the curve to rise at later ages, and the constant captures residual background mass. We do not include the Wilson student-migration extension (Wilson, 2010), because the available data are grouped into five-year age bands beginning at 15–19 and the model is applied to migrant stocks rather than single-year migration propensities. In this setting, the additional student curve is not separately identifiable from the working-age component, so adding it would increase flexibility without a clear substantive interpretation (Figure 2).

The components are defined as follows:

1. *Working-age* curve: A left-skewed unimodal curve describing the migration intensity for individuals of working age.
2. *Postretirement* curve: Typically used for postretirement migration, but here adapted to represent the increase in older migrants due to ageing or families bringing elderly parents.
3. *Constant* component: Represents background migration levels.

To fit the Rogers–Castro model, the R package migraR (Ruiz-Santacruz, 2019), available on GitHub, was modified to accommodate the required parameters. The model was fitted simultaneously to data from the Facebook Advertising Platform and the LFS. During exploratory analysis, three distinct country groups were identified based on age profiles:

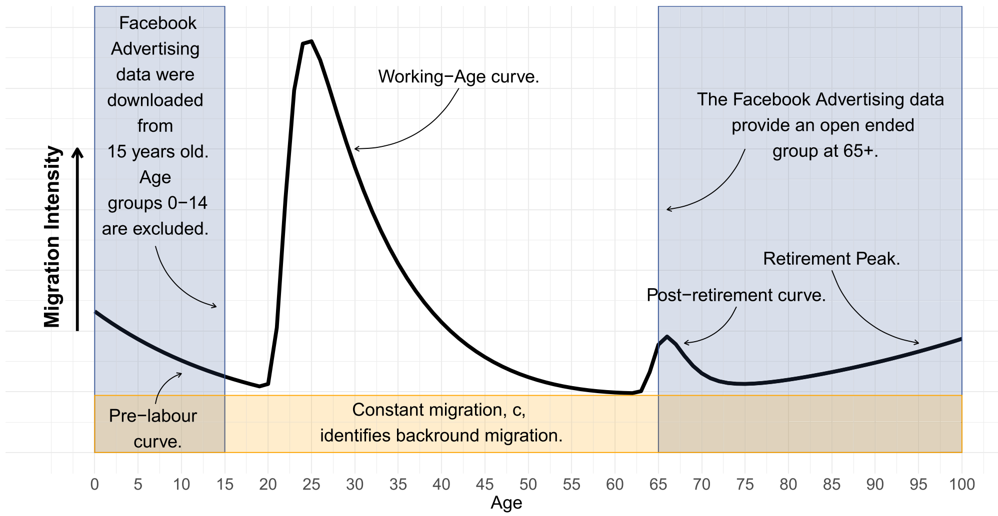

Figure 2. Reduced Rogers–Castro model components and parameter rationale. The figure is a schematic illustrating the three components of the reduced schedule (working-age curve, older-age curve, and constant background level) over age. No empirical data points are plotted in this schematic; observed age-specific shares from the LFS and Facebook data are shown alongside the fitted schedules in Figure 3. Age groups 0–14 are excluded from the analysis and the open-ended group 65+ is treated as a single category. LFS = Labour Force Surveys.

1. *Group 1*: Western and Southern European countries (e.g. France, Germany, Italy, Portugal, and Spain).
2. *Group 2*: Republic of Ireland, reflecting unique age distributions in the LFS data.
3. *Group 3*: Central and Eastern European countries (e.g. Latvia, Lithuania, Poland, and Romania).

Each group demonstrates distinct age profiles of migrants (see Figure 3). For group 1 (Western and Southern Europe), Facebook data indicate younger migrants, whereas LFS data suggest older-age distributions. Group 2 (Republic of Ireland) reflects a long history of migration, with many migrants aged 65+ years, consistent with historical connections between Ireland and the UK. Group 3 (Central and Eastern Europe) represents more recent migration, with patterns emerging after EU accession in 2004. Here, the Facebook and LFS data show similar age distributions, with younger age groups slightly more represented in Facebook data.

The fit of these schedules is assessed below in the validation and sensitivity checks. This placement reflects the role of the Rogers–Castro model in the analysis: it is a first-stage smoothing input whose adequacy must be checked, rather than the main inferential target.

6 The MDD model for age schedules

Given the structure of the data, a statistical model for proportions is required. Let $a = 1, \ldots, A$ index the 11 age groups from 15–19 to 65+, and let $s = 1, \ldots, S$ index the 40 country-sex-year strata (10 origin countries × 2 sexes × 2 years) included in the age-disaggregation analysis. Each stratum s is uniquely determined by the triple (i, j, t) for country, sex, and year defined in the previous section; we use the compact index s here to keep the composition equations readable. Let $r \in \{L, F\}$ index the two sources, and let $g \in \{1, 2, 3\}$ index the three exploratory country groups (Western/Southern Europe, Ireland, and Central/Eastern Europe). The goal is to estimate a latent age-composition vector

$$\boldsymbol{\pi}_s = (\pi_{s1}, \ldots, \pi_{sA}),$$

where $\sum_a \pi_{sa} = 1$, and then to multiply this vector by the posterior sex-specific stock total y_s from the first component of the model.

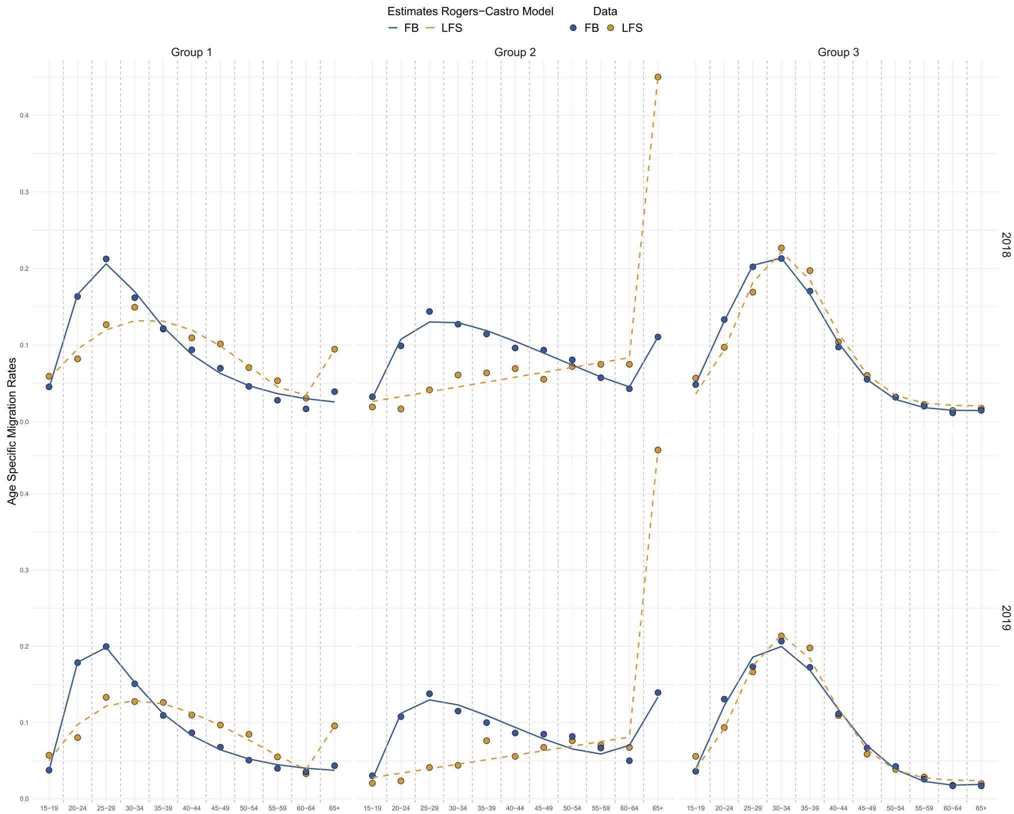

Figure 3. Age-specific migrant-stock shares (proportions, summing to one across age groups within each source–group panel) for EU migrants by country group, 2018–2019. The *y*-axis reports the share of the migrant stock in each five-year age group, computed by dividing each age group count by the source-specific total stock; this share is unitless and does not refer to a migration rate (we do not attempt to compute rates because no consistent population-at-risk denominator is available across the two sources). The Rogers–Castro smoothed schedules (lines) are compared to the observed shares from the Facebook Advertising Platform and LFS (points). Results are shown separately for the three identified country groups: Western and Southern Europe, Republic of Ireland, and Central and Eastern Europe. LFS = Labour Force Surveys.

A natural distribution for age-composition vectors is the Dirichlet distribution, with density

$$\text{Dir}(\boldsymbol{\Theta}|\boldsymbol{\eta}) = \frac{\Gamma\left(\sum_{a=1}^{A}\eta_a\right)}{\prod_{a=1}^{A}\Gamma(\eta_a)}\prod_{a=1}^{A}\theta_a^{\eta_a-1}, \quad \boldsymbol{\eta} = (\eta_1, \ldots, \eta_A).$$

The Dirichlet distribution is well-suited for demographic studies, as demonstrated by Caussinus and Courgeau (2010), where it was used to estimate past (historic and prehistoric) population age structures. Its flexibility allows for redistributing population components into categories constrained by a total. In Bayesian contexts, it offers the additional advantage of being a conjugate prior for multinomial likelihoods, which offers a natural and computationally efficient framework for modelling proportions.

This study employs a MDD model to harmonize migration age profiles from the LFS and Facebook Advertising data. The model is estimated by including variables for country, age, sex, and year, borrowing strength across these parameters.

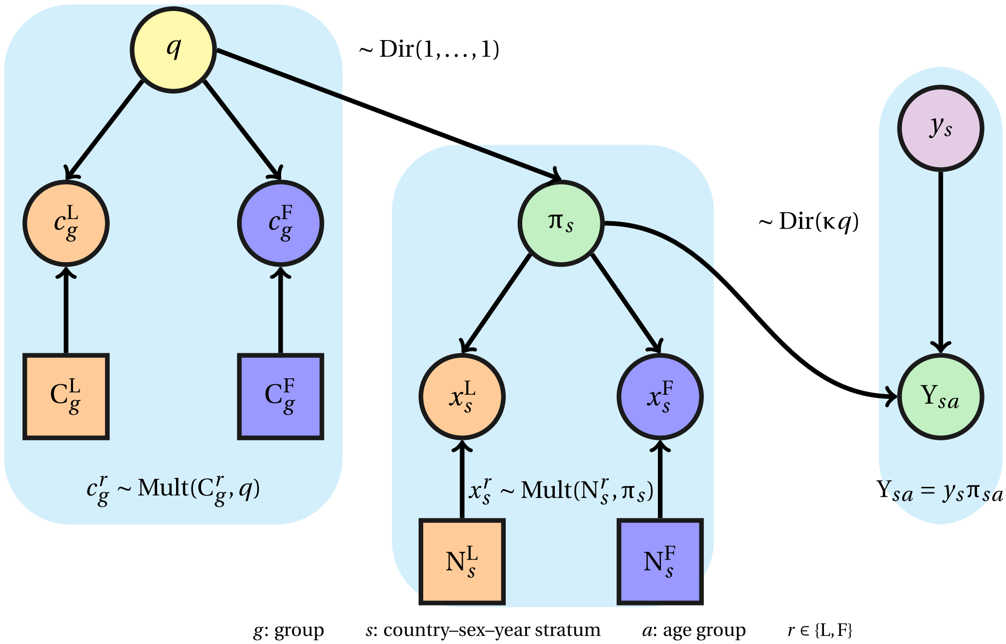

Figure 4. Hierarchical structure of the multinomial-Dirichlet–Dirichlet age-composition model. The shared age profile $\boldsymbol{q}$ generates Rogers–Castro pseudo-counts $\boldsymbol{c}_g^r$ and the stratum-specific age profile $\boldsymbol{\pi}_s$, with $r \in \{L, F\}$ indexing the data source. Observed LFS and Facebook age-count vectors $\boldsymbol{x}_s^L$ and $\boldsymbol{x}_s^F$ are modelled as multinomial draws from $\boldsymbol{\pi}_s$. The final age-specific stock is given by $Y_{sa} = y_s \pi_{sa}$, where y_s is the sex-specific total from the first model component. LFS = Labour Force Surveys.

6.1 Model hierarchy

The hierarchical structure of the MDD model is illustrated in Figure 4. The notation below uses distinct symbols for the three quantities that were easy to confuse in the previous version of the manuscript: $\mathbf{q}$ is the shared age profile, $\mathbf{c}_g^r$ are Rogers–Castro-derived pseudo-counts for source r and country group g, and $\boldsymbol{\pi}_s$ is the final age profile for stratum s.

The shared age profile has a symmetric Dirichlet prior:

$$\mathbf{q} \sim \mathrm{Dir}(1, \ \ldots, 1),$$

which does not favour any age group a priori. The Rogers–Castro smoothing step produces age-specific pseudo-counts $\mathbf{c}_g^L$ and $\mathbf{c}_g^F$ for each source and country group g. If $C_g^r = \sum_a c_{ga}^r$ is the corresponding pseudo-count total, these vectors are modelled as multinomial draws around the shared age profile:

$$\mathbf{c}_g^r \sim \mathrm{Multinomial}(C_g^r, \mathbf{q}), \qquad r \in \{L, F\}.$$

This is the point at which the Rogers–Castro schedules enter the Bayesian model. They enter only through pseudo-count vectors constructed before Just Another Gibbs Sampler (JAGS) is run; the Rogers–Castro parameters themselves are not estimated jointly with the MDD hierarchy. The schedules therefore act as a prior smoothing device, not as fixed final estimates: the observed LFS and Facebook age-count vectors below update the resulting stratum-specific profiles wherever data are available. The posterior intervals from the MDD model should be interpreted as conditional on this first-stage smoothing step. We assess the adequacy of the Rogers–Castro step and the robustness of the resulting age profiles, in the validation and sensitivity checks below.

For each country-sex-year stratum s, the latent age profile is then drawn from a Dirichlet distribution centred on the shared profile:

$$\boldsymbol{\pi}_s \sim \mathrm{Dir}(\kappa \mathbf{q}).$$

Here, $\kappa > 0$ is the concentration parameter that controls how tightly the stratum-specific age profile $\boldsymbol{\pi}_s$ is shrunk towards the shared profile $\mathbf{q}$. Larger values of κ imply stronger pooling, in the sense that all strata are pulled close to the shared age profile, while smaller values allow each country-sex-year combination to depart further from the common shape. We therefore interpret κ as a *reliability/smoothing* parameter rather than as an algorithmic tuning knob: it directly regulates the prior plausibility of the stratum-specific profiles. In the submitted model, κ is fixed implicitly by the concentration of the vector used in JAGS; in the sensitivity checks, we assess whether the results are robust to alternative concentration values. The observed LFS and Facebook age profiles for each stratum are treated as multinomial observations around the same latent age profile:

$$\begin{aligned} \mathbf{x}_s^L &\sim \mathrm{Multinomial}(N_s^L, \boldsymbol{\pi}_s), \\ \mathbf{x}_s^F &\sim \mathrm{Multinomial}(N_s^F, \boldsymbol{\pi}_s), \end{aligned}$$

where $\mathbf{x}_s^L$ and $\mathbf{x}_s^F$ are the observed age-count vectors from the LFS and Facebook, and N_s^L and N_s^F are their source-specific totals. Finally, the estimated stock in age group a for stratum s is

$$Y_{sa} = y_s \pi_{sa}.$$

This final multiplication is why the age-specific estimates sum to the sex-specific stock totals by construction. The substantive contribution of the MDD layer is therefore not to change the total stock, but to estimate how that stock is distributed across age groups while reconciling the LFS and Facebook age profiles. The Dirichlet composition layer is chosen for transparency, conjugacy, and computational efficiency. We acknowledge that the Dirichlet imposes a single concentration parameter and a restricted (negative) dependence structure across age groups, which can be limiting in higher-dimensional or more dispersed compositions. More flexible alternatives such as the Beta–Liouville distribution (Bakhtiari & Bouguila, 2016; Fang et al., 1990) offer richer dependence structures and could be substituted into the same hierarchy; we view the Dirichlet specification used here as a deliberate, pragmatic trade-off that is appropriate for the size and structure of the present application, rather than a claim of maximal flexibility.[2]

The harmonized population pyramids derived from this model are presented in Figure 5, which compares the migration proportions by age and sex from the Facebook Advertising data (blue lines) and LFS data (yellow lines) with the harmonized results (grey lines). These lines show the migration profiles for three groups of countries, reflecting the combined age and sex distributions derived from the Rogers–Castro estimates.

7 Results

In this section, the results of the models for the three groups of countries across two years are presented. The models were estimated in R using JAGS (Plummer et al., 2016), with three chains, a burn-in/adaptation period, and a thinning interval of 10. The model was run simultaneously for both years (2018 and 2019) and all 10 origin countries; results are discussed separately for Western and Southern European, Irish, and Central and Eastern European migrants.

[2] In Figure 4, square nodes denote scalar source-specific totals or effective sample sizes, such as C_g^r and N_s^r, while circular nodes denote age profile vectors, latent composition parameters, or derived stock quantities. Orange marks LFS quantities, indigo marks Facebook quantities, yellow marks the shared age profile prior $\mathbf{q}$, green marks the stratum-specific age profile $\boldsymbol{\pi}_s$ and the derived age-specific stock Y_{sa}, and pink marks the sex-specific stock total y_s imported from the first model component.

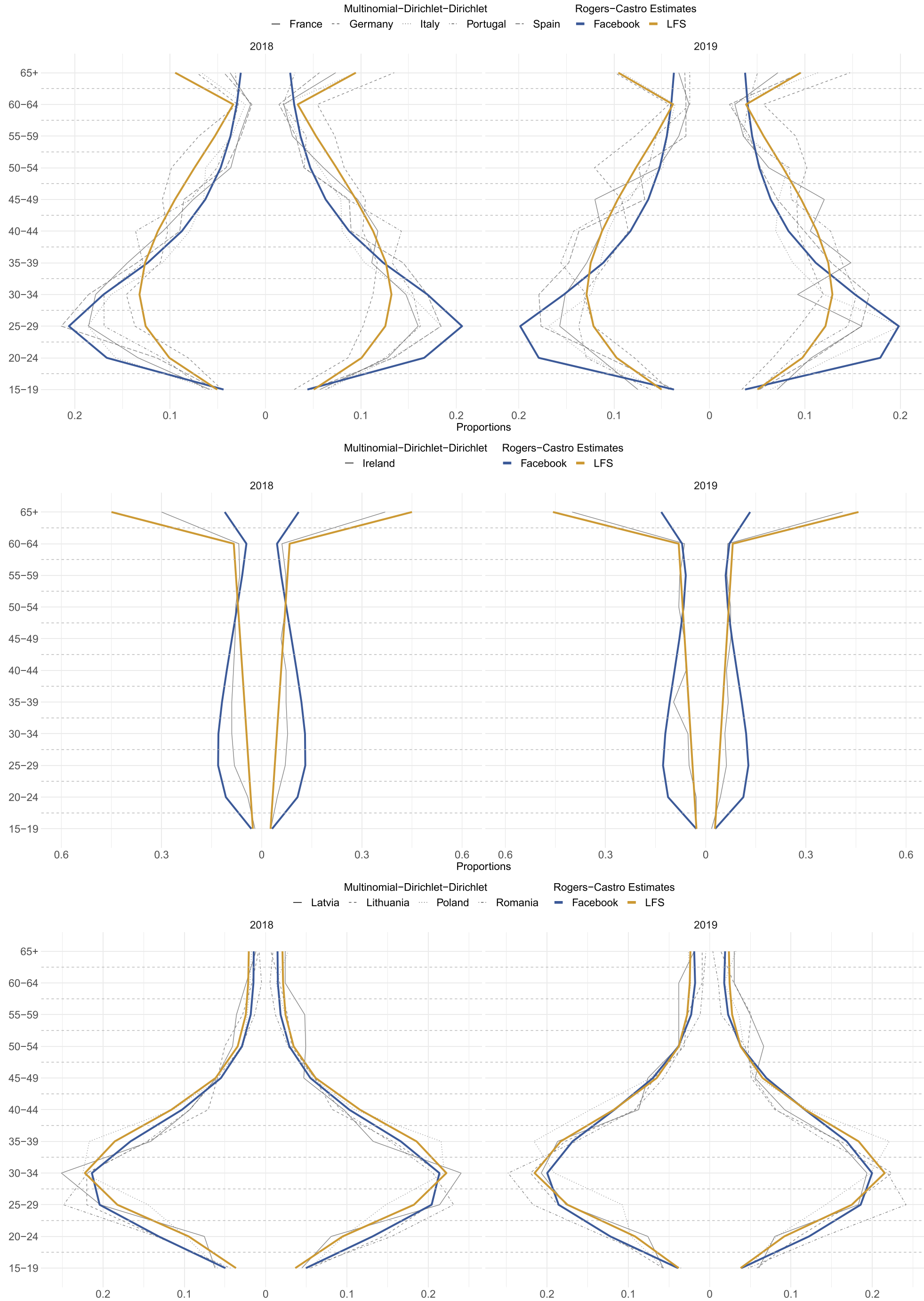

Figure 5. Harmonized population pyramids based on the first-stage Rogers–Castro smoothed shares, before the MDD model is run. This figure is illustrative and precedes the Bayesian estimation: the pyramids are not posterior MCMC outputs of the MDD model. They show migration proportions by age and sex for three groups of countries derived from the reduced Rogers–Castro fits described above. Solid blue lines represent Facebook Advertising Platform data, solid yellow lines represent LFS data, and dashed and solid grey lines show the harmonized age profiles combining both sources. The corresponding posterior MCMC estimates from the full multinomial-Dirichlet–Dirichlet model are reported later in Figures 6–8. LFS = Labour Force Surveys; MDD = multinomial-Dirichlet–Dirichlet.

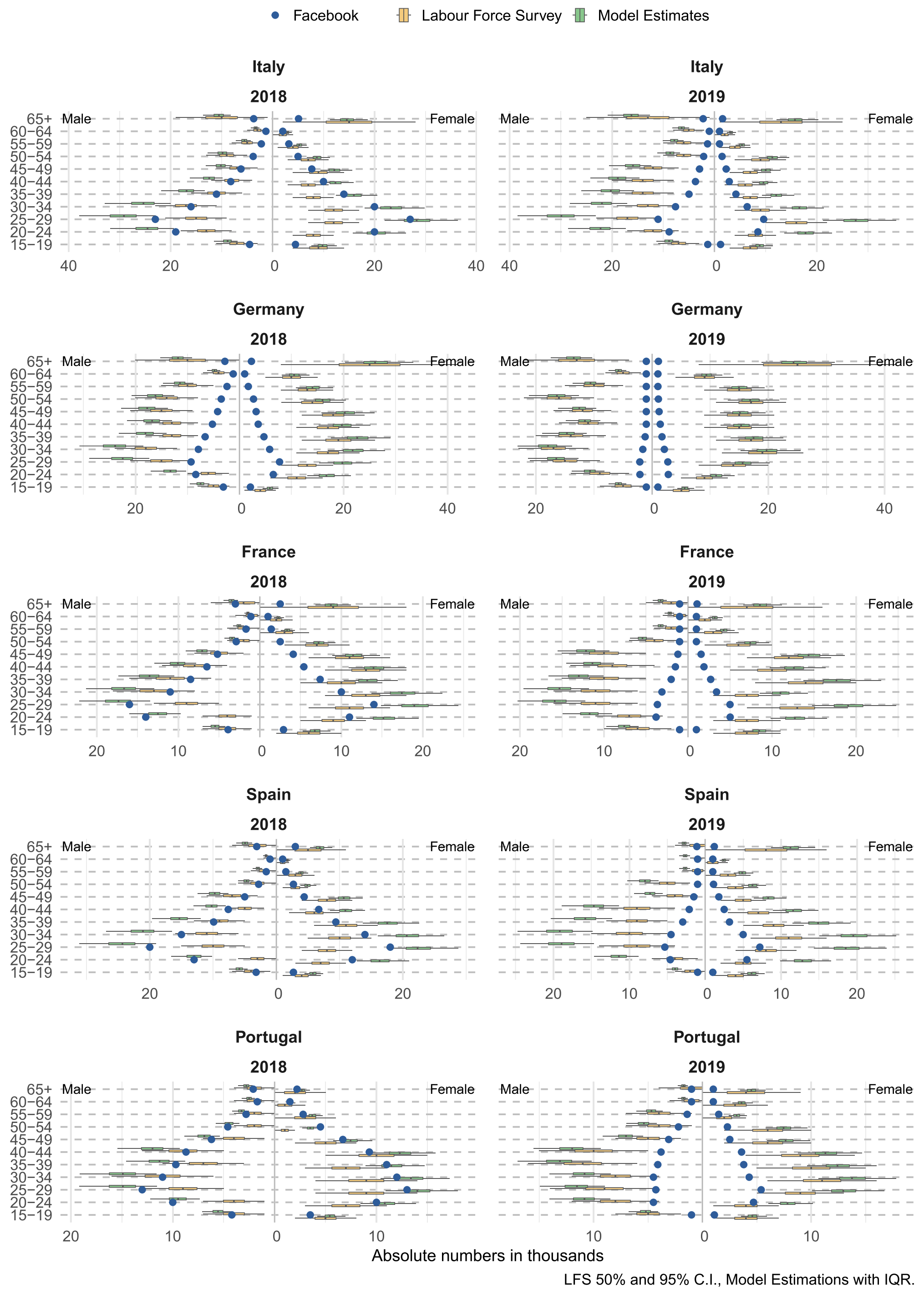

Figure 6. Age–sex distribution for Western and Southern European migrants in the UK. Population pyramids use box plots to compare model estimates (green), Labour Force Survey (yellow), and Facebook data (blue, represented as points with no intervals) for 2018 and 2019. Box plots illustrate median values, interquartile ranges, and variability across sources.

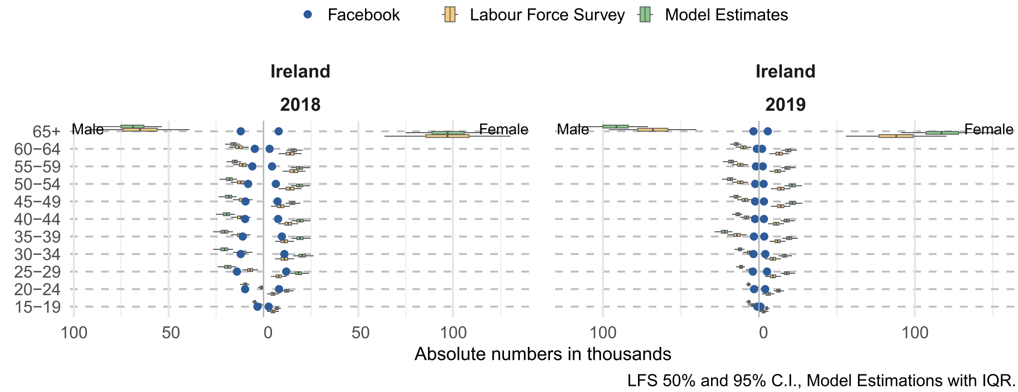

Figure 7. Age–sex distribution for Irish migrants in the UK. Population pyramids use box plots to represent age–sex profiles of Irish migrants, showing model estimates (green), Labour Force Survey (yellow), and Facebook data (blue, represented as points with no intervals) for 2018 and 2019. The box plots highlight the unique age structure of Irish migration, particularly among older-age groups.

Figures 6–8 present the results. Green box plots show the interquartile range of the model estimates, the LFS data are shown as yellow box plots with 50% and 95% intervals, and Facebook Advertising Platform values are shown as blue points. All results are expressed as counts of migrants by age and sex.

Convergence was assessed using the potential scale reduction factor $\hat{R}$ (Gelman–Rubin statistic) and effective sample sizes (Gelman et al., 2013). The main MDD quantities show satisfactory convergence; the diagnostic run reported below has maximum $\hat{R} = 1.002$ and minimum effective sample size 7,295 across the monitored age profile, age–sex count, stock, and sum-proportion blocks. Brief posterior predictive p-values for cross-age variance, chi-square count discrepancy, and the share of stock at ages 20–34 are summarized in the validation section, with full diagnostics in the online supplementary material.

The model estimates for Western and Southern European migrants are shown in Figure 6. A reduction in the number of migrants between 2018 and 2019 is evident. For countries in this group, the largest number of migrants is concentrated in the 20–29 age group. However, there is greater uncertainty in age groups with larger numbers of migrants. The model estimates indicate more migrants across the age groups (except for Germany in 2019) compared to the LFS.

The population pyramids for the Irish migrants in the UK are presented in Figure 7. The estimates of the number of migrants are highest for the age group 65+ years. 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 at this time. There is higher uncertainty in the estimates at older ages, and generally, the model estimates more total migrants than the LFS.

As shown in Figure 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 is older than the Western and Southern European migrants, being between 25–29 and 30–34 years. 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.

8 Validation and sensitivity checks

We report four validation and robustness checks: external validation, posterior predictive checks, comparison with the area model of Wiśniowski et al. (2016) for age–sex estimation, and sensitivity to prior specification.

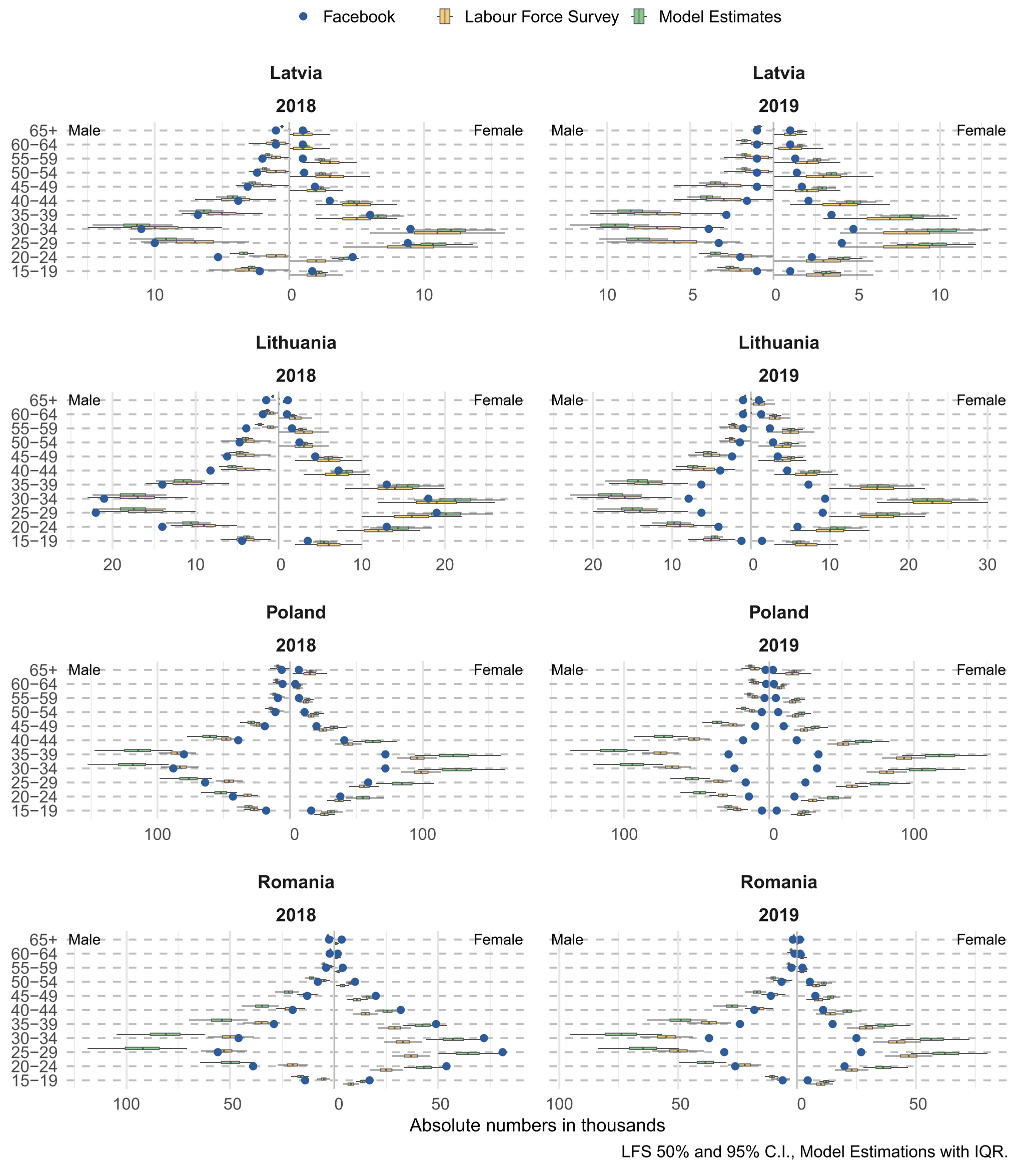

Figure 8. Age–sex distribution for Central and Eastern European migrants in the UK. Population pyramids use box plots to compare age–sex profiles derived from model estimates (green), Labour Force Survey (yellow), and Facebook data (blue, represented as points with no intervals) for 2018 and 2019. Across the four origin countries shown, working-age stocks are concentrated in the 25–34 age range; the median ages are slightly older than those for Western and Southern European migrants but markedly younger than those for Irish migrants.

First, we compared the 2019 posterior age–sex profiles with Census 2021 country-of-birth profiles for England and Wales, using the closest available external benchmark for the selected origin countries. Although the Census refers to a later date, the age–sex composition of migrant stocks is not expected to change abruptly over a two-year period, so the main comparability limitation is geographical: our model estimates refer to the UK, whereas the Census benchmark used here covers England and Wales only. We therefore treat the Census comparison as an external check on age–sex *shape*, not as a direct validation of absolute stock totals or posterior interval coverage. We did not

Table 1. External validation against Census 2021 age–sex country-of-birth profiles

Metric	MDD model	Wiśniowski et al. (2016)
Mean absolute age-share error	0.017	0.032
RMSE age-share error	0.022	0.055
Median absolute age-share error	0.013	0.020

Note. MDD = multinomial-Dirichlet–Dirichlet.

extend the comparison to Eurostat country-of-birth tables, because in this study Eurostat appears upstream as an input to the first-component theory-based regression rather than as an independent benchmark, but cross-country validation against Eurostat is a natural extension that we discuss in Section 9. Second, we used posterior predictive checks to assess whether the model replicates observed features of the LFS and Facebook age profiles. Third, we implemented an age–sex specification inspired by the area model of Wiśniowski et al. (2016) and compared it with the MDD specification. Fourth, we assessed sensitivity to the prior on the Dirichlet concentration parameter κ, so that the dual roles of κ as a pooling/smoothing control and as a determinant of the prior on stratum-specific dispersion are made transparent.

8.1 Census 2021 external validation

We then compared the estimated 2019 age–sex profiles with the relevant Census 2021 country-of-birth table, which reports population by country of birth, sex, age, and date of arrival in England and Wales. The comparison is made on age shares within each country-sex profile, rather than on absolute counts, because the model input data refer to UK migrant stocks in 2018–2019, while the Census benchmark refers only to England and Wales in 2021 and therefore excludes Scotland and Northern Ireland. The two-year timing difference is unlikely to explain large changes in the age structure of established migrant stocks; the more important caveat is this geographical mismatch, especially if the age composition of a migrant group differs between England and Wales and the rest of the UK. Apparent out-of-interval differences in absolute counts should therefore be read primarily as a data-comparability artefact, while remaining discrepancies in age shares indicate the limits of the model and its input data in recovering the externally observed profile. This validation therefore asks whether the estimated *shape* of the age–sex profile is externally plausible.

The MDD model is consistently closer to the Census benchmark than the age–sex specification inspired by the area model of Wiśniowski et al. (2016). The mean absolute age-share error is 0.0166 for MDD and 0.0325 for the sensitivity model; the corresponding Root Mean Square Error (RMSE) values are 0.0222 and 0.0546. The median absolute age-share error is also lower for MDD (0.0131) than for the sensitivity model (0.0205). These results indicate that the MDD model better recovers the externally observed shape of migrant age–sex profiles.

The population-pyramid figures below show the MDD posterior age–sex profiles against Census 2021 points (Figures 9–11). Because the points and intervals refer to different geographical populations, nonoverlap in absolute counts should not be interpreted as posterior interval failure. The figures instead provide a visual check on whether the model captures the main country-specific age–sex shape, complementing the numerical comparison with the age–sex specification inspired by the area model of Wiśniowski et al. (2016) in Table 1.

We assess robustness of the MDD age–sex profiles by comparing them with an age–sex specification inspired by the area model of Wiśniowski et al. (2016). This specification uses a lognormal multiplicative age–sex structure. In the comparison reported here, sex-specific totals are fixed to the MDD posterior stock totals, so this specification is treated as a sensitivity check for the conditional age profile within sex, rather than as a fully independent estimate of total migrant stocks. For the comparable

Table 2. Comparison of convergence, external Census validation, and posterior predictive checks for the MDD model and the age–sex specification inspired by the area model of Wiśniowski et al. (2016)

Model	Convergence	Census Mean Absolute Error	Census RMSE	Posterior Predictive Check extremes	Main interpretation
MDD model	Max Rhat 1.002; min ESS 7295	0.017	0.022	102/240	Best external validation; clean convergence
Wiśniowski et al. (2016)	Primary max Rhat 1.007; min ESS 1292	0.032	0.055	64/240	Better source replication; weaker Census fit; sex-share block not interpreted

Note. MDD = multinomial-Dirichlet–Dirichlet.

quantities, the comparison specification converges satisfactorily: the conditional age profile and age–sex count blocks have maximum $\hat{R} = 1.007$. The full joint sex-share block is weakly identified under this fixed-sex-total setup, with $\hat{R}$ reaching 3.122, and is therefore not interpreted substantively.

Table 2 summarizes the model diagnostics. The MDD model has clean convergence, with maximum $\hat{R} = 1.002$ and minimum effective sample size of 7,295 in the diagnostic run. It also provides a closer match to the Census benchmark: the mean absolute age-share error is 0.0166 for the MDD model, compared with 0.0325 for the comparison specification; the corresponding RMSE values are 0.0222 and 0.0546. The comparison specification gives fewer extreme posterior predictive checks for the observed LFS and Facebook summaries (64 of 240, compared with 102 of 240 for the MDD model), especially for the chi-square discrepancy statistic, indicating that it reproduces some source-specific within-sample summaries more closely. The two specifications are therefore complementary: the comparison specification is informative about source-specific dispersion, while the MDD model is closer to the external Census benchmark and converges more cleanly.

The posterior predictive checks focus on three features of the observed age profiles: cross-age variance (T1), chi-square discrepancy on the count scale (T2), and the share of the profile concentrated at ages 20–34 (T3). Bayesian posterior predictive *p*-values close to 0.5 indicate that the replicated data resemble the observed data with respect to the statistic in question, whereas values below 0.05 or above 0.95 indicate an extreme discrepancy. The MDD model performs more strongly in external validation against Census data, while the comparison specification reproduces some within-sample replication statistics more closely. The comparison therefore serves as a sensitivity check on the MDD specification.

8.2 Rogers–Castro fit diagnostics and sensitivity to priors

Finally, we assessed the adequacy of the reduced Rogers–Castro schedules used as first-stage smoothing inputs. The fit is mixed, which is expected because the schedules are applied to migrant stocks rather than flows: the strongest fits occur for the Irish LFS profiles ($R^2 = 0.973$ in 2018 and 0.975 in 2019), where the older-age concentration is very pronounced; Facebook profiles fit reasonably well for Western/Southern Europe ($R^2 = 0.779$–0.784); LFS profiles for the same group have weaker fit ($R^2 = 0.369$–0.399); and Central and Eastern European profiles show moderate fit ($R^2 = 0.507$–0.670). These diagnostics support using the Rogers–Castro schedule as a smoothing prior on age shares, while reinforcing that it should not be interpreted as a fully adequate generative model for every stock profile; the observed age-count data can still pull the posterior profiles away from this prior shape. Full

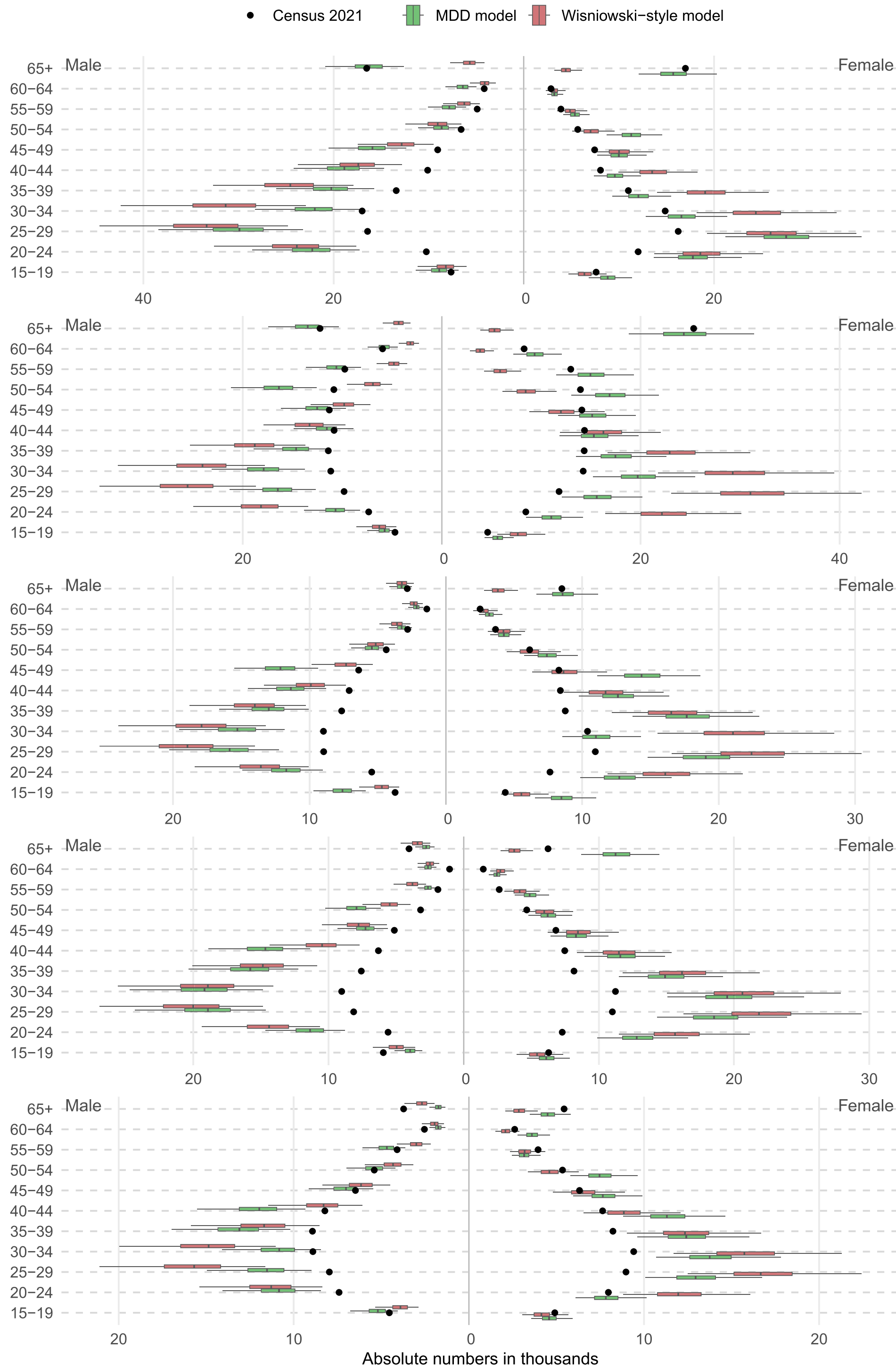

Figure 9. Population-pyramid comparison with Census 2021 for Western and Southern European countries. Solid black dots show Census 2021 counts in thousands. Box plots show Bayesian estimates in thousands for the MDD model (green) and the Wisniowski-style age–sex sensitivity model (red). Male profiles are mirrored to the left and female profiles to the right; *x*-axes are allowed to vary by country panel so that differences in country magnitude do not obscure the age–sex pattern. MDD = multinomial-Dirichlet–Dirichlet.

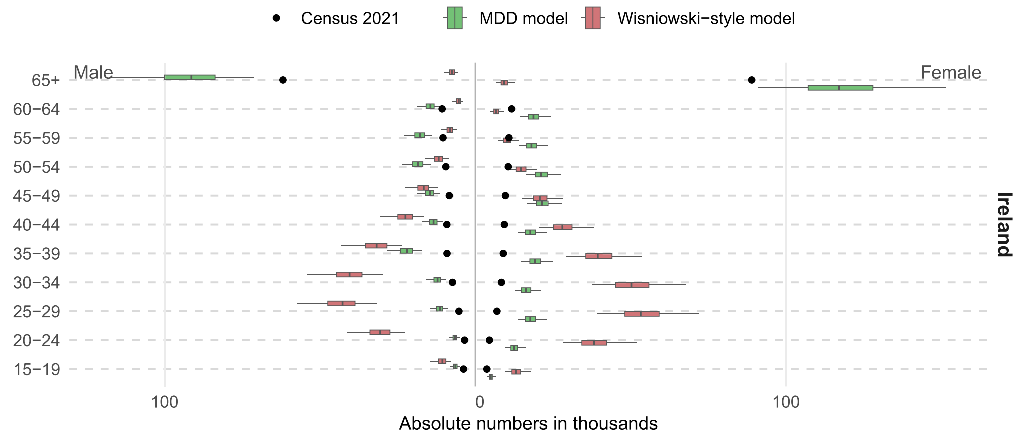

Figure 10. Population-pyramid comparison with Census 2021 for Ireland. Solid black dots show Census 2021 counts in thousands. Box plots show Bayesian estimates in thousands for the MDD model (green) and the Wisniowski-style age–sex sensitivity model (red). Male profiles are mirrored to the left and female profiles to the right; *x*-axes are allowed to vary by country panel so that differences in country magnitude do not obscure the age–sex pattern. MDD = multinomial-Dirichlet–Dirichlet.

parameter estimates with bootstrap standard errors, the goodness-of-fit table, the fitted-schedule diagnostic figure (see Figure S1, online supplementary material), and the country-level residual plot (see Figure S2, online supplementary material) are provided in the online supplementary material.

To assess robustness of the MDD age–sex profiles to the prior on the Dirichlet concentration parameter $\varkappa$, we refitted the MDD model with $\varkappa \in \{0.1, 1, 10, 100\}$. The posterior age–sex profiles are stable across these values: extreme $\varkappa$ values change the apparent precision of the stratum-specific profiles but leave the substantive ranking of countries and the location of the working-age peak essentially unchanged. The corresponding figure is reported in Figure S3, online supplementary material. The posterior predictive checks reported above further indicate that a single Dirichlet-style dependence and concentration structure cannot reproduce every aspect of the observed source-specific dispersion. Richer composition models, including logistic-normal specifications or the distributional family noted above, could allow heterogeneous dispersion across age groups and more flexible covariance patterns. We retain the MDD specification here because it has stronger convergence and better external validation against Census 2021.

9 Conclusions

This article extended the estimates from Rampazzo et al. (2021) to include age and sex disaggregation by proposing a hierarchical MDD model. The model uses reduced Rogers–Castro schedules as first-stage smoothing inputs for age-specific stock shares, before harmonizing the LFS and Facebook profiles in the Bayesian hierarchy. This use is deliberately pragmatic: the Rogers–Castro schedule provides a parsimonious demographic shape constraint, but is not interpreted as a complete generative model for migrant stocks.

In substantive terms, the results revealed distinct age and sex profiles among migrant groups. Western and Southern European migrants in the UK tend to be younger, with the largest numbers concentrated in the 20–29 age group. Eastern European migrants are only slightly older on average, with relatively fewer older adults. In contrast, Irish migrants exhibit a significantly older profile, reflecting a long history of migration to the UK and the resulting accumulation of older adults within this group.

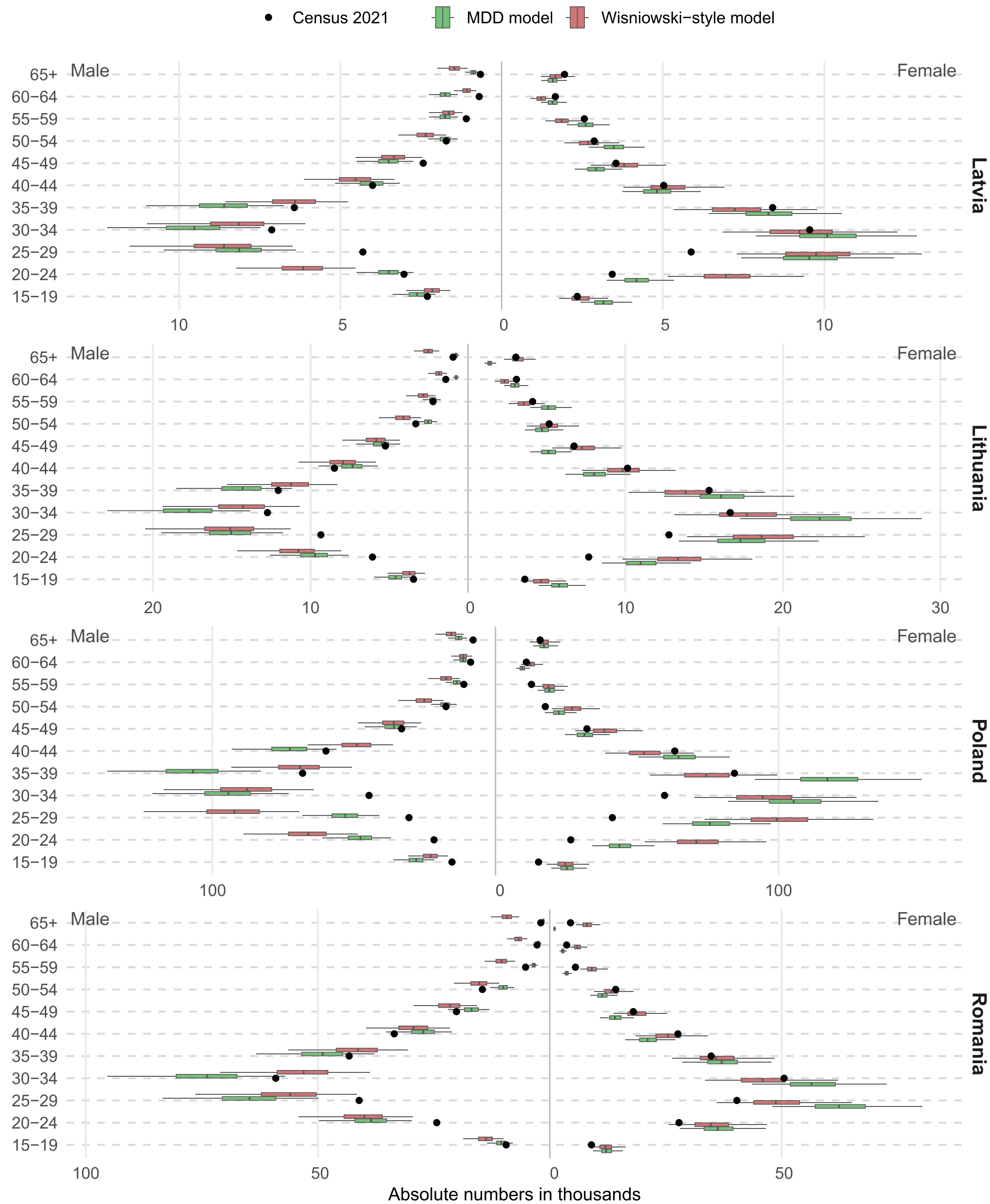

Figure 11. Population-pyramid comparison with Census 2021 for Central and Eastern European countries. Solid black dots show Census 2021 counts in thousands. Box plots show Bayesian estimates in thousands for the MDD model (green) and the Wisniowski-style age–sex sensitivity model (red). Male profiles are mirrored to the left and female profiles to the right; *x*-axes are allowed to vary by country panel so that differences in country magnitude do not obscure the age–sex pattern. MDD = multinomial-Dirichlet–Dirichlet.

The inclusion of data from the Facebook Advertising Platform enhanced the coverage of younger migrants, addressing gaps in traditional surveys such as the LFS. This integration highlights the utility of combining digital trace data with traditional sources to provide a more comprehensive understanding of migrant demographics. While European migrants are predominantly younger, Irish migrants represent a notable exception due to historical migration patterns. The application of the Rogers–Castro model to stock data demonstrated its limitations in capturing the unique age profile of older migrant populations, such as those from Ireland. Although demonstrated with Facebook data, the model itself is data-agnostic. Any source providing age-structured proportions, such as other social media advertising platforms (e.g. TikTok Ads, Google Ads, Instagram), mobile-phone metadata, email-provider demographic panels, or administrative records, can be integrated into the same modelling framework. The MDD hierarchy allows these sources to be combined in a computationally efficient and statistically coherent manner.

By harmonizing age and sex profiles from two data sources, this study demonstrated a framework for addressing gaps in migration estimates from traditional data sources. The differing profiles observed suggest varied motives for migration and lengths of stay, with Western and Southern European migrants appearing to move to the UK for temporary periods, while Eastern European migrants seem more likely to settle, as evidenced by their slightly older-age distribution.

The MDD framework and the age–sex specification inspired by the area model of Wiśniowski et al. (2016) share the goal of disaggregating migrant counts by age and sex, but they sit on different points of a within-sample/external validation trade-off. The MDD model is computationally efficient because it relies on conjugate Dirichlet priors and a parsimonious parametrization; in this application, it converges cleanly and is closer to the external Census 2021 age–sex benchmark. The comparison specification reproduces some observed LFS and Facebook within-sample summaries more closely, particularly the source-specific dispersion captured by the chi-square discrepancy statistic, which is useful for diagnostic purposes. We therefore view the two models as complementary rather than substitutable, and read the comparison as a balanced robustness check on the MDD estimates. Future extensions may incorporate richer dispersion structures, for example, Beta–Liouville or logistic-normal composition layers or accuracy measures of the kind discussed in Wiśniowski et al. (2016), combining the within-sample flexibility of the comparison specification with the external accuracy and computational efficiency of the MDD model.

The analysis, conducted on data for 2018 and 2019, revealed shifts in migration estimates between the two years. Understanding the impact of significant events, such as Brexit, on the number and characteristics of European migrants in the UK remains a crucial area for further research (Di Iasio & Wahba, 2023). The integration of digital trace data and traditional sources offers a valuable path forward for improving migration estimates and informing effective policy and service planning for diverse migrant populations. However, caution is needed regarding the varying popularity of different social media platforms and the availability of data, which may change over time. It is essential to develop models that can be adapted to cultural and normative changes in the use of social media, and to be critical when interpreting their results, which are very likely to be context-specific.

Facebook's demographic composition and reporting practices continue to change over time, particularly among younger users. Nevertheless, the modelling framework presented here is not platform-dependent. The data pipeline deriving age–sex proportions, fitting parametric age schedules, and harmonizing them within a MDD hierarchy is fully reproducible with any data source that provides comparable age and sex breakdowns. As new digital platforms emerge or existing ones decline, the same approach can be adapted to integrate alternative sources while retaining the statistical structure of the model.

Acknowledgements

We thank the editor and three anonymous reviewers for their constructive feedback, which helped improve the paper. We thank Arkadiusz Wiśniowski for comments on an early draft of this paper. We also thank Maciej Dańko for sharing the JAGS code for the model developed in Wiśniowski et al. (2016).

Conflicts of interest

The authors declare no conflict of interest.

Funding

Francesco Rampazzo conducted most of his research for this article while a Ph.D. student at the Department of Social Statistics and Demography at the University of Southampton, with a scholarship from the Economic and Social Research Council through the South Coast Doctoral Training Partnership (Project ES/P000673/1). Jakub Bijak was supported by the Leverhulme Trust through the Leverhulme Centre for Demographic Science (Grant RC-2018-003). Ingmar Weber's work was supported by funding from the Alexander von Humboldt Foundation and its founder, the Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung).

Data availability

The code and replication materials underlying this article are available in the GitHub repository at https://github.com/chiccorampazzo/age_extension_model. Source data obtained from third-party providers, including the LFS and the Facebook Advertising Platform, are subject to the access conditions and terms of use of the original data providers.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series A*.

References

Alexander, M., Polimis, K., & Zagheni, E. (2022). Combining social media and survey data to nowcast migrant stocks in the United States. *Population Research and Policy Review*, *41*(1), 1–28. https://doi.org/10.1007/s11113-020-09599-3

Araujo, M., Mejova, Y., Weber, I., & Benevenuto, F. (2017). Using Facebook Ads audiences for global lifestyle disease surveillance: Promises and limitations. Proceedings of the 2017 ACM on web science conference *(WebSci '17*) (pp. 253–257). ACM. https://doi.org/10.1145/3091478.3091513

Aref, S., Zagheni, E., & West, J. (2019). The demography of the peripatetic researcher: Evidence on highly mobile scholars from the web of science. In I. Weber, K. M. Darwish, C. Wagner, E. Zagheni, L. Nelson, S. Aref, & F. Flöck (Eds.), *Social informatics* (pp. 50–65). Springer International Publishing. https://doi.org/10.1007/978-3-030-34971-4_4

Avramescu, A., & Wiśniowski, A. (2021). Now-casting Romanian migration into the United Kingdom by using google search engine data. *Demographic Research*, *45*(40), 1219–1254. https://doi.org/10.4054/DemRes.2021.45.40

Bakhtiari, S., & Bouguila, N. (2016). A latent Beta-Liouville allocation model. *Expert Systems with Applications*, *45*, 260–272. https://doi.org/10.1016/j.eswa.2015.09.044

Bernard, A., Bell, M., & Charles-Edwards, E. (2014). Life-course transitions and the age profile of internal migration. *Population and Development Review*, *40*(2), 213–239. https://doi.org/10.1111/j.1728-4457.2014.00671.x

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. https://doi.org/10.1080/02681102.2011.643209

Caussinus, H., & Courgeau, D. (2010). Estimating age without measuring it: A new method in paleodemography. *Population (English Edition, 2002)*, *65*(1), 117–144. https://doi.org/10.3917/pope.1001.0117

Chi, G., Abel, G. J., Johnston, D., Giraudy, E., & Bailey, M. (2025). Measuring global migration flows using online data. *Proceedings of the National Academy of Sciences*, *122*(18), Article e2409418122. https://doi.org/10.1073/pnas.2409418122

Courgeau, D. (1985). Interaction between spatial mobility, family and career life-cycle: A French survey. *European Sociological Review*, *1*(2), 139–162. https://doi.org/10.1093/oxfordjournals.esr.a036382

Dańko, M. J., Wiśniowski, A., Jasilionis, D., Jdanov, D. A., & Zagheni, E. (2024). Assessing the quality of data on international migration flows in Europe: The case of undercounting. *Migration Studies*, *12*(2), Article mnae014. https://doi.org/10.1093/migration/mnae014

Di Iasio, V., & Wahba, J. (2023). Expecting Brexit and UK migration: Should I go? *European Economic Review*, *157*(August), Article 104484. https://doi.org/10.1016/j.euroecorev.2023.104484

Disney, G. (2015). *Model-based estimates of UK immigration* [PhD thesis]. University of Southampton.

Fang, K.-T., Kotz, S., & Ng, K. W. (1990). Symmetric multivariate and related distributions. In *Monographs on statistics and applied probability* (Vol. *36*). Chapman and Hall.

Fiorio, L., Zagheni, E., Abel, G. J., Hill, J., Pestre, G., Letouzé, E., & Cai, J. (2021). Analyzing the effect of time in migration measurement using georeferenced digital trace data. *Demography*, *58*(1), 51–74. https://doi.org/10.1215/00703370-8917630

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data analysis* (3rd ed.). CRC Press.

Grow, A., Perrotta, D., Del Fava, E., Cimentada, J., Rampazzo, F., Gil-Clavel, S., Zagheni, E., Flores, R. D., Ventura, I., & Weber, I. (2022). Is Facebook's advertising data accurate enough for use in social science research? Insights from a cross-national online survey. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *185*(S2), S343–S363. https://doi.org/10.1111/rssa.12948

Hargittai, E. (2018). Potential biases in big data: Omitted voices on social media. *Social Science Computer Review*, *38*(1), 10–24. https://doi.org/10.1177/0894439318788322

HESA. (2020). *Where do HE students come from?*|HESA. https://www.hesa.ac.uk/data-and-analysis/students/where-from

Jacobs, E., Theile, T., Perrotta, D., Zhao, X., Anastasiadou, A., & Zagheni, E. (2025). Global gender gaps in the international migration of professionals on LinkedIn. *Population and Development Review*, *51*(3), 963–994. https://doi.org/10.1111/padr.70012

Kierans, D. (2020). Who migrates to the UK and why? *Migration Observatory*. https://migrationobservatory.ox.ac.uk/resources/briefings/who-migrates-to-the-uk-and-why/

Leasure, D. R., Kashyap, R., Rampazzo, F., Dooley, C. A., Elbers, B., Bondarenko, M., Verhagen, M., Frey, A., Yan, J., Akimova, E. T., Fatehkia, M., Trigwell, R., Tatem, A. J., Weber, I., & Mills, M. C. (2023). Nowcasting daily population displacement in Ukraine through social media advertising data. *Population and Development Review*, *49*(2), 231–254. https://doi.org/10.1111/padr.12558

Office for National Statistics. (2018a). *Labour force survey user guidance—Office for national statistics.* Office for National Statistics.

Office for National Statistics. (2018b). *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

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

Office for National Statistics. (2019b). *Understanding different migration data sources—Office for National Statistics*. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/understandingdifferentmigrationdatasources/juneprogressreport

Office for National Statistics. (2019c). *Understanding different migration data sources: August Progress Report—Office for National Statistics*. https://www.ons.gov.uk/releases/understandingdifferentmigrationdatasourcesaugustprogressreport

Office for National Statistics. (2020). *Migration statistics quarterly report—Office for National Statistics*. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/bulletins/migrationstatisticsquarterlyreport/may2020

Office for National Statistics. (2021). *Methods for measuring international migration using RAPID administrative data—Office for National Statistics*. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/methodologies/methodsformeasuringinternationalmigrationusingrapidadministrativedata

Office for National Statistics. (2022). *International migration, England and Wales—Office for National Statistics*. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/bulletins/internationalmigrationenglandandwales/census2021

Office for National Statistics. (2025). *Provisional long-term international migration estimates: Technical user guide*. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/methodologies/provisionallongterminternationalmigrationestimatestechnicaluserguide

Palotti, J., Adler, N., Morales-Guzman, A., Villaveces, J., Sekara, V., Herranz, M. G., Al-Asad, M., & Weber, I. (2020). Monitoring of the Venezuelan exodus through Facebook's advertising platform. *PLoS One*, *15*(2), Article e0229175. https://doi.org/10.1371/journal.pone.0229175

Plummer, M., Stukalov, A., Denwood, M., & Plummer, M. M. (2016). *'Package "Rjags"'*. R Foundation for Statistical Computing. https://cran.r-project.org/web/packages/rjags/index.html

Poulain, M., Perrin, N., & Singleton, A. (2006). *THESIM: Towards harmonised European statistics on international migration*. Presses universitaires de Louvain.

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. https://doi.org/10.1215/00703370-9578562

Rampazzo, F., Bijak, J., Vitali, A., Weber, I., & Zagheni, E. (2025). Assessing timely migration trends through digital traces: A case study of the UK before Brexit. *International Migration Review*, *59*(1), 119–140. https://doi.org/10.1177/01979183241247009

Rampazzo, F., & Weber, I. (2020). Facebook advertising data in Africa. In *Migration in west and north Africa and across the Mediterranean: Trends, risks, development and governance* (pp. 32–38). International Organization for Migration. https://publications.iom.int/books/migration-west-and-north-africa-and-across-mediterranean

Raymer, J., Wiśniowski, A., Forster, J. J., Smith, P. W. F., & Bijak, J. (2013). Integrated modeling of European migration. *Journal of the American Statistical Association*, *108*(503), 801–819. https://doi.org/10.1080/01621459.2013.789435

Rogers, A., & Castro, L. J. (1981). *Model migration schedules*. Monograph. http://pure.iiasa.ac.at/id/eprint/1543/

Rogers, A., & Watkins, J. (1987). General versus elderly interstate migration and population redistribution in the United States. *Research on Aging*, *9*(4), 483–529. https://doi.org/10.1177/0164027587094002

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. In *Papers de Demografia* (*Vol. 463*, pp. 1–69). Centre d'Estudis Demogràfics.

Sanlitürk, E., Samin, A., Zagheni, E., & Billari, F. C. (2024). Homecoming after Brexit: evidence on academic migration from bibliometric data. *Demography*, *61*(6), 1897–1921. https://doi.org/10.1215/00703370-11679804

State, B., Rodriguez, M., Helbing, D., & Zagheni, E. (2014). Migration of professionals to the U.S. In L. M. Aiello, & D. McFarland (Eds.), *Social informatics* (*Vol. 8851*, pp. 531–543). Springer International Publishing. https://doi.org/10.1007/978-3-319-13734-6_37

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. SEC.report*. https://sec.report/Document/0001326801-20-000013/fb-12312019x10k.htm

Wilson, T. (2010). Model migration schedules incorporating student migration peaks. *Demographic Research*, *23*(July), 191–222. https://doi.org/10.4054/DemRes.2010.23.8

Wiśniowski, A., Forster, J. J., Smith, P. W. F., Bijak, J., & 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. https://doi.org/10.1111/rssa.12177

Zagheni, E., Weber, I., & Gummadi, K. (2017). Leveraging Facebook's advertising platform to monitor stocks of migrants. *Population and Development Review*, *43*(4), 721–734. https://doi.org/10.1111/padr.12102