Jakub Bijak

European migration scenarios with probabilistic uncertainty assessment

Deliverable 9.4

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	30 May 2023	Issued for Consortium Review
1.1	30 June 2023	First version submitted as official deliverable to the EC

Suggested citation

Bijak J (2023) European migration scenarios with probabilistic uncertainty assessment. QuantMig Project Deliverable 9.4. 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 document reflects the author's views and the Research Executive Agency of the European Commission are not responsible for any use that may be made of the information it contains. I am very grateful for useful comments and advice from Jason Hilton, Nico Keilman, Michaela Potančoková, Ann Singleton and Peter WF Smith. Needless to say, any errors or interpretations are exclusively mine.

Cover photo: iStockphoto.com/Guenter Guni

Abstract

There are important gaps in the current methodology and practice of constructing the future migration scenarios. In particular, even though a lot of attention in scenario-setting is paid to the underlying narratives and drivers, operationalisation of the link between these drivers and migration scenarios is very weak and highly uncertain. The interactions between the drivers as such, and within the broader driver environments, are largely ignored. This problem is strengthened by the theoretical fragmentation of migration studies and a lack of a general high-level framework that could adequately explain a broad enough spectrum of migration processes. In this report, for methodological as well as practical reasons, we suggest a shift in perspective, inspired by approaches used in civil contingency planning. Instead of building the scenarios from the presumed driver trajectories, we use the harmonised information about origin-destination-specific flows to derive levels of migration corresponding to certain frequencies of occurrence, such as once-in-a-decade, or twice-in-a-century. In probabilistic terms, these quantities can be approximated by quantiles 0.9 and 0.98 from a suitably heavy-tailed probability distribution fitted to the harmonised migration estimates. The proposed approach relies on applying a simplified version of the statistical theory of modelling extreme values to approximate the magnitude of the rare events of varying frequencies. The approach is illustrated with examples of immigration to Europe from the eight other major world regions, for which we estimate the magnitudes of once-in-a-decade and twice-in-a-century flows. Such estimates, even if approximate, can then serve as a basis for providing a different form of an uncertainty assessment for scenarios and facilitate communicating them to their users.

Table of Contents

1. Introduction

The literature on methods and approaches for setting migration scenarios has been steadily increasing since the beginning of the 20th century, mirroring the increasing public and policy interest in migration. A recent review (Boissonneault et al. 2020) proposed a simple typology of existing work on scenarios along two dimensions: the purpose of the scenario construction and its focus. Three main types of purpose include predictive, exploratory, or normative studies, the last one related to setting or monitoring of migration-related targets. At the same time, the focus can be either migration itself, or broader socio-economic processes. The review identified the mostly quantitative nature of existing work, typically based on past and current data, as well as expert opinion, although with qualitative, narrative-based approaches also gaining prominence in the recent years.

Across the scenario literature, the quantification of assumptions and narratives has been found challenging (Boissonneault et al. 2020). Even though a lot of attention in scenario-setting is paid to the underlying narratives and drivers, their operationalisation, and formalising the links between drivers and migration, is typically rather tenuous and highly uncertain. Of course, drivers – economic, political, conflict-related, environmental, and many other – as well as their complex environments and interactions (see e.g. Czaika et al. 2021) are difficult to include in scenarios. At the same time, from the point of view of the uses of scenarios in policy and practice, this may not matter that much, as long as scenarios deliver their promise in terms of horizon scanning and aiding preparedness.

At the same time, the role of uncertainty in scenarios, even though often acknowledged as important, is rarely formally quantified, although with some important exceptions, such as Acostamadiedo et al. (2020), who used expert opinion for assessing the relative probabilities of the four proposed scenario pathways. The lack of focus on the sources of uncertainty is an important theoretical and conceptual limitation of most of the existing approaches. In addition, a majority of the scenarios concentrate on the total inflow or overall net migration – contemporary political priorities – and do not look into scenarios for individual types of flows or specific routes or corridors, which can widely differ in terms of their levels of uncertainty and unpredictability (see e.g. de Beer 2008 or Bijak et al. 2019). It is worth noting here that while the term 'uncertainty' relates to all unknown aspects of future migration, 'unpredictability' relates more specifically to those aspects that are unknowable in advance – the intrinsic *aleatory* uncertainty – and cannot be therefore predicted (Bijak and Czaika 2020).

Against this backdrop, the aim of this report is to propose an alternative approach and to illustrate it on the example of scenarios of migration flows into Europe from different regions of the world. Taking inspiration from planning for civil contingencies, so explicitly adopting a preparedness perspective, the focus of the proposed method is on the frequency and magnitude of various possible migration events (e.g. twice-in-a-century or once-in-a-decade). The proposed method is based on statistical approaches to modelling rare events, including the extreme value theory (Coles 2001). It has been designed to be simple in terms of design and construction, and yet applicable to a broad range of migration contexts, not limited to immigration, but possibly including all other types of flows, planning for which may be of interest for policy makers and public planners.

The remainder of this report is structured as follows. Following on the discussion in this Introduction, Section 2 makes the theoretical and conceptual case for designing scenarios in a probabilistic fashion. In Section 3, a methodological framework for setting migration scenarios is proposed, framed in terms of the frequency of rare events, estimated with statistical techniques using heavy-tailed distributions. In Section 4, this approach is illustrated with examples related to immigration into Europe in the 2020s, with scenarios created on the basis of a novel, harmonised dataset for European migration flows in 2009–19 (Aristotelous et al. 2022). Finally, the concluding Section 5 contains a discussion of the key findings and limitations of the proposed approach.

2. Case for probabilistic scenarios

As mentioned in Section 1, existing examples of scenarios utilising probabilistic concepts are rare, and yet the use of stochastic approaches in scenario setting – or at least acknowledging the migration uncertainty in some probabilistic form – can have broader appeal to the users. Even when the scenarios are based on specific assumptions or narratives, information about the relative probabilities of different pathways versus one another, or including expert-based uncertainty around the envisaged migration trajectories, as done by Acostamadiedo et al. (2020), provides the users with important value added, and cautions against being overconfident in any particular pathway. At the same time, the probabilities of a small number of individual scenarios (typically, four) are by necessity relative to the pre-defined universe of all these possibilities, rather than being general. As the number of possible migration futures is (nearly) infinite, the probability of each of them occurring is (almost) zero.

In other work, we looked at producing probabilistic scenarios by using complex models, for example macroeconomic dynamic stochastic general equilibrium (DSGE) models (Barker and Bijak 2021). Such approaches allow for stochasticity by construction, and can also include other probabilistic elements – typically, through assessing responses to 'shocks' to different variables included in the models. These shocks can be operationalised as one- or two-standard deviations departures from past trends of specific variables, such as migration or its drivers. In Barker and Bijak (2021), we studied technological shocks related to job automations. The effects of such shocks can be examined with impulse-response functions, showing how their effects propagate through the entire system, and affect a range of other variables (*idem*). This approach is promising, but is at the same time context-dependent and resource intensive: each individual situation requires constructing and calibrating a dedicated complex model. Still, with the exception of very specific case studies, this approach is very labour-intensive and not really well suited for modelling whole multi-country migration systems interlinked by bilateral migration flows (Potančoková et al. 2023).

At the same time, the modelling of rare or extreme events for complex systems – such as migration processes – which have a tendency to generate unpredictable and surprising 'Black Swan' events (Taleb 2007) – remains an important gap in setting migration scenarios. Trying to anticipate such events – not in precise terms, such as exact timing and magnitude, but at least in terms of the possible order of magnitude relative to the frequency of occurrence, can become a key component of migration preparedness. As flagged in Bijak and Czaika (2020, p17), one – as of yet unexplored – promising pathway of setting migration scenarios is based on the statistical theory of extreme values (Coles 2001). In this context, and because scenarios are typically prepared for longer terms than plausible horizons of predictability, the dominant part of uncertainty is aleatory – the unknowable intrinsic randomness. In the light of the taxonomy of Boissonneault et al. (2020), this shifts the purpose of scenarios, from predictive to exploratory, but with a normative aim (greater preparedness) in mind.

To make use of the option to build exploratory scenarios related to frequencies of various migration events would require a change in perspective both on making and using migration scenarios. Instead of constructing them based on assumptions, narratives, or presumed driver trajectories, the link with which is bound to remain highly uncertain, we suggest a pragmatic alternative – an approximate data-driven approach. Rather than looking for explanatory or causal links with other variables, our approach looks at the processes themselves, with a similar rationale as for using atheoretical models (such as autoregressive ones) for migration forecasting – theories and explanations are too uncertain to offer a strong predictive capability (Bijak 2010).

Even though examples of models using additional, theory-based explanatory variables to predict migration exist, also in the context of official statistics (e.g. Cappelen et al. 2015), they typically do not address the problem of predicting the predictors, and coherently propagating their uncertainty to migration forecasts. Those models that do that, such as vector autoregressive approaches (Bijak 2010; Barker and Bijak 2021), end up with very high predictive uncertainty very quickly – too high to be useful in most practical applications and offer meaningful decision support.

The main premise of the approach proposed in this report is inspired by civil contingency planning for such events as earthquakes, floods or other rare events, which are unpredictable with respect to their exact timing and magnitude of occurrence, but can have high impact when they happen. In other words, we know rare events happen, but we do not know when the next one will occur, or how large it will be. Despite the link and inspiration by the civil contingencies planning, there is an important difference of interpretation when it comes to migration. In particular, for civil contingencies, the events requiring preparedness are typically destructive and have mainly negative impact, being something that mainly requires protection against.

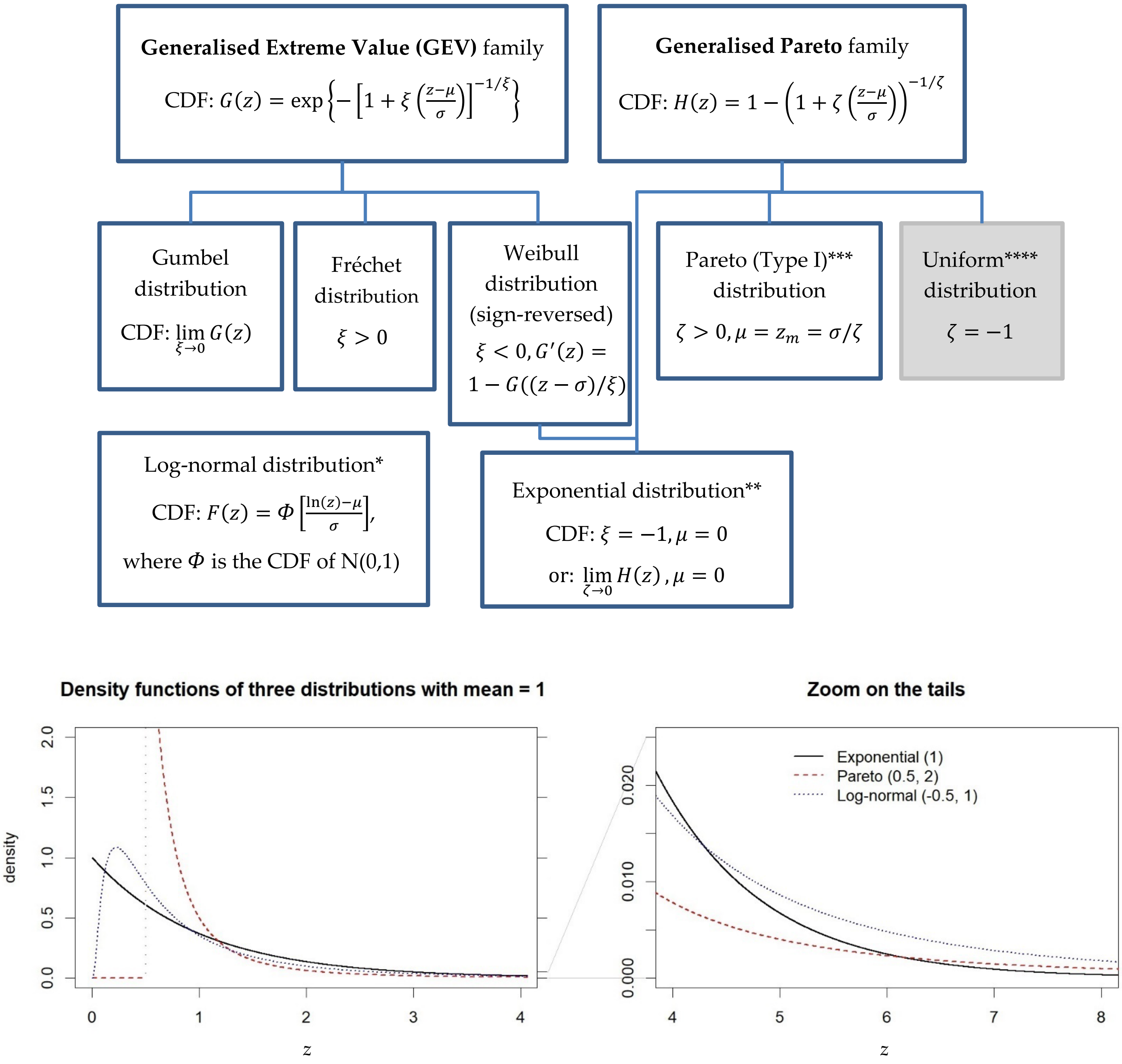
In contrast, in this work we consider migration as a neutral phenomenon – a normal fact of life – that, in certain circumstances, such as wars, environmental disasters, or profound shifts in political or socio-economic circumstances, can exhibit large-scale and high-intensity characteristics. In such instances, large and rapidly-changing migration processes can be challenging to accommodate, especially in the short term. From this perspective, we understand preparedness not as a way of protecting host societies against migration, but rather as a reflective and proactive element of the process of decision making related to migration. To be effective, preparedness needs to be coupled with the commitment of resources for migration to be better managed, both for the sake of migrants and the receiving societies. In addition, it has to be stressed that the proposed approach is not limited to immigration – its methodological premise is sufficiently flexible to cover emigration, onward mobility, and other movement types as well. The details of the proposed method are outlined in Section 3.

3. Preparing for rare events: Methodological Framework

In the proposed data-driven methodology of scenario setting, we rely on applying a simplified version of the statistical theory of modelling extreme values (Coles 2001) to approximate the magnitude of the rare migration events of varying frequencies. For illustrative purposes, throughout this report we examine two such frequencies – once-in-a-decade events, corresponding to quantile $q_{0.9}$ from an underlying probability distribution for yearly migration counts[1], and twice-in-the-century events, corresponding to quantile $q_{0.98}$. Due to the presence of many unforeseen events, statistical distributions describing migration are likely to be heavy tailed (Bijak 2010), which we define as declining more slowly than exponential distributions. In this light, heavy-tailed distributions, fit to existing observations, are a natural choice for modelling rare-event quantiles. We fit models to 11 years of migration estimates (2009–2019) in order to make full use of the information available from the harmonisation of European migration data (Aristotelous et al. 2022), and not to overly rely on any single data point, even though these are also described by their own probability distributions.

In this report, we consider distributions from the Generalised Extreme Value (GEV) and Generalised Pareto distribution families (Coles 2001), as well as a log-normal distribution, commonly used for modelling positively skewed phenomena that can only assume positive values, and the exponential distribution, used as a benchmark. The formulae for the cumulative distribution functions, as well as the relationships between the individual distributions and their families, are presented in Figure 1. In addition to the relationships between distribution families, Figure 1 also shows two stylised examples of heavy-tailed distributions, log-normal and (Type I) Pareto, both having the same mean (one), as well as the corresponding exponential distribution. In the zoomed-in bottom-right panel, the relationship between the tails of various distributions becomes apparent. Subsequently, in Figure 2, the two illustrative quantiles of interest from these three distributions – $q_{0.9}$ and $q_{0.98}$ – are presented.

[1] For a given distribution, the probability of exceeding quantile q_a is 1–a, so we can expect the events to exceed $q_{0.9}$ about 10% of the time (once-in-a-decade), and $q_{0.98}$ – 2% of the time (twice-in-a-century).



Generalised Extreme Value (GEV) family

CDF: $G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}$

Generalised Pareto family

CDF: $H(z) = 1-\left(1+\zeta\left(\frac{z-\mu}{\sigma}\right)\right)^{-1/\zeta}$

Gumbel distribution

CDF: $\lim_{\xi\to 0} G(z)$

Fréchet distribution

$\xi > 0$

Weibull distribution (sign-reversed)

$\xi < 0, G'(z) = 1 - G((z-\sigma)/\xi)$

Pareto (Type I)*** distribution

$\zeta > 0, \mu = z_m = \sigma/\zeta$

Uniform**** distribution

$\zeta = -1$

Log-normal distribution*

CDF: $F(z) = \Phi\left[\frac{\ln(z)-\mu}{\sigma}\right]$,

where Φ is the CDF of N(0,1)

Exponential distribution**

CDF: $\xi = -1, \mu = 0$

or: $\lim_{\zeta\to 0} H(z)\,, \mu = 0$

Figure 1. Selected heavy-tailed distributions and their cumulative distribution functions (CDF): Fragment of taxonomy (top) and examples of the probability distribution functions (bottom panel).

Notes: * The log-normal distribution belongs to the same five-parameter generalised beta and exponential beta families as many other distributions shown here (McDonald and Xu 1995), with the full taxonomy omitted for transparency. *** Closed form of the CDF for the exponential distribution with parameter (expected value) σ is $F(z) = 1 - e^{-z/\sigma}$. **** CDF for the Pareto (Type I) distribution with parameters (z_m, ζ), defined for $z > z_m$ is $F(z) = 1 - (z_m/z)^{\zeta}$, which for $\zeta > 1$ has expected value $(\zeta\, z_m)/(\zeta - 1)$. For simplicity, other possible Pareto distributions are not considered here (e.g. Type II / Lomax, with $F(z) = 1 - (z_m/(z + z_m))^{\zeta}$). **** The uniform distribution is not considered in this work, as its properties – constant density across of the domain – do not fit the migration context.

Source: Taxonomy – Coles (2001: pp 47, 75, 77); Examples – own elaboration in R (R Core Team 2022).

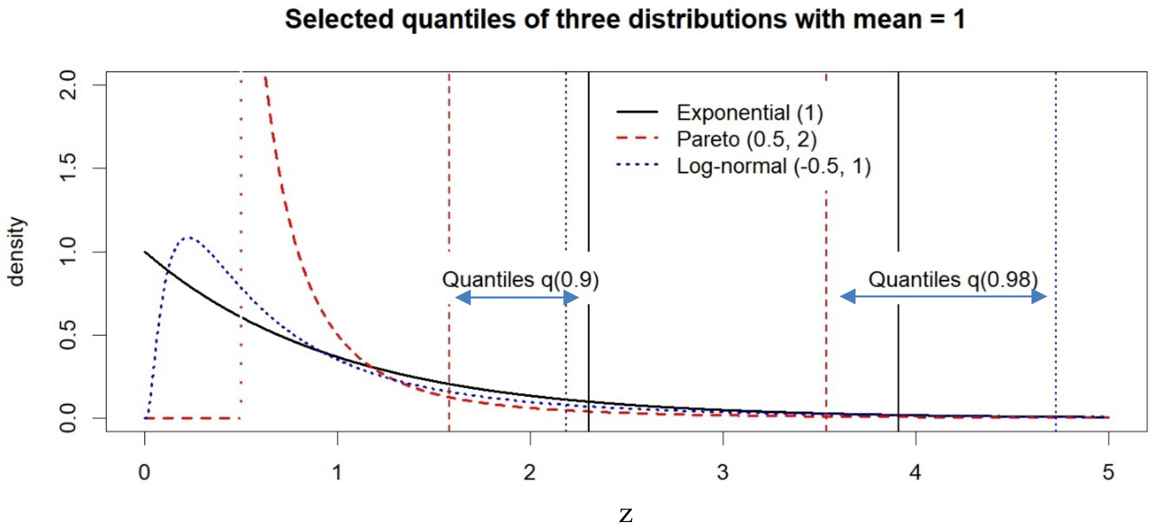


Figure 2. Stylised examples of quantiles $q_{0.9}$ and $q_{0.98}$ from the exponential distribution and two heavy-tailed distributions: log-normal and Pareto. Note that the tails of the exponential distribution are below those of the log-normal and Pareto distribution only for suitably large values of z (see the zoomed part of Figure 1). *Source: Own elaboration in R (R Core Team 2022)*

With respect to GEV, in this work we allow the most flexible form of the distribution, which can yield the specific cases (Gumbel, Fréchet or Weibull, see Figure 1) depending on the precise estimates of the model parameters. At the same time, the GEV distributions are designed for modelling the maximum values of observations within certain period, such as maximum annual levels of the process in question (in the case of civil contingencies, these could be for example water levels, rainfall, seismic activity, and so on). In our case, we will use it just for modelling annual data, given the availability of harmonised migration flows for yearly frequency, but noting that this does not utilise the full potential of the GEV distribution family. We come back to this issue in the discussion in Section 5, charting some possibilities for utilising the GEV models more fully in specific migration and data contexts.

As the proposed approach is based on using uncertain migration estimates, which themselves are described through probability distributions, one important consideration is, what the heavy-tailed distributions should be fitted to. In our illustrative example, we use the *medians* of the distributions for individual migration flows, for the sake of consistency with reporting elsewhere in the QuantMig project, especially in scenarios (e.g. Potančoková et al. 2023). At the same time, other quantities, such as means or higher quantiles are possible choices, too – they would additionally allow for including measurement error explicitly in the assessment of the future uncertainty for scenario setting. In the examples shown in the next section, we do not include this additional error, to focus just on the effect of the uncertainty of migration processes, but in real life applications, the selection of an appropriate underlying measure can also be a matter of deliberate choice, depending on the requirements and resource constraints of particular users. In our examples, we fit the distributions to the median estimates by using the maximum likelihood method implemented in a suite of functions of the R package EnvStats (Millard 2013; Millard and Kowarik 2022).

In addition, in a probabilistic framework, such as the one employed in this study, quantiles are also represented by probability distributions. Hence, in the estimation process, full distributions of both $q_{0.9}$ and $q_{0.98}$ are obtained. This gives rise to another question: which summary measure of these distributions to choose for scenario setting? In our example, we opt for the *mean* values of the quantiles, given that they are meant to serve as basis for scenarios, are additive, and are typically somewhat higher than the corresponding medians, making the scenarios a bit more conservative. In Appendix B, we present both mean and median values of the quantiles $q_{0.9}$ and $q_{0.98}$.

Of course other choices of measures are also possible. Such measures, for example including higher quantiles from the $q_{0.9}$ and $q_{0.98}$ distributions, have different implications in terms of the corresponding attitudes to risk, and imply various levels of preparedness. Formally, alternative choices of summary measures correspond to different *loss functions* describing the possible costs of overpredicting versus underpredicting migration – for example, using the top quartile of the $q_{0.98}$ distribution would imply that underpredicting of the twice-in-a-century migration level would be three times more costly than overpredicting (see Bijak 2010). Alternative measures could also, in principle, account for higher-order uncertainty – the uncertainty about uncertainty. If the precautionary principle – planning for a broad range of uncertain possibilities – was to be followed, such an approach would imply even higher levels of migration in the derived scenarios.

4. Illustration: European migration scenarios

In this section, we illustrate the approach introduced before with examples of high-immigration events to Europe from the eight other regions of the world, for which we estimate the magnitudes of once-in-a-decade and twice-in-a-century flows. Even though the focus of this exercise is on immigration, for the purpose of feeding into migration scenario setting in Potančoková et al. (2023), the process is generic and can be equally applied to emigration, as well as to intra-EU or other flows. From that point of view, the proposed methodology is agnostic with respect to the data used for approximating the relevant quantiles from the fitted probability distributions.

In the estimation process, four distributions have been fitted to the medians and means of the QuantMig estimates of migration into the whole EU+ system in 2009–2019 (Aristotelous et al. 2022): three heavy-tailed distributions – GEV, log-normal and Pareto – as well as the exponential distribution, used as a benchmark for comparison. A similar exercise could be of course carried out for each country separately, as well as for different migration flows – emigration as well as immigration. In our case, the choice of inflows into the whole EU+ was driven by the input requirements of the scenarios of impacts of migration on population and labour force resources in Europe (Potančoková et al. 2023).

In addition, for 2015–19, given the unavailability of German data in Eurostat (Aristotelous et al. 2022), coupled with known high levels of immigration, especially from outside Europe, reported in national sources (the Federal Statistical Office, DESTATIS, www.destatis.de), an additional additive correction has been manually applied. Thus, thee estimates for immigration from the rest of the world regions into Germany were adjusted by a ratio between the total immigration inflow of persons born outside the EU into Germany as reported by the Eurostat[2] and the median flows from the rest of the work regions into Germany for 2015-2019 by Aristotelous et al. (2022). The process is explained in more detail in Marois et al. (2023), and the correction values, which have been subsequently added to the initial QuantMig estimates for Germany (Aristotelous et al. 2022), are listed in Table 1.

Table 1. Additive correction for migration to Germany from outside the EU+ system, 2015–19

Immigration from:	**2015**	**2016**	**2017**	**2018**	**2019**
East Asia	16,356	6,641	18,493	16,996	14,441
Latin America	10,582	4,652	13,851	16,395	18,407
North Africa	19,849	5,025	8,351	9,964	11,577
North America and Oceania	–2,258	–8,525	2,231	–48	–2,809
Other Europe	186,328	67,271	104,007	107,516	112,331
South-Southeast Asia	111,260	48,410	13,971	17,581	19,555
Sub-Saharan Africa	47,149	9,469	9,784	10,819	11,067
West Asia	414,151	177,515	49,651	23,495	1,326
Total correction	**803,417**	**310,458**	**220,339**	**202,718**	**185,895**

Source: Own elaboration based on DESTATIS data

[2] Eurostat table `migr_imm3cb` was used for that purpose.

For all four distributions, a basic goodness-of-fit assessment was carried out, through a visual analysis of quantile-quantile (QQ) plots and formal Shapiro-Wilk tests, the detailed results of which are reported in Appendix A. The QQ plots in Figure A1 show the median migration estimates and their 90-per cent credible intervals against the fitted values for the four distributions under study: exponential (benchmark), GEV, log-normal and Pareto. The results of the Shapiro-Wilk tests, computed within the R package EnvStats (Millard 2013; Millard and Kowarik 2022), are shown in Table A1 in terms of the p-values both for the median migration estimates, as well as for the corresponding mean estimates, as a consistency check. For the purpose of current analysis, as discussed above, the *median* estimates have been used, for consistency with other element of the migration scenario exercise, although the fit for the four distributions was similar for the mean estimates as well.

The final choice of the distribution for scenario setting has been partially driven by the goodness of fit, and partially by the plausibility of the results obtained. The full results reported in Table B1 in Appendix B for all four distributions, both in terms of means and medians of the quantiles of interest – $q_{0.9}$ and $q_{0.98}$. Amongst the heavy-tailed distributions, the main contenders for selection were the Pareto and GEV distributions, each of which was the best-fitting for roughly half of the regions, judging by the highest p-values in the Shapiro-Wilk tests. At the same time, the results for the GEV distribution were found to be less stable numerically, especially given the small sample size (eleven years) and the three parameters that need estimating for a GEV distribution, as opposed to two for a Pareto. In particular, the GEV yielded very high twice-in-a-century ($q_{0.98}$) values for migration from Latin America, North America and Oceania, and West Asia (see Table B1 for details). On balance, we have selected the quantiles from Pareto distribution as a basis for quantifying the uncertain immigration scenarios.

In summary, for scenario setting we have therefore selected posterior *means* of the quantiles $q_{0.9}$ and $q_{0.98}$ (the 90th and 98th percentiles) from the Pareto distributions fitted to the *median* QuantMig flow estimates for 2009–2019 (Aristotelous et al. 2022), with corrections related to German immigration manually added for 2015–2019, as described above (Marois et al. 2023). The quantile means have been chosen mainly for their additivity across different scenario settings. In scenarios, these two sets of quantiles would serve as input into charting four types of migration trajectories: for each quantile, one trajectory would involve a high-impact event recorded for one year only, after which the trajectory would instantly return to the baseline scenario, and one trajectory, where the return to the baseline would be gradual, over the course of a decade (*idem*). Numerical results for aggregated immigration into the whole EU+

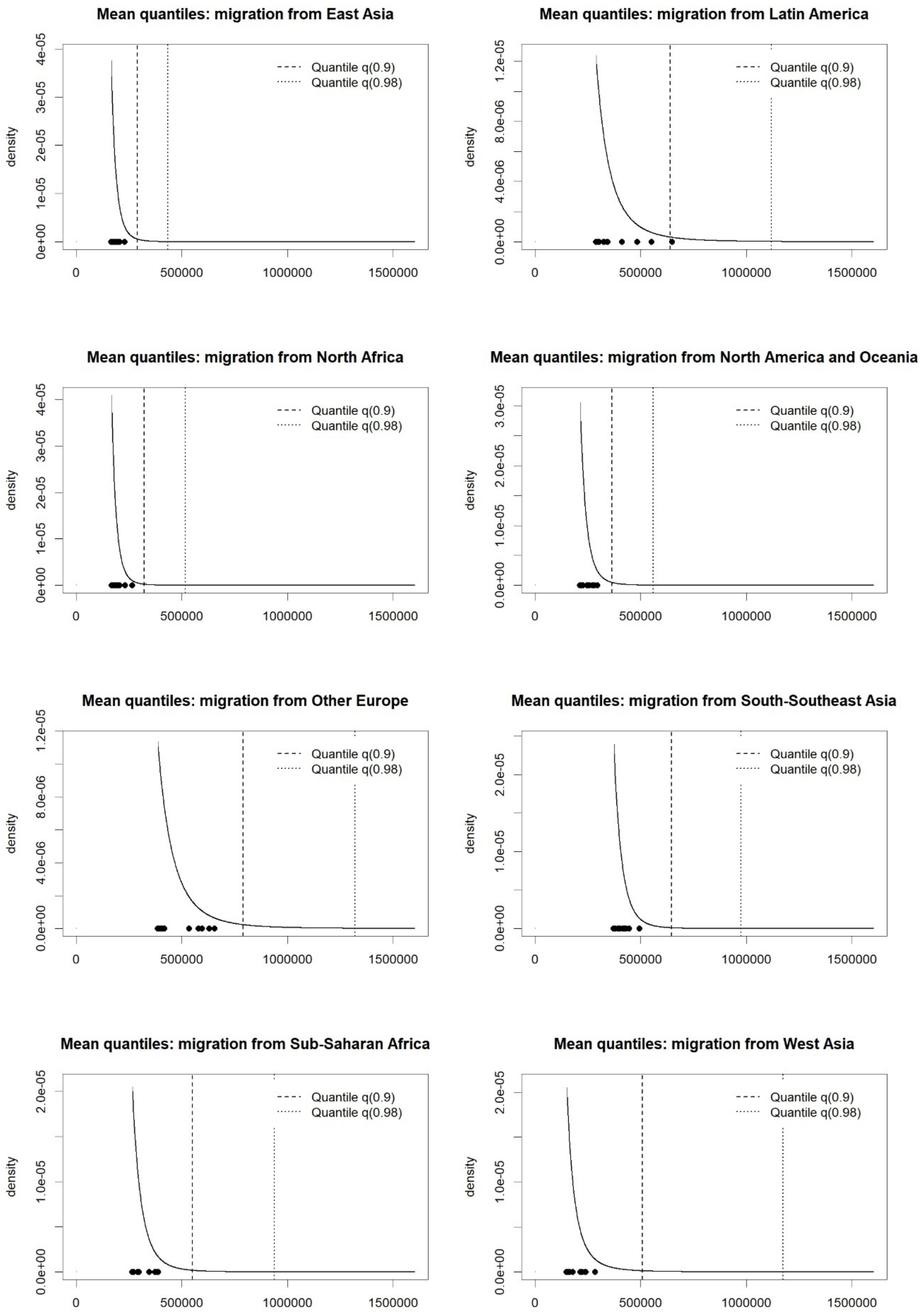
system are shown in Table 2. It is apparent that the values of $q_{0.9}$ are typically around 1.7 times higher than last decade's average, and $q_{0.98}$ – three times higher.

Table 2. Annual average immigration into the EU+ system, 2009–19, and the Pareto distribution quantiles corresponding to rare (once-in-a-decade and twice-in-a-century) events

Immigration from:	Average 2009–19	Pareto q_{90}	Pareto q_{98}
East Asia	192,450	290,973	434,868
Latin America	391,958	639,134	1,118,823
North Africa	194,248	322,715	516,837
North America and Oceania	247,130	364,185	558,724
Other Europe	438,703	790,758	1,318,817
South-Southeast Asia	414,850	645,593	973,279
Sub-Saharan Africa	320,199	549,965	936,962
West Asia	208,352	507,621	1,173,796
Total Rest of the World	**2,407,890**	**4,110,944**	**7,032,106**

Source: Estimates: QuantMig database (Aristotelous et al. 2022); quantiles: own calculations in R package EnvStats (Millard 2013; Millard and Kowarik 2022)

The fitted Pareto distributions and their two selected tail quantiles ($q_{0.9}$ and $q_{0.98}$) are illustrated in Figure 3. Of course, from a purely statistical point of view, fitting heavy-tailed distributions to eleven data points can seem like a heroic exercise, which can only provide approximate magnitudes of possible once-in-a-decade or twice-in-a-century events. This additionally underscores the uncertainty involved in the scenario setting task. At the same time, such quantile values, even if only approximate, can serve as a useful basis for providing a different form of an uncertainty assessment for scenarios and facilitate communicating them to their users, as long as their limitations are made clear. The promises of the contingency-based approach for migration scenario setting, alongside the caveats and health warnings are discussed in the next section.



Mean quantiles: migration from East Asia

Mean quantiles: migration from Latin America

Mean quantiles: migration from North Africa

Mean quantiles: migration from North America and Oceania

Mean quantiles: migration from Other Europe

Mean quantiles: migration from South-Southeast Asia

Mean quantiles: migration from Sub-Saharan Africa

Mean quantiles: migration from West Asia

Figure 3. Selected results for the Pareto distribution: fitted Pareto distributions (solid lines); underlying data – median estimates for 2009–19 (dots •); $q_{0.9}$ and $q_{0.98}$ quantiles – dashed and dotted lines. *Source: Own calculations in R package EnvStats (Millard 2013; Millard and Kowarik 2022)*

5. Discussion and Conclusion

In this report, we have proposed to borrow an approach to preparedness familiar from the area of civil contingency planning, and to apply it to the context of migration scenario setting. The benefits of doing so include a shift of perspective for migration scenarios, departing from the necessity to frame the scenario narratives in terms of the underlying drivers and driver-based migration trajectories, while retaining the probabilistic framing of the outcomes. In our approach, drivers enter into the scenarios only marginally and indirectly, being used in the estimation model to help estimate flows with completely missing observations (Aristotelous et al. 2022). The proposed approximate data-driven approach is less resource-consuming, easier to implement, and can potentially respond to a range of various user needs in terms of the different levels of preparedness required in various areas of migration policy and operations. The proposed framework is also more flexible and can be equally well applied to other flows besides immigration, allowing scenarios to move away from the presumption of Europe solely as the migration destination.

The proposed approach is also applicable more generally, with uses reaching beyond scenario setting. In the context of broadly-understood migration preparedness, wherever higher-frequency data are available, the suggested framework can augment the existing early warning models (see Barker and Bijak 2022). This can be particularly promising in areas, where greater precision of results is important, and where longer series of higher-frequency data are available. One important example is asylum and other types of humanitarian responses, where the availability of monthly, weekly, or sometimes even daily data, boosting the sample size considerably, can increase the precision of the estimation of selected quantiles from heavy-tailed distributions and more fully utilise their analytical potential.

The availability of sub-yearly data is particularly important for making use of distributions from the Generalised Extreme Value (GEV) family. When looking at, for example, annual maxima based on weekly data, or monthly maxima based on daily data – questions that are very important for planning operational response and securing appropriate resources – by construction, GEV naturally becomes the distribution family of choice. With more precise data, not only in terms of time granularity, but also spatial and contextual detail, the presented models can be extended to approaches recognising the spatial, and possibly also categorical dependence between different migration types, origins, trajectories and destinations, as identified by Czaika et al. (2021). Methods for such problems already exist (see, e.g., Heffernan and Tawn 2004 or Towe et al. 2018, in the context of preparedness for civil contingencies), and could be adapted to the migration context. In addition, given that the choice of an appropriate probability distribution is far from unambiguous, such approaches as (Bayesian) model averaging or model selection could be additionally applied to synthesise information across a range of model specifications (for a discussion of migration applications, see e.g. Bijak 2010).

Still, for the presented illustrative examples, as demonstrated by the results shown in Appendix B, given the high sensitivity of the results to the choice of the underlying probability distribution, largely due to the short series of available annual data, it is difficult to select the best model. With such high levels of model uncertainty, the numerical results can be only seen as approximate, broadly exact to the order of magnitude, rather than as precise values.

The approximate nature of the generated quantiles can be illustrated by two most prominent examples of immigration into Europe from the past decade. On the one hand, the 1 million of immigrants coming into Europe from Syria (in our classification, grouped under 'West Asia') in the 2010s are broadly in line with the twice-in-a-century magnitude estimated for that region. On the other hand, between 4 and 6 million migrants from Ukraine since the Russian invasion in 2022 far exceed the corresponding value for 'Other Europe'[3]. Of course, this reflects the relatively low variability of the historical data series when compared to migration in 2022–23, but also underscores the sensitivity of the model choice. At the same time, the possibility of interpreting the recent migration from Ukraine as being of an even higher magnitude than the twice-in-a-century designation would imply, cannot be excluded either.

In addition, for some probability distributions, such as GEV, the results may be also sometimes numerically unstable for short samples, with the quantiles of interest being far out in the distribution tails. This only adds to the overall uncertainty, highlighting the approximate nature of the resulting migration scenarios. In our examples, this model-based layer of uncertainty is not formally included in the estimates, nor taken explicitly into account in the production of final quantiles and scenarios, but needs to be borne in mind when interpreting the numerical results. Still, with increased length

[3] Information from UNHCR, via: https://www.unhcr.org/cy/2021/03/18/syria-refugee-crisis-globally-in-europe-and-in-cyprus-meet-some-syrian-refugees-in-cyprus (citing "over 1 million Syrian asylum seekers and refugees" in the EU as of 18 March 2021) and the most recent Ukraine Situation Report available at the time of writing, https://data2.unhcr.org/en/situations/ukraine (with the reports of 6.3 million refugees based on the differences in border crossings, and as of 26 June 2023). For the EU, Eurostat reports just under 4 million people under temporary protection in April 2023 (data as of 26 June 2023, https://ec.europa.eu/eurostat/databrowser/view/MIGR_ASYTPSM/default/table?lang=en).

and frequency (quarterly) of the data series, sensitivity of the results to the data sample and model choice can hopefully reduce.

Still, at the current state of knowledge, the proposed approach for migration scenario-setting, for all its approximate character, can offer a 'good enough' view of possible migration futures, helping the decision makers improve the levels of preparedness. As long as the uncertainty is properly communicated, and the limitations listed above are clearly mentioned, to avoid the illusory precision of the scenario levels, the proposed tools can be useful in offering indications of the possible orders of magnitude of migration events of different frequencies. At the same time, should better and higher-frequency data become available, this is one important area of migration policy and practice where the epistemic uncertainty – related to limited knowledge about the past and present trends, can be slightly reduced thanks to more and better information. To that end, the proposed framework needs to be seen as work in progress, in need of further research and refinement once more information becomes available. The current report is intended as the first step on the path towards hopefully more robust methodology for setting and using forward-looking migration scenarios in the future.

References

Acostamadiedo E, R Sohst, J Tjaden, G Groenewold and HAG de Valk (2020) Assessing Immigration Scenarios for the European Union in 2030 – Relevant, Realistic and Reliable? Geneva: IOM and The Hague: NIDI.

Arango J (2000) Explaining Migration: A Critical View. *International Social Science Journal*, 52, 283–296.

Aristotelous G, PWF Smith and J Bijak (2022) Technical report: Estimation Methodology. QuantMig Project Deliverable D6.3.

Barker ER and J Bijak (2021) Uncertainty in Migration Scenarios. QuantMig Project Deliverable D9.2.

Barker ER and J Bijak (2022) Could We Have Seen it Coming? Towards an Early Warning System for Asylum Applications in the EU. QuantMig Project Deliverable D9.3.

Bijak J (2010) *Forecasting International Migration in Europe: A Bayesian View*. Dordrecht: Springer.

Bijak J and M Czaika (2020) Assessing Uncertain Migration Futures - A Typology of the Unknown. QuantMig Project Deliverable D1.1.

Bijak J, et al. (2021) *Towards Bayesian Model-Based Demography: Agency, Complexity and Uncertainty in Migration Studies*. Cham: Springer.

Bijak J, G Disney, AM Findlay, JJ Forster, PWF Smith and A Wiśniowski (2019) Assessing time series models for forecasting international migration: Lessons from the United Kingdom. *Journal of Forecasting*, 38(5), 470–487.

Boissonneault M, M Mooyaart, P de Jong and HAG de Valk (2020) QuantMig: The Use of Migration Scenarios in Future Characterisations: A systematic Review and Typology. QuantMig Project Deliverable D7.1.

Cappelen Å, T Skjerpen and M Tønnessen (2015). Forecasting Immigration in Official Population Projections Using an Econometric Model. *International Migration Review*, 49(4), 945–980.

Coles S (2001) *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics. London: Springer.

Czaika M, H Bohnet and A Soto-Nishimura (2021) Spatial and Categorical Dependence of European Migration Flows. QuantMig Project Deliverable 5.2.

Czaika M, MB Erdal and C Talleraas (2021) Theorising the Interaction Between Migration Relevant Policies and Migration Driver Environments. QuantMig Project Deliverable 1.4.

de Beer J (2008) Forecasting international migration: Time series projections vs argument-based forecasts. In: J Raymer and F Willekens (eds), *International migration in Europe: Data, models and estimates*. Chichester: Wiley (pp. 283–306).

Heffernan E and JA Tawn (2004) A conditional approach to modelling multivariate extreme values (with discussion), *Journal of the Royal Statistical Society, Series B*, 66(3), 497–547.

McDonald JB and YJ Xu (1995) A generalization of the beta distribution with applications, *Journal of Econometrics*, 66(1–2), 133–152.

Millard SP (2013) *EnvStats: An R Package for Environmental Statistics*. New York: Springer.

Millard SP and A Kowarik (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).

Marois, G, Potančoková M and M González-Leonardo (2023) Technical report: QuantMig-mic microsimulation population projection model. QuantMig Project Deliverable D8.2.

Potančoková M, G Marois and M González-Leonardo (2023) Demographic and labour force implications of high-immigration events scenarios. QuantMig Project Deliverable D10.1.

Raymer J, A Wiśniowski, JJ Forster, PWF Smith and J Bijak (2013) Integrated Modeling of European Migration. *Journal of the American Statistical Association*, 108(503), 801–819.

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

Taleb NN (2007) *The Black Swan: The impact of the highly improbable*. New York: Random House.

Towe R, JA Tawn and R Lamb R (2018) Why extreme floods are more common than you might think? *Significance*, 15(6), 16–21.

All QuantMig project deliverables listed above are available via www.quantmig.eu

Appendix A. Assessment of the goodness of fit of various probability distributions: Immigration into Europe from eight world regions

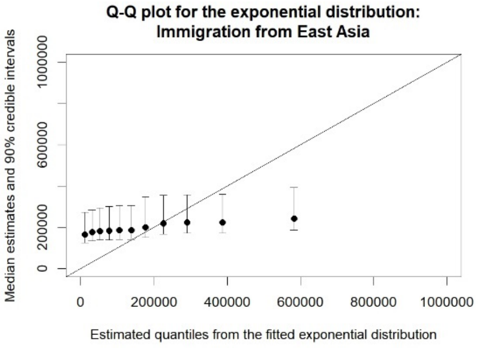
Note: the Q-Q plots use the same scales for all regions, showing estimated and fitted values up to 1,000,000, so occasional large values may be off the scale

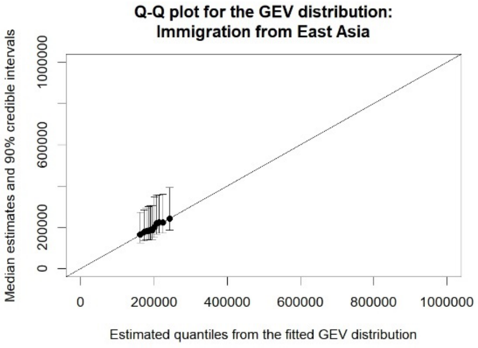
Exponential distribution

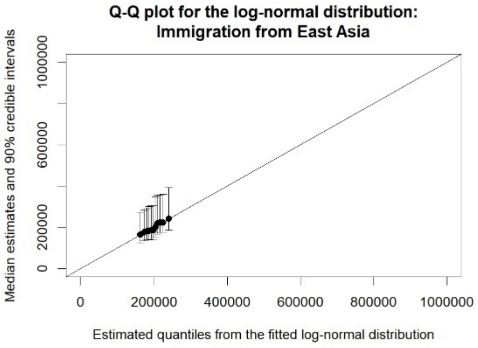
GEV distribution

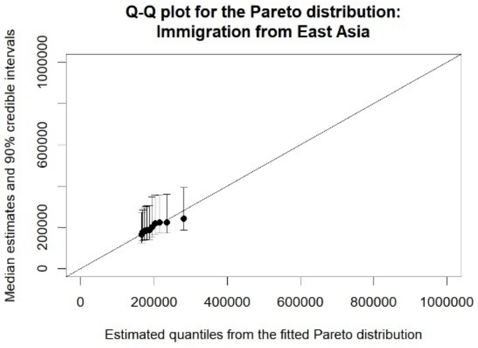
Log-normal distribution

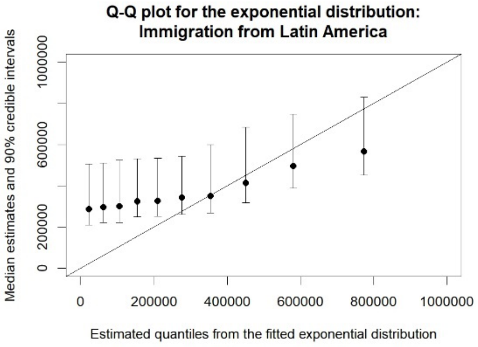
Pareto distribution

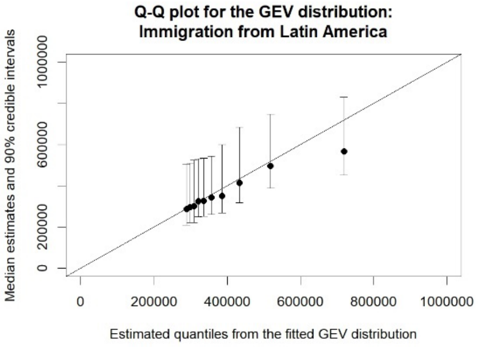

Q-Q plot for the exponential distribution: Immigration from East Asia

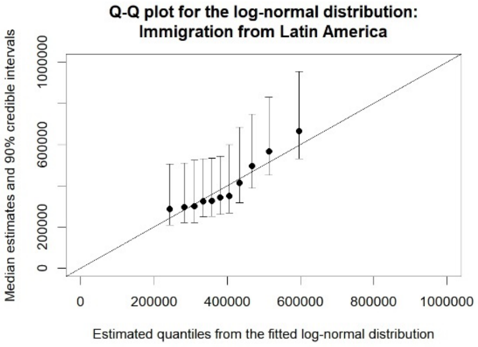

Q-Q plot for the GEV distribution: Immigration from East Asia

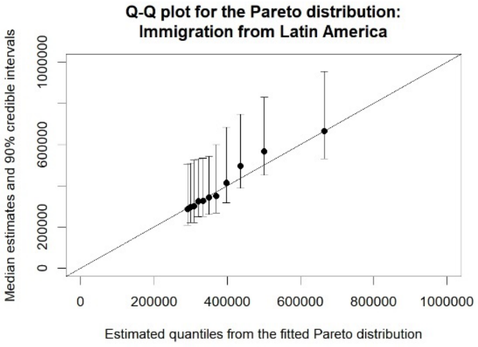

Q-Q plot for the log-normal distribution: Immigration from East Asia

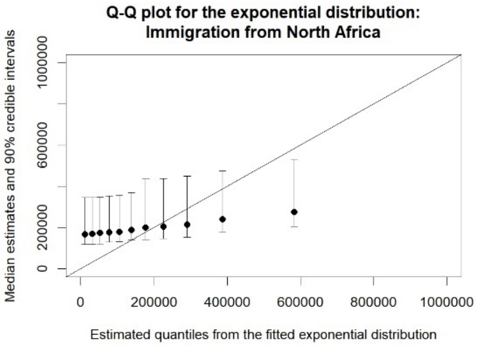

Q-Q plot for the Pareto distribution: Immigration from East Asia

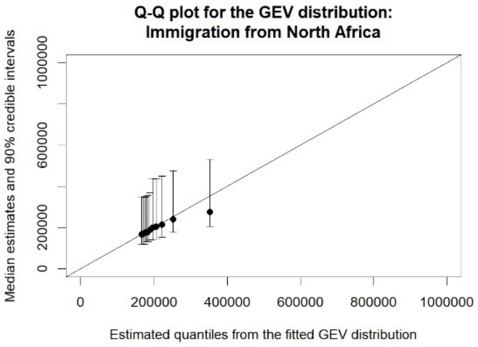

Q-Q plot for the exponential distribution: Immigration from Latin America

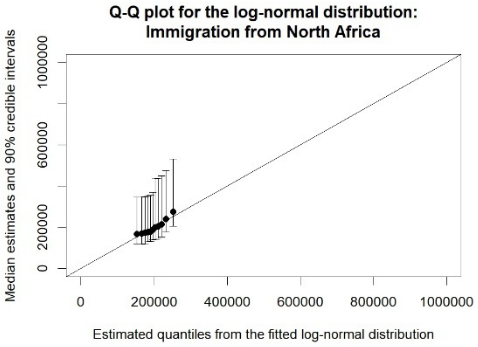

Q-Q plot for the GEV distribution: Immigration from Latin America

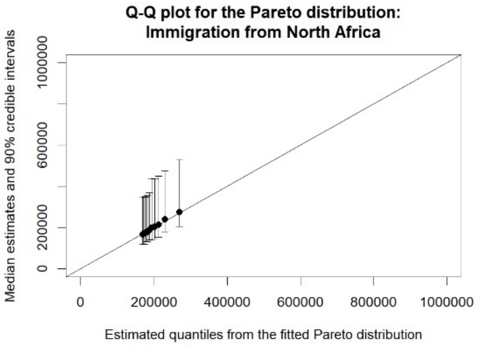

Q-Q plot for the log-normal distribution: Immigration from Latin America


Q-Q plot for the Pareto distribution: Immigration from Latin America

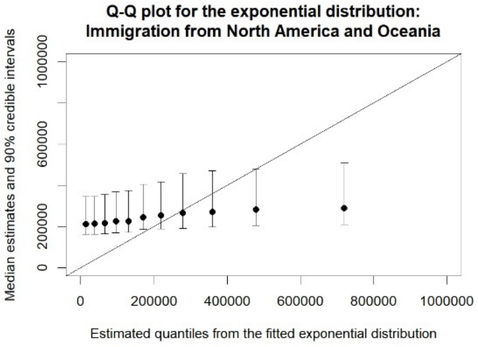

Q-Q plot for the exponential distribution: Immigration from North Africa

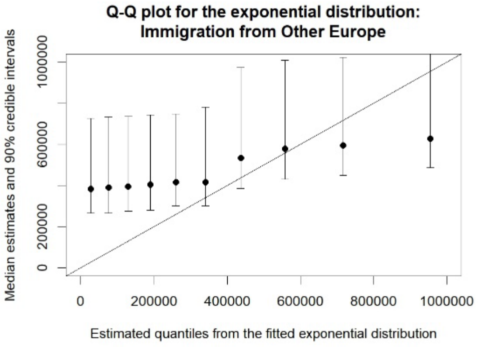

Q-Q plot for the GEV distribution: Immigration from North Africa

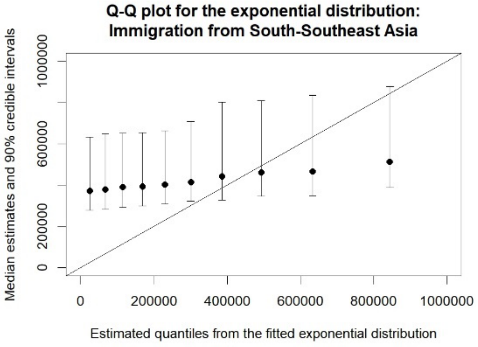

Q-Q plot for the log-normal distribution: Immigration from North Africa


Q-Q plot for the Pareto distribution: Immigration from North Africa

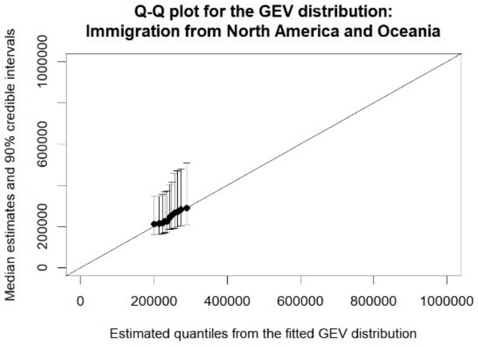
Exponential distribution

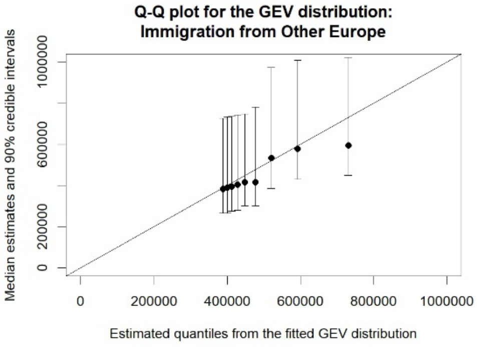


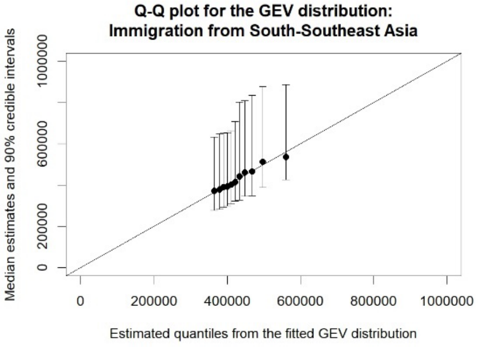




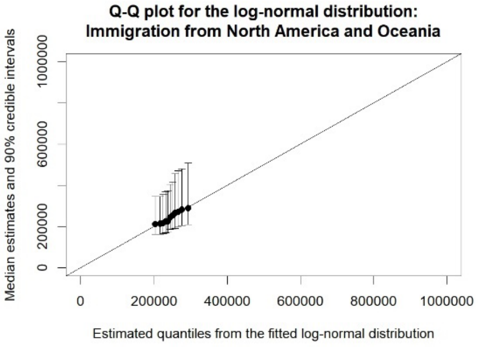
GEV distribution

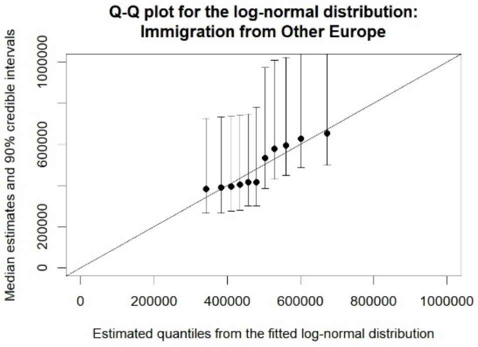


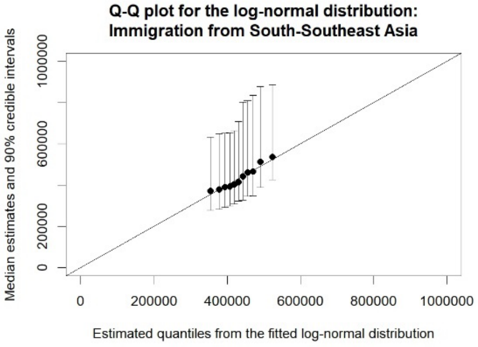




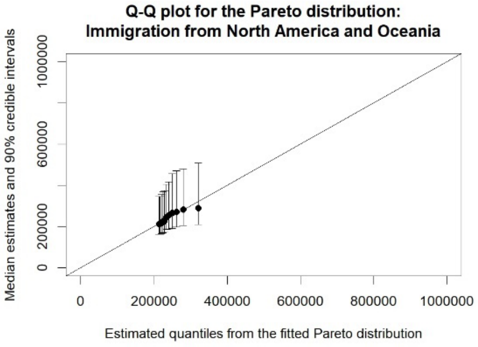
Log-normal distribution

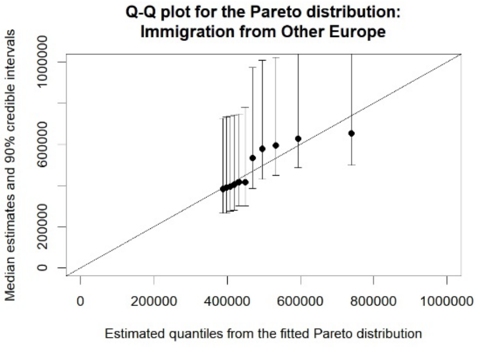


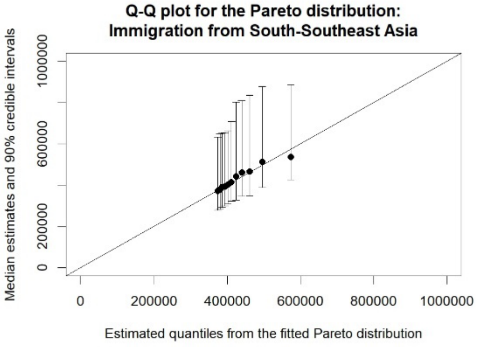




Pareto distribution







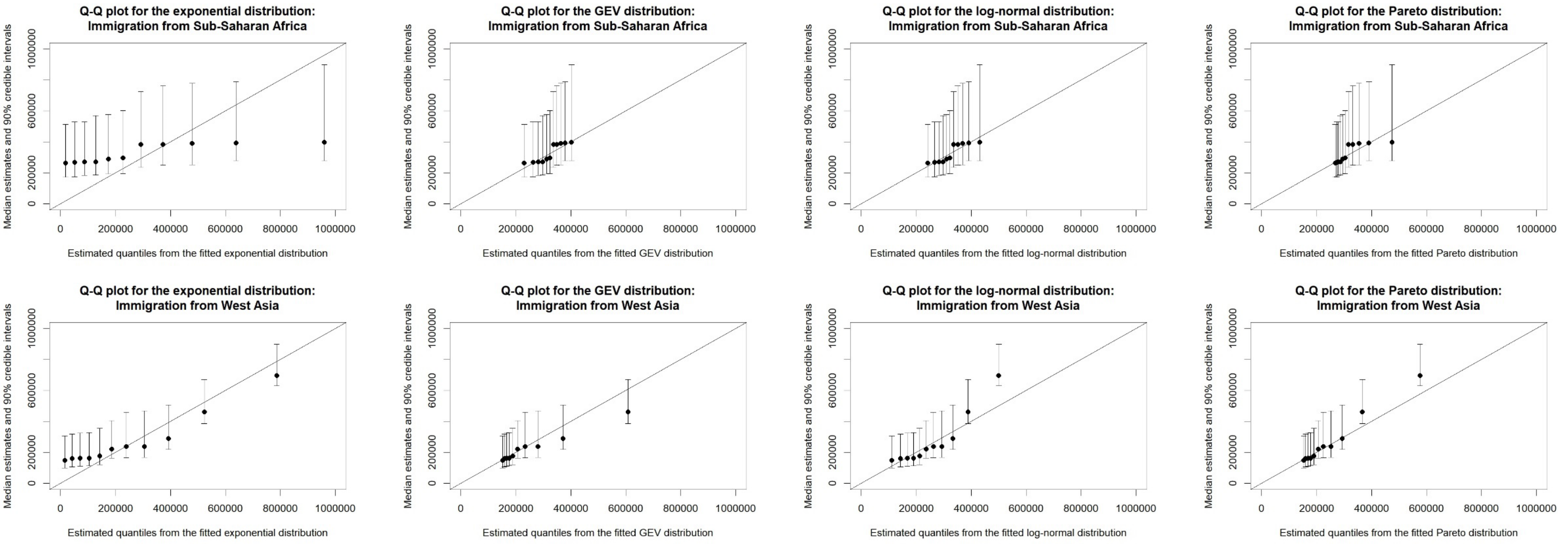


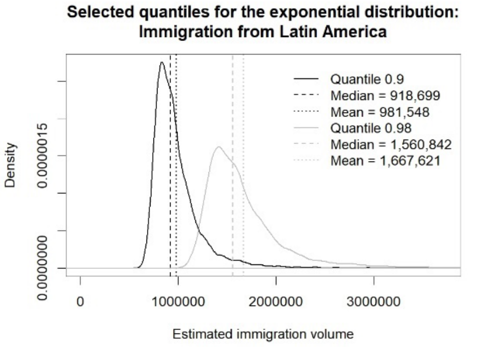
Table A1. P-values from the Shapiro-Wilk test of the goodness of fit for median and mean estimates (H$_0$: distribution as stated). Largest values for each region **in bold**.

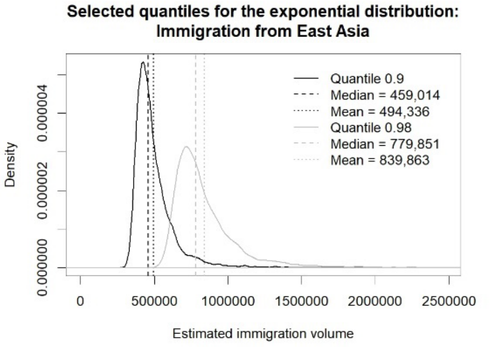
Distributions For MEDIANS	East Asia	Latin America	North Africa	North America and Oceania	Other Europe	South-Southeast Asia	Sub-Saharan Africa	West Asia
H$_0$: Exponential	0.489152	0.044752	0.082316	0.228612	0.016688	0.177308	0.007871	0.014937
H$_0$: GEV	**0.602847**	**0.805418**	**0.674212**	0.241963	**0.103633**	0.393746	0.008634	**0.701651**
H$_0$: Log-normal	0.520943	0.064252	0.102877	0.229417	0.016393	0.193993	0.008112	0.043473
H$_0$: Pareto	0.424391	0.744304	0.530512	**0.263825**	0.060021	**0.575256**	**0.053544**	0.428119
For MEANS (for sensitivity checks only)								
H$_0$: Exponential	0.507897	0.048632	0.123123	0.289967	0.022279	0.334488	0.004001	0.011799
H$_0$: GEV	**0.647404**	**0.810111**	0.917291	0.309663	**0.141393**	0.697045	0.003921	**0.779300**
H$_0$: Log-normal	0.544703	0.072731	0.167644	0.291138	0.022378	0.374408	0.004215	0.037798
H$_0$: Pareto	0.444304	0.723944	**0.988584**	**0.321483**	0.104766	**0.815814**	**0.011716**	0.501131

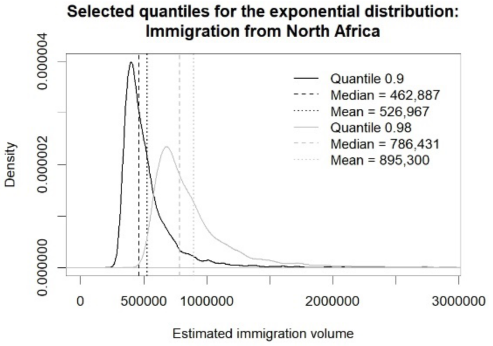
Source: Own calculations in R package EnvStats (Millard 2013; Millard and Kowarik 2022)

Appendix B. Selected quantiles (0.9 and 0.98) from the fitted distributions: Immigration into Europe from eight world regions

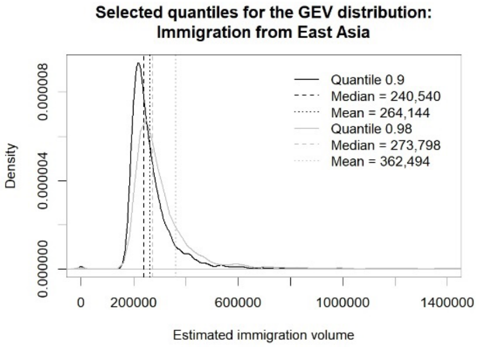
Exponential distribution

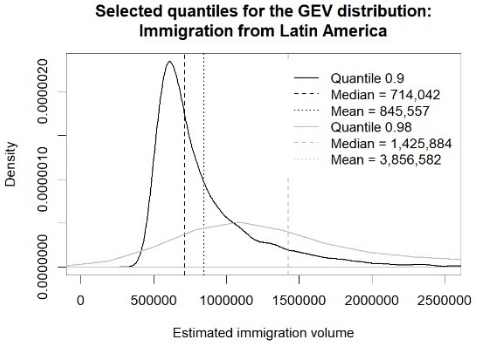

Selected quantiles for the exponential distribution: Immigration from Latin America

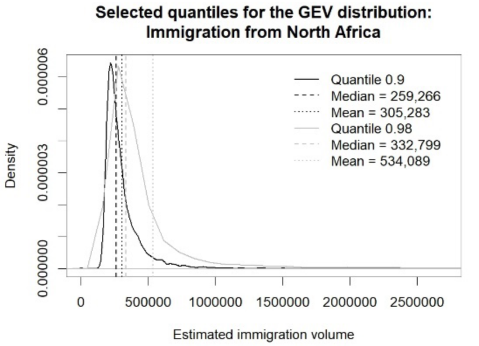

Selected quantiles for the exponential distribution: Immigration from East Asia


Selected quantiles for the exponential distribution: Immigration from North Africa

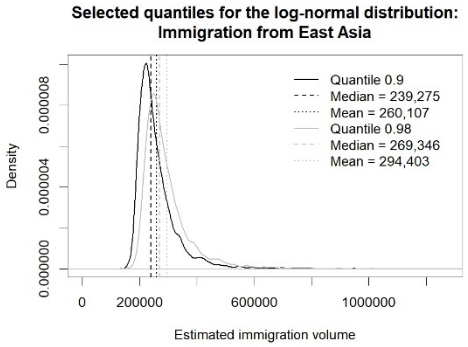
GEV distribution

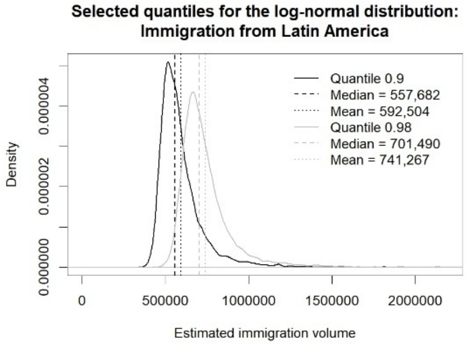

Selected quantiles for the GEV distribution: Immigration from East Asia

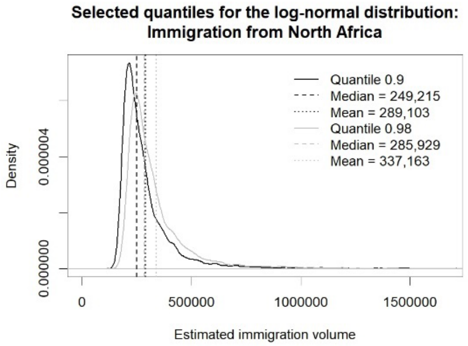

Selected quantiles for the GEV distribution: Immigration from Latin America


Selected quantiles for the GEV distribution: Immigration from North Africa

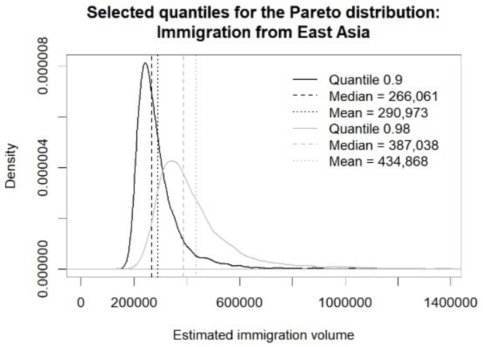
Log-normal distribution

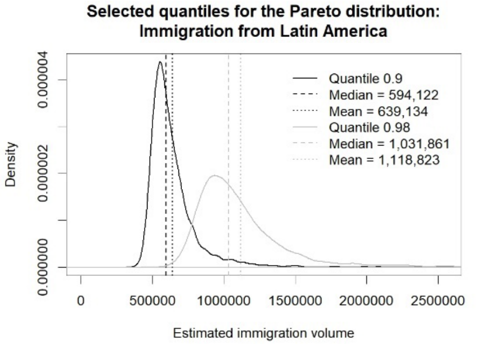

Selected quantiles for the log-normal distribution: Immigration from East Asia

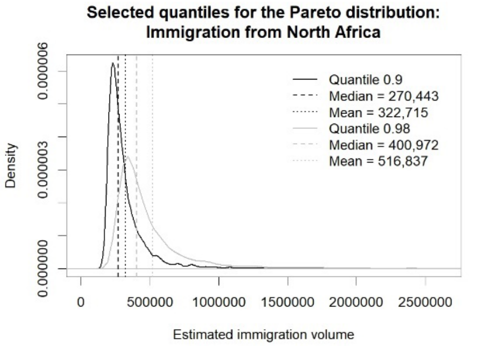

Selected quantiles for the log-normal distribution: Immigration from Latin America


Selected quantiles for the log-normal distribution: Immigration from North Africa

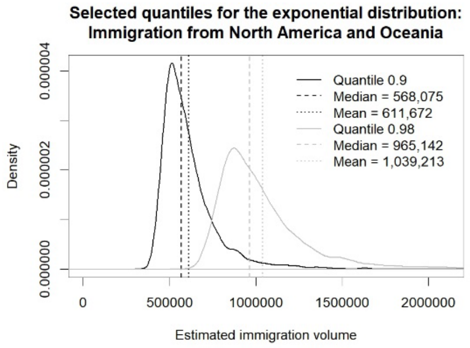
Pareto distribution

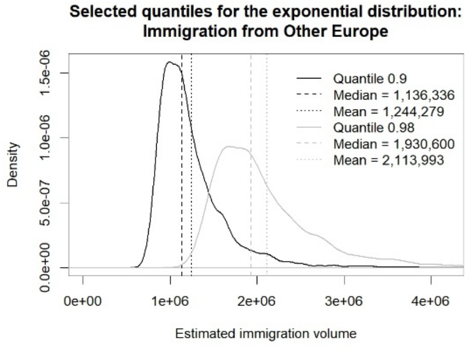

Selected quantiles for the Pareto distribution: Immigration from East Asia

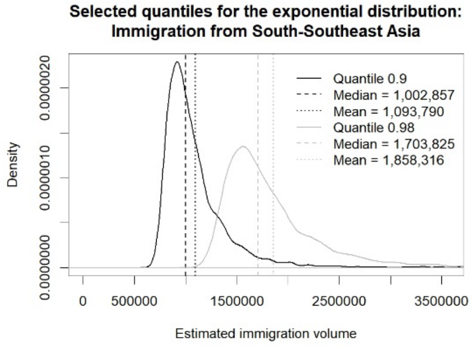

Selected quantiles for the Pareto distribution: Immigration from Latin America


Selected quantiles for the Pareto distribution: Immigration from North Africa

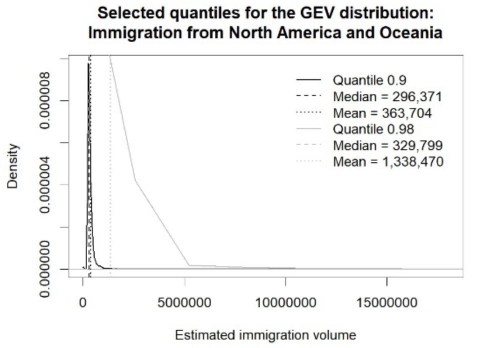
Exponential distribution

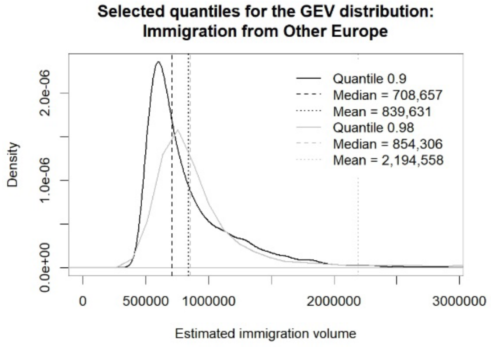


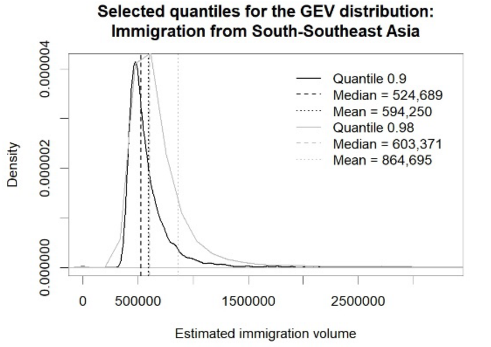




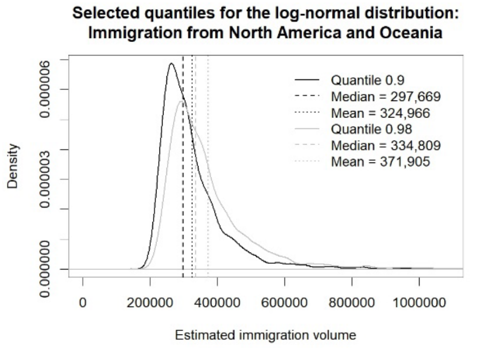
GEV distribution

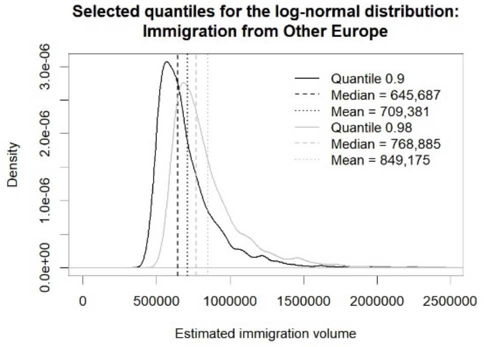


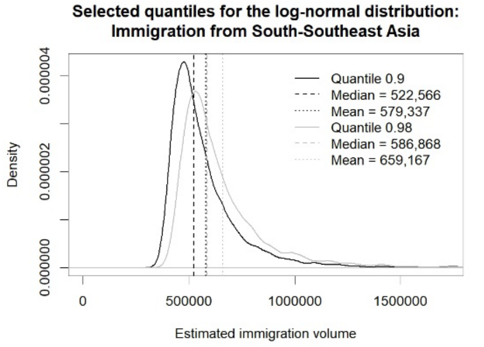




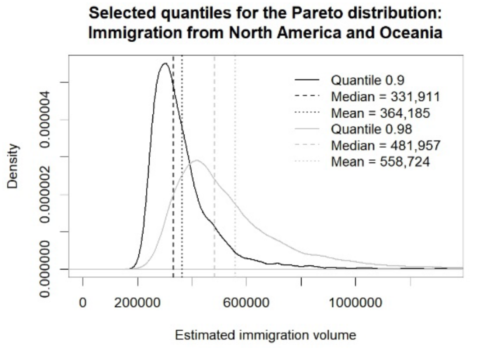
Log-normal distribution

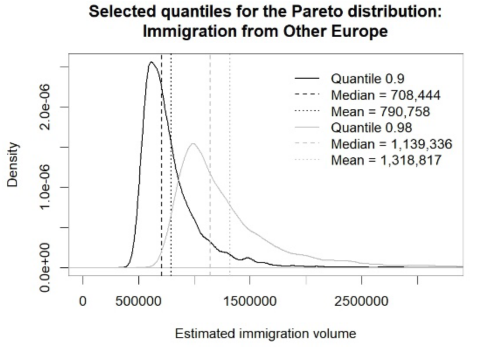


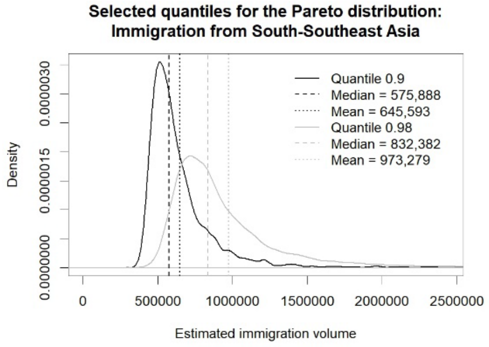


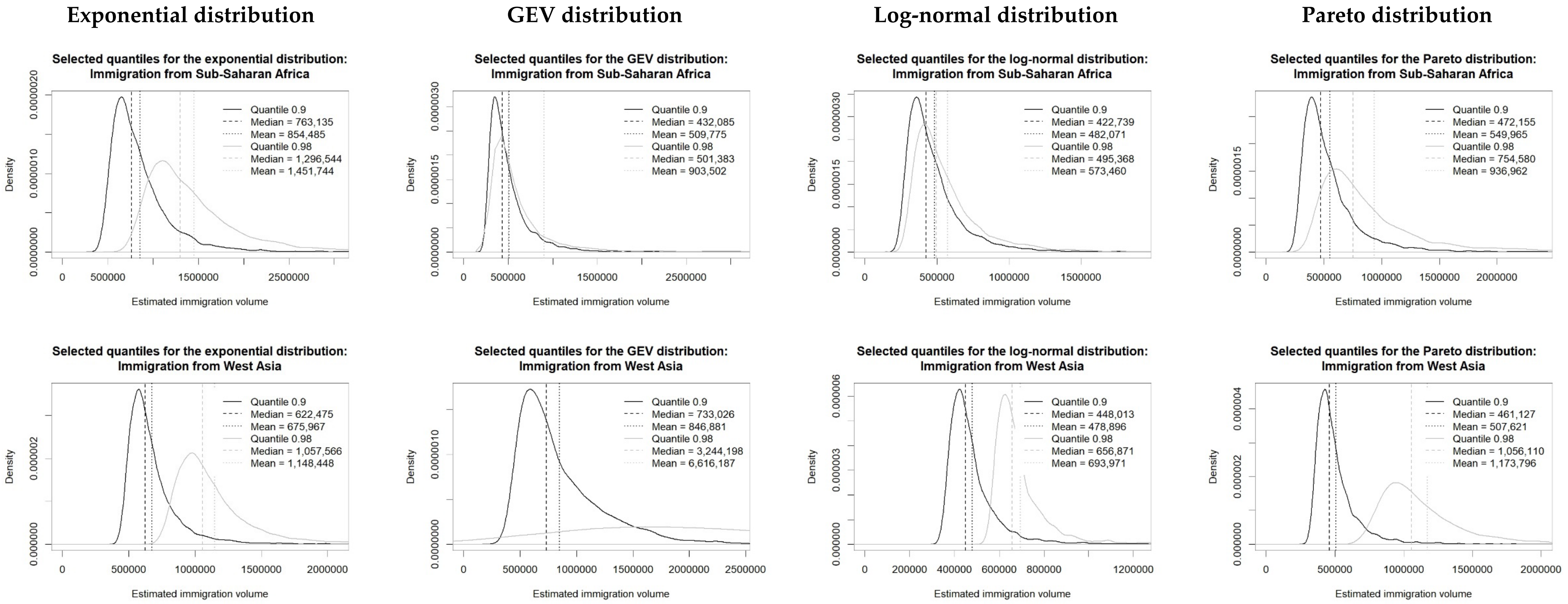


Pareto distribution









Source: Own calculations in R package EnvStats (Millard 2013; Millard and Kowarik 2022)

Table B1. Selected values of the quantiles q_{90} and q_{98} for four distributions fitted to the posterior samples of estimates of annual immigration to Europe. Values for the chosen distribution (Pareto) highlighted **in bold**.

a) Posterior **medians** of the distributions of the selected quantiles

Distribution and quantile	East Asia	Latin America	North Africa	North America and Oceania	Other Europe	South-Southeast Asia	Sub-Saharan Africa	West Asia
q_{90}: Exponential	459,014	918,699	462,887	568,075	1,136,336	1,002,857	763,135	622,475
q_{98}: Exponential	779,851	1,560,842	786,431	965,142	1,930,600	1,703,825	1,296,544	1,057,566
q_{90}: GEV	240,540	714,042	259,266	296,371	708,657	524,689	432,085	733,026
q_{98}: GEV	273,798	1,425,884	332,799	329,799	854,306	603,371	501,383	3,244,198
q_{90}: Log-normal	239,275	557,682	249,215	297,669	645,687	522,566	422,739	448,013
q_{98}: Log-normal	269,346	701,490	285,929	334,809	768,885	586,868	495,368	656,871
q_{90}: Pareto	**266,061**	**594,122**	**270,443**	**331,911**	**708,444**	**575,888**	**472,155**	**461,127**
q_{98}: Pareto	**387,038**	**1,031,861**	**400,972**	**481,957**	**1,139,336**	**832,382**	**754,580**	**1,056,110**

b) Posterior **means** of the distributions of the selected quantiles

Distribution and quantile	East Asia	Latin America	North Africa	North America and Oceania	Other Europe	South-Southeast Asia	Sub-Saharan Africa	West Asia
q_{90}: Exponential	494,336	981,548	526,967	611,672	1,244,279	1,093,790	854,485	675,967
q_{98}: Exponential	839,863	1,667,621	895,300	1,039,213	2,113,993	1,858,316	1,451,744	1,148,448
q_{90}: GEV	264,144	845,557	305,283	363,704	839,631	594,250	509,775	846,881
q_{98}: GEV	362,494	3,856,582	534,089	1,338,470	2,194,558	864,695	903,502	6,616,187
q_{90}: Log-normal	260,107	592,504	289,103	324,966	709,381	579,337	482,071	478,896
q_{98}: Log-normal	294,403	741,267	337,163	371,905	849,175	659,167	573,460	693,971
q_{90}: Pareto	**290,973**	**639,134**	**322,715**	**364,185**	**790,758**	**645,593**	**549,965**	**507,621**
q_{98}: Pareto	**434,868**	**1,118,823**	**516,837**	**558,724**	**1,318,817**	**973,279**	**936,962**	**1,173,796**

c) Posterior **means** of the distributions of the selected quantiles, all-region sums

Distribution and quantile	Exponential	GEV	Log-normal	Pareto
q_{90}: Sum	6,483,044	4,569,225	3,716,365	**4,110,944**
q_{98}: Sum	11,014,498	16,670,577	4,520,511	**7,032,106**

Notes. The q_{90} values correspond to once-in-a-decade events, and q_{98} to twice-in-a-century events. Means are additive, so the sum of the mean values is the mean of the sum. The same property does not hold for medians, or other quantile-based measures. *Source: Own calculations in R*

■