`singleRcapture`: an R package for estimation of population size using single-source capture-recapture models

Chlebicki Piotr[*], Beręsewicz Maciej[†]

30.04.2023
Paper prepared for the ESRA Early Career Researcher Award 2023

Abstract

Estimating population size is an important issue in official statistics, social sciences and natural sciences. One way to approach this problem is to use capture-recapture methods, which can be classified according to the number of sources used, the main distinction being between methods based on one source and those based on two or more sources.

In this paper, we focus on the former group, i.e. single-source capture-recapture (SSCR). SSCR models assume that observed counts follow truncated count distributions (e.g. zero-truncated Poisson, one-inflated zero-truncated geometric), and this assumption is used to estimate missing (hidden) zero counts. The literature includes applications of SSCR methods for estimating the number of undocumented migrants, cases of domestic violence or homeless people.

In this presentation we will introduce the `singleRcapture` R package for fitting SSCR models. The package implements state-of-the-art models as well as some new models proposed by the authors (e.g. extensions of zero-truncated one-inflated and one-inflated zero-truncated models). The software is intended for users interested in estimating the size of populations, particularly those that are difficult to reach or for which information is available from only one source and dual/multiple system estimation cannot be used.

The paper contains two empirical examples of the size of two hard-to-reach populations: drivers under the influence of alcohol or drugs and drivers with revoked driving licence.

[*]Corresponding author piochl@st.amu.edu.pl, Adam Mickiewicz University in Poznań, Faculty of Mathematics and Computer Science, ul. Wieniawskiego 1, 61-712 Poznań

[†](1) Poznań University of Economics and Business, Department of Statistics, Al. Niepodległości 10, 61-875 Poznań; (2) Statistical Office in Poznań, Poland.

Acknowledgements

This work is supported by the National Science Center, OPUS 22 grant no. 2020/39/B/-HS4/00941.

Data and codes are freely available at Github repository: https://github.com/ncn-foreigners/paper-esra-conf. The `singleRcapture` package can be installed from repository: https://github.com/ncn-foreigners/singleRcapture.

We would like to thank Polish Police for providing data for the example. Additionally, we would like to thank Peter van der Heijden, Maarten Cruyff and Dankmar Böhning for helpful comments on the early stages of development of the package and methods

Contents

1 Intoduction

Estimating population size is an important issue in official statistics, social sciences and natural sciences. One way to approach this problem is to use capture-recapture methods, which can be classified according to the number of sources used, the main distinction being between methods based on one source and those based on two or more sources.

In this paper, we focus on the former group, i.e. single-source capture-recapture (SSCR). SSCR models assume that observed counts follow truncated count distributions (e.g. zero-truncated Poisson, one-inflated zero-truncated geometric), and this assumption is used to estimate missing (hidden) zero counts. The literature includes applications of SSCR methods for estimating the number of undocumented migrants, cases of domestic violence or homeless people.

In this presentation we will introduce the `singleRcapture` R package for fitting SSCR models. The package implements state-of-the-art models as well as some new models proposed by the authors (e.g. extensions of zero-truncated one-inflated and one-inflated zero-truncated models). The software is intended for users interested in estimating the size of populations, particularly those that are difficult to reach or for which information is available from only one source and dual/multiple system estimation cannot be used.

The structure of the paper is as follows. In section 2 we present SSCR methods based on zero- and zero-one truncated count distributions, and recently introduced extensions that allow for one-inflation. We also present extensions proposed by the authors and describe methods for estimating the variance of the population size. Section 3 contains the description of the *singleRcapture* package, which implements state-of-the-art SSCR methods as well as post-hoc procedures to verify the quality of the estimates. The package is developed using S3 methods, allowing users to easily use the package without having to learn new functions. Section 4 contains a description of two case studies relating to drunk drivers and drivers with revoked licences. Section 5 contains results of the estimated population size, which in the case of drunk or drugged drivers may be about 500k (about 2% of all drivers in Poland) and 20k drivers with a revoked driving licence. The paper ends with conclusions.

2 Single-source capture-recapture methods

2.1 Zero-truncated count distributions

The most popular common approach to estimating the unknown population size from multiple counts within a source is based on zero-truncated count distributions, see Cruyff and van der Heijden (2008), van der Heijden, Bustami et al. (2003) and van der Heijden, Cruyff et al. (2003). The zero-truncated probability mass function is defined as

$$\mathbb{P}(Y_k = y_k | Y_k > 0) = \frac{\mathbb{P}(Y_k = y_k)}{1 - \mathbb{P}(Y_k = 0)} \text{ for: } y_k > 0 \tag{2.1}$$

where Y_k is a random variable denoting the number of times the k-th unit was observed in the available register, with a count distribution such as geometric, Poisson, negative

binomial, etc., as defined below:

$$\mathbb{P}(Y=y)=\begin{cases}\frac{\lambda^y e^{-\lambda}}{y!} & \text{Poisson,}\\ \frac{\Gamma(y+\alpha^{-1})}{\Gamma(\alpha^{-1})y!}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda}\right)^{\alpha^{-1}}\left(\frac{\lambda}{\alpha^{-1}+\lambda}\right)^{y} & \text{negative binomial,}\\ \frac{\lambda^y}{(1+\lambda)^{y+1}} & \text{geometric.}\end{cases} \tag{2.2}$$

In the SSCR problems, we do not have information about the missing units, i.e. for how many units $y=0$ holds. In this case we can use generalised linear models assuming a count distribution (2.2) with zero truncation (2.1). Maximum likelihood can be used to estimate parameters (including covariate information) and then the Horwitz-Thompson type estimator of population size can be used:

$$\hat{N}=\sum_{k=1}^{N_{obs}}\frac{1}{1-\mathbb{P}\left(Y_k=0|\hat{\boldsymbol{\beta}},\boldsymbol{x}_k\right)} \tag{2.3}$$

where $\hat{\boldsymbol{\beta}}$ is a vector of fitted regression coefficients and $\boldsymbol{x}_k$ is a vector of covariates. The quantity $\frac{1}{1-\mathbb{P}\left(Y_k=0|\hat{\boldsymbol{\beta}},\boldsymbol{x}_k\right)}$ is sometimes referred to as the contribution of the k-th unit.

One of the assumptions that is often violated when using (2.1) is that the process leading to the observation of a given unit in principle always allows for more than one observation, which may not be the case if, for example, the 'capture' of a unit is meant to represent the police apprehending an undocumented immigrant, in which case there is a substantial probability that this person will be deported, or a drunk driver, in which case we cannot ignore the possibility that this person will lose their driving licence (at least temporarily).

For such scenarios Godwin and Böhning, 2017 (among others) recommended using a mixture distribution of (2.2) and Bernoulli and considered two cases:

$$\mathbb{P}_{new}(Y=y)=\begin{cases}\mathbb{P}(Y=0) & y=0,\\ \omega\left\{1-\mathbb{P}(Y=0)\right\}+(1-\omega)\mathbb{P}(Y=y) & y=1,\\ (1-\omega)\mathbb{P}(Y=y) & y>1,\end{cases} \tag{2.4}$$

$$\mathbb{P}_{new}(Y=y)=\begin{cases}\omega+(1-\omega)\mathbb{P}(Y=1) & \text{when: } y=1,\\ (1-\omega)\mathbb{P}(Y=y) & \text{when: } y\neq 1,\end{cases} \tag{2.5}$$

which, after truncation, can be expressed as zero-truncated one-inflated (2.6) and one-inflated zero-truncated (2.6) distributions:

$$\mathbb{P}_{new}(Y=y|Y>0)=\begin{cases}\omega+(1-\omega)\frac{\mathbb{P}(Y=y)}{1-\mathbb{P}(Y=0)} & y=1,\\ (1-\omega)\frac{\mathbb{P}(Y=y)}{1-\mathbb{P}(Y=0)} & y>1.\end{cases} \tag{2.6}$$

$$\mathbb{P}_{new}(Y=y|Y>0)=$$
$$\begin{cases}\frac{\omega}{1-(1-\omega)\mathbb{P}(Y=0)}+\frac{(1-\omega)}{1-(1-\omega)\mathbb{P}(Y=0)}\mathbb{P}(Y=1) & \text{when: } y=1\\ \frac{(1-\omega)}{1-(1-\omega)\mathbb{P}(Y=0)}\mathbb{P}(Y=y) & \text{when: } y>1\end{cases} \tag{2.7}$$

Then the estimation procedure is the same, i.e. the estimated probabilities of $Y_k = 0$ are substituted into the Hortwitz-Thompson type estimator presented in (2.3).

The main differences between (2.6) and (2.7) are:

- (2.6) corresponds to first truncating the distribution and then mixing it with binomial distribution, whereas in (2.7) we first mix the count data distribution and then truncate it.
- In the (2.6), the latent variable, which may partially correspond to, for example, undocumented immigrants being expelled from the country in which they reside, does not alter the probability of first detection, which is not the case in the (2.7) distribution.

In our particular example, it seems that the process generating the latent variable should only take effect after the driver has come into contact with the traffic police.

2.2 Other approaches

There are also other approaches to estimating population size in SSCR based on extending some lower bound estimates, such as

$$\hat{N} = N_{obs} + \frac{\boldsymbol{f}_1^2}{2\boldsymbol{f}_2} \qquad \text{Chao's estimator}$$

$$\hat{N} = \frac{N_{obs}}{1 - \exp\left(-2\frac{\boldsymbol{f}_2}{\boldsymbol{f}_1}\right)} \qquad \text{Zelterman's estimator}$$

where $\boldsymbol{f}_l$ denotes the number of units sampled exactly l times (i.e. $l = 1$ or $l = 2$). An extension of this approach for covariate information is described in Böhning and van der Heijden (2009) and Böhning et al. (2013). It is based on the use of logistic regression with auxiliary variables:

$$Z = \begin{cases} 0 & \text{if } Y = 1, \\ 1 & \text{if } Y = 2, \\ \text{ommited} & \text{otherwise}, \end{cases}$$

with link:

$$\text{logit}(p_k) = \ln\left(\frac{\lambda_k}{2}\right) = \boldsymbol{\beta}\mathbf{x}_k,$$

whose use is justified by the following property of the Poisson probability mass function (PMF):

$$\frac{\mathbb{P}(Y = 2|\lambda)}{\mathbb{P}(Y = 1|\lambda) + \mathbb{P}(Y = 2|\lambda)} = \frac{\frac{1}{2}\lambda^2 e^{-\lambda}}{\lambda e^{-\lambda} + \frac{1}{2}\lambda^2 e^{-\lambda}} = \frac{\frac{\lambda}{2}}{1 + \frac{\lambda}{2}}.$$

The generalised Chao's estimator of the population size is given by

$$\begin{aligned}\hat{N} &= N_{obs} + \sum_{k=1}^{\boldsymbol{f}_1+\boldsymbol{f}_2} \frac{1}{2\exp\left(\boldsymbol{x}_k\hat{\boldsymbol{\beta}}\right) + 2\exp\left(2\boldsymbol{x}_k\hat{\boldsymbol{\beta}}\right)} \\ &= N_{obs} - (\boldsymbol{f}_1 + \boldsymbol{f}_2) + \sum_{k=1}^{\boldsymbol{f}_1+\boldsymbol{f}_2} \left(1 + \frac{\mathbb{P}\left(y_k = 0|\hat{\lambda}_k\right)}{\mathbb{P}\left(y_k = 1|\hat{\lambda}_k\right) + \mathbb{P}\left(y_k = 2|\hat{\lambda}_k\right)}\right)\end{aligned} \tag{2.8}$$

where the summation from $k = 1$ to $\boldsymbol{f}_1 + \boldsymbol{f}_2$ is understood as a summation over units observed either once or twice, and $\mathbb{P}$ is the Poisson probability function. The generalised Zelterman estimator is given by

$$\hat{N} = \sum_{k=1}^{N_{obs}} \frac{1}{1 - \exp\left(-2\exp\left(\boldsymbol{x}_k\hat{\boldsymbol{\beta}}\right)\right)} \tag{2.9}$$

where the summation is over all units, which is just the estimate in zero-truncated Poisson with the non-standard method of estimating $\hat{\boldsymbol{\beta}}$. Note that:

1. Extended Zelterman and Chao's estimators can no longer be interpreted as lower bound estimates, since they are usually higher than zero-truncated Poisson estimates.

2. Extended Zelterman and Chao estimates usually suggest similar estimates (for a detailed comparison, see Böhning, 2010).

3. These models are particularly sensitive to the presence of one-inflation.

4. As expected, when we model logistic regression on Z as an intercept only, we get the usual Zelterman/Chao estimates.

2.3 Further extensions

We have extended relevant methods to *vector generalised linear models* (Yee, 2015) that allow for heterogeneity in α, in negative binomial-based models, or ω, in zero-truncated one-inflated (ZTOI) and zero-inflated zero-truncated (OIZT) type models, parameters, and implemented them in the package using the extended iteratively reweighted least squares (IRLS) algorithm, whose detailed explanation can be found in (Yee, 2015), with an additional option of non-standard link functions.

Another way of accounting for the modified probability of a single capture is to model the first capture as structurally different from the probabilities of other counts. This is achieved by modifying $\mathbb{P}(Y = 1)$ or $\mathbb{P}(Y = 1|Y > 0)$, the same problem encountered in one-inflated distributions, to an arbitrary probability $\pi \in (0, 1)$. The resulting probability is what we call a *a pseudo hurdle model*, since the idea is very similar to the hurdle models sometimes encountered in modelling count data whose PMF before truncation is described by:

$$\mathbb{P}_{new}(Y=y)=\begin{cases}\frac{1}{1-\mathbb{P}(Y=1)}\mathbb{P}(Y=0), & \text{when } y=0,\\ \pi\left(1-\frac{\mathbb{P}(Y=0)}{1-\mathbb{P}(Y=1)}\right) & \text{when } y=1,\\ (1-\pi)\frac{\mathbb{P}(Y=y)}{1-\mathbb{P}(Y=1)} & \text{when } y>1.\end{cases} \tag{2.10}$$

$$\mathbb{P}_{new}(Y=y)=\begin{cases}\pi & \text{when } y_i=1,\\ (1-\pi)\frac{\mathbb{P}(Y=y)}{1-\mathbb{P}(Y=1)} & \text{when } y\neq 1.\end{cases} \tag{2.11}$$

which after truncation become respectively:

$$\mathbb{P}_{new}(Y=y|Y>0)=\begin{cases}\pi & \text{when } y=1,\\ (1-\pi)\frac{\mathbb{P}(Y=y)}{1-\mathbb{P}(Y=0)-\mathbb{P}(Y=1)} & \text{when } y>1.\end{cases} \tag{2.12}$$

$$\mathbb{P}_{new}(Y=y|Y>0)=\begin{cases}\pi\frac{1-\mathbb{P}(Y=1)}{1-(1-\pi)\mathbb{P}(Y=0)-\mathbb{P}(Y=1)} & \text{when } y=1,\\ (1-\pi)\frac{\mathbb{P}(Y=y)}{1-(1-\pi)\mathbb{P}(Y=0)-\mathbb{P}(Y=1)} & \text{when } y>1.\end{cases} \tag{2.13}$$

We use the same naming scheme for hurdle models, i.e. ztHurdle and hurdleZt, as for one-inflated models to express a similar idea about the order of truncation and modification of $\mathbb{P}(Y=1)$ or $\mathbb{P}(Y=1|Y>0)$. The main advantages of pseudo-hurdle models over one-inflated models are:

- one-deflation and lack of change in probability of $Y=1$, are naturally included in hurdle models, whereas if we were to consider them in the context of a one-inflated model, we would run into a boundary problem, since not all values of ω result in (2.5), (2.4) and (2.7), where (2.6) are probability mass functions,
- Some preliminary simulations suggest that pseudo-hurdle models are more robust to violations of the model assumptions.
- It is possible to fit (2.12) by a two-step procedure, first fitting a binary outcome regression on $I(Y=1)$ and then fitting a zero one truncated regression on $Y|Y>1$, which is significantly less computationally demanding than the rest of the described models which take into account altered $\mathbb{P}(Y=1|Y>0)$. While this is not much of a concern if we have a package that does the fitting internally, it can be a significant advantage if, for example, one were to extend SSCR models to include regression with functional data.

While the main disadvantage stems from the interpretability of the 'inflation factor' ω versus the interpretation of the probability parameter π and how its presence changes the interpretation of the parameters of the 'base distribution'. It seems that, at least in the context of this paper, the one-inflated model interpretation is more convenient.

2.4 Model and variable selection and interpretation of parameters

If we fit regressions to all available distributions with covariate information on all relevant parameters, a selection is required. In this paper we use the Bayesian Information

Criterion (BIC) to select the appropriate model and covariates.

The recommended procedure associated with parameter interpretation is to choose log link for λ parameters and complementary-log-log link function for ω parameters. The use of complementary-log-log on ω requires elaboration, as it has not been used (to the authors' knowledge) in this setting before.

From properties of `cloglog` link we have:

$$\exp(\boldsymbol{\beta}\mathbf{x}) = -\ln(1-\omega) = -\log(\text{Survival function}(\mathbf{x}))$$

$$-\frac{d}{dx_k}\log(\text{Survival function}(\mathbf{x})) = \frac{d}{dx_k}\exp(\boldsymbol{\beta}\mathbf{x}) = \exp(\beta_k x_k),$$

for some survival function, here the second equation describes the hazard function. Since in ZTOI and OIZT models the parameter ω corresponds to a binary latent variable describing the 'death' of the process (e.g. Poisson process) that generates counts of the dependent variable Y. For example, if, after an encounter with the police, a drunk driver loses their ability to drive (e.g. loses their driving licence) or changes in some way that makes it impossible for the police to arrest him or her again (e.g. refrains from drinking), the process 'dies'. Again, the difference between the ZTOI and OIZT models is that this survival process begins at different stages in ZTOI after the first arrest and in OIZT before the first observation.

Regression coefficients associated with λ can be interpreted in much the same way as they are interpreted using ordinary Poisson regression for cases where the 'death' of the Poisson-like process does not occur, i.e. the latent variable with distribution $b(\omega)$ takes the value 0. Here there is no difference between ZTOI and OIZT, since they both collapse to their respective zero-truncated distributions when we know that the Poisson-like process does not halt.

2.5 Confidence interval construction and bootstraps

The commonly used analytical approximation for the variance of the estimate $\hat{N}$ is done in two steps. First, the variance is decomposed into two components by the law of total variance:

$$\text{var}(\hat{N}) = \mathbb{E}\left(\text{var}\left(\hat{N}|I_1,\ldots,I_n\right)\right) + \text{var}\left(\mathbb{E}(\hat{N}|I_1,\ldots,I_n)\right), \tag{2.14}$$

where I_k are binary variables indicating whether the kth unit was recorded at least once. Assuming $\hat{N}$ is normally distributed, we can obtain the first part of (2.14) using the multivariate δ method:

$$\mathbb{E}\left(\text{var}\left(\hat{N}|I_1,\ldots,I_n\right)\right) = \left(\frac{\partial(N|I_1,\ldots,I_N)}{\partial\boldsymbol{\beta}}\right)^T \text{cov}\left(\hat{\boldsymbol{\beta}}\right)\left(\frac{\partial(N|I_1,\ldots,I_N)}{\partial\boldsymbol{\beta}}\right)\Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$$

Second part of (2.14) is accurately approximated by:

$$
\begin{aligned}
\operatorname{var}\left(\mathbb{E}(\hat{N}|I_1,\ldots,I_n)\right) &= \operatorname{var}\left(\sum_{k=1}^{N}\frac{I_k}{\mathbb{P}(Y_k>0)}\right)\\
&= \sum_{k=1}^{N}\operatorname{var}\left(\frac{I_k}{\mathbb{P}(Y_k>0)}\right)\\
&= \sum_{k=1}^{N}\frac{1}{\mathbb{P}(Y_k>0)^2}\operatorname{var}(I_k)\\
&= \sum_{k=1}^{N}\frac{1}{\mathbb{P}(Y_k>0)^2}\mathbb{P}(Y_k>0)(1-\mathbb{P}(Y_k>0))\\
&= \sum_{k=1}^{N}\frac{1}{\mathbb{P}(Y_k>0)}(1-\mathbb{P}(Y_k>0))\\
&\approx \sum_{k=1}^{N}\frac{I_k}{\mathbb{P}(Y_k>0)^2}(1-\mathbb{P}(Y_k>0))\\
&= \sum_{k=1}^{N_{obs}}\frac{1-\mathbb{P}(Y_k>0)}{\mathbb{P}(Y_k>0)^2},
\end{aligned}
$$

where the second line is an approximation instead of equality, since in the 5th line of the sequence above we were summing over all units including those we did not observe, we used an unbiased estimate to approximate this quantity.

If the distribution of $\hat{N}$ is indeed asymptotically normal, this leads to a studentized confidence interval of the form, for the significance level α:

$$
\left(\max(N_{obs},\hat{N}-z\left(1-\frac{\alpha}{2}\right)\operatorname{sd}(\hat{N}),\hat{N}+z\left(1-\frac{\alpha}{2}\right)\operatorname{sd}(\hat{N}))\right) \tag{2.15}
$$

An important drawback of this confidence interval is its symmetry around $\hat{N}$. It is often argued in the literature, see Böhning and van der Heijden, 2019; Chao, 1989, on SSCR that statistic $\hat{N}$ may have skew distribution. An alternative method of constructing confidence interval assumes normality of $\ln\left(\hat{N}-N_{obs}\right)$ which leads to confidence interval of the following form:

$$
\left(N_{obs}+\frac{\hat{N}-N_{obs}}{G},N_{obs}+\left(\hat{N}-N_{obs}\right)G\right), \tag{2.16}
$$

where $G=\exp\left(z\left(1-\frac{\alpha}{2}\right)\sqrt{\ln\left(1+\frac{\hat{\operatorname{var}}(\hat{N})}{\left(\hat{N}-N_{obs}\right)^2}\right)}\right)$. But this approach does not adjust the variance estimation for lack of normality and more importantly it is argued that the variance estimation itself underestimates variance of $\hat{N}$.

Apart from the standard non-parametric bootstrap, which needs no introduction, there are two other bootstrap types implemented in the `singleRcapture` package:

Algorithm 1 Parametric Bootstrap

1: $B \leftarrow \{\}$
2: Compute $\hat{N}$ for the model and given data and round it to nearest integer, as well as each units contribution to $\hat{N}$ marked as N_i for i-th unit
3: "Integeraze" all N_i's by transformation

$$N_i \leftarrow \lfloor N_i \rfloor + \mathcal{T}_i \text{ where: } \mathcal{T}_i \sim b(N_i - \lfloor N_i \rfloor)$$

4: $\boldsymbol{P} \leftarrow \left(N_i/\hat{N}\right)_{i=1}^{N_{obs}}$
5: **for** $k \in 1 : M$ **do**
6: Choose $\hat{N}$ vectors $\boldsymbol{x}_k$ from N_{obs} of vectors of additional information according to vector of probabilities $\boldsymbol{P}$, and additional weights associated with observations chosen if any were given.
7: Generate data $\boldsymbol{y}_{new}$ depending on probability model associated with chosen estimator $\hat{N}$ given the linear predictors $\boldsymbol{\eta}_{new} = \boldsymbol{X}_{new}\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}}$ is a vector of fitted regression parameters, using functions like `rpois`.
8: Truncate any zeros obtained in generation,
9: Fit regression on new data and obtain new estimate B_k of N for current bootstrap sample based on new fitted vector of regression parameters $\hat{\boldsymbol{\beta}}_{new}$ and append results to B.
10: **end for**
11: Given pseudo-sample B estimate variance and construct confidence interval for a given significance level depending on type of confidence interval chosen.

which is a single source extension of procedures described in Norris and Pollock, 1996; Zwane and van der Heijden, 2003.

An even more common bootstrap procedure that was utilised in Böhning and van der Heijden, 2019 (among others) is semi-parametric bootstrap which can be roughly described by the following algorithm:

Algorithm 2 Semi-parametric bootstrap

1: Define probability distribution by:

$$\frac{|\hat{\boldsymbol{f}}_0|}{|\hat{N}|}, \frac{\boldsymbol{f}_1}{|\hat{N}|}, \frac{\boldsymbol{f}_2}{|\hat{N}|}, \ldots, \frac{\boldsymbol{f}_{\max(\boldsymbol{y})}}{|\hat{N}|} \tag{2.17}$$

2: Generate $\lfloor \hat{N} \rfloor$ units according to distribution in (2.17) and delete those for which $y = 0$ holds.
3: Sample vectors of independent variables $\boldsymbol{x}$ for each unit with uniform distribution from units that have been captured exactly as many times as the generated y value.
4: Estimate $\hat{\boldsymbol{\beta}}^*$ and subsequently $\hat{N}^*$ using generated data.
5: Repeat until the desired number of statistics $\hat{N}^*$ is reached.

When using the 2 algorithm we assume that the estimate $\hat{N}$ is approximately correct, while when using the 1 algorithm we assume that the entire probabilistic model we have constructed is approximately correct. If these assumptions are not violated, bootstrapping will provide a more reliable confidence interval.

3 Implementation in the `singleRcapture` package

3.1 Description of functionalities

We implemented the following features in `singleRcapture` package (Chlebicki and Beręsewicz, 2023) in the R programming language (R Core Team, 2022). The main function `estimatePopsize` is responsible for some data cleaning regression fitting and population size estimation similar to `glm` function from the base R.

The `estimatePopsize` function accepts the following models as `model` argument:

- Zero-truncated regression – poisson, geometric, negative binomial
- Zero one truncated regression (Böhning and van der Heijden, 2019) – poisson, geometric, negative binomial
- Zero-truncated one-inflated regression – poisson, geometric
- One-inflated zero-truncated regression – poisson, geometric
- Pseudo hurdle zero-truncated regression – poisson, geometric
- Zero-truncated Pseudo hurdle regression – poisson, geometric
- Generalized Chao (Böhning et al., 2013) and Zelterman (Böhning and van der Heijden, 2009) models based on logistic regression

with negative binomial extensions for the remaining models currently being constructed.

For confidence interval construction and variance estimation (argument `popVar`):

- Analytical approxmiation – the default option chosen if no `popVar` argument was provided on call
- Nonparametric bootstrap – which requires no introduction
- Semiparametric bootstrap
- Parametric bootstrap

with commonly used types of bootstrap confidence intervals (eg. percentile, studentized, basic) that can be specified in `controlPopVar` argument.

Post-hoc procedures such as:

- Marginal frequencies and testing fit of marginal distributions by G and χ^2 tests in function `marginalFreq` and `summary.singleRmargin` method respectively.
- Estimation of population size for user defined sub-populations in `stratifyPopsize` function.

Other notable features consist of:

- A plethora of control options explained in the package documentation in functions that create control arguments: `controlMethod, controlModel, controlPopVar`.
- Extraction of information criteria such as `AIC`, `BIC`, `AICc` and leave-one-out diagnostics for regression and population size in `dfbeta` and `dfpopsize` functions respectively.
- Rootograms to assess the fit (Kleiber & Zeileis, 2016),
- Various diagnostic plots in `plot.singleRmethod` and `glm` like detailed output in `summary.singleR` method.
- Methods for `sandwich` (described in Zeileis, 2004 for `glm` class) estimates of cov $\left(\hat{\boldsymbol{\beta}}\right)$ and `redoPopEstimation` function for updating population size estimation with these updated covariances in. Robust regression being planned for the near future.

3.2 Example usage

To demonstrate `singleRcapture` in practice, we recreate a zero-truncated Poisson model from van der Heijden, Bustami et al. (2003). The data from this publication included in `singleRcapture` concern undocumented immigrants in four cities in the Netherlands, the `capture` column denotes a number of times a given immigrant was apprehended by the police `age` is the age of a given person, coded as either below or above 40 years old, `reason` is information on why a person was apprehended by the police, either for being in the Netherlands illegally or for some other reason, and finally `gender` is self-explanatory (here with only two levels, male and female). More information on the data can be found in the original paper.

The code and output for creating this model in `singleRcapture` framework is presented below:

```
model <- estimatePopsize(
  formula = capture ~ gender + age + nation + reason,
  data = netherlandsimmigrant,
  popVar = "analytic",
  model = "ztpoisson",
  method = "IRLS"
)
summary(model)
#>
#> Call:
#> estimatePopsize(
#>     formula = capture ~ gender + age + nation + reason,
#>     data = netherlandsimmigrant, model = "ztpoisson",
#>     method = "IRLS", popVar = "analytic")
#>
#> Pearson Residuals:
#>      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
#> -0.488779 -0.486085 -0.297859  0.002075 -0.210439 13.921578
```

```
#>
#> Coefficients:
#> -----------------------
#> For linear predictors associated with: lambda
#>                       Estimate Std. Error z value  P(>|z|)
#> (Intercept)          -1.33179    0.25486  -5.226 1.74e-07 ***
#> gendermale            0.39741    0.16305   2.437 0.014796 *
#> age>40yrs            -0.97463    0.40824  -2.387 0.016969 *
#> nationAsia           -1.09241    0.30164  -3.622 0.000293 ***
#> nationNorth Africa    0.18997    0.19400   0.979 0.327471
#> nationRest of Africa -0.91129    0.30097  -3.028 0.002463 **
#> nationSurinam        -2.33665    1.01357  -2.305 0.021146 *
#> nationTurkey         -1.67453    0.60291  -2.777 0.005479 **
#> reasonOther reason   -0.01093    0.16153  -0.068 0.946048
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> AIC: 1714.896
#> BIC: 1764.747
#> Residual deviance: 1128.549
#>
#> Log-likelihood: -848.4481 on 1871 Degrees of freedom
#> Number of iterations: 8
#> -----------------------
#> Population size estimation results:
#> Point estimate 12691.45
#> Observed proportion: 14.8% (N obs = 1880)
#> Std. Error 2809.508
#> 95% CI for the population size:
#>           lowerBound upperBound
#> normal      7184.917   18197.99
#> logNormal   8430.749   19723.38
#> 95% CI for the share of observed population:
#>           lowerBound upperBound
#> normal     10.330814   26.16592
#> logNormal   9.531836   22.29932
```

Section `For linear predictors associated with:lambda` refers to zero-truncated Poisson regression parameters, and section `Population size estimation results` provides results on the size of undocumented residing immigrants in Netherlands in 1995.

4 Data on drivers in Poland

For this study, we obtained anonymised unit-level data from Polish Police on two hard-to-identify populations i.e. drunk or intoxicated drivers and drivers after revocation

of driving licence. Members of this population were identified during roadside checks or accidents.

According to the law, these individuals violated the following provisions of the Polish Criminal Code:

- Article 178a. – driving while intoxicated or under the influence of an intoxicant,
- Article 180a. – driving after revocation of driving licence.

Table 1 shows the number of people identified by the Polish Police for driving after revocation, or being drunk or intoxicated by the number of process registrations in 2021.

Table 1: Number of people identified by the Polish Police for driving after revocation, or being drunk or intoxicated by the number of process registrations in 2021

Number of registrations (captures)	Number of people	
	revocation	drunk or intoxicated
0	?	?
1	4,255	53,384
2	400	1,876
3	63	188
4	20	27
5	5	8
6	1	–
7	1	–
8	1	–
9	1	1

Note: Legal basis: Criminal Code: 1) Article 180a. Driving after revocation; 2) Article 178a. Driving while drunk or under the influence of drugs. Source: Krajowy System Informacyjny Policji (KSIP). Symbols: – means that the phenomenon did not occur, and ? means unknown quantity.

In 2021 in Poland, 4,255 people driving after a revocation were identified by Polish Police only once, 400 twice, and so on. While over 53k for driving while being drunk or intoxicated only once, over 1,8k twice, and so on. The unknown quantity is so-called *dark number* i.e. how many members of these populations were not identified by the Police.

The aim of the study is to estimate the size of these two populations based on police data and SSCR models.

5 Results

5.1 Results for drunk or intoxicated driving

Based on BIC we selected zero-truncated one-inflated geometric model with the following formulas for distribution parameters depending on two variables: gender and age group defined as below:

$$\begin{aligned}\ln \lambda &= \beta_{00} + \beta_{01} \cdot I(\text{gender} = \text{male}) \\ \ln(-\ln(1-\omega)) &= \beta_{10} + \beta_{11} \cdot I(30 < \text{age} \leq 40) + \beta_{12} \cdot I(40 < \text{age} \leq 50) \\ &\quad + \beta_{13} \cdot I(50 < \text{age} \leq 60) + \beta_{14} \cdot I(60 < \text{age})\end{aligned}$$

The following code was used to fit the final model:

```
modelZtoiGeom <- estimatePopsize(
  formula = counts ~ gender,
  model = ztoigeom(omegaLink = "cloglog"),
  # log link is the default link for lambda
  # and can be ommited when specifying distribution
  data = ex1,
  method = "IRLS",
  popVar = "bootstrap",
  controlMethod = controlMethod(verbose = 5,
                                saveIRLSlogs = TRUE),
  controlModel  = controlModel(omegaFormula = ~ age),
  controlPopVar = controlPopVar(
    bootType = "semiparametric",
    B = 5000,
    bootstrapVisualTrace = TRUE,
    traceBootstrapSize = TRUE
  )
)
summary(modelZtoiGeom)
```

Since χ^2 and G tests suggest[1] that our model does not accurately model marginal frequencies:

```
> summary(marginalFreq(modelZtoiGeom), dropl5 = "group", df = 1)
Test for Goodness of fit of a regression model:

                 Test statistics df P(>X^2)
Chi-squared test           16.32  1 5.3e-05
G-test                     12.49  1 4.1e-04

-----------------------------------------------------------------
Cells with fitted frequencies of < 5 have been grouped
Names of cells used in calculating test(s) statistic: 1 2 3 4
```

We opted for the use of semi-parametric bootstrap with 5000 samples and the results were as follows[2]:

[1] We specified that cells with low fitted frequencies should be merged into one cell 4+ in `dropl5` parameter supplied to method for `summary` function.

[2] Here we omit the part of output which describes 'Call' since it was already provided

```
Pearson Residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.20295 -0.20295 -0.19527  0.00009 -0.16392 33.43030

Coefficients:
-----------------------
For linear predictors associated with: lambda
            Estimate Std. Error z value  P(>|z|)
(Intercept)  -2.9365     0.1438 -20.414  < 2e-16 ***
gender: Male         0.9149     0.1324   6.909 4.89e-12 ***
-----------------------
For linear predictors associated with: omega
                  Estimate Std. Error z value  P(>|z|)
(Intercept)     0.006536   0.069092   0.095    0.925
age (30,40]    -0.033629   0.055131  -0.610    0.542
age (40,50]     0.039918   0.057775   0.691    0.490
age (50,60]     0.320366   0.064458   4.970 6.69e-07 ***
age 60+         0.501869   0.076941   6.523 6.90e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

AIC: 19425.81
BIC: 19488.26
Residual deviance: 13330.21

Log-likelihood: -9705.903 on 110763 Degrees of freedom
Number of iterations: 8
-----------------------
Population size estimation results:
Point estimate 513452.9
Observed proportion: 10.8% (N obs = 55385)
Boostrap sample skewness: 0.3442703
0 skewness is expected for normally distributed variable
---
Bootstrap Std. Error 34340.56
95% CI for the population size:
lowerBound upperBound
  453699.7   589443.4
95% CI for the share of observed population:
lowerBound upperBound
  9.396153  12.207414
```

Analytical approximation of variance suggests a very similar confidence interval for population size as presented below:

```
Population size estimation results:
Point estimate 513452.9
Observed proportion: 10.8% (N obs = 55385)
```

```
Std. Error 30387.65
95% CI for the population size:
          lowerBound upperBound
normal      453894.2   573011.6
logNormal   457661.1   576982.5
95% CI for the share of observed population:
          lowerBound upperBound
normal      9.665599   12.20218
logNormal   9.599078   12.10175
```

Finally, a visual inspection of the rootogram presented in Plot 1 suggests that the model fits the data well.

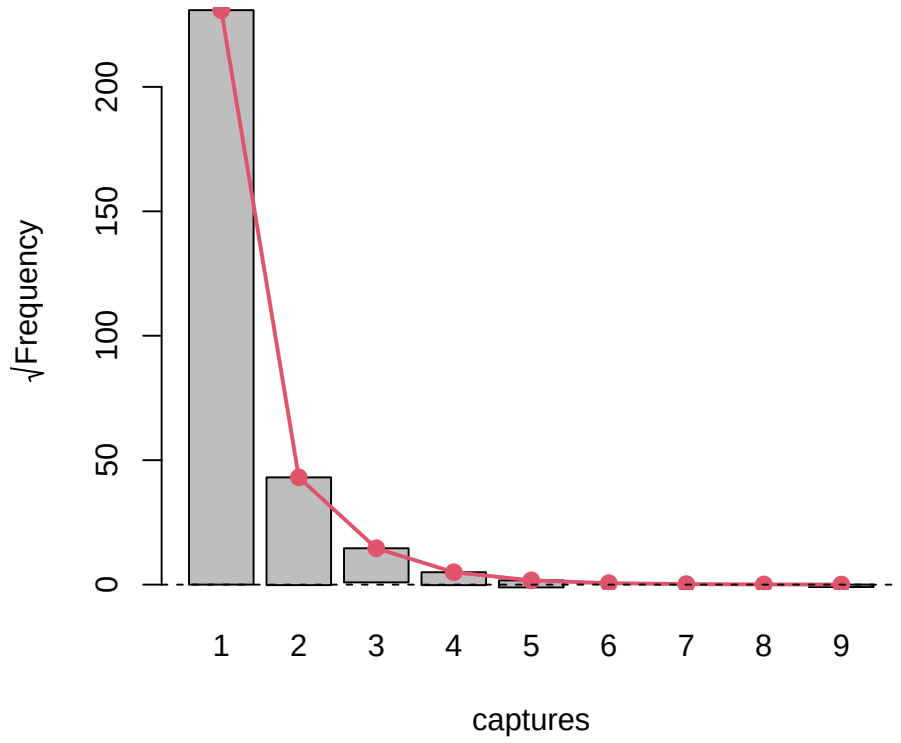

Figure 1: Rootogram suggests that despite χ^2 and G tests rejecting the hypothesis that marginal distribution of `count` is accurately described by out model, the difference are not very noticeable.

Based on this output, we estimate that the population of drunk or intoxicated drivers in Poland in 2021 is 95% between 454k and 573k with the point estimate around 513k. This means that this population represents around 2% of all drivers in Poland[3].

[3]We compare to the number of driving licences in Poland as we do not have an estimate of the number of active drivers.

There are two significant age groups in terms of the likelihood of losing one's driving ability (e.g. losing one's driving licence) after being arrested by the police, or changing one's behaviour in a way that makes it impossible to be arrested again by the police (e.g. refraining from driving under the influence of alcohol): people aged 50 to 60 and people over 60. Since:

$$\exp(0.320366) \cdot 100\% \approx 137.76\% \quad \exp(0.501869) \cdot 100\% \approx 165.18\%$$

Our model suggests that the probability of a process described above occurring is greatest for people over 60 years of age and is increased by about 65.18% of what it is for bellow 50 year olds, the other group that differs from the under 50's are the 50 to 60 year olds whose probability is increased by about 37.76% of what it is for bellow 50 year olds.

5.2 Results for driving after a revocation

For this example, we were unable to distinguish between zero-truncated one-inflated (*ztoigeom*) and one-inflated zero-truncated (*oiztgeom* geometric by information criteria. Codes to fit these models is presented below:

```
modelZtoiGeom <- estimatePopsize(
  formula = counts ~ 1,
  model = ztoigeom(omegaLink = "cloglog"),
  data = ex2,
  method = "IRLS",
  popVar = "bootstrap",
  controlMethod = controlMethod(verbose = 5, saveIRLSlogs = TRUE),
  controlModel  = controlModel(omegaFormula = ~ age),
  controlPopVar = controlPopVar(
    bootType = "semiparametric",
    B = 5000,
    bootstrapVisualTrace = TRUE,
    traceBootstrapSize = TRUE
  )
)
```

```
modelOiztGeom <- estimatePopsize(
  formula = counts ~ 1,
  model = oiztgeom(omegaLink = "cloglog"),
  data = ex2,
  method = "IRLS",
  popVar = "bootstrap",
  controlMethod = controlMethod(verbose = 5, saveIRLSlogs = TRUE),
  controlModel  = controlModel(omegaFormula = ~ age),
  controlPopVar = controlPopVar(
    bootType = "semiparametric",
    B = 5000,
    bootstrapVisualTrace = TRUE,
```

```
    traceBootstrapSize = TRUE
  )
)
```

Hence, we decided to report both of them since the choice between these two must be made using more subtle criteria

```
> summary(modelZtoiGeom)

Pearson Residuals:
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-0.358476 -0.319967 -0.281719 -0.001699 -0.245724 15.322633

Coefficients:
-----------------------
For linear predictors associated with: lambda
            Estimate Std. Error z value P(>|z|)
(Intercept) -1.30609    0.09775  -13.36  <2e-16 ***
-----------------------
For linear predictors associated with: omega
                 Estimate Std. Error z value  P(>|z|)
(Intercept)   -0.8355     0.2386  -3.502 0.000462 ***
age (30,40]    0.3853     0.2072   1.859 0.062973 .
age (40,50]    0.6974     0.2145   3.251 0.001151 **
age (50,60]    0.9548     0.2341   4.078 4.55e-05 ***
age 60+        1.1936     0.2506   4.763 1.91e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

AIC: 3773.169
BIC: 3811.959
Residual deviance: 2279.36

Log-likelihood: -1880.584 on 9486 Degrees of freedom
Number of iterations: 7
-----------------------
Population size estimation results:
Point estimate 22266.95
Observed proportion: 21.3% (N obs = 4746)
Boostrap sample skewness: 0.6137086
0 skewness is expected for normally distributed variable
---
Bootstrap Std. Error 2078.284
95% CI for the population size:
lowerBound upperBound
  18678.40   26834.82
95% CI for the share of observed population:
lowerBound upperBound
```

```
  17.68597   25.40902

> summary(modelOiztGeom)

Pearson Residuals:
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-0.358476 -0.319967 -0.281719 -0.001699 -0.245724 15.322633

Coefficients:
-----------------------
For linear predictors associated with: lambda
            Estimate Std. Error z value P(>|z|)
(Intercept) -1.30609    0.09775  -13.36  <2e-16 ***
-----------------------
For linear predictors associated with: omega
                 Estimate Std. Error z value  P(>|z|)
(Intercept)   -2.2119     0.3362  -6.579 4.75e-11 ***
age (30,40]    0.4633     0.2454   1.888 0.059055 .
age (40,50]    0.8625     0.2566   3.361 0.000776 ***
age (50,60]    1.2116     0.2883   4.202 2.65e-05 ***
age 60+        1.5535     0.3196   4.861 1.17e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

AIC: 3773.169
BIC: 3811.959
Residual deviance: 2279.36

Log-likelihood: -1880.584 on 9486 Degrees of freedom
Number of iterations: 6
-----------------------
Population size estimation results:
Point estimate 13250.41
Observed proportion: 35.8% (N obs = 4746)
Boostrap sample skewness: 0.9612116
0 skewness is expected for normally distributed variable
---
Bootstrap Std. Error 1909.052
95% CI for the population size:
lowerBound upperBound
  10418.77   17893.71
95% CI for the share of observed population:
lowerBound upperBound
  26.52329   45.55239
```

However, based on the interpretation of the distribution we again opted for the zero-truncated one-inflated geometric model with the following formulas for distribution parameters:

$$\begin{aligned} \ln \lambda &= \beta_{00} \qquad \text{Intercept-only} \\ \ln(-\ln(1-\omega)) &= \beta_{10} + \beta_{11} \cdot I(30 < \text{age} \leq 40) + \beta_{12} \cdot I(40 < \text{age} \leq 50) \\ &\quad + \beta_{13} \cdot I(50 < \text{age} \leq 60) + \beta_{14} \cdot I(60 < \text{age}) \end{aligned}$$

Firstly, it seems that the halting of data collection process, described previously, in case of police data is more in line with zero-truncated one-inflated model and secondly here estimates in zero-truncated one-inflated model are less affected by deletion of units from the original data making the estimates more reliable.

Since again we see that marginal distribution is not well approximated by our model (and hence it cannot be correct in modeling the whole distribution):

```
> summary(marginalFreq(modelZtoiGeom), df = 1, dropl5 = "group")
Test for Goodness of fit of a regression model:

                 Test statistics df P(>X^2)
Chi-squared test            7.60  1  0.0058
G-test                      7.62  1  0.0058

------------------------------------------------------------
Cells with fitted frequencies of < 5 have been grouped
Names of cells used in calculating test(s) statistic: 1 2 3 4
```

We performed semi-parametric bootstrap to construct confidence interval for population size. Here we see that analytic approximation likely underestimates the standard error of $\hat{N}$ when compared to semi-parametric bootstrap.

```
Population size estimation results:
Point estimate 22266.95
Observed proportion: 21.3% (N obs = 4746)
Std. Error 1736.54
95% CI for the population size:
          lowerBound upperBound
normal      18863.39   25670.50
logNormal   19180.41   26013.48
95% CI for the share of observed population:
          lowerBound upperBound
normal      18.48815   25.15985
logNormal   18.24439   24.74399
```

Similarly as in the previous example, rootogram presented in 2 suggest that the selected distribution fits the data well.

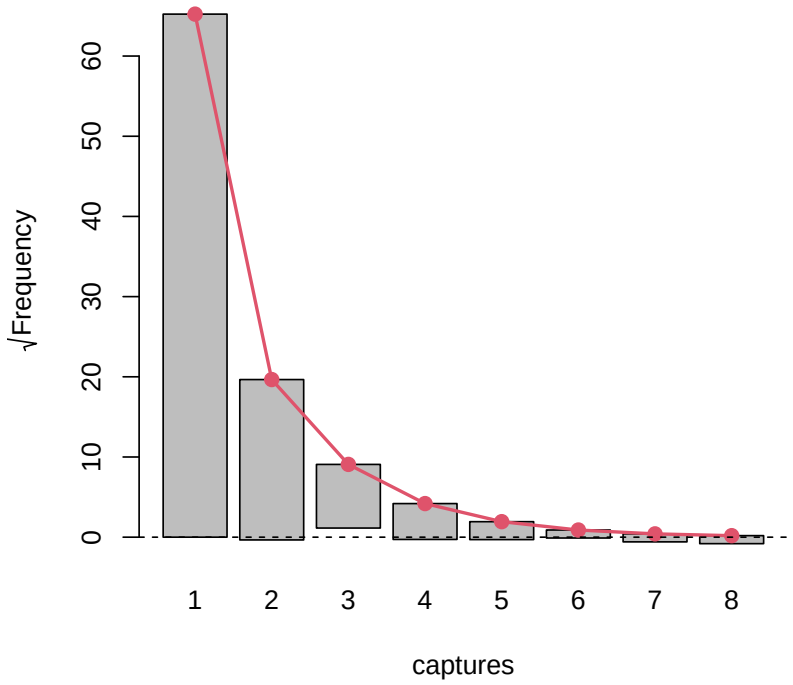

Figure 2: Rootogram for our model on people who drive without a valid licence. Here we see more of a discrepancy between empirical and fitted marginal distribution than in previous case.

This model suggests the around 22k people in Poland drives after their driving licence was revoked and the 95% confidence interval is between 18-26k. Unfortunately, we do not have information how many driving licences were revoked in 2021 or in previous years. Therefore, we cannot indicate what share of these people drive without driving licence. Based on this model we can indicate that the police identified about $1/5$ of members of this population.

Additionally, we see that no available covariate information significantly impacts the Poisson parameter and that advanced age significantly increases the probability that the Poisson like process halts either by behavioral change occurring in the observed unit or as a result of some more intrusive procedure such as arrest.

People in the $(40, 50]$ age bracket are more than twice as likely to undergo such a halting procedure ($\approx 236\%$ of what it is for bellow 30 year olds) for $(40, 50]$ age bracket the probability is around $\approx 335\%$ of what it is for bellow 30 year olds (which is about 10%) and this number comes to about $\approx 473\%$ for the elderly (above 60 years old).

6 Conclusions

The paper presents the `singleRcapture` – an R package for fitting SSCR models. It should be emphasised that the package implements not only state-of-the-art models, but also new models proposed by the authors (e.g. extensions of zero-truncated one-inflated and one-inflated zero-truncated models) but not yet published[4]. The software is intended for users interested in estimating the size of populations, particularly those that are difficult to reach or where information is available from only one source and dual/multiple system estimation cannot be used.

In addition, this study used two previously unpublished datasets on drunk or intoxicated drivers and drivers with revoked licence. Knowing the size of these populations is crucial for assessing the safety of Polish roads and for informing the police about the extent of this phenomenon. Based on the SSCR models, we estimate that the first population is about 510k and the second population is about 22k. These estimates are the first estimates of these populations in Poland based on sound and justified statistical models.

References

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.

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. *The Annals of Applied Statistics*, *3*(2), 595–610. https://doi.org/10.1214/08-AOAS214

Böhning, D., & 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. *The Annals of Applied Statistics*, *13*(2), 1198–1211. https://doi.org/10.1214/18-AOAS1232

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

Chao, A. (1989). Estimating population size for sparse data in capture - recapture experiments. *Biometrics*, 427–438.

Chlebicki, P., & Beręsewicz, M. (2023). *Singlercapture: A package for single-source capture-recapture models* [R package version 0.1.4]. https://github.com/ncn-foreigners/singleRcapture/tree/0.2-development

Cruyff, M. J. L. F., & van der Heijden, P. G. M. (2008). Point and interval estimation of the population size using a zero-truncated negative binomial regression model. *Biometrical Journal*, *50*(6), 1035–1050. https://doi.org/https://doi.org/10.1002/bimj.200810455

Godwin, R. T., & Böhning, D. (2017). Estimation of the population size by using the one-inflated positive poisson model. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *66*(2), 425–448. Retrieved April 28, 2023, from http://www.jstor.org/stable/44682582

Kleiber, C., & Zeileis, A. (2016). Visualizing count data regressions using rootograms. *The American Statistician*, *70*(3), 296–303.

[4]Papers on these methods are in preparation.

Norris, J. L., & Pollock, K. H. (1996). Including model uncertainty in estimating variances in multiple capture studies. *Environmental and Ecological Statistics*, *3*(3), 235–244.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

van der Heijden, P. G., Bustami, R., Cruyff, M. J., Engbersen, G., & van Houwelingen, H. C. (2003). Point and interval estimation of the population size using the truncated poisson regression model. *Statistical Modelling*, *3*(4), 305–322. https://doi.org/10.1191/1471082X03st057oa

van der Heijden, P. G., Cruyff, M., & Van Houwelingen, H. C. (2003). Estimating the size of a criminal population from police records using the truncated poisson regression model. *Statistica Neerlandica*, *57*(3), 289–304.

Yee, T. W. (2015). *Vector generalized linear and additive models: With an implementation in r* (1st). Springer Publishing Company, Incorporated.

Zeileis, A. (2004). Econometric computing with hc and hac covariance matrix estimators. *Journal of Statistical Software*, *11*(10), 1â€“17. https://doi.org/10.18637/jss.v011.i10

Zwane, E., & van der Heijden, P. (2003). Implementing the parametric bootstrap in capture–recapture models with continuous covariates. *Statistics & probability letters*, *65*(2), 121–125.

Appendix A
Fitting logs for models described in section 5

Drunk driving:

```
  iterationNumber halfStep Log-likelihood
1               1        0     -10716.204
2               2        0     -10053.103
3               3        0      -9787.109
4               4        0      -9716.946
5               5        0      -9706.158
6               6        0      -9705.903
7               7        0      -9705.903
8               8        0      -9705.903
9               8        1      -9705.903
```

Drivin without a licence:

```
  iterationNumber halfStep Log-likelihood
1               1        0      -1951.356
2               2        0      -1893.047
3               3        0      -1881.852
4               4        0      -1880.604
5               5        0      -1880.584
6               6        0      -1880.584
7               7        0      -1880.584
```