

# Reality and risk: a refutation of S. Rendón’s analysis of the Peruvian Truth and Reconciliation Commission’s mortality study

DANIEL MANRIQUE-VALLIER\*

Department of Statistics, Indiana University  
dmanriqu@indiana.edu

PATRICK BALL†

Human Rights Data Analysis Group  
pball@hrdag.org

January 26, 2019

## Abstract

We refute Rendón (2019)’s assessment of the 2003 Peruvian Truth and Reconciliation Commission (TRC) conflict mortality study. We first show that his most important result, an alternative estimate of the mortality due to the Maoist guerrillas of *Shining Path* is lower than existing observed data and is therefore impossible. We then analyze his statistical approach and find that it is affected by a subtle form of selection bias. We contrast his approach to the TRC’s using tools from statistical decision theory, and determine that his method is inadequate for this problem—and that the TRC’s approach is, at minimum, better. Without advocating for the TRC’s original results, we conclude that Rendón’s approach and methods are inferior to the TRC’s original work.

**Keywords:** Casualty count estimation, Capture-Recapture, Estimation Risk, Peruvian Truth Commission

## I. INTRODUCTION

Peru experienced a terrible internal armed conflict during the period 1980–2000 between the Maoist guerrillas of *Sendero Luminoso* (SLU) and agents of the Peruvian state (EST). In 2003, a team at Peru’s Truth and Reconciliation Commission (TRC) estimated conflict mortality due to violence using Capture-Recapture (CR) methods by combining the TRC’s information with five other databases. Estimates were stratified by location and perpetrator. Killings committed by SLU were infrequently documented by the non-TRC databases. Therefore, instead of obtaining direct CR estimates for SLU, the TRC first estimated a total including both EST and SLU, then estimated EST alone, and estimated SLU by subtraction. Rendón (2019) disqualifies this procedure as “unusual,” and

proposes instead estimating the few estimable strata and extrapolate to the rest.

We agree that there are aspects of the TRC’s approach that should be improved—we have been working on this for several years. However, our response here focuses specifically on Rendón’s proposal: we show that his methods are substantially weaker than the TRC’s, and that his results are unsound.

There are three bases for our rejection of Rendón’s methods and findings: first, his estimates are inconsistent with observed data. By combining the data used by the TRC with data published by the Peruvian government between 2004 and 2006 (MIMDES 2004, 2006), we see that Rendón’s estimates for SLU are, in most strata and in the aggregate, *lower* than the number of *observed* SLU victims—without considering victims who continue to be undocumented. This fact alone is enough to dismiss Rendón’s conclusions. The TRC’s estimates do not suffer from this problem.

\*Daniel Manrique-Vallier is Assistant Professor of Statistics, Indiana University, Bloomington, IN, USA.

†Patrick Ball is Director of Research at the Human Rights Data Analysis Group, San Francisco, CA, USA.

Second, Rendón’s methods are flawed. His approach consists of estimating only those strata where overlaps among datasets are frequent enough to allow direct log-linear CR estimation of SLU. This sounds innocuous until we note that this selection criterion is directly related to the potential outcomes. Indeed, CR methods rely on overlap information—and CR methods tend to produce smaller estimates when overlaps are relatively larger. Thus, by selecting these exceptional (9 of 59) strata, Rendón is affecting the outcomes, likely underestimating.

Third, Rendón’s claim that his approach is an improvement over the TRC’s is unsupported. He reasons as follows: the TRC did not employ the most obvious, direct approach, and merely for that, their estimates are dubious; an application of the obvious approach produces different results; therefore the new results discredit both the TRC’s approach and conclusions. This is fallacious argumentation. Furthermore, showing mere differences is not the appropriate way to compare statistical methods.

In this paper, we refute Rendón’s approach and conclusions. We first show that his estimates fail basic external validity tests. We then investigate why his methods fail. We compare his approach to the TRC’s using appropriate tools from statistical decision theory, and show that his methods are unsuitable for this application.

## II. EXTERNAL VALIDITY: ESTIMATES SHOULD NOT CONTRADICT REALITY

The TRC’s original 2003 data is not the only list of fully-identified victims of the Peruvian conflict. Between 2001 and 2006, the Peruvian *Ministry of Women and Social Development* (MIMDES) conducted a similar, but independent, survey of victims. 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). As part of an as-yet-unfinished project, we have re-digitized these

In Data TRC 2003	In Data MIMDES	Observed (Stratum 11)	Observed (Global)
No	Yes	590	8,351
Yes	No	332	5,245
Yes	Yes	261	4,091
Observed (2018)		1,183	17,687
$\hat{N}_{\text{direct}}$		692	15,089

**Table 1:** Known cases of deaths attributed to SLU by the time of TRC report (2003), and after adding MIMDES data. We report global totals and an illustrative stratum (11). In both cases Rendón’s estimates of the total,  $\hat{N}_{\text{direct}}$ , are smaller than the number known cases.

data.

In order to perform the record linkage, we had to obtain the confidential TRC data through direct agreement with its current steward, the Peruvian Ombudsman office. However, we note that the MIMDES lists and project have been publicly known at least since 2006 (and used in academic investigations since at least Fermi Blanco, 2012). Reflection on the scale and distribution by perpetrator in the MIMDES data—both similar to the TRC—should have raised questions about whether Rendón’s estimates are plausible. Furthermore, as we note below, there is sufficient information even in the TRC’s own publicly-reported data to doubt Rendón’s approach.

The MIMDES dataset contains a total of 20,468 unique records, from which 13,011 were not already present in the TRC’s dataset. Importantly, it adds a total of 8,351 new records attributed to SLU. This raises the number of *known* victims of SLU to 17,687. Table 1 summarizes this information together with Rendón’s estimates,  $\hat{N}_{\text{direct}}$ . We also include details of one of Rendón’s selected strata (stratum 11) as an illustration.

Inspection of Table 1 makes clear that Rendón’s estimates are illogical: in both the aggregate and the example stratum, his projections are smaller than the *known* absolute lower bound.

Analysis of the TRC’s data alone would have sufficed to detect this problem. Stratum 25 (Chungui district in Ayacucho) is the only re-

gion for which there is an extra data source (CEDAP) with plenty of SLU cases. We can test Rendón’s method by that excluding that database and checking if the result is consistent with the observed lower bound. The resulting estimate is  $\hat{N}_{\text{direct}} = 206$  ( $\text{CI}_{95\%} = [204, 208]$ ). This is smaller than the observed (as of 2003) total of 422.

### III. TECHNICAL VALIDITY: TO SELECT IS TO SAMPLE

With the exception of stratum 25, most of the deaths attributed to SLU in the 2003 dataset originated from the TRC’s own data collection. Fifty out of 59 geographic strata lacked sufficient overlaps between sources to allow direct SLU log-linear CR estimates. The remaining 9 strata in principle allow the calculation of at least one estimate. Because the strata that would allow a direct estimate were so few, the TRC chose not to attempt direct calculations, and instead used an indirect approach. Rendón argues that estimating these nine strata and extrapolating from them to the whole country is a better procedure.

Rendón’s approach sounds sensible and intuitive. However, this approach involves *selecting strata*: a stratum will be selected whenever it is amenable to estimation, that is, when it contains enough overlaps. This is a random event which depends on the particular sample that resulted from the (random) data collection process.

We can identify two scenarios. If the unobserved population were actually small, as estimated by Rendón, then samples would tend to overlap, thus making the existence of estimable strata a relatively common event. In the converse case, if the dark figure were larger, overlaps and therefore estimable strata would be rarer. In any case, the probability of selecting a stratum is related inversely to its underlying size. This is what Rubin (1987) calls a *non-ignorable* observation mechanism. Rendón’s procedure ignores it.

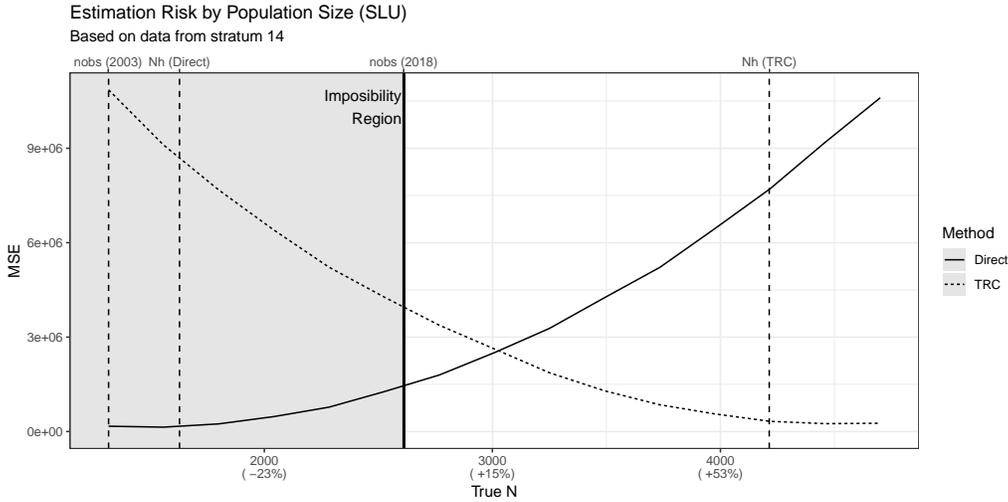
There are two elements that make us believe that we are in the second of the two outlined

scenarios, i.e., the larger dark figure. First, we have already shown that Rendón has underestimated so severely that his estimates contradict observed data. Second, the fact that only 9 out of 59 strata are estimable suggests that being estimable is indeed a low-probability event. If this were the case, Rendón’s method would result in underestimation.

### IV. COMPARING STATISTICAL METHODS: ESTIMATION RISK

Intuitively, an estimator is better than another if it produces estimates that are closer to the truth. Unfortunately, these direct comparisons are not usually possible or even meaningful: true values are usually unknown, and different samples lead to different estimates. Statistical decision theory avoids these pitfalls through the study of *estimation risk*: the average deviation of estimates, over all possible samples, with respect to the true value of the parameter (see e.g. Schervish, 1996; Shao, 2003). Since true values are usually unknown, statisticians calculate risks for relevant sets of values. A method will be considered better for estimating one given true value whenever its risk at that value is smaller than the risk of the competing method at the same value. In many cases, one of two competing estimators will be better for estimating only a subset of the parameter space. Thus determining which of the two is “the best” for an application needs to take into consideration which values of the parameter are relevant to that application.

We have calculated the risk of both Rendón’s “direct” approach and the TRC’s method. Our selected risk measure is the mean squared error,  $\text{MSE}(\hat{N}, N) = \mathbb{E}_N[(\hat{N} - N)^2]$ , a common choice for these comparisons (Shao, 2003). In order for the comparison to be meaningful, we have made sure that the distribution of data—which determines the distributions of both  $\hat{N}_{\text{direct}}$  and  $\hat{N}_{\text{TRC}}$ —is such that 1) the induced distribution on the *observable* cells is similar to that of 2003 data, and 2) the probability of the unobservable cell is related to the others through the condition assumed by both meth-



**Figure 1:** Estimation risk (MSE) vs. true population size for TRC and direct methods. 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.

ods (no-second-order interaction; see Bishop et al., 1975, Ch.6). We have constructed such distributions by, starting from an assumed  $N$ , estimating multinomial probabilities on the complete  $2 \times 2 \times 2$  contingency table using real stratum data completed to  $N$ , and a no-second-order-interaction model; see Supplement for details. This approach is similar to the non-parametric bootstrap, but gives the direct method a better chance of success by avoiding the truncation of the support of the distribution due to the presence of random zeros (Bishop et al., 1975). For both estimators, we included the determination of their existence and the model selection as part of their definition.

Figure 1 shows the risk of both methods for estimating SLU casualties, based on data from stratum 14. We have evaluated the risks for  $N$  ranging from the observed count as of 2003,  $\text{nobs}(2003)$ , to the observed count as of 2018 plus 80%,  $\text{nobs}(2018) + \%80$ . Even though any  $N$  smaller than  $\text{nobs}(2018)$  is impossible, we have found that inspecting these values can help understanding why and how Rendón’s method went astray. Aside from the fact that Rendón’s method has resulted in an impossible estimate ( $N_h$  direct), a salient feature

in Figure 1 is that both methods can behave both adequately and poorly depending on the true  $N$ . However Rendón’s method only behaves better than TRC’s for small values of  $N$ , most of which are either implausible or, as already shown, impossible. This behavior is consistent with our analysis in the previous section. The TRC’s method behaves poorly for low (but implausible) values of  $N$ , but it overtakes Rendón’s at approximately  $\text{nobs}(2018) + 15\%$  and improves from there. This shows that, although far from perfect, the TRC’s method is likely to produce better estimates for values of  $N$  which are consistent with the full set of observed data. Analysis of most of other 8 selected strata (in supplement) shows a similar picture. An interesting exception is stratum 47, where both methods have similar risk profiles—and produce similar estimates.

## V. DISCUSSION

The TRC’s study has several limitations. In particular, estimates stratified by perpetrator require accounting for records with missing perpetrator labels, as Rendón does. In addition, we note that both the TRC’s estimates,

and Rendón's, rely on log-linear models which have important limitations (see e.g. Manrique-Vallier, 2016).

This said, existence of limitations in the TRC's work does not by itself justify any alternative. Alternatives should at minimum stand on their own: they have to be statistically sound, and should produce plausible results. In addition, if they are to contribute anything to the discussion, they should also have some advantage over the original other than merely appearing more obvious. Rendón's work fails all three of these requirements.

We agree with Rendón's observation that the magnitude and distribution by perpetrator of killings during the Peruvian conflict are of great importance for Peruvian historical memory. We agree further that these questions merit considerable additional scientific attention. However, the flaws in this article make it unsuitable for this discussion.

## VI. ACKNOWLEDGMENTS

The authors thank Professor Andrew Womack for his thoughtful suggestions.

## REFERENCES

- Bishop, Y., Fienberg, S., and Holland, P. (1975), *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA: MIT Press, reprinted in 2007 by Springer-Verlag, New York.
- Fermi Blanco, E. (2012), "Las relaciones entre las Ciencias Sociales y la Política: Un ejercicio metodológico sobre los hallazgos de la comisión de la Verdad y la Reconciliación," Sociology professional licensing thesis, Pontificia Universidad Católica del Peru.
- Manrique-Vallier, D. (2016), "Bayesian Population Size Estimation Using Dirichlet Process Mixtures," *Biometrics*, 72, 1246–1254.
- 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).
- Rendón, S. (2019), "Capturing Correctly: A Reanalysis of the Indirect Capture-Recapture Methods in the Peruvian Truth and Reconciliation Commission," *Research and Politics*, TBD.
- Rubin, D. B. (1987), *Multiple Imputation for Non-response in Surveys*, New York: John Wiley & Sons.
- Schervish, M. (1996), *Theory of Statistics*, Springer Series in Statistics, Springer New York.
- Shao, J. (2003), *Mathematical Statistics*, Springer-Verlag New York Inc, 2nd ed.