Measuring Lethal Counterinsurgency Violence in Amritsar District, India Using a Referral-based Sampling Technique

Romesh Silva, Jeff Klingner & Scott Weikart

Abstract

When evaluating the magnitude and pattern of conflict-related mortality during armed conflicts, demographers rely on either available population census data, convenience sample data, retrospective mortality data or a combination of these data sources. By its nature, each source is vulnerable to bias and error. Existing quantitative claims about enforced disappearances and extrajudicial executions in Punjab (India) between 1984 and 1996 are not based on defensible demographic or statistical methods. We present preliminary results from a retrospective mortality survey in the rural region of Amritsar District designed to measure lethal counterinsurgency violence. The survey uses a hybrid sampling design which combines probability-based random sampling with referral-based sampling. Villages are sampled via probability proportional to size sampling. Within villages, village administrators and village elders are sampled, and asked to provide referrals to families who are resident within the village and who lost a colineal relative as a result of conflict-related violence during the counterinsurgency period. These referred families are then interviewed, documenting the conflict-related death of their colineal relative and asking for referrals to any other families they know who also directly experienced conflict-related violence during this period. All referrals within the village from village administrators, village elders and affected families are followed-up. This paper explores the dynamics of referral-based sampling through a presentation of the social-network-analysis metrics and the referral chains. This analysis is the basis for determining the coverage rate of the referral-based sampling in villages within which caste and socioeconomic status may influence social knowledge and social relations.

1 Existing Research on the Problem of Incomplete Data Sources

When evaluating the magnitude and pattern of conflict-related mortality during armed conflicts, social scientists often rely on either available population census data, convenience sample data, retrospective mortality data or a combination of these data sources. By its nature, each convenience source is vulnerable to bias and error. Exclusive reliance on a single convenience sample can produce misleading conclusions about both the actual pattern and the actual magnitude of direct-conflict deaths. As an example, Ball et al. showed that direct-conflict deaths reported by convenience-sample data systems, such as media reports, were inversely correlated with violent deaths reported

in interviews in Guatemala during the most intense years of that conflict.[1] It has also been shown that retrospective mortality surveys of conflict-related deaths that draw on classical survey sampling methods are effective in estimating relatively common conflict-related deaths among the general population. However, they are inefficient and prone to considerable sampling error in estimating deaths that are relatively rare and hidden amongst the general population.[2] Further evidence of the incomplete and biased nature of data on conflict-related mortality during armed conflict was generated in Timor-Leste by Silva and Ball. They showed that classical survey methods were effective in retrospectively estimating the magnitude and pattern of famine-related deaths common among the general population, but subject to considerable error and uncertainty when used to estimate the highly-targeted phenomena of killings and disappearances.[3] Lastly, population census data on communities in affected areas is subject to substantial data quality and coverage defects, which leads to incomplete records.[4] Indirect estimation techniques that draw on population census data are difficult to adjust for the considerable internal and external conflict-related migration that may have occurred. Further, such national-level census data by themselves rarely provide the level of precision needed to examine conflict-related mortality which is concentrated in a particular region of the country or experienced almost exclusively by a small minority or subpopulation.

2 The Importance of Developing Quantitative Estimates of Conflict-Related Mortality

When the world considers mass human rights violations, among the first questions is "How many?" Scientifically-defensible answers about the statistical patterns and overall magnitude of massive violence can depoliticize arguments about the past. Rigorous statistics about violence help to overcome partisan arguments about blame and victimhood, instead fostering honest debates about truth, accountability and reconciliation. The moral legitimacy of international human rights and humanitarian work rests on practitioners' claims that they seek to uncover the truth about complex, politicized situations, and to ensure that the human rights and humanitarian response to such situations is guided by an accurate assessment of what happened.

International human rights norms and international humanitarian law have grown over the last 60 years to provide an important set of widely-accepted principles and norms. As Seltzer (2009) notes, these bodies of law provide a detailed framework for understanding and organizing evidence

[1] Ball, Patrick, Paul Kobrak, and Herb Spirer. 1999. *State Violence in Guatemala: a Quantitative Reflection.* Washington, DC: AAAS.

[2] Spiegel, P.B., and P. Salama. 2000. "War and Mortality in Kosovo, 1998–1999: An Epidemiological Testimony." *Lancet* 355:2206–2211.

[3] Silva, Romesh and Patrick Ball. "The Demography of Conflict-Related Mortality in Timor-Leste (1974–1999): Empirical Quantitative Measurement of Civilian Killings, Disappearances & Famine-Related Deaths" In *Statistical Methods for Human Rights*, J. Asher, D. Banks and F. Scheuren, eds., Springer (New York) (2007).

[4] Heuveline, Patrick. 1998. "Between One and Three Million in Cambodia: Toward the Demographic Reconstruction of a Decade of Cambodian History (1970–1980)." *Population Studies* 52(1): 49–65.

from situations where large-scale human rights violations have taken place. In this vein, a diverse array of professionals from many disciplines have begun to apply quantitative methods to engage basic questions of magnitude and pattern. Such questions include: "How many conflict-related deaths resulted from armed hostilities?", "Does the pattern of conflict-related mortality suggest that these acts were part of a plan or policy by those responsible?" and "Were certain subpopulations of the community more vulnerable, or even targeted by those responsible for the resultant human suffering?" Specialists engaging such questions include statisticians, demographers, epidemiologists, sociologists, political scientists, anthropologists, and economists. The context for this work has been wide-ranging, and includes efforts to organize and improve humanitarian and public-health interventions during armed conflict, the historical clarification of what happened during the conflict, the organizing of evidence for domestic and international prosecutions of those most responsible, and efforts to shape institutional reform of organizations which either directly or indirectly contributed to conflict-related mortality. However, regardless of the varied contexts and forums in which such work is developed, the scientific quantification of mass human rights violations seeks to move beyond partisan rhetoric and informal anecdotes. Such analyses seek to identify errors and biases and quantify the uncertainty. Rigorous quantitative conclusions can be tested according to clear and well-established scientific standards. When quantitative data in the service of human rights fail these standards, the credibility of the human rights enterprise and of the entire victim population can be brought into question. Statistics in the service of human rights must be transparent, reproducible and defensible.

3 Previous Survey Based Approaches to Measuring Elusive Phenomena such as Direct Conflict Deaths

Survey implementation is often haphazard and vulnerable to considerable bias during complex humanitarian emergencies, which leads to several methodological limitations. In most crises, the population at risk is rarely stable. Lists of households are non-existent and the residential layout is chaotic, making simple or systematic random sampling difficult. As a result, survey researchers often resort to multi-stage cluster sampling when trying to estimate conflict-related mortality in such situations. Cluster sampling, although easier to implement, is vulnerable to large imprecision and high design effects, which yields large confidence intervals, hampering the interpretation of estimates.

Conflict-related mortality can usually be broken down into direct and indirect conflict-related deaths. Direct deaths refer to those deaths which result from acts of violence such as killings, disappearances and armed hostilities. Whereas indirect deaths refer to those deaths which do not result directly from violent trauma in a conflict, but by the deterioration in health services and food access plus increased risk of disease attributable to the violence. These two types of deaths are rarely distributed in similar ways. Direct-violence deaths tend to be elusive, and targeted at certain particular subpopulations. Whereas indirect-violence deaths, particularly in communities with only

basic transportation, health and communications infrastructure, tend to occur at a population-level.

There are considerable challenges involved in using statistical survey methods to measure elusive populations (Sudman et al., 1988).[5] Elusive populations are defined by two key characteristics. First, they are rare at the "global" population level. Second, there is lack of member registries that could be utilized as a sampling frame to aid estimation of the total magnitude and nature of the hidden population. Sudman et al. discuss the challenges involved in sampling hidden populations.[6] In particular, they note that the use of classical, probability-based sampling techniques to measure hidden populations are inefficient, expensive and subject to very large standard errors. As a consequence of the weakness of classical sampling methods, we will use adaptive methods.

Silva and Ball (2006), in their study of Timor-Leste, have noted the particular issues related to conflict-related mortality in conflict-zones.[7] In particular, they find that classically-designed sample surveys are appropriate for estimating population-based phenomena (such as famine deaths, in the case of Timor-Leste), but are subject to considerable uncertainty for highly targeted and elusive phenomena (such as direct political violence deaths in Timor-Leste).

The development of sample survey techniques to estimate conflict-related mortality in an unsettled environment poses two major challenges:

- Designing a sampling strategy that is able to efficiently document the "hidden" phenomena of conflict-related mortality whereby the selection probability of an individual can be explicitly calculated;
- Developing appropriate survey estimation techniques to estimate conflict-related mortality by sampling surviving relatives of the victims of conflict-related deaths.

Recent advances in the statistical sampling of elusive phenomena seem particularly promising for conflict-related mortality researchers. In particular, adaptive sampling designs and respondent-driven sampling have shown much promise when measuring hidden, vulnerable populations. In contrast to conventional sampling designs, adaptive sampling makes use of values observed in the sample to determine additional units to be sampled.

In adaptive sampling, the sampling statistician specifies

- the initial sampling design (prior to any adaptive sampling)
- the initial sample size
- the description of the neighborhood for a sampling unit

[5]Sudman Seymour, Sirken Monroe G., Cowan Charles D. (1988) "Sampling Rare and Elusive Populations." *Science* 240 (4855): 991–6.

[6]Sudman S, M.G. Sirkin, and C.D. Cowan .1988. "Sampling Rare and Elusive Populations." *Science* 240:991–95.

[7]Silva, Romesh and Patrick Ball. "The Demography of Conflict-Related Mortality in Timor-Leste (1974–1999): Empirical Quantitative Measurement of Civilian Killings, Disappearances & Famine-Related Deaths." In Statistical Methods for Human Rights, J. Asher, D. Banks and F. Scheuren, eds., Springer (New York) (2007)

- the condition that triggers or initiates adaptive sampling at a sampled unit

The neighborhood is specified in advance of sampling and is not adaptive. If the y-value (response) of a sampled unit satisfies the condition for adaptive sampling, say, $y \in C$, then the unit's neighborhood is added to the sample. If any other units in that neighborhood satisfy C, then their neighborhoods are also added to the sample. The process continues until a cluster of units is obtained that contains a "boundary" of edge units that do not satisfy C. The final sample consists of n_1, not necessarily distinct, clusters, one for each unit selected in the initial sample. If many of the units satisfy the condition, then the sample could consist of most of the units in the population, and hence be very costly. Thus, the design is most appropriate when the characteristic of interest is highly aggregated or clustered.[8]

In respondent-driven sampling, the analyst uses the social networks of the reference population and models the referral process as a first-order Markov chain. Heckathorn (1997) and Heckathorn & Salganik (2002) developed Respondent-Driven Sampling (RDS) to overcome some of the limitations and biases to which convenience sampling and snowball sampling are subject. RDS is specifically designed to measure elusive populations which are often vulnerable as a result of the stigma attached to membership in the population, and for which no comprehensive sampling frame exists. RDS integrates the idea of chain-referral, which earlier snowball sampling techniques used, with a Markov model of the peer recruitment process, a theory of steering-incentives and a post-stratification weighting system. The modeling of the referral and recruitment process as a first-order Markov process attempts to overcome the potential bias associated with the choice of the initial respondents (often referred to as "seeds"). The reformulation of the recruitment process around a dual incentive system, whereby respondents recruit other respondents (as opposed to recruitment being implemented by the survey team on the basis of names and contact information reported by respondents), is designed to overcome the potential bias of "volunteerism" and the effects of "masking." [9] The post-stratification weighting system is designed to control for potential biases which result from the over-representation of individuals who are members of the social networks of individuals who have large personal social networks.

RDS, in contrast to traditional snowball sampling methods, is based on direct peer-based recruitment. This peer-based recruitment system uses a dual incentive structure: respondents are given financial incentives to respond to the survey, as well as additional financial incentives to recruit their peers to become survey respondents. By elucidating the social networks of members of the population of interest, RDS is able to reach individuals who may not be directly observable in the general population using traditional sampling techniques.

[8] Thompson, Steven K. (1991) "Adaptive Cluster Sampling: Designs with Primary and Secondary Units." *Biometrics* 47(3):1103–1115

[9].

4 Towards Two Approaches of Estimating Conflict-Related Mortality during Counterinsurgency Operations in Amritsar District

When multiple data sources are available, they can be triangulated to estimate the total magnitude and pattern of conflict-related mortality during an armed conflict. Triangulation of such data sources can overcome some of the considerable limitations of a single individual data source and control for its associated biases. Demographers have pioneered work in this area, drawing particularly on recent advances in data mining techniques and the demographic technique of Multiple Systems Estimation (MSE). However, past work in this area has relied heavily on data systems which each measure the same geographic area or reference years. Yet in some cases, like that of Punjab, some existing data sources sometimes are confined to certain years or certain geographic regions (notably Amritsar District, in the case of existing data on conflict-related mortality in Punjab).[10] Hence, in order to engage questions about conflict-related mortality and mass cremations across the entire state of Punjab, extension of existing data sources, data modeling and estimation techniques is required, and the development of new sampling techniques is needed.

Previous research on conflict-related mortality in Timor-Leste showed how MSE estimates and survey estimates complement each other (Silva and Ball, 2006). MSE and survey techniques measure the same phenomenon by drawing on different types of data: administrative records and incomplete registrations lists vs. sample-survey data. They use very different methods of statistical inference: statistical modeling of probabilities of inclusion/exclusion in multiple registration lists vs. design-based population estimators. Scientific results which are reproducible using different methods and different data sources tend to have higher levels of reliability and validity relative to other kinds of estimation. And, better estimates have greater potential to impact institutions which are developing future human rights policy.

This paper presents preliminary results from a household survey of rural Punjab designed to measure lethal counterinsurgency violence. The survey uses a hybrid sampling design which combines the strengths of classical survey sampling and adaptive sampling techniques to improve the coverage rate of existing data sources. Adaptive sampling techniques are particularly attuned to the sampling of "hidden populations" for which researchers do not have a reliable sampling frame. Adaptive sampling is a new sampling design in which sampling regions, defined as "units," are selected based on values of the variables of interest observed during a sampling survey. In a conventional sampling design, the selection for a sampling unit does not depend on previous observations made during an initial survey; entire sampling units are selected before any physical sampling in the field ever takes place. Therefore, conventional sampling guarantees that the calculated statistics will be unbiased. To use the adaptive sampling technique, however, new estimators must be implemented to guarantee unbiasedness. In our hybrid sampling plan, villages are sampled via probability

[10] Existing data sources are mostly focused on Amritsar District, as all past official investigations and processes into mass forced disappearances in Punjab between 1984 and 1996 have had their geographic mandate limited to this particular district.

proportional to size cluster sampling. Within villages, village administrators and village elders are sampled and asked to provide referrals to families who are resident within the village and who lost a colineal relative as a result of conflict-related violence during the counterinsurgency period. These families are then interviewed; they're asked to document the conflict-related death of their colineal relative, and also asked to provide referrals to any other families they know who also directly experienced conflict-related violence during this period. All referrals within the village, from village administrators, village elders and affected families, are followed-up. This paper explores the dynamics of referral-based sampling through a presentation of the social-network-analysis metrics and the referral chains. This analysis is the basis for determining the coverage rate of the referral-based sampling in villages within which caste and socioeconomic status may influence social knowledge and social relations. Consideration of this coverage-rate analysis is important when making survey-based estimates which do not require the strong assumption of uniform exhaustive coverage of all sampled villages in the survey, but rather account for the heterogeneity across villages due to variation in social knowledge and social networks between and amongst village administrators, village elders and the families of victims of lethal violence.

The survey data presented in this paper provide means for making both survey-based estimates and inferences using multiple systems estimation (MSE). The MSE-based estimates can be derived by triangulating these survey data with other existing data sources which are essentially independently-collected, incomplete registers of lethal violence deaths. By triangulation, we mean using multiple data sources to produce estimates, to overcome the considerable limitations of a single individual data source and its idiosyncratic biases. There is a rich and growing literature on using such methods in the demography of armed conflict.[11] Multiple systems estimation (MSE), or capture-recapture, is a technique that uses two or more separately collected but incomplete lists of a population to estimate the total population size. Demographers have used the MSE technique in human rights cases to project the total number of violations, including those that were never documented. This information is vital to producing a complete and accurate historical record of the violations. Silva, Marwaha and Klingner (2009) present a preliminary analysis of these existing data sources.[12] These existing data sources include (1) a relational database compiled between 1997 and

[11] Ball, Patrick, Jana Asher, David Sulmont, and Daniel Manrique. 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." Washington, DC: AAAS.
Guzmàn, Daniel, Tamy Guberek, Amelia Hoover, and Patrick Ball. 2007. "Missing People in Casanare." Benetech. Available online at `http://www.hrdag.org/resources/publications/casanare-missing-report.pdf`
Ball, Patrick and Wendy Betts, Fritz Scheuren, Jana Dudukovich, and Jana Asher. "Killings and Refugee Flow in Kosovo March – June 1999: A Report to the International Criminal Tribunal for the Former Yugoslavia." 3 January 2002.
Brunborg, Helge, Torkild Hovde Lyngstad, and Henrik Urdal, 2003. "Accounting for Genocide: How Many Were Killed in Srebrenica?" *European Journal of Population* 19(3): 229–248.

[12] See Silva, Romesh, Jasmine Marwaha & Jeff Klingner, (2009) "Violent Deaths and Enforced Disappearances During the Counterinsurgency in Punjab, India: A Preliminary Quantitative Analysis" A Joint Report by Benetech's Human Rights Data Analysis Group & Ensaaf, Inc. Available online at `http://www.hrdag.org/about/india-punjab.shtml`.

2001 by a coalition of non-governmental organizations and community-based organizations, called the Committee for Coordination on Disappearances in Punjab, which includes approximately 1,800 disappearance events; (2) a list of 500 disappearance cases documented by the Peoples Commission on Human Rights in 1999; (3) more than 14,000 encounter killings reported between 1988 and 1996 in the Tribune ()the major daily English newspaper in Punjab), and the Ajit newspaper (the major Punjabi-language newspaper); (4) three lists of illegal cremations, containing approximately 2,000 cases, compiled by the Central Bureau of Investigation in 1996 and submitted to the National Human Rights Commission; and (5) official logbooks, compiled by state crematoria officials, which document transfers of corpses of the disappeared from the police to the crematoria, and purchase orders for firewood. The MSE-based estimates which combine both the survey-based data and existing convenience-sample data will be presented in a subsequent, forthcoming paper.

5 Quasi-Adaptive Sampling to Measure the Elusive Population of Those Forcibly Disappeared

Adaptive sampling techniques are particularly effective for sampling "hidden populations" for which researchers do not have a reliable sampling frame. Adaptive sampling is a new survey sampling design in which sampling regions, defined as "units," are selected based on values of the variables of interest observed during a sampling survey. In a conventional sampling design, the selection for a sampling unit does not depend on previous observations made during an initial survey; entire sampling units are selected before any physical sampling in the field ever takes place. Therefore, conventional sampling guarantees that the calculated statistics will be unbiased. To use the adaptive sampling technique, however, different estimators must be implemented to guarantee unbiasedness.

The development of sample survey techniques to estimate conflict-related mortality across rural areas of Amritsar District poses two major challenges. First, we must design a sampling strategy that is able to efficiently document the "hidden" phenomena of conflict-related mortality. This requires that selection probability of an individual death can be explicitly calculated. Then we develop survey techniques to estimate conflict-related mortality from responses of surviving relatives of the victims of conflict-related deaths.

Demographers refer to populations such as those individuals who were disappeared between 1984 and 1996 in Punjab as "hidden populations" or "elusive populations." Hidden populations are defined by two key characteristics. First, they are rare at the "global" population level. Second, there is a lack of comprehensive member registries that could be utilized as a sampling frame to aid estimation of the total magnitude and nature of the hidden population. Sudman et al. discuss the challenges involved in sampling hidden populations.[13] In particular, they note that the use of classical, probability-based sampling techniques to measure hidden populations are inefficient, expensive and subject to very large standard errors. As a consequence of the weakness of classical

[13]Sudman S, M.G. Sirkin, and C.D. Cowan. 1988. "Sampling Rare and Elusive Populations." *Science* 240:991–95.

sampling methods, we use quasi-adaptive sampling methods.[14]

In this paper, we develop a quasi-adaptive sampling strategy, which combines both probability-based random sampling (at the first sampling stage) and adaptive sampling (at the second stage), to measure the random data omissions of lethal violence in Amritsar. Such a design circumvents the inefficiencies of classical, probability-based random sampling for measuring a hidden population by exploiting the strengths of adaptive sampling.[15] Furthermore, this design strategy ensures that we can use classical, probability-sampling estimators that are consistent and unbiased, unlike some of the available adaptive sampling estimators. This design utilizes the post-conflict stability of administrative regions in Punjab and the strong social networks of the families of the victims of enforced disappearances.

This paper explores the dynamics of the adaptive part of the survey sampling design, namely the referral process within villages between village administrators, village elders and families of the victims of lethal violence. We present a series of social-network-analysis metrics and visualizations which show the heterogeneity of the referral process across different dimensions. In particular we study the centrality of different referral points across sampled villages by exploring "betweenness", "closeness", and "degree".[16] These explorations serve to quantitatively characterize the structure of the referral networks. We also study the demographic characteristics (such as age, sex, religion, caste and socioeconomic status) associated with the referral chains within the sampled villages. This analysis of referral dynamics is the basis for determining the coverage rate of the referral-based sampling in villages within which caste and socioeconomic status may influence social knowledge and social relations. We conclude the paper by comparing the strengths and weaknesses of this

[14] For information on major developments on adaptive/snowball sampling see:
Goodman, L.A. 1961. "Snowball sampling." *Annals of Mathematical Statistics* 32:148–70.
Klovdahl, A.S. 1989. *Urban social networks: Some methodological problems and possibilities.* Ablex Publishing Corp. 176–210.
Frank, O., and T. Snijders. 1994. "Estimating the size of hidden populations using snowball sampling." *Journal of Official Statistics* 10:53–67.
Spreen, Marinus and Ronald Zwaagstra.1994. "Personal Network Sampling, outdegree analysis and multilevel analysis: introducing the network concept in studies of hidden populations." *International Sociology* Vol 9, 4:475–491.
Heckathorn, D. D. 1997. "Respondent-driven sampling: A new approach to the study of hidden populations." *Social Problems* 44:174 1999
Weiss, Robert, Jason Wang, Pamina Gorbach, and Steve Shoptaw. 2008. "Respondent-Driven Sampling in an HIV at-risk Population: The analogy with Markov Chain Monte Carlo." Forthcoming in Proceedings of the 2008 Joint Statistical Meetings of the American Statistical Association, al Association, the International Biometric Society (ENAR and WNAR), the Institute of Mathematical Statistics, and the Statistical Society of Canada.

[15] Felix-Medina, M.H. and S.K. Thompson. 2004. "Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations." *Journal of Official Statistics* 20:19–38.

[16] *Betweenness* is the extent to which a referral point lies between other referral points in the social network of the sampled village. This measure takes into account the connectivity of the referral point's neighbors, giving a higher value for referral points which bridge clusters of referrals within the village. The measure reflects the number of people who a person is connecting indirectly through their direct links. *Closeness* is the degree an individual is near all other individuals in a sampled village (directly or indirectly). It reflects the ability of the sampling process to access information about lethal violence deaths which occurred in the sampled village through the social network of the referral points. The *degree* is the count of the number of links to other referral points in the social network.

quasi-adaptive sampling design to estimate conflict-related mortality in war zones.