Smoothed population size estimation of human rights violations in Colombia 1985–2018

Alessio Farcomeni

Journal of the Royal Statistical Society Series A: Statistics in Society, 2026, 00, 1–22
https://doi.org/10.1093/jrsssa/qnag077
Published: 7 July 2026
Original article

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

Smoothed population size estimation of human rights violations in Colombia 1985–2018

Alessio Farcomeni

Department of Enterprise Engineering, University of Rome Tor Vergata, Via del Politecnico, 1, 00133 Rome, Italy
Address for correspondence: Alessio Farcomeni, Department of Enterprise Engineering, University of Rome Tor Vergata, Via del Politecnico, 1, 00133 Rome, Italy. Email: alessio.farcomeni@uniroma2.it

Abstract

We estimate the extent of conflict-related homicides, kidnappings, and forced recruitments in Colombia between 1985 and 2018; also at the gender, department, and year level. In order to do so, we first introduce general principles for stratified estimation. We then implement Ridge and fused Ridge type penalties to smooth generalized Chao and Zelterman estimators. The fused Ridge type penalties are particularly apt at pooling information across time and space, after specification of an opportune adjacency matrix. Penalty parameters are selected through an empirical Focused Information Criterion strategy. A simulation study illustrates how penalized estimators outperform competitors in terms of mean squared error, and stability, both overall and at the stratified level. We estimate more than 640,000 conflict-related homicides, almost 30,000 forced recruitments, and more than 60,000 conflict-related kidnappings in the period; with clear temporal and spatial patterns. The Antioquia department, and years between 2000 and 2005, were most severely affected by violence. We also estimate that after the peace agreements of 2016 in some departments the conflict experienced a slight resurgence in killings and, more broadly, in human rights violations.

Keywords Chao estimator, human rights violations, penalized likelihood, population size estimation, Zelterman estimator

1 Introduction

The study of human rights violations in conflict settings is an integral part of the response to crises (Checchi & Roberts, 2008), it can guide the humanitarian response, and be used to exert pressure on belligerents. After the conflict, it is helpful for policy evaluation, accountability, and memory; which is a crucial pillar for a durable peace. However, documented data on human rights violations usually provide only a partial account of the violence that occurred (Ball & Price, 2019; Price & Ball, 2014). Population size estimation methods are frequently used to correct for undercount, through modelling observation patterns of individuals across multiple lists. Reviews of population size estimation methods for conflict data can be found, for example, in Seybolt et al. (2013), Ball and Price (2019), and Manrique-Vallier et al. (2022); but see also Lum et al. (2013), Mitchell et al. (2013), and Sadinle (2018). Some examples of applications of these models include Bales et al. (2015), Manrique-Vallier et al. (2019), Farcomeni and Maruotti (2024), and Jamaluddine et al. (2025).

Received: March 21, 2026. **Revised:** May 27, 2026. **Accepted:** June 8, 2026

We wish to study three types of human rights violations (conflict-related homicides, conflict-related kidnappings, and forced recruitments) that occurred in Colombia between 1985 and 2018. The conflict in Colombia has been ongoing since 1964. During the last six decades, multiple armed groups—including criminal syndicates, guerrilla groups, paramilitaries, and state forces—have used violence in pursuit of organizational and political goals (Karl, 2019). In doing so, different armed groups have engaged in broad repertoires of lethal and non-lethal violence (Gutiérrez-Sanín & Wood, 2017), mainly against civilian populations (Comisión de la Verdad, 2022). The 2016 peace accords signed between the *Fuerzas Armadas Revolucionarias de Colombia—Ejército del Pueblo* (FARC), the largest guerrilla group, and the Colombian government established the country's transitional justice mechanism, the *Sistema Integral de Verdad, Justicia, Reparación y No Repetición* (SIVJRNR). As part of the transitional justice process, a collaborative project was undertaken between the *Jurisdicción Especial para la Paz* (JEP), the *Comisión para el Esclarecimiento de la Verdad, la Convivencia y la No Repetición* (CEV)—two of the bodies that form part of the SIVJRNR—and the Human Rights Data Analysis Group (HRDAG) to gather statistical information about victims of certain types of human rights violations (Amado et al., 2025) occurring during the period 1985–2018. Subsequently, anonymized data from the project have been made publicly available for analysis; with individual observation patterns across $S = 34$ sources, and some additional information. More details will be given in Section 2. Our main goal is to estimate the total number of victims and its spatio-temporal distribution, accounting for incomplete reporting.

The most commonly used approaches for population size estimation in this context involve incomplete log-linear models (Cormack, 1989; Fienberg, 1972; Madigan & York, 1997; Overstall et al., 2014; Rivest & Baillargeon, 2007). Log-linear formulations can directly model list dependence and, under identifying assumptions (e.g. the absence of the highest-order interaction), provide an estimate for the number of individuals that were not reported by any list. With $S = 34$ lists, the full contingency table has 2^{34} possible capture patterns, most of which are empty. Even models restricted to pairwise interactions become unstable, and may also not capture the true dependency patterns. An elegant alternative is proposed in Manrique-Vallier (2016) in a Bayesian framework, where lists are assumed to be independent conditionally on a discrete latent variable. A Dirichlet process mixture avoids the need to specify the number of latent classes. While effective, the resulting non-parametric latent class model cannot explicitly take into account covariates, and it is computationally intense. Similarly to many other approaches, it might yield unstable/extreme estimates. Furthermore, it is often sensitive to list selection (that is, estimates might change substantially if one list is removed). A direct comparison with our proposals in terms of mean squared errors will be given in the simulation study below.

A key part of the study of human rights violations in conflicts involves an examination of victimization estimates stratified by covariates such as time, space, age, sex, and race/ethnicity; in order to identify patterns, assess responsibility, and modulate intervention (e.g. Nason & Bailey, 2008; Racek et al., 2026). Unfortunately, as in our motivating application, stratified population size estimates are often unstable and/or extreme. Instability, even for the overall estimate, is often observed due to sparsity and relatively low sampling fraction (i.e. large discrepancy between observed and true population sizes). Some attempts to stabilize the overall estimates are available in the literature, involving, for example, thresholding (Silverman, 2020); and Bayesian methods in general are partly protected by the prior input. In this work, we focus on generalized Chao and Zelterman estimators (Böhning, 2010; Böhning & van der Heijden, 2009; Böhning et al., 2013; Farcomeni, 2018). Both families of estimators are based on the subset of individuals observed at most twice, and the same truncated likelihood specification. Truncation brings about some robustness to misspecification and low sensitivity to list selection. Generalized estimators can take into account covariates. Chao and Chao-like estimators are often conservative in finite samples under any form of unobserved heterogeneity; a desirable guarantee in the study of human rights violations, which directly tackles possible critiques about exaggerated estimates. Zelterman type estimators are described for completeness, but we will see that they might not be completely reliable in the present setting. They are nonetheless commonly useful in other contexts (e.g. with low levels of unobserved heterogeneity).

We give two contributions to the methodological literature before applying our proposed methods to the Colombian case. First, we introduce simple general principles for stratified estimation, whose only requirement is that the method used for population size estimation is compatible with the Horvitz–Thompson estimator. Secondly, we derive Ridge and fused Ridge penalized generalized Chao and Zelterman estimators. The penalty is introduced in order to shrink the covariate effects towards zero (Ridge penalty) or towards each other (fused Ridge penalty), thereby reducing standard errors and greatly improving stability; especially in sparse multi-list settings, and when considering stratified estimates. Fused Ridge type penalties (e.g. Goeman, 2008; Lettink et al., 2023) are particularly apt at pooling information across time and space. Temporal smoothing is introduced by penalizing the squared differences between parameters referring to consecutive time occasions. Spatial smoothing is introduced by shrinking towards each other parameters corresponding to neighbouring areas, after specification of an opportune adjacency matrix. As a result, information is directly pooled across neighbouring areas and adjacent time periods. Notably, penalized Chao and Zelterman estimators share several desirable properties with non-penalized methods. Importantly, the penalized Chao estimator is still conservative in finite samples under the same conditions of the non-penalized one. The penalty parameter will be selected by minimizing a proxy for the expected mean squared error, namely, an empirical version of the Focused Information Criterion.

A simulation study will be used to argue how penalized estimators outperform competitors in terms of estimation error and stability, both overall and at the stratified level. In the real data example, standard estimators (including generalized Chao and Zelterman estimators) are rather unstable and provide extreme estimates after stratification over time and/or space. However, penalized estimators show no instability or extreme values even after stratification (e.g. for department-year-specific estimates).

The rest of the article is as follows: in the next section, we give more details on our motivating data set. In Section 3, we set up the notation and present our proposed penalized Chao and Zelterman estimators for spatio-temporal analysis. We evaluate the performance of our proposals in a simulation study in Section 4. In Section 5, we evaluate the total and stratified number of conflict-related homicides, kidnappings, and forced recruitments in Colombia between 1985 and 2018. Some concluding remarks are reported in Section 6. An R implementation of our method and the code for reproducing the data analysis is available online as supplementary material.

2 Data

The JEP-CEV-HRDAG joint project received more than 12.8 million records of victims, and finally focused on disappearance, homicide, kidnapping, and forced recruitment. This analysis leverages data on homicides, kidnappings, and forced recruitment for the period 1985–2018; we do not consider disappearance due to the high likelihood of overlap with the other human rights violations (e.g. many people that disappeared were probably killed) and lack of cross-violation identifiers.

A brief description of data preparation follows, but the reader is referred to Amado et al. (2025) for complete details. More than 100 databases were contributed by more than 40 among state institutions, civil society organizations, and victim groups. The initial work of the group involved data cleaning, standardization, deduplication. The sources were also merged and/or discarded, to a final $S = 34$ lists, whose documentation systems are heterogeneous (including government registries, forensic systems, NGO databases, and human rights monitoring initiatives). Individuals could be listed in any (at least one) of these. A preliminary work involves matching, dealing with uncertainty in the identification of the same individual by more than one source (e.g. Sadinle, 2018; Tancredi et al., 2020). The linkage process combined deterministic and probabilistic techniques. A blocking step was used to group potentially matching records into candidate sets, in order to reduce computational complexity. Within blocks, features capturing similarities (names, demographic attributes, and event information) were used to predict matches. For each individual reported by at least one source, we finally observe a binary capture vector indicating the presence or absence in each list.

While by definition all documented instances of forced recruitment are related to the armed conflict, in the case of homicide and kidnapping not all sources focused exclusively on instances of violence that occurred in the context of the conflict. For example, the *Instituto Nacional de Medicina Legal y Ciencias Forenses* is a public institution charged with conducting forensic investigations into homicides of all

Table 1. Descriptive statistics of homicide, kidnapping, and forced recruitment between 1985 and 2018 from the JEP-CEV-HRDAG joint project (v2). f_j denotes the number of individuals observed exactly j times, and $f_{>j}$ the number observed more than j times.

Violation	n	f_1	f_2	f_3	f_4	f_5	$f_{>5}$	% Male
Homicide	441,869	185,140	92,071	90,997	38,369	20,412	14,880	90.5
Kidnapping	50,864	16,559	13,311	6,485	6,687	3,676	4,146	77.7
Forced recruitment	19,223	9,275	5,238	2,237	1,436	736	301	69.2

kinds, and their data about homicide victims reflects this broad mandate. As such, the data compiled by the JEP-CEV-HRDAG Joint Project includes a binary conflict-relatedness indicator. In some cases, it was not possible to determine whether an event occurred in the context of the armed conflict (17% of the homicides, 9% of the kidnappings). In these cases, the indicator was assigned a missing value and, along with other variables considered in the database, was imputed using Multiple Imputation via Chained Equations (MICE) (van Buuren & Groothuis-Oudshoorn, 2011). MICE was also used to complete records with respect to sex, ethnicity, age, perpetrator, municipality. The imputation models assumed a missing at random (MAR) mechanism and relied on both observed variables and a set of support variables generated through a neural-network classifier trained on the observed data.

The JEP-CEV-HRDAG joint project finally published 100 multiply imputed data sets. Our primary analysis will be based on data from a single replicate in order to simplify the exposition and interpretability. A sensitivity analysis that considers all replicate files will be reported in Section 5 and will show that the presented results are not affected by this choice. The JEP-CEV-HRDAG Joint Project data is available from https://hrdag.org/jep-cev-colombia/. The publicly released data set contains no identifying information for individuals. This analysis uses the most recent Version 2 of the data. Table 1 summarizes the available data on homicide, kidnapping, and forced recruitment. In the table, we denote as f_j the number of individuals observed exactly j times in any of the lists, as $f_{>j}$ the number of individuals observed $j+1$ times or more, while $n = f_{>0}$ is the total number of individuals observed at least once.

3 Penalized Chao and Zelterman estimators

In this section, we first review generalized Chao and Zelterman estimators and establish notation. We then introduce a framework for stratified population size estimation and propose penalized estimators that stabilize estimates across covariates, time, and space.

Suppose n subjects are observed in at least one of S lists. For the purposes of Chao and Zelterman methodologies, we select the n_{1+2} subjects that are observed at most twice. For ease of notation and without loss of generality, we assume that the first n_{1+2} subjects are observed at most twice, while for $i = n_{1+2} + 1, \ldots, n$ subjects are observed three or more times. The outcome is a binary indicator X_i, where $X_i = 1$ if the ith subject has been observed in exactly two lists. Hence, $X_i = 0$ indicates that the ith subject appears in exactly one of the S lists for $i = 1, \ldots, n_{1+2}$.

The population is made of $N \geq n$ individuals, and if $N > n$, there are $N - n$ subjects never included in any list. For the observed subjects, we possibly collect a vector of individual covariates Z_i. We also record year y_i and department d_i. In order to model spatial dependence, we fix a department adjacency matrix C.

Generalized Chao and Zelterman estimators (Böhning & van der Heijden, 2009; Böhning et al., 2013; Farcomeni, 2018) are based on (i) a conditional Poisson approximation and (ii) a truncated likelihood. That is, it is assumed that $\Pr(X_i = 0 \mid Z_i)$ is proportional to the probability that a Poisson with parameter λ_i yields a 1, and $\Pr(X_i = 1 \mid Z_i)$ is proportional to the probability that a Poisson with parameter λ_i yields a 2. Clearly, λ_i can be interpreted as the expected number of times the i-th individual would be observed. It is now a matter of parameterizing $\lambda_i > 0$ in terms of the covariates, e.g. through a log-linear model of the kind

$$\log(\lambda_i) = \beta_0 + \beta' Z_i. \tag{1}$$

After some algebra, the truncated likelihood can be expressed as

$$\prod_{i=1}^{n_{1+2}} \left(\frac{2}{2+\lambda_i}\right)^{1-X_i} \left(\frac{\lambda_i}{2+\lambda_i}\right)^{X_i}. \tag{2}$$

Parameter estimates $\hat{\beta}_0$ and $\hat{\beta}$ are obtained through optimization of (2) after plug-in of (1). Expected counts $\hat{\lambda}_i$ are computed after plug-in of $\hat{\beta}_0$ and $\hat{\beta}$ in (1). The generalized Chao estimator is obtained as

$$\hat{N}_C = n + \sum_{i=1}^{n_{1+2}} \frac{2}{2\hat{\lambda}_i + \hat{\lambda}_i^2}. \tag{3}$$

The generalized Zelterman estimator on the other hand corresponds to an Horvitz–Thompson type estimator, with expression

$$\hat{N}_Z = \sum_{i=1}^{n} \frac{1}{1 - \exp(-\hat{\lambda}_i)};$$

where the denominator can be seen to correspond to the estimated inclusion probability for the ith individual.

These estimators, with and without covariates, have some desirable properties. Zelterman estimator is partially robust to misspecification, since it relies on Poisson assumptions only on the singletons and doubletons. Chao's estimator is also robust to misspecification, provides a lower bound for the true population size (i.e. the bias of $\hat{N}_C$ is non-positive), and it is unbiased if all individual heterogeneity is captured by the covariates. More precise statements are given below.

If no covariates are used, $\hat{N}_C$ and $\hat{N}_Z$ reduce to the usual Chao and Zelterman estimators, whose expressions are

$$\hat{N}_C = n + \frac{f_1^2}{2f_2},$$

and

$$\hat{N}_Z = \frac{n}{(1 - e^{-2f_2/f_1})}.$$

3.1 Stratified estimators

Suppose that a researcher is (also) interested in estimators conditionally on a covariate Z. If Z is a categorical covariate, let N_z denote the number of individuals in the population restricted to $Z = z$. In case Z is a continuous variable, $N_z = Nf(z)$, where $f(z)$ denotes the density of Z. A generalized Chao estimator for N_z can be defined as

$$\hat{N}_{Cz} = \sum_{i=1}^{n} w_i(z) + \sum_{i=1}^{n_{1+2}} \frac{2w_i(z)}{2\hat{\lambda}_i + \hat{\lambda}_i^2},$$

whereas a generalized Zelterman estimator corresponds to

$$\hat{N}_{Zz} = \sum_{i=1}^{n} \frac{w_i(z)}{1 - \exp(-\hat{\lambda}_i)}, \tag{4}$$

where $w_i(z) > 0$ is a weight function. The weights should arise from appropriate estimates of the probability mass function or probability density function (PDF) of Z. When Z is categorical, this corresponds to $w_i(z) = I(Z_i = z)$, where $I(\cdot)$ denotes the indicator function. In this case, the fact that $\sum_z w_i(z) = 1$ implies that $\sum_z \hat{N}_{Zz} = \hat{N}_Z$ and $\sum_z \hat{N}_{Cz} = \hat{N}_C$, that is, that stratified estimators satisfy the appropriate 'benchmarking' constraint: stratified estimates aggregate exactly to the overall estimate, avoiding inconsistencies that may arise when subgroups are estimated separately. It is also straightforward to show that, with this choice of weights, stratified estimators enjoy the same asymptotic theoretical properties (e.g. consistency) of the aggregated ones.

In the case of continuous Z the appropriate PDF can be estimated through parametric inference or through kernel density estimation methods. After specification of an appropriate kernel $K(\cdot, \cdot)$ (e.g. a standard Gaussian density) and bandwidth $h > 0$, the weights are given by

$$w_i(z) = \frac{K((z - z_i)/h)}{h}.$$

It can be shown that in this case $\int_z \hat{N}_{Cz}\, \mathrm{d}z = \hat{N}_C$ and $\int_z \hat{N}_{Zz} = N_Z$. Similar asymptotic theoretical results as the categorical case hold as soon as $w_i(z)$ is consistent to the true PDF of Z.

We note here that (4) establishes a general principle for stratified estimation based on the Horvitz–Thompson formulation. In the discrete case, (4) corresponds to a generalized Zelterman estimator where the summation is restricted to a subgroup, but parameters are estimated based on the entire sample. The idea applies to any Horvitz–Thompson type estimator, as soon as the denominator is replaced with the opportune inclusion probabilities and parameters are estimated using the entire sample.

These strategies are easily generalized to stratification by more than one covariate, but many stratified estimators might quickly become unstable and/or very small as the number of covariates increases. Even conditionally on a single categorical covariate stratified estimators might have an unreasonably small sampling fraction locally. For example, in Figure 1, we show the generalized Chao estimator of the number of homicides, based on gender and a polynomial time trend up to the third order, stratified by year, together with the observed numbers. We report the observed conflict-related homicides both with and without the imputed indicators of the relationship to the conflict. Estimates are unreasonably large in recent years, possibly due to the fact that many sources have started paying less attention to the phenomenon from the period before the 2016 peace accords signature. In the next section, we propose a possible solution to the issue of extreme or unstable stratified population size estimates, which is definitely common.

3.2 Penalized estimators

In Figure 1, the variability in the sampling fraction is clearly due to the variability in parameter estimates for the time trend. If all coefficients related to time were zero, the ratio between the estimated population size and observed counts would be constant. Instability often arises because small changes in estimated covariate effects can induce large changes in estimated sampling fractions and, consequently, in the estimated population size. More stable results can thus be obtained by shrinking slopes, through an appropriate penalization term.

It is furthermore reasonable to expect penalization not only to improve stability and overall expected error of stratified estimators, but also to improve upon the marginal one. This will be demonstrated in the simulations below.

A general formulation for the penalized truncated likelihood is

$$\prod_{i=1}^{n_{1+2}} \left(\frac{2}{2 + \exp(\beta_0 + \beta' Z_i)} \right)^{1 - X_i} \left(\frac{\exp(\beta_0 + \beta' Z_i)}{2 + \exp(\beta_0 + \beta' Z_i)} \right)^{X_i} - P(\beta), \tag{5}$$

where $P(\beta)$ is component-wise increasing in $|\beta|$ and usually depends on a penalty parameter c. This can be selected through a comparison of the results over an opportunely specified grid. In our context, we

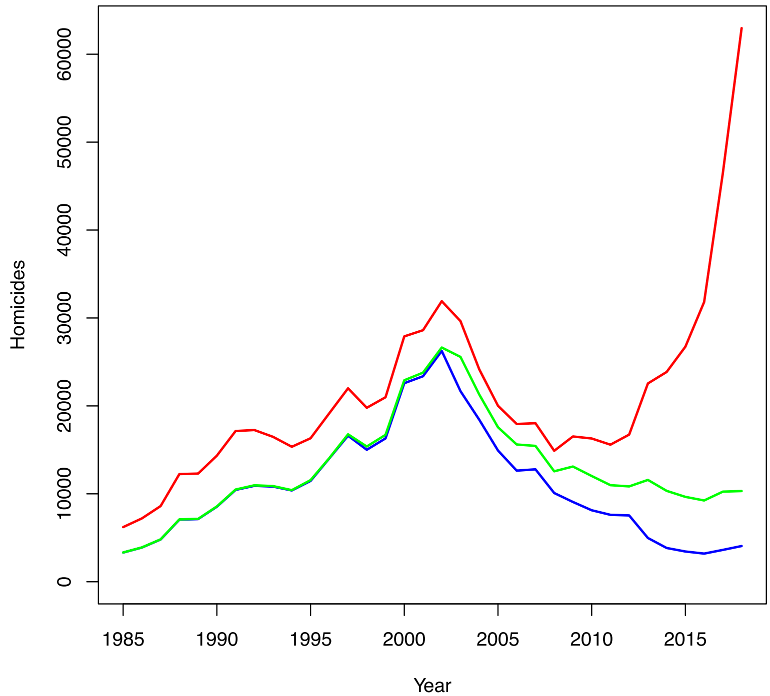

Figure 1. Observed, imputed, and estimated number of homicides in Colombia by year. Clearly, observed ≤ imputed ≤ estimated. Estimates are obtained using a generalized Chao estimator conditional on gender and a cubic time trend.

propose to select c by minimizing an empirical version of the focused information criterion (FIC), which for each c corresponds to the estimated expected mean squared error of $\hat{N}$, where $\hat{N}$ is a short notation for any population size estimator. More details are given below. The penalized truncated likelihood can be maximized numerically through an augmented Lagrangian algorithm, as the number of covariates is usually small. Other approaches can be considered for higher dimensional cases (e.g. Shi et al., 2010; Zhu & Liu, 2021).

We now give some possible alternative specifications for the penalty term.

Since in population size estimation the main target is prediction, rather than model selection, and the number of candidate covariates is usually limited, we propose using a Ridge type penalty

$$P(\beta) = c \sum_j \beta_j^2,$$

for an appropriate penalty parameter $c \geq 0$.

The penalty can also accommodate spatio-temporal smoothing. A simple scenario involves the separation of temporal and spatial effects, with parameters η_y, $y = 1, 2, \ldots$ and γ_d associated with year and department indicators. In this case, temporal smoothing is obtained by encouraging η-parameters associated with consecutive years to be close to each other, resulting in a fused Ridge type penalty of the kind

$$P(\beta) = c \sum_{y>1} (\eta_y - \eta_{y-1})^2.$$

See, for instance, Goeman (2008). Similarly, for spatial smoothing a fused Ridge type penalty can be used to penalize differences among γ_d parameters in neighbouring departments, with

$$P(\beta) = c \sum_c \sum_{d>c} C_{cd}(\gamma_d - \gamma_c)^2,$$

where C_{cd} is the cd-th element of the adjacency matrix, so that $C_{cd} = 0$ if two departments are not neighbours. This penalty corresponds to smoothing over the spatial adjacency graph defined by the departments.

Penalties can be easily combined. For instance, if there are external covariates (like gender) and year indicators in the model, after opportune standardization of any continuous covariates, the penalty is specified as

$$P(\beta) = c\left(\sum_{y>1}(\eta_y - \eta_{y-1})^2 + \sum_j \beta_j^2\right);$$

which is a generalized Ridge penalty (e.g. Obakrim et al., 2024; van Wieringen, 2019).

Efficient optimization algorithms are available (Lettink et al., 2023; Obakrim et al., 2024), which allows us to also set up bootstrap procedures for the estimation of standard errors. Surprisingly enough, bootstrap is not sufficient to accurately estimate the standard error for the population size estimators, as it proceeds conditionally on the observed sample size. In Böhning (2008) it is shown that the total variance formula should be used, where

$$\mathrm{Var}(\hat{N}) = \mathrm{Var}_n(E[\hat{N}|n]) + \mathrm{E}_n(V[\hat{N}|n]). \tag{6}$$

The resampling variance of the estimates is a good estimate for the second addend in the formula above. In order to estimate the first term, it can be shown (Böhning, 2008; Farcomeni & Scacciatelli, 2013) that a good approximation is given by

$$\sum_{i=1}^{n} \frac{\exp(-\hat{\lambda}_i)}{(1-\exp(-\hat{\lambda}_i))^2},$$

for generalized Zelterman estimators; while

$$\sum_{i=1}^{n_{1+2}} \frac{2\hat{\lambda}_i + \hat{\lambda}_i^2 - 2}{\hat{N}_C\left(2\hat{\lambda}_i + \hat{\lambda}_i^2\right)} \sum_{i=1}^{n_{1+2}} \frac{2}{2\hat{\lambda}_i + \hat{\lambda}_i^2},$$

for generalized Chao estimators. A possible alternative to (6) is given by the imputed bootstrap (Anan et al., 2017; Sangnawakij et al., 2026), which would have very little computational overhead with respect to the proposed approach.

We report here a few considerations on the properties of the proposed estimators, which are a natural consequence of the properties of the classical Chao/Zelterman estimators and penalized generalized linear models. We make the usual assumptions that (i) the population is closed, (ii) indicators of being in two lists X_i are conditionally independent and (iii) all true probabilities are bounded away from zero and one. First of all, the penalty parameter controls the relative variability of stratified estimators, with $c = 0$ corresponding to the non-penalized approach, and $c \to \infty$ implying a constant sampling fraction, namely

$$\hat{N}_v / \sum_i w_i(v) = \hat{N}/n,$$

where $\hat{N}_v$ is a short notation for any stratified estimator. This follows immediately from the fact that $\|\hat{\beta}\|$ is infinitesimal with c.

For Zelterman estimators, the underlying assumption involves that singletons and doubletons are conditionally Poisson distributed, without any requirement on the remaining counts. Use of the truncated likelihood makes Zelterman type estimators robust with respect to departures from the Poisson assumption (e.g. heavier tails) for counts larger than two. There are no other guarantees, however.

For Chao type estimators, we can give formal statements for the case of categorical covariates, with strata defined by their combination, as in Theorem 3 of Böhning et al. (2013). We assume that the covariates included are a subset of the true ones, and that the number of times each subject is observed arises from an arbitrary mixture of Poisson distributions. It is already known that when $c = 0$ (generalized Chao estimators) and $c \to \infty$ (classical Chao estimators) both marginal and stratified estimators are negatively biased (Böhning et al., 2013; Chao, 1987; Farcomeni, 2018; Farcomeni & Dotto, 2021). For penalized estimators a key result is the fact that, under the assumptions above, linear predictors and consequently $\hat{\lambda}_i(c)$ are monotone in c (e.g. Friedman et al., 2010); where we temporarily make explicit the dependence on the penalty parameter c. Monotonicity implies that $\hat{\lambda}_i(c)$ must be in the interval whose limits are defined by $\hat{\lambda}_i(0)$ and $\hat{\lambda}_i(\infty)$. A direct application of (3) leads to the claim that both marginal and stratified penalized estimators are negatively biased. It should be remarked that in the absence of unobserved heterogeneity and with a correctly specified model, generalized Chao estimators without any penalty ($c = 0$) are unbiased; while the penalized estimators will still exhibit some bias. Finally, the properties of penalized regression models guarantee that, if $c \propto n^{\alpha}$ with $\alpha < 1$, the estimators are asymptotically normal under the usual conditions.

Model selection is a final important issue. In addition to choosing the appropriate covariates to be included (since some even relevant covariates might lead to an increase in the standard error of the estimates as seen in Farcomeni, 2018), a crucial step in model selection involves fixing the penalty parameter. In this work, we minimize the empirical FIC. FIC corresponds to the expected mean squared error for a specific parameter of interest, which is estimated for each candidate value of c. The latter, in our context, is clearly the population size N (Bartolucci & Lupparelli, 2008). The empirical FIC proposed in Farcomeni (2018) simply corresponds to an empirical estimate of the mean squared error, which is a direct by-product of the resampling procedure already used for standard error estimation. For penalized generalized Chao estimators, we use the biased FIC proposed in Farcomeni (2018), which estimates bias as the average distance from the maximum estimate observed. While in this work, we base model selection on minimization of the expected mean squared error for the overall population size estimator, other choices are possible, including for instance the sum of mean squared errors of stratified estimators, or a weighted average between overall and stratified. The proposed selection approach will be validated in the simulation study below.

4 Simulation study

In this section, we present the results of a brief simulation study. Our main aim is to show how and when the penalty can lead to an improvement of marginal and/or stratified estimators.

We let $N = \{5{,}000,\ 10{,}000\}$, with $S = 30$ lists. Data generation from such a high dimensional log-linear model would be cumbersome. We therefore induce dependence (and unobserved heterogeneity) assuming that lists are independent conditionally on a discrete latent variable. We use a two-masses latent variable with support points $(-\tau, \tau)$ and probability masses $(2/3, 1/3)$ (see, e.g. Pledger, 2005). We let $\tau = \{1, 3, 5\}$. We also generate three covariates from a zero-centred multivariate normal with unit variances and correlations equal to -0.2. The first covariate is rounded to the first decimal point and truncated at -1.5 and 1.5. Stratified estimates are then obtained with respect to the 31 resulting levels of this covariate. The three covariates are included in the data generating model with equal coefficients, fixed at $\{1, 2\}$ We finally set the intercept so that the overall sampling fraction is $E[n]/N = \{0.4, 0.8\}$. Overall, we then evaluate 24 scenarios (two true population sizes, three latent locations, two slopes, and two expected sampling fractions).

For each scenario, we generate data $B = 1{,}000$ times and compare (i) Manrique-Vallier (2016) Bayesian non-parametric approach (BNP), (ii) Chao estimator, (iii) Zelterman estimator (Zelt), (iv) generalized Chao estimator (GChao), (v) generalized Zelterman estimator (GZelt), (vi) our proposed penalized generalized Chao estimator (PGChao), and (vii) our proposed penalized generalized Zelterman estimator (PGZelt). Penalties are selected as discussed above. Log-linear estimators are not included because the large number of lists makes even reduced interaction models unstable.

We report the median of the squared errors for the overall population size in Table 2. For each configuration we then estimate the median (across the B replicates) squared error $MeSE_{x_1}$ for each level of the stratification covariate X_1, and report the median of $MeSE_{x_1}$ over x_1 in Table 3. We report the

Table 2. Simulation study. Median squared error for $\hat{N}$ at different levels of true population size N, sampling fraction $E[n]/N$, covariate effects β, and latent locations τ; for seven different methodologies. Results are based on $B = 1{,}000$ replicates.

N	$E[n]/N$	β	τ	BNP	Chao	GChao	PGChao	Zelt	GZelt	PGZelt
5,000	0.40	1	1	0.40	0.34	0.32	0.33	0.29	0.28	0.29
5,000	0.40	1	3	0.45	0.24	0.17	0.21	1.40	2.47	1.41
5,000	0.40	1	5	0.42	0.20	0.17	0.13	1.70	5.82	1.72
5,000	0.40	2	1	0.46	0.39	0.33	0.34	0.30	0.29	0.32
5,000	0.40	2	3	0.55	0.50	0.47	0.48	0.19	0.14	0.19
5,000	0.40	2	5	0.52	0.43	0.21	0.24	0.40	22.08	0.40
5,000	0.80	1	1	0.10	0.06	0.04	0.04	0.06	0.06	0.04
5,000	0.80	1	3	0.09	0.05	0.01	0.01	0.10	0.22	0.10
5,000	0.80	1	5	0.09	0.05	0.01	0.01	0.10	0.22	0.10
5,000	0.80	2	1	0.12	0.09	0.06	0.07	0.06	0.02	0.02
5,000	0.80	2	3	0.13	0.10	0.02	0.03	0.11	0.47	0.07
5,000	0.80	2	5	0.13	0.10	0.02	0.03	0.11	0.48	0.07
10,000	0.40	1	1	0.37	0.34	0.32	0.33	0.28	0.28	0.29
10,000	0.40	1	3	0.43	0.27	0.24	0.26	1.47	1.69	1.47
10,000	0.40	1	5	0.38	0.21	0.12	0.10	1.82	5.99	1.84
10,000	0.40	2	1	0.43	0.39	0.34	0.35	0.31	0.30	0.33
10,000	0.40	2	3	0.52	0.49	0.47	0.47	0.17	0.14	0.17
10,000	0.40	2	5	0.51	0.43	0.15	0.20	0.40	20.16	0.40
10,000	0.80	1	1	0.08	0.06	0.03	0.04	0.06	0.06	0.04
10,000	0.80	1	3	0.08	0.05	0.01	0.01	0.10	0.22	0.10
10,000	0.80	1	5	0.08	0.05	0.01	0.01	0.10	0.22	0.09
10,000	0.80	2	1	0.12	0.10	0.06	0.07	0.06	0.01	0.02
10,000	0.80	2	3	0.12	0.10	0.02	0.03	0.11	0.47	0.07
10,000	0.80	2	5	0.12	0.10	0.02	0.03	0.11	0.47	0.07

median squared errors instead of the mean, in order to reduce the influence of extreme estimates on competing methodologies. A comparison based on mean squared errors would be overwhelmingly in favour of the proposed penalized approaches. We also report the results about bias (as a proportion of the true population size) for the overall and stratified estimators in Tables 4 and 5, respectively.

From the comparison among the estimators considered, it can be seen that the Manrique-Vallier (2016) approach is consistently outperformed by Zelterman and/or Chao type estimators. Penalized estimators often bring about the best performance within their model class, that is, penalized Zelterman estimators often outperform both classical and generalized Zelterman estimators; and penalized Chao estimators often outperform both classical and generalized Chao estimators. This can be seen especially when the (observed and/or unobserved) heterogeneity is high. Covariate inclusion, without any penalty, in Chao and Zelterman estimators might either be helpful or detrimental, as already noted in the simulation study of Farcomeni (2018). The generalized Zelterman estimator has inflated median squared errors in some scenarios, a problem that is corrected by the penalty only to some extent. This fact strengthens our message that Chao type estimators should be preferred in settings similar to those of this work. Ridge type penalties improve generalized estimators when their performance is worse than the classical ones also for Chao type estimators, which are generally more stable. The advantage of penalized estimators becomes even more evident after stratified estimation.

Chao type estimators are negatively biased or approximately unbiased; with some positive values for the non-penalized generalized ones. As expected, Zelterman type estimators are much less frequently negatively biased, regardless of their formulation.

Table 3. Simulation study. Median of cross-strata median squared error for $\hat{N}_{x_1}$ at different levels of true population size N, sampling fraction $E[n]/N$, covariate effects $\boldsymbol{\beta}$, and latent locations $\boldsymbol{\tau}$; for seven different methodologies. Results are based on $B = 1{,}000$ replicates.

N	$E[n]/N$	$\boldsymbol{\beta}$	τ	BNP	Chao	GChao	PGChao	Zelt	GZelt	PGZelt
5,000	0.40	1	1	2.90	1.70	1.69	1.75	1.58	1.50	1.59
5,000	0.40	1	3	3.06	3.44	1.14	1.32	17.52	13.20	7.34
5,000	0.40	1	5	3.08	6.05	1.13	0.96	34.59	30.96	9.02
5,000	0.40	2	1	3.06	1.97	1.77	1.85	1.79	1.55	1.83
5,000	0.40	2	3	3.14	2.34	2.70	2.67	2.60	1.78	1.33
5,000	0.40	2	5	3.11	3.15	1.36	1.28	8.18	82.69	2.03
5,000	0.80	1	1	0.75	0.24	0.17	0.19	0.43	0.28	0.38
5,000	0.80	1	3	0.66	0.24	0.14	0.14	0.65	1.05	0.56
5,000	0.80	1	5	0.65	0.24	0.14	0.14	0.65	1.06	0.56
5,000	0.80	2	1	0.83	0.27	0.20	0.22	0.47	0.19	0.29
5,000	0.80	2	3	0.84	0.26	0.14	0.14	0.66	1.30	0.41
5,000	0.80	2	5	0.87	0.28	0.14	0.14	0.69	1.30	0.42
10,000	0.40	1	1	5.47	3.18	3.43	3.50	2.81	3.04	3.19
10,000	0.40	1	3	6.17	4.86	3.12	3.31	27.43	17.97	15.44
10,000	0.40	1	5	6.16	6.00	1.54	1.43	48.72	64.65	19.54
10,000	0.40	2	1	5.84	3.81	3.62	3.75	3.29	3.23	3.71
10,000	0.40	2	3	6.23	4.13	5.27	5.25	5.14	3.42	2.57
10,000	0.40	2	5	6.26	6.10	2.04	2.04	24.25	154.16	4.14
10,000	0.80	1	1	1.27	0.38	0.29	0.33	0.69	0.52	0.70
10,000	0.80	1	3	1.04	0.34	0.20	0.20	1.26	2.28	1.08
10,000	0.80	1	5	1.04	0.34	0.20	0.20	1.26	2.27	1.08
10,000	0.80	2	1	1.49	0.52	0.46	0.51	0.80	0.35	0.57
10,000	0.80	2	3	1.68	0.45	0.20	0.21	1.23	2.59	0.72
10,000	0.80	2	5	1.67	0.43	0.20	0.21	1.20	2.53	0.72

5 Data analysis

In this section, we report estimates of penalized generalized Chao and Zelterman estimators for our motivating real data example. For ease of presentation, we use the same covariates for all data sets; conditioning on gender, department, and a polynomial time trend of the third order. Time and department of the event are the most reliably and consistently recorded variables across sources. Other covariates (especially possibly important ones like age and ethnicity) often suffer from measurement error and high rates of missing values. The third order for the polynomial for the time variable is selected using FIC, which also confirms that gender is another important covariate.

The penalty parameter is selected separately for Chao and Zelterman estimators, minimizing the estimator-specific empirical FIC on a grid of values for c. Let $c_{\max} = \min_c \{c : \max_j |\hat{\beta}_j| < 10^{-38}\}$, that is, $c_{\max}$ is the smallest penalty parameter for which all coefficients are practically equal to zero. The starting penalty value is fixed as $10^{-6}c_{\max}$, and then increased. Subsequent values are equally spaced on the logarithmic scale. After the 80th value, we quintuple the shifting constant. The search is stopped as soon as the estimated FIC starts to increase with the penalty parameter. For each c, FIC is estimated using 100 bootstrap replicates.

In Table 6, we report estimates for the number of homicides, also stratified by gender, together with their 95% confidence intervals. Standard errors are computed using (6), after 500 bootstrap replicates. As stated in Section 2, we restrict ourselves to conflict-related homicides, using one imputed data set (a sensitivity analysis will be reported below).

There is a clear discrepancy between Chao type and Zelterman type estimators, which could be expected. We stress that Zelterman type estimators rely on the assumption that counts for singletons

Table 4. Simulation study. Relative bias for $\hat{N}$ at different levels of true population size N, sampling fraction $E[n]/N$, covariate effects $\boldsymbol{\beta}$, and latent locations $\boldsymbol{\tau}$; for seven different methodologies. Results are based on $B = 1{,}000$ replicates.

N	$E[n]/N$	$\boldsymbol{\beta}$	τ	BNP	Chao	GChao	PGChao	Zelt	GZelt	PGZelt
10,000	0.40	1	1	−0.37	−0.34	−0.32	−0.33	−0.28	−0.28	−0.29
10,000	0.40	1	3	−0.43	−0.26	−0.23	−0.25	1.49	1.86	1.49
10,000	0.40	1	5	−0.37	−0.20	0.14	−0.04	1.87	6.80	1.90
10,000	0.40	2	1	−0.43	−0.39	−0.34	−0.35	−0.31	−0.30	−0.33
10,000	0.40	2	3	−0.52	−0.48	−0.47	−0.47	−0.16	−0.13	−0.17
10,000	0.40	2	5	−0.50	−0.43	0.15	−0.17	0.41	25.35	0.41
10,000	0.80	1	1	−0.08	−0.06	−0.03	−0.04	0.06	0.05	0.04
10,000	0.80	1	3	−0.08	−0.05	0.01	−0.00	0.10	0.22	0.09
10,000	0.80	1	5	−0.08	−0.05	0.01	−0.00	0.10	0.22	0.09
10,000	0.80	2	1	−0.12	−0.10	−0.06	−0.07	0.06	−0.01	−0.02
10,000	0.80	2	3	−0.12	−0.10	0.01	−0.03	0.11	0.48	0.07
10,000	0.80	2	5	−0.12	−0.10	0.01	−0.02	0.11	0.47	0.07
5,000	0.40	1	1	−0.70	−0.67	−0.66	−0.67	−0.64	−0.64	−0.65
5,000	0.40	1	3	−0.72	−0.61	−0.53	−0.59	0.24	1.12	0.24
5,000	0.40	1	5	−0.71	−0.59	−0.32	−0.51	0.39	4.69	0.41
5,000	0.40	2	1	−0.73	−0.69	−0.66	−0.67	−0.65	−0.64	−0.66
5,000	0.40	2	3	−0.77	−0.75	−0.74	−0.74	−0.59	−0.56	−0.60
5,000	0.40	2	5	−0.76	−0.71	−0.22	−0.59	−0.28	20.65	−0.28
5,000	0.80	1	1	−0.55	−0.53	−0.52	−0.52	−0.47	−0.47	−0.48
5,000	0.80	1	3	−0.54	−0.53	−0.49	−0.50	−0.45	−0.39	−0.45
5,000	0.80	1	5	−0.54	−0.53	−0.49	−0.50	−0.45	−0.39	−0.45
5,000	0.80	2	1	−0.56	−0.55	−0.53	−0.53	−0.47	−0.50	−0.51
5,000	0.80	2	3	−0.57	−0.55	−0.49	−0.51	−0.44	−0.25	−0.46
5,000	0.80	2	5	−0.56	−0.55	−0.49	−0.51	−0.44	−0.24	−0.46

and doubletons are conditionally Poisson, which might not be the case here. On the other hand, Chao estimator can be expected to provide a reliable lower bound, that is, we can expect the actual toll to be at least the reported estimate. The estimate is almost 60% higher than the recorded number of homicides, corresponding to a sampling fraction of approximately 69%. The male to female ratio is between 7 and 8, which is in line with what is expected in a war due to combatant targeting and gender-selective recruitment patterns.

Before discussing further results, we would like to provide evidence that the penalty gives substantial advantages. The selected penalty for both estimators is $c = 796.03$. First of all, we note that the results are rather stable with respect to the choice of penalty, with Chao type estimates between 650,000 and 640,000 for $c \geq 20$; and Zelterman type estimates between 720,000 and 710,000 for $c \geq 4$. When $c = 0$ the estimates are 15% and 9.5% higher, respectively; with some extreme and implausible estimates in specific department/years, an indication that both spatial and temporal smoothing are important in this application (see also the discussion below). The penalty reduces the overall uncertainty, with a standard error of the penalized estimator that is 40% of the non-penalized one for the Chao type estimator; and 56% of the non-penalized estimator for the Zelterman type one. Importantly, FIC values at zero penalty are 10.5 and 5 times higher, respectively, indicating strong evidence against the choice $c = 0$.

In Figure 2, we report the estimated number of homicides by department, excluding San Andrés island.

The departments that seem to be most affected in absolute terms are Antioquia ($\hat{N}_{Cz} = 170{,}947$ 95% CI: 169,841–172,338), and Valle del Cuaca ($\hat{N}_{Cz} = 59{,}663$, 95% CI: 58,785–60,657); with substantially lower levels of violence in the South-East departments of Vichada, Guainía, Vaupes, and Amazonas.

Table 5. Simulation study. Median relative bias for $\hat{N}_{x_1}$ at different levels of true population size N, sampling fraction $E[n]/N$, covariate effects β, and latent locations τ; for seven different methodologies. Results are based on $B = 1{,}000$ replicates.

N	$E[n]/N$	β	τ	BNP	Chao	GChao	PGChao	Zelt	GZelt	PGZelt
10,000	0.40	1	1	−0.02	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01
10,000	0.40	1	3	−0.02	0.00	−0.01	−0.01	0.09	0.06	0.05
10,000	0.40	1	5	−0.02	0.01	0.00	−0.00	0.15	0.20	0.06
10,000	0.40	2	1	−0.02	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01
10,000	0.40	2	3	−0.02	−0.01	−0.02	−0.02	0.00	−0.01	−0.01
10,000	0.40	2	5	−0.02	−0.00	0.00	−0.01	0.08	0.48	0.01
10,000	0.80	1	1	−0.00	−0.00	−0.00	−0.00	0.00	0.00	0.00
10,000	0.80	1	3	−0.00	−0.00	0.00	−0.00	0.00	0.01	0.00
10,000	0.80	1	5	−0.00	−0.00	0.00	0.00	0.00	0.01	0.00
10,000	0.80	2	1	−0.00	−0.00	−0.00	−0.00	0.00	0.00	0.00
10,000	0.80	2	3	−0.01	−0.00	0.00	−0.00	0.00	0.01	0.00
10,000	0.80	2	5	−0.01	−0.00	0.00	−0.00	0.00	0.01	0.00
5,000	0.40	1	1	−0.01	−0.01	−0.01	−0.01	−0.00	−0.00	−0.00
5,000	0.40	1	3	−0.01	0.00	−0.00	−0.00	0.05	0.04	0.02
5,000	0.40	1	5	−0.01	0.02	0.00	−0.00	0.11	0.10	0.03
5,000	0.40	2	1	−0.01	−0.01	−0.01	−0.01	−0.01	−0.00	−0.01
5,000	0.40	2	3	−0.01	−0.01	−0.01	−0.01	0.00	−0.00	−0.00
5,000	0.40	2	5	−0.01	−0.00	−0.00	−0.00	0.03	0.26	0.01
5,000	0.80	1	1	−0.00	−0.00	−0.00	−0.00	0.00	0.00	0.00
5,000	0.80	1	3	−0.00	−0.00	0.00	−0.00	0.00	0.00	0.00
5,000	0.80	1	5	−0.00	−0.00	0.00	−0.00	0.00	0.00	0.00
5,000	0.80	2	1	−0.00	−0.00	−0.00	−0.00	0.00	0.00	0.00
5,000	0.80	2	3	−0.00	−0.00	0.00	−0.00	0.00	0.00	0.00
5,000	0.80	2	5	−0.00	−0.00	0.00	−0.00	0.00	0.00	0.00

Table 6. Estimated number of homicides overall, and by gender, with 95% confidence intervals. Estimates are obtained using a penalized generalized Chao ($\hat{N}_C$) or Zelterman ($\hat{N}_Z$) estimator conditional on gender, a cubic time trend, and department indicators.

	$\hat{N}_C$	2.5% CI	97.5% CI	$\hat{N}_Z$	2.5% CI	97.5% CI
Overall	642,964	640,943	644,986	715,186	712,048	718,906
Females	72,517	71,465	73,643	81,972	80,528	83,423
Males	570,447	568,777	572,786	633,214	630,372	636,473

The Antioquia department is by far the most affected by lethal violence, with almost three times the estimated number of homicides with respect to the Valle del Cuaca department. It has historically been a hub for drug trafficking and illegal mining, which has led to armed disputes for territorial control. Smoothing prevents extreme estimates in sparsely observed departments such as Guainía, Bogotá, and Amazonas. For instance, the non-penalized stratified estimate for Guainía is 520 homicides (95% CI: 327–1013), almost twice the optimal penalized estimate of 271 (95% CI: 211–342). The penalty also drastically reduces the standard error of the stratified estimates, as expected.

The estimates stratified by year are shown in Figure 3. The continuous red line of Figure 3 should be compared with Figure 1. It is clear that the penalty yields much more stable results for the estimates in the most recent years, through temporal smoothing. The discrepancy between the completely

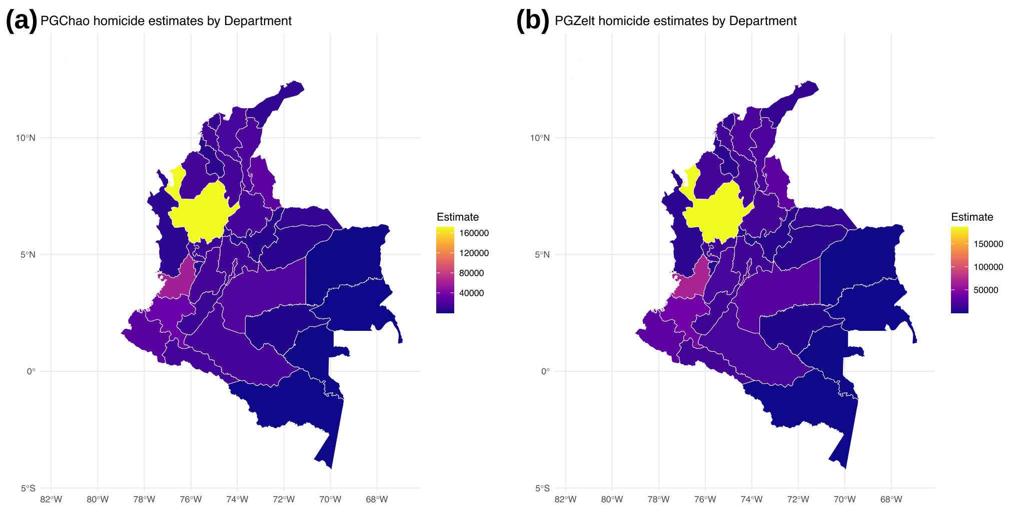

Figure 2. Estimated number of homicides by department. Estimates are obtained using a penalized generalized Chao (left panel) or Zelterman (right panel) estimator conditional on gender, a cubic time trend, and department indicators.

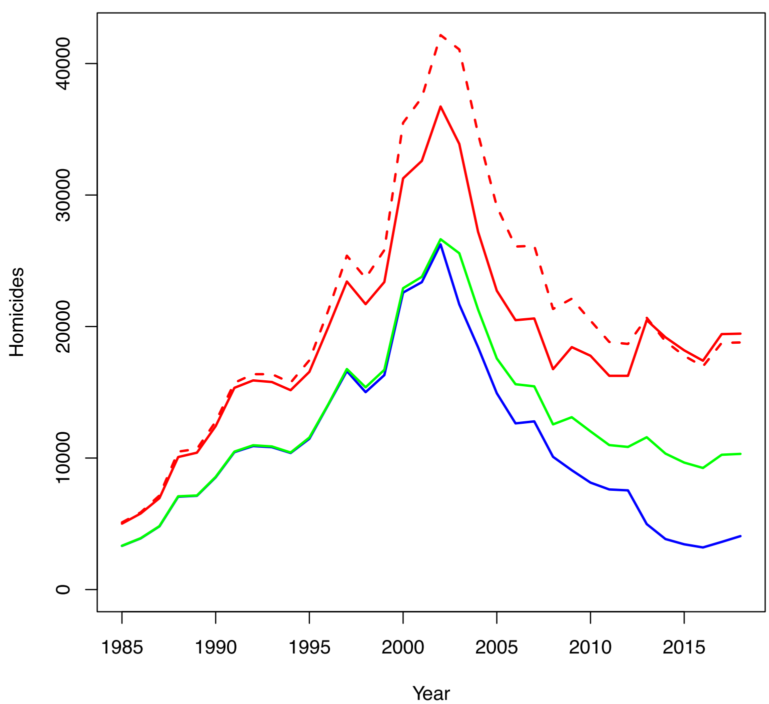

Figure 3. Observed, imputed, and estimated number of homicides by year. Clearly, observed ≤ imputed ≤ estimated. Estimates are obtained using a penalized generalized Chao (continuous red line) or Zelterman (dashed red line) estimator conditional on gender, a cubic time trend, and department indicators.

observed and estimated counts in recent years is only partly due to imputation. Two phenomena likely occur after 2016 then: on the one hand, a decreased sampling fraction due to documentation decline after the official declarations of peace; on the other hand, fragmentation of armed groups and resurgence due to lower centralized control and dissatisfaction with the peace conditions. While there has

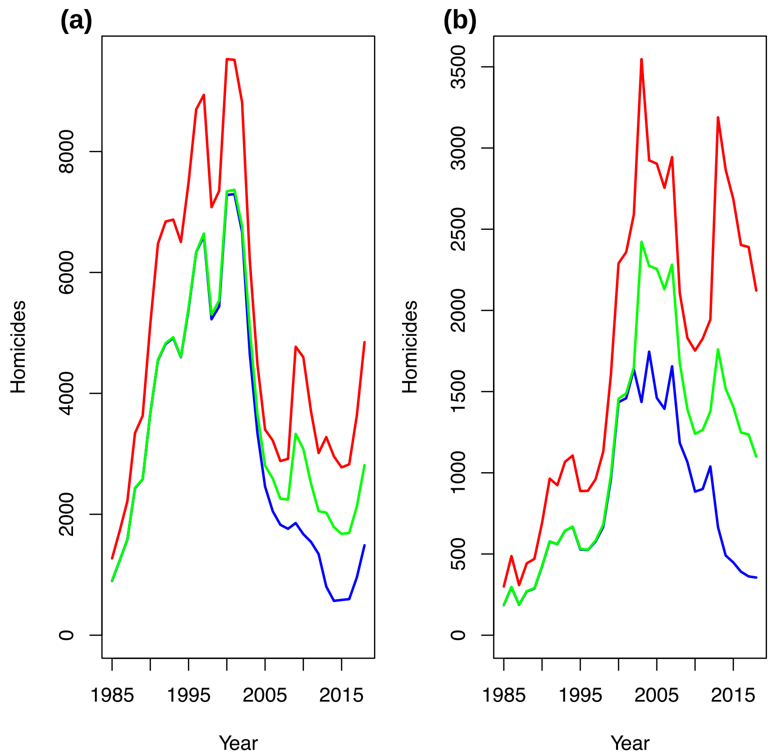

Figure 4. Observed, imputed, and estimated number of homicides by year in Antioquia (left panel) and Valle del Cuaca (right panel). Clearly, observed ≤ imputed ≤ estimated. Estimates are obtained using a penalized generalized Chao estimator conditional on gender, a cubic time trend, and department indicators.

been a substantial decrease in violence with respect to the worst years, there might be a newly increasing trend in the most recent years. This should also be associated with the lack of an increase in deaths just before the signature, as would sometimes be expected before a durable ceasefire (e.g. Williams et al., 2024 and references therein).

In order to support the last statements and further illustrate the potential of penalized estimates for smoothing and stabilization even in subgroups, we show in Figure 4 penalized Chao estimates stratified by year and department for the two most affected departments: Antioquia and Valle del Cuaca. It can be seen that in Antioquia there was both an observed and an estimated resurgence after 2016; while Valle del Cuaca seems to have experienced a slight increase in deaths before the ceasefire, and a lack of resurgence afterwards.

Between 1% and 2% of the Colombian population has been killed in relation to the conflict in the period considered, a figure that is even more striking when one observes that many homicides occurred in a relatively short time frame and a small subset of departments. Indeed, an incidence about three times the national one is estimated for Antioquia; and an overall incidence in the worst year (2002) that is about twice the year-specific mean, and more than seven times the minimal.

We now focus on forced recruitment. In Table 7, we report estimates overall, and stratified by sex, together with their 95% confidence intervals. The optimal penalties are selected as $c = 32.58$ for the Chao type estimator and $c = 65.17$ for the Zelterman type estimator. Once again, penalized estimates seem to be fairly stable, and definitely better (in terms of FIC, stability, and standard errors) than the non-penalized ones.

The ratio between the observed sample size and the population size is estimated at 68%, very close to the coverage for conflict-related homicides. The male to female ratio is about 2.5, reflecting a lower level of gender-selective recruitment patterns compared to other forms of recruitment. Forcefully

Table 7. Estimated number of victims of forced recruitment overall, and by gender, with 95% confidence intervals. Estimates are obtained using a penalized generalized Chao ($\hat{N}_C$) or Zelterman ($\hat{N}_Z$) estimator conditional on gender, a cubic time trend, and department indicators.

	$\hat{N}_C$	**2.5% CI**	**97.5% CI**	$\hat{N}_Z$	**2.5% CI**	**97.5% CI**
Overall	28,235	27,837	28,793	29,584	29,113	30,425
Females	8,122	7,923	8,419	8,306	8,047	8,632
Males	20,113	19,762	20,660	21,278	20,843	22,041

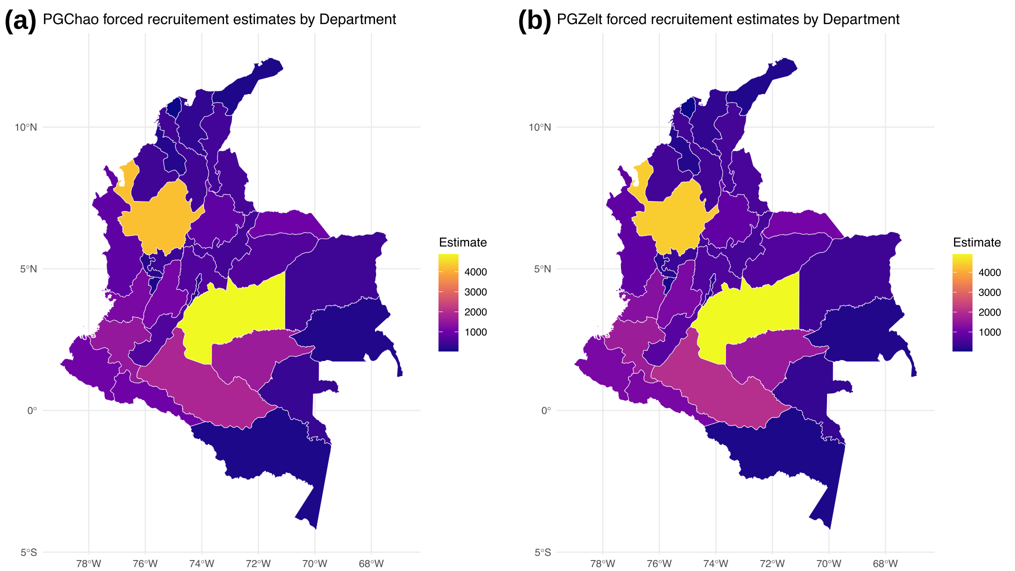

Figure 5. Estimated number of victims of forced recruitment by department. Estimates are obtained using a penalized generalized Chao (left panel) or Zelterman (right panel) estimator conditional on gender, a cubic time trend, and department indicators.

recruited individuals are often not selected prominently for their training or abilities in combat. A few females were forcefully recruited to be raped by their supposedly fellow combatants. Before or after recruitment, some were subjected to forced abortion.

Finally, in Figure 5, we report the estimated number of victims of forced recruitment by department. The highest number of forced recruitments is seen in the Meta department, closely followed by Antioquia.

We report our estimates for kidnappings in Table 8, together with their 95% confidence intervals. Once again, we restrict to conflict-related kidnappings, that could occur for instance for financing the conflict through ransoms. The penalty parameter for the Chao type estimator is selected as $c = 4{,}192.55$, and $c = 1{,}048.14$ for the Zelterman type estimator. In order to evaluate sensitivity with respect to the search strategy, we have also selected the penalty parameters on a pre-specified grid made of twice as many values equally spaced on the logarithmic scale, with no stopping rules. The selected penalties were approximately the same. Despite the seemingly large values for the selected penalty parameters, stratified estimates retained a reasonable variability, with standard deviation for the time-specific estimates of 2,331.3 and 2,488.4 for the penalized generalized Chao and Zelterman approaches, respectively. Similar values are seen at the department-specific level. Penalized estimates also seem to be remarkably stable, with Chao type estimators between 62,000

Table 8. Estimated number of kidnappings overall, and by gender, with 95% confidence intervals. Estimates are obtained using a penalized generalized Chao ($\hat{N}_C$) or Zelterman ($\hat{N}_Z$) estimator conditional on gender, a cubic time trend, and department indicators.

	$\hat{N}_C$	2.5% CI	97.5% CI	$\hat{N}_Z$	2.5% CI	97.5% CI
Overall	61,559	61,212	61,939	64,682	64,055	65,471
Females	14,516	14,260	14,813	15,174	14,870	15,534
Males	47,043	46,724	47,379	49,508	48,948	50,102

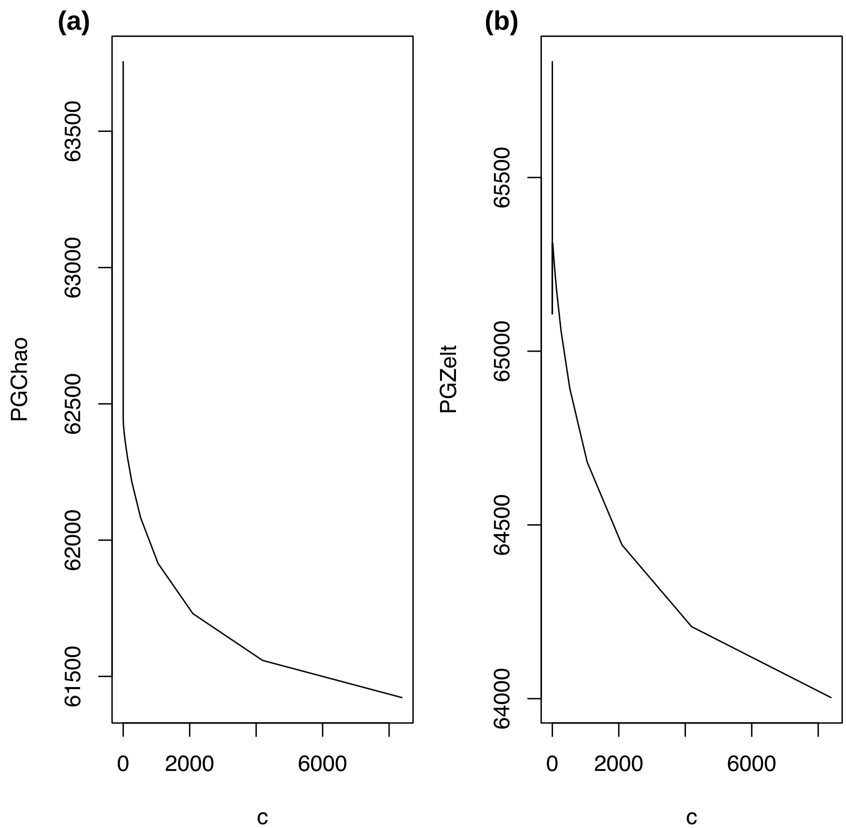

Figure 6. Estimated number of victims of kidnapping as a function of *c*. Estimates are obtained using a penalized generalized Chao (left panel) or Zelterman (right panel) estimator conditional on gender, a cubic time trend, and department indicators.

and 61,400 for $c \geq 0.07$, and Zelterman type estimators below 66,000 for all values of c; as testified by Figure 6.

Despite stability with respect to c, when $c = 0$ Chao and Zelterman type estimators have a FIC that is at least twice the penalized ones. Furthermore, they seem to be unstable at the resampling stage, with estimates as large as 400,000. On the other hand, resampled penalized Chao estimators have a range of 61,100–62,310, and penalized Zelterman estimators a range of 63,852–65,620.

The estimated coverage is approximately 83%, slightly larger than the previous ones. A male to female ratio of approximately 3 might be due to several reasons. We speculate that one of the main reasons is simply availability; as a few kidnapped individuals were combatants. Finally, in Figure 7 we report the estimated number of kidnappings by department.

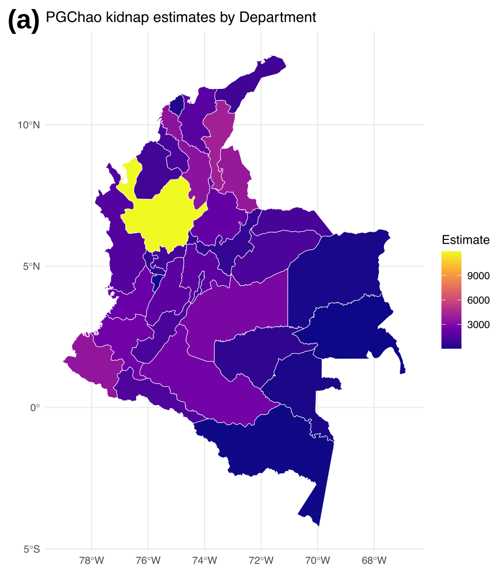

Figure 7. Estimated number of kidnappings by department. Estimates are obtained using a penalized generalized Chao (left panel) or Zelterman (right panel) estimator conditional on gender, a cubic time trend, and department indicators.

Table 9. MAD for estimates across 100 multiply imputed data sets. Estimates are obtained using a penalized generalized Chao ($\hat{N}_C$) or Zelterman ($\hat{N}_Z$) estimator conditional on gender, a cubic time trend, and department indicators.

	$\hat{N}_C$	$\hat{N}_Z$
Homicide	381.73	523.99
Forced recruitment	2.84	5.03
Kidnapping	52.87	87.45

We conclude with a brief sensitivity analysis with respect to missing data. All results presented are based on a single imputed data set. In Table 9, we report the median absolute deviation (MAD) for the overall estimates, based on the same penalty parameters, over the 100 imputed data sets. Note that there are two different sources of variability: on the one hand, individuals with missing conflict-relatedness indicator might enter or exit the sample due to imputation. Additionally, covariates (e.g. gender) might change due to imputation.

It can be seen that sensitivity is always reasonable. Results cannot be expected to be strongly affected by imputation.

6 Concluding remarks

We have proposed a general framework for stratified population size estimation with respect to continuous and categorical variables. The variables used for stratification do not necessarily have to be used at the estimation stage. We have also proposed a simple approach for estimation of the standard errors, and consequently confidence intervals, for population size estimators. It shall be remarked that Chao type estimators are in general biased, hence the nominal coverage will be achieved at the estimable lower bound; not at the true population size.

Population size estimates often show instability, and stratified analyses particularly so, which prompted us to propose penalized estimators. Generalized penalized Chao and Zelterman type estimators have been seen to be stable, with a decreased standard error overall and for the stratified estimates. They are robust and scale well with respect to both the sample size n and, more importantly, the number of lists S. While there are other possible strategies to tackle the same problem (e.g. Bayesian approaches), we believe the proposed method is valuable due to computational scalability with many lists, simple and efficient optimization, robustness to misspecification and unobserved heterogeneity. For the homicide data set, estimates are obtained in about thirty seconds with non-optimized R code on a personal laptop. A non-computationally intensive approach like the one proposed might foster the applicability of population size estimation methods.

The evaluation of the extent of lethal and non-lethal violence is an important component of the peace process and, at the very least, a responsibility towards the victims. The resulting estimates provide new quantitative evidence on the scale and temporal evolution of human rights violations during the Colombian armed conflict, identifying a clear time pattern, most and least affected departments, and the gender ratio of victims. Importantly, there is some heterogeneity in department-specific time trends, with the Antioquia department (which is also by far the department with the highest number of conflict-related homicides) that might have experienced a resurgence in violence after the signature of the peace treaties in 2016. Peace and stability require addressing the underlying drivers of conflict; which in Colombia include illegal mining, drug trafficking, economic inequality, poverty, and lack of justice. The Colombian peace process was indeed conceived also to pursue these structural changes (Fabra-Zamora et al., 2021). While important challenges remain, recent developments indicate substantial progress, and provide grounds for cautious optimism of further future improvements.

While we have reported both Chao and Zelterman estimators for completeness, we stress that in our context Chao type estimators might be more appropriate as they are often conservative in finite samples. This is especially useful in the area of conflict, in our opinion. A common critique to estimates (as opposed to observed counts) is that they might exaggerate the impact of the conflict by exhibiting a positive bias. This critique is directly handled by the use of (penalized, generalized) Chao estimators.

Another important issue involves avoiding possible contradictions in stratified estimates (i.e. that they do not sum to the estimated total), which might prompt the lay public and politicians to claim errors in data analysis. This is avoided using our proposed strategy for stratified estimation, even if it is not the only option. A poor man's solution might be to separately estimate population size of subgroups and then sum the estimates. It is straightforward nonetheless to see that this strategy would be sub-optimal, prone to instability in subgroups with small sample size or low sampling fraction, and with larger standard error with respect to the proposal. The penalty brings about some additional bias, but with clear advantages.

A limitation of the proposed approaches is that they account for list dependence indirectly, through heterogeneity in capture frequencies. Direct modelling might lead to lower standard errors and bias. Nonetheless, in our application there are $S = 34$ lists, which are definitely too many to consider log-linear models. A study of how list dependence maps to conditional counting distributions is left for further work. An alternative route might be to extend the non-parametric latent class methodology of Manrique-Vallier (2016) to the case of covariates. Another important open issue involves missing values. A carefully performed multiple imputation is definitely an appropriate approach. The scarce sensitivity found in the motivating application indicates that the results are credible. However, a promising route seems related to doubly robust estimators (Das et al., 2024; Rubio et al., 2026), which would give theoretical guaranties under MAR assumptions. The development of doubly robust generalized Chao and Zelterman estimators is left for further work.

Acknowledgments

The author thanks Maria Gargiulo for explanations about the data and comments on the results; and three anonymous reviewers for constructive and useful remarks.

Conflicts of interest

The author has no potential conflicts of interests to disclose.

Funding

The author received no funding in connection with this research.

Data availability

The data underlying this article are available at https://hrdag.org/jep-cev-colombia/

Supplementary material

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

References

Amado P., Acero W., Argoty C., Babativa G., Bernal L. K., Castro A., Durán M. J., Espinosa E., Gómez V., González A., Ortíz M. A., Ball P., Gargiulo M., Rozo V., Shah T., Dueñas J. G., Jaimes L. E., Lozano C., Murillo M. J., & Rodríguez H. (2025). *Methodological report of the JEP-CEV-HRDAG joint project on data integration and statistical estimation (Technical report)*. Human Rights Data Analysis Group.

Anan O., Böhning D., & Maruotti A. (2017). Uncertainty estimation in heterogeneous capture-recapture count data. *Journal of Statistical Computation and Simulation*, *87*(10), 2094–2114. https://doi.org/10.1080/00949655.2017.1315668

Bales K. B., Hesketh O., & Silverman B. W. (2015). Modern slavery in the UK: How many victims? *Significance*, *12*(3), 16–21. https://doi.org/10.1111/j.1740-9713.2015.00824.x

Ball P., & Price M. (2019). Using statistics to assess lethal violence in civil and inter-state war. *Annual Review of Statistics and Its Application*, *6*(1), 63–84. https://doi.org/10.1146/annurev-statistics-030718-105222

Bartolucci F., & Lupparelli M. (2008). Focused information criterion for capture-recapture models for closed populations. *Scandinavian Journal of Statistics*, *35*(4), 629–649. https://doi.org/10.1111/j.1467-9469.2008.00604.x

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

Böhning D. (2010). Some general comparative points on Chao's and Zelterman's estimators of the population size. *Scandinavian Journal of Statistics*, *37*(2), 221–236. https://doi.org/10.1111/j.1467-9469.2009.00676.x

Böhning D., & van der Heijden P. G. M. (2009). A covariate adjustment for zero-truncated approaches to estimating the size of hidden and elusive populations. *Annals of Applied Statistics*, *3*(2), 595–610. https://doi.org/10.1214/08-AOAS214

Böhning D., Vidal-Diez A., Lerdsuwansri R., Viwatwongkasem C., & Arnold M. A. (2013). A generalization of Chao's estimator for covariate information. *Biometrics*, *69*(4), 1033–1042. https://doi.org/10.1111/biom.12082

Chao A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, *43*(4), 783–791. https://doi.org/10.2307/2531532

Checchi F., & Roberts L. (2008). Documenting mortality in crises: What keeps us from doing better? *PLoS Medicine*, *5*(7), e146. https://doi.org/10.1371/journal.pmed.0050146

Comisión de la Verdad (Ed.). (2022). *Hallazgos y recomendaciones: de la Comisión de la Verdad de Colombia*. Number Tomo 2 in Hay futuro si hay verdad. Comisión de la Verdad, Bogotá, primera edición edition.

Cormack R. M. (1989). Log-linear models for capture-recapture. *Biometrics*, *45*(2), 395–413. https://doi.org/10.2307/2531485

Das M., Kennedy E. H., & Jewell N. P. (2024). Doubly robust capture-recapture methods for estimating population size. *Journal of the American Statistical Association*, *119*(546), 1309–1321. https://doi.org/10.1080/01621459.2023.2187814

Fabra-Zamora J. L., Molina-Ochoa A., & Doubleday N. C. (2021). *The Colombian peace agreement: A multidisciplinary assessment*. Routledge.

Farcomeni A. (2018). Fully general Chao and Zelterman estimators with application to a whale shark population. *Journal of the Royal Statistical Society: Series C, Applied Statistics*, *67*(1), 217–229. https://doi.org/10.1111/rssc.12219

Farcomeni A., & Dotto F. (2021). A correction to make Chao estimator conservative when the number of sampling occasions is finite. *Statistics & Probability Letters*, *176*, 109154. https://doi.org/10.1016/j.spl.2021.109154

Farcomeni A., & Maruotti A. (2024). Estimating the number of attacks to civilians in Ukraine: A quantitative analysis from media sources. *Peace Review*, *36*(1), 37–52. https://doi.org/10.1080/10402659.2023.2294063

Farcomeni A., & Scacciatelli D. (2013). Heterogeneity and behavioural response in continuous time capture-recapture, with application to street cannabis use in Italy. *The Annals of Applied Statistics*, *7*(4), 2293–2314. https://doi.org/10.1214/13-AOAS672

Fienberg S. E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika*, *59*, 591–603. https://doi.org/10.1093/biomet/59.3.591

Friedman J., Hastie T., & Tibshirani R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01

Goeman J. J. (2008). Autocorrelated logistic ridge regression for prediction based on proteomics spectra. *Statistical Applications in Genetics and Molecular Biology*, *7*(2), 2. https://doi.org/10.2202/1544-6115.1344

Gutiérrez-Sanín F., & Wood E. J. (2017). What should we mean by “pattern of political violence”? Repertoire, targeting, frequency, and technique. *Perspectives on Politics*, *15*(1), 20–41. https://doi.org/10.1017/S1537592716004114

Jamaluddine Z., Abukmail H., Aly S., Campbell O. M. R., & Checchi F. (2025). Traumatic injury mortality in the Gaza Strip from Oct 7, 2023, to June 30, 2024: A capture–recapture analysis. *The Lancet*, *405*(10477), 469–477. https://doi.org/10.1016/S0140-6736(24)02678-3

Karl R. A. (2019). Colombia’s history with “genocide”. *Journal of Genocide Research*, *21*(2), 255–262. https://doi.org/10.1080/14623528.2019.1599592

Lettink A., Chinapaw M., & van Wieringen W. N. (2023). Two-dimensional fused targeted ridge regression for health indicator prediction from accelerometer data. *Journal of the Royal Statistical Society: Series C, Applied Statistics*, *72*(4), 1064–1078. https://doi.org/10.1093/jrsssc/qlad041

Lum K., Price M. E., & Banks D. (2013). Applications of multiple systems estimation in human rights research. *The American Statistician*, *67*(4), 191–200. https://doi.org/10.1080/00031305.2013.821093

Madigan D., & York J. (1997). Bayesian methods for estimation of the size of a closed population. *Biometrika*, *84*(1), 19–31. https://doi.org/10.1093/biomet/84.1.19

Manrique-Vallier D. (2016). Bayesian population size estimation using Dirichlet process mixtures. *Biometrics*, *72*(4), 1246–1254. https://doi.org/10.1111/biom.12502

Manrique-Vallier D., Ball P., & Sadinle M. (2022). Capture-recapture for casualty estimation and beyond: recent advances and research directions. In A. L. Carriquiry, J. M. Tanur, W. F. Eddy, & M. L. Smykla (Eds.), *Statistics in the public interest: In memory of Stephen E. Fienberg* (pp. 15–31). Springer Nature.

Manrique-Vallier D., Ball P., & 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’, arXiv, arXiv:1906.04763, not peer reviewed.

Mitchell S., Ozonoff A., Zaslavsky A. M., Hedt-Gauthier B., Lum K., & Coull B. A. (2013). A comparison of marginal and conditional models for capture-recapture data with application to human rights violations data. *Biometrics*, *69*(4), 1022–1032. https://doi.org/10.1111/biom.12089

Nason G. P., & Bailey D. (2008). Estimating the intensity of conflict in Iraq. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *171*(4), 899–914. https://doi.org/10.1111/j.1467-985X.2008.00540.x

Obakrim S., Ailliot P., Monbet V., & Raillard N. (2024). EM algorithm for generalized Ridge regression with spatial covariates. *Environmetrics*, *35*(6), e2871. https://doi.org/10.1002/env.2871

Overstall A. M., King R., Bird S. M., Hutchinson S. J., & Hay G. (2014). Incomplete contingency tables with censored cells with application to estimating the number of people who inject drugs in Scotland. *Statistics in Medicine*, *33*(9), 1564–1579. https://doi.org/10.1002/sim.6047

Pledger S. (2005). The performance of mixture models in heterogeneous closed population capture-recapture. *Biometrics*, *61*(3), 868–873. https://doi.org/10.1111/j.1541-020X.2005.00411_1.x

Price M., & Ball P. (2014). Big data, selection bias, and the statistical patterns of mortality in conflict. *SAIS Review of International Affairs*, *34*(1), 9–20. https://doi.org/10.1353/sais.2014.0010

Racek D., Thurner P. W., & Kauermann G. (2026). Capturing the spatio-temporal diffusion effects of armed conflict: a nonparametric smoothing approach. *Journal of the Royal Statistical Society Series A: Statistics in Society*, qnaf120. https://doi.org/10.1093/jrsssa/qnaf120

Rivest L.-P., & Baillargeon S. (2007). Rcapture: Loglinear models for capture-recapture in R. *Journal of Statistical Software*, *19*(05), 1–31. https://doi.org/10.18637/jss.v019.i05

Rubio M. D., Kennedy E. H., & Jewell N. P. (2026). 'Doubly robust machine learning for population size estimation with missing covariates: Application to Gaza conflict mortality', arXiv, arXiv:2602.09911, preprint: not peer reviewed.

Sadinle M. (2018). Bayesian propagation of record linkage uncertainty into population size estimation of human rights violations. *The Annals of Applied Statistics*, *12*(2), 1013–1038. https://doi.org/10.1214/18-AOAS1178

Sangnawakij P., Lerdsuwansri R., Pijitrattana P., Schlattmann P., Maruotti A., & Böhning D. (2026). Two-way capture-recapture methods with emphasis on the bootstrap. *Statistical Methods & Applications*, *35*(1), 1–22. https://doi.org/10.1007/s10260-026-00840-5

Seybolt T. B., Aronson J. D., & Fischhoff B. (2013). *Counting civilian casualties: An introduction to recording and estimating nonmilitary deaths in conflict*. Oxford University Press.

Shi J., Yin W., Osher S., & Sajda P. (2010). A fast hybrid algorithm for large-scale L1-regularized logistic regression. *Journal of Machine Learning Research*, *11*, 713–741. https://doi.org/10.5555/1756006.1756029.

Silverman B. W. (2020). Multiple systems analysis for the quantification of modern slavery: Classical and Bayesian approaches (with discussion). *Journal of the Royal Statistical Society Series A: Statistics in Society*, *183*(3), 691–736. https://doi.org/10.1111/rssa.12505

Tancredi A., Steorts R. C., & Liseo B. (2020). A unified framework for de-duplication and population size estimation (with discussion). *Bayesian Analysis*, *15*(2), 323–384. https://doi.org/10.1214/19-BA1146

van Buuren S., & Groothuis-Oudshoorn K. (2011). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*, 1–67. https://doi.org/10.18637/jss.v045.i03

van Wieringen W. N. (2019). The generalized ridge estimator of the inverse covariance matrix. *Journal of Computational and Graphical Statistics*, *28*(4), 932–942. https://doi.org/10.1080/10618600.2019.1604374

Williams J. P., Hermansen G. H., Strand H., Clayton G., & Nygard H. M. (2024). Bayesian hidden Markov models for latent variable labeling assignments in conflict research: Application to the role ceasefires play in conflict dynamics. *The Annals of Applied Statistics*, *18*(3), 2034–2061. https://doi.org/10.1214/23-AOAS1869

Zhu Y., & Liu R. (2021). An algorithmic view of L2 regularization and some path-following algorithms. *Journal of Machine Learning Research*, *22*, 1–62. https://doi.org/10.5555/3546258.3546396