Supplemental Materials to : "Reality and risk: a refutation of S. Rendón's analysis of the Peruvian Truth and Reconciliation Commission's mortality study"

Daniel Manrique-Vallier and Patrick Ball*

January 25, 2019

1 Explaining the MIMDES counts

Between 2001 and 2006, the *Program de Apoyo al Repoblamiento* at the Peruvian *Ministry of Women and Social Development* (MIMDES) conducted its own survey of victims of the Peruvian conflict. The results of this "Census for Peace" were published by the Peruvian Government as a series of pdf documents which included lists of fatal victims (see, e.g., MIMDES, 2003, 2007). The 2003 MIMDES report documents 10,470 deaths, of which 66% were attributed to SLU. An update, in 2006, reported an additional 8,227 deaths and 2,390 disappearances, of which 71% were attributed to SLU.

As part of an as-yet-unfinished project, we re-digitized these data and matched them with the TRC's data. For our record linkage, we extracted the names of the victims from four pdf files. There were approximately 51,000 names in these files, which contained many duplicate records. After removing a small number of records that were not full names (e.g., "son of," "wife of") and linking the duplicate records, we found 20,468 unique, fully-identified records of victims of fatal violations in the MIMDES surveys. This is 619 records fewer than MIMDES reported.

Since we needed to link the MIMDES records to the TRC's, we had to obtain a version of the TRC's confidential data including the names of the victims. We did so through direct agreement with its current steward, the Peruvian Ombudsman Office (*Defensoría del Pueblo*). This database has some minimal differences with the anonymized data published by the TRC in 2003. This is due to an additional quality control step performed by the TRC which was not ready in time for the publication of the commission's final report.

2 Comparison between $\hat{N}_{\mathbf{direct}}$ and $\hat{N}_{\mathbf{TRC}}$

The general multinomial Capture-Recapture (CR) problem can be formalized as follows. Consider a multinomial population $\mathbf{n} \sim \text{Multinomial}(N, \mathbf{p})$ where $\mathbf{p} = (p_c)_{c \in \mathcal{C}}$ is a vector of discrete probabilities, and $\mathcal{C} = \{0,1\}^J$ is an index set that represents cells in a J-dimensional contingency table formed by the cross-classification of J binary categorical with values in $\{0,1\}$. Let

*Daniel Manrique-Vallier is Assistant Professor of Statistics, Indiana University, Bloomington, IN 47405 (e-mail dmanriqu@indiana.edu);

$\mathbf{0} = (0, ..., 0) \in \mathcal{C}$, and $\mathbf{n}^* = (n_c)_{c \in \mathcal{C} \setminus \{\mathbf{0}\}}$. The CR problem consists in estimating N using a single observation of $\mathbf{n}^*$. Usually, estimation of N involves the estimation of the nuisance parameter $\mathbf{p}$. In order to allow identifiability we usually model $\mathbf{p}$ through a parametric model $p_c = p_c(\theta)$ with $c \in \mathcal{C}$ and $\dim(\theta) < J$. CR methods differ in which model they impose on $\mathbf{p}$. Most methods (e.g. log-linear models) specify parametric families from which the analyst has to choose a model using some model-selection technique.

In the case of the estimation of SLU victims, most elements of the set of observations $\mathbf{n}^*_{SLU}$ were either zero or had small counts, except for the count corresponding to elements only in the TRC list, $n_{SLU(1,0,0)}$. This makes CR estimation of $\mathbf{p}$, and thus N, either impossible or unreliable. Data for EST did not have this problem. Knowing this, the TRC opted for an indirect approach: use data $\mathbf{n}^*_{EST+SLU} = (n_{SLU,c} + n_{EST,c})_{c \in \mathcal{C} \setminus \{\mathbf{0}\}}$ to obtain a CR estimate $\hat{N}$ of $N = N_{EST} + N_{SLU}$; obtain a CR estimate $\hat{N}_{EST}$ of N_{EST} using data $\mathbf{n}^*_{EST}$; and take $\hat{N}_{TRC} = \hat{N} - \hat{N}_{EST}$ as an estimator of N_{SLU}.

A simple application of Slutsky's theorem (see e.g. van der Vaart, 1998, p11) shows that $\hat{N}_{TRC} = \hat{N} - \hat{N}_{EST}$ is consistent, provided that estimators $\hat{N}$ and $\hat{N}_{EST}$ are also consistent (Sanathanan, 1972, shows that this is the case provided that the model for $\mathbf{p}$ is correct). Rendón argues that this approach is "unusual" and that somehow this implies that it is incorrect. He proposes to select the few geographic regions where $\mathbf{n}^*_{SLU}$ allows the identification of at least one log-linear model for $(N, \mathbf{p})$ *no matter how small the counts* $\mathbf{n}^*_{SLU}$ *are*, and apply CR directly. We call this estimator $\hat{N}_{\text{direct}} = \hat{N}_{SLU}$.

Both approaches are in principle sensible, yet both have important disadvantages. $\hat{N}_{TRC}$ ensures larger sample sizes thus reducing the risk of breakdown of asymptotic properties, bias due to small counts, and of unidentifiability. However it relies strongly on the ability of the CR procedure $\hat{N}$ to pick up the structure in $\mathbf{p}$—which due to the aggregation of perpetrators may be complex. On the other hand, $\hat{N}_{direct}$ simplifies the structure of $\mathbf{p}$ through stratification (Sekar and Deming, 1949), but severely reduces sample sizes.

In order to compare the performance of the the two proposed methods we calculate their mean squared error (MSE) as a measure of estimation risk. Let $\mathbf{n} = (\mathbf{n}_{EST}, \mathbf{n}_{SLU}) \sim F$ be the *full* data—including $n_{SLU,\mathbf{0}}$ and $n_{EST,\mathbf{0}}$—, and F a measure on the discrete set $\Omega = \mathbb{N}^{2^{J+1}}$. Then the MSE of the direct estimator is:

$$\begin{aligned} MSE(\hat{N}_{\text{direct}}, N_{SLU}(F)) &= \mathbb{E}_F[(\hat{N}_{SLU} - N_{SLU}(F))^2] \\ &= \int_\Omega \left(\hat{N}_{SLU}(\mathbf{n}) - N_{SLU}(F)\right)^2 F(d\mathbf{n}) \\ &= \int_\Omega \left(\hat{N}_{SLU}(\mathbf{n}^*_{SLU}) - N_{SLU}(F)\right)^2 F(d\mathbf{n}) \end{aligned}$$

where $N_{SLU}(F) = \mathbb{E}_F\left[\sum_{c \in \mathcal{C}} n_{SLU,c}\right]$. The MSE of $\hat{N}_{TRC}$ is slightly more complex:

$$\begin{aligned} MSE(\hat{N}_{\text{TRC}}, N_{SLU}(F)) &= \mathbb{E}_F[(\hat{N} - \hat{N}_{EST} - N_{SLU}(F))^2] \\ &= \int_\Omega \left(\hat{N}(\mathbf{n}) - \hat{N}_{EST}(\mathbf{n}) - N_{SLU}(F)\right)^2 F(d\mathbf{n}) \\ &= \int_\Omega \left(\hat{N}(\mathbf{n}^*_{EST+SLU}) - \hat{N}_{EST}(\mathbf{n}^*_{EST}) - N_{SLU}(F)\right)^2 F(d\mathbf{n}) \end{aligned}$$

In order to perform a meaningful comparison, we need to constrain the set of measures F to a subset relevant to our problem. We want to analyze, for each of the 9 strata where $\hat{N}_{direct}$ exists, how the estimators would have behaved had the actual $N_{SLU}(F)$ being equal to some plausible or interesting hypothetical value, while simultaneously the induced distribution of $\mathbf{n}^*$ is such that it could plausibly had generated the actual observed data. Let $F_{OBS}(\cdot)$ be the actual distribution of $\mathbf{n}^*$, induced by $F(\cdot)$. Therefore we are interested in the set of measures

$$\mathcal{F}(N_{SLU}, N_{EST}) = \Bigg\{ F : F\Bigg(\sum_{c\in\mathcal{C}} n_{SLU,c} = N_{SLU}, \sum_{c\in\mathcal{C}} n_{EST,c} = N_{EST} \Bigg) = 1, \\ F(\mathbf{n}^* \in A) = F_{OBS}(\mathbf{n}^* \in A) \Bigg\}$$

Since Rendón does not question the TRC's estimates of N_{EST}, we will fix N_{EST} at its estimated (by the TRC) value $\hat{N}_{EST}$. Regarding F, we will use an estimate obtained from the observed data plus the constraints. For this, we estimate $\hat{F}_{N_{SLU}} \in \mathcal{F}(N_{SLU}, \hat{N}_{EST})$ using a model on the complete table and the observed data completed with $n_{SLU,\mathbf{0}} = N_{SLU} - \sum_{c\in\mathcal{C}\setminus\mathbf{0}} n^{OBS}_{SLU,c}$ and $n_{EST,\mathbf{0}} = \hat{N}_{EST} - \sum_{c\in\mathcal{C}\setminus\mathbf{0}} n^{OBS}_{EST,c}$. Thus we have

$$MSE(\hat{N}_{\text{direct}}, N_{SLU}) = MSE(\hat{N}_{\text{direct}}, N_{SLU}(\hat{F}_{N_{SLU}})) \\ MSE(\hat{N}_{\text{TRC}}, N_{SLU}) = MSE(\hat{N}_{\text{TRC}}, N_{SLU}(\hat{F}_{N_{SLU}}))$$

A minimal-assumption alternative for $\hat{F}_{N_{SLU}}$ is to use the empirical distribution (Efron and Tibshirani, 1993). We have discarded this idea because, since many entries in the contingency tables are zero, the empirical distribution would necessarily assign probability zero to those cells. This would cause the direct approach, which relies in sparsely populated tables, to fail often. Instead, we have opted for a no-second-order-interaction log-linear model. This alternative has three advantages: it assigns positive probability to all cells in the contingency table; it is a very flexible model, but is not saturated; and it satisfies the main assumption on which log-linear CR rests. We consider the last point an advantage because it allows us to compare the methods under the most favorable conditions for both of them.

3 Risk of $\hat{N}_{direct}$ and $\hat{N}_{TRC}$ in Rendón's chosen strata

Figures 1 to 9 show the results of the risk comparison between $\hat{N}_{\text{TRC}}$ and $\hat{N}_{\text{direct}}$ described in the previous section for the strata selected by Rendón. We have calculated the expectations using Monte Carlo integration. Each graph shows the computed risk for different levels of the true number of SLU victims ("True N", x-axis), from the observed count as of 2003 ("nobs(2003)") to either the observed count as of 2018 ("nobs(2018)") plus 80% or the largest estimate, whichever is larger. We have shaded the regions for which N_{SLU} is less than the number of total victims *observed* (as of 2018) which we have labeled the "impossiblity region". As discussed, an estimate of the population size cannot be smaller than the size of an observed sample. On the right, in the unshaded region, there are a range of possible true values for N_{SLU} at levels greater than the

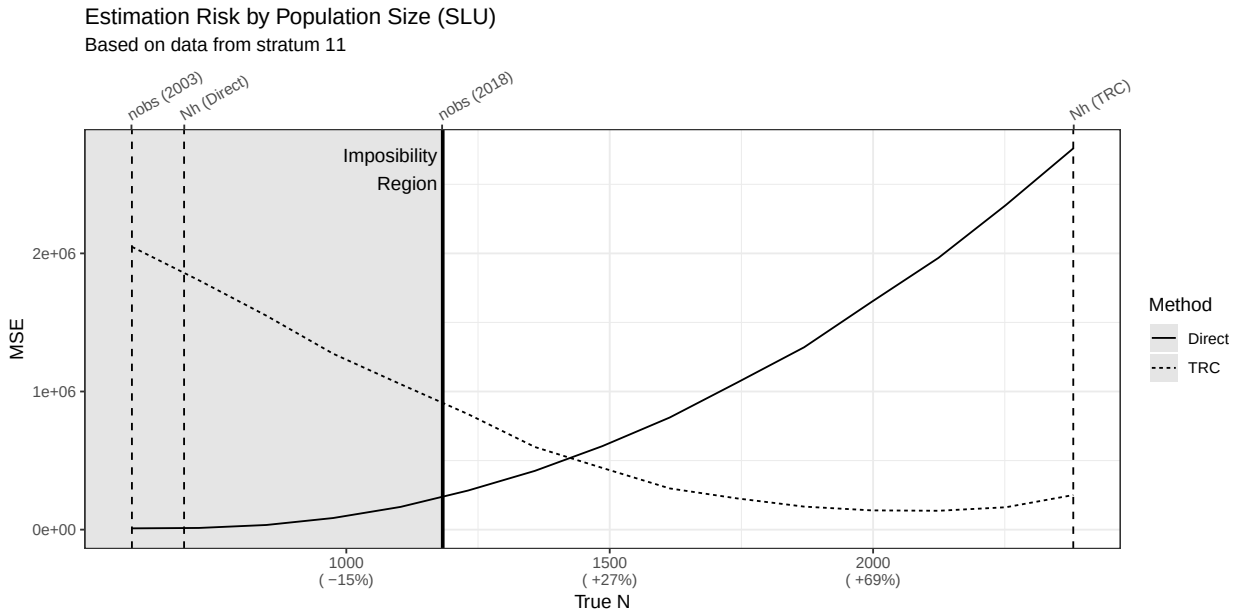

Figure 1: Estimation risk (MSE) vs. true population size for TRC and direct methods for stratum 11. Smaller is better. Shaded region correspond to values of N smaller than the known minimum as of 2018. Percentages in x-axis are with respect to the known minimum.

count of observed victims. The direct estimate is shown as "Nh(direct)" and the TRC's estimate is shown as "Nh(TRC)".

In most of the strata, the TRC's approach has high risk in the impossibility region, and the risk declines as N_{SLU} increases, while the Direct approach shows the inverse pattern of low risk in the impossible values and higher risk at larger possible values of the true number of SLU victims. Some observations:

1. In 5 out of the 9 strata Nh(Direct) is in the imposibility region.Nh(TRC) is never there.
2. In 3 out of the 9 strata Nh(Direct) risk is higher than the risk of Nh(TRC) for all possible values.
3. In 6 out of the 9 strata Nh(Direct) risk is higher that the risk of Nh(TRC) for all values larger than nobs(2018) plus 30%.
4. Two noteworthy exceptions, where the risk of Nh(direct) is smaller than that of of Nh(TRC) until values past nobs(2018) plus 30% are stratum 36, where estimates are similar and stratum 51, where Nh(Direct) falls well into the imposibility region.

References

Efron, B. and Tibshirani, R. (1993), *An Introduction to the Bootstrap*, London: Chapman & Hall/CRC.

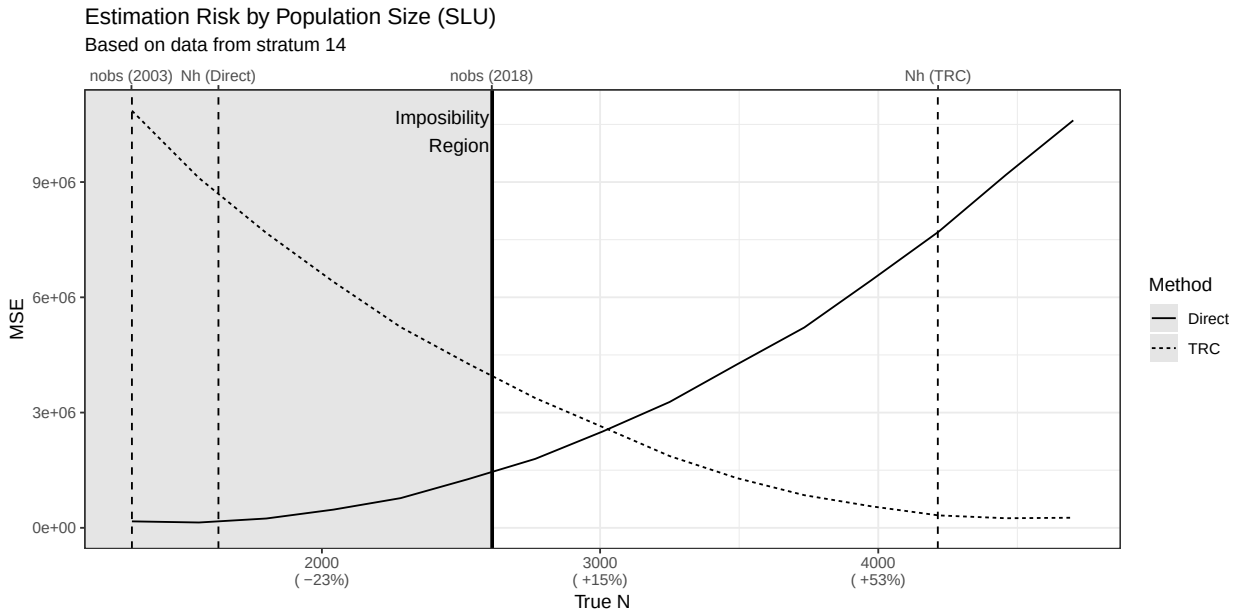

Figure 2: Estimation risk (MSE) vs. true population size for TRC and direct methods for stratum 14. Smaller is better. Shaded region correspond to values of N smaller than the known minimum as of 2018. Percentages in x-axis are with respect to the known minimum.

MIMDES (2003), "Censo por la Paz 2001–2003: Relación preliminar de personas muertas por el conflicto armado interno de acuerdo con el Censo por la Paz 1980–2000," Tech. rep., Ministerio de la Mujer y Desarrollo Social (MIMDES).

— (2007), "Censo por la Paz 2006: Centros poblados rurales afectados por la violencia ocurrida en el periodo 1980–2000," Tech. rep., Ministerio de la Mujer y Desarrollo Social (MIMDES).

Sanathanan, L. (1972), "Estimating the size of a multinomial population," *Annals of Mathematical Statistics*, 43, 142–152.

Sekar, C. C. and Deming, W. E. (1949), "On a Method of Estimating Birth and Death Rates and the Extent of Registration," *Journal of the American Statistical Association*, 44, 101–115.

van der Vaart, A. (1998), *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

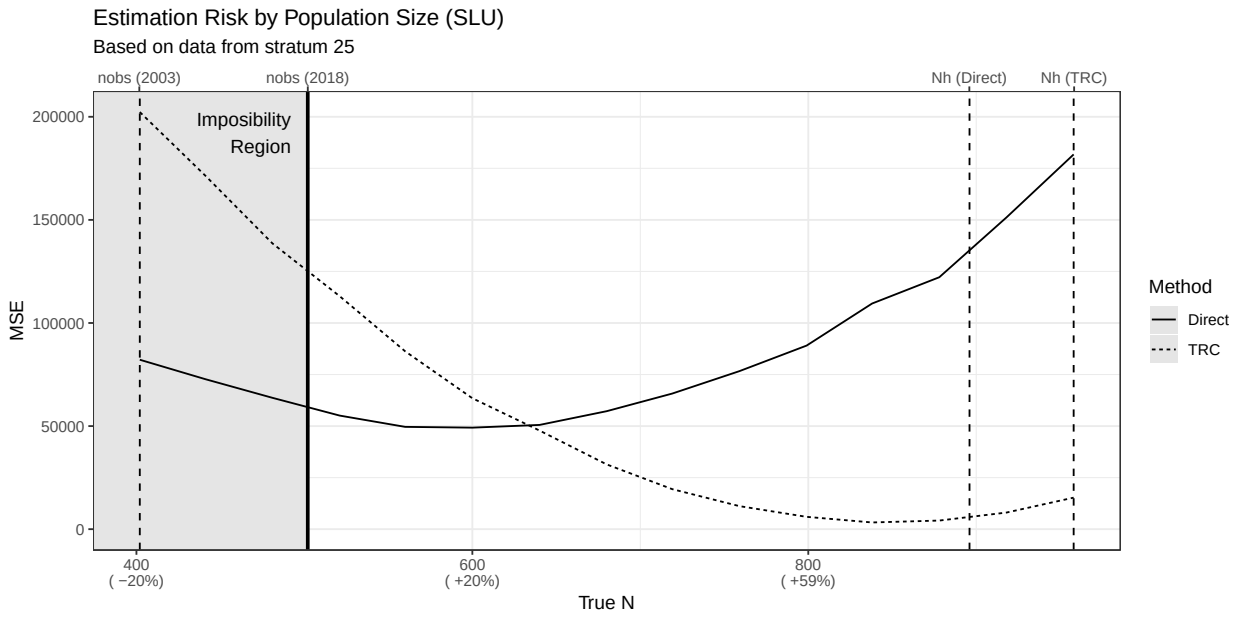

Figure 3: Estimation risk (MSE) vs. true population size for TRC and direct methods for stratum 25. Smaller is better. Shaded region correspond to values of N smaller than the known minimum as of 2018. Percentages in x-axis are with respect to the known minimum.

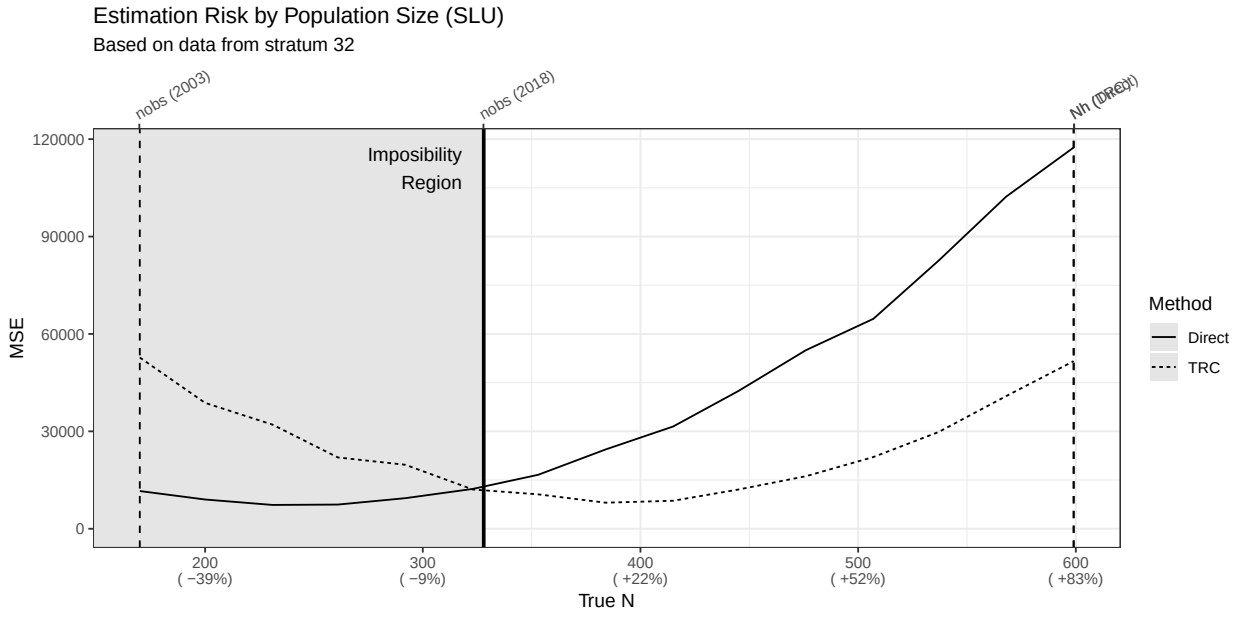

Figure 4: Estimation risk (MSE) vs. true population size for TRC and direct methods for stratum 32. Smaller is better. Shaded region correspond to values of N smaller than the known minimum as of 2018. Percentages in x-axis are with respect to the known minimum.

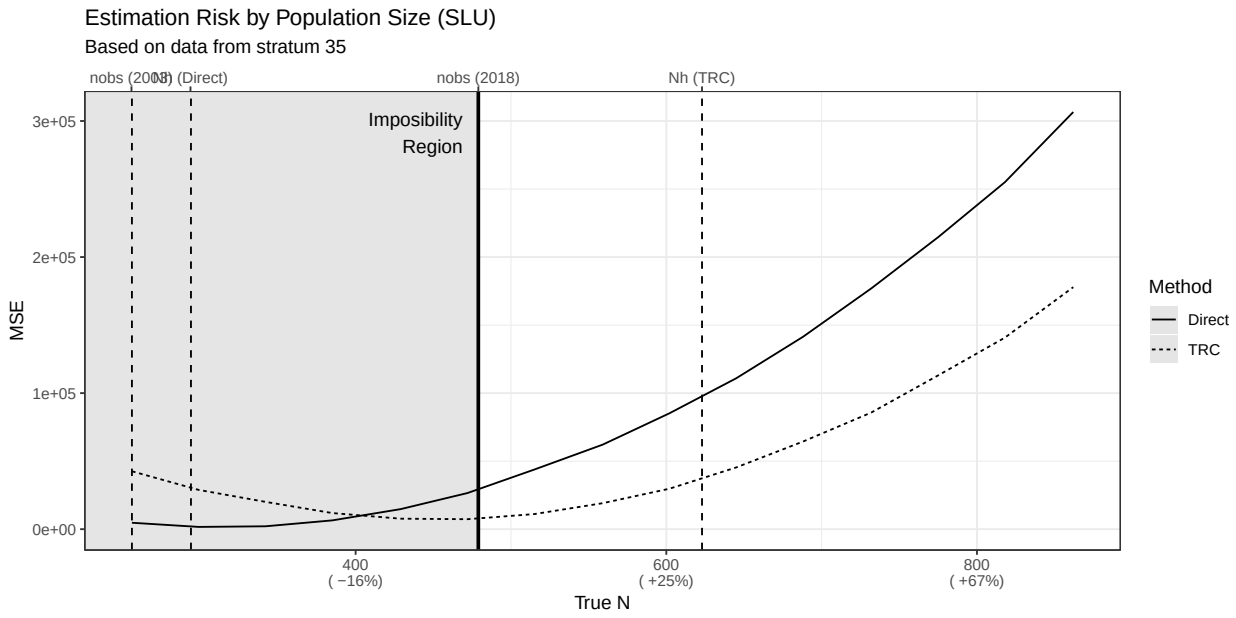

Figure 5: Estimation risk (MSE) vs. true population size for TRC and direct methods for stratum 35. Smaller is better. Shaded region correspond to values of N smaller than the known minimum as of 2018. Percentages in x-axis are with respect to the known minimum.

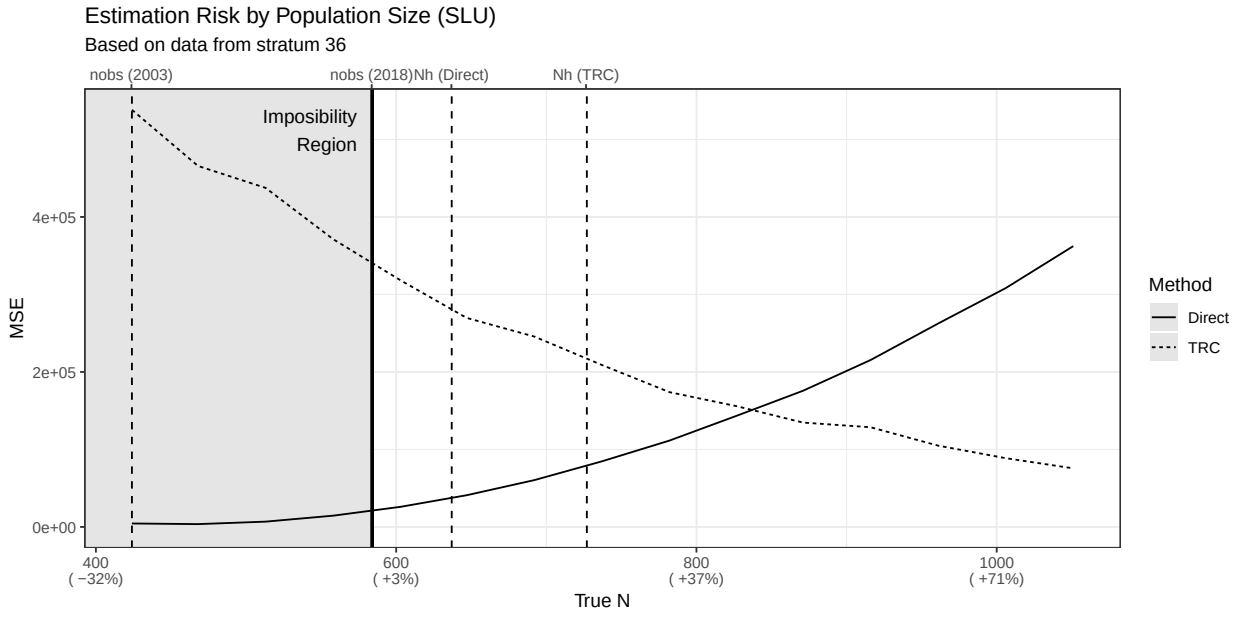

Figure 6: Estimation risk (MSE) vs. true population size for TRC and direct methods for stratum 36. Smaller is better. Shaded region correspond to values of N smaller than the known minimum as of 2018. Percentages in x-axis are with respect to the known minimum.

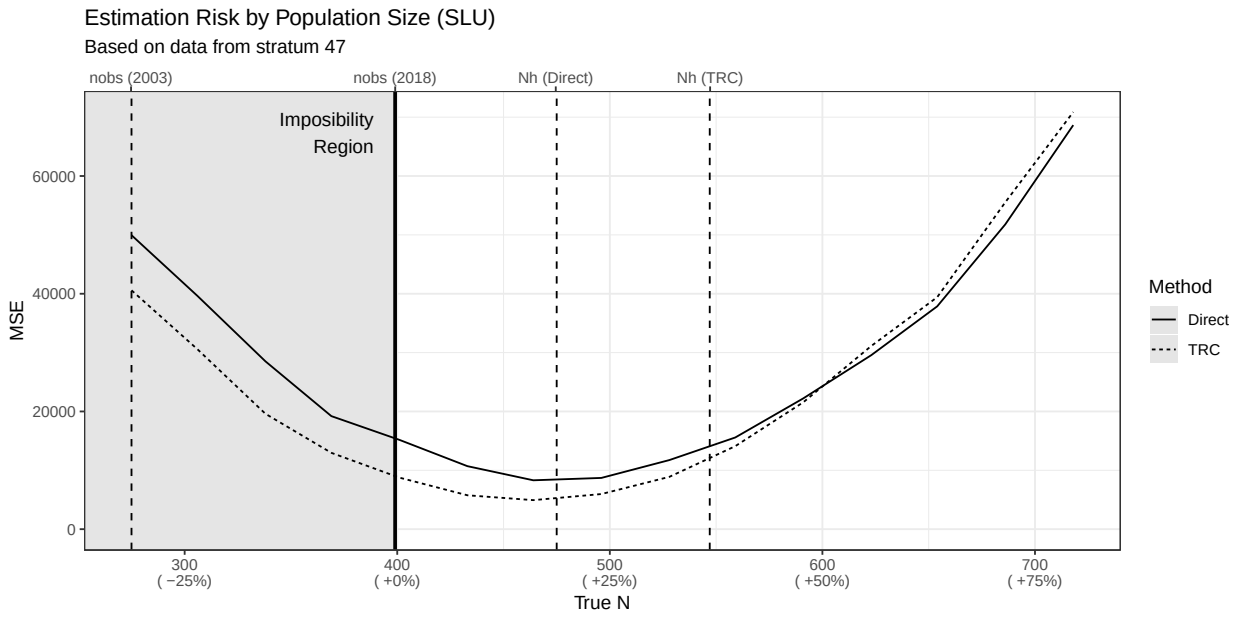

Figure 7: Estimation risk (MSE) vs. true population size for TRC and direct methods for stratum 47. Smaller is better. Shaded region correspond to values of N smaller than the known minimum as of 2018. Percentages in x-axis are with respect to the known minimum.

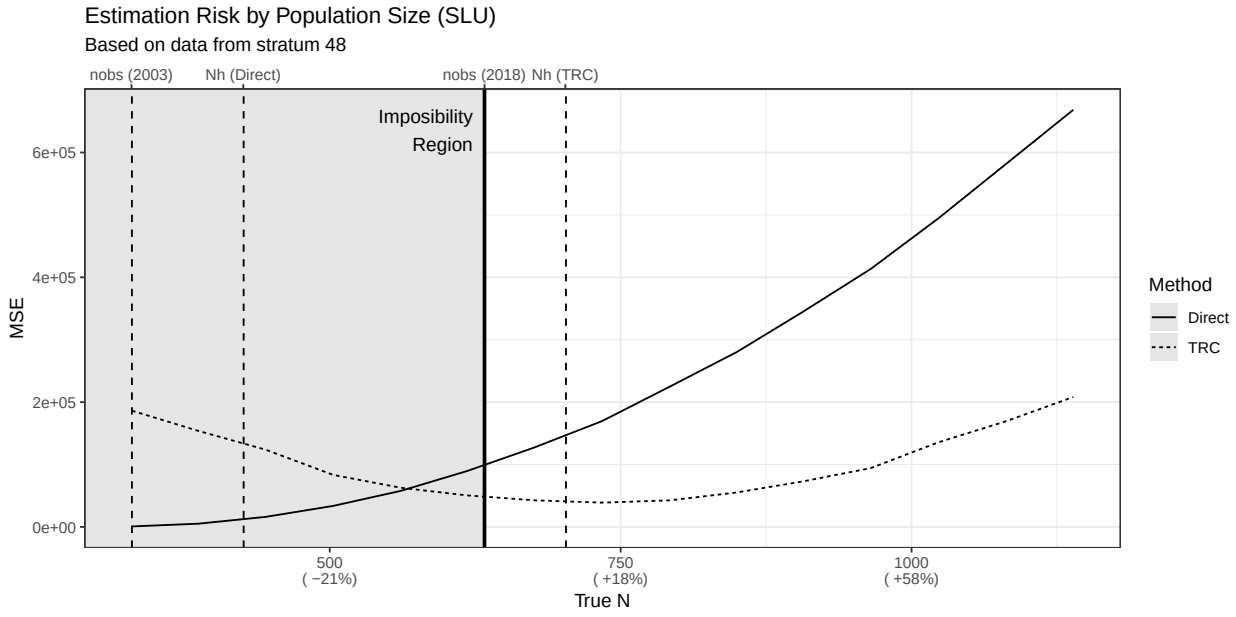

Figure 8: Estimation risk (MSE) vs. true population size for TRC and direct methods for stratum 48. Smaller is better. Shaded region correspond to values of N smaller than the known minimum as of 2018. Percentages in x-axis are with respect to the known minimum.

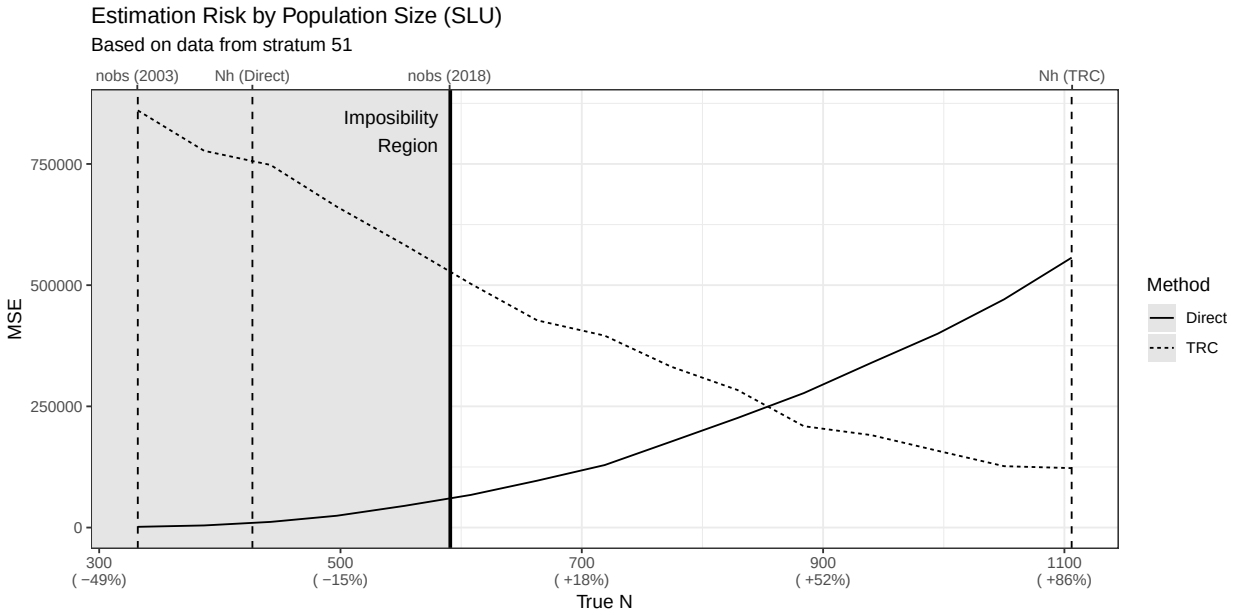

Figure 9: Estimation risk (MSE) vs. true population size for TRC and direct methods for stratum 51. Smaller is better. Shaded region correspond to values of N smaller than the known minimum as of 2018. Percentages in x-axis are with respect to the known minimum.