Georgios Aristotelous

Database: Migration estimates

Deliverable 6.4

QuantMig has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 870299. Website: **www.quantmig.eu**

History of changes

Version	Date	Changes
1.0	23 June 2022	Issued for Consortium Review
1.1	29 June 2022	First version submitted as official deliverable to the EC
1.2	7 August 2023	Public version submitted to the open data repository (Zenodo)
1.3	20 November 2023	Amended version submitted to the EC after the final Review Meeting

Suggested citation

Aristotelous G (2022) Database: Migration estimates. QuantMig Project Deliverable D6.4. Southampton: University of Southampton.

Dissemination level

PU Public

Acknowledgments

Comments made by Peter W. F. Smith and Jakub Bijak are gratefully acknowledged. This document reflects the author's view and the Research Executive Agency of the EC are not responsible for any use that may be made of the information it contains.

Cover photo: iStockphoto.com/Guenter Guni

Database:
Migration estimates*

QuantMig Deliverable 6.4

Georgios Aristotelous †

This document is a 'read me' type of cover note for the files that form the database of migration estimates, presented in the QuantMig Migration Estimates Explorer, available at https://bit.ly/quantmig-estimates. The document is provided 'as is', and is intended to serve only as a general manual for the database, rather than providing extensive documentation of the underlying data and methodology, which can be found elsewhere (Aristotelous et al., 2022). The user manual for the QuantMig Migration Estimates Explorer is included in Potančoková et al. (2023).

The QuantMig Migration Estimates Explorer database consists of three sets of migration estimates. The first set of estimates are disaggregated by origin (O), destination (D) and time (T), a breakdown which we denote as ODT. The second and third sets of estimates are again disaggregated by origin, destination and time, but they are additionally disaggregated by other factors. The second set is further disaggregated by age (A) and sex (S) while the third set by a broad region of birth (B). We respectively denote these breakdowns as ODAST and ODBT. The model that produces the ODT estimates is described in detail in Aristotelous et al. (2022) while the methodology followed to further disaggregate these estimates and produce the ODAST and ODBT estimates is described in Wiśniowski et al. (2016). The sourcing and cleaning of the migration flow data used in the modelling process are described in Aristotelous et al. (2020).

*Deliverable 6.4 - QuantMig is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 870299. All the interpretations and errors are ours alone.

†University of Southampton, UK. Email: G.Aristotelous@soton.ac.uk. Changes in Version 1.3 have been implemented by Jakub Bijak, J.Bijak@soton.ac.uk.

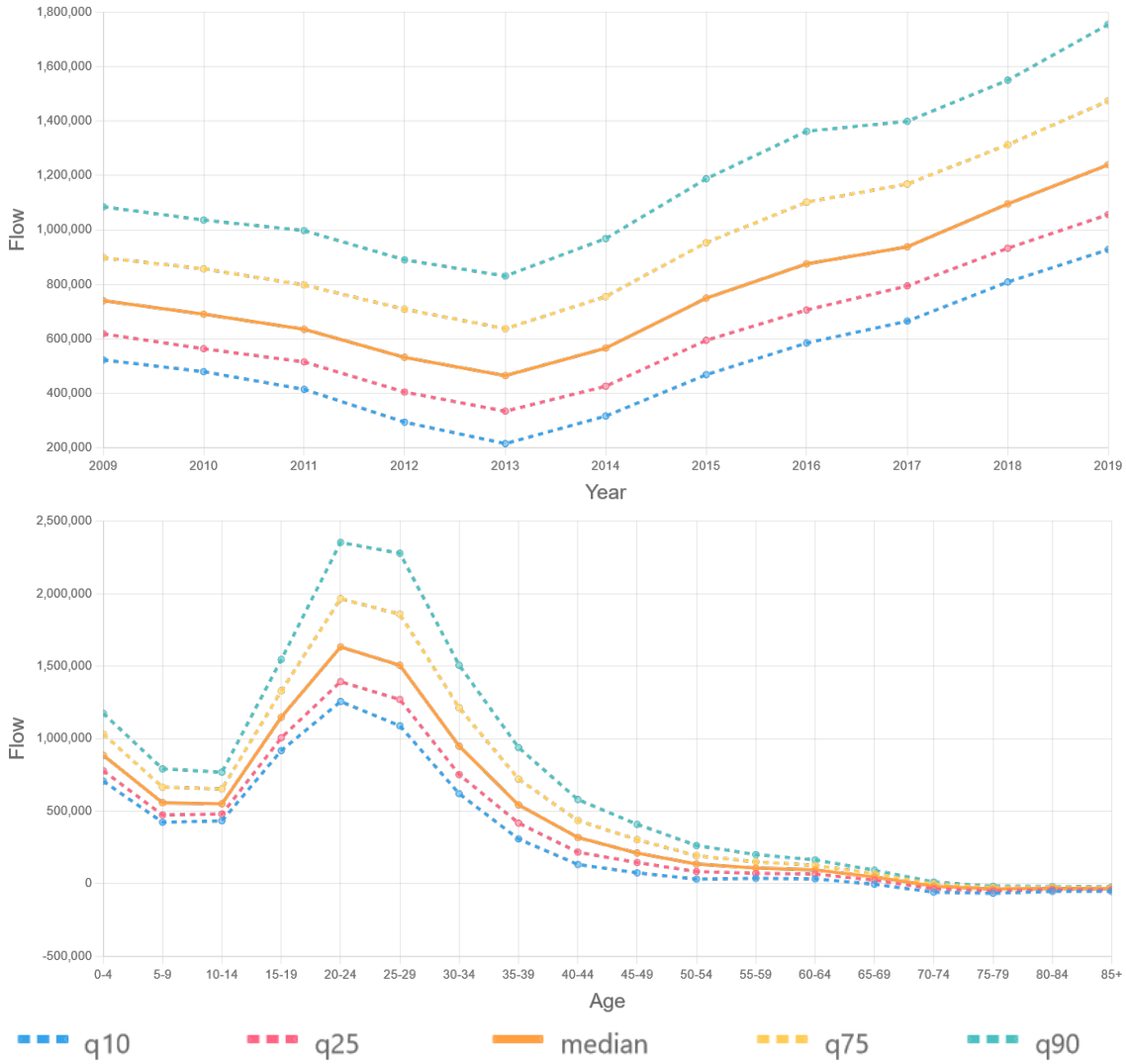
The three sets of estimates, ODT, ODAST and ODBT, are stored in three `csv` files respectively named `flows_ODT_long.csv`, `flows_ODAST_long.csv` and `flows_ODBT_long.csv`. These three files form the migration flow database. All the output files are available from the Zenodo repository, at https://zenodo.org/doi/10.5281/zenodo.8224827 (Smith and Aristotelous 2023b). The files are all provided in long format and have as variables (i.e. columns) the disagreggation factors and the migration estimates, so that each row provides the estimates for each combination of the factor levels. For example, in the `flows_ODT_long.csv` file, the second row provides the migration estimates for the flow where the origin is Belgium, the destination is Austria and the year is 2009. Similarly, in the `flows_ODAST_long.csv` file, the second row provides the migration estimates for the flow where the origin is Belgium, the destination is Austria, the age group is less than 5 years old, the sex group is female and the year is 2009.

In all files we provide three migration estimates, namely the median estimate along with the 2.5% lower and the 97.5% upper percentile estimates, respectively named `flow_50%`, `flow_2.5%` and `flow_97.5%`. These three values are providing a central estimate (`flow_50%`), a lower bound estimate (`flow_2.5%`) and an upper bound estimate (`flow_97.5%`) of a given flow. Furthermore, the interval (`flow_2.5%`, `flow_97.5%`) is a 95% probability interval meaning that, according to the model, the true flow lies in (`flow_2.5%`, `flow_97.5%`) with 95% probability. Illustrative results from the ODT and ODAST tables are presented in Figure 1. The top panel shows an aggregate from the ODT table, with derived net migration for EU-27 countries for 2009–19, presenting the estimates alongside their 50- and 80-percent intervals. The bottom panel shows the age profile of net migration for EU-27 in 2009–19 from the ODAST table, aggregated over time and sex.

The age definition used for describing age groups in the estimates is one of the *age reached* during the calendar year. This corresponds to period-cohort parallelograms on the Lexis diagram (Figure 2). Such a methodological choice was motivated by the better alignment of the period-cohort observational plan with the needs of cohort-component population projections, as well as greater availability of data by age reached in Eurostat. Where only data by completed age (age-period rectangles) were available, they were used

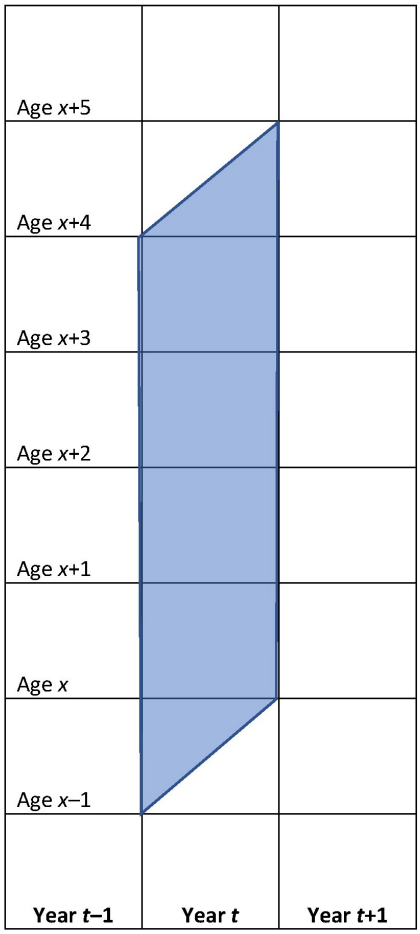
for imputing the age-reached estimates by dissecting the rectangles into Lexis triangles under the usual assumption of uniformity across time and age. Methodological details for the choice of the age definition and the imputation are available in the R code used for downloading and processing Eurostat data, accessible on the Zenodo repository, at https://zenodo.org/doi/10.5281/zenodo.8215060 (Smith and Aristotelous 2023a).

Figure 1: Examples of outputs from the ODT and ODAST tables, generated by the QuantMig Migration Estimates Explorer: Net migration in EU-27 in 2009–19, presented as trends over time (top panel), and aggregate profile by age (bottom panel)



Source: QuantMig Migration Estimates Explorer, https://bit.ly/quantmig-estimates

Figure 2: Visualisation of the adopted age definition (years reached) on a Lexis diagram. Age reported in the period-cohort observational plan: annual data for five-year age groups



Source: Own elaboration

We now provide a description of the levels of the factors for each of the three output files listed above. For completeness and clarity we also summarize this information in Tables 1, 2 and 3. In all files time is in years (named in the files as `year`) and its levels are the years 2009 to 2019, accordingly named `2009`, `2010`, ..., `2019`. In all files the destination factor is named `dest`. Where applicable, age is expressed in years reached during the calendar year, as discussed above.

Table 1: Disaggregation factors and their levels for the ODT estimates (file `flows_ODT_long.csv`)

factor	**levels**
origin and destination	32 EU+ countries, North Macedonia and 8 RW regions
time	years 2009 to 2019

Table 2: Disaggregation factors and their levels for the ODAST estimates (file `flows_ODAST_long.csv`)

factor	**levels**
origin and destination	32 EU+ countries, North Macedonia and 1 RW region
age	18 non-overlapping five-year age groups (age reached)
sex	female and male
time	years 2009 to 2019

Table 3: Disaggregation factors and their levels for the ODBT estimates (file `flows_ODBT_long.csv`)

factor	**levels**
origin and destination	32 EU+ countries and 1 RW region
birth region	broad aggregates only: EU-born and non-EU-born
time	years 2009 to 2019

For the ODT estimates (file `flows_ODT_long.csv`), the levels for origin and destination are the 32 EU+ countries, North Macedonia and 8 rest of the world (RW) regions, in total 41 levels. By EU+ we mean the 28 EU-member countries[1], and the 4 EFTA countries, namely Iceland, Liechtenstein, Norway and Switzerland. The 8 RW regions are Other Europe, North Africa, Sub-Saharan Africa, West Asia, East Asia, South-Southeast Asia, North America and Oceania and Latin America. The countries which comprise each of the 8 RW regions are provided in the file `RW_regions.xlsx` for completeness.

[1]For our considered time period, 2009-2019, the United Kingdom was an EU-member country and is therefore one of the 28 EU-member countries.

For the ODAST estimates (file `flows_ODAST_long.csv`), the levels for origin and destination are the 32 EU+ countries, North Macedonia and a single RW region, in total 34 levels. The difference compared to the ODT estimates is that in the ODAST estimates the 8 RW regions are combined into a single RW region (named `Rest World` in the file). This is because of data availability and more specifically because migration data disaggregated by age and sex are not available for specific RW regions but rather only for RW in general. Regarding age, the levels are 18 non-overlapping 5-year age groups, namely less than five years old, 5-9, 10-14, and so on up until 80-84 and finally the greater than or equal to 85 years old group. In the file these levels are respectively named `Y_LT5`, `Y5-9`, `Y10-14`, ...,`Y80-84`, and `Y_GE85`. As far as sex, the levels are female and male, respectively named `F` and `M` in the file.

Finally, for the ODBT estimates (file `flows_ODBT_long.csv`), the levels for origin and destination are the 32 EU+ countries and a single RW region, a total of 33 levels. Same as for the ODAST estimates, and for the same reasons, the 8 RW regions are combined into a single RW region (named `Rest World` in the file). Note also that North Macedonia is included into the RW, unlike in the ODT and the ODAST files. Regarding birth region (named `birth_region` in the file), the levels are EU-born and non-EU-born, and are respectively named `EU` and `non-EU` in the file. We also note that in all output files flow estimates of same-origin-destination entries are recorded as NA (Not Applicable) since we do not consider within-country migration. We also note that all the estimates are rounded to one decimal point for clarity.

The detailed description of all input and output files used in or generated by the estimation process is beyond the scope of this cover note, but all the specifics are listed in both associated Zenodo repositories (Smith and Aristotelous 2023ab). The individual input data files, most notably from the Eurostat collections (table `migr_imm5prv`), as well as the routines for data processing and use, are explained more thoroughly as comments in the R software code. The output files are available both in the R and `csv` formats in Smith and Aristotelous (2023b), together with a version of this cover note.

The R files in Smith and Aristotelous (2023b) contain full samples of posterior distributions of various model parameters, as well as of the quantities of interest (migration estimates), stored as large lists. The number of samples varies from 5,000 for the ODT table (file size 1.96GB) to 2,500 for the ODBT table (450MB) and 1,000 for the ODAST table (6.58GB). Even though the QuantMig Migration Estimates Explorer only presents a selection of posterior quantiles, the user can generate other summary statistics from the original data files in R – quantiles, moments, etc. The models generate far too many outputs to enable them to be summarised in a tabular format, but advanced users can generate the required quantities themselves following the model description in Aristotelous et al. (2022) and meta-information included in the individual data files and software code. At the same time, the `csv` files contain high-level summaries, as described above. Examples of the information content of the few first records of these files are given in Table 4. Overall, the ODT table contains 18,491 records ($41 \times 41 \times 11$), the ODBT table – 23,958 records ($33 \times 33 \times 2 \times 11$), and the ODAST table – 457,776 records ($34 \times 34 \times 18 \times 2 \times 11$).

Table 4: Examples of the estimate outputs available in the `csv` files

Orig	**Dest**	**Year**	**Age**	**Sex**	**Birth reg.**	**2.5%**	**median**	**97.5%**
ODT								
AT	AT	2009	-	-	-	NA	NA	NA
BE	AT	2009	-	-	-	221.9	344.3	528.8
BG	AT	2009	-	-	-	1695.2	2611.9	3975
CY	AT	2009	-	-	-	22.4	36.5	59.4
ODBT								
AT	AT	2009	-	-	non-EU	NA	NA	NA
BE	AT	2009	-	-	non-EU	19.3	51.1	120.2
BG	AT	2009	-	-	non-EU	37.4	106	279.6
CY	AT	2009	-	-	non-EU	1.2	3.3	8.3
ODAST								
AT	AT	2009	< 5	F	-	NA	NA	NA
BE	AT	2009	< 5	F	-	6.3	11.9	22.4
BG	AT	2009	< 5	F	-	28.6	52.2	98.7
CY	AT	2009	< 5	F	-	0.6	1.4	2.9

Source: Smith and Aristotelous (2023b). Note: the column and row headers are edited for brevity.

References

Aristotelous, G., Smith, P. W. F., and Bijak, J. (2020). Database: Flows, stocks and quality for modelling. QuantMig Deliverable D5.3, University of Southampton, Southampton.

Aristotelous, G., Smith, P. W. F., and Bijak, J. (2022). Technical report: Estimation methodology. QuantMig Deliverable D6.3, University of Southampton, Southampton.

Potančoková, M., Sadler, J., Bijak, J., González-Leonardo, M., and Soto-Nishimura, A. (2023). Documentation of QuantMig web tools for data and simulations. QuantMig Deliverable D10.4, IIASA, Laxenburg.

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.

Zenodo deposits

Smith, P. W. F., and Aristotelous, G. (2023a). R script to download and pre-process Eurostat migration data. Version 1, 4 August 2023. Zenodo. https://doi.org/10.5281/zenodo.8215060

Smith, P. W. F., and Aristotelous, G. (2023b). MCMC simulations from the posterior distributions of European migration flows [Data set]. Version 1, 8 August 2023. Zenodo. https://doi.org/10.5281/zenodo.8224827