The Annals of Applied Statistics
2022, Vol. 16, No. 4, 2339–2351
https://doi.org/10.1214/21-AOAS1593

HOW MANY REFUGEES AND MIGRANTS DIED TRYING TO REACH EUROPE? JOINT POPULATION SIZE AND TOTAL ESTIMATION

BY ALESSIO FARCOMENI[a]

Department of Economics and Finance, University of Rome "Tor Vergata", [a]*alessio.farcomeni@uniroma2.it*

We estimate the number of migrants and refugees that died while trying to enter the European Union, during a period of 25 years. Only a subset of attempts with at least one casualty are reported by at least one media source. In order to obtain the estimate, we propose a regression-extrapolation approach, for joint estimation of population size (here, the number of deadly individual or group attempts) and the sum of an accompanying trait (here, the number of deaths) over the population. The trait is measured only for a biased sample of individuals, that are repeatedly observed. Closed-form expressions are derived for the estimator and its standard error. Our findings are that about 40,000 have died from January 1993 to March 2019, during about 5500 attempts to enter the European Union. The number of deaths has been steadily increasing over time, and so has the number of deaths per attempt. About 20% of attempts with at least one casualty have not been recorded by any media source, and slightly less than 10% of deaths have thus been overlooked by media.

1. Introduction. Migration regulation is a controversial and politically sensitive theme. Border control is central in the political debate (Buonfino (2004), Marino (2016), Celata and Coletti (2016)), which is also strongly shaped by media (e.g., Herbers (2016), Vieira (2016)). According to the 1951 Refugee Convention, every human has the right to look for a shelter safe from war, ungrounded persecutions, life threats for their beliefs, political views, or for love. In this work, we focus on the specific phenomenon of refugees and migrants trying to enter "Fortress Europe" (Marino and Dawes (2016), Junemann, Scherer and Fromm (2017)). Restrictive policies of European states make it difficult for many refugees and migrants to enter Europe legally and safely. This makes movement of refugees also a public health issue (Smith and Daynes (2016)). Data from UNITED for Intercultural Action (see Section 2) collect information about attempts to enter the European Union (EU) that lead to at least one casualty, the number of casualties of each event, and the media sources that reported the event.

In our application, unlike similar problems, our aim is not only to estimate a population size (the number events in which there was at least one death in trying to reach the EU), but also to estimate the *total* of an accompanying covariate: the number of deaths. Here, we define total as the sum of the covariate over observed and unobserved elements of the population. To the best of our knowledge, capture–recapture methods have never been used for this scope, and qualitative and quantitative measurements taken during sightings have only been used to improve population size estimates (e.g., by conditioning on covariates).

Population size estimation methods are on the other hand not new to be used for investigating social phenomena. A recent review can be found in Silverman (2020). Examples of population size estimation methods for social phenomena include human rights violations (Mitchell et al. (2013), Sadinle (2018)), modern slavery (Bales, Hesketh and Silverman (2015)), drug abuse (e.g., Farcomeni and Scacciatelli (2013), Overstall et al. (2014), Huggins,

Received September 2020; revised November 2021.
Key words and phrases. Capture–recapture, Chao estimator, Horvitz–Thompson estimator, migration, total estimator.

Yip and Stoklosa (2016); a review can be found in Hay and Richardson (2016)), homelessness (Coumans et al. (2015)), drink and driving (Böhning and van der Heijden (2019)), counting victims during conflicts (Manrique-Vallier, Ball and Sulmont (2019)), adolescent pregnancy (Sukraz et al. (2020)), human trafficking (Bales, Murphy and Silverman (2020)).

In order to perform simultaneous estimation of a population size and sum of an accompanying trait, we propose a regressive approach based on conditioning the distribution of the target trait on the number of times an event is reported. At the first stage, population size is estimated as usual. Note that, unlike many papers focused on multiple systems estimation, we will have a preference for Chao's estimator (Chao (2001), Chao and Colwell (2017)) rather than loglinear models; but any population size estimator can be used at the first stage. At the second stage a regression model is specified for a function of the expected value of the target trait. Extrapolation allows then to obtain a total estimator. The law of large numbers guarantees consistency as long as the population size estimator is consistent and the model is well specified. While the flexible framework allows us to specify models that are more complex than classical polynomial generalized linear models (e.g., include spline smoothers or specify general distributions for the outcome), we will see that in our application there seems to be little sensitivity to model specification.

The rest of the paper is as follows: in the next section, we give a brief description of the available data. In Section 3, we describe our proposed regression-extrapolation approach for two-stage estimation of population size and total, mentioning also some possible direct extensions. A closed-form expression for the standard error of the estimator is given in Section 3.1. A simulation study is reported in Section 4, where we illustrate superiority with respect to two possible natural alternatives. The migrants' data is analysed in Section 5. Finally, Section 6 gives some concluding remarks and outlines routes for further work.

The data, an R implementation of our method, and code for reproducing the data analysis is available online as supplementary material (Farcomeni (2022)).

2. Data description.

Data have been collected by UNITED for Intercultural Action, http://www.unitedagainstracism.org, a European network of more than 560 organizations against nationalism, racism, fascism and in support of migrants, refugees, and minorities. The campaign "Fortress Europe No More Deaths" has collected data about 36,570 deaths of refugees and migrants that occurred while trying to enter Europe irregularly, between January 1993 and March 2019. The list is about single or group attempts to enter EU, in most cases by sea, but also by land (e.g., while trying to climb the fence of one of European enclaves in Africa), or air. The causes of death are not detailed in the data. It is reported by the network that most deaths are due to drowning in the Mediterranean. Others are due to shootings at borders, killings by traffickers, starvation, and other reasons.

From January 1993 to March 2019, $n = 4333$ events were recorded by at least one of a list of sources (newspapers, news channels, NGOs, etc.). We define an event as a single or group attempt to enter the EU, during which at least one death occurred. The target population size $N \geq n$ is the number of events that occurred in the period, also those that were not recorded by any sources. We have recorded n_i, the number of media sources reporting the occurrence of the ith event, for $i = 1, \dots, 4333$. Additionally, for each event, we also have recorded a target trait: the number of deaths X_i. In Figure 1, we report the scatter plot of n_i versus X_i, after jittering both count variables. Our main task is joint estimation of N, the total number of events occurring between January 1993 and March 2019; and $S = \sum_{i=1}^{N} X_i$, the total number of deaths associated to those events.

The median number of casualties is 1, with a mean of 8 and a standard deviation of 28. A total of 57 recorded events are associated with more than 100 casualties, with the maximum of 1050 on April, 19, 2015, due to a major shipwreck in the Sicily canal. Unsurprisingly,

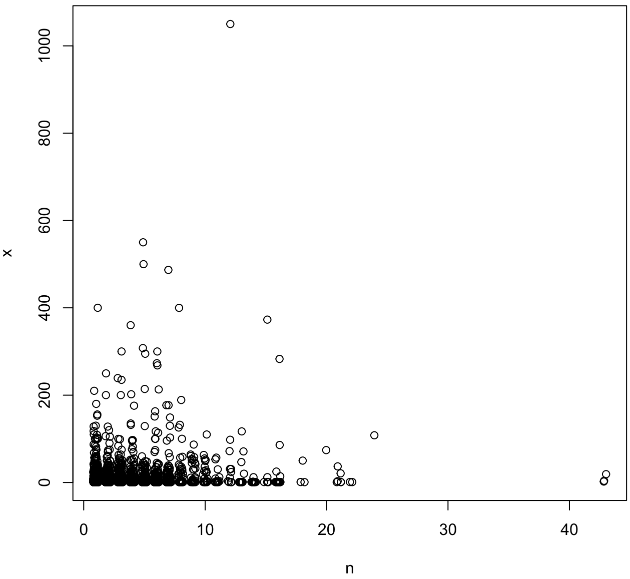

FIG. 1. *Number of deaths as a function of the number of sources reporting each event, between January* 1993 *to March* 2019. *Both count variables are jittered.*

there is a positive association between the number of sources and the number of casualties ($p < 0.001$), but with a small effect: a truncated Poisson regression estimates an increase in the number of sources reporting the event of only 2% per each 10 additional casualties. The figure increases to 3% after removing the outlier.

Obviously, and admittedly, not all events have been recorded by UNITED for Intercultural Action. For those, we do not have information on the number of casualties. If our task was only that of estimating the total number of events (e.g., shipwrecks), we would have several methods at our disposal. On the other hand, our main task is that of estimation of the total number of casualties associated to observed and unobserved events.

3. A regression-extrapolation approach. Let $n_i \geq 0$ denote the number of sources reporting the ith event, $i = 1, \dots, N$, where without loss of generality $n_i > 0$ for $i = 1, \dots, n$. Let also X_i denote a quantitative trait (here, the number of deaths) recorded for the ith event, $i = 1, \dots, N$. We observe couples (n_i, X_i), for $i = 1, \dots, n$. The quantity of primary interest is the vector $(N, S = f(N, X))$ for some known function $f(\cdot, \cdot)$. In our application (and in several other cases), $S = \sum_{i=1}^N X_i$, which we will use henceforth for ease of notation. Generalization to any function $f(\cdot, \cdot)$ for which contributions of observed and unobserved individuals can be separated is straightforward. As noted below, consistency is guaranteed (under certain assumptions) as long as this function only involves a sum of the unobserved values of X_i. Note further that in our application $X_i \geq 1$ for all $i = 1, \dots, N$, since we are interested in events with at least one casualty.

The fundamental problem with our task is that not only N is unknown and we only have observed a biased sample, but also X_i is unobserved for all cases for which $n_i = 0$. We have observed $s = \sum_{i=1}^n X_i$.

An obvious naíve approach would be to proceed by estimating event-specific observation probabilities $p_i = \Pr(n_i > 0)$ and then the sum of interest S through the Horvitz–Thompson (HT) estimator

$$\hat{S}_{\text{HT}} = \sum_{i=1}^n X_i / \hat{p}_i. \tag{3.1}$$

There are several problems with this naive approach, linked mainly to the fact that also X_i is random. Consequently, even if p_i is estimated without bias, the naive HT estimator ignores the (conditional) distribution of X_i, and (3.1) might be improved upon by conditioning. We illustrate this in simulation below.

Another possibility would be to exploit the Royle (2009) method of augmentation in a Bayesian framework. This method involves sampling the unobserved values of X_i conditionally on the current value for the population size N. Clearly, a by-product of this approach is an estimate of $\hat{X}_i$ for each unobserved unit, which can be directly used to obtain $\hat{S}$. A limitation is that this approach involves specification of a parametric marginal distribution for X_i, and while the posterior mean for the population size might not be too sensitive to misspecification, it is intuitive that the posterior mean for the total might be quite sensitive instead. Furthermore, it can be seen in simulation that even when the model is correctly specified, augmentation leads in general to slightly larger estimation errors.

Our proposal stems from the following reasoning: let $s_0 = E[X_i|n_i = 0]$. Under the assumption that $N - n$ is large, the law of large numbers guarantees that $S \cong \sum_{i=1}^{n} X_i + (N - n)s_0$. Consequently, one could proceed by estimating N and s_0 and obtain

$$\hat{S} = (\hat{N} - n)\hat{s}_0 + \sum_{i=1}^{n} X_i. \tag{3.2}$$

In order to put forward a joint estimate for N and s_0, one should deal with the joint distribution of X_i and n_i, $\pi(X_i, n_i)$. The latter can be parameterized as $\pi(X_i|n_i)\pi(n_i)$. Clearly, $\pi(n_i)$ can be used as usual to estimate N *without using X_i as a covariate*. This is particularly convenient, as readily available software can be used for $\hat{N}$. The second term can be used to estimate $s_0 = \int x\pi(x|n_i = 0)\,dx$, if $\pi(x|n_i = 0)$ is an absolutely continuous density, or, as in our case

$$s_0 = \sum_{x=1}^{\infty} x \Pr(x|n_i = 0)$$

if it is discrete. Estimation of s_0 is not at first sight straightforward since it involves conditioning on an event, $n_i = 0$, that is never observed. A simple solution is given by a regression-extrapolation approach. A generalized linear model can be estimated if we assume X_i follows an exponential family distribution conditionally on n_i, and that

$$g\bigl(E[X_i|n_i, \phi]\bigr) = \phi(n_i), \tag{3.3}$$

where $g(\cdot)$ is a known link function and $\phi(\cdot)$ is known up to a finite dimensional parameter (e.g., a polynomial linear transformation $\alpha + \beta_1 n_i + \beta_2 n_i^2$ or even an additive component $s(n_i)$, which can be estimated via splines and then extrapolated after Taylor expansion). Then, by extrapolation, $\hat{s}_0 = g^{-1}(\hat{\phi}(0))$. It is straightforward to check that in case $\phi(n_i) = \alpha$ for all $n_i \geq 1$, then (3.2) recovers the naive Horvitz–Thompson estimator. In our case, for instance, one could estimate a truncated Poisson GLM with log link predicting X_i as a polynomial function of n_i, and then extrapolate a prediction at $n_i = 0$, which can be plugged in (3.2).

Note that extension of this simple approach to more general (continuous or discrete, and even multivariate) outcomes is straightforward. The approach is also easily generalized to take into account overdispersion, for instance, due to unobserved heterogeneity. In general, one could specify

$$g\bigl(E[X_i|n_i, \phi, \alpha]\bigr) = \phi(n_i) + \alpha_i,$$

with $\alpha_i \sim F(\theta)$ for, some known distribution function $F(\cdot)$. Common choices include Gaussian assumptions (e.g., $\alpha_i \sim N(\mu, \sigma^2)$). A more flexible specification could be based on a latent class model, according to which $\alpha_i = \xi_j$, with probability π_j, for $j = 1, \dots, k$; where $\xi_j \in \mathcal{R}$ and k is a user-specified number of latent classes. More simply, one could specify a parametric distribution for X_i, which is flexible enough to capture overdispersion and other features of the data. Examples include the Conway–Maxwell–Poisson distribution (e.g., Anan, Böhning and Maruotti (2017)) or mixtures that can deal with one-inflaction (e.g., Böhning and van der Heijden (2019)).

3.1. *Standard errors.* Approaches for standard error estimation of $\hat{N}$ are well known and depend on the method used. Let now $V(\hat{N})$ denote the variance of the estimator of N used in the first step. In this section, we derive the variance estimator for $\hat{S}$, $V(\hat{S})$. To do so, note that (van der Heijden et al. (2003), Böhning (2008))

$$V(\hat{S}) = E_n\big[V(\hat{S}|n)\big] + V_n\big[E(\hat{S}|n)\big]. \tag{3.4}$$

To derive $V(\hat{S}|n)$, we use the law of total variance once again, and obtain

$$V(\hat{S}|n) = E_{\hat{N}}\left[V\left((\hat{N}-n)\hat{s}_0 + \sum_{i=1}^{n} X_i \middle| \hat{N}\right)\right] + V_{\hat{N}}\left[E\left((\hat{N}-n)\hat{s}_0 + \sum_{i=1}^{n} X_i \middle| \hat{N}\right)\right].$$

After some algebra, it is straightforward to check that the first addend of (3.4) can be consistently estimated by the expression

$$(\hat{N}-n)^2 V(\hat{s}_0|n) + \hat{s}_0^2 V(\hat{N}|n),$$

where $V(\hat{s}_0|n)$ is a direct by-product of the regression estimator (in case of identity transformation it will just correspond to the variance of the intercept in a linear regression model) and $V(\hat{N}|n)$, as discussed above, is a standard by-product of population size estimators.

We now focus on the second addend of (3.4). First, we let $\delta_i = I(n_i > 0)$ denote the indicator that the ith unit is observed, $i = 1, \dots, N$; where without loss of generality $\delta_i = 1$ for $i = 1, \dots, n$. Let also $w_i = \Pr(\delta_i = 1)$, where an estimate $\hat{w}_i$ is also a by-product of classical methods for population size estimation. Furthermore, given that the HT estimator is consistent one can approximate $N = \sum_{i=1}^{n} \delta_i / w_i$.

Now, expressing $\hat{S} = \sum_{i=1}^{N} (\delta_i/\hat{w}_i - \delta_i)\hat{s}_0 + \sum_{i=1}^{N} \delta_i X_i = \sum_{i=1}^{n} (\delta_i/\hat{w}_i - \delta_i)\hat{s}_0 + \sum_{i=1}^{n} \delta_i X_i$, and approximately considering $\hat{s}_0$ as a constant with respect to n, after some algebra one can approximate the second addend of (3.4) by

$$\sum_{i=1}^{n} \left(\frac{\hat{s}_0}{\hat{w}_i} - \hat{s}_0 + X_i\right)^2 \hat{w}_i(1-\hat{w}_i).$$

4. Simulation study. In this section we report a brief simulation study where we compare our approach with two alternatives. Furthermore, we validate the approach for estimation of the standard error of $\hat{S}$ proposed in (3.4).

We proceed by generating N X_i values from a Poisson with parameter λ_x. We then generate n_i, $i = 1, \dots, N$, from a Poisson with parameter

$$\lambda \exp\big(\beta\big(\log(X_i + 1) - \log(\lambda_x + 1)\big)\big). \tag{4.1}$$

After data generation, we ignore cases with $n_i = 0$ and proceed to estimation of $\hat{S}$ using a naive HT approach (ignoring dependence between counts and totals), Royle (2009) augmentation method, and our proposal. Given that the performance of estimators for S depends on $\hat{N}$, to make the comparison more fair we use $\hat{N}$ obtained from augmentation in (3.2). We compute four regression-extrapolation estimates, with $g(\cdot)$ being the identity link function and $\phi(n_i)$ being constant, linear, quadratic, or cubic. The final estimate is selected by maximising the adjusted R^2. The procedure is repeated $B = 1000$ times, for any combination of $N = \{1000, 5000\}$, $\lambda = \{0.5, 1\}$, $\lambda_x = \{0.5, 1\}$, and $\beta = \{-1, 0, 1\}$. Furthermore, we investigate also the case in which the relationship between n_i and X_i is reversed, so that n_i is generated from a Poisson with parameter λ_x and X_i from (4.1). Consequently, we evaluate a total of 48 scenarios.

For each scenario, we report the square root of the Median Square Error (MSE) in Tables 1 and 2.

It shall be noted that our approach always outperforms augmentation and HT when $\beta \neq 0$. When $\beta = 0$ the HT approach is optimal since the model is well specified, but our proposal is associated with only a minor increase in the MSE; while augmentation often leads to a much larger MSE.

In order to validate our estimator for the standard error of $\hat{S}$ we report in Table 3, for each scenario, the standard deviation of the estimates and the mean estimated standard error. For reasons of space, we restrict to linear and quadratic specifications for $\phi(\cdot)$.

It can be seen that average estimated standard errors and actual standard deviations of the estimates are fairly close in all scenarios, taking into account the fact that there is some extra variability due to the fact that S itself is not fixed across iterations within the same scenario.

TABLE 1

Square root of median squared errors for estimating S though our approach ($\hat{S}$), Horvitz–Thompson (HT), and augmentation in simulated scenarios, when n_i is generated from (4.1). Results are based on $B = 1000$ replicates

N	λ	λ_x	β	$\hat{S}$	HT	Augmentation
1000	0.5	0.5	−1	113.34	207.32	170.98
5000	0.5	0.5	−1	342.57	1040.02	946.49
1000	1.0	0.5	−1	59.93	163.62	139.11
5000	1.0	0.5	−1	159.83	794.00	711.95
1000	0.5	1.0	−1	183.95	385.99	314.09
5000	0.5	1.0	−1	535.22	1807.75	1524.90
1000	1.0	1.0	−1	89.52	256.39	179.40
5000	1.0	1.0	−1	292.35	1225.88	1021.40
1000	0.5	0.5	0	42.83	39.22	88.57
5000	0.5	0.5	0	94.10	88.42	189.75
1000	1.0	0.5	0	20.80	19.83	39.67
5000	1.0	0.5	0	46.11	43.83	92.19
1000	0.5	1.0	0	75.66	70.91	174.29
5000	0.5	1.0	0	172.74	163.61	373.43
1000	1.0	1.0	0	36.71	35.81	78.03
5000	1.0	1.0	0	83.25	79.81	182.80
1000	0.5	0.5	1	152.38	175.43	287.87
5000	0.5	0.5	1	452.19	982.48	1308.34
1000	1.0	0.5	1	59.54	151.47	230.33
5000	1.0	0.5	1	182.32	851.15	1042.42
1000	0.5	1.0	1	203.94	224.83	427.74
5000	0.5	1.0	1	681.67	1319.25	1937.42
1000	1.0	1.0	1	80.05	231.77	378.07
5000	1.0	1.0	1	233.39	1302.29	1667.77

TABLE 2

Square root of median squared errors for estimating S though our approach ($\hat{S}$), Horvitz–Thompson (HT) and augmentation in simulated scenarios, when X_i is generated from (4.1). Results are based on $B = 1000$ replicates

N	λ	λ_x	β	$\hat{S}$	HT	Augmentation
1000	0.5	0.5	−1	202.00	242.55	189.57
5000	0.5	0.5	−1	736.95	1225.55	1095.12
1000	1.0	0.5	−1	339.19	485.63	374.40
5000	1.0	0.5	−1	1295.21	2446.80	2189.12
1000	0.5	1.0	−1	115.19	210.41	177.89
5000	0.5	1.0	−1	479.57	1066.24	991.61
1000	1.0	1.0	−1	200.85	422.65	359.43
5000	1.0	1.0	−1	921.98	2132.27	1985.97
1000	0.5	0.5	0	43.39	40.43	89.36
5000	0.5	0.5	0	96.55	89.02	199.37
1000	1.0	0.5	0	81.62	77.98	177.48
5000	1.0	0.5	0	171.71	165.41	395.52
1000	0.5	1.0	0	21.34	20.63	41.76
5000	0.5	1.0	0	48.92	46.88	89.79
1000	1.0	1.0	0	38.61	37.07	80.89
5000	1.0	1.0	0	87.28	83.85	187.17
1000	0.5	0.5	1	275.92	263.24	388.43
5000	0.5	0.5	1	747.32	1304.22	1586.38
1000	1.0	0.5	1	395.67	525.81	777.16
5000	1.0	0.5	1	1010.36	2605.01	3162.70
1000	0.5	1.0	1	68.51	148.70	197.12
5000	0.5	1.0	1	152.16	733.87	842.52
1000	1.0	1.0	1	100.03	297.10	393.97
5000	1.0	1.0	1	230.43	1466.76	1684.33

5. Data analysis. In this section, we estimate the number of casualties of refugees and migrants that occurred in the period 01/1993–03/2019 while trying to enter the European Union.

We define an event as an attempt to irregularly enter the European Union by a single or a group of persons, leading to at least one death. As stated in Section 2, these events have been reported by at least one of several sources during the observation period.

We compare our method with the HT approach (3.1), and with Royle augmentation method. For the HT approach and for our method, the number of events $\hat{N}$ is estimated using Chao lower bound estimator (Chao (1987)), which (using function `closedp.0` in `R` package `Rcapture`, see Rivest and Baillargeon (2019)) is seen to be the preferred estimator in terms of Bayesian Information Criterion (BIC) among a series of possible ones. Chao's estimator is $\hat{N} = 5508$, with a standard error of 77.4. For Royle's method, as noted above, an estimator for the number of events is the primary target. For these data, the posterior mean is 5543, with a 95% credibility interval of $(5365, 5583)$. In both cases, it can be concluded that over the period more than one thousand events might have been overlooked. We also implement a ratio regression approach as in Böhning (2016), based on the Conway–Maxwell–Poisson distribution, to find a similar estimate $\hat{N} = 5535$. Finally, a referee asked to check for one-inflation in the counting distribution (Böhning and van der Heijden (2019), Böhning and Friedl (2021)), which might be ruled out. For instance, for the One-Inflated-Zero-Truncated Negative Binomial model (Godwin (2017)), the proportion of one-inflation is estimated as 10^{-7}, with a standard error of the same magnitude.

In Table 4, we report the estimates for the number of casualties $\hat{S}$ obtained as a by-product of Royle's augmentation method (*Royle)*, with the Horvitz–Thompson (*HT)* estimator, using

TABLE 3

Standard deviation $\mathrm{sd}(\cdot)$ *and average estimated standard error* $\sqrt{V(\cdot)}$ *for linear* ($\hat{S}_1$) *and quadratic* ($\hat{S}_2$) *regression-extrapolation estimators, when* n_i *is generated from* (4.1). *Results are based on* $B = 1000$ *replicates*

N	λ	λ_x	β	$\mathrm{sd}(\hat{S}_1)$	$\sqrt{V(\hat{S}_1)}$	$\mathrm{sd}(\hat{S}_2)$	$\sqrt{V(\hat{S}_2)}$
1000	0.5	0.5	−1	56.19	52.54	88.04	97.33
5000	0.5	0.5	−1	120.84	113.07	173.86	198.21
1000	1.0	0.5	−1	38.17	29.14	47.45	40.58
5000	1.0	0.5	−1	84.47	66.05	102.72	90.41
1000	0.5	1.0	−1	89.83	87.91	121.43	132.46
5000	0.5	1.0	−1	198.34	194.78	255.43	281.18
1000	1.0	1.0	−1	65.71	54.85	77.26	68.88
5000	1.0	1.0	−1	147.37	124.37	169.55	154.78
1000	0.5	0.5	0	89.18	79.32	183.66	184.54
5000	0.5	0.5	0	174.09	168.35	371.09	359.75
1000	1.0	0.5	0	41.75	35.57	60.91	57.17
5000	1.0	0.5	0	92.44	76.94	129.23	120.25
1000	0.5	1.0	0	142.98	133.42	269.15	270.94
5000	0.5	1.0	0	285.27	284.21	535.23	531.92
1000	1.0	1.0	0	66.49	61.37	90.38	88.35
5000	1.0	1.0	0	150.33	132.64	200.87	186.18
1000	0.5	0.5	1	62.70	59.03	157.67	138.00
5000	0.5	0.5	1	139.83	126.52	288.55	265.72
1000	1.0	0.5	1	33.35	25.41	48.37	42.26
5000	1.0	0.5	1	76.00	56.29	103.27	90.21
1000	0.5	1.0	1	96.93	96.99	194.43	182.90
5000	0.5	1.0	1	231.38	211.63	395.99	362.82
1000	1.0	1.0	1	53.49	44.89	68.56	61.89
5000	1.0	1.0	1	122.88	100.13	150.68	133.19

TABLE 4

Estimates of the number of casualties between 1993 *and March* 2019, *with* 95% *CI and BIC. For the Royle method, we report posterior mean and* 95% *credibility interval. Our proposal is based on a polynomial specification for* $\phi(\cdot)$, *where the degree of the polynomial is reported in the second column. LM indicates a linear regression model for the log-counts, NB a negative binomial assumption for the same*

Method	Degree	$\hat{S}$	2.5% CI	97.5% CI	BIC
Royle	–	47,762	47,551	48,028	–
HT	–	46,486	44,005	48,967	–
Proposal(LM)	0	39,286	37,482	41,091	14,309.35
Proposal(LM)	1	39,002	37,215	40,789	14,282.33
Proposal(LM)	2	38,858	37,079	40,637	14,280.13
Proposal(LM)	3	38,615	36,840	40,391	14,273.00
Proposal(LM)	4	38,869	37,077	40,660	14,273.26
Proposal(LM)	5	40,228	38,393	42,063	14,216.61
Proposal(LM)	6	39,792	37,967	41,617	14,222.89
Proposal(NB)	0	45,311	42,981	47,642	18,278.98
Proposal(NB)	1	41,356	39,334	43,378	18,164.44
Proposal(NB)	2	40,815	38,806	42,823	18,157.75
Proposal(NB)	3	40,372	38,347	42,397	18,162.84

a simple polynomial regression on $\log(x_i)$ (*Proposal (LM)*) and using a polynomial GLM after assuming that $x_i - 1$ is a negative binomial (*Proposal (NB)*). For methods based on our proposed approach we report also the degree of the polynomial predictor, where a 0 indicates that the regression model includes only the intercept, a 1 that right-hand side of the model is $\alpha + \beta_1 n_i$, a 2 indicates $\alpha + \beta_1 n_i + \beta_2 n_i^2$, and so on. We additionally report 95% confidence intervals (*CI*) and BIC, where appropriate. Our proposal gives fairly stable results with respect to the degree of the polynomial and model specification. This is confirmed also using other specifications for our regression-extrapolation approach, which are not reported for ease of presentation.

Our final estimate for $\hat{S}$ is smaller than the one obtained using the competitors by about 15%. We speculate this is due to the presence of a fraction of events with many deaths. To support this claim, it shall be noted that when data are restricted to 2015, when a major shipwreck occurred in the Sicily canal, Royle and HT methods indicate that the total number of deaths in that year is about twice the recorded ones. This does not occur with our proposed method (see Figure 2, right upper panel), regardless of how we specify it.

The observed number of casualties per event is estimated as about 7. It shall be noted that the estimated number of casualties per nonreported event is just about 3, which is reasonable as events with lower number of casualties are less likely to be recorded.

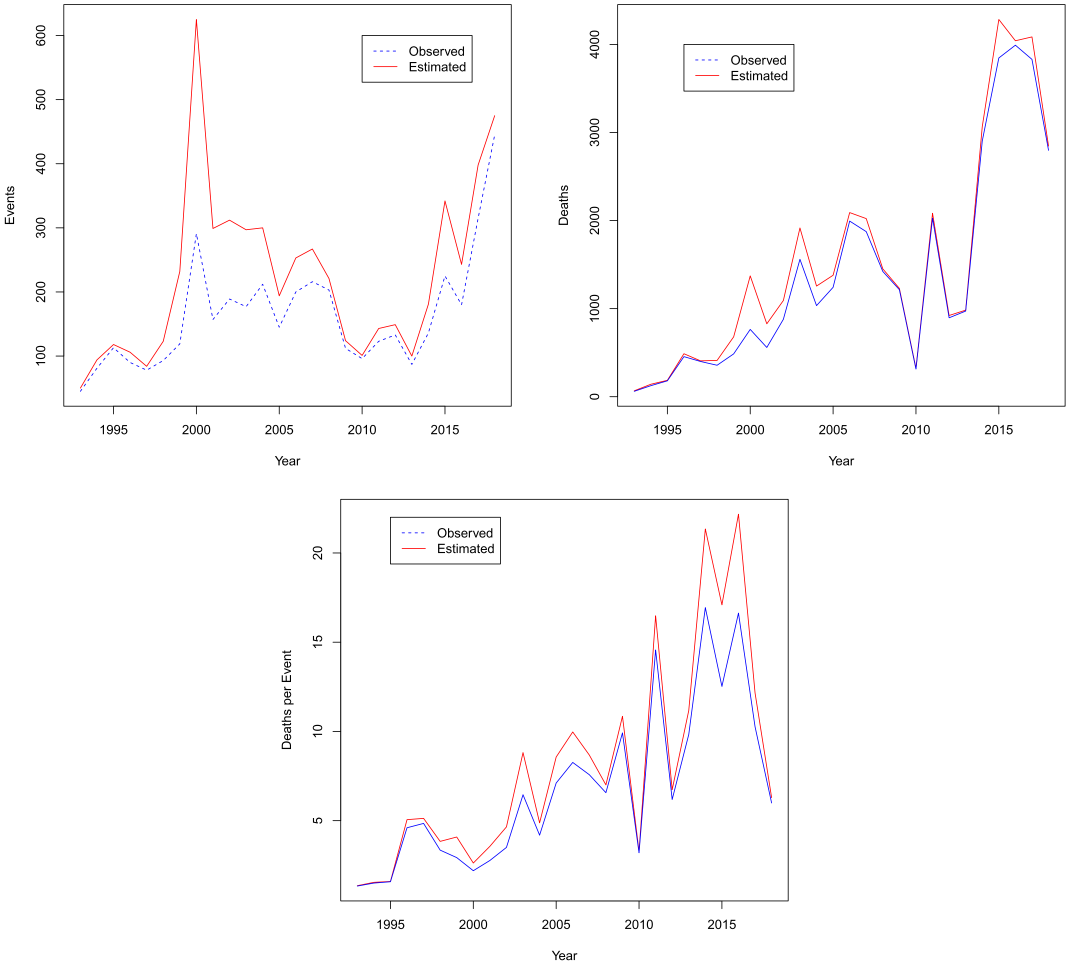

FIG. 2. *Observed and estimated events, deaths, and casualties per event by year. Events are estimated using Chao's estimator, casualties using our proposal with logarithmic transformation, and polynomial regression. The degree of the polynomial is chosen using BIC.*

We conclude with an evaluation of data and estimates over time, for the years 1993 to 2018. We exclude 2019 as we have only partial data for that year. In Figure 2, we report observed and estimated events and deaths by year, and their ratio. Events are estimated using Chao's estimator, casualties using our proposal with logarithmic transformation and polynomial regression. The degree of the polynomial is chosen using BIC. It can be seen that in the last few years the number of deaths has increased both due to an increase in the number of events with at least one casualty, and an increase in the number of deaths per event. The latter has been steadily increasing over time.

In conclusion, it can be estimated that about 40,000 people have died in the 25 years between 1993 and 2019. Number of events per year, number of deaths per year, and number of deaths per event have all been clearly and steadily increasing in the recent past, indicating that the problem is becoming worse over time.

6. Conclusions. It is stated in Last and Spijkerboer (2014) that "existing estimates are insufficient for documenting how many people have died trying to cross the southern EU external borders". Critiques involve both data collection mechanisms and methods (Laczko et al. (2016)). It is therefore widely acknowledged that the true number of deaths is likely to be substantially higher than the official estimates. For example, the official UNHCR estimate for the period 1998–2011 is 13,500 deaths, while our data record 15,727, and our estimate is as high as 18,134. The increase in number of events and deaths after 2012 has likely widened this gap. Indeed, in Esperti (2019) an estimate of 15,062 is reported for the period 2014–2018, while our data record 17,363, and our estimate is 18,320. The main issues with official estimates in our opinion are that they (i) clearly underestimate the number of casualties, (ii) are often not up to date, (iii) are often not accompanied by a confidence interval. This work is an attempt at providing adequate estimates, with sufficiently narrow confidence intervals. Our estimates have good theoretical properties under reasonable assumptions. The use of current media content as source of data partially solves issues with costs and accuracy of data collection mechanisms; and would allow institutions to constantly have timely estimates. Data collection and analysis could in principle be set up in real time.

We estimated that about 40,000 human beings have died trying to enter the European Union, during about 5500 tragic attempts, in the period between January 1993 and March 2019. This is a staggering number. The resulting figure of 1600 per year is furthermore misleading, as recently the number of deaths per year has increased. As we could see from Figure 2, the number of events (e.g., shipwrecks) with at least one casualty has been recently increasing over time *and* the number of deaths per event has simultaneously increased over time. This is a trend that must be reverted. Refugee lives matter, and something must be done to drastically decrease the risks of injuries or death associated with migration or asylum seeking. This should be a widely accepted idea, regardless of the migration policy of each nation, or of the fact that refugee status is finally granted or denied. A general discussion about policy is given in Amenta, Di Betta and Ferrara (2021). A list of possible actions has been put forward by Shilhav (2017), and these include but are not limited to: the design of laws that aim at increasing the benefit that migration can carry for those that are moving across borders as well as for the communities and countries of origin, transit, and destination; the promotion of development aid in the countries of origin; avoidance of agreements that reduce EU responsibility for hosting and protecting asylum seekers and refugees. Importantly, both at a EU and member state level, it would be crucial to provide regular and safe pathways for refugees and migrants, with mechanisms for relocation that respect their needs and choices. The role of nongovernmental actors like Migrant Offshore Aid Station, Emergency, Médecins Sans Frontières, Save the Children, SOS Med, Sea-Watch, and Pro-Activa Open Arms might also be crucial until institutions do not act systemically (e.g., Cusumano (2017), Stierl (2018)).

To the best of our knowledge, we have proposed the first method for joint estimation of a population size and total based on a biased sample. Nonetheless, we have compared with two approaches (HT and Royle method) for which the total is a direct by-product of the population size estimator. These have been seen in simulation to have larger MSE, and to overestimate the total in the real data application when the target trait had a fraction of outliers. HT is a special case of our proposal. Our approach is flexible in that any population size estimation method can be used at the first stage, and it requires only the specification of a regressive equation linking the expected value of the observed trait and the number of repeated observations of each individual. The idea to use regression is clearly also related to the ratio regression method for estimation of population size (Böhning (2016)). All in all, our regression-extrapolation approach relies on assumptions about the distribution of X_i conditional on n_i; but in our application results are stable with respect to this choice. A limitation of our method, which relies on the law of large numbers for its validity, is that it is somehow restricted to estimation of the total, or mean, of a continuous or binary trait. Generalization to general summaries/functions of the target trait is straightforward only as long as they involve a summation might involve quantile regression; but the properties of the resulting estimator are grounds for further work. Another open issue is model choice, in those cases in which some sensitivity is found with respect to model specification. A promising but not yet fully explored approach in the population size estimation area seems to be use of the Focused Information Criterion (Bartolucci and Lupparelli (2008), Farcomeni (2018)).

Acknowledgments. The author is grateful to UNITED for Intercultural Action for access to their database. He is also grateful to a referee, an Associate Editor and the Editor (Professor T. B. Murphy) for constructive comments.

SUPPLEMENTARY MATERIAL

Supplement A: Data and code (DOI: 10.1214/21-AOAS1593SUPP; .zip). Zip file with list of deaths as published by UNITED for Intercultural Action, and R code for each proposed estimator and for reproducing data analysis.

REFERENCES

AMENTA, C., DI BETTA, P. and FERRARA, C. (2021). The migrant crisis in the Mediterranean Sea: Empirical evidence on policy interventions. *Socio-Econ. Plan. Sci.* **78** 101038. https://doi.org/10.1016/j.seps.2021.101038

ANAN, O., BÖHNING, D. and MARUOTTI, A. (2017). Population size estimation and heterogeneity in capture–recapture data: A linear regression estimator based on the Conway–Maxwell–Poisson distribution. *Stat. Methods Appl.* **26** 49–79. MR3610181 https://doi.org/10.1007/s10260-016-0358-7

BALES, K. B., HESKETH, O. and SILVERMAN, B. W. (2015). Modern slavery in the UK: How many victims? *Significance* **12** 16–21.

BALES, K. B., MURPHY, L. T. and SILVERMAN, B. W. (2020). How many trafficked people are there in Greater New Orleans? Lessons in measurement. *J. Hum. Traffick.* **6** 375–387.

BARTOLUCCI, F. and LUPPARELLI, M. (2008). Focused information criterion for capture–recapture models for closed populations. *Scand. J. Stat.* **35** 629–649. MR2468866 https://doi.org/10.1111/j.1467-9469.2008.00604.x

BÖHNING, D. (2008). A simple variance formula for population size estimators by conditioning. *Stat. Methodol.* **5** 410–423. MR2528565 https://doi.org/10.1016/j.stamet.2007.10.001

BÖHNING, D. (2016). Ratio plot and ratio regression with applications to social and medical sciences. *Statist. Sci.* **31** 205–218. MR3506100 https://doi.org/10.1214/16-STS548

BÖHNING, D. and FRIEDL, H. (2021). Population size estimation based upon zero-truncated, one-inflated and sparse count data. *Stat. Methods Appl.* **30** 1197–1217. MR4324408 https://doi.org/10.1007/s10260-021-00556-8

BÖHNING, D. and VAN DER HEIJDEN, P. G. M. (2019). The identity of the zero-truncated, one-inflated likelihood and the zero-one-truncated likelihood for general count densities with an application to drink-driving in Britain. *Ann. Appl. Stat.* **13** 1198–1211. MR3963568 https://doi.org/10.1214/18-AOAS1232

BUONFINO, A. (2004). Between unity and plurality: The politicization and securitization of the discourse of immigration in Europe. *New Polit. Sci.* **26** 23–49.

CELATA, F. and COLETTI, R. (2016). Beyond fortress Europe. Unbounding European normative power and the neighbourhood policy. *Geogr. Compass* **10** 15–24.

CHAO, A. (1987). Estimating the population size for capture–recapture data with unequal catchability. *Biometrics* **43** 783–791. MR0920467 https://doi.org/10.2307/2531532

CHAO, A. (2001). An overview of closed capture–recapture models. *J. Agric. Biol. Environ. Stat.* **6** 158–175.

CHAO, A. and COLWELL, R. K. (2017). Thirty years of progeny from Chao's inequality: Estimating and comparing richness with incidence data and incomplete sampling. *SORT* **41** 3–54. MR3665906

COUMANS, A. M., CRUYFF, M. J. L. F., VAN DER HEIJDEN, P. G. M., WOLF, J. and SCHMEETS, H. (2015). Estimating homelessness in the Netherlands using a capture–recapture approach. *Soc. Indic. Res.* **130** 189–212.

CUSUMANO, E. (2017). Emptying the sea with a spoon? Non-governmental providers of migrants search and rescue in the Mediterranean. *Mar. Policy* **75** 91–98.

ESPERTI, M. (2019). Rescuing migrants in the central Mediterranean: The emergence of a new civil humanitarianism at the maritime border. *Amer. Behav. Sci.* **64** 436–455.

FARCOMENI, A. (2018). Fully general Chao and Zelterman estimators with application to a whale shark population. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **67** 217–229. MR3758763 https://doi.org/10.1111/rssc.12219

FARCOMENI, A. (2022). Supplement to "How many refugees and migrants died trying to reach Europe? Joint population size and total estimation." https://doi.org/10.1214/21-AOAS1593SUPP

FARCOMENI, A. and SCACCIATELLI, D. (2013). Heterogeneity and behavioral response in continuous time capture–recapture, with application to street cannabis use in Italy. *Ann. Appl. Stat.* **7** 2293–2314. MR3161723 https://doi.org/10.1214/13-AOAS672

GODWIN, R. T. (2017). One-inflation and unobserved heterogeneity in population size estimation. *Biom. J.* **59** 79–93. MR3593722 https://doi.org/10.1002/bimj.201600063

HAY, G. and RICHARDSON, C. (2016). Estimating the prevalence of drug use using mark-recapture methods. *Statist. Sci.* **31** 191–204. MR3506099 https://doi.org/10.1214/16-STS553

HERBERS, M. R. (2016). StandUpMigranten: The role of television comedy for the discourse on migration in Germany. *Netw. Knowl.* **9**.

HUGGINS, R. M., YIP, P. S. F. and STOKLOSA, J. (2016). Nonparametric estimation of the number of drug users in Hong Kong using repeated multiple lists. *Aust. N. Z. J. Stat.* **58** 1–13. MR3499148 https://doi.org/10.1111/anzs.12149

JUNEMANN, A., SCHERER, N. and FROMM, N. (2017). *Fortress Europe? Challenges and Failures of Migration and Asylum Policies*. Springer, Wiesbaden, Germany.

LACZKO, F., SINGLETON, A., BRIAN, T. and RANGO, M. (2016). Migrant arrivals and deaths in the Mediterranean: What do the data really tell us? *Forced Migr. Rev.* **51** 30–31.

LAST, T. and SPIJKERBOER, T. (2014). Tracking deaths in the Mediterranean. In *Fatal Journeys. Tracking Lives Lost During Migration* (T. Brian and F. Laczko, eds.) 85–106. International Organization for Migration, Geneva.

MANRIQUE-VALLIER, D., BALL, P. and SULMONT, D. (2019). Estimating the number of fatal victims of the Peruvian internal armed conflict, 1980–2000: An application of modern multi-list capture–recapture techniques. Preprint. Available at arXiv:1906.04763.

MARINO, S. (2016). What are we going to do about them? The centrality of borders in fortress Europe. *Netw. Knowl.* **9**.

MARINO, S. and DAWES, S. (2016). Introduction to fortress Europe: Media, migration and borders. *Netw. Knowl.* **9**.

MITCHELL, S., OZONOFF, A., ZASLAVSKY, A. M., HEDT-GAUTHIER, B., LUM, K. and COULL, B. A. (2013). A comparison of marginal and conditional models for capture–recapture data with application to human rights violations data. *Biometrics* **69** 1022–1032. MR3146797 https://doi.org/10.1111/biom.12089

OVERSTALL, A. M., KING, R., BIRD, S. M., HUTCHINSON, S. J. and HAY, G. (2014). Incomplete contingency tables with censored cells with application to estimating the number of people who inject drugs in Scotland. *Stat. Med.* **33** 1564–1579. MR3240769 https://doi.org/10.1002/sim.6047

RIVEST, L.-P. and BAILLARGEON, S. (2019). Rcapture: Loglinear models for capture–recapture experiments. R package version 1.4-3.

ROYLE, J. A. (2009). Analysis of capture–recapture models with individual covariates using data augmentation. *Biometrics* **65** 267–274. MR2665889 https://doi.org/10.1111/j.1541-0420.2008.01038.x

SADINLE, M. (2018). Bayesian propagation of record linkage uncertainty into population size estimation of human rights violations. *Ann. Appl. Stat.* **12** 1013–1038. MR3834293 https://doi.org/10.1214/18-AOAS1178

SHILHAV, R. (2017). Beyond fortress Europe: Principles for a humane EU migration policy.

SILVERMAN, B. W. (2020). Multiple-systems analysis for the quantification of modern slavery: Classical and Bayesian approaches. *J. Roy. Statist. Soc. Ser. A* **183** 691–736. MR4114463

SMITH, J. and DAYNES, L. (2016). Borders and migration: An issue of global health importance. *Lancet Glob. Health* **4** 85–86.

STIERL, M. (2018). A fleet of Mediterranean border humanitarians. *Antipode* **50** 704–724.

SUKRAZ, B., THAKKINSTIAN, A., OKASCHAROEN, C., AEKPLAKORN, W., SAEJENG, K., BÖHNING, D. and ARUNAKUL, J. (2020). Estimation of the adolescent pregnancy rate in Thailand 2008–2013: An application of capture–recapture method. *BMC Pregnancy Childbirth* **20** 120.

VAN DER HEIJDEN, P. G. M., BUSTAMI, R., CRUYFF, M. J. L. F., ENGBERSEN, G. and VAN HOUWELINGEN, H. C. (2003). Point and interval estimation of the population size using the truncated Poisson regression model. *Stat. Model.* **3** 305–322. MR2012155 https://doi.org/10.1191/1471082X03st057oa

VIEIRA, I. (2016). The construction of the Mediterranean refugee problem from the Italian digital press (2013–2015): Emergencies in a territory of mobility. *Netw. Knowl.* **9**.