A Discrete Choice Approach to Estimating Armed Conflicts' Casualties: Revisiting the Numbers of a 'Truth Commission'[*]

Silvio Rendon[†]

Stony Brook University

August 2012

Abstract.- I discuss the application of capture-recapture methods to estimating the total number of deaths in armed conflicts, and propose an alternative method based on a trivariate discrete choice model. Data come from the 'Truth and Reconciliation Commission' (TRC) of Peru, around 25000 deaths, classified by three sources of information, geographical strata, and perpetrator: the State and the Shining Path. In these data many killings have been only documented by one source, which makes a projection of killings unfeasible . TRC consultants Ball et al. (2003) tried to overcome this problem by means of a 'residual estimation,' consisting of merging data for different perpetrators. I show theoretically and empirically that this method over-estimates the number of deaths. Using a conditional trivariate Probit I estimate the total number of deaths in around 28000, 60% by the State, 40% by the Shining Path. This number is substantially lower and has a different composition than the around 69000 deaths, 30% by the State, 46% by the Shining Path, and 24% by 'other perpetrators,' calculated by Ball et al.

Keywords: Armed Conflict, Capture-Recapture, Count Data, Discrete Choice, Human Rights, Maximum-Likelihood Estimation, Poisson Regression.
JEL Classification: D74, C35, C4, O54, P16.

1 Introduction

In many countries that suffered ruthless armed conflicts 'Truth Commissions' (TCs) were established to, among other goals, estimate the number of deaths caused by those conflicts. In some cases these commissions did not limit themselves to document these deaths, but also tried to estimate the total number of deaths caused by these conflicts using capture-recapture methods (CRMs). This is the first research article to discuss the foundations of this approach, analyze in detail its particular implementation in a country, Peru, and propose alternative methods. While Peruvian TC officials estimated around 69000 deaths, the largest number of all TCs given observed deaths, I find a number of only around 28000 estimated deaths.

In 1980, on election day, the Shining Path, a Maoist armed group, irrupted in Chuschi, a very poor town in the highlands of Peru, with violent actions: a small brigade burned electoral ballots. It was the beginning of a war that involved the Peruvian police and army, paramilitary, self-defense groups, and insurgent groups, and lasted until around 2000 (CVR 2003). In 2001 a 'Truth and Reconciliation Commission'(TRC) was established by the government to find out the facts of this war and, in particular, to make an account of the casualties. The TRC created a dataset of around 25000 observations, fatal victims, classified by three sources of information, geographical strata, and perpetrator.[1]

A salient feature of these data is that the State is the only perpetrator whose deaths have been documented by all three sources of information: the TRC, the Ombudsman's Office, and human rights Non-governmental Organizations (NGOs). Deaths caused by the Shining Path and other perpetrators have many geographical strata that have only been documented by one source of information, the TRC. This feature of the data makes CRMs inapplicable, as they require that the different sources of information have a fair number observations in common. Ball, Asher, Sulmont and Manrique (2003) (BASM) , the authors of the TRC estimations, tried to overcome this limitation in an ingenious way: they performed an estimation by

---

[1]These data have been used by several economists to measure the long term effects of the armed conflict on education and health or to analyze discrimination issues in Peru (Castillo and Petrie 2007, León 2009, Galdo 2010, Laszlo and Grimard 2010).

CRMs for merged data of the State and the Shining Path, and from the resulting number they substracted the number of deaths that they estimated for the State, for which CRMs are applicable. This substraction is then claimed to be an accurate estimate of the deaths of the Shining Path. I show formally that this procedure, which I call a 'residual estimation,' is inadequate: it is based on the aggregation of heterogeneous groups and happens to bias results upward.

Another source of bias is that the popular saturation specification based on Poisson regressions tacitly implies independence of sources, another source of an upward bias. As an alternative, I only perform estimations when there are enough intersections of observations between the three available sources. If the approach of BASM was oriented to try to perform as many estimations as possible, my approach is to be more prudent, considering that we are dealing with a very sensitive subject, people killed whose cases have to be judicialized. I propose a discrete choice estimation based on a conditional trivariate normal or Gauss distribution that matches the observed counts of the three sources as an alternative to Poisson regressions, which are usually used on count data. Then, I compare the results obtained by BASM with Poisson and Gauss estimations on the TRC data.

My main finding is that with a trivariate normal distribution the estimated total number of deaths is around 28000, approximately 60% by the State and 40% by the Shining Path. This number is substantially lower and has a different composition than the around 69000 deaths, 30% by the State, 46% by the Shining Path, and 24% by 'other perpetrators,' calculated by BASM. In this article I also show that 'other perpetrators' are actually either State or Shining Path agents, so that the estimated number for these two perpetrators should already contain them. This duplication is another source of over-estimation of the total number of casualties in the Peruvian armed conflict by BASM.

The present article calls attention on the fact that Truth Commissions' estimations can be biased as well, which defeats its declared purpose of correcting for biases resulting from taking figures emerging from observed data at face value (See Ball 2012). With the authority of a 'truth' commission, biased figures become official and solidify as a 'truth' in sensitive post-war

social environments without much room for a fair scientific discussion of their foundations. Quantitative researchers in a sensitive subject such as human rights violations want to be extremely cautious in their statements. Hopefully this lesson is useful for the new TCs that will be created in the future around the world.

The rest of this article is organized as follows. The next section explains the quantitative work of Truth Commissions, their estimation methods, and its particular implementation in Peru. Section 3 describes the data used in this article, their selection and basic descriptive statistics. Section 4 discusses estimation methods. Section 5 presents the main results of applying several estimation methods to the data, and Section 6 concludes.

2 Background

2.1 The numbers of 'Truth Commissions'

In several countries that experienced bloody armed conflicts, TCs were established to document human rights abuses and violations of humanitarian laws, give historical truth to victims, seek to punish human rights violators, and create an environment of reconciliation (See Hayner 1994, 2001 for a detailed description and analysis of TCs). Most TCs in the world were created by post-war governments and had a fact finding mission, accomplished by numerous interviews with the several participants in the conflicts: authorities, victims, perpetrators, war lords, police, etc. A key part of that mission was to quantify the number of deaths occurred during these armed conflicts.

Most TCs just reported documented numbers of deaths, a 'body count,' as it is usually done in armed conflicts. However, starting in 1999 in Guatemala, some TCs performed estimations of the total number of deaths, notably in Kosovo, Peru and East Timor. In recent years, this type of estimation was applied in Colombia for two investigations, in the Casanare region and for killings of trade-unionists (Ball 2000, Ball et al. 2002, Ball et al. 2003, Silva and Ball 2006, Guzmán et al. 2007, Guzmán et al. 2012). For this quantitative work, all of these

TCs hired a consultant, Patrick Ball of the Human Rights Data Analysis Group (HRDAG),[2] who introduced CRMs or a Multiple Systems Estimation (MSE) in the computation of armed conflicts' casualties.

[Table 1 here]

Table 1 shows TCs and research on human rights advised or inspired by Patrick Ball, in which estimations using CRMs were performed. The Peruvian TC performs the largest projections from observed cases: a ratio of estimated over observed deaths of 3.16. This projection rate is 2.76 for Guatemala, 2.35 for Kosovo, and 1.87 for East Timor. Two estimations for Colombia exhibit ratios of 1.68, for the Casanare region, and of 1.27 for trade unionists.[3] As CRMs are becoming increasingly used in human rights quantitative analysis, is utmost necessary to assess the accuracy of these methods and their applications to this subject .

2.2   Accuracy of Capture-Recapture Methods

John Graunt was the first to use a CRM to estimate the London human population in 1662 (Krebs 1999). Since then, CRMs have been used mostly by wildlife biologists and ecologists to estimate the size of rare and elusive populations, which are difficult to find and count or are highly mobile and cannot be counted at one time. The technique has also been applied to human populations that are difficult to count, such as the homeless, children on medical support, rates of chronic diseases, dog bites, injuries, among other studies.

Wildlife is captured, marked and released; then, in another capture some animals are found to be already marked, that is, they were recaptured. Humans are also 'captured' in that they report an event to a survey, and 'recaptured' in that they may report the event again to another survey. The capture-recapture procedure uses the collected information to make an inference about the size of the whole population. A minimum requirement for the

---

[2]The HRDAG, originally based on the American Association for the Advancement of Science (AAAS), is part of Benetech, a non-profit organization that makes consulting in human rights issues.

[3]The reader will notice that the expansion by TCs goes down over time: younger TCs have lower projection rates, being the Peruvian TC the only outlier in this trend.

CRM to be unbiased, regardless of whether it is about wildlife or humans, is that the number of recaptures is not too small; otherwise the CRM will over-estimate the size of the population (Krebs 1999). A related crucial assumption for the unbiasedness of CRM population estimates is that subjects are homogenous and can be captured without selectivity. In Pollock's (1981) words:

> *If animals are "trap shy" there will be too few recaptures resulting in overestimation of population size whereas underestimation will result from "trap happy" animals.*

Behavioral differences between captured and missed individuals, also called heterogenous catchability (or capturability) or 'correlation bias,' is a serious problem for CRMs, and consequently there is an extensive and growing literature on techniques to solve this issue.[4] The problem is not just the dependency between sources of information, but that inferences based on captured or self-reported individuals are not valid for unobserved different individuals.

A clear source of bias when applying CRMs is the spurious aggregation of heterogeneous individuals. Suppose that we have two species coexisting in a geographical area. A wildlife researcher has traps that can capture raccoons and possums in, say, Long Island, NY. He captures and recaptures lots of raccoons, while he captures lots of possums, but without any recapture. What numbers would be obtained if a researcher merges data of possums and raccoons, estimates by CRM their joint population, computes by CRM the population of raccoons, and then calculates the population of possums by just substracting the latter from the former? In the next sections I will analyze this issue in detail.

Finally, an important difference between 'capturing' or 'recapturing' animals and interviewing human beings is that the latter decide on their participation and response to a survey, based on present and expected incentives, or disincentives, which are objective and subjective

---

[4]One possible approach is to relax the assumption that all individuals have the same response probability, and assume a catchability distribution over individuals. Some authors distinguish 'correlation bias' from heterogeneity, which is reserved for response differences over geographic areas (See, for instance, Darroch et al. 1993).

rewards and expectations, or fears.[5] These well-known issues are likely to also affect behavior of agents responding surveys after bloody armed conflicts. A respondent who already denounced to two human rights organization the death of a relative, without justice being made, may be less likely to denounce it to a third organization. On the contrary, a victim that is approached for an interview for the first time by a new more sympathetic organization may be very likely to denounce her case for the first time. This consideration will be very important for understanding the generating process of human rights data, particularly of the TRC.

2.3   The numbers of the Peruvian TC

Before the TRC was created there were around 9000 deaths, documented by human rights NGOs and the Ombudsman office. The State was responsible for more than 90% of those deaths. The TRC was a new source of information that changed this picture dramatically: it added around 15000 more documented cases, in which the Shining Path was the main perpetrator, 57% of all deaths. According to Ball (2012) at the end of the data collection process the picture of the armed conflict was that the State and the Shining Path were responsible for an almost equal number of deaths.[6] This picture changed further when the TRC under the advice of Patrick Ball estimated a total number of 69000 deaths, being the State responsible for 30% of them and the Shining Path for around 46% of them.[7]

---

[5]It is straightforward to set up a dynamic model of reporting of rational forward-looking agents, with a Bellman equation. However, data are not always available for identifying behavioral parameters of such possible models, so this endeavor is left for future research.

[6]As we will see, of around 24692 observations, the State is responsible for 47%, the Shining Path for 37%, and unknown perpetrators for the rest. If observations without a geographical identifier are excluded, 21950 observations remain in the sample and the State and the Shining Path become responsible for almost the same number of deaths: the State is responsible for 44%, the Shining Path for 41%, and unknown perpetrators for the rest.

[7]According to Ball (2012) the estimation by CRMs 'corrects population census undercounts,' while direct observations are misleading and biased: "*In Peru, working with the Truth and Reconciliation Commission, my colleagues and I observed that there were approximately the same number of deaths reported that were attributed to the Peruvian Army and the Shining Path, the Sendero Luminoso, the Maoist guerrillas. So if we had just concluded, on the basis of what was reported, that the two parties committed equal numbers of deaths, we would have misunderstood the deaths that were not reported. Once we corrected in Peru for the deaths that had not been reported, we realized that Sendero Luminoso was responsible for half again more deaths, and this dramatically changed the understanding of the conflict in Peru.*"

After the publication of the Final Report of the Peruvian TC big political debates followed. The TRC estimated number almost tripled previous death estimates, which was therefore 'greater than the number of human losses suffered by Peru in all of the foreign and civil wars that have occurred in its 182 years of independence.' CVR 2003). The TRC claim that the Shining Path was the main perpetrator[8] made Peru a unique case in the world in which a defeated insurgency kills more than the victorious State. It was a surprising statement for many.[9]

More 'technical' discussions happened in some media, but mostly in social networks, particularly by economists, sociologists, historians and statisticians. The main methodological criticism, by economist Hugo Ñopo, was that the assumptions of random sampling, independence, and homogeneity required for the CRM to be applicable were not fulfilled. Individuals' heterogeneity biased population estimates upward. BASM [10]replied to these objections stating that stratification by geographical strata and perpetrator, by building homogeneous groups, takes care of heterogeneity, while log-linear models correct the estimate for underlying non-zero correlations in the probability of capture among the sources.[11]

As discussed in the previous subsection, if some assumptions, such as individuals' homogeneity, are not fulfilled, CRMs are likely to produce biased population estimates, even with geographical stratification or interaction terms in log-linear models. However, because nobody offered alternative methods nor estimates, BASM's estimations remained uncontested until now as the only quantitative figure of the Peruvian armed conflict.

---

[8]*'This high degree of responsibility on the part of the PCP-SL* [the Shining Path] *is an exceptional case among subversive groups in Latin America, and one of the most notable unique features of the process that the TRC has had to analyze'* (CVR 2003).

[9]Although many TRC officials made several anouncements advancing that the expected number of deaths of 30000 was going to double and that the main perpetrator was the Shining Path (see Appendix A.1), the announced final numbers of the TRC were surprising to many people.

[10]Ball defended his method of estimation and his application thereof: *'What is at stake here is not whether the method is appropriate (...)The questions should be about whether we have applied the methods in a technically competent manner.'*

[11]Ñopo replied that the issue of heterogeneity is not solved by stratification, as the same problem will be present within a stratum, nor the issue of dependency is corrected by a log-linear regression. The interested reader can check these debates in Manrique and Sulmont (2007) and Rendon (2010).

3 Data

The TRC data are a non-random sample of victims, collected by 'mobile teams' sent to pre-assigned areas with target numbers from each area. Data are based on retrospective information provided by victims or their families, which were further reviewed, verified and qualified by TRC officials. That is, these are self-reported data with a further qualification by researchers (See Appendix A.1 for details).

The sample consists of 24692 observations of fatal victims, collected by three sources: the TRC, the Ombudsman's office, and Human Rights NGOs.[12] They registered information about perpetrator, sex and age of the victim, year of the event, which can be death or disappearance, and location of the event. For their estimation BASM partitioned the data by geographical strata, which implied excluding data without geographical identifiers. However, surprisingly, they did not use a similar criterion for data without perpetrator identifiers.

[Table 2 here]

Table 1 shows the effect of this sample selection on the final data. BASM classed perpetrators in three groups: the State, the Shining Path and 'Others.' The State includes State agents, that is, police and army units, 'self-defense' groups, which are State-sponsored citizens' armed groups that fought the insurgents, and paramilitary groups, i.e., State-sponsored or protected armed units that do not legally belong to any official armed branch. The Shining Path is a group on its own; while the other, less powerful, insurgent group, the 'Tupac Amaru' Revolutionary Movement (MRTA) is classed as part of a third group, called 'Other perpetrators.' This group contains mostly unidentified perpetrators: there are no data on the perpetrator, the perpetrator is undetermined, or it is known that the death happened in a clash with an armed insurgent group, but there is no information on who killed the victim. Only 263 casualties have the MRTA as a perpetrator and only around 223 observations are

---

[12]The data were collected by these three institutions at different times. Human Rights NGOs collected data from the 80s onwards, while the Ombudsman's Office is a public institution created in 1996 (that took over earlier official data collection about human rights violations by prosecutors of the Public Ministry). The TRC, as explained, started its work in 2001.

reported to have 'other perpetrators.' That is, there are only 486 casualties of known perpetrators, who are not the State nor the Shining Path. Most observations of this third group are likely to be victims or the State or the Shining Path. In what follows, I will keep this group as a separate partition, following BASM's criterion, but I will call it 'unknown perpetrators.'

A selection by geographical stratum keeps 98.2% of the observations of the Shining Path but selects out many observations of the State and of unknown perpetrators: 82.7% and 85.2%, respectively, remain in the sample. This implies an important change in the composition of perpetrators. In the original sample, 46.8% of the victims are caused by the State, 37.4% by the Shining Path, and 15.7% by unknown perpetrators; in the final sample, these percentages change to 43.6% by the State, 41.3% by the Shining Path, and 15.1% by unknown perpetrators. This selection mainly affects observations documented by sources other than the TRC: 75.5% of the excluded observations come from the Ombudsman's Office or Human Rights NGOs.

Excluding observations without identifiers for geographical strata is sensible for the purpose of this research, as they can be compensated by the estimated population. A similar criterion applies for observations without an identifier for perpetrator: performing an estimation for this group would imply double-counting projected populations that are already contained in the estimations for the State and the Shining Path. If observations without a geographical identifier were grouped as an extra geographical stratum, double-counting would occur as well.[13]

[Table 3 here]

---

[13]In Ball et al. (2003) 'self-defense' and paramilitary groups are classed as 'Other perpetrators,' which conveyed the idea that some violence was not caused by the State or the Shining Path, but it was the initiative of particular citizens' groups. In the discussions that followed the presentation of the Final Report of the TRC in August 2003, this classification was strongly criticized, as self-defense and paramilitary groups were indeed sponsored by the State, as asserted by the TRC report itself. In October of 2003 TRC officials stated that by a typo in the published report 'self-defense' and paramilitary groups were wrongly classed as part of 'Others;' however, the estimated numbers were actually computed having these two groups correctly classed as part of the State (CVR 2003, Appendix 10). This erratum was only done in the TRC report, in Spanish, not in Ball et al. (2003). It is sensible to class those two perpetrators as part of the State, however, the authors of this reclassification did not seem to realize that the group named 'Other perpetrators' was practically left without any other perpetrators.

Table 2 shows a summary of the available information in the selected sample, by perpetrator and intersected sources. The TRC has exclusively 68.9% of all observations, but only observations for the State are fairly evenly contained in all three sources: for the State the TRC has only 39.0% exclusive observations and 16.3% of the observations are contained in all three sources. On the contrary, the TRC has 94.8% exclusive observations of the Shining Path and 84.0% exclusive observations of unknown perpetrators, leaving just room for small intersections of other sources, mainly between the TRC and the Ombudsman's Office: at most 2.7% of the observations of the Shining Path and 12.4% of the observations of unknown perpetrators.

[Table 4 here]

Table 3 shows the distribution of observations over geographical strata by source and perpetrator. Most source intersections by the State contain between 1 and 20 observations, except for the exclusive observations documented by the TRC, 26 strata of which contain between 21 and 100 observations. The configuration for the Shining Path, in contrast, is really concentrated: there are lots of exclusive observations for the TRC and no observations for other sources. Concretely, 48 of 57 geographical strata with exclusive observations for the TRC contain more than 20 observations, while most other intersections contain no observations. A similar picture occurs with unknown perpetrators: for most geographical strata the source intersections have no observations, whereas 33 of 58 strata contain more than 20 observations. In terms of populations this will imply that for 5319 out of 9075 observations of the Shining Path, 58.6%, it is not feasible to perform estimations based on intersecting sources. For unknown perpetrators,1860 out of 3309 observations, 56.2%, do not intersect enough to perform any estimation.

[Figures 1, 2, and 3 here]

Figures 1, 2, and 3 illustrate the incidence of the armed conflict over time by source and perpetrator. Again, deaths by the State are fairly reported in all three sources, see Figure 1,

whereas deaths by the Shining Path, in Figure 2, and unknown perpetrators, in Figure 3, are overwhelmingly reported only to the TRC.

In sum, the available data are shared fairly by all three sources only for deaths by the State; for most geographical strata deaths by the Shining Path and unknown perpetrators are overwhelmingly documented exclusively by the TRC. It is important to explain this feature of the data, as it imposes a big challenge to the estimation. BASM assert that the Ombudsman's Office and the human rights NGOs focused particularly on acts of the State not on the insurgent groups 'because the State had signed international agreements that obliged them to respect human rights and to sanction those who committed violations.' On the contrary, the TRC was 'mandated to investigate all crimes and human rights violations irrespective of the alleged perpetrator.' This suggests that victims of the Shining Path and of unknown perpetrators had a clear incentive to approach this new friendly institution, the TRC, created by the Peruvian State, which would listen to them for the first time. On the contrary, many victims of the State, who already denounced their case to other sources, may have been understandably less likely to denounce the State for a second or third time.[14] This feature will be crucial for the fulfillment of the assumptions required for the CRM to render unbiased estimations of the total number of deaths by perpetrator.

4   Methods

In this section, I describe the capture-recapture method (CRM) or multiple system estimation (MSE), first when there are two sources and then when there are three sources of information.

4.1   Two sources

The analysis of only two sources of information illustrates very well the basic intuition of CRMs, as well as the cases of interdependent sources and of a 'residual estimation.'

---

[14]Some TRC officials proposed that only their collected data should be official, while data from other sources and their emerging figures were biased (See Appendix A.1.).

4.1.1   Independent sources

We have a dataset of observations that come potentially from two sources, some of them from both, and some of them from only one:

	Source 2	
Source 1	Yes	No
Yes	a	b
No	c	d

   The number of observations contained in both datasets is a, the number contained in Source 1 and not in Source 2 is b, and the number contained in Source 2 and not in Source 1 is c, while the number not contained in either source is d, which is obviously unknown. The way to compute this unknown is simply by 'cross-multiplication' or the 'Rule of Three,' that is, realizing the following proportionality rule :

$$\frac{a}{b} = \frac{c}{d},$$

which determines the usual formula for the unknown value:

$$d = \frac{bc}{a}.$$

Changing notation and using subscripts to indicate whether the number is in one source or the other, $a = m_{11}$, $b = m_{10}, c = m_{01}$, and $d = m_{00}$, the formula for this estimator is

$$\widehat{m}_{00} = \frac{m_{10}m_{01}}{m_{11}}. \tag{1}$$

So, the 'Rule of Three' implies directly what in the statistical literature is known as the Lincoln-Petersen estimator.[15] It is straightforward to show that this estimator implies that

---

[15]The name of the estimator honors C.G.J. Petersen, who used the method to estimate fish populations in 1896, and F.C. Lincoln, who used the method to estimate ducks from band returns in 1930.

the total estimated population is $\widehat{N} = \frac{(m_{10}+m_{11})(m_{01}+m_{11})}{m_{11}}$. We can exemplify this with some numbers:

	Source 2	
Source 1	Yes	No
Yes	2587	15139
No	4327	m_{00}?

Total counts are $n = m_{11} + m_{10} + m_{01} = 2587 + 15139 + 4327 = 22053$ in this realistic example.[16] Applying the 'Rule of Three' we compute the unknown variable: $\widehat{m}_{00} = \frac{4327 \times 15139}{2587} = 25321$, so that $\widehat{N} = n + \widehat{m}_{00} = 22053 + 25321 = 47374$.

This simple result depends crucially on the assumption of independence of sources. Were data not gathered independently, or were there heterogeneous agents, this result would not hold in general.

4.1.2 Interdependent sources

Suppose that sources are interdependent and let ρ be the correlation coefficient between sources. Allow it to change from -1 to 1, and for each of these values calculate the total population of the previous numerical example (Appendix A.2 details the method for these). This exercise is graphed in Figure 4.

[Figure 4 here]

A correlation of -1 implies that the estimated total population just equals the total number of counts 22053. Total population is increasing in this correlation coefficient. Naturally, if $\rho = 0$, we obtain the same result as in the previous numerical example, 47374. The larger the correlation coefficient, the larger the estimated population, and as ρ approaches 1, the estimated total population goes to infinity. Biases of assuming independence are potentially

---

[16]Source 1 is actually the sum of data collected by Human Rights NGOs and the Ombudsman's Office, while Source 2 are the data collected by the TRC.

huge: when there is negative correlation, assuming independence implies over- estimating the total population, when there is positive correlation, it implies under-estimating total population. Say, if $\rho = -0.2$, then $\widehat{N} = 36334$, but the researcher who assumed independence would have computed 30.4% more than that, while if $\rho = 0.2$, then $\widehat{N} = 70313$, that is, but the researcher who assumed independence would have computed 32.6% less.

A negative correlation means that observations that appear in one source are less likely to appear in the other source, say, sources tend to cover different areas where individuals hang out. In this case, even a large number of recaptures is not indicative of a large population. At the limit, a correlation of -1 implies that the number of missing observations is simply zero, so that the total estimated population is just the total number of counts. In contrast, a positive correlation means that observations that appear in one source are more likely to appear in the other source, which implies that recaptures are indicative of a large population, so that the total estimated population can be very high. This understanding is important when the researcher applies CRMs to a given dataset, particularly consisting of people that are likely to decrease their report rate to different sources over time and to regard some sources as more sympathetic than others.

Unfortunately, with only two sources of information there is no way to estimate a correlation coefficient from the data. One would have to use information from other sources or just assume scenarios for ρ, bearing in mind that assuming 0 is as arbitrary as assuming any other possible value.

4.1.3 Residual estimation

Suppose that there are two groups, say E and S. For E there is information from all sources, but for S there is information exclusively from one source, that is, $m_{10} > 0$ and $m_{11}^S = m_{01}^S = 0$, so that it is unfeasible to estimate the total population. BASM attempt to overcome this problem by estimating the population of the sum of two groups and then substracting the estimated population of one group for which the estimation is feasible, that is, $\widetilde{m}_{00}^S = \widehat{m}_{00}^{E+S} -$

$\widehat{m}_{00}^{E}$. I call this a residual estimation.[17]

Proposition 1 *A residual estimation $\widetilde{m}_{00}^{S} = \widehat{m}_{00}^{E+S} - \widehat{m}_{00}^{E}$ when there are no recaptures in one group, $m_{11}^{S} = m_{01}^{S} = 0$, implies that the residual missing population estimate is proportional to observed counts:*

$$\widetilde{m}_{00}^{S} = \frac{m_{10}^{S} m_{01}^{E}}{m_{11}^{E}}.$$

Proof: In Appendix A.3.

For $\widetilde{m}_{00}^{S}$ we have the same Eq. (1) just with the unavailable counts of group S replaced with those of group E.

Corollary 1 *A residual estimation when there are no recaptures in one group implies that*
i) the ratio m_{10}^{S}/m_{10}^{E} indicates the ratio for the residual missing population of S over the estimated missing population of E:

$$\widetilde{m}_{00}^{S} = \frac{m_{10}^{S}}{m_{10}^{E}} \widehat{m}_{00}^{E}.$$

Proof: $\widetilde{m}_{00}^{S} = \frac{m_{10}^{S} m_{01}^{E}}{m_{11}^{E}} = \frac{m_{10}^{S} m_{01}^{E} m_{10}^{E}}{m_{11}^{E} m_{10}^{E}} = \frac{m_{10}^{S}}{m_{10}^{E}} \widehat{m}_{00}^{E}$ ∎
ii) the ratio $\frac{m_{10}^{S}}{m_{10}^{E} + m_{11}^{E}}$ determines the ratio of he estimated total population of S over the estimated total population of E:

$$\widetilde{N^{S}} = \frac{m_{10}^{S}}{m_{10}^{E} + m_{11}^{E}} \widehat{N}^{E}.$$

Proof: $\widetilde{N}^{S} = m_{10}^{S} + \widetilde{m}_{00}^{S} = m_{10}^{S} + \frac{m_{10}^{S} m_{01}^{E}}{m_{11}^{E}} = \frac{m_{10}^{S}\left(m_{11}^{E} + m_{01}^{E}\right)}{m_{11}^{E}} = \frac{m_{10}^{S}}{m_{10}^{E} + m_{11}^{E}} \frac{\left(m_{10}^{E} + m_{11}^{E}\right)\left(m_{11}^{E} + m_{01}^{E}\right)}{m_{11}^{E}} = \frac{m_{10}^{S}}{m_{10}^{E} + m_{11}^{E}} \widetilde{N}^{E}$ ∎
iii) the fraction of the estimated population of S over total estimated population is larger than

---

[17]Ball et al. (2003) compute confidence intervals for residual estimations using a jackknife variance estimator.

the fraction of observed counts over total counts

$$\frac{\widetilde{N^S}}{\widetilde{N^S} + \widehat{N^E}} = \frac{m_{10}^S}{m_{10}^S + m_{10}^E + m_{11}^E} > \frac{m_{10}^S}{m_{10}^S + m_{10}^E + m_{11}^E + m_{01}^E} = \frac{n^S}{n^S + n^E}.$$

So, if $m_{10}^S > m_{10}^E$, then $\widetilde{m}_{00}^S > \widehat{m}_{00}^E$; and if $m_{10}^S > m_{10}^E + m_{11}^E$, then $\widetilde{N^S} > \widehat{N^E}$. At any rate, the fraction of S over both populations is larger always for the total estimated population than for the total observations; and this fraction is larger, the larger is m_{01}^E. Thus, a residual estimation artificially expands the size of a group, by mixing up heterogenous groups. It only gives correct results if groups are proportional to each other.

Proposition 2 *If all counts of S are proportional to all counts of E, that is, $m_{ij}^S = km_{ij}^E$, $\forall i, j$, then a residual estimation coincides with a direct estimation, that is, $\widetilde{m}_{000}^S = \widehat{m}_{000}^S$ and $\widetilde{N^S} = \widehat{N^S}$. Proof: In Appendix A.3.*

So, what the residual estimation does in fact is to impose an inexistent homogeneity between heterogenous populations. If there are no recaptures, a residual estimation does not makes it feasible, but it just introduces a proportionality that is borrowed from other populations.

4.2 Three sources

With three sources of information there is no need to assume independence of sources as above. However, as we will see in this subsection, even with three sources some particular independence of sources happens to arise in some specifications.

4.2.1 Poisson regression

With three sources of data, there are more possible intersections and it is practical follow Fienberg's (1972, 1975) approach of performing a Poisson regression. Let m_{ijk} be the number of observations recorded by source i, j and k, where $i, k, j \in \{0, 1\}$, and 1 denotes that the

source contains those observations, and 0 otherwise. For example, m_{101} is the number of observations contained in source i, not contained in source j, and contained in source k.

The model is linear in parameters, $\widehat{y} = x\beta$, and is hereby expressed in the usual econometric notation:

$$\ln\left(\widehat{m}_{ijk}\right) = \beta_0 + \beta_1 i + \beta_2 j + \beta_3 k + \beta_4 ij + \beta_5 ik + \beta_6 jk.$$

That is, regressors in this model are just dummy variables indicating whether the source contains those observations. The missing unobserved cell is $\ln\left(\widehat{m}_{000}\right) = \beta_0$, which is the goal of the estimation.

Let l index each binary triplet (i, j, k) by $l = i + 2j + 4k$. Thus, we have seven counts indexed by $l = 1, .., 7$, with which we want to estimate at most seven parameters β. The probability for each count is given by a Poisson distribution: $\Pr\left(m_l\right) = \frac{\lambda^{m_l} e^{-\lambda}}{m_l!}$. Let $\lambda = \widehat{m}_l = e^{x_l \beta}$, then the likelihood function for all counts is

$$\mathcal{L}\left(\beta | x, m\right) = \prod_{l=1}^{7} \frac{e^{m_l(x_l\beta)} e^{-e^{x_l\beta}}}{m_l!},$$

and the log-likelihood is $\ln\mathcal{L}\left(\beta | x, m\right) = \sum_{l=1}^{7} \left(m_l\left(x_l\beta\right) - e^{x_l\beta} - \ln\left(m_l!\right)\right).$

The estimated βs will be the maximizers of this log-likelihood function: $\widehat{\beta} = \arg\max \ln\mathcal{L}\left(\beta | x, m\right).$

Certainly, estimating seven parameters with seven observations implies a perfect fit and little predicting power. So, there are seven constrained versions of this estimation: one without interaction terms, three with one interaction term and three with two interaction terms.[18] The researcher estimates all of these seven constrained versions and chooses the one with the lowest chosen target statistic, say, $\chi^2 = \sum_{l=1}^{7} \frac{(m_l - \widehat{m}_l)^2}{\widehat{m}_l}$.[19] Once a model is chosen the

---

[18]The seven possible specifications are $1 : \beta_5 = \beta_6 = \beta_7 = 0$; $2 : \beta_6 = \beta_7 = 0$; $3 : \beta_5 = \beta_7 = 0$; $4 : \beta_5 = \beta_6 = 0$; $5 : \beta_7 = 0$; $6 : \beta_6 = 0$; $7 : \beta_5 = 0$.

[19]Other popular target statistics are the Akaike information criterion, AIC $= -2\ln(\mathcal{L}) + 2K$, or the Bayesian information criterion BIC $= -2\ln(\mathcal{L}) + 2K\ln(n)$, where K is the number of parameters.

estimated missing population is $\widehat{m}_{000} = \exp\left(\widehat{\beta}_0\right)$, and a 95% confidence interval can be built from the computed standard deviation, the lower bound being $\underline{\widehat{m}_{000}} = \exp\left(\widehat{\beta}_0 - 1.96\sigma_{\widehat{\beta}_0}\right)$ and the upper bound being $\overline{\widehat{m}_{000}} = \exp\left(\widehat{\beta}_0 + 1.96\sigma_{\widehat{\beta}_0}\right)$.

4.2.2 Analytical solutions, independence and residual estimations

Fienberg (1972) also provides some explicit formulae for the Poisson regressions. I concentrate on specifications with one and two interactions, as almost all of the models selected by the TRC.[20] I just illustrate one of the possible three combinations of sources.

For one interaction, there is an explicit solution

$$\widehat{m}_{000} = \frac{(m_{010} + m_{110} + m_{100})\, m_{001}}{m_{011} + m_{101} + m_{111}},$$

which is equivalent to having independence, as the following proposition shows.

Proposition 3 *The estimated missing population with three sources and one interaction, $\widehat{m}_{000} = \frac{(m_{010}+m_{110}+m_{100})m_{001}}{m_{011}+m_{101}+m_{111}}$, is the same as the estimated missing population when two sources are added and treated as one.*
Proof: Notice that $\widehat{m}_{000} = \frac{(m_{010}+m_{110}+m_{100})m_{001}}{m_{011}+m_{101}+m_{111}}$ is equivalent as $\widehat{m}_{000} = \frac{m_{10}m_{01}}{m_{11}}$ when source one and source two are treated as single one∎

Figure 4 illustrates the result of this Proposition. An estimation with one interaction is equivalent to merging the first and second sources, and combining this union with the third source, assuming independence.

---

[20]If all three sources are independent, the estimated total population is implicitly defined from this equation $(N - m_{1**})(N - m_{*1*})(N - m_{**1}) = N^2(N-n)$, where $m_{1**} = \sum_i \sum_j m_{1ij}$, $m_{*1*} = \sum_i \sum_j m_{i1j}$, $m_{**1} = \sum_i \sum_j m_{ij1}$, and $n = \sum_i \sum_j \sum_k m_{ijk}$. Actually, for this specification there is an explicit solution: $\widehat{N} = \frac{1}{2A}\left(B + \sqrt{(B^2 + 4AC)}\right)$, where $A = m_{1**} + m_{*1*} + m_{**1} - n$, $B = m_{1**}m_{*1*} + m_{1**}m_{*1*} + m_{*1*}m_{**1}$, $C = m_{1**}m_{*1*}m_{**1}$.
If there is non-zero correlation between all three sources we have another explicit formula

$$\widehat{m}_{000} = \frac{m_{100}m_{010}m_{001}m_{111}}{m_{110}m_{011}m_{101}}.$$

[Figure 4 here]

With two interactions the explicit formula is

$$\widehat{m}_{000} = \frac{m_{100}m_{010}}{m_{110}},$$

which implies a particular form of independence between sources, as the following proposition shows.

Proposition 4 *Estimating a missing population $\widehat{m}_{000}$ with three sources allowing for two interactions, $\widehat{m}_{000} = \frac{m_{100}m_{010}}{m_{110}}$, is the same as estimating it with only two sources excluding the intersections with the third source.*

Proof: Notice that in this case $\widehat{m}_{000} = \frac{m_{100}m_{010}}{m_{110}}$ is equivalent to $\widehat{m}_{000} = \frac{m_{10}m_{01}}{m_{11}}$ when the third index is zero, that is, excluding observations that are in the third source$\blacksquare$

Figure 5 graphs the result of this Proposition. All observations in the third source are excluded from the estimation; only observations in the first and second source are used to estimate the missing population.

[Figure 5 here]

These propositions prove that the missing population is not really derived from three interdependent sources, but from *only two independent sources*, by the unification or exclusion of some sets.

This property has powerful consequences for residual estimations, as the previous analysis also applies to three sources. A residual estimation with three sources is also implicitly obtained just replacing counts of one group with counts of the other group or at most adding terms that depend only on data of the available group. These insights are discussed formally in Appendix A.4. As with two groups, residual estimations systematically increase the participation of groups that only have once source of information.

4.3 A discrete choice alternative

The previous approach, of a saturated model, does a good job in fitting the observed counts. However, besides the undesired implication of independence that it entails, it is less telling in reflecting the decision process by individual agents of reporting events to different sources. As an alternative, I propose a discrete multivariate choice model, or qualitative response model, in the spirit of the models that integrate economic theory and econometrics developed by McFadden in the seventies and eighties (see, e.g., McFadden 1981, 1984, Amemiya 1981, or Maddala, 1983 for surveys of this topic).

I adapt a conditional trivariate Probit to the problem of estimating the size of a closed population from correlated binary data contained in three overlapping sources. The purpose of this estimation is not to quantify the effect of explanatory variables on outcome variables, as other Probit models, but to predict a counterfactual, the triplet $(0, 0, 0)$, that is, the number of counts that was not captured by any source. This Gaussian specification does not have the restrictive Independence of Irrelevant Alternatives (IIA) property, undesirable for computing a population, and is thus more flexible in its covariance matrix than a logistic specification; a drawback, though, may be the difficulty of integrating a multivariate Gaussian density. In the literature this issue has been solved practically by simulation methods[21] or by a numerical integration (Genz 2004), which is the computational method followed in this article.

Using the same notation as in the previous subsection, suppose that an agent reports an event to a source when his or her utility of reporting, y, exceeds a threshold value normalized at 0,[22] that is,

$$i = \begin{cases} 1, \text{ if } y_1 \geq 0, \\ 0, \text{ if } y_1 < 0; \end{cases} \qquad j = \begin{cases} 1, \text{ if } y_2 \geq 0, \\ 0, \text{ if } y_2 < 0; \end{cases} \qquad k = \begin{cases} 1, \text{ if } y_3 \geq 0, \\ 0, \text{ if } y_3 < 0; \end{cases}$$

where $(y_1, y_2, y_3) \sim N(\mu, \Sigma)$ is a trivariate normal distribution with a variance-covariance

---

[21]Such as the Geweke–Hajivassiliou–Keane (GHK) multivariate normal simulator (Geweke 1989, Hajivassiliou and McFadden 1998, and Keane 1994).

[22]As described in Appendix A.1., data are reported by relatives of victims and further reviewed and qualified by TRC officials. As there is no information on how this review proces changed the original collected data, its modelling is left for future research.

matrix with ones on the main diagonal and correlations ρ_{12}, ρ_{13}, ρ_{23}. From this specification, we obtain probability statements for each binary triplet. For instance, counts m_{011} correspond to $\Pr(0,1,1) = \Pr(y_1 < 0, y_2 \geq 0, y_3 \geq 0) = \Phi_3(-\mu_1, \mu_2, \mu_3; -\rho_{12}, -\rho_{13}, \rho_{23})$. And, in general, $\Pr(i,j,k) =$

$\Phi_3(q_i\mu_1, q_j\mu_2, q_k\mu_3; q_iq_j\rho_{12}, q_iq_k\rho_{13}, q_iq_j\rho_{23})$, where $q_l = 2l - 1$, $l = 1, ..., 7$ is associated to each binary triplet (i, j, k) as it was done with the Poisson regression, and Φ_3 denotes the trivariate normal distribution.

Then, the conditional likelihood function for the reported data is

$$\mathcal{L}(\Theta|m) = \left[\frac{\Pr(1,1,1)}{1 - \Pr(0,0,0)}\right]^{m_{111}} \left[\frac{\Pr(0,1,1)}{1 - \Pr(0,0,0)}\right]^{m_{011}} \cdots \left[\frac{\Pr(0,0,1)}{1 - \Pr(0,0,0)}\right]^{m_{001}},$$

where $\Theta = \{\mu_1, \mu_2, \mu_3, \rho_{12}, \rho_{13}, \rho_{23}\}$. Certainly, each probability statement is conditional on the event that observations are at least in one of the three sources. The conditional log-likelihood function is then

$$
\begin{aligned}
\ln\mathcal{L}(\Theta|m) &= m_{111}\ln\Pr(1,1,1) + m_{011}\ln\Pr(0,1,1) + m_{101}\ln\Pr(1,0,1) \\
&\quad + m_{110}\ln\Pr(1,1,0) + m_{100}\ln\Pr(1,0,0) + m_{010}\ln\Pr(0,1,0) \\
&\quad + m_{001}\ln\Pr(0,0,1) - n\left[1 - \Pr(0,0,0)\right],
\end{aligned}
$$

where n is the total number of counts. The estimated parameters $\widehat{\Theta}$ will be the maximizers of this log-likelihood function.

Conditional probabilities of being in intersections of the three sources are identified from their empirical counterparts. As with the Poisson regression, there are seven constrained versions of this estimation just with interaction terms replaced by correlation coefficients: one with all zero correlations, three with one non-zero correlation, and three with two non-zero correlation.[23] And, as before, one selects the specification that yields the lowest χ^2-statistic (or AIC or BIC). The estimated missing population is then $\widehat{m}_{000} = n\frac{\Pr(0,0,0)}{1-\Pr(0,0,0)}$, and the estimated

---

[23]The seven possible specifications are $1 : \rho_{12} = \rho_{13} = \rho_{23} = 0$; $2 : \rho_{12} = \rho_{13} = 0$; $3 : \rho_{13} = \rho_{23} = 0$; $4 : \rho_{12} = \rho_{23} = 0$; $5 : \rho_{12} = 0$; $6 : \rho_{13} = 0$; $7 : \rho_{23} = 0$.

missing population's variance is

$$Var\left(\widehat{m}_{000}\right) = \frac{\partial \widehat{m}_{000}}{\partial \Theta}' Var\left(\Theta\right) \frac{\partial \widehat{m}_{000}}{\partial \Theta}.$$

A 95% confidence interval is built from the computed standard deviation, with a lower bound $\underline{\widehat{m}_{000}} = \widehat{m}_{000} - 1.96\sigma_{\widehat{m}_{000}}$ and an upper bound $\overline{\widehat{m}_{000}} = \widehat{m}_{000} + 1.96\sigma_{\widehat{m}_{000}}$.

4.4 Methods for the TRC data

BASM's approach in the TRC was to perform residual estimations whenever direct estimations were not feasible. My approach in those cases is to report the observed cases, both in Poisson and trivariate Gaussian estimations. If the estimation is feasible, to select a specification I adopt exactly the same criterion of BASM: I select specifications with p-values between 0.01 and 0.5, and from them I keep the one that produces the lowest χ^2-statistic over degrees of freedom.[24] I also report the number that fits the observed counts best, that is, the Poisson or Gauss model, that produces the lowest χ^2-statistic over degrees of freedom. I compare these results to the residual estimations of BASM.

5 Results

In this section, I report the results for estimating the total number of deaths using a Poisson Regression and using a tri-variate normal distribution function, stratifying the data by geographical area and by perpetrator.

[Table 5 here]

Table 5 shows total observed and estimated deaths, and their respective percentages by perpetrator, availability of geographical identifiers, data source and estimation method. The

---

[24]The first four Gaussians estimations coincide perfectly with the first four Poisson estimations. These are the specification with three zero correlations and the three specifications with one non-zero correlation and two zero correlations. However, Gaussian estimations generate different results than Poisson estimations when there are two or more non-zero correlations.

first rows show the number of observations by each of the three available sources, as well as the total number of observations. The next rows show point estimates, lower and upper bounds for the results of BASM, Poisson regressions, and Gaussian estimations. The numbers for the point estimates that give the best fit between Poisson and Gauss estimations are reported in the last row.[25] The first column shows the estimated total number of deaths assuming homogeneity by perpetrator; the second column assumes heterogeneity by perpetrator, that is, computing numbers of deaths by perpetrator and then adding them up. The next columns show the numbers by perpetrator, the State and the Shining Path, followed by the numbers for unknown perpetrators and for unknown locations.

As BASM estimated total deaths and those by the State using a Poisson regression, these two point estimates are very similar to the Poisson results.[26][27] Under homogeneity by perpetrator, the total number computed by the BASM is 69283 deaths, which is now the 'official' number of deaths in Peru caused by the armed conflict. The number estimated by a Poisson regression under homogeneity of perpetrator, without any adjustment, is 68229 deaths. An estimation with a Gaussian specification yields a total number of around 56000 deaths, which is substantially lower, about 20% less. The reason for this reduction is that the Gaussian estimation allows for correlation between sources, while the   Poisson regression, as shown previously, implies independence of sources.

For the State the figures of BASM are also fairly close to the Poisson estimation, both are around 20500 deaths from data of around 9500 observations. However, the Gaussian estimate for the State is only around 16500 deaths, which is again substantially lower.[28] A negative correlation between sources, people who already reported to one or two sources being less

---

[25]Detailed tables with estimations by geographical statum and perpetrator are available from the author upon request.

[26]Although Ball et al. use Poisson regressions, their confidence intervals are based on a jackknife variance estimator, whereas I use the inverse of the information matrix, as is usual in maximum likelihood estimations. Accordingly their point estimates coincide with the Poisson estimates, but not their lower and upper bounds.

[27]Discrepancies occur mostly because they applied some adjustment to some strata, choosing second-best specifications to allow the total number of deaths to be greater than the sum of the estimates by perpetrator.

[28]It is interesting to note that without any geographical statification, a Poisson estimation results in 18964 deaths, while a Gaussian results in 15114 deaths. These numbers are very similar to those obtained assuming geographical homogeneity, which suggests that data for the State are pretty homogeneous, in that partitions of the data have little impact on the total estimated number.

likely to report to a third source, under a Gaussian specification generates a lower estimate than a Poisson regression that assumes implicitly that sources are independent.[29]

For the Shining Path BASM perform residual estimation of around 31000 deaths out of around 9100 observations. This is, as explained in Section 3, inadequate, as patterns of an available group, in this case the State, are artificially attributed to another group, the Shining Path. The TRC contains more exclusive observations for the Shining Path than for the State. As seen, the ratio m_{10}^S/m_{10}^E is a crucial determinant of the expansion rate in a residual estimation, which happens to be in total 8607/3734 (see Table 3), it is not surprising that a residual estimation applied by geographical strata to the Shining Path data tends to give larger estimates than those of the State. However, if a direct Poisson estimation is performed only on strata in which it is feasible, one obtains only around 11500 casualties. By a Gaussian approach the figure is similar, around 11000 casualties.

Unknown perpetrators computed by BASM, once again using a residual estimation, are around 16000 from around 3300. However, a direct estimation only of those strata that have enough intersections yields around 7100, a lower number, yet more than twice the number of observations. A similar number is estimated using a Gaussian specification, around 7500. In a similar fashion, I also performed estimations for observations with unknown location. Poisson and Gaussian estimates are very similar, around 19000 casualties. As discussed above, these numbers should already be contained, respectively, in the estimations for the State or the Shining Path, or in the identified geographical strata, with the sole exception of MRTA and 'other perpetrators,' which amount to 486 cases. These few cases cannot be used to perform other estimations, because they are very heterogenous, as they happened in several locations.[30]

There is a trade off between fit and predicting power. The best fit between Poisson and Gauss tends to be closer to Poisson figures, as these are really a saturated model; however, the parameterized Gaussian specification accounts better for correlation between sources, when

---

[29]The expansion of 16500 from 9500 observations should contain the almost 2000 deaths by the State that were excluded from the sample, because they did not have geographical identifiers, plus part of the around 4000 observations with unkonwn perpetrators.

[30]I add the total number of deaths by the MRTA and 'other perpetrators,' 486, to the total estimated numbers by each method. They do not affect rounded numbers, except those for the Gaussian estimation, which become 27793 and are thus rounded to 28000.

there are two non-zero correlations.

Adding up numbers by perpetrator, we obtain the total under heterogeneity by perpetrator: 52000 for BASM, 32000 for the Poisson regression, and 28000 for the Gaussian estimation (these numbers become, respectively, 68000, 39000, and 35000, if 'unknown perpetrators' are added to them). The results of the Poisson regression, 32000, could have been also obtained by BASM, had they not performed residual estimations and classed 'unknown perpetrators' as a separate group. We can decompose the overestimation by BASM: considering 'unknown perpetrators' as a separate group increased the estimated number from 52000 to 68000, adopting a residual estimation increased the estimated number of deaths from 32000 to 52000, and adopting a specification that implicitly neglects correlation between sources increased the estimated number from 28000 to 32000.

As regards as percentages, the figures of BASM are 40% for the State and 60% for the Shining Path. This figure assigns the State a lower percentage not only than the data, 51%, but even than the TRC, 41%. A direct Poisson estimation implies moving in the opposite direction of BASM by estimating State's share at 64%, which lies between the 41% and the more than 92% of the TRC data and of data coming from other sources, respectively, and is an increase over the 56% reported in the sum of the data from all sources. A Gauss estimation, with its more moderate expansion of State's deaths, increases this percentage only to 60%.

[Figures 8, and 9 here]

Figure 8 offers a comparison of the total numbers of deaths in the data by geographical strata with those estimated by BASM and by a Gauss estimation. It illustrates that expansions by strata of direct Gauss estimations are lower than the residual Poisson estimations of BASM Figure 9 show a similar picture, but with the shares by the State. Once again, these figures illustrate how BASM's estimations of the State's share are smaller than the State's share in the data, while the share estimated by a Gaussian direct estimation is smaller than the State's share in the data. The opposite happens with the Shining Path for which BASM's estimations reduce their share, while the Gauss increases it.

In sum, the estimation of BASM tends to compute a relatively large number of total deaths while increasing the share therein of the Shining Path. The explanation for these results is that on the one hand they are double-counting deaths actually caused by the State and the Shining Paths as caused by 'other' perpetrators. On the other hand, residual estimations performed on data in which one source has a large number of exclusive observations, as it happens for the Shining Path, tends to produce large estimates. These are two objectionable decisions in the implementation of Poisson regressions. Another reason for an over-estimation of casualties is that Poisson regressions do not really account for correlations between interdependent sources, particularly with expansions of the deaths caused by the State. A conditional multivariate normal direct estimation shows that estimates are substantially lower than those obtained by Poisson regressions.

6 Conclusions

In this article I show that the implementation of CRMs by BASM over-estimates the total number of the armed conflict's casualties in Peru from 1980 until 2000. This over-estimation comes from including observations of unknown perpetrators that should be contained in identified perpetrators and from performing 'residual estimations,' that is, merging data for different perpetrators. I also show that, besides these implementation problems, Poisson regressions do not really account for source interdependence. Using a trivariate normal distribution I estimate the total number of deaths in around 27000, 60% by the State and 40% by the Shining Path. This number is substantially lower and has a different composition than the around 69000 deaths, 30% by the State, 46% by the Shining Path, and 24% by 'other perpetrators,' which BASM calculated for the Peruvian TC.

At a methodological level, in this article I also show that choice-based models are applicable to the problem of estimating missing reported populations from overlapping sources of information. They are flexible to account for correlated binary variables reported voluntarily by rational agents and produce results that are comparable to those obtained from Poisson regressions. The goal of the present application is not to estimate causal effects, as it is done

usually with discrete choice models, but to predict a counter-factual, the missing population.

One important point of this article is that a prediction of a population is a counterfactual, not a factual. A 'Truth Commission,' already with its authoritative name, may give the impression of an actually inexistent factuality in their predictions.[31] Truth Commissions have a very important responsibility toward the victims of violence of all sides. These may not examine the 'fine print' with the technicalities, but they certainly will put lots of expectations on their quantitative figures, which will become the basis for the narratives about the armed conflict. I consider that a cautious quantitative approach, with clear warnings about the limitations of statistical methods, would suit better the fulfillment of the goals of truth and reconciliation in post-war environments than an 'expansion-happy' approach that bears the risk of running into serious implementation problems, as we have seen in this article.

---

[31]This impression is reinforced by the use of misleading terms, such as 'undocumented' for predicted cases and 'documented' for observed cases. Even the word 'truth' is too big and particularly unusual for a careful quantitative researcher.

Appendix

A.1. The quantitative work of the TRC

The TRC was established formally on July 13, 2001 by the Peruvian government. It was led by twelve government-designated commissioners, half of which were well- known politicians, mostly of center-left wing orientation. To accomplish their mission they were given a maximum of 23 months. On October 2001, they started to work on the design of he data collection process.[32] Initial announcements at this point did not indicate that the TRC was giving much emphasis to a quantitative work: 'it is impossible to discover in only one year and a half what occurred with more than 20 thousand deaths and 5 thousand disappeared,' therefore the TRC is going to 'only investigate the most significant cases' (TRC President Salomón Lerner, October 10, 2001 in newspapers La República, Liberación, and El Peruano.).

On April 2002, the TRC determined the design of the data collection process, which included the collection of 12000 testimonies by October 31, 2002.[33] This collection consisted of sending 'mobile teams' to remote areas to interview witnesses for around one hour and a half. These teams filed a succinct report of the interview and filled out forms with socio-demographic information of the testimony. These reports went from the regional offices to the headquarters in Lima where other teams keypunched the information into the dataset and then analyzed the information trying to clarify facts, particularly trying to avoid duplication or repetition of the same case by several witnesses, as well as and eliminating inconsistencies. All these tasks were done simultaneously by several teams.

The data collection was done under the explicit criterion of maximizing the number of testimonies that the TRC could collect given its resources and time constraints. Accordingly, the TRC collected much more testimonies than its target, and kept collecting them after its deadline.

Parallel to the process of data collection several TRC official made announcements to the press. On January 2002, the construction of the TRC data set had not even started, but commissioners already announced that the number of 30 000 deaths was going to increase substantially (Commissioner Sofía Macher, January 1 2002, in the official newspaper El Peruano.). On July 15, 2002, a commissioner anticipated that 'the number of victims of the political violence between 1980 and 1992 could be twice as large as it is officially thought' (Commissioner Sofía Macher, July 15, 2002, in newspapers Gestión and Liberación.). On July, the TRC had collected 7,980 testimonies out of their target of 12,000. At most there were internal confidential reports briefing commissioners about preliminary descriptive statistics emerging from the current dataset.[34]

However, on October 2002 an internal, more assertive, report, stated that trends in the data were more stable.[35] The TRC was proposed to be the official source of the State, not

---

[32]From October 2001 until April 2002 the TRC was in the process of defining the database. For that purpose it was assisted by international consultant Michael Yard hired by USAID-OTI and Oscar Mazariegos of Guatemala hired by the ICTJ. They advised the TRC on issues such as how many people will key-punch data into the dataset, type of data, software, estimated number of records, type of reports (CVR 2002).

[33]The geographical distribution of testimonies was the following: 4320 from the South-Central region, 1980 from the Central region, 2640 from the North-Eastern region, 1200 from the South-Andean region, and 1860 from Lima.

[34]There was a short report dated July 10, 2002 with the warning that data could be biased and that they should not be cited.

[35]A TRC official mentioned to me that the initial testimonies clearly had the Shining Path as the main perpetrator, but as the TRC visited more areas and collected more testimonies the share of State agents'

other sources, otherwise, it was said, its report and recommendations will lose importance and credibility. In terms of conclusions, the 'advanced hypothesis' was that

> 'The Shining Path and State agents were responsible almost in equal proportion of the elimination (by execution, or definite disappearance of people) of a considerable number of Peruvians. The Shinning Path is responsible for a larger proportion of victims eliminated than the State. This practice was focalized in provinces and rural zones'.

In that report, it was already announced that victims of the State reported by the TRC overlapped in 41% with the data of the Ombudsman Office, while victims of the Shining Path only in 3% (Sulmont 2002). Around those dates, the TRC over-accomplished its target of collecting 12000 testimonies.

On October, the President of the TRC announced in an interview to the press that the estimated number of deaths was going to be larger than what it was thought, and the responsibility of the Shining Path in human rights violations was going to be larger (TRC President Salomón Lerner, October 9, 2002, in an interview in newspaper La República.). At the end of November and beginning of December 2002 another Commissioner announced that according to the collected information the Shining Path was the largest violator of human rights during the internal conflict (Commissioner Carlos Tapia, November 22, 2002 in newspaper Liberación and December 1, 2002 in La República.).

It is likely that these public announcements were based on internal preliminary and confidential briefings, constructed from the preliminary dataset. Certainly the dataset that was finally used in the estimations by BASM was far from being fully constructed. From November onwards the active collection of testimonies stopped, and TRC teams were devoted to verify and analyze the accuracy of the testimonies collected. As interviewers were trying to do as many interviews as possible, there were delays in transcribing those interviews into written reports. Thus, the TRC gave priority to processing over collecting more testimonies. For this stage the TRC adopted as a guiding criterion the 'presumption of veracity': victims' depositions did not need to be documented to be accepted as true (Alvarez-Perez 2003).[36] However, TRC officials still had to decide on many unclear testimonies, which were complemented with 'an analysis of the context in which violent events happened,' 'behavior patterns of armed parties,' and 'the regional historical process.' For instance, witnesses may not have reported a perpetrator, but, following the mentioned criteria, a TRC evaluator may have reported one. An important effort of this qualification process was to avoid victim duplication, as several different witnesses frequently referred to the same victim.

Patrick Ball went to Lima to advise on data collection procedures in March 2003 and on April 2003, by his recommendation, the emphasis shifted to the construction of a dataset that is useful for a quantitative account of the internal war.[37] At that moment, it was stated that the CRM was applicable to the victims caused by the State, *but not to those of the Shining*

responsibilty increased. Moreover, 'initially people did not come to the TRC, but almost at the end they lost fear to declare and more people came to submit their testimony.'

[36]Because of this qualification process, the TRC data cannot be totally considered as self-reported data. For example, in many cases in which a witness did not declare who killed his or her relative TRC officials may have used the mentioned criteria to deduce who was the perpetrator. This qualification was considered to be one of the missions of the TRC: 'to find the truth.'

[37]The TRC decision of giving priority to the construction of the dataset generated some disagreeing opinions, like the following: 'The Commission does not work with numbers, but with people. Maybe the only really tangible thing for most victims are those two or three hours that they spent with the interviewers feeling, for the first time heard and accompanied. It is people that expects with anxiety the arrival of the Commission

Path, for which the TRC was going to do simulations based on a set of hypothetical scenarios (Sulmont 2003). The estimations performed under the guidance of Ball and Asher must have been done in late May or early June 2003.

On June 2003 the President of the TRC advanced results at the United Nations in New York. The announced number of victims was between 40 000 and 60 000, and the Shining Path was the perpetrator of almost half of them (TRC President Salomón Lerner, June 18, 2003, in newspaper La República). On August 28 2003 the final full report was presented to the country, provoking a big discussion on the accuracy of the estimated figures.

After the presentation of the final report, TRC officials revealed that there were internal discussions on which numbers would be presented to the public. They preferred to report estimated rather than observed numbers, because many victims were not included in the final account of numbers due to resource and time constraints. Many late witnesses could not give their testimony or their testimony was not included in the closed dataset. TRC officials alleged that mostly poor and indigenous people were going to be excluded from the total accounts without being even noticed. The estimation, also adopted in TCs in Guatemala and Kosovo by Patrick Ball, was recommended also by the President of the American Association of Statistics and UN experts. TRC Commissioners adopted the estimations as a way to 'call attention on the drama suffered in Peru.'[38] This choice of the estimations over the data may have been also influenced by the precedent of several public announcements by TRC officials about large numbers of deaths, even before the dataset was fully constructed. That was an incentive to favor a method that did not contradict previous announcements.

A.2. Capture-recapture with a bivariate normal

We have two variables

$$i = \left\{ \begin{array}{l} 1, \text{ if } y_1 \geq 0, \\ 0, \text{ if } y_1 < 0; \end{array} \right. \quad j = \left\{ \begin{array}{l} 1, \text{ if } y_2 \geq 0, \\ 0, \text{ if } y_2 < 0. \end{array} \right.$$

Variables y_1 and y_2 follow a bivariate normal distribution with means μ_1 and μ_2, respectively, variances of one and correlation coefficient ρ. We can match two conditional probabilities to their empirical counterparts:

$$\Pr(1,1|j=1) = \frac{\Phi_2(\mu_1,\mu_2;\rho)}{\Phi(\mu_2)} = \frac{a}{a+b}, \Pr(1,1|i=1) = \frac{\Phi_2(\mu_1,\mu_2;\rho)}{\Phi(\mu_1)} = \frac{a}{a+c},$$

where Φ stands for the standard normal and Φ_2 stands for the standard bivariate normal. Here, we have only two equations for three unknowns. With independence, that is, assuming that $\rho = 0$, obviously $\frac{\Phi_2(\mu_1,\mu_2;\rho)}{\Phi(\mu_2)} = \Phi(\mu_1) = \frac{a}{a+b}$. Then $\widehat{N} = \frac{a+c}{\Phi(\mu_1)} = \frac{(a+b)(a+c)}{a}$, as the rule of three would indicate. However, we can also assume non-zero correlation between random

---

because they feel that it is their only hope. And if we can do it in places where we have not arrived yet inside the time terms that I mention, why not? Why denying those pople that opportunity and reduce the work of the TRC to a number imposed by the database. Let's not miss the ethical and humanitarian dimension of our mission. The Commission is a hopeful encounter with those who suffered, beyond the thousands of pages that will be written and whose reading we cannot guarantee' (Luque-Mogrovejo 2003).

[38]See the interventions of TRC officials Javier Ciurlizza and Daniel Manrique in Rendon (2010, p.146 and pp. 150-151.).

variables, just fixing ρ. Furthermore, we can simplify the problem realizing that

$$\frac{\Phi\left(\mu_2\right)}{\Phi\left(\mu_1\right)} = \frac{a+b}{a+c},$$

$$\mu_2\left(\mu_1\right) = \Phi^{-1}\left(\frac{a+b}{a+c}\Phi\left(\mu_1\right)\right),$$

and plugging it in one of the previous equations:

$$\Phi_2\left(\mu_1, \mu_2\left(\mu_1\right); \rho\right) = \Phi\left(\mu_1\right)\frac{a}{a+c}.$$

We have just one equation that implicitly defines μ_1, which determines $\widehat{N} = \frac{a+c}{\Phi(\mu_1)}$.

A.3. Proofs or Propositions

Proof of Proposition 1 *Let $\widehat{m}_{00}^E = \frac{m_{10}^E m_{01}^E}{m_{11}^E}$, then*

$$\widehat{m}_{00}^{ES} = \frac{\left(m_{10}^E + m_{10}^S\right) m_{01}^E}{m_{11}^E} = \frac{m_{10}^E m_{01}^E + m_{10}^S m_{01}^E}{m_{11}^E} = \widehat{m}_{00}^E + \frac{m_{10}^S m_{01}^E}{m_{11}^E},$$

$$\widetilde{m}_{00}^S = \widehat{m}_{00}^{ES} - \widehat{m}_{00}^E = \frac{m_{10}^S m_{01}^E}{m_{11}^E}\blacksquare$$

Proof of Proposition 2 *The aggregate missing population is*

$$\widehat{m}_{00}^{E+S} = \frac{\left(m_{10}^E + m_{10}^S\right)\left(m_{01}^E + m_{01}^S\right)}{m_{11}^E + m_{11}^S} = \frac{\left(m_{10}^E + km_{10}^E\right)\left(m_{01}^E + km_{01}^E\right)}{m_{11}^E + km_{11}^E},$$

$$= (1+k)\frac{m_{10}^E m_{01}^E}{m_{11}^E} = (1+k)\widehat{m}_{00}^E.$$

The residual estimation is $\widetilde{m}_{000}^S = \widehat{m}_{00}^{E+S} - \widehat{m}_{00}^E = k\widehat{m}_{00}^E$, which coincides with a direct estimation:

$$\widehat{m}_{00}^S = \frac{m_{10}^S m_{01}^S}{m_{11}^S} = \frac{km_{10}^E km_{01}^E}{km_{11}^E} = k\widehat{m}_{00}^E = \widetilde{m}_{000}^S.$$

Since $\widehat{m}_{00}^S = \widetilde{m}_{000}^S = k\widehat{m}_{00}^E$ and all observed $m_{ij}^S = km_{ij}^E$, then $\widetilde{N}^S = \widehat{N}^S = k\widehat{N}^E$∎

A.4. Residual estimations with three sources

If the two estimations use the same specification i, with one interaction we have the following:

$$\widehat{m}_{000}^{E,i} = \frac{\left(m_{010}^E + m_{110}^E + m_{100}^E\right) m_{001}^E}{m_{011}^E + m_{101}^E + m_{111}^E},$$

$$\widehat{m}_{000}^{E+S,i} = \frac{\left(m_{010}^E + m_{110}^E + m_{100}^E + m_{100}^S\right) m_{001}^E}{m_{011}^E + m_{101}^E + m_{111}^E} = \frac{m_{100}^S m_{001}^E}{m_{011}^E + m_{101}^E + m_{111}^E} + \widehat{m}_{000}^{E,i},$$

$$\widetilde{m}_{000}^{S,i,i} = \frac{m_{100}^S m_{001}^E}{m_{011}^E + m_{101}^E + m_{111}^E} = \frac{m_{100}^S \widehat{m}_{000}^{E,i}}{m_{010}^E + m_{110}^E + m_{100}^E}.$$

With three sources, the same specification i and two interactions, we have the following:

$$\widehat{m}_{000}^{E,i} = \frac{m_{100}^E m_{010}^E}{m_{110}^E},$$

$$\widehat{m}_{000}^{E+S,i} = \frac{\left(m_{100}^E + m_{100}^S\right) m_{010}^E}{m_{110}^E} = \frac{m_{100}^S m_{010}^E}{m_{110}^E} + \widehat{m}_{000}^{E,i},$$

$$\widetilde{m}_{000}^{S,i,i} = \frac{m_{100}^S m_{010}^E}{m_{110}^E} = \frac{m_{100}^S}{m_{100}^E} \widehat{m}_{000}^{E,i}.$$

As it occurs with two sources, a residual estimation implies just replacing counts of one group with counts of the other group. If the residual estimation is based on different specifications, we have similar results, just with more terms that depend only on the information of group E. When the aggregate estimation E+S is performed with a specification i with one interaction and the estimation E is done with any other specification $j \neq i$, $\widetilde{m}_{000}^{S,i,j} = \frac{m_{100}^S}{m_{010}^E + m_{110}^E + m_{100}^E} \widehat{m}_{000}^{E,i} + \widehat{m}_{000}^{E,i} - \widehat{m}_{000}^{E,j}$. When the aggregate estimation E+S is done with a specification i with two interactions and the estimation E with any other specification $j \neq i$, $\widetilde{m}_{000}^{S,i,j} = \frac{m_{100}^S}{m_{100}^E} \widehat{m}_{000}^{E,i} + \widehat{m}_{000}^{E,i} - \widehat{m}_{000}^{E,j}$.

As these formulae show, residual estimations are determined strongly by the data of the available group.

References

Alvarez-Perez, V. (2003), Informe de las actividades y resultados de la unidad de calificación de eventos y víctimas. August 26, 2003. Memorando Interno No. 048-2003-CVR/SI-UCEV. CIMC. Archivo CVR. SCO59702. Sistemas de Información. Unidad de Calificación de casos y víctimas. Procedimientos de calificación.

Amemiya, T. (1981), 'Qualitative Response Models: A Survey', *Journal of Economic Literature* **19**(4), 1483–1536.

Ball, P. (2000), The Guatemalan Commission for Historical Clarification: Intersample Analysis, *in* P. Ball, H. Spirer and L. Spirer, eds, 'Making the Case: Investigating Large Scale Human Rights Violations using Information Systems and Data Analysis', Washington., American Association for the Advancement of Science, chapter 11.

Ball, P. (2012), Calculating Body Counts. Brooke Gladstone interviews Patrick Ball. Friday, March 16, 2012. On the Media. http://www.onthemedia.org/2012/mar/16/calculating-body-counts/transcript/.

Ball, P., Asher, J., Sulmont, D. and Manrique, D. (2003), How many Peruvians have died? An estimate of the total number of victims killed or disappeared in the armed internal conflict between 1980 and 2000. American Association for the Advancement of Science, Washington, DC.

Ball, P., Betts, W., Scheuren, F., Dudukovich, J. and Asher, J. (2002), Killings and Refugee Flow in Kosovo, March June 1999: A Report to the International Criminal Tribunal for the Former Yugoslavia. Washington, DC: AAAS/ABA-CEELI.

Bishop, Y., Fienberg, S. and Holland, P. (1975), *Discrete Multivariate Analysis*, MIT Press, Cambridge MA.

Castillo, M. and Petrie, R. (2007), Discrimination in the Warplace: Evidence from a Civil War in Peru, Experimental Economics Center Working Paper Series 2007-10, Experimental Economics Center, Andrew Young School of Policy Studies, Georgia State University.

CVR (2002), Documentos de la Base de datos. CIMC. Archivo CVR. SCO58201. Sistemas de Información. Coordinación.

CVR (2003), Comisión de la Verdad y Reconciliación. Informe Final. Lima.

Darroch, J. N., Fienberg, S. E., Glonek, G. F. V. and Junker, B. W. (1993), 'A Three-Sample Multiple-Recapture Approach to Census Population Estimation with Heterogeneous Catchability', *Journal of the American Statistical Association* **88**(423), 1137–1148.

Fienberg, S. E. (1972), 'The Multiple Recapture Census for Closed Populations and Incomplete 2k Contingency Tables', *Biometrika* **59**(3), 591–603.

Galdo, J. (2010), The Long-Run Labor-Market Consequences of Civil War: Evidence from the Shining Path in Peru, IZA Discussion Papers 5028, Institute for the Study of Labor (IZA).

Genz, A. (2004), 'Numerical computation of rectangular bivariate and trivariate normal and t probabilities', *Statistics and Computing* **14**, 251–260.

Geweke, J. (1989), 'Bayesian Inference in Econometric Models Using Monte Carlo Integration', *Econometrica* **57**(6), 1317–39.

Guzmán, D., Guberek, T., Hoover, A. and Ball, P. (2007), Missing People in Casanare. Benetech Human Rights Program.

Guzmán, D., Guberek, T. and Price, M. (2012), Unobserved Union Violence: Statistical Estimates of the Total Number of Trade Unionists Killed in Colombia, 1999-2008. Benetech Human Rights Program.

Hajivassiliou, V. A. and McFadden, D. L. (1998), 'The Method of Simulated Scores for the Estimation of LDV Models', *Econometrica* **66**(4), 863–896.

Hayner, P. (1994), 'Fifteen Truth Commissions - 1974 to 1994: A Comparative Study', *Human Rights Quarterly* **16**, 597–655.

Hayner, P. (2001), *Unspeakable Truths*, Routledge, New York.

Keane, M. P. (1994), 'A Computationally Practical Simulation Estimator for Panel Data', *Econometrica* **62**(1), 95–116.

Krebs, C. J. (2009), *Ecology (6th ed.)*, Benjamin-Cummings, San Francisco.

Laszlo, S. and Grimard, F. (2010), Long Term Effects Of Civil Conflict On Women's Health Outcomes In Peru, Departmental Working Papers 2010-05, McGill University, Department of Economics.

León, G. (2009), Civil Conflict and Human Capital Accumulation: The Long Term Effects of Political Violence in Per, Technical report, University of California, Berkeley.

Luque-Mogrovejo, R. (2003), Informe de las actividades y resultados de la unidad de calificación de eventos y víctimas. August 26, 2003. Memorando Interno No. 048-2003-CVR/SI-UCEV. CIMC. Archivo CVR. SCO60103. Sistemas de Información. Secretaria Ejecutiva-Comite de Gestión.

Maddala, G. S. (2001), *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge University Press, Cambridge, Massachusetts.

Manrique, D. and Sulmont, D. (2007), FAQ estadísticas CVR. Preguntas y críticas frecuentemente formuladas sobre las estadísticas de la Comisión de la Verdad y Reconciliación del Per. http://cifrascvr.wikidot.com/.

McFadden, D. (1984), Econometric analysis of qualitative response models, *in* Z. Griliches and M. Intriligator, eds, 'Handbook of Econometrics', Vol. 2 of *Handbook of Econometrics*, North Holland, Amsterdam, pp. 1395–1457.

McFadden, D. L. (1981), Structural discrete probability models derived from theories of choice, *in* C. F. Manksi and D. L. McFadden, eds, 'Structural Analysis of Discrete Data with Econometric Applications', MIT, chapter 5, pp. 198–272.

Pollock, K. H. (1981), Capture-recapture models: a review of current methods, assumptions and experimental design, *in* C. J. Ralph and J. M. Scott, eds, 'Estimating the numbers of terrestrial birds', number 6 *in* 'Studies in Avian Biology', Allen Press Inc., Lawrence, Kansas., pp. p. 426–435.

Rendon, S. (2010), Los debates por las cifras de la CVR peruana en la Internet. Gran Combo Club: http://grancomboclub.com/wp-content/uploads/2010/08/DebateCVRagosto.pdf.

Silva, R. and Ball, P. (2006), A Report by the Benetech Human Rights Data Analysis Group to the Commission on Reception, Truth and Reconciliation of Timor-Leste. Benetech. https://www.hrdag.org/resources/Benetech-Report-to-CAVR.pdf.

Sulmont, D. (2002), Commisioners' Briefing. Preliminary and confidential advancement of results. october 17, 2002. CIMC. Archivo CVR. SCO58201. Sistemas de Información. Coordinación. Documentos de la Base de datos.

Sulmont, D. (2003), Plan de procesamiento de testimonios. 9 de abril de 2003. CIMC. Archivo CVR. SCO57801. Sistemas de Información. Coordinación. Informes de Actividades.

Table 1. Fatal victims observed and estimated by CRMs
by HRDAG consultants

Country	Year	Observed	Estimated	$\frac{\text{Estimated}}{\text{Observed}}$
TC Guatemala	1999	47803	132174	2.76
TC Kosovo	2003	4400	10356	2.35
TC Peru	2003	21950	69283	3.16
TC East Timor	2006	55043	102800	1.87
Colombia*	2007	1524	2553	1.68
Colombia**	2012	1321	1674	1.27

* Casanare, ** Trade Unionists.

Table 2. Sample selection: composition of the data by geographical and
perpetrator identifiers

Perpetrator	Numbers by geographical id			Percentages by geographical id		
	With	Without	Total	With	Without	Total
State						
State agents	9022	1939	10961	41.1	70.7	44.4
Self-defense groups	476	50	526	2.2	1.8	2.1
Paramilitary	68	9	77	0.3	0.3	0.3
Total	9566	1998	11564	43.6	72.9	46.8
Percentage	82.7	17.3	100.0			
Shining Path	9075	168	9243	41.3	6.1	37.4
Percentage	98.2	7.3	100.0			
Total: State+Shining Path	18641	2166	20807	84.9	79.0	84.3
Percentage	89.6	10.4	100.0			
'Other'						
No data on perpetrator	1291	357	1648	5.9	13.0	6.7
Undetermined	1208	80	1288	5.5	2.9	5.2
MRTA	250	13	263	1.1	0.5	1.1
Clashes with the Shining Path	243	13	256	1.1	0.5	1.0
Clashes with unknown insurgents	136	3	139	0.6	0.1	0.6
Other perpetrators	115	108	223	0.5	3.9	0.9
Clashes with MRTA	66	2	68	0.3	0.1	0.3
Total	3309	576	3885	15.1	21.0	15.7
Percentage	85.2	14.8	100.0			
Total	21950	2742	24692	100.0	100.0	100.0
Percentage	88.9	11.1	100.0			

Table 3. Number of observations with geo id by source and perpetrator
Percentages by row and by column below, in small fonts

Perpetrator	Sources							Total
	C	D	O	CD	CO	DO	CDO	
State	3734	959	1561	551	620	966	1175	9566
Row	39.0	10.0	16.3	5.8	6.5	10.1	12.3	100.0
Column	24.7	93.8	70.4	88.9	81.7	96.1	97.6	43.6
Shining P	8607	34	247	59	105	1	22	9075
Row	94.8	0.4	2.7	0.7	1.2	0.0	0.2	100.0
Column	56.9	3.3	11.1	9.5	13.8	0.1	1.8	41.3
Unknown Perpe	2781	29	410	10	34	38	7	3309
Row	84.0	0.9	12.4	0.3	1.0	1.1	0.2	100.1
Column	18.5	2.8	18.5	1.6	4.5	3.8	0.6	15.1
Total	15122	1022	2218	620	759	1005	1204	21950
Row	68.9	4.7	10.1	2.8	3.5	4.6	5.5	100.0
Column	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Note: C: TRC; D: Ombudsman Office; O: Human Rights NGOs.

Table 4. Distribution of observations over 58 geographical strata
by source and perpetrator. (Number of geographical strata in interval)

Perpetrator	Nobs	Sources						
		C	D	O	CD	CO	DO	CDO
State	0	1	2	2	1	9	8	0
	1-20	20	46	45	41	41	43	42
	21-100	26	8	10	13	7	6	14
	101+	11	2	1	3	1	1	2
Shining Path	0	0	47	57	48	34	41	46
	1-20	10	11	1	9	24	16	12
	21-100	27	0	0	0	0	0	0
	101+	21	0	0	0	0	1	0
Unknown	0	2	46	40	22	49	44	52
	1-20	23	12	18	31	9	14	6
	21-100	26	0	0	4	0	0	0
	101+	7	0	0	1	0	0	0

Note: C: TRC; D: Ombudsman Office; O: Human Rights NGOs.

Table 5. Numbers and percentages of deaths by perpetrator, data source and estimation method

Data source or estimation method	Numbers						Percentage by Perpetrator	
	Total		Perpetrator		Unknown*			
	Homog	Heterog.	State	S Path	Perpet.	Geo.	State	S.Path
Data Sources								
Human Rights NGOs	5186		4322	375	489	1366	92.0	8.0
Ombudsman's Office	3851		3651	116	84	927	96.9	3.1
Truth Commission	17705		6080	8793	2832	692	40.9	59.1
Data with Geo ID	21950		9566	9075	3309		51.3	48.7
Data without Geo ID	2742		1998	168	576	2742	92.2	7.8
All data	24692		11565	9243	3885		55.6	44.4
Estimations								
Ball et al.								
Lower Bound	39599	29051	13000	16051	6966		44.7	55.3
Point Estimate	69283	51793	20460	31333	15968		39.5	60.5
Upper Bound	104953	84301	30570	53731	30634		36.3	63.7
Poisson								
Lower Bound	46893	24742	14776	9966	4416	12094	59.7	40.3
Point Estimate	68229	31890	20507	11383	7169	18572	64.3	35.7
Upper Bound	129075	62049	43952	18097	19474	34777	70.8	29.2
Gauss								
Lower Bound	39931	22883	13443	9440	3372	10565	58.7	41.3
Point Estimate	55828	27307	16486	10821	7478	18689	60.4	39.6
Upper Bound	72122	32543	19702	12841	13187	27892	60.5	39.5
Best of Poisson-Gauss	66964	31585	20205	11380	7318	18572.0	64.0	36.0

* Includes altogether 486 observations for MRTA and 'other perpetrators.'

Table A1. Descriptive Statistics. Number of observations by geographical stratum and perpetrator. All sources of data.

Stratum	Perpetrator		Unknown	All
	State	Shining Path	Perp.*	
1. Loreto, Ucayali, Madre de Dios	367	151	139	657
2. Callao	42	1	5	48
3. Ancash (Orcos, Bolognesi)	29	23	14	66
4. Ancash (Rest)	57	46	40	143
5. Arequipa, Moquegua, La Libertad, Lambayeque, Piura, Ica, Tacna, Tumbes	71	105	58	234
6. Lima (Oyón, Cajatambo)	22	18	3	43
7 Lima (Huaura, Huarochirí, Huaral, Canta, Barranca)	40	12	19	71
8 Lima (Lima [Rimac, San Martín de Porres, Villa el Salvador, San Juan de Miraflores])	23	5	11	39
9. Lima (Lima [Rest])	185	36	53	274
10. Lima (Cañete, Yauyos)	19	21	0	40
11. Apurímac, Cusco	675	570	234	1479
12. Amazonas, Cajamarca, San Martín (Rioja, El Dorado, Moyobamba, Lamas, Huallaga)	141	32	124	297
13. San Martín (Picota, Bellavista)	12	12	33	57
14. Huanuco, San Martin (Mariscal Cáceres, Tocache)	1260	1270	689	3219
15. Pasco	68	95	101	264
16. Ayacucho (Parinacochas, Páucar del Sara Sara)	26	52	9	87
17 Ayacucho (Lucanas[Aucara, Chavipa, Chipao, Laramate, Carmen Salcedo, Cabanal])	47	57	13	117
18. Ayacucho (Lucanas [Rest]	17	54	6	77
19. Ayacucho (Huanta [Ayahuanco, Huamanguilla, Llochegua])	168	166	29	363
20 Ayachuco (Huanta [Huanta])	341	347	89	777
21 Ayacucho (Huanta [Iguain])	68	23	21	112
22 Ayacucho (Huanta [Luricocha])	68	60	8	136
23 Ayacucho (Huanta [Santullana])	199	174	26	399
24. Ayacucho (Huanta [Sivia])	37	130	45	212
25. Ayacucho (La Mar [Chungui, Luis Carranza])	956	422	31	1409
26. Ayacucho (La Mar [San Miguel])	156	205	64	425
27. Ayacucho (La Mar [Anco])	68	127	46	241
28 Ayacucho (La Mar [Ayna])	144	82	28	254
29 Ayacucho (La Mar [Chilcas])	47	98	6	151

Table A1. Descriptive Statistics. Number of observations by geographical stratum and perpetrator. All sources of data. (Cont.)

Stratum	Perpetrator		Unknown Perp.*	All
	State	Shining Path		
30 Ayacucho (La Mar [Santa Rosa])	72	132	60	264
31 Ayacucho (La Mar [Tambo])	161	252	34	447
32 Ayacucho (Cangallo)	378	163	57	598
33. Ayacucho (Huanca, Sancos)	89	190	33	312
34. Ayacucho (Sucre)	62	114	8	184
35. Ayacucho (Victor Fajardo)	436	246	103	785
36. Ayacucho (Vilcashuamán)	623	426	75	1124
37. Ayacucho (Huamanga [Quinua, Ayacucho, Pacaycasa, Jesús Nazareno, Orcos])	489	169	74	732
38. Ayacucho (Huamanga [Acos Vinchos])	17	12	2	31
39. Ayacucho (Huamanga [Carmen Alto])	32	21	2	55
40. Ayacucho (Huamanga [Santiago de Pischa, Acocro])	137	225	30	392
41. Ayacucho (Huamanga [Chiara])	101	21	23	145
42. Ayacucho (Huamanga [San José de Ticllas])	55	81	20	156
43. Ayacucho (Huamanga [San Juan Bautista])	79	28	14	121
44. Ayacucho (Huamanga [Socos])	43	6	2	51
45. Huancavelica (Huancavelica, Tayacaja, Acobamba, Churcampa	362	409	120	891
46. Ayacucho (Huamanga [Vinchos])	70	123	11	204
47. Puno	91	258	81	430
48. Huancavelica (Huaytara, Angaraes)	145	320	51	516
49. Ayacucho (Huamanga [Tambillo])	23	8	5	36
50. Huancavelica (Castrovirreyna)	35	82	6	123
51. Junín (Junín, Tarma, Yauli,Chanchamayo)	111	321	140	572
52. Junín (Jauja, Concepción, Chupaca)	170	159	119	448
53. Junín (Huancayo [Huancayo])	94	27	21	142
54. Junín (Huancayo [Chilca])	19	7	8	34
55. Junín (Huancayo [El Tambo])	48	26	15	89
56. Junín (Huancayo [Rest]	60	72	38	170
57 Junin (Satipo)	139	776	167	1082
58. San Martin (San Martín)	72	7	46	125
Sub-Total	9566	9075	3309	21950
Unknown geographical stratum	1998	168	576	2742
Total	11564	9243	3885	24692

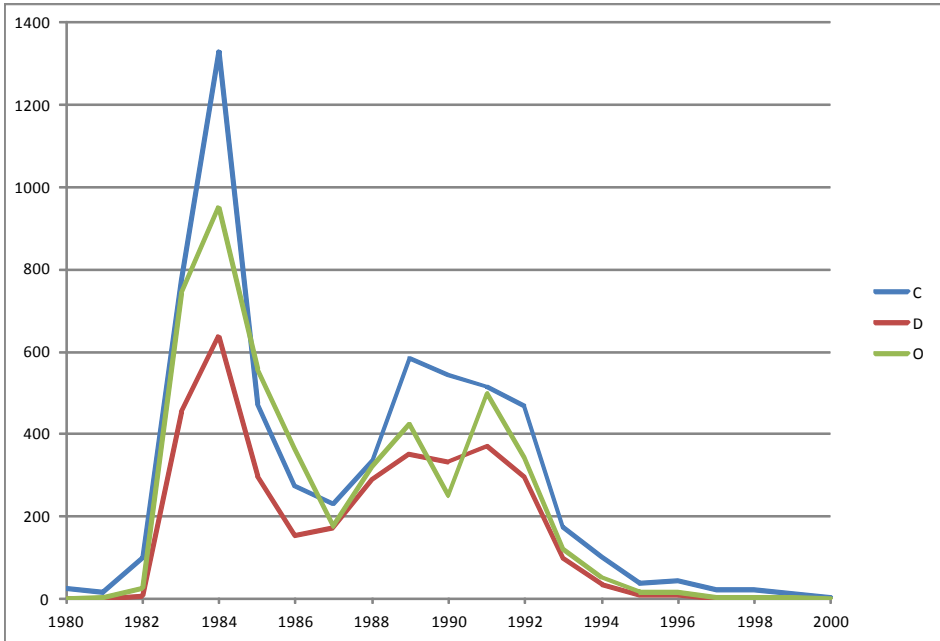* Includes observations for MRTA and 'other perpertrators.'

Figure 1. Fatal victims of the State by year and source

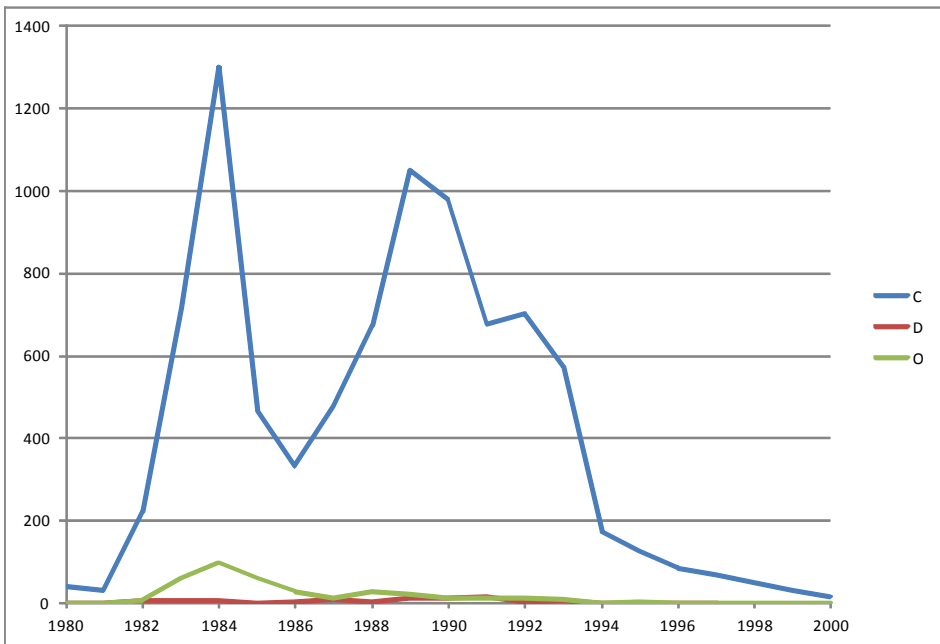C: TRC; D: Ombudsman Office; O: Human Rights NGOs.

Figure 2. Fatal victims of the Shining Path by year and source

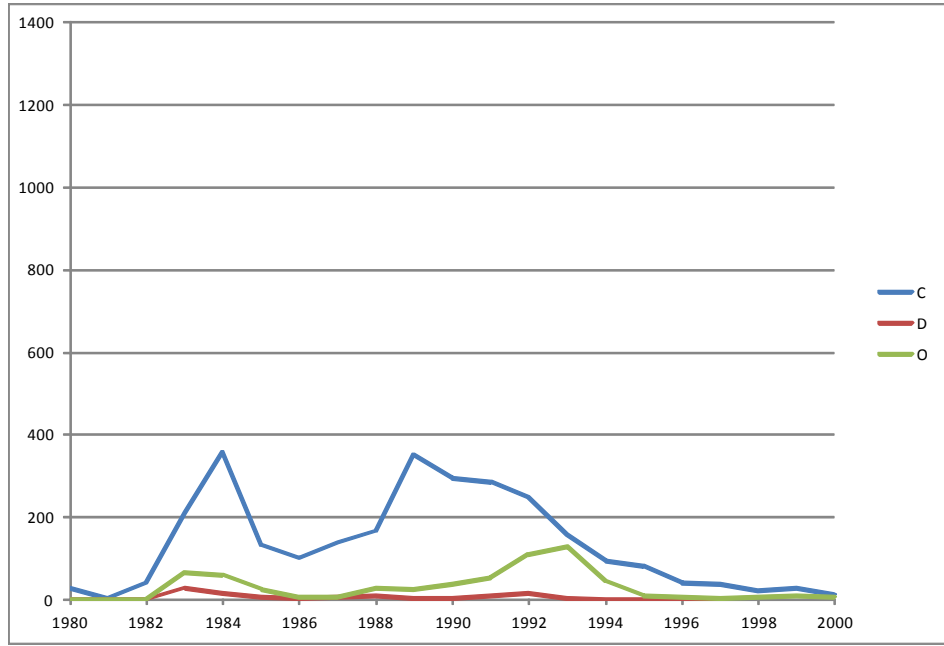C: TRC; D: Ombudsman Office; O: Human Rights NGOs.

Figure 3. Fatal victims of unknown perpetrators by year and source

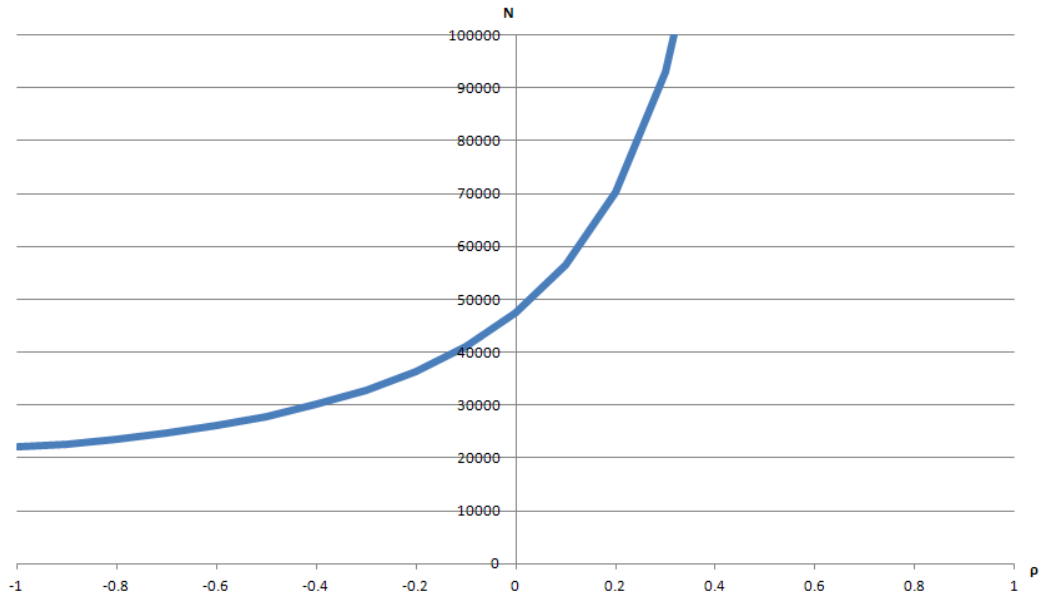C: TRC; D: Ombudsman Office; O: Human Rights NGOs.

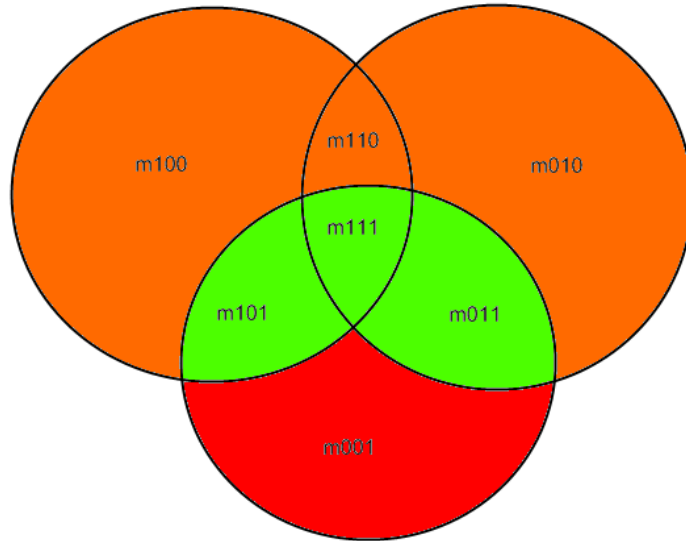Figure 4. Estimation of $\widehat{N}$ with two sources as a function of ρ.

Figure 5. Estimation with three sources and one interaction equals collapsing two sources into one
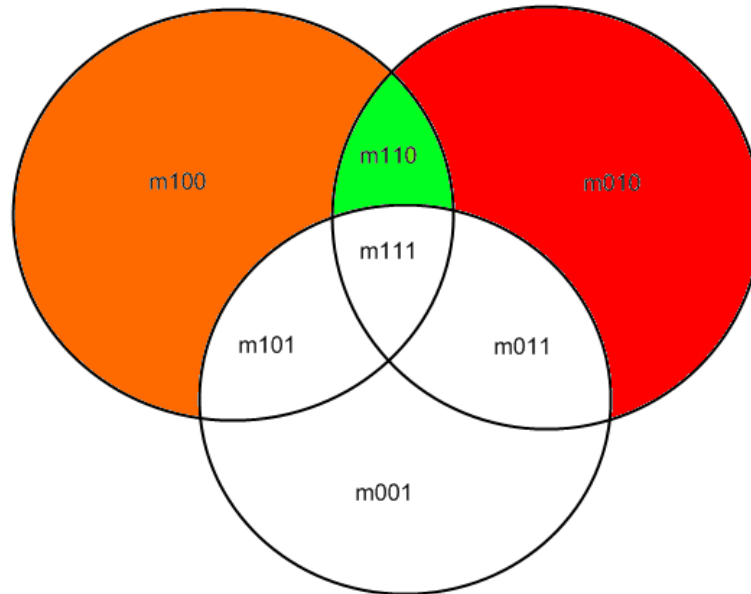
Figure 6. Estimation with three sources and two interactions equals estimation
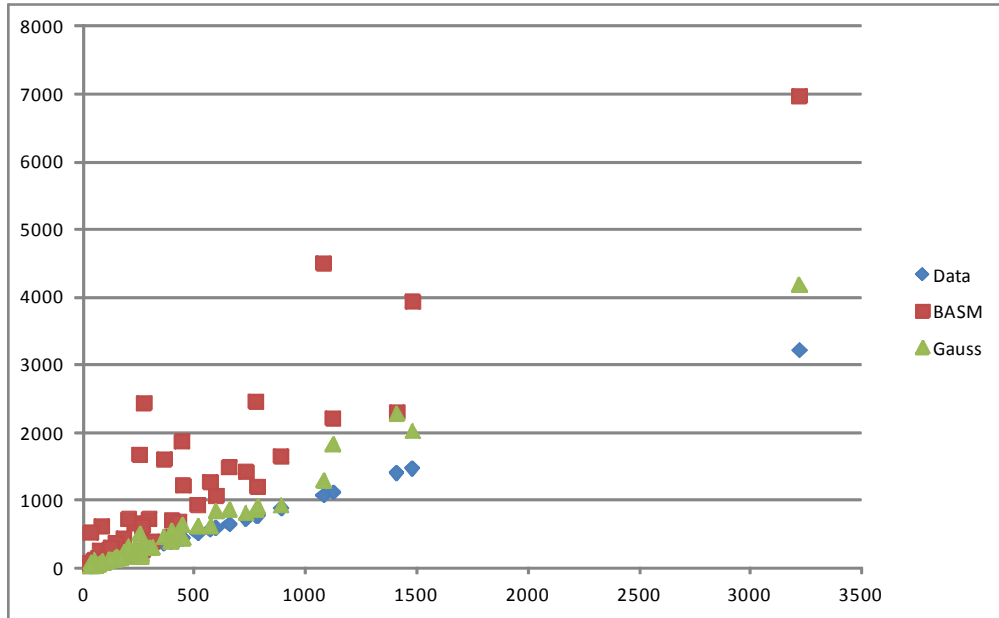of two sources excluding intersections witht the third source.

Figure 7. Data, BASM's estimation and Gaussian estimation by stratum. All.
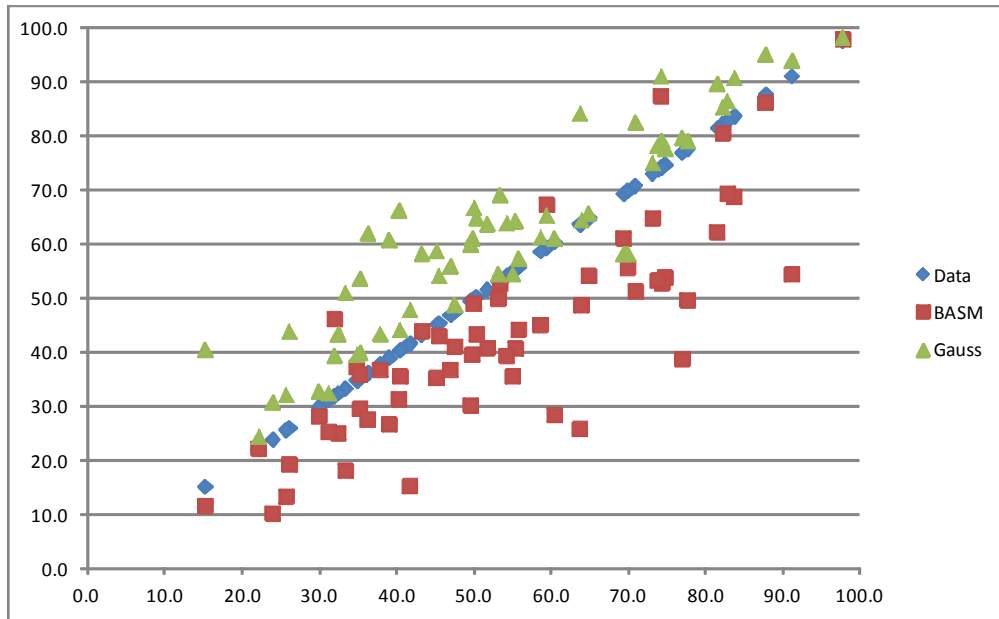

Figure 8. Data, BASM's estimation and Gaussian estimation by stratum.
Percentage of the State.