Applications of Multiple Systems Estimation in Human Rights Research

Kristian Lum Megan Emily Price David Banks

Kristian Lum, Universidade Federal do Rio de Janeiro, kristian@dme.ufrj.br
Megan Emily Price, Benetech, meganp@benetech.org
David Banks, Duke University, banks@stat.duke.edu

The authors thank Patrick Ball and Jeff Klingner for their helpful suggestions.

1 Introduction

Human rights work generally conjures images of field workers risking their lives to help people in exotic countries who are suffering horrendous abuse at the hands of sinister warlords. Less dramatic, but nonetheless increasingly vital in this effort, is the role of the statistician.

Recent trends have forced human rights workers to adopt stronger methodology in counting the dead, the disappeared, and the damaged. One reason is that rigorous estimates significantly strengthen criminal prosecution, and can cast light on the role and responsibility of leaders who do not seem to be directly accountable [Ball et al., 2002]. A second reason is that when the fieldwork estimates are unreliable, public attention is distracted from the conversation that should occur, regarding inappropriate use of force; the discussion becomes sidetracked onto numbers games, and the fog of uncertainty gives cover to parties seeking to avoid criticism and correction (see, for example, the controversy over the number of people who have died in the conflict in Darfur - `http://news.bbc.co.uk/2/hi/africa/6951672.stm`).

A third reason is that modern Truth and Reconciliation Commissions are often tasked with apportioning blame among multiple culpable parties. If the statisticians cannot defensibly infer estimates for the casualties attributable to these parties, then neither truth nor reconciliation is achieved [Ball et al., 2003]. Indeed, truth is arguably the most important thing a rigorous statistical analysis has to offer. For the human rights community to be successful, claims made by activists must be well-supported. Otherwise, they can be easily dismissed or ignored by those in power.

A fourth reason is that trends and patterns in violence need to be determined to serve a variety of purposes - for example, to identify geographic, ethnic, or other areas with higher or lower levels of violence for improved resource allocation and to

evaluate intervention strategies. Such comparisons (over time, space, demographic groups) require defensible estimates of magnitude.

To these ends, statisticians have powerful tools. In this paper, we focus on multiple systems estimation (MSE) and describe several cases in which this particular method played an indispensable part in assessing large-scale human rights violations. MSE is closely related to existing literature on capture-recapture methodology [Peterson, 1896], but we also show its connection to recent ideas in graphical models and model averaging. The methodology allows statisticians to make compelling inferences about human rights abuses that could not be obtained from simple descriptive tabulations [Jewell et al.].

Looking further ahead, the modern world needs a science of atrocity. In order to avert genocides, arrest violent outbreaks, and rebuild societies that have been riven by conflict, we must create models for the mechanisms that cause cultures to tip into chaos. Clearly, there are many pathways—Nazi Germany and Charles Taylor's Liberia had different trajectories. Useful models will be hard to build. But MSE, in conjunction with emerging work in economics [Far, 2005], sociology [Turner, 2006], public health [Cairns et al., 2009, Roberts et al., 2010], and political science [Davenport, 2007], will help identify the timing of the processes that lead to widespread violence.

2 Case Studies

In this section, we provide the historical setting of three situations in which MSE was used to estimate the magnitude of conflict-related deaths, the data available for each analysis, and the questions that motivated each project. The first case is an analysis of the number of killings that occurred in Guatemala during that country's

36 years of internal armed conflict. This analysis marks the first time that MSE was used to address the extent of under-counting of human rights violations. The second case is from Peru, where MSE estimates tripled the previously generally accepted estimate of killings. Finally, in Colombia the number and diversity of available datasets posed a novel challenge for established MSE methodology. But a modified MSE technique developed for this situation indicated that there had been a significant number of undocumented killings and disappearances.

2.1 Guatemala

Beginning with the attempted overthrow of the government of General Miguel Ydígoras Fuentes by army officers in 1960, an internal armed conflict raged in Guatemala from 1960 to 1996. During this time, the Guatemalan government carried out a massive counter-insurgency campaign of extra-judicial killings and disappearances. The government's main opponents in the conflict were guerilla groups, which formed to combat government coercion, workers' rights groups, and university intellectuals, who spoke out against the oppressive policies. [Ball et al., 1999]

Three databases describe the violence in Guatemala between 1960 and 1996 based primarily on victim and witness testimonies. These are from projects conducted by the Commission for Historical Clarification (CEH), the International Center for Human Rights Investigations (CIIDH), and the Recovery of Historical Memory (REMHI). The CEH requested that researchers at the American Association for the Advancement of Science (AAAS) analyze the three datasets to answer the question "How many people were killed in Guatemala during the period of the CEH mandate, 1960-1996?" [Ball, 1999, Ch. 11]

The CIIDH database contains multiple sources, but only testimonies were used in MSE analyses Ball [2001]

2.2 Peru

Between 1980 and the mid-1990s, the Sendero Luminoso, or "Shining Path", waged an internal revolution in Peru. They began in Ayacucho (a Peruvian state), as a committee in the pro-Chinese faction of the communist party. By recruiting university students who became teachers in rural areas, the Sendero Luminoso were able to indoctrinate the peasants. This allowed them to create a paramilitary group consisting mainly of rural Peruvian youth.

Through a combination of force and persuasion, the Sendero Luminoso wrestled authority from the local government in many rural areas. Sendero Luminoso exercised its new, sometimes unwanted, power over the former domain of the ousted local government. In the cities, the Sendero Luminoso engaged in terrorist activity. The Peruvian government eventually deployed the army to try to arrest the spreading power of the Sendero Luminoso, which resulted in many civilian casualties in what were described as "peasant massacres." Since it was often unknown whether a person was a senderista, the army's initial strategy was simply to kill anyone who was suspected. Thus, there were many casualties attributable to both the Sendero Luminoso and to the army's countermeasures [Ball et al., 2003] [Sulmont, 2005].

The Peruvian analysis had seven available databases which had been collected by very different organizations: the Peruvian Truth and Reconciliation Commission (TRC), the National Coalition of Human Rights, the Agricultural Development Center, the Human Rights Commission, the Defender of the People, and the International Committee of the Red Cross. The Peruvian TRC's data was collected with the purpose of answering questions about the political, social, and economic factors that contributed to both state and Sendero Luminoso violence and to clarify which crimes were committed by the state, which by the terrorist organizations, and who were the victims of each sets of crimes.

2.3 Colombia

Colombia is burdened with ongoing unrest. Data on conflict-related mortality is currently being collected by many different groups, despite the difficulties posed by the unsettled and often insecure circumstances. Government forces, left-wing insurgents, and right-wing paramilitaries are the actors in this, the longest-running internal armed conflict in South America.

Many groups, both within Colombia and the broader international community, have a declared interest in measuring the number of homicides, or lethal violations, occurring in Colombia. More specifically, patterns of lethal violations over time and space must be assessed to evaluate claims that demobilization of paramilitaries and amnesty laws have led to decreases in violence.

Compared to the previous examples, the statistical analysis in Colombia had access to the largest number and most diverse kinds of datasets. The 15 data sources used for the Colombian study come from state agencies (including government, security, forensic and judicial bodies) and from civil society organizations. Focusing specifically on Casanare, a single department in Colombia (similar to a US state), Guberek et al. [2010] used multiple systems estimation to make inference on the number of killed or disappeared individuals between 1998 and 2007.

3 Multiple Systems Estimation

In all three of our case studies, the researchers needed to understand patterns of violence that were not fully apparent in the *reported* data. There was reason to believe that the reported data failed to record a number of deaths in such a way that conclusions based on this data would be biased [Davenport and Ball, 46]. So they relied upon a method called multiple systems estimation (MSE) to make

inferences about the extent and kind of under-registration of the violations that were committed. As first introduced in Peterson [1896], the method was originally developed to estimate the number of fish in a pond (those that had been caught and counted, and those that had not). But despite this early application, the strategy is relevant to human rights applications as well, and enables quantification of the uncertainty regarding the number of people whose deaths were not reported.

There are many reasons why a lethal violation might not be reported. As noted in Ball et al. [2003], it may happen that victims live in remote regions where recording agencies do not exist. Some victims or the families of the victims may fear retaliation if the violation is reported. Some violations may occur without any witnesses, and some may entail other kinds of violence, such as rape, which the witness could want to conceal. For all these reasons and more, the recorded conflict mortalities represent only a lower bound on the true count.

As a simple example of two-system estimation, consider Peterson's original application, estimating the number of fish in a pond. On Monday, one casts a net and obtains n_M fish, which are tagged and released. On Tuesday, a second netting is done, which yields n_T fish, of which n_{MT} are tagged. Under the assumptions that

1. there are no births, deaths, emigrations, or immigrations (i.e., the pond is a closed population)),

2. all fish are equally likely to be caught (homogeneous capture probability),

3. being captured on the first netting does not affect the probability of being captured on the second; in particular, fish that are caught do not become more cautious (independent systems), and

4. the tagging is error free (perfect matching),

Assuming a successful de-duplication effort.

then it is simple to obtain an estimate of n, the number of fish in the pond, and the variance of that estimate. Obviously, these assumptions fail significantly in the human rights applications we consider, but careful modeling can address many of the concerns.

In the fishpond example, the assumptions imply that the Monday fish and the Tuesday fish are simple random samples of the population. The estimated probability that the Monday sample captures a particular fish is $\hat{p}_M = n_M/n$, and similarly the estimated capture probability for Tuesday is $\hat{p}_T = n_T/n$. By assuming independence, the expected number caught on both days is np_Mp_T, so the plug-in estimate solves $n_{MT} = n(n_M/n)(n_T/n)$ to get $\hat{n} = n_Mn_T/n_{MT}$. This traditional estimate is not entirely satisfactory; it is biased and the variance is not explicit. As noted in Chapman [1951], a less biased estimate of N is given by

$$\hat{n} = \frac{(n_M + 1)(n_T + 1)}{n_{MT} + 1} - 1$$

with variance

$$var[\hat{n}] = \frac{(n_M + 1)(n_T + 1)(n_M - n_{MT})(n_T - n_{MT})}{(n_{MT} + 1)^2(n_{MT} + 2)}.$$

But people are not fish. Any naïve attempt to estimate the conflict mortality from overlaps in different casualty report lists runs aground because of the failure of critical assumptions. For retrospective studies, the closed population assumption is usually valid, because the dead stay dead. In the following subsections, we discuss ways in which assumptions (2) and (3) can be relaxed and how assumption (4) is typically handled.

3.1 Heterogeneous Capture Probabilities

The assumption of homogeneous capture probability means that we believe that each data collection effort (or system) records each of the violations with equal

probability. Of course, different systems may have different probabilities of recording violations (e.g., a well-funded survey will capture more violations than an underfunded one), but within a given system, each violation has the same chance of being recorded.

This assumption is rarely defensible in the context of human rights violations. As mentioned in Bishop et al. [1975], "social visibility" is often a concern–victims who are socially very well connected are much more likely to be reported to each of the lists, whereas victims with a weaker social network are less likely to be reported to any of the data collecting organizations. This would violate the assumption that each of these two types of individuals has the same probability of being captured for each list. And it is easy to think of other mechanisms that affect the probability that a death is reported.

In the Guatemala case study, for example, it was found that the estimated probability of capture varied by location; rural killings were more likely to be reported than urban disappearances (personal correspondence with Patrick Ball). In the Guatemala case, and for other cases in which there are obvious grouping variables such as ethnicity, one possible solution is stratification. One constructs stata within which the assumption of homogeneous capture probabilities is plausible, and generates a separate estimate of conflict mortality for each stratum, which is then summed over strata to give a total. A side benefit of this is that one gains fine-grained insight into ethnic, regional and/or class aspects of the conflict.

An alternative to stratification is a formal model-based treatment of the heterogeneity of individual capture probabilities. For example, Chao [1987] addresses individual heterogeneity of capture probability by assuming a parametric form for the distribution of the individual capture probabilities and letting each individual have the same capture probability for each system, with the systems being independent. Chao et al. [1992] builds on this work by relaxing the assumption that all

of the lists have the same capture probability for each individual. Instead, Chao et al. [1992] assumes that the probability of capture for the ith person on the jth list is the product of the ith individual's capture probability and an inclusion probability for the jth list. Alternative models that explicitly model individual capture heterogeneity based on Grade of Membership, as introduced in Woodbury et al. [1978], are exemplified by Manrique-Vallier and Fienberg [2008]. In these models, each individual's capture probability on each list is modeled as a mixture of latent class variables. Basu and Ebrahimi [2001] also employs a (two-component) mixture model that allows for two different classes with distinct capture probability for each class by each system. However, it should be noted that these models have thus far been under-utilized in the context of human rights research.

3.2 System dependence

System dependence among administrative lists can be of great concern in obtaining estimates of the total number of violations. This occurs if the appearance of a reported death on one list implies that the same person is more likely or less likely to be reported on another list.

In the case of Colombia, for example, lists were collected from both government records and NGOs. It is possible that people who were inclined to trust and report a violation to an NGO were less likely to trust the government (and vice versa), which would result in negative dependence between these two systems. Positive list dependence is possible as well. For example, again in Colombia, the National Police are one of several groups that contribute to the records of homicides kept by the Vice Presidency. However, the National Police also keep their own list of homicides. Therefore, these two lists are positively correlated—records that appear on the National Police list are almost certain to appear on the Vice Presidency's list.

Positive list dependence can result in underestimation of the number of violations, whereas negative dependence results in overestimation. Although field workers may be able to predict certain dependencies from their experience with the various data collecting agencies, such contextual knowledge can typically only address pairwise dependence. Judging three-way or higher order dependence structure among systems is much more difficult to assess.

3.2.1 Log-linear models

In the motivating example with exactly two systems, the assumption of independence between the two samples was necessary, but with more than two systems there is some flexibility in modeling the dependence among the lists through the use of a log-linear model.

Using the notation of Bishop et al. [1975], imagine there are d systems, which are not necessarily independent, each of which has sampled from the population of lethal violations. Let $\{i_1, i_2, ..., i_d\}$ define an intersection of the lists, where for each individual $i_j = 1$ if the individual is present in the jth sample and 0 otherwise. For example, with $d = 3$, the intersection pattern $\{1, 0, 1\}$ corresponds to names that appear in the first system's sample, do *not* appear in the second sample, and appear in the third. Similarly, Y_{101} is the total number of names that appear on the first and third lists but not the second, and $\boldsymbol{Y}$ is the collection of counts for all possible intersections. Also let $m_{i_1 i_2 \ldots i_d}$ be the expected number of individuals that have the list intersection $\{i_1, i_2, ..., i_d\}$.

When there are more than two lists, one can use log-linear models to estimate the dependence structure among the lists. For example, define the expected counts as

$$\begin{aligned} \log m_{i_1 i_2 \ldots i_d} = \quad & u \quad + u_{1(i_1)} + u_{2(i_2)} + \cdots + u_{d(i_d)} + \sum u_{\alpha\beta(i_\alpha i_\beta)} \\ & + \quad \sum u_{\alpha\beta\gamma(i_\alpha i_\beta i_\gamma)} + \cdots + u_{12\ldots d(i_1 i_2 \ldots i_d)} \end{aligned}$$

or any subset of the above terms, where the $u_{1(i_1)}$ only appears for the $m_{i_1 i_2 \ldots i_d}$ for which $i_1 = 1$ and $u_{\alpha\beta(i_\alpha i_\beta)}$ only appears in the sum for $m_{i_1 i_2 \ldots i_d}$ for which $i_\alpha = 1$ and $i_\beta = 1$. The inclusion of such terms as the $u_{\alpha\beta(i_\alpha i_\beta)}$ provides a model that accounts for dependence between list α and β.

Because u appears regardless of which other terms and cross-terms appear in the equation, we see that the unobserved count, $m_{00\ldots0}$, would be expressed in this framework as $\log m_{00\ldots0} = u$. Bishop et al. [1975] gives several examples of closed form estimates for the maximum likelihood estimate, $\hat{N} = n + \hat{m}_{00\ldots0}$, and the associated uncertainty for each.

Notice how this relates to the simple two list estimate. Equating the current notation with that of Section 3.2, $\boldsymbol{Y} = \{Y_{11} = n_{MT}, Y_{10} = n_M, Y_{01} = n_T\}$. Then, under the log-linear model with an intercept and one coefficient for each list, the fitted value of n_{MT} is $\exp\{u + u_1 + u_2\}$. Thus, $\exp\{u\}$ corresponds to n in the two-system example, $\exp\{u_1\}$ corresponds to p_M, and $\exp\{u_2\}$ corresponds to p_T.

3.2.2 Graphical models

An alternative way to incorporate system dependence is through graphical models. In this framework, we write $Y_{i_1 i_2 \ldots i_d} \sim Multinomial(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{C_1}, \boldsymbol{\theta}_{C_2}, ..., \boldsymbol{\theta}_{C_J}\}$ is the collection of probabilities of a violation occurring in each cell of the multinomial distribution. Each $\boldsymbol{\theta}_{C_j}$ is the vector of capture probabilities for clique C_j in the decomposable graphical model describing the dependence structure of the systems. In the Bayesian framework, we let each $\boldsymbol{\theta}_{C_j} \sim Dirichlet(\boldsymbol{\delta})$, which is the conjugate prior for the multinomial likelihood and allows for easy marginalization over the nuisance parameters, $\boldsymbol{\theta}$, for posterior inference about N. Thus one can obtain closed form expressions for the posterior probability, $P(N|\boldsymbol{Y})$, free of all other unknown model parameters. For a detailed description of using graphical models to represent list dependence, see Madigan et al. [1995], and for an introduction to

graphical models generally, see Lauritzen [1996].

3.2.3 Model selection and averaging

This expanded flexibility in modeling now raises the question of which model to choose. In the log-linear setting, Bishop et al. [1975] discusses fit statistics for determining which log-linear model best describes the dependence structure exhibited by the data. The Peruvian report [Ball et al., 2003] uses this χ^2 test statistic to choose among several three-system models. Other standard model-fit statistics, such as Akaike or Bayesian Information Criterion, may also be used.

In some cases, such as the Colombia example discussed previously, these fit statistics can be almost indistinguishably similar for models that produce quite different estimates of N. In this case, it would be undesirable to completely ignore the models that are nearly as good as the best model but which tell a very different story. One solution to this is to employ Bayesian model averaging. York and Madigan [1992] shows how to use model averaging, as opposed to model selection, in the graphical model setting to average across models for list dependence based upon the marginal likelihood of each model. While each of these are reasonable options for handling all of the options for modeling the dependence structure, these methods generally assume an enumerable model space.

If one does not restrict the set of models under consideration, aside from requiring that the model contain an intercept and not be saturated, then the number of possible models is $2^{2^d-1}-2$. For the case in Colombia, in which $d = 15$, this number is incalculably large and so not all models can even be considered. Madigan et al. [1995] explains how to use Markov Chain Monte Carlo model composition to average across a universe of models which may not be easily enumerable by stochastically moving through the model space, switching from the current model to a new proposed model with probability that is a function of the ratio of the two

marginal likelihoods. Lum et al. [2010] introduce a method that considers only three systems at a time and averages over all models applied to each three-system partition with the additional layer of also averaging over all three-list partitions to create a tractable model-partition space.

3.3 Record-linkage

The fourth critical assumption for MSE calculations is perfect matching. This is necessary because the size of each list and the size of the overlaps among lists must be known. Therefore individual records must be both identifiable as unique within a list (e.g., duplicate records must be reconciled and eliminated within a single list) and identifiable when identical across lists (e.g., if the same record appears on multiple lists, it must be recognized as the same).

This can be particularly challenging within the context of estimating killings and disappearances, since in these instances we are matching individual victims between lists. In many societies, names may not be a useful personal identifier, because of common given names or surnames. Often, complete victim names are unknown, or may be intentionally withheld. Typically, victim names and additional information, such as the date or location of the violation, are necessary to generate confidence in matched records.

This problem is compounded since imperfect matching occurs both through false negative matching and through false positive matching. False positive matches are usually caused when records of two people with identical names are mistakenly thought to refer to the same person. Because both the size of the lists and the size of the overlaps influence the final estimate of the full population size, imperfect matching could either decrease or increase an MSE estimate, according to which records appear in which subsets of the lists. Thus, it is impossible even to take a "safe" strategy of conservative matching, because the effect of such a matching

policy on the final estimate is unpredictable.

Record matching is complex. The initial statistical strategy for record linkage was the model of Fellegi and Sunter [1969]. What follows is a generalization of this model. Given two records, $\mathbf{x}$ and $\mathbf{y}$, one calculates some (usually vector-valued) statistic $\mathbf{h}(\mathbf{x}, \mathbf{y})$ and estimates $\mathrm{P}(\mathrm{match} \,|\, \mathbf{h}(\mathbf{x}, \mathbf{y}))$ according to an appropriate model. If this estimated probability exceeds some user-specified value γ, then the algorithm declares that the records are a match.

To make this concrete, suppose there are two records:

$$\begin{aligned} \mathbf{x} &= \{ \text{ R. Munson, 710 Duckpin Lane, Bowling Green, KY } \} \\ \mathbf{y} &= \{ \text{ Roy Munson, 710 Duckling Street, Bowling Green, KY } \}. \end{aligned}$$

The analyst would use their best judgment to create a function $\mathbf{h}$ that extracts information from these records (e.g., match on first initial, equal street number, similar street name, same city name, same state) and build a model which describes the probability that these records refer to the same person. That model would be calibrated against, say, a corrupted version of a known database. In principle, depending on the modeling effort, this approach could take account of the fact that "Munson" is a rarer name than "Smith", increasing the estimated match probability, or that Bowling Green is relatively small, also increasing the estimated probability.

Much of the art in this science derives from the selection of the function h. Phillips [1990] let h reflect phonetic similarity: Duckpin and Duckling sound a bit alike, so if the information for either record were collected aurally, it could explain the discrepancy. Other strategies look for matches on alphanumeric strings separated by punctuation, or focus on distances between short character strings (so that misspellings have little influence); these may involve metrics that are based on "edit distance" or "affine gaps". Elmagarmid et al. [2007] provide a recent survey

of this literature. In terms of implementation, the WEKA project [Hall et al., 2009] provides machine learning algorithms that are useful for record linkage.

Additional art results from the selection of the statistical model for estimating match probabilities. The default analysis uses logistic regression, but more sophisticated record-linkage techniques employ generalized additive models or other strategies. The success of such models can be assessed by application to a real or artificial database in which known amounts and kinds of corruption have been introduced. A good assessment should be able to duplicate the kinds of mistakes that occur in the actual collection process. One advantage of such assessment is that the analyst can tune the simulation to the particulars of their circumstance; record linkage with American names and addresses is a very different problem from linking, say, Japanese names and addresses.

In the context of human rights applications, the challenge is more difficult than linking names and addresses. A report might indicate that a male body was found at a given intersection on a given date; a potential match might indicate that a specific person was abducted by armed men a day earlier at a nearby location. Building appropriate summary functions $\mathbf{h}$ and matching models is an on-going challenge. Each analysis must be hand-fit to the situation; important issues include whether the person and place names have been transliterated, whether certain names are common or rare in the relevant location, and whether recall for date is accurate.

These uncertainties should be propagated into the MSE analysis. For each pair of records, the analysis produces an estimated probability that they match. From this, one can simulate realizations of the match sets among the the different systems used to produce the MSE inference. By chance, one simulation, in which records are matched according to their estimated probabilities, might yield a relatively high estimate of duplication; a second simulation might produce a lower estimate.

By repeating these simulations many times and producing the corresponding point estimate through MSE for each, one can obtain a confidence interval on the number of casualties that takes account of the probabilistic uncertainty in the matching procedure.

4 Summary of Findings

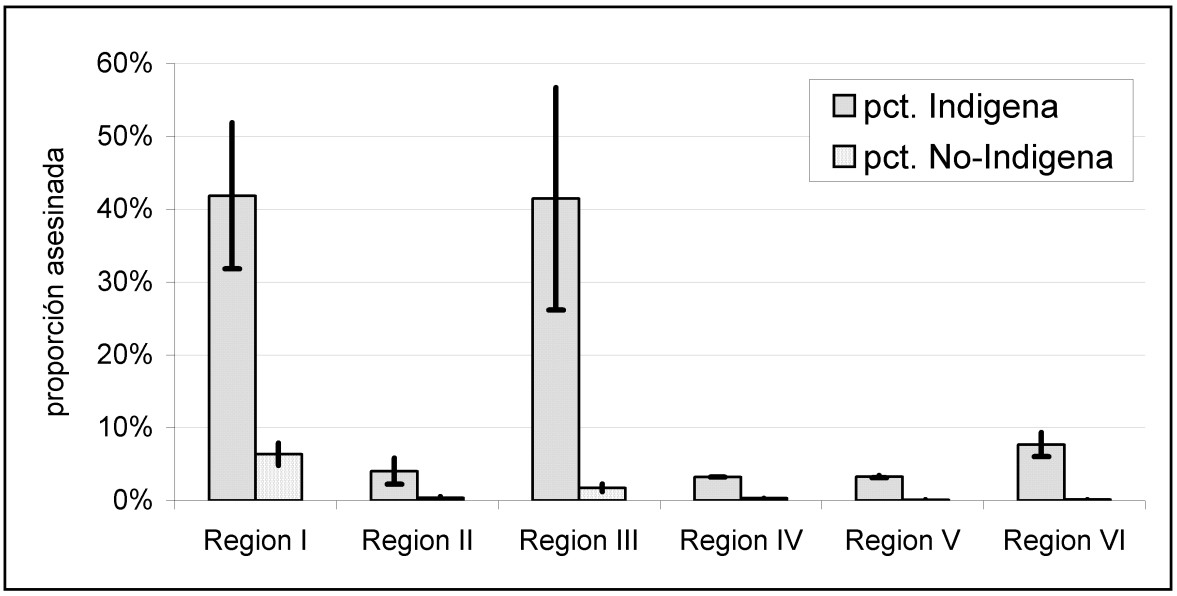
By using multiple systems methods, for each of the case studies described in this paper, analysts uncovered patterns in the conflict violence that would have been mis-represented had they relied upon traditional descriptive statistics.

In the Guatemala case, after matching and de-duplicating, Ball [1999, Ch. 11] estimate a total of 47,803 *reported* killings across the three databases described in Section 2.1. Matching was done on a random sample from each database, with matching rates used to estimate the total number of matches across combinations of entire databases. See Chapter 11 of Ball et. al. 2000 for more details. Using an expansion of the two-systems estimation technique in Section 3 [developed in Marks et al., 1974], Ball [1999, Ch. 11] estimated that approximately 84,468 killings were not reported to any project, for a total of 132,174 killings in Guatemala between 1978-1996 (SE = 6,568).

Arguably even more important than the estimated magnitude of total killings was the calculation of this estimate by ethnic group and region. The Commission for Historical Clarification (CEH) used secondary sources and qualitative evidence to identify six geographic regions during 1981-1983 in which they believed state violence was concentrated against indigenous peoples. Using the MSE method, separate estimates were calculated for each ethnic group in each region for the years 1981-1983. The 1981 census was then used to calculate a killing rate. In at least two of the six regions, indigenous peoples were killed at a significantly higher

rate than non-indigenous (see Figure 1). In fact, in these regions it was estimated that 40% of indigenous people had been killed, which was five to eight times greater than the rate for non-indigenous people. The Commission for Historical Clarification used this, among other evidence, as an indication that acts of genocide were committed against the indigenous people of Guatemala [Ball, 1999, Ch. 11].

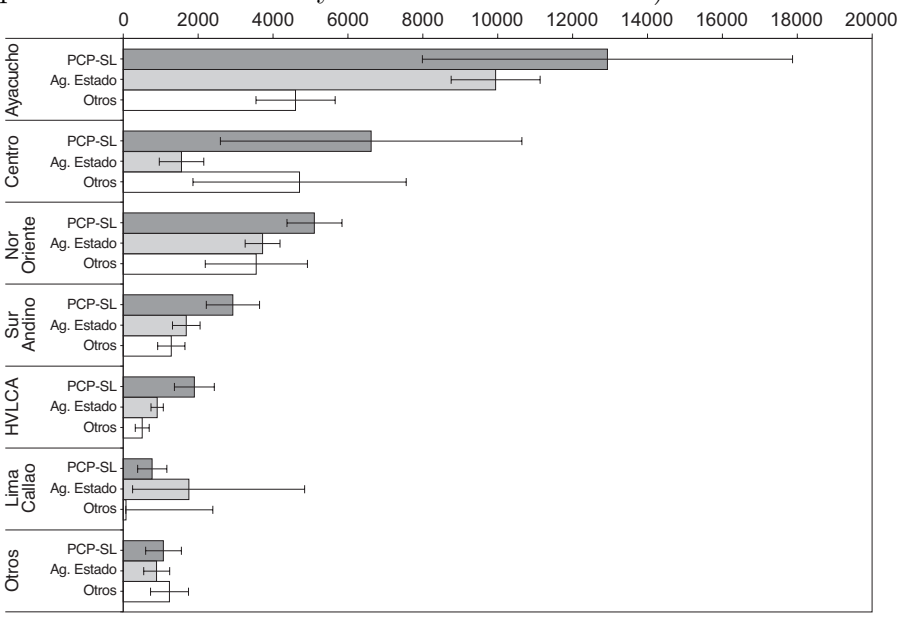
Figure 1: Estimated Proportions of Ethnic Groups Killed in Guatemala 1981-1983 (originally published in 'The Guatemalan Commission for Historical Clarification: Generating Analytic Reports,' Chapter 11 of *Making the Case: Investigating Large Scale Human Rights Violations Using Information Systems and Data Analysis*)



In Peru, MSE analyses were conducted using the log-linear model approach described in Section 3.2.1. These analyses resulted in an estimate of 69,280 deaths between 1980 and 2000, more than three times the previously accepted estimate of 25,000 deaths often reported by human rights NGOs and newspapers [Ball et al., 2003]. Additionally, the common local narrative of violence was that government

forces had been the main perpetrators. MSE analyses revealed that violations committed by the Sendero Luminoso were disproportionately under-reported, and that in fact this group was responsible for between 41% and 48% of killings and disappearances [Ball et al., 2003]. Figure 2 shows the estimated total number of victims by perpetrator group and geographic location. The perpetrator group labels are in Spanish - *PCP-SL* are the Sendero Luminoso, *Ag. Estado* are State agents, and *Otros* are others. As Figure 2 shows, the estimated number of victims of the Sendero Luminoso is greater than the estimated number of victims of state agents in every geographic region except Lima Callao.

Figure 2: Estimated Number of Victims by Region and Perpetrator (originally published in 'How Many Peruvians Have Died?')



The large amount of data available to the Colombia project required the use of both the graphical modeling approach described in Section 3.2.2 and model averaging and partitioning of data, as described in Section 3.2.3. This project

generated specific point estimates: 2,653 disappearances (95% credible interval: 1,270, 5,552) between 1998 and 2005 and 6,215 killings (3,944, 9,983) between 2000 and 2007. These estimates were also calculated for specific geographic regions based on grouping the following municipalities within Casanare:

- Geographic Region *D* - *Center*: Yopal and Aguazul
- Geographic Region *E* - *Piedemonte*: Sacama, La Salina, Tamara, Recetor, Chameza and Nunchia
- Geographic Region *F* - *South*: Tauramena, Monterrey, Villanueva, Mani and Sabanalarga
- Geographic Region *G* - *Plains*: Hato Corozal, Paz de Ariporo, Pore, San Luis de Palenque, Trinidad and Orocue

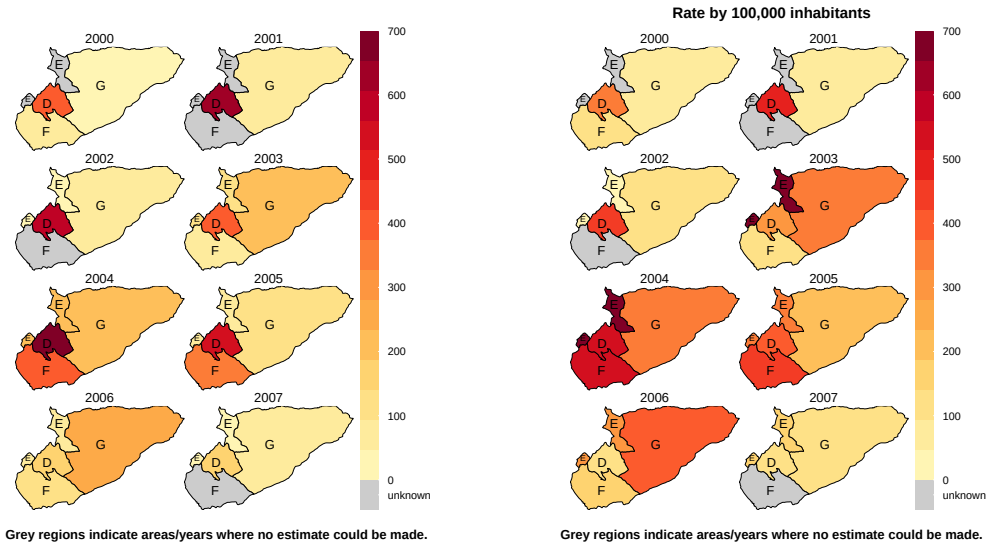
Figure 4 shows the point estimates for killings for each geographic region and year (a) and population-adjusted estimates (b).

The major findings of this research were that 1) even with access to 15 datasets, a statistically significant number of events are unreported (i.e., the number of unobserved violations is not zero), 2) the pattern of unreported events varies over time and space, and 3) descriptive comparisons of the 15 datasets revealed that even groups which might have comparable reporting mechanisms, in practice document distinct subsets of cases (such as the Vice Presidency and the National Police). It is important to note that this last point is not a criticism of the data collection mechanisms of the Vice Presidency or National Police (or any of the

We used the projections for the total population of Casanare by year as calculated by the Colombian National Bureau of Statistics. See (`http://www.dane.gov.co/daneweb_V09/index.php?option=com_content&view=article&id=238&Itemid=121`. Since we were unable to find the yearly projections disaggregated by municipality, we assigned the same proportions by municipality as the proportions in the 2005 census.

other organizations that provided datasets for the Casanare analysis) but rather a key reality that different organizations capture different proportions of the universe of interest, even when contextual knowledge leads us to expect similar samples.

Figure 3: Estimated Killings by Region and Year in Casanare (originally published in "To Count the Uncounted: An Estimation of Lethal Violence in Casanare")



(a) Killings

(b) Population-Adjusted Killings

5 Conclusion

The increase in data collection about human rights violations has allowed political scientists, policy makers, historians, and other researchers in traditionally non-quantitative fields to ask and answer questions about the patterns and extent of violence in times of conflict. Unfortunately, many of these answers are derived from convenience samples, resulting in biased conclusions. Although data collection efforts are constantly improving, it is probable that, in many situations, there are still large numbers of violations that are not recorded by any of the data-collecting

groups.

By bringing modern statistical tools to human rights research, otherwise inaccessible features of the conflict can be illuminated. These methods can potentially aid a retrospective understanding of which policies were effective in curtailing violence, who the main perpetrators were, and who the victims were, based on estimates (with quantified uncertainty) of the actual number of events, not just the reported number. As the case studies of violence in Guatemala, Peru, and Colombia show, this often results in substantially larger estimates of total conflict mortality. Furthermore, using statistical methods to separate the number of crimes reported from the number committed, the statisticians were able to scientifically support a claim of genocide in Guatemala. In Peru, the analysis discovered approximately that Sendero Luminoso were responsible for the majority of atrocities (46% compared to 30% of violations committed by agents of the Peruvian state) Ball et al. [2003]. And in Colombia, new methodology found a much larger number of victims than expected, and exposed gaps in the record collection systems. The indispensable role of statistics in the field of human rights–and in many other seemingly unquantitative fields–should not be underestimated.

References

Pathologies of Power: Health, Human Rights, and the New War on the Poor. University of California Press, Berkeley, CA, 2005.

P. Ball. *Making the Case: Investigating Large Scale Human Rights Violations Using Information Systems and Data Analysis*, volume 12. American Academy for the Advancement of Science, Washington, DC, 1999.

More precisely, a legal argument that the Army committed acts of genocide against certain Mayan groups in Guatemala citeCEH

P. Ball. Making the case: The role of statistics in human rights reporting. *Statistical Journal of the United Nations Economic Commission for Europe*, 18(2-3): 163–174, 2001.

P. Ball, P. Kobrak, and H. Spirer. *State violence in Guatemala, 1960-1996: a quantitative reflection*. American Association for the Advancement of Science (AAAS) Science and Human Rights Program International Center for Human Rights Research, 1999.

P. Ball, W. Betts, F. Scheuren, J. Dudukovich, and J. Asher. Killings and refugee flow in kosovo march - june 1999: A report to the international criminal tribunal for the former yugoslavia. American Association for the Advancement of Science, Washington, DC, 2002.

P. Ball, J. Asher, D. Sulmont, and D. Manrique. 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. Report to the Peruvian Commission for Truth and Justice (CVR), Washington, DC, August 2003.

S. Basu and N. Ebrahimi. Bayesian capture-recapture methods for error detection and estimation of population size: Heterogeneity and dependence. *Biometrika*, 88(1):269–279, 2001.

Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete multivariate analysis: theory and practice [by] Yvonne M. M. Bishop, Stephen E. Fienberg and Paul W. Holland. With the collaboration of Richard J. Light and Frederick Mosteller*. MIT Press Cambridge, Mass., 1975.

K. L. Cairns, B. A. Woodruff, M. Myatt, L. Bartlett, H. Goldberg, and L. Roberts. Cross-sectional survey methods to assess retrospectively mortality in humanitarian emergencies. *Disasters*, 33(4):503–521, 2009.

A. Chao. Estimating the population size for capture-recapture data with unequal catchabilty. *Biometrics*, 43:783–791, 1987.

A. Chao, S.-M. Lee, and S.-L. Jeng. Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal. *Biometrics*, 48:201–216, 1992.

D. Chapman. Some properties of the hypergeometric distribution with applications to zoological censuses. *University of California Publications on Statistics*, 1:131–160, 1951.

C. Davenport. State repression and political order. *Annual Review of Political Science*, 10:1–23, 2007.

C. Davenport and P. Ball. Views to a kill: Exploring the implications of source selection in the case of guatemalan state terror, 1977-1996. *Journal of Conflict Resolution*, 3(427-49):2002, 46.

A. Elmagarmid, P. Ipeirotis, and V. Verykios. Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1-16), 2007.

I. Fellegi and A. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64:1183–1210, 1969.

T. Guberek, D. Guzmán, M. Price, K. Lum, and P. Ball. To count the uncounted: An estimation of lethal violence in casanare. Technical report, Benetech Human Rights Progream, 2010.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11(1), 2009.

N. Jewell, M. Spagat, and B. Jewell. *Multiple Systems Estimation and Casualty Counts: Assumptions, Interpretation and Challenges.*

S. L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

K. Lum, M. Price, T. Guberek, and P. Ball. Measuring elusive populations with bayesian model averaging formultiple systems estimation: A case study on lethal violations in casanare, 1998-2007. *Statistics, Politics, and Policy*, 1(1), 2010.

D. Madigan, J. York, and D. Allard. Bayesian graphical models for discrete data. *International Statistical Review / Revue Internationale de Statistique*, 63(2):215 – 232, August 1995.

D. Manrique-Vallier and S. E. Fienberg. Population size estimation using individual level mixture models. *Biometrical Journal*, 50(6):1051–1063, 2008.

E. S. Marks, W. Seltzer, and K. J. Krótki. Population growth estimation: A handbook of vital statistics measurement. The Population Council, New York, 1974.

C. Peterson. The yearly immigration of young plaice into the limfjord from the german sea. *Report of the Danish Biological Station(1895)*, 6:5–84, 1896.

L. Phillips. Hanging on the metaphone. *Computer language Magazine*, 8:39–44, 1990.

B. Roberts, O. W. Morgan, M. G. Sultani, P. Nyasulu, S. Rwebangila, M. Myatt, E. Sondorp, D. Chandramohan, and F. Checchi. A new method to estimate mortality in crisis-affected and resource-poor setting: validation study. *International Journal of Epidemiology*, pages 1–13, 2010.

D. Sulmont. The peruvian truth commission (comisión de la verdad y reconciliación-cvr). Unpublished notes for the Conference at the Latin American Centre, University of Oxford, February 2005.

B. Turner. *Vulnerability and Human Rights*. Penn State Press, University Park, PA, 2006.

M. A. Woodbury, J. Clive, and J. Arthur Garson. Mathematical typology: A grade of membership technique for obtaining disease definition. *Computers and Biomedical Research*, 11:277–298, 1978.

J. C. York and D. Madigan. Bayesian methods for estimating the size of a closed population. Technical Report 234, University of Washington, July 1992.