Journal of the Royal Statistical Society Series A: Statistics in Society, 2025, **00**, 1–21
https://doi.org/10.1093/jrsssa/qnaf196

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

Original Article

Asylum-seekers at the extremes

Mohammad Noori and Marco Bee

Department of Economics and Management (DEM), University of Trento, 5, Via Inama, Trento 38122, Italy

Address for correspondence: Mohammad Noori, Department of Economics and Management (DEM), University of Trento, 5, Via Inama, Trento 38122, Italy. Email: mohammad.noori@unitn.it, mohammadenoori@gmail.com

Abstract

This article conducts a first-hand study on forecasting asylum-seekers' applications across seven major host countries, i.e. the USA, Germany, France, Italy, Spain, Austria, and Greece by 2027. Poisson regression is used for flow forecasting and Extreme Value Theory for the maximum monthly number of applications. As for the latter, we first employ a static version of the Peaks-over-Threshold approach, before modelling the possible non-stationarity of the exceedances via a Generalized Pareto Distribution with time-varying parameters. Overall, the findings are in line with one another and can provide a useful road map to policymakers.

Keywords: extremes, forecasting, migration, refugees, statistical modelling

JEL codes: C1; C53; E6; J15

1 Introduction

According to United Nations High Commissioner for Refugees (UNHCR) 'An asylum-seeker is someone who is seeking international protection. Their request for refugee status, or complementary protection status, has yet to be processed, or they may not yet have requested asylum but they intend to do so'. Wars, natural crises, extreme destitution, political instability, unfair persecution, violations of human rights are among the primary reasons, one seeks asylum.

UNHCR reports[1] that, as of mid-2023, there are over 110 million forcibly displaced people all around the globe, of which 34 millions are refugees[2] and over 6.1 million registered asylum-seekers. This is a worrisome trend given the approximate 46 million cases back in 2012. Syria (6.3 million), Venezuela (6.2 million), Afghanistan (6.1 million), and Ukraine (6.1 million) are the four major source countries with the largest groups of refugees across the worlds, while Iran (3.8 million), Turkey (3.1 million), and Columbia (2.8 million) are the three major host countries for refugees.[3] According to the UNHCR, 'Not all asylum-seekers will be found to be refugees, but all refugees were once asylum-seekers'. Figure 1 shows the major causes of forced displacement around the world from 1975 to the end of 2022.[4]

War and regional armed conflicts are the major causes of forced displacement. Many asylum-seekers lose their lives in reaching their destinations. Such stories and news have created public grievance. The death of Alan Kurdi, the 2-year-old Syrian refugee in September 2015, or the

1 https://www.unhcr.org/refugee-statistics/

2 Note that, in this study, *refugees* and *asylum-seekers* are used interchangeably, while referring to the first-time asylum applicants.

3 https://www.unhcr.org/refugee-statistics

4 https://www.unhcr.org/global-trends-report-2022.

Received: November 30, 2024. **Revised:** November 4, 2025. **Accepted:** November 18, 2025

© The Royal Statistical Society 2025.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/jrsssa/advance-article/doi/10.1093/jrsssa/qnaf196/8369180 by Universita' di Trento - Biblioteca di Ateneo user on 11 December 2025

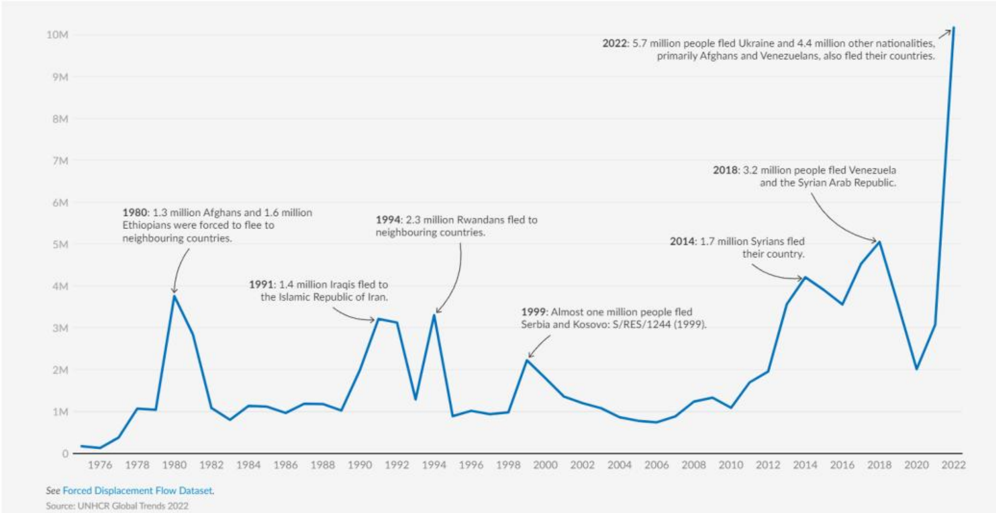

Figure 1. Forced displacement 1975–2022.

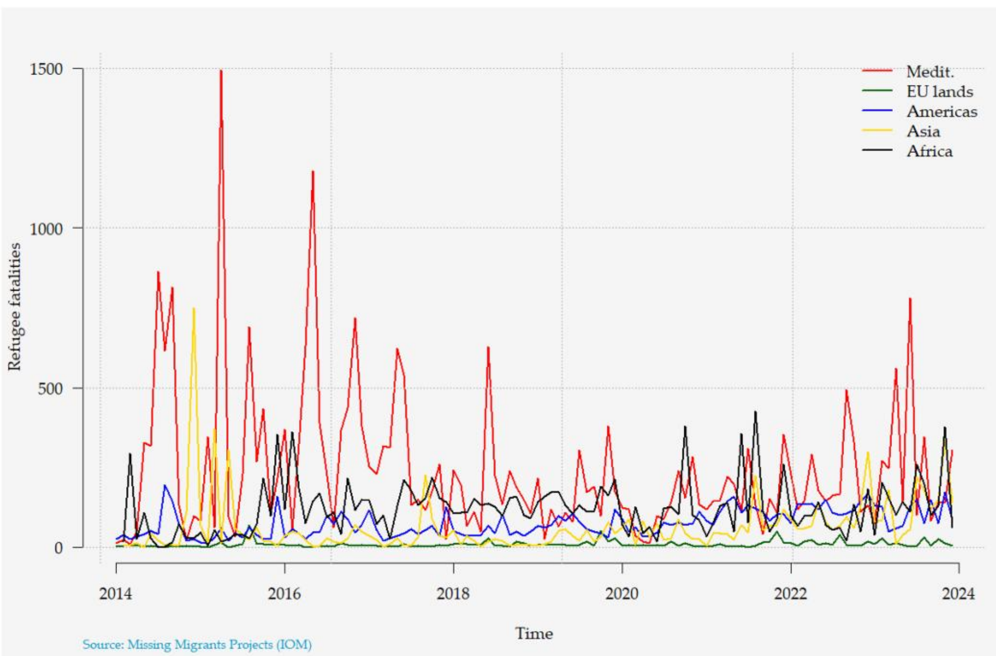

Figure 2. Fatalities across the world regions 2014–2023.

Greece shipwreck incident (June 2023), where around 80 dead bodies were located in the Ionian Sea, are only a few instances among the many. Figure 2 exhibits the monthly refugee fatalities across different world regions. As the red line suggests, the Mediterranean region has the highest death toll across the world regions, followed by Africa (black).

Motivated by the lack of literature in forecasting refugee flows, this article models and forecasts the expected asylum-seekers' flow across seven major host countries including the USA, Germany,

Downloaded from https://academic.oup.com/jrsssa/advance-article/doi/10.1093/jrsssa/qnaf196/8369180 by Universita' di Trento - Biblioteca di Ateneo user on 11 December 2025

France, Italy, Spain, Austria, and Greece by 2027,[5] using Poisson regression methods and Extreme Value Theory (EVT). As concerns the latter, we first employ the Peaks-over-Threshold (POT) approach (Embrechts et al., 1997, Section 6.5), and estimate the generalized Pareto distribution (GPD) by means of maximum likelihood estimation (MLE) and Bayesian techniques; then, we also consider non-stationary models with time-dependent parameters. The main goal of Poisson regression is flow forecasting: we find that apprehensions in the US south border can surpass 350,000 and reach 460,000 cases per month by 2027. In case of Germany, the forecast shows that asylum applications would be within 27,000–52,000 cases, while France and Spain can potentially receive between 17,500 to 20,000 and 33,000 to 43,000, respectively. In terms of maximum applications per month, stationary POT using MLE allows us to find out that there is no upper bound (i.e. an unbounded, heavy-tailed distribution) for the USA, Germany, and Austria. In particular, the findings indicate that the asylum applications in case of Germany will most likely exceed 21,700, and there is a 25% chance that they surpass 39,000. On the other hand, France, Austria, Italy, Spain, and Greece have an upper bound for asylum applications, due to the estimated negative value of the shape parameter in the POT modelling.

We then utilize Bayesian methods to estimate GPD parameters. Overall, the outcomes predict a slightly greater number of asylum applications across different countries compared to predictions obtained by the MLE GPD, especially at the tails (including the upper bound). The longer right tail (for most countries), means that chances are a larger number of asylum applicants land in a host country.

Given the observed non-stationarity in the exceedances, the final section of the article is dedicated to the non-stationary GPD model, with time-varying threshold and regime-dependent shape parameter. The non-stationary results are comparable to the stationary ones, particularly for Germany, Austria, and Spain. On the other hand, in case of the USA, the non-stationary models yield more reliable predictions.

The rest of the article is organized as follows. In Section 2., we summarize the background and some literature. In Section 3., we describe the data. In Section 4., we predict the asylum applications by 2027 using Poisson regression and discuss possible alternative approaches, while in Section 5. we investigate the maximum number of asylum applications by means of both the static GPD, estimated via MLE and Bayesian methods, and the non-stationary GPD. Section 6. discusses our study limitations, and Section 7. concludes.[6]

2 Background and literature

Refugees are mostly perceived as a major cause of concern across the world. For instance, a sudden massive flow of migrants to a country can put the infrastructure at stress. This is because the current state of economy is at equilibrium, and a demographic shock can indeed result in unemployment (especially in the low skilled sectors), food shortages or housing crisis, among others (Araya et al., 2021; Jaafar et al., 2020; Jacobson, 2021; Peters, 2022). Yet, this is only one side of the coin: the right to asylum is a human right under Article 14 of the Universal Declaration of Human Rights, so that the welfare of refugees is also, or perhaps more, crucial. Accordingly, it is important to predict their arrival, both to minimize potential adverse effects and to set up immigration policies and facilities. Borjas and Monras (2017) state that refugee supply shocks are truly exogenous, and both the timing and size of supply shocks have nothing to do with the economic state of the destination country. The consequences of refugee shocks are not limited to labour market. Rozo and Vargas (2021) find that the Venezuelan humanitarian crises and shocks have significantly affected Colombian voters political preferences from left-wing to right-wing, and an increase in political participation. Alhawarin et al. (2021) evidence the negative impact of Syrian refugees on local rentals market, whereas Tanrıkulu (2021) witnesses positive impact of refugees on economic growth in Turkiye. All in all, these studies, among many others, accentuate the importance and the

[5] We also include a tentative 2030–forecast in the online supplementary material. Since the nature of the refugees flow is quite dynamic and dependent on many variables, the distributional properties of the refugee arrivals can change. Even an economic shock in one of the major host countries can change the (statistical) distribution of the refugees. Thus, caution is advised in interpreting and relying on this longer forecast.

[6] The forecasts by 2030, forecast accuracy results including the outcomes of a Gamma-distributed Generalized Linear Model, and the diagnostics about the fit of non-stationary EVT are all provided in the online supplementary material.

Downloaded from https://academic.oup.com/jrsssa/advance-article/doi/10.1093/jrsssa/qnaf196/8369180 by Universita' di Trento - Biblioteca di Ateneo user on 11 December 2025

need for forecasting refugee arrivals more than ever. To cite just the best known cases, the recent and ongoing world conflicts from the Russian invasion of Ukraine (in 2014 and 2022), or the Israel clashes with Gaza and Lebanon since 2023, will have a crystal-clear outcome, which is forced displacement. Accordingly, the motivation for this work is to conduct a first-hand forecasting analysis on asylum-seekers across major host countries in the world (with publicly available data) using methods from EVT.

Refugees differ from regular migrants who (mainly for job opportunities) choose to reside in a country other than their own. An instance of that are the points-based immigration systems in Canada, Australia, the US H1 visa, or the EU blue card among many other various forms. Legal migrants enter a host country following the local laws and regulations. They do not create pressure on the host infrastructure, economy, or the demographic, since their movements are managed and planned. On the other hand, refugees' arrival is unexpected, and not fully manageable. The ongoing conflict in the Sahel region of Africa, the Russian invasion of Ukraine, the humanitarian crisis in Gaza and Lebanon, Latin American political instability and poverty, Afghan, Syrian, or Kurd minorities freedom issues among others, all would translate to an increased number of refugees.

According to the Missing Migrants Project of the International Organization for Migration (IOM), the death toll in the Mediterranean region surpasses by far any other region across the world. Nearly 30,000 refugees have either died or gone missing passing through the Mediterranean sea where over 27,500 drowned in the water. The Africa region, and in particular North and North-West Africa has had the highest fatality in the second place. Harsh environment (e.g. high temperature, water shortage) is the major cause of fatalities in that region. The Americas region comes in third as the mostly deadly route for refugees. From 2021 to 2023, the annual refugee fatalities in the region has nearly doubled compared to previous periods. Figure 2 illustrates the IOM's annual refugee fatalities across different regions.

EVT has been a jack of all trades when it comes to predicting and characterizing not only natural crises including floods (Canfield et al., 1980; Morrison & Smith, 2002), tornadoes (Caldera et al., 2018; Tippett et al., 2016), pandemic diseases (Campolieti, 2021; Nadarajah & Ojo, 2023), wildfires (Abreu et al., 2022; Jiang & Zhuang, 2011; Keyser & Westerling, 2019) but also buildings and structures (Chen et al., 2020; Ngo & Nguyen, 2024), financial markets' doom days, insurance losses (Bee et al., 2017; Cotter, 2006; Rohrbeck et al., 2018; Uppal, 2013) and even issues in urban studies (Auerbach & Wan, 2020).

EVT is a set of probability methods that deals with extreme observations. Unlike most classical statistical approaches, instead of fitting a single distribution to the entire dataset, it deliberately investigates only the tails of the data, via general asymptotic results requiring mild assumptions. Probabilities of rare events, possibly larger than the most extreme observations, are then evaluated by extrapolation. In terms of implementation, we need to select a suitable threshold to identify the point where the tail starts. In particular, the POT method suggests to choose a high enough threshold, such that any data points above that cutoff will approximately follow a GPD. By exploiting this procedure, we will be able to identify the distributional characteristics of the tail and predict the extremes of asylum requests in the near future. Note that the design of our study and the available data point towards this method: since we have monthly data over 16 years, and we need to predict large masses of asylum applicants, other EVT approaches, such as block maxima, would not be appropriate.

The refugee flow forecasting literature is completely depleted. This is directly linked to the lack of comprehensive and detailed databases. The closest work to this study is Carammia et al. (2022), where the authors develop an adaptive machine learning framework model to predict asylum-seekers movements. The model is then used to forecast refugee arrivals up to one-month ahead. Thus, it is crucial to develop models capable of predicting refugee arrivals at longer dates, which is the main motivation of this article. While Carammia et al. (2022) forecast refugee arrivals only for the EU host countries, in the present study we also include the USA south border (refugee crisis) in our analysis,[7] in addition to six major European host countries. To predict the asylum-applications flow (based on the existing trends), we rely on quasi-Poisson regression. This approach suits our integer, non-negative and over-dispersed data. Essentially, with the quasi-Poisson modelling we exploit the pre-existing flow, while with the POT method we forecast the extremes via the GPD.

[7] https://www.theguardian.com/

Downloaded from https://academic.oup.com/jrsssa/advance-article/doi/10.1093/jrsssa/qnaf196/8369180 by Universita' di Trento - Biblioteca di Ateneo user on 11 December 2025

Table 1. Asylum-seeker applications descriptive statistics

Country	Mean	SD	Skewness	Kurtosis	Maximum	Obs.
USA	83,229	71,441	1.3	3.2	301,983	147
Germany	17,032	16,647	2.2	8.5	94,350	192
Austria	2,730	2,942	2.7	11.1	18,450	192
France	7,964	3,619	0.6	2.2	16,770	192
Italy	4,930	3,646	0.8	2.7	15,485	192
Spain	3,643	4,448	1.1	3.25	17,600	192
Greece	2,710	2,381	1.3	4.2	11,045	192

3 Data

Our data come from the European Statistical Office (Eurostat) database. The monthly data starts from January 2008 up to December 2023. Based on the European Council 2023 report,[8] over 1,129,000 non-EU refugees applied for asylum in the EU in 2023. The six major host countries including Germany, France, Spain, Austria, Italy, and Greece, received and hosted the highest number of asylum-seeker applications, amounting to 75% of all applications.

Regarding the USA asylum cases, our data are sourced to the aggregate data on south border apprehensions by the US Customs and Border Patrol (CBP).[9] We use the monthly total apprehensions since October 2011 up to December 2023. We consider apprehension as the first-time asylum application, since this is the only publicly available data.

Table 1 displays some descriptive statistics. The positive skewness indicates that data are mainly located on the left side of the mean; hence, there are a few large outliers. The bigger the skewness (like that of Austria or Germany), the greater the sudden asylum waves. The rather large kurtosis in case of Germany and Austria is another piece of evidence pointing in the same direction.

Note that the series shows autocorrelation.[10]

4 Asylum-seekers flow: a Poisson regression approach

We start our empirical analysis of the asylum-seekers' flow using a Poisson regression approach. Since we are dealing with count data, i.e. positive integer observations, an appropriate method is based on the Poisson distribution with time-varying parameter. In this regard, the number of asylum-seekers at time t or N_t is assumed to follow a Poisson process with time-varying parameter $\lambda_t = \alpha + \beta t$, where α and β are the intercept and the growth coefficient, respectively. From this assumption, it follows that the marginal distribution at time t is Poisson: $N_t \sim \text{Pois}(\lambda_t)$ and that the expected amount of asylum-seekers at time t is modelled as:

$$E[N_t] = \exp(\alpha + \beta t).$$

From a theoretical point of view, Poisson regression is based on the assumption of iid observations (Dobson, 1990), which in our case is not met. However, the Poisson distribution arises as an approximation of the binomial when the product of the number of trials and of the probability of success converges to a finite limit, a result sometimes known as *law of rare events*, which remains valid under weak dependence (Freedman, 1974). While the first-time asylum applications are autocorrelated, and the flows from different countries at different times are uncorrelated, it is difficult to quantify the dependence among people who migrate at the same time from a given country. Here, we assume it to be sufficiently weak to justify the Poisson limit. At least in periods with no major conflicts, this assumption may be justified, but we acknowledge that it is a rather strong requirement.

The model can be estimated, for each country, by means of the commands of the `R` `stats` package. However, a key assumption in Poisson regression modelling is the equality of the distribution mean

8 https://www.consilium.europa.eu/en/infographics/asylum-applications-eu.

9 https://www.cbp.gov/newsroom/stats/cbp-public-data-portal.

10 Results are not reported due to space constraints. We delve into this matter later.

Downloaded from https://academic.oup.com/jrsssa/advance-article/doi/10.1093/jrsssa/qnaf196/8369180 by Universita' di Trento - Biblioteca di Ateneo user on 11 December 2025

and variance. Since in many applications this condition is not met, two alternative distributions have been proposed in the literature, i.e. quasi-Poisson (also known as over-dispersion model with quasi-likelihood) and negative binomial (also known as gamma-Poisson); see Ver Hoef and Boveng (2007).

Accordingly, given the evident over-dispersion (see Table 1), we build a model based on the quasi-Poisson distribution. Figure 3 shows the observed monthly asylum applications across different countries, as well as the 3-year forecast with the 95% prediction interval.

In general, there are three main evident patterns including the 2014–2015 EU refugee crisis, the 2022 Ukrainian mass refugees (EU hosts) and the 2020 COVID-19 period across several host countries (Figure 3). If the COVID-19 pandemic had never occurred (as a buffer), we conjecture that the pre-COVID trend would have probably continued following the Poisson regression fitted line. These three shocks changed both the trends and refugee policies across the host countries.

The estimated growth rates $\hat{\beta}$ in Table 2 suggest that the asylum applications across the USA, France, and Spain grow by 1.6%, 0.7%, and 2% respectively, while in case of Germany, Austria, Italy and Greece they are equal to 0.6%, 0.8%, 0.7%, and 0.9%. Accordingly, Spain and the USA have the highest asylum applicants growth rate among others. Figure 3 (top left) shows that asylum applications in the USA might surpass 350,000 and reach 460,000 (within the 95% prediction interval) cases per month by 2027 if the existing trend continues. In case of Germany, we expect the asylum applications to go beyond 27,000 up to 52,000 cases, while for Austria 6,000 is the lowest range of our 2027-forecast. France, Spain, and Italy are considered the major refugee gates to the EU territories (in line with mortality rate in the Mediterranean region in Figure 2). The north African countries including Morocco, Algeria, Libya, and Tunis are the departure points for refugees to reach the southern EU countries. We find that France and Spain, with rather low uncertainty band in the forecasts, might witness 17,500–20,000, and 33,000–43,000 asylum applicants by 2027, respectively.

The volatile political agenda regarding refugee intake in Italy and Greece (Figure 3 bottom-left and bottom) is reflected in their asylum-seeker trends. As a result, the Poisson regression forecasts for Italy and Greece might not show what is expected to happen in the coming years. The state of the economy, high unemployment rates, housing crisis, language barriers and strict immigration policies are among the major drivers, an asylum-seeker chooses to pass through (and not settle in) these countries, with the exception of Ukrainian and Russian refugees who potentially have more wealth, education, and skill compared to the African, Afghan, or Syrian refugees. Additionally, many African refugees are well aware of the German, French and Spanish culture and languages, since in the past the home countries of these refugees were a colony of the host countries. Overall, we can assume that in the near future, due to ongoing crises such as the Sudan hunger crisis,[11] the political instability like the 2023 Niger coup as part of the wider Sahel region, the ISIS terrorist activities in Burkina Faso and Mali, or Israel's conflict with Gaza and Lebanon, the southern European countries should expect to see an uprise in refugee arrivals.

4.1 Forecast accuracy and alternative model

In this section, we evaluate the forecast accuracy of the Poisson regression model using an out-of-sample validation approach. Detailed accuracy measurements based on cross-validation are reported in the online supplementary material (Appendix B).

In our investigation, we keep the last 10% of our data as out-of-sample, and use the remaining 90% (in-sample) to fit the Poisson regression. These out-of-sample data correspond to nearly 19 months for the European countries and 15 months for the USA. To evaluate the forecast accuracy, we employ the mean absolute scaled error (MASE) and the root mean squared error (RMSE), which are respectively defined as (Hyndman & Athanasopoulos, 2021, Section 5.8):

$$\mathrm{MASE} = \frac{\frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|}{\frac{1}{N-1}\sum_{i=2}^{N} |y_i - y_{i-1}|},$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2},$$

[11] https://news.un.org/en/story/2024/03/1147287.

Downloaded from https://academic.oup.com/jrsssa/advance-article/doi/10.1093/jrsssa/qnaf196/8369180 by Universita' di Trento - Biblioteca di Ateneo user on 11 December 2025

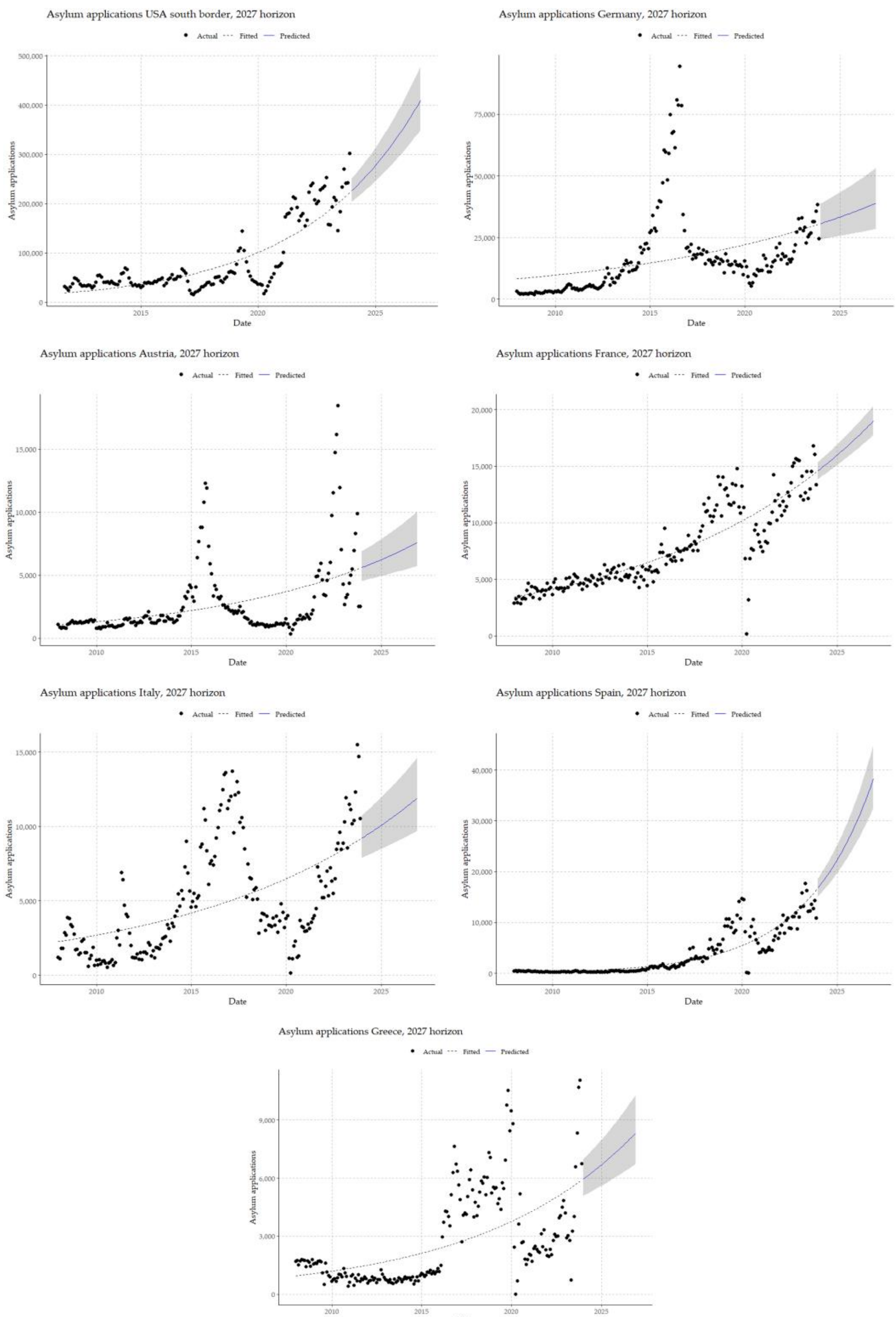

Figure 3. Asylum applications – A 2027 scenario.

Downloaded from https://academic.oup.com/jrsssa/advance-article/doi/10.1093/jrsssa/qnaf196/8369180 by Universita' di Trento - Biblioteca di Ateneo user on 11 December 2025

Table 2. Poisson regression results

Country	Intercept $\hat{\alpha}$	Growth rate $\hat{\beta}$
USA	9.84	0.016
Germany	9.014	0.006
Austria	6.965	0.008
France	8.136	0.007
Italy	7.714	0.007
Spain	5.196	0.02
Greece	6.843	0.009

where y_i is the ith data point and $\hat{y}_i$ is the corresponding predicted value. Figure 4 displays the results.

When comparing forecasts to out-of-sample data, a visual inspection suggests a reasonable fit in the case of the USA, Germany, France, Austria, and Greece, while for Italy the forecasts underestimate the asylum application flows; finally, the forecast for Spain overestimates the observed data. As discussed earlier, the Poisson assumption has been justified by the nature of our count data. However, the target variable—the aggregate monthly asylum-seeker applications - takes very large values, so it is acceptable to approximate it using a non-negative continuous distribution. Among the various size distributions available in the literature (e.g. Kleiber & Kotz, 2003), we use the Gamma distribution. A Gamma density function $f_\Gamma(.)$, is often parameterized by a shape parameter α and an inverse scale parameter β as rate parameter, with the following specification:

$$f_\Gamma(y; \alpha, \beta) = \frac{y^{\alpha-1} e^{-\beta y} \beta^\alpha}{\Gamma(\alpha)}, \quad y > 0.$$

The model is therefore a generalized linear model (GLM) with Gamma-distributed responses: see McCullagh and Nelder (1989, Chapter 8) for details. For the sake of brevity, we have included the plots and other numerical results such as MASE and RMSE in the online supplementary material (Appendix B). The results show that, overall, the Poisson-based approach outperforms the Gamma-based one, though the two provide similar results.[12]

5 Maximum asylum application: a Peaks-over-Threshold approach

In this section, we estimate the maximum monthly asylum-seekers' application per each country by 2027 using the stationary GPD implied by EVT for the distribution of exceedances over a high threshold. Our motivation is to potentially evaluate the expected maximum asylum application across different host countries.

We model the monthly (aggregate) asylum-seekers' flow N_t using the stationary GPD limit. Letting $M_t = \max\{N_1, \ldots, N_t\}$, under the assumptions of the extremal types theorem (Coles, 2001, Theorem 3.1), it holds that:

$$\lim_{t\to\infty} P(M_t \le z) \approx G(z),$$

where

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

[12] We also employed the Holt's linear trend method—an exponential smoothing approach—to evaluate and forecast asylum applications flow. The predictions and the 95% intervals are significantly out of range, with respect to the Poisson or Gamma regressions.

Downloaded from https://academic.oup.com/jrsssa/advance-article/doi/10.1093/jrsssa/qnaf196/8369180 by Universita' di Trento - Biblioteca di Ateneo user on 11 December 2025

Downloaded from https://academic.oup.com/jrsssa/advance-article/doi/10.1093/jrsssa/qnaf196/8369180 by Universita' di Trento - Biblioteca di Ateneo user on 11 December 2025

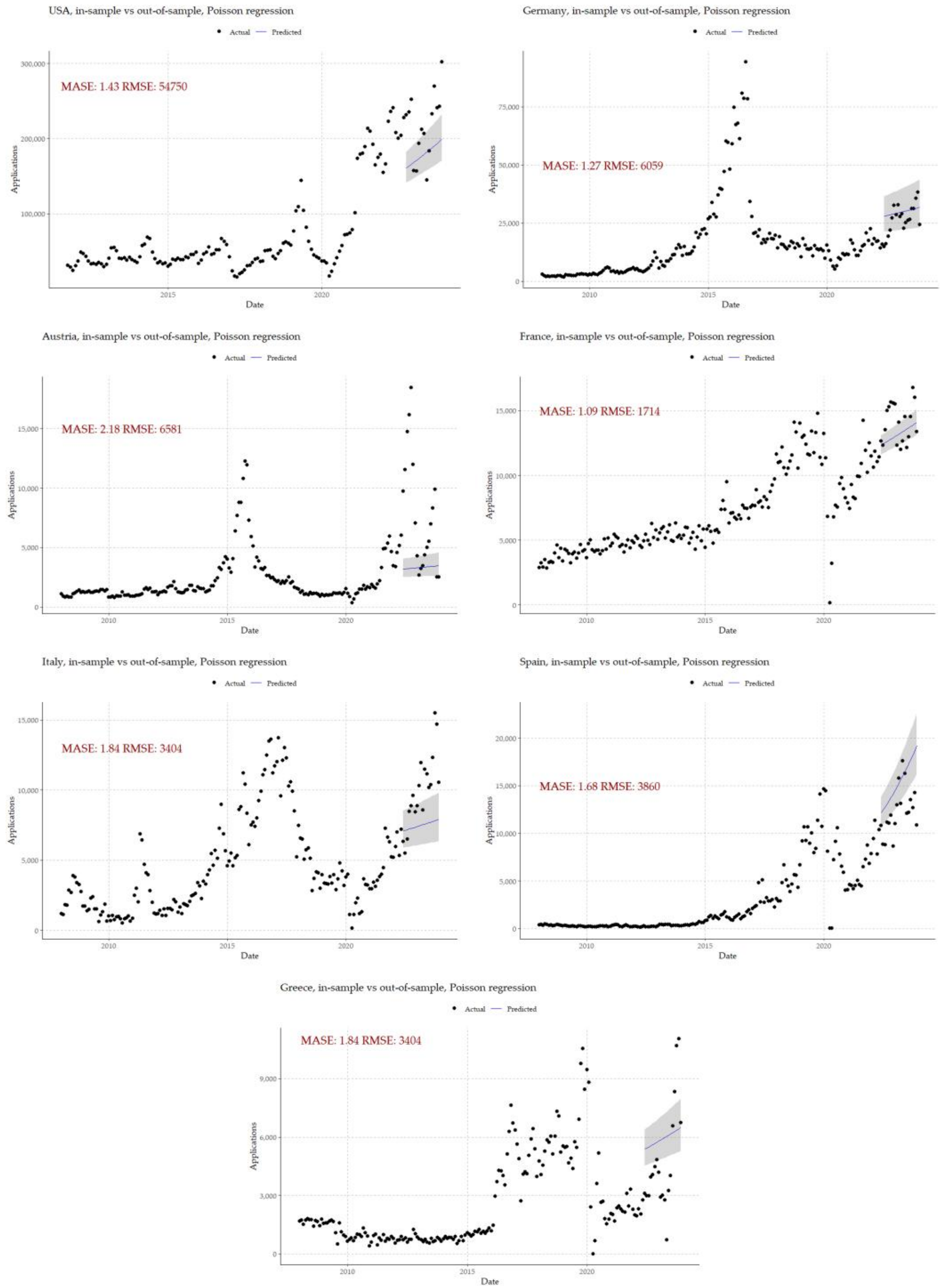

Figure 4. Poisson regression forecast accuracy – In-sample vs. out-of-sample analysis.

is the cumulative distribution function (cdf) of the Generalized Extreme Value distribution with parameters μ, σ, and ξ.

However, from a statistical point of view, it is more convenient to work with exceedances over a high threshold. Accordingly, we consider asylum-seekers' applications as the exceedances over the documented legal migrants' applications. Let $Y_t = N_t - u$ denote the exceedance over a

Table 3. Chosen threshold, *p*-value of AD and ADF tests computed with the corresponding exceedances, MLE vs. Bayesian (mean) GPD parameters

	Threshold			ML GPD			Bayesian GPD	
Country	$\hat{u}$	AD test	ADF	$\hat{\sigma}_{ml}$	$\hat{\xi}_{ml}$	AIC	$\hat{\sigma}_{ba}$	$\hat{\xi}_{ba}$
USA	42,000	0.00	0.35	60,860	0.20	2,153	56,726	0.30
Germany	18,000	0.13	0.53	10,023	0.28	1,288	11,823	0.37
Austria	1,690	0.40	0.11	2,438	0.22	1,484	2,359	0.29
France	13,100	0.39	0.64	2,257	−0.57	330	1,981	−0.35
Italy	10,000	0.79	0.39	2,693	−0.42	444	2,486	−0.28
Spain	11,000	0.35	0.14	3,155	−0.38	334	2,612	−0.08
Greece	3,850	0.49	0.00	2,443	−0.18	969	2,349	−0.11

threshold $u > 0$. Following Coles (2001), as $u \to \infty$, the cdf of Y_t, conditional on $X_t > u$, converges to the GPD with cdf

$$H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-\frac{1}{\xi}}, \quad y > 0,$$

with $(1 + \xi y/\tilde{\sigma}) > 0$, where $\tilde{\sigma} = \sigma + \xi(u - \mu)$ (Balkema & De Haan, 1974; Pickands, 1975). In other words, regardless of the distribution of the original data, under suitable normalization of the threshold excesses, as the threshold u moves towards the right endpoint of the data, the limiting distribution of the exceedances follows the GPD. The shape parameter ξ plays the main role in determining the qualitative behaviour of the GPD. In particular, its sign determines the following two cases:

$$\begin{cases} \text{if } \xi \geq 0: & \text{The distribution of exceedances is unbounded;} \\ \text{if } \xi < 0: & \text{The distribution of exceedances has an upper bound } u - \tilde{\sigma}/\xi. \end{cases} \tag{1}$$

When $\xi = 0$, we need to take the limit $\xi \to 0$, such that the cdf becomes:

$$H(y) = 1 - \exp\left(-\frac{y}{\tilde{\sigma}}\right), \quad y > 0,$$

which corresponds to an exponential distribution with parameter $1/\tilde{\sigma}$. As a result, inference about the distribution of the exceedances is carried out by exploiting the properties of the fitted GPD.

5.1 Stationary GPD: maximum likelihood estimation

Regarding the estimation, the GPD log-likelihood has to be maximized numerically. A preliminary critical decision is the choice of a threshold u such that a balance between bias and variance exists. A too low threshold will lead to large bias, while a too high threshold will leave the investigator with very few exceedances, leading to high variance in our parameter estimates. In this regard, two general approaches exist, one is exploratory analysis to find the appropriate threshold based on model validations, while parameter stability analysis is the other that we rely on this former approach essentially.

Two approaches will be employed for threshold selection. First, we test the null hypothesis that the exceedances follow a GPD by means of the Anderson–Darling (AD) test; see Table 3; then we use the Q–Q plot (Figures 5 to 11) for goodness-of-fit verification. The Q–Q plots display empirical vs theoretical quantiles of the exceedances, in order to check how close the exceedances are to the fitted GPD. The closer the exceedances are located to the 45° line, the better is the fit to the GPD. For instance, in the case of Germany, the chosen threshold is located at 18,000, with

Downloaded from https://academic.oup.com/jrsssa/advance-article/doi/10.1093/jrsssa/qnaf196/8369180 by Universita' di Trento - Biblioteca di Ateneo user on 11 December 2025

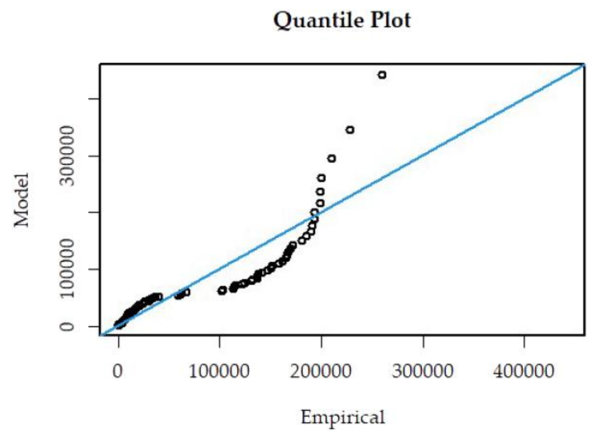

Figure 5. Q-Q plot – USA.

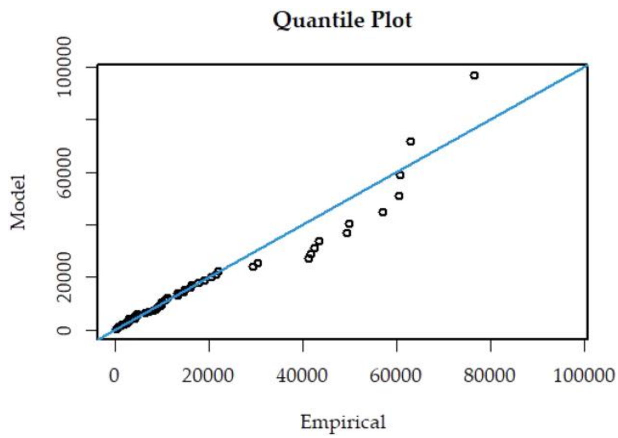

Figure 6. Q-Q plot – Germany.

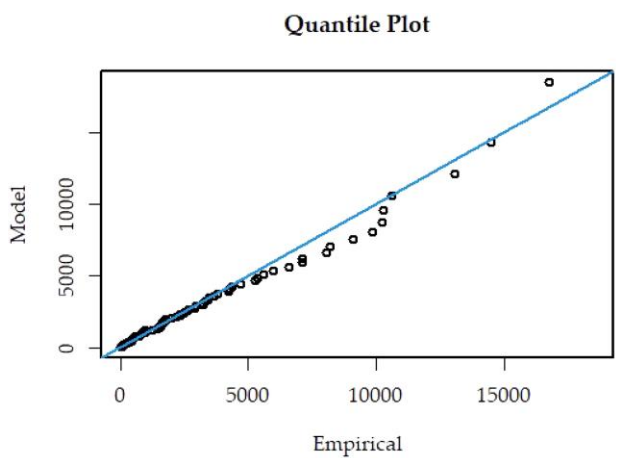

Figure 7. Q-Q plot – Austria.

Downloaded from https://academic.oup.com/jrsssa/advance-article/doi/10.1093/jrsssa/qnaf196/8369180 by Universita' di Trento - Biblioteca di Ateneo user on 11 December 2025

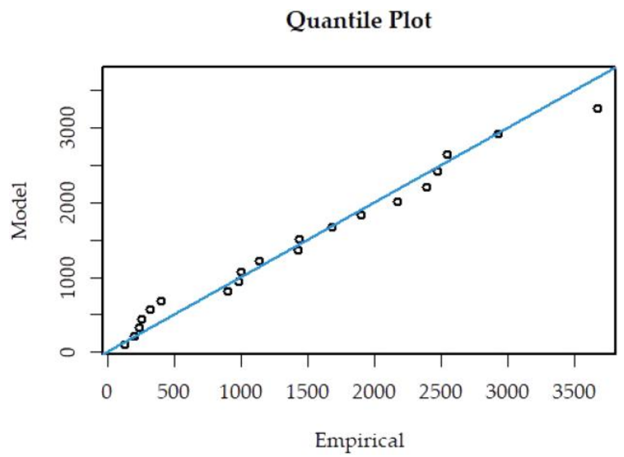

Figure 8. Q-Q plot – France.

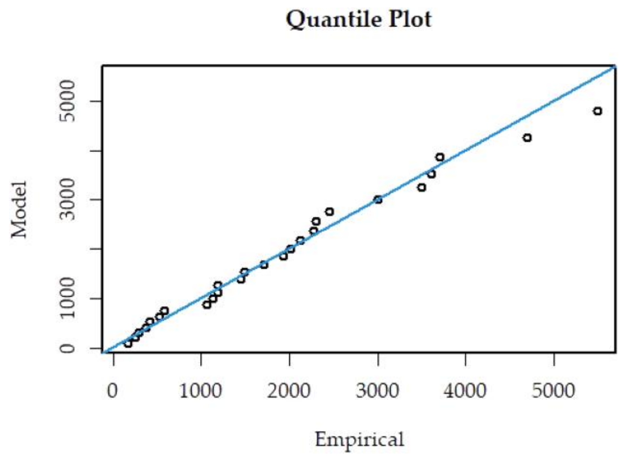

Figure 9. Q-Q plot – Italy.

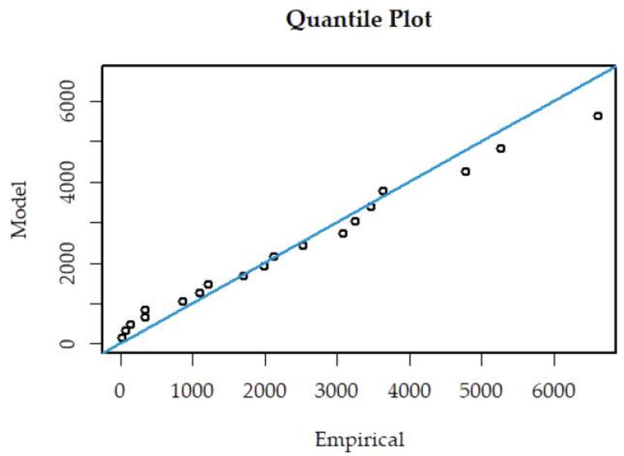

Figure 10. Q-Q plot – Spain.

Downloaded from https://academic.oup.com/jrsssa/advance-article/doi/10.1093/jrsssa/qnaf196/8369180 by Universita' di Trento - Biblioteca di Ateneo user on 11 December 2025

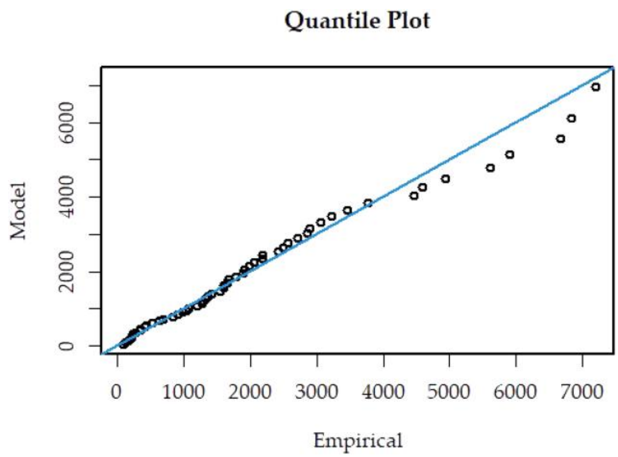

Figure 11. Q-Q plot – Greece.

60 data points above this value. In Figure 6, except for a few large values, the remaining observations lie on the theoretical line. This suggests that overall the fit is good and the data approximately follow the GPD. Table 3 reports a p-value of 0.13, hence we cannot reject the null hypothesis that the exceedances follow a GPD. We use the same approach in choosing a suitable threshold for every other country with the p-values of the AD test reported in Table 3 and the Q-Q plots in Figures 5–11. The only country where the fit is poor, even after experimenting with different thresholds, is the USA. Figure 5 shows that the Q-Q has a sinusoidal shape, with the AD test p-value essentially equal to zero. Table 3 also reports the p-value of the augmented Dickey–Fuller (ADF) test of stationarity for the exceedances. Except for Greece, the exceedances of all other countries are non-stationary (up to 3rd lag).

Table 3 (MLE GPD), summarizes the GPD parameters estimated by MLE GPD based on the selected threshold for each country. Note that the scale ($\hat{\sigma}_{ml}$) and shape ($\hat{\xi}_{ml}$) estimates are all significant (up to 10% level). The positive shape parameter for Austria, Germany and the US exceedances indicates a heavy-tailed GPD, while in case of France, Italy, Greece, and Spain, we are dealing with a bounded GPD (see Eq. (1)). The positive sign of the shape parameter for those three countries somehow resembles the existing immigration policies. In particular, the USA and Germany are among the largest economies in the world, and their 'open-border' policies might indicate that in the future they can indeed attract and accommodate much more refugees than ever. This might be reflected in their positive ξ parameter. It is worth noting that the positive sign of the shape parameter in our study for the three countries does not mean that a potential 6 or 5 million refugees will either apply or enter these countries in the future in each month. In fact, to step out of the theoretical EVT grounds, there is a certain threshold for each of these countries where refugees reception will be tolerated. By the time of conducting this research, there were agreements under negotiations by the EU with some non-EU countries to slow down the refugee flows, like the one-billion Euro deal package between EU and Lebanon, or the Italian-Albanian 5-year deal for hosting the refugees in Albania. The range and efficiency of these policies yet needs to be evaluated in the upcoming years.

Table 4 (MLE GPD) exhibits the distribution of the expected maximum monthly refugee applications across the seven host countries. The estimates are obtained by random draws[13] from a GPD with the estimated parameters (Table 3) and the chosen threshold u; we report the first and third quartile, the median and the upper bound.[14] The findings overall agree with the Poisson regression results in Section 4. In this regard, we would expect that Germany witnesses 27,500 asylum applications per month by 2027, on average. There is a 75% chance that the

[13] Thirty-six draws, bootstrapped 10^5 times.

[14] Recall that, for the three countries with positive ξ, i.e. the USA, Germany, and Austria, there is no upper bound theoretically.

Downloaded from https://academic.oup.com/jrsssa/advance-article/doi/10.1093/jrsssa/qnaf196/8369180 by Universita' di Trento - Biblioteca di Ateneo user on 11 December 2025

Table 4. Maximum asylum applications distribution statistics—MLE GPD vs. Bayesian GPD

	ML GPD				Bayesian GPD			
Country	Q_1	Median	Q_3	Up. bound	Q_1	Median	Q_3	Up. bound
USA	60,000	87,000	140,000	–	59,064	85,606	139,664	–
Germany	21,696	27,450	38,957	–	21,591	27,374	39,544	–
Austria	2,416	3,521	5,666	–	2,401	3,508	5,716	–
France	13,700	14,391	15,262	17,055	13,640	14,317	15,275	18,715
Italy	10,729	11,618	12,826	16,987	10,692	11,568	12,857	18,800
Spain	11,858	12,921	14,400	19,300	11,743	12,761	14,411	33,636
Greece	4,535	5,440	6,840	16,400	4,515	5,419	6,871	19,811

monthly applicants exceed 21,400 and a 25% probability that they exceed (roughly) 39,000. These results show that the recent asylum applications in late 2023 in fact were on the upper range of the distribution. In case of Austria, France, Italy, Spain, and Greece, there is a 50% chance that monthly applications surpass 3,521, 14,400, 11,610, 12,290, and 5,440, respectively.

To give some motivation for the extensions in the following sections, it is worth noting that there are three main potential issues with our results.

1. Compared to most datasets commonly utilized in EVT studies, our data is short. Hence, in Section 5.2, we will employ Bayesian techniques.
2. Unlike most EVT applications that have rather stable underlying DGP, asylum applications rely on a larger number of latent variables. While our EVT forecasts might be breached in the coming years, especially for the countries with and upper bound, the previous analysis suggests that the bulk of the expected asylum applications shall remain between the first and third quartiles for a foreseeable future, unless major shocks such as wars, famine, floods, or other natural disasters happen. Another possibility would be that one or two of these major host countries get into a severe crisis. In that case, not only the expected refugee flows would divert to another host country but also potentially the previous asylum applicants would have to transit to another host. Such occurrences could potentially change the GPD parameters in one country. We would identify such major events as shocks in the distributional properties.
3. Possible non-stationarity of the exceedances might influence our findings. For this reason, in Section 5.3, we use a non-stationary GPD for this inquiry.

5.2 Stationary GPD: Bayesian estimation

As Coles (2001) points out, there are several reasons why a Bayesian analysis of extreme values is interesting. One primary motivation is data scarcity, a setup where information in prior distributions may be a valuable support to empirical evidence. The second advantage of Bayesian techniques is related to forecasting. Since EVT looks to project the extreme cases, posterior distributions of the parameters will be quite useful in that case. The third advantage is the independence of the Bayesian approach from asymptotic issues that may arise with MLE, for example when $\xi < -0.5$. Usually Bayesian analysis relies on Monte Carlo Markov Chain (MCMC) analysis, although other approaches, such as the generalized ratio-of-uniforms method (Wakefield et al., 1991), have been shown to generate robust results.

To conduct the Bayesian analysis in this section, we utilize the `texmex` R package (Southworth et al., 2024). Following Coles (2001), the prior for each parameter θ_i, $i = 1, 2$, where $\theta = \{\sigma, \xi\}$, is assumed to be normal with $\theta_i \sim \mathcal{N}(0, 10^4)$, thus $f(\theta_i)$ is the $\mathcal{N}(0, 10^4)$ density function. The initial values are the MLEs based on an arbitrary threshold, which we chose based on parameter stability discussed in the previous section.[15] The transition (function) from θ_i to θ_{i+1} is a the $\mathcal{N}(\theta_i, 1)$

[15] Note that the threshold (u) is the same in both estimation methods for each country.

Downloaded from https://academic.oup.com/jrsssa/advance-article/doi/10.1093/jrsssa/qnaf196/8369180 by Universita' di Trento - Biblioteca di Ateneo user on 11 December 2025

Downloaded from https://academic.oup.com/jrsssa/advance-article/doi/10.1093/jrsssa/qnaf196/8369180 by Universita' di Trento - Biblioteca di Ateneo user on 11 December 2025

distribution, based on the Metropolis algorithm. We draw nearly 41,000 realizations of the posterior distribution of the parameters. Following the literature, and for computational stability, we use $\phi = \log(\sigma)$. Finally, we study the simulated posterior distribution of the parameters $f(\theta \mid x)$. Now a prediction about the (distribution of) future values can be obtained as follows:

$$\mathrm{P}\{Z \leqslant z \mid x_1, \ldots, x_n\} = \int_\Theta \mathrm{P}\{Z \leqslant z \mid \theta\} f(\theta \mid x)\, \mathrm{d}\theta,$$

where Z denotes the future maximum value, $Z \sim GPD(\sigma_{ba}, \xi_{ba})$. T In this way, we allow for both randomness and uncertainty in future observations. As a result of the uncertainty in the model parameters, the predictions that we get from the Bayesian analysis are expected to be greater than the MLE GPD estimates (Coles, 2001).

Table 3 (right) reports the Bayesian estimates of the parameters for each country.[16] The Bayesian estimates show less variability, i.e. $\hat{\sigma}_{ml} > \hat{\sigma}_{ba}$, while tending towards fatter tails, i.e. $\hat{\xi}_{ml} < \hat{\xi}_{ba}$.

The right side of Table 4 summarizes the distribution of the random draws[17] from a GPD using the mean values of the simulated distributions for the shape and scale parameters. While the first quartile (Q_1) and the median are slightly smaller than the ML estimates, the upper tail represented by the third quartile (Q_3) is greater than its ML counterpart for all countries except the USA. In line with that, Bayesian estimation in all host countries shows a slightly smaller value for the left tail of the data around the first quartile, while the upper tails in the vicinity of the third quartile are greater than the ML counterpart. It is worth noting that Bayesian models are often empirically successful in forecasting (Keilman, 2020; Koop, 2017; Lynch & Bartlett, 2019). Hence, the Bayesian outcomes are likely to be quite reliable. According to the results in this section, Germany will host and receive the highest amount of the asylum applications among the EU hosts. The positive shape parameter for this country reveals that, in case of further migration shocks in the EU, Germany will host the refugees primarily. This, of course, would be consistent with the existing refugee policies and the greater economic power of this country in the EU.

Figure 12 depicts the GPD density estimated by the ML and Bayesian methods, which represents the possible maximum asylum applications in 2027 across different countries. The solid line representing the Bayesian distribution has a longer tail than the ML distribution (dashed line) nearly across all countries. For USA, Germany, and Austria, since there is no upper bound theoretically, we display a truncated plot, focusing only on the leftmost portion of the support. On the other hand, France, Italy, Spain, and Greece have (either ML and Bayesian) density plots with an upper bound limit. All in all, Bayesian and ML GPD predictions are very similar to one another.

5.3 Non-stationary GPD

Based on the augmented Dickey–Fuller test (Table 3), the exceedances across numerous countries, are non-stationary. This might lead to bias in the results, even though in some studies non-stationary models do not outperform their stationary counterparts and produce analogous outcomes. A stationary POT analysis assumes that the parameters are time-independent, or $\theta = \{\mu, \sigma, \xi\}$. When the exceedances are non-stationary, the time dependencies will be modelled directly in the estimation procedure by means of covariates. In this way, the parameters are now time-varying: $\theta_t = \{\mu_t, \sigma_t, \xi_t\}$. We partition the countries into three categories (Table 5): the USA, Germany, France, Spain, and Greece seem to have time-varying threshold μ_t; Austria is deemed to have time-dependent shape ξ_t; and finally Italy, for which both the threshold and the shape parameter are likely to be time-dependent. Time dependencies in the threshold and shape parameters are modelled as follows:

$$\mu_t = \mu_0 + \mu_1 t, \quad \xi_t, \ \text{with } t = \{\text{crisis}, \ \text{non-crisis}\}.$$

[16] Detailed estimation procedures and plots are not reported for space scarcity but are available upon request. As usual, since Bayesian estimation returns a distribution for each of the GPD parameters, we employ the mean as estimate of each parameter in our analysis.

[17] Thirty-six draws, bootstrapped 10^4 times.

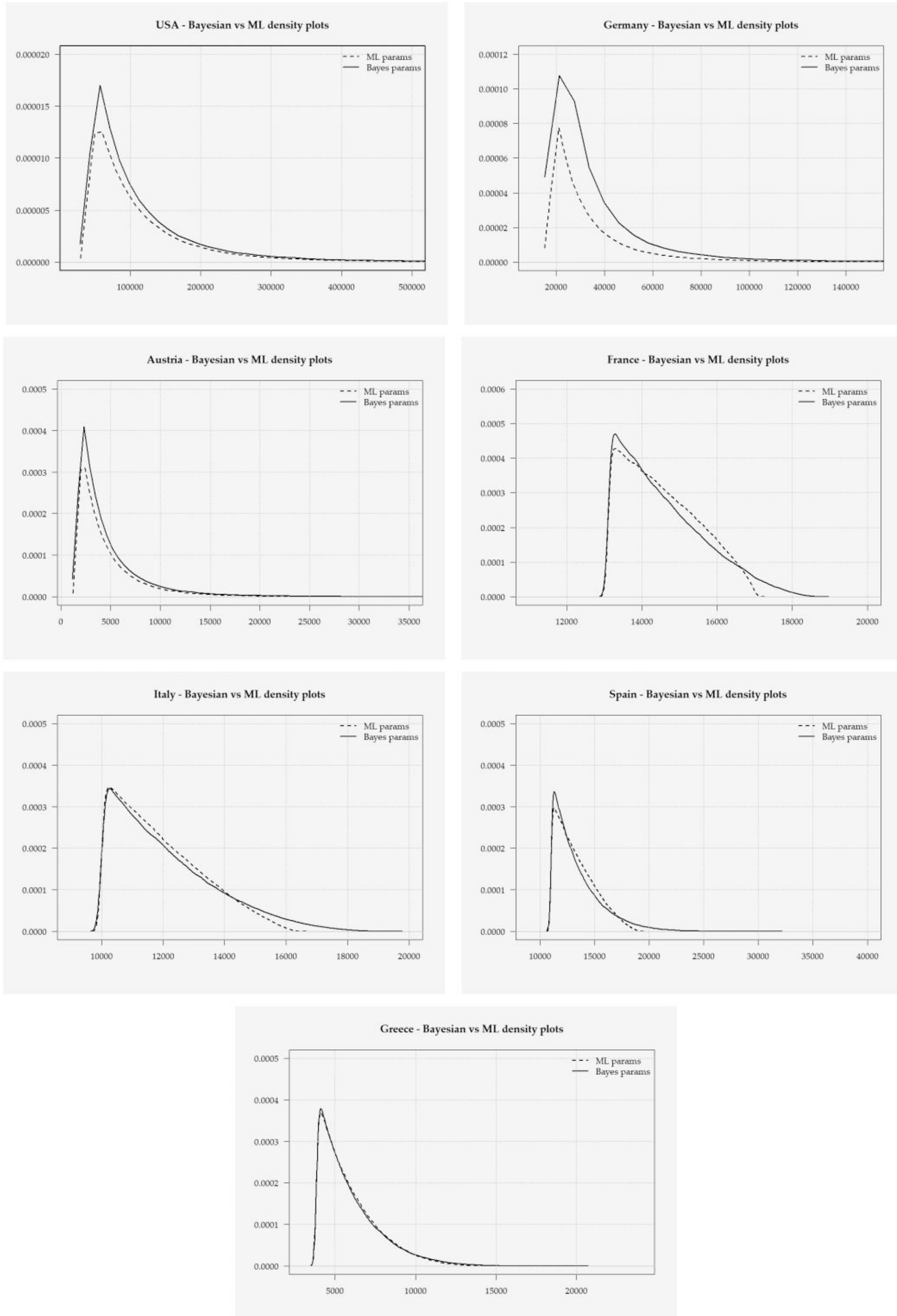

Figure 12. Density plots based on Bayesian and ML estimates.

Downloaded from https://academic.oup.com/jrsssa/advance-article/doi/10.1093/jrsssa/qnaf196/8369180 by Universita' di Trento - Biblioteca di Ateneo user on 11 December 2025

Table 5. Non-stationary GPD specifications

Country	Non-stationary GPD parameters
USA, Germany, France, Spain, Greece	$\mu_t = \mu_0 + \mu_1 t, \sigma, \xi$
Austria	μ, σ, ξ_t = Regime {crisis, non-crisis}
Italy	$\mu_t = \mu_0 + \mu_1 t, \sigma, \xi_t$ = Regime {crisis, non-crisis}

Table 6. Non-stationary GPD modelling

Country	Model	$\hat{\sigma}$	$\hat{\xi}_1$	$\hat{\xi}_2$	SD $\hat{\sigma}$	SD $\hat{\xi}_1$	SD $\hat{\xi}_2$	AIC
USA	$\mu_t = 2 + 1.1t + 0.05t^2$	41,678	−0.78	–	4,844	0.09	–	1,806
Germany	$\mu_t = 10 + 0.1t$	8,975	0.35	–	1,412	0.13	–	2,387
Austria	$\xi_t = 6 * dummy\_crises$	1,899	0.25	0.08	444	0.44	0.07	1,482
France	$\mu_t = 1 + 0.13t$	2,225	−0.44	–	237	0.06	–	1,920
Italy	$\mu_t = 7 + 0.009t$ and $\xi_t = 1 * dummy\_crises$	5,342	−0.39	−0.37	894	0.19	0.09	1,519
Spain	$\mu_t = 18 + 0.00006t^2$	2,979	−0.30	–	886	0.22	–	438
Greece	$\mu_t = 15 + 0.04t$	2,632	−0.28	–	574	0.16	–	794

Since the shape parameter has a pivotal role on the tails, migration crises like those of the 2014–2015 and the 2022 create significant outliers in the asylum applications data. For that reason, we allow the shape parameter to depend on a crisis dummy. This dummy variable takes value 1 when the year is equal to 2014, 2015, 2022, and 0 for other years. The decision to divide the countries into three groups is taken according to the gradual increasing trends in the asylum applications. On one side, countries like the USA, Germany, France, Spain, and Greece exhibit increasing threshold over the sample period. These countries do not show patterns similar to that of Austria, where the 2014–2015 and 2022 migration crises created big outliers. Finally, Italy shows time dependencies in both threshold and shape parameters.[18]

Table 6 reports the non-stationary models and estimated parameters for each country, obtained via the R `extRemes` package by Gilleland and Katz (2016). In the table, t refers to the month t.[19] Given the estimated parameters for each country in Table 6, we calculate the corresponding distribution statistics, similar to those of Table 4. Table 7 summarizes them accordingly. For the countries that have regime-dependent shape parameter, two scenarios are realized, based on the crisis period. Overall, the distribution statistics for Germany, Austria and Spain are in line with the stationary and Bayesian GPD in the previous section.[20] The non-stationary model statistics for the USA are greater than the stationary ones. They show that, by 2027, the number of asylum applications will most likely exceed 386,500 (75% probability) in a given month, while in the upper end of the distribution, there is a 25% chance that applications exceed $\approx 411,000$. Unlike the stationary model case, where we do not find an upper bound for the applications, the non-stationary model has an upper bound of 429,183. It is important to note that no stationary model fits the US exceedances well (based on QQ plots for the residuals or exceedances).

[18] We ran all possible scenarios for all countries. That is, we also let both of the threshold and shape parameters to depend on time, and then choose the model with the best fit. For instance, in case of Germany, a GPD model with only time-varying threshold function outperforms another with time-varying threshold and shape functions, based on both fit and information criteria. We also experimentally chose the coefficients for the time-varying threshold function.

[19] For the European countries, $t = \{1, 2, 3, \ldots, 192\}$, and for the USA, $t = \{1, 2, 3, \ldots, 147\}$, which represents the total number of months (observations) in our sample.

[20] Unfortunately, estimation results are not meaningful in case of Italy. We employed numerous functions for threshold and shape time-variations, including polynomial, quadratic, logarithmic, exponential, linear, and co/sin. But none of these models did converge. Part of this problem relates to our small sample size. Specifically, Italy's asylum applications shows cyclicality over time. We encourage future works to focus on this issue, possibly with larger datasets.

Downloaded from https://academic.oup.com/jrsssa/advance-article/doi/10.1093/jrsssa/qnaf196/8369180 by Universita' di Trento - Biblioteca di Ateneo user on 11 December 2025

Table 7. Non-stationary GPD distribution statistics

Country	Q_1	Median	Q_3	Up. bound
USA	386,529	398,058	411,069	429,183
Germany	22,909	27,153	36,403	–
Austria (Crisis)	2,266	3,171	5,030	–
Austria (Non-crisis)	2,258	3,133	4,843	–
France	16,423	17,151	18,828	20,856
Italy†	–	–	–	–
Spain	11,884	12,929	14,441	20,813
Greece	5,751	6,682	8,047	14,342

Note. †Results are not reported due to inconsistent GPD parameters.

As mentioned earlier, distribution statistics for Germany, Austria, and Spain are similar to those in Table 4. Q_1, in case of Germany, is greater in the non-stationary model compared to the stationary counterpart. That is, most likely (75% probability) the number of asylum applicants will exceed 22,900 cases per month in Germany. The distribution statistics for France and Greece are larger than their stationary (ML and Bayesian) counterparts. For France, on average, the non-stationary model adds 3,000 further cases on top of the stationary model. On the other hand, non-stationary models predict larger asylum applications for Greece, roughly 1,000 additional cases with respect to the stationary models. Since we do not get meaningful results for Italy, we do not report distribution statistics for this country. Finally, while in the stationary analysis the USA, Germany, and Austria do not have an upper bound, in the non-stationary models only Germany and Austria keep that property.

6 Study limitations

The results in Section 4. are obtained from temporally dependent data. Further, our predictions are based on the 'existing patterns' or on the assumption that the natural flow goes on without intervention for the foreseeable future (3 years) at the time of writing the article (mid-2024).

In an effort to obviate the issues caused by temporal dependence, we tried to employ different approaches, possibly better suited to the data. As mentioned repeatedly throughout the article, our data is positive, integer, non-zero with overdispersion, leading to a rather limited number of appropriate methods. We implemented autoregressive (AR), ARIMA, GARCH, Poisson-AR, integer-valued AR (INAR), generalized linear ARMA, integer-valued GARCH (INGARCH), generalized additive models (GAMs), and the Holt's method, but the predictions were disappointing for a number of reasons. For instance, out-of-sample forecasts are out of the expected range, and the parameter estimates are often non-significant. The problem with AR and its variants is the iterative predictions which quickly reverts to the unconditional mean. De-trending the data based on demeaning, natural log changes, or normalization, leads to poor results, such as negative forecasts with huge prediction intervals. In any case, the codes for these models are included in the online appendix. Overall, the quasi-Poisson regression deals with the time-dependence to some degrees. That is the over-dispersion estimation and the time trend in the regression lessen the potential bias. As a result based on the existing data, the quasi-Poisson can provide the most probable scenario. On the other side, since we are unaware of the asylum applications' DGP, the POT approach is a better tool to predict the maximum asylum-applicants regardless of the flow.

7 Conclusion and discussion

This study aims at forecasting the refugee arrivals across seven major host countries, including the USA (south border), Germany, Austria, France, Italy, Spain, and Greece, using Poisson regression and EVT methods. These seven hosts are among the countries that receive the highest amount of the asylum applications. Using publicly available monthly data, for the first-time asylum

Downloaded from https://academic.oup.com/jrsssa/advance-article/doi/10.1093/jrsssa/qnaf196/8369180 by Universita' di Trento - Biblioteca di Ateneo user on 11 December 2025

applicants in the EU, and the monthly apprehensions in the US south border posts, we try to predict the potential asylum-seekers' flow by 2027.

First, we employ Poisson regression to model a dynamic growth analysis across the seven countries. Among them Spain (2%) and then the USA (1.6%) have the highest applications growth rate at which we extrapolate the existing trends to continue until 2027. We also witnessed the massive effect of the local policies (Italian government) or migration shocks (2015 crisis, COVID-19, the Russian invasion of the Ukrainian, Afghan mass refugee, etc.) across countries. We then rely on the generalized Pareto distribution to evaluate the maximum range of the asylum applications in the host countries. Compared to MLE, the Bayesian analysis is indicative of a larger possible number of asylum applications. That is, Bayesian estimates imply more extreme events. Both the ML and the Bayesian estimates show that, except for the USA, Germany, and Austria (where there is no theoretical upper bound on maximum applications), other countries including France, Italy, Spain, and Greece have an upper bound. This can also posit that the first three countries will potentially host a larger portion of the asylum applicants if another migration crisis is set to occur. Finally, given the non-stationarity in our data, we utilize non-stationary GPD, running a dozen models to fit the time-variations in the parameters. The findings agree with both Bayesian and ML GPD results in case of Germany, Austria, and Spain, and are slightly different for others. In comparison with stationary models, non-stationary models predict larger asylum applications for France and Greece.

Overall, our results are in line with one another. This study is limited to those countries where reliable data are publicly accessible. Several other host countries, including Iran, Turkey and Columbia, where there is an on-going refugee crisis, should be investigated in the future. It is worth noting that EVT performs best when large datasets are available. In fact, the greatest challenge when employing EVT methods is directed at the data quantity and quality.

The findings can provide a road map particularly to policymakers across the host countries. Identifying and understanding the main drivers of asylum-seekers' flows in the recent decade is of utmost priority. While some countries can indeed host and accommodate these unexpected guests, other smaller countries struggle. Even though, in this study, we focus on major host countries, other smaller countries can benefit from our predictions. Since asylum-seekers take long roots to get to the host country, for some reasons they might land in a country they were not planning for. As a result, the predictions in this study are insightful to the authorities in other countries as well. The fact that we expect asylum-seekers' flow to break new records by 2027, requires thorough analyses and preparations. These plans can range from halting or slowing refugee flows, to setting emergency budgets for security forces at the borders, or funding training facilities aimed at integrating refugees into the workforce to address labour shortages. Based on initial analyses in our work, three countries, including Germany, Austria, and the USA, do not have an upper bound in receiving asylum-seekers applications, at least theoretically. While this does not imply they can accommodate millions of asylum-seekers, it means that it is highly plausible for them to host more refugees than other host countries in this study, especially at times of social disasters like wars, famine, political unrest, and so on.

In this work, we aim at forecasting the number of asylum-seekers using a statistical approach, based on models whose parameters are estimated from historical data. Such a perspective is mostly useful to handle the effects of significant people influxes on host countries. A more thorough investigation of the causes of large migrations would also be important, both for trying to prevent the events that trigger mass movements and to build models that may incorporate auxiliary information possibly useful to improve the models' performance. While conflicts and social disruptions are likely to be the main issues, extreme poverty should be mentioned as another reason. Building causal models that account for these explanatory variables is a problem that requires further research.

Similarly, it would be relevant to understand whether political issues, in a broad sense, affect the decisions to migrate. Whereas political instability in the origin country resulting in social or economic disruptions is likely to have some effects, the outcomes of authoritarian drifts are more difficult to predict. Also in these cases, incorporating politically related elements into the statistical models may be highly beneficial for a better understanding of the phenomenon and for the development of more accurate forecasting procedures. Political orientation in the host country might also be a factor that affects the influx of migrants. However, this seems difficult to model, both

Downloaded from https://academic.oup.com/jrsssa/advance-article/doi/10.1093/jrsssa/qnaf196/8369180 by Universita' di Trento - Biblioteca di Ateneo user on 11 December 2025

because actions taken by right-wing parties are not always effective and because they may be vastly different depending on the origin countries of the migrants: for example, the current Italian cabinet can be seen as anti-asylum, but kept the doors open for all Ukrainian asylum-seekers. Moreover, from the statistical point of view, in countries like France, Spain, or even Germany, the trend is monotonically increasing, apparently with no political party effect.

As concerns the last two remarks, it should be noted that a major difficulty related to their implementation is data availability. In our opinion, reliable data are a key building block for the use of more sophisticated statistical models, a better interpretation of the phenomenon, and more accurate forecasts.

Acknowledgments

We would like to thank the participants at the 3rd Political Economy workshop, Radboud University, 2024, three anonymous reviewers and an associate editor for helpful comments and suggestions.

Conflicts of interest: None declared.

Funding

This study was funded by the European Union-Next Generation EU, within the scope of the PRIN2022 project 'Prediction and causal inference on the tail index for policy decisions', Grant Number: (2022NA2C8Z) - CUPE53D23006380006.

Data availability

The data and the codes are available at the corresponding author's web page at https://sites.google.com/view/mikenoori/research.

Supplementary material

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

References

Abreu M. C., Lyra G. B., de Oliveira-Júnior J. F., Souza A., Pobočíková I., de Souza Fraga M., & Abreu R. C. R. (2022). Temporal and spatial patterns of fire activity in three biomes of Brazil. *Science of the Total Environment*, *844*(11), 157138. https://doi.org/10.1016/j.scitotenv.2022.157138

Alhawarin I., Assaad R., & Elsayed A. (2021). Migration shocks and housing: Short-run impact of the Syrian refugee crisis in Jordan. *Journal of Housing Economics*, *53*(1), 101761. https://doi.org/10.1016/j.jhe.2021.101761

Araya F., Faust K. M., & Kaminsky J. A. (2021). Agent-based model of hosting communities' perceptions of water and wastewater infrastructure during the German refugee crisis. *Journal of Management in Engineering*, *37*(4), 04021035. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000938

Auerbach J., & Wan P. (2020). Forecasting the urban skyline with extreme value theory. *International Journal of Forecasting*, *36*(3), 814–828. https://doi.org/10.1016/j.ijforecast.2019.09.004

Balkema A. A., & De Haan L. (1974). Residual life time at great age. *Annals of Probability*, *2*(5), 792–804. https://doi.org/10.1214/aop/1176996548

Bee M., Riccaboni M., & Trapin L. (2017). An extreme value analysis of the last century crises across industries in the US economy. *Journal of Economic Dynamics and Control*, *81*(1), 65–78. https://doi.org/10.1016/j.jedc.2017.01.012

Borjas G. J., & Monras J. (2017). The labour market consequences of refugee supply shocks. *Economic Policy*, *32*(91), 361–413. https://doi.org/10.1093/epolic/eix007

Caldera H. J., Wirasinghe S., & Zanzotto L. (2018). Severity scale for tornadoes. *Natural Hazards*, *90*(3), 1051–1086. https://doi.org/10.1007/s11069-017-3084-z

Campolieti M. (2021). Tail risks and infectious disease: Influenza mortality in the US, 1900–2018. *Infectious Disease Modelling*, *6*, 1135–1143. https://doi.org/10.1016/j.idm.2021.09.001

Canfield R. V., Olsen D., Hawkins R., & Chen T. (1980). Use of extreme value theory in estimating flood peaks from mixed populations.

Carammia M., Iacus S. M., & Wilkin T. (2022). Forecasting asylum-related migration flows with machine learning and data at scale. *Scientific Reports*, *12*(1), 1457. https://doi.org/10.1038/s41598-022-05241-8

Downloaded from https://academic.oup.com/jrsssa/advance-article/doi/10.1093/jrsssa/qnaf196/8369180 by Universita' di Trento - Biblioteca di Ateneo user on 11 December 2025

Chen Z. Q., Zheng S. X., Zhou Q., Chen Z. W., & Li X. (2020). Extreme value distribution and dynamic reliability estimation of high-pier bridges subjected to near-fault impulsive ground motions. *Advances in Structural Engineering*, *23*(7), 1367–1382. https://doi.org/10.1177/1369433219894245

Coles S. (2001). *An introduction to statistical modeling of extreme values* (Vol. *208*). Springer.

Cotter J. (2006). Extreme value estimation of boom and crash statistics. *The European Journal of Finance*, *12*(6-7), 553–566. https://doi.org/10.1080/13518470500460111

Dobson A. (1990). *An introduction to generalized linear models*. Chapman and Hall.

Embrechts P., Klüppelberg C., & Mikosch T. (1997). *Modelling extremal events for insurance and finance*. Springer.

Freedman D. (1974). The Poisson approximation for dependent events. *Annals of Probability*, *2*(2), 256–269.

Gilleland E., & Katz R. W. (2016). extRemes 2.0: An extreme value analysis package in R. *Journal of Statistical Software*, *72*(8), 1–39. 10.18637/jss.v072.i08. https://doi.org/10.18637/jss.v072.i08

Hyndman R., & Athanasopoulos G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts.

Jaafar H., Ahmad F., Holtmeier L., & King-Okumu C. (2020). Refugees, water balance, and water stress: Lessons learned from Lebanon. *Ambio*, *49*(6), 1179–1193. https://doi.org/10.1007/s13280-019-01272-0

Jacobson S. B. (2021). *Squatting to make ends meet: Southern Italian migrants and the right to a home in 1970s Italy and West Germany* [PhD dissertation]. Michigan State University.

Jiang Y., & Zhuang Q. (2011). Extreme value analysis of wildfires in Canadian boreal forest ecosystems. *Canadian Journal of Forest Research*, *41*(9), 1836–1851. https://doi.org/10.1139/x11-102

Keilman N. (2020). Uncertainty in population forecasts for the twenty-first century. *Annual Review of Resource Economics*, *12*(1), 449–470. https://doi.org/10.1146/resource.2020.12.issue-1

Keyser A. R., & Westerling A. L. (2019). Predicting increasing high severity area burned for three forested regions in the western United States using extreme value theory. *Forest Ecology and Management*, *432*(42), 694–706. https://doi.org/10.1016/j.foreco.2018.09.027

Kleiber C., & Kotz S. (2003). *Statistical size distributions in economics and actuarial sciences*. Wiley.

Koop G. (2017). Bayesian methods for empirical macroeconomics with big data. *Review of Economic Analysis*, *9*(1), 33–56. https://doi.org/10.15353/rea.v9i1

Lynch S. M., & Bartlett B. (2019). Bayesian statistics in sociology: Past, present, and future. *Annual Review of Sociology*, *45*(1), 47–68. https://doi.org/10.1146/soc.2019.45.issue-1

McCullagh P., & Nelder J. (1989). *Generalized linear models* (2nd ed.). Chapman and Hall.

Morrison J. E., & Smith J. A. (2002). Stochastic modeling of flood peaks using the generalized extreme value distribution. *Water Resources Research*, *38*(12), 41–1. https://doi.org/10.1029/2001WR000502

Nadarajah S., & Ojo O. O. (2023). An extreme value analysis of daily new cases of COVID-19 for sixteen countries in west Africa. *Scientific Reports*, *13*(1), 10814. https://doi.org/10.1038/s41598-023-37722-9

Ngo D. Q., & Nguyen H. C. (2024). Monitoring and analysis of temperature distribution in reinforced concrete bridge box girders in Vietnam. *Case Studies in Construction Materials*, *20*(No. 3), e02857. https://doi.org/10.1016/j.cscm.2024.e02857

Peters M. (2022). Market size and spatial growth-evidence from Germany's post-war population expulsions. *Econometrica: Journal of the Econometric Society*, *90*(5), 2357–2396. https://doi.org/10.3982/ECTA18002

Pickands J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics3*, 119–131.

Rohrbeck C., Eastoe E. F., Frigessi A., & Tawn J. A. (2018). Extreme value modelling of water-related insurance claims. *The Annals of Applied Statistics*, *12*(1), 246–282. https://doi.org/10.1214/17-AOAS1081

Rozo S. V., & Vargas J. F. (2021). Brothers or invaders? How crisis-driven migrants shape voting behavior. *Journal of Development Economics*, *150*(5), 102636. https://doi.org/10.1016/j.jdeveco.2021.102636

Southworth H., Heffernan J. E., Southworth M. H., & Suggests M. (2024). Package 'texmex'.

Tanrıkulu F. (2021). The political economy of migration and integration: Effects of immigrants on the economy in Turkey. *Journal of Immigrant & Refugee Studies*, *19*(4), 364–377. https://doi.org/10.1080/15562948.2020.1810840

Tippett M. K., Lepore C., & Cohen J. E. (2016). More tornadoes in the most extreme US tornado outbreaks. *Science*, *354*(6318), 1419–1423. https://doi.org/10.1126/science.aah7393

Uppal J. Y. (2013). Measures of extreme loss risk–an assessment of performance during the global financial crisis. *Journal of Accounting and Finance*, *13*(3), 105–117.

Ver Hoef J., & Boveng P. (2007). Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology*, *88*(11), 2766–2772. https://doi.org/10.1890/07-0043.1

Wakefield J. C., Gelfand A. E., & Smith A. F. (1991). Efficient generation of random variates via the ratio-of-uniforms method. *Statistics and Computing*, *1*(2), 129–133. https://doi.org/10.1007/BF01889987

Downloaded from https://academic.oup.com/jrsssa/advance-article/doi/10.1093/jrsssa/qnaf196/8369180 by Universita' di Trento - Biblioteca di Ateneo user on 11 December 2025