



Jakub Bijak

European migration scenarios with probabilistic uncertainty assessment

Deliverable 9.4 - Data Description



QuantMig has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 870299.

History of changes

Version	Date	Changes
1.0	21 May 2023	Final draft (embargoed) for internal circulation only
1.1	30 June 2023	Public version issued for open data depositing

Suggested citation

Bijak J (2023) European migration scenarios with probabilistic uncertainty assessment. QuantMig Project Deliverable D9.4 Data Description. Southampton: University of Southampton.

Dissemination level

PU Public

Key words

Civil contingencies, Extreme value theory, Heavy-tailed distributions, Migration scenarios, Quantiles, Rare events, Uncertainty

Acknowledgments

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 870299 QuantMig: Quantifying Migration Scenarios for Better Policy. All the remaining errors and inaccuracies are ours. This document reflects the authors' view and the Research Executive Agency of the European Commission are not responsible for any use that may be made of the information it contains.

Cover photo: [iStockphoto.com/Guenter Guni](https://iStockphoto.com/GuenterGuni)

European migration scenarios with probabilistic uncertainty assessment: Data Description

Preamble

The current data deposit (DOI: [10.5281/zenodo.7954150](https://doi.org/10.5281/zenodo.7954150)) contains source data, R code and output used in the analysis reported in QuantMig project deliverable 9.4 “European migration scenarios with probabilistic uncertainty assessment” (Bijak 2023). This cover note summarises the content of individual input and output files, and the R code (**D9-4_migration_extreme_quantiles.R**) used for the analysis. For replicating the results, the required source files (**D9-4_de_correction.csv** and **QuantMig_Migration_Estimates_2009-19.RData**) need to be stored in the same directory, which needs to be specified in the R code below. The source data come from a dedicated set of migration estimates within, into and out of Europe produced as a part of the QuantMig output (Aristotelous et al. 2023), which will be openly available both on www.quantmig.eu, as well as on Zenodo.

List of files

D9-4_de_correction.csv

An auxiliary **input** file used by the R code (**D9-4_migration_extreme_quantiles.R**), containing the manual additive correction for the immigration data to Germany, 2015–2019 (see Bijak 2023).

D9-4_migration_extreme_quantiles.csv

The key **output** file, produced by the R code (**D9-4_migration_extreme_quantiles.R**), containing a selection of estimated quantiles for immigration from eight rest of the world regions (East Asia, Latin America, North Africa, North America and Oceania, Other Europe, South-Southeast Asia, Sub-Saharan Africa, and West Asia) to the aggregate of 32 EU+ countries (European Union, EFTA, and the United Kingdom). All results are based on the median flows – medians of the posterior estimates calculated on the basis of the MCMC samples. The output includes: means and medians of the q_{90} and q_{98} quantiles from the fitted exponential, generalised extreme value (GEV), log-normal and Pareto distributions, as well as the p-values from the Shapiro-Wilk test, used as a goodness-of-fit measure (Millard and Kowarik 2022). In Bijak (2023), p-values are also provided for distributions fitted to the mean flows, as a sensitivity check. **Please note:** the output version included in the package is the one reported in Bijak (2023) – due to a small sample (short series), the estimates for the GEV distribution may be numerically unstable, which is one of the reason why the GEV quantiles were not used for scenario-setting.

D9-4_migration_extreme_quantiles.R

The **R code** for the analysis, was developed and tested using R version 4.2.2 (R Core Team 2022). The code requires the R packages `EnvStats` (Millard and Kowarik 2022), for estimations of the heavy-tailed distributions, and `stringr` (Wickham 2022) for presentational purposes. The code uses the full set of QuantMig estimates (from **QuantMig_Migration_Estimates_2009-19.RData**) fine-tuned by applying the additive Germany correction (**D9-4_de_correction.csv**). The code is commented, but has not been optimised, as it has been used both for estimation and for preparing graphical output for the report, and is therefore provided ‘as is’.

QuantMig_Migration_Estimates_2009-19.RData

An extract from the original QuantMig dataset (Aristotelous et al. 2022) for individual migration flows between 32 EU+ countries (European Union, EFTA, and the United Kingdom), and eight rest of the world regions (see above), by origin, destination, and year, for 11 years (2009–19). For each of the flows, 5,000 MCMC samples are included in the main data list `output.MCMC`. The whole extract is only loaded for the purpose of calculating the median flows from eight rest-of-the-world origins to the whole EU+ area, for the 11-year period, aggregating North Macedonia (region 33) with the Other Europe (region 34), and after applying the additive Germany correction (see above) is then removed from the active R environment to save memory.

D9-4_migration_extreme_quantiles.RData

A full output dataset in R format, containing information detailed in Table 1 below.

Table 1. Elements included in the D9-4_migration_extreme_quantiles.RData output file

Element	Type	Description
migdata	Numeric [8x11x5000]	Full set of flow estimates of immigration from eight regions of the world, 11 years (2009–2019), and 5000 MCMC samples
correction	Numeric [8x11]	Correction of the Germany (and hence aggregate) data by region of the world and year, from D9-4_de_correction.csv
data_median	Numeric [8x11]	Median flow estimates of immigration from each of the eight regions of the world across 11 years (2009–2019)
data_05, data_95	Numeric [8x11]	5th and 95th percentiles of the immigration flow estimates, used for depicting error bars on QQ plots
flab_[dist], flab_q[dist]	Text [8]	File labels for graphs of QQ plots and quantile from each of the distributions: [dist] = {"exp", "gevd", "lnorm", "pareto"}
Labels	Text [8]	Labels with the names of the eight rest of the world regions. Note a different (proximity) order to the one in Bijak (2023)
q_[dist]	Numeric [2x8x5000]	Estimated quantiles from the fitted distributions [dist] (as above): $q_{0.9}$ (first index = 1) and $q_{0.98}$ (first index = 2), for each of the eight origin regions, for each of the 5000 MCMC samples
x_[dist]	Numeric [8x11]	Values of the quantiles from the fitted distributions [dist] corresponding to the median estimates for each of the eight world regions and 11 years (used for the QQ plots)
mig_mean	Numeric [8x8]	Means of the quantiles $q_{0.9}$ and $q_{0.98}$ for four distributions, indexed from 1 to 8: 1 = $q_{0.9}$ exponential, 2 = $q_{0.98}$ exponential, 3 = $q_{0.9}$ gevd, etc., for the eight origin regions, based on q_[dist]
mig_quantiles	Numeric [8x8]	Medians of the quantiles $q_{0.9}$ and $q_{0.98}$ for four distributions and eight regions of the world, based on q_[dist] , indexed as above
sw_pval, sw_test	Numeric [4x8]	P-values and test statistics for the Shapiro-Wilk test used for goodness of fit, for four distributions and eight world regions
i, k, t	Numeric	Indices for region of origin (i), MCMC sample (k), and year (t)
domain, newdomain	Numeric [N]	Domains for plotting, with N=5000 (Figs. 1 and 2 in Bijak 2023) or N=1600 (Fig. 3) points
collab, rowlab	Text	Labels for columns and rows in the output csv file
parloc, parshape	Numeric [8]	Location and shape parameters of the Pareto distributions estimated for immigration from the eight regions of the world

References

- Aristotelous G, Smith PWF and Bijak J (2022) Technical report: Estimation methodology. QuantMig Project Deliverable D6.3. Southampton: University of Southampton.
- Bijak J (2023) European migration scenarios with probabilistic uncertainty assessment. QuantMig Project Deliverable 9.4. Southampton: University of Southampton.
- Coles S (2001) *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics. London: Springer.
- Millard SP and Kowarik A (2022) EnvStats: Package for Environmental Statistics, Including US EPA Guidance. Version 2.7.0 (7 March 2022). Available from CRAN: <https://CRAN.R-project.org/package=EnvStats> (accessed on 19 May 2023).
- R Core Team (2022) R: A language and environment for statistical computing. Version 4.2.2 “Innocent and Trusting”. Vienna: R Foundation for Statistical Computing. Via <https://www.R-project.org>
- Wickham H (2022) stringr: Simple, Consistent Wrappers for Common String Operations. Version 1.5.0 (2 December 2022). Available from CRAN: <https://CRAN.R-project.org/package=stringr> (accessed on 19 May 2023).

■