Multiple Systems Estimation Techniques for Estimating Casualties in Armed Conflicts

Daniel Manrique-Vallier
Social Science Research Institute and
Department of Statistical Science, Duke University.

Megan E. Price
Benetech Human Rights Data Analysis Group.

Anita Gohdes
University of Mannheim and
Benetech Human Rights Data Analysis Group.

1 *Introduction*

During and after armed conflicts, different groups attempt to gather information on the extent to which violence has claimed human lives. Depending on the reasons why the group or institutions record this information, lists of casualties are compiled with more or less detailed data and coverage. For example, humanitarian organizations collect information on cases known to them, death registries try to keep track of the deceased, and press agencies report on victims fallen in battle. Retrospectively, it is often national Truth Commissions and human rights non-profit organizations that document atrocities committed, in order to help recollect the past and provide justice to the victims of conflict.

All of these ’casualty lists’ are prone to incomplete registration, be it for institutional, financial, geographical or political reasons (see Krüger et al.; Landman and Gohdes in this volume). Answers to questions about the real magnitude and characteristics of the conflict cannot be obtained from any single ’found’ data source in a direct way. However, with basic infrastructure and security oftentimes lacking in conflict or post-conflict settings, researchers and practitioners attempting to determine the actual number of casualties that resulted from the conflict commonly find themselves in a situation where they have to rely on these data sources as the basis of their inquiries. Statistical methods that make it possible to draw conclusions about the entire population, based on these incomplete data sources, are thus desirable.

This paper offers an introduction to one such statistical tool, Multiple Systems Estimation (MSE) methods for the estimation of casualties in armed conflicts. These methods provide a way to quantify the probability that a death will be missed, i.e. that no enumeration effort will record it, and therefore a way to estimate the undercount. MSE methods comprise a rather broad family of statistical techniques specifically designed to estimate undercounts in situations like the one described here, where multiple intersecting but incomplete lists are available. They offer an opportunity to take advantage of the multiplicity of data sources often found in armed-conflict situations for drawing conclusions about the underlying extent of the conflict they document.

MSE methods date back into the 19th century, when they were developed to estimate the size of animal populations[1]. For this reason much of the language associated with MSE methods frequently refers to 'captures,' such as 'capture probabilities.' This has been carried over from studies in which animals were captured, tagged, and released (Petersen, 1896). Since then they have been adapted to deal with human populations, in applications that range from census undercount correction (Sekar and Deming, 1949), problems in epidemiology (Hook and Regal, 1999; International Working Group for Disease Monitoring and Forecasting, 1995a, 1995b), and casualty estimation (Ball et al., 2002, 2003), among others (see Jewell et al in this volume for additional references). Due to the development of these methods across different fields, a variety of terminologies has simultaneously developed to describe essentially the same class of methods. Besides MSE, these methods are, among other names, also known as Multiple Recapture Estimation, Multiple-Record Systems Estimation and, in the particular case of two systems, Capture-Recapture and Dual Systems Estimation. While we have aimed for consistency favoring the name MSE (the preferred term for the method applied to human populations), all of them may be used interchangeably.

We begin with an overview of the statistical intuition that underlies MSE methods, as it has some particularities that set it apart from more traditional and well known statistical techniques. Section
2.1 deals with the two-list case, which is then developed into a general multi-list framework in Section 2.2. Section 2.2.1 provides a deeper reflection on two of the classic assumptions of the basic two-list model, and the challenge of interpreting and testing these assumptions in the general case. We address the question of representing unobserved individuals in Section
2.2.2. We occasionally rely on mathematical notation to refer back to concepts that otherwise would require lengthy—and ambiguous—prose to describe. While comfort with mathematical notation and basic probability theory is beneficial, it is not indispensable to understand this chapter. Finally, we present two case studies in Sections 3.1 and
3.2, from Kosovo and Peru, to further illustrate applications, challenges, and successes of MSE techniques. The paper concludes with a discussion of the opportunities and limitations that these methods offer to the field of casualty estimation in armed conflicts .

2 Basics of Multiple Systems Estimation

Any effort to enumerate casualties will likely miss some individuals. Certain geographic areas may be too remote to access or still too violent and unstable for researchers to safely collect data. In some areas wide-sweeping violence may not leave behind any witnesses to tell researchers about what happened, or existing witnesses may choose not to tell their story.

In general, MSE methods attempt to estimate the number of cases that were not included in lists that partially enumerate a closed population. In this context, consider a conflict situation where an unknown number N, of individuals were killed. Now, assume that different 'counting teams', working independently, have already attempted to enumerate them. Each team will have counted a part of the casualty population; some of the individuals will have been counted by more than one team, and some will not have been counted at all.

[1] See Goudie and Goudie (2007) for an account of the origin of these techniques.

If we had access to all these lists, we could try to pool all of them into a single comprehensive list. Since some individuals will have been counted more than once and some left out, it is likely that the combination of all lists will also be incomplete. If the teams recorded some identifying information on the individuals, we could remove the duplicates by comparing the datasets and noting who and how many of those individuals were included on more than one list. As we will see, this inclusion, or capture pattern, i.e. which lists included and which missed an individual, is the crucial piece for the MSE calculations. For now, we can safely state that the de-duplication (i.e., identification and removal of duplicates) of individuals who were included on one or more lists allows us to compute a lower bound on the total size of the population of interest, assuming that de-duplication efforts were successful. The question that remains is: 'How many individuals were not counted by any of the teams? '

Table 1 shows an example of such a de-duplication and matching of different lists into one dataset.[2] Every listed individual now only appears in one row. The last three columns indicate which list recorded the case. This example shows how binary information indicating 'included' (1) or 'not included' (0) in list A, B or C creates an 'inclusion pattern'. The diagram in Figure 1 presents the same information in a graphical form. Note how each inclusion pattern unequivocally refers to a location in the Venn diagram.[3] Again, each individual appears just once in the diagram. Inclusion patterns represent the link between the concepts of “unique individual" and “records on a list". Also note that since real-life de-duplicated lists can only be composed of individuals that have been observed at least once, no individual with a capture pattern consisting of only 0s can be listed.

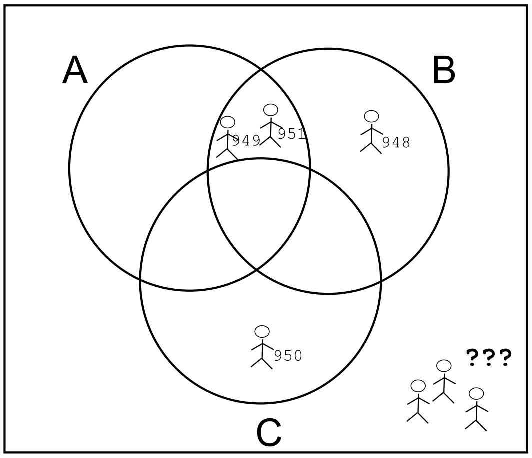
Table 1: Example: ’clipping’ of a matched and de-duplicated list of casualties

ID	**Sex**	**Age**	**Location**	**Date**	**List A**	**List B**	**List C**
..	..	..	..	..	..	..	..
948	M	32	south-west	1999	0	1	0
949	M	45	West	1995	1	1	0
950	F	30	south-west	1990	0	0	1
951	M	?	West	1991	1	1	0
..	..	..	..	..	..	..	..

[2] This is a fictitious table that was created for exemplary purposes. In real-life cases, the information existing for each individual varies profoundly. The co-variables included here (age, sex, location) merely describe some of the most frequently encountered victim/incidence characteristics.

[3] Since Table 1 only shows four individuals, not all inclusion patterns are populated in the diagram. A complete list could potentially include individuals captured in all locations of the diagram, except outside all circles.

Figure 1: Schematic representation of de-duplicated list in Table 1.



MSE techniques attempt to provide an answer to the question of who was *not* included in any of the lists by looking at the inclusion pattern of each individual that *was* included. Intuitively, it is easy to see how inclusion patterns can help us learn the size of the population. Let us assume that one of the teams in our example did a particularly thorough job, so its list is already close to a full enumeration. Then, any additional comparable list by another team is likely to have a considerable number of individual records in common with the first list. Conversely, if the team only covered a very small part of the population, any additional similar list can only be expected to share a few individual records with it, if any at all. Note that if the list were a full enumeration, any additional list could only be a subset of it. Thus, subsequent lists could only enumerate individuals already recorded.

If we formalize the relationships between the probabilities of these inclusion patterns, then we can—with the help of some additional assumptions—estimate the probability distribution of multiple inclusions in lists, and thus ultimately estimate the probability of an individual not being included in *any* of the lists. The following section illustrates this procedure in the two-list case.

2.1 Classic Two-Systems Estimation

Consider the number of casualties in a conflict to be N—more generally, N is the unknown size of a finite population of interest. We assume that we have available two incomplete lists, A and B, that enumerate some of the individuals in the population. If we consider an arbitrary individual, he or she will necessarily fall into one of the following four cases: the individual is included in the first list, but not in the second (this case will be represented by O_{10} for the remainder of this chapter), *or*, in the second list, but not in the first (O_{01}), *or* in both of the lists (O_{11}). Finally, the individual might not be included in either list (O_{00}). We call these cases *inclusion patterns*.

It is important to note that although every single individual in the population must fall into one (and only one) of these categories, we can only have information on the individuals that were included in at least one of the lists. We thus only observe individuals that correspond to the first three inclusion patterns, O_{10}, O_{01} or O_{11}. Pattern O_{00} is unobservable by definition.

Now consider an arbitrary individual, i, from the population.[4] We can assign a probability to the event that any one individual (i) falls into each of the categories described above. For example, we can consider the probability that individual i is on list A but not B. We denote this probability $p_{10}^{i} = P$(Individual i in category O_{10}). We then associate each individual with an array (a 4-dimensional probability vector) detailing their probability of falling into each of the four inclusion patterns. This vector, $(p_{00}^{i}, p_{01}^{i}, p_{10}^{i}, p_{11}^{i})$, fully describes the probabilities of individual i being documented (or not) according to each possible inclusion pattern.

As is common when using statistical models to describe real-world situations, some assumptions are necessary to estimate the quantities of interest. As the simplest version of this method, the two-list situation requires some strong—and untestable—assumptions. We will revisit them in the next section in the context of the existence of multiple lists.

While not the only possibility, the usual assumptions in the two-list case are

1. *Closed System*: The lists refer to a closed system: N must refer to the same population in each dataset.[5]
2. *Homogeneity*: For each list, every individual must have the same probability of being included, or captured.[6]
3. *Independence*: The probability that an individual is (or is not) included on list A is not affected by being included (or not) on list B, and vice versa.

Intuitively, the homogeneity assumption refers to the fact that no individual should be intrinsically ”more listable" than the others due to individual traits. Independence, requires lists that have been created without influencing one another (i.e., inclusion on one list does not affect the likelihood of inclusion on another list). Going back to the example, these conditions require that (1) no victim has any distinctive characteristic (e.g. age) that makes her more likely than the others to be in any list, and (2) each team worked without receiving information from any other team. Both the homogeneity and the independence assumptions pose demands to the data that are unlikely to be fulfilled—even approximately—in the casualty estimation context.[7] However, as is the case with other statistical methods, there are means by which we can detect departures from these assumptions and alternatively account for them. We will discuss this in more detail in Section 2.2.1.

[4] i could equal any number between 1 and N, indicating the first, second, etc. individual in the population; in mathematical notation we write $i = 1,.,N$.

[5] For the case discussed here, this assumption is generally met: individuals that were killed in armed conflict add to the number of casualties and cannot ’leave’ this population.

[6] This definition of homogeneity is simple and appropriate in this two-list context. In more complex settings, however, it can be too restrictive to be really useful. A more general definition has to be stated with respect to a model: the recording of each individual can be well described by the same model.

[7] A fourth assumption usually made is *perfect matching*: Each of the included individuals must be perfectly categorized to be either only included in system A, only included in system B, or included in both system A and B (see Lum et al., 2010: 3).

To estimate the unknown population size, the above assumptions must be mathematically formalized and combined with some statistical technique that allows the calculation of such numbers (an “estimator", in statistical terminology). To illustrate this process, the rest of this section shows the derivation of one specific two-list estimator, the Petersen Estimator, which is the best known of the two-list estimators (Petersen, 1896; Bishop et al., 1975; International Working Group for Disease Monitoring and Forecasting, 1995a). Readers without elementary knowledge of probability and statistics can safely skip the rest of this section.

Elementary probability calculations show that the above stated assumptions have two immediate consequences. First, the same set of pattern probabilities apply to every individual—that is, we can drop the index i in the probability vector, making $(p^i_{00}, p^i_{01}, p^i_{10}, p^i_{11}) = (p_{00}, p_{01}, p_{10}, p_{11})$. Second, these probabilities must satisfy the condition $p_{01}p_{10} / p_{11}p_{00} = 1$. These two conditions, together with the fact that the population is finite, define a *model*. Standard statistical techniques can then be used to estimate the parameters of the model, including the population size.

One simple way of obtaining the Petersen estimator is the following. Under the proposed model, the expected values of the number of individuals in each capture category is $m_{ab} = E[n_{ab}] = N \cdot p_{ab}$ ($a = 0,1; b = 0,1$),[8] where n_{ab} are the actual (observed and unobserved) counts, N is the true (unknown) population size, and p_{ab} is the general form of the probabilities defined above. Then, the condition above that $p_{01}p_{10} / p_{11}p_{00} = 1$ can be rewritten as

$$m_{00} = \frac{m_{01}m_{10}}{m_{11}}.$$

We can use the observed counts as estimators for their expected values, $\hat{m}_{01} = n_{01}$, $\hat{m}_{10} = n_{10}$ and $\hat{m}_{11} = n_{11}$ and use them to estimate the undercount:

$$\hat{n}_{00} = \frac{\hat{m}_{01}\hat{m}_{10}}{\hat{m}_{11}} = \frac{n_{01}n_{10}}{n_{11}}.$$

The Petersen estimator, or Dual-System estimator, is the best known of the two-system population size estimators and is often used to illustrate the core Capture-Recapture ideas. It is usually motivated as an extrapolation of capture ratios through a process that implicitly requires the independence and homogeneity assumptions (see e.g. International Working Group for Disease Monitoring and Forecasting (1995a)). It can also be shown to be a conditional Maximum Likelihood estimate for the two-variable log-linear quasi independence model under multinomial sampling (Fienberg, 1972; Bishop et al., 1975).

2.2 Multiple Systems Estimation (MSE): The general case

MSE techniques draw inferences about the size of the population of interest from the classification of observed (listed) individuals into multiple inclusion patterns. In the case of 2 lists, there are only 4 possible classifications or ’partitions’, of which one, O_{00}, is unobservable.

[8] The subscripts “ab” are a shortcut to refer to any of the four inclusion patterns: 00, 01, 10, 11.

With only three partitions, the analyst has very limited information on which to base her estimations. This is why strong assumptions like those stated in the previous section are unavoidable.

More than two lists, in contrast, provide a much richer set of information from which we can draw inferences. Every additional list has the effect of doubling the number of capture patterns. We therefore gain *observable* information at an exponential rate, while the *unobservable* pattern (individuals not included in any list) remains fixed at one. For instance, 3 lists produce $2^3 - 1 = 7$ observable partitions (O_{001}, O_{010}, O_{011}, O_{100}, O_{101}, O_{110} and O_{111}), while 15 lists (see e.g. Lum et al. (2010)), produce $2^{15} - 1 = 32,767$ observable partitions. In both cases, we have only one unobservable pattern (O_{000}, and $O_{000000000000000}$, i.e., those that were missed by all three and all 15 lists respectively).

An increasing number of partitions offer more information. This enables the use of more sophisticated models that rely on weaker assumptions than the ones needed for the two-system model. Unfortunately, the modeling of that increasing number of inclusion patterns also requires increasingly sophisticated statistical techniques.

Several families of models have been proposed to deal with the multiple-list situation, each of which specifies different sets of assumptions about the population and the list creation processes. For example, log-linear models (Fienberg, 1972; Bishop et al., 1975) take advantage of multiple lists to account for the effect of list interdependence, by modeling dependence patterns explicitly. Other techniques, that respond to different assumptions, include Bayesian versions of standard approaches (George and Robert, 1992), discrete mixture models (Basu and Ebrahimi, 2001), Grade of Membership models (Manrique-Vallier and Fienberg, 2008), and Rasch models (Fienberg et al., 1999), to name just a few.

A generalization of the framework from the two-list case to the multi-list case is straightforward: if we have a number, J, of lists, then any individual in the population can be classified into one and only one inclusion pattern $O_{X_1 X_2 X_3 . X_J}$, where $X_j = 1$ indicates presence in list j and $X_j = 0$, absence. For each individual, i, we can then model the probability distribution over the inclusion patterns. The information about the $2^J - 1$ observable patterns is then used to estimate the probability of the single unobserved pattern, $O_{000.000}$. The specific way this information is used to estimate the probability of the single unobserved pattern varies across the general class of MSE methods.

In addition to allowing weaker—and hopefully, more realistic—assumptions, the use of more than two lists also provides the means to *test* the plausibility of some of these assumptions. This feature is crucial since successful estimation of the unobserved pattern depends on accurately modeling the inclusion-exclusion structure present in the population.

Analysts must examine the data to determine appropriate methods and reasonable assumptions. This presents an additional challenge since this process often results in several plausible models. When a set of inclusion patterns appears to be adequately described by more than one model, a ’best’ model may often be chosen (an example of this can be found in the Kosovo case study in

Section 3.1). Another possibility is to produce estimates based on an average of multiple models (see Lum et al. 2010).[9]

It is important to keep in mind that to test the plausibility of any set of assumptions we can only rely on observed data (Fienberg, 1972; Bishop et al., 1975; International Working Group for Disease Monitoring and Forecasting, 1995a). This means that no matter how well our models describe the observable patterns, the only way of estimating the undercount is through an untestable assumption: the inclusion pattern of unobserved individuals can be described by the same model that describes the inclusion pattern of those we get to observe. This is an inescapable limitation that must be taken seriously. We will elaborate on this idea in Section 2.2.2.

2.2.1 A Closer Look at Heterogeneity and Dependence

Homogeneity and independence are intuitive assumptions that are sometimes reasonable in applications such as very simple animal capture-recapture experiments. However, in more sophisticated settings such as casualty estimation we cannot expect them to hold.

In general, victims of violence are likely to be a heterogeneous group. Lists documenting them are unlikely to be independent. People have different social visibility due to networks, geographic location and other traits. These characteristics, which often influence the outcome to some extent, are called covariates in statistics[10]. Different documentation projects have different objectives and may have different propensities to record victims with particular characteristics/covariates. Projects will sometimes collaborate or share information with each other, directly inducing dependence between lists.[11]

In cases where covariate information on the particularities of the victim and context is available, stratification can be used as a means to reduce the effect of heterogeneity. In short, stratification works by partitioning the population into separate and relatively more homogeneous subsets, where the modeling assumptions can be expected to hold better.[12] Estimates are then calculated within each stratum, using the MSE method of choice. For example, if we think that the place of death can strongly influence capture probabilities (i.e., that it is an important covariate), as is likely to be true in many cases, we can divide the combined sample into geographic subgroups. More precisely, introducing a stratification scheme is equivalent to assuming that, if two individuals, i and j, belong to the same stratum (e.g. the same geographic region), then the probability of each pattern of inclusion is the same for both or, more formally, that $\Pr(\text{individual i in } O_X) = \Pr(\text{individual j in } O_X)$.

[9] The “model selection" problem is a major subject in statistical methodology, common to a wide range of applications. Although it is a crucial problem in MSE applications, due to space constraints we will not elaborate it further.

[10] Examples of covariates include sex, age, location, and date from Table 1 above.

[11] Although this may seem a pervasive problem, in actual applications, the situation is not as dire. For instance, in the Peruvian study (Ball et al., 2003), shared information was clearly labeled and could be separated without problem. As we will see when discussing the applications (Sections 3.1 and 3.2), thorough knowledge of the data is essential.

[12] In other words, where the inclusion patterns can be described reasonably well by a particular model.

When relevant covariate information is not available, the only visible effect of heterogeneity is the emergence of dependences between lists. Take for example the case of a conflict with two main perpetrators and three available lists documenting casualties. Assume that no list registered the perpetrators of the killings. If two lists were more prone to register people that were victims of a particular perpetrator, while the other one proceeded in a more balanced way, then—assuming for now that no other sources of heterogeneity are as important—the observable effect will be the emergence of (positive) dependence between the first two lists, while the third will remain relatively independent. In cases like these, with no covariate information that could fully explain it, the only way we can learn about the heterogeneity structure is through these induced dependencies.

This observable dependence between lists can, however provide enough information to successfully account for the effects of the heterogeneity. Furthermore, sometimes accounting for the induced dependence amounts to directly controlling for the effects of the heterogeneity. For example, some general patterns of heterogeneity can be successfully represented or approximated by interactions between lists with the aid of log-linear models.[13]

2.2.2 Representing Unobserved Individuals

The ultimate goal of MSE techniques is to estimate undercounts. This requires estimating probabilities of non-inclusion in lists based on information about non-inclusion contained in the inclusion patterns of individuals observed at least once. This seems to be a paradox. However, we must note that in a combined J-list sample, many individuals are likely to have been missed by one or more lists. This means that the observable inclusion patterns contain a great deal of information about individuals who are not included on a given list. For instance, if we had six lists, we would have $2^6 - 1 = 63$ observable inclusion patterns, from which 62 of them describe ways of *not being on lists*.[14] Not being in *all* lists simultaneously is just one more way of not being in lists.

The arguably most basic assumption in MSE is that the non-inclusion of the fully unobserved individuals (those not included in any list) can be represented by the same model that represents the inclusion (and non-inclusion) of those we get to observe in at least one list. This is a strong and untestable condition. However, we argue that it is far less demanding than it initially may seem to be.

To better understand this requirement, let us examine its violation. Assuming that non-observed individuals (those with inclusion pattern O_{00}) differ substantially from observed ones (those with all the other $2^J - 1$ patterns), amounts to assuming that the event of *simultaneously* not appearing

[13] In general, heterogeneity that affect groups of lists, but not every list, can be directly accounted for as interactions between lists in log-linear models. Other, more sophisticated, patterns that simultaneously affect all lists can also have estimable log-linear representations (see e.g. Darroch et al. (1993) for more details).

[14] For instance, the observable pattern O_{000100} in a 6-list situation not only gives us information about being in list 4, but also about *not being* in lists 1, 2, 3, 5 and 6.

in all *those particular lists* is somehow an intrinsic attribute of those individuals. This means, for example, that being missed by five lists but not by a sixth, in a six-list case, is qualitatively different from being missed by those six lists; and that, if we added another seventh list, being missed by all previous six lists but not by the new one is also substantially different from being missed by all seven. Except for a few situations, it appears to us more difficult to reject this requirement than to accept it.

One of these problematic situations is the case of erroneously assumed coverage. This would, for instance, occur if we had a situation where our lists were specifically designed so that they only reported events on a particular region—and ignored any other report—but we assumed we obtained estimates for a wider region.[15] This would lead to the existence of two classes of subjects: those who are, at least in principle, listable and those who are not. In the language of capture probabilities, the first group has a positive probability of inclusion, while for the second, that probability is exactly zero. Any individual who has an intrinsic attribute which *causes* him or her to be unobservable by any list has, by definition, a capture probability of zero. Individuals with capture probabilities of zero cannot be represented by *any* data collection mechanism. In contrast, individuals with non-zero capture probabilities, who just happen to be un-observed by every list (not due to some intrinsic attribute but rather by chance) are likely to be represented by other, observed individuals (who are also missed by some subset of the lists).

Related to concerns regarding how best to represent unobserved individuals is the belief that MSE techniques can only produce valid inferences when based on lists that are random samples from the target population (e.g. Jewell et al. in this volume). This point is of particular importance, since arguably most lists of casualties that can be found in practice are unable to meet such a standard. Fortunately, except in some truly problematic cases, this belief is not correct. The key is that the only information that matters for MSE estimation is the relationship between lists, and not the lists' composition. It can be shown that the only requirement is that the collection of *inclusion patterns* is representative of the relationships between the lists, not for each list to represent the underlying population. As an extreme example, consider an organization that collected information giving preference to individuals that lived close to their headquarters, and another one that did so giving preference to older people over younger people. This is an example where none of the partial samples is representative of the characteristics of the target population and where the "homogeneity" assumption, understood as the same probability of inclusion within each list, is clearly violated. However, if we assume, as seems reasonable, that age is uncorrelated with how close to the organization headquarters the victim lived, we can show that even the simple 2-list Petersen estimator is valid. The example is, of course, artificial,[16] but serves well to illustrate the point. As long as we can approximately model the characteristics of the resulting aggregated pattern—independence in the example—the internal characteristics of the lists turn out to be irrelevant.[17]

[15] This situation is analogous to a survey with a sampling frame that does not cover the totality of the target population.

[16] And, depending on the specific situation at hand, it could also be argued that the age distribution could be correlated with geographic location.

[17] This is one reason why we believe that understanding homogeneity as equal probability of being in each list can be misleading in this discussion.

Real complications may arise, however, if the underlying data structure is such that a “wrong model" can successfully account for the observable part of the inclusion patterns, but not for the full-exclusion one ($O_{0.0}$), and we are led to choose it over other more appropriate ones. In theory, some heterogeneity patterns could lead to such a situation. As an exteme example, consider lists from distinct age groups with little or no overlap. This scenario could plausibly result in most MSE procedures considerably overestimating the actual total counts[18]. This may be an example of a heterogeneity pattern in the population that could plausibly induce a pattern of dependences between lists that would be consistent with an identifiable model for *all* the ($2^J -1$) observable patterns, but somehow not for the unobserved category $O_{000.000}$. In this general case, the risk is that we would be led to accept and rely on a model that does not correctly represent the non-inclusion of the unobserved individuals, and therefore poses a risk for producing biased estimates. However, it is not well understood which plausible patterns of heterogeneity can induce such outcomes and more research is needed on this topic. In the authors’ experience, populations where the same source of heterogeneity strongly affects all lists simultaneously can sometimes generate observable data with these characteristics.

However, while acknowledging such extreme situations as real limitations, we should bear in mind that, although plausible, they are unlikely to be completely unknown to researchers. In the example, for instance, a simple tabulation of the counts broken down by age would immediately reveal this special heterogeneity. This would allow researchers to account for it, for example, by stratifying by age group. Moreover, if researchers were able to secure any other list that was less sensitive to that particular source of heterogeneity—even if it was extremely sensitive to any other uncorrelated source—such a list could potentially provide enough information to overcome the problem through direct modeling of the dependence patterns—in theory, even without stratification.

As will be shown in more detail for the case of Peru (below and in Chapter XX by Landman and Gohdes), the researchers’ knowledge of the situation is crucial here. In Peru, two of the lists used for the three-system estimation of killed and disappeared people across the 20 year conflict gravely underreported acts committed by the rebel group Sendero Luminoso. With the help of the third, largest, list provided by the Truth and Reconciliation Commission, the research team was able to use the information provided by all three lists to account for this ‘perpetrator heterogeneity’ and estimate credible levels of casualties committed by both the state and insurgent group.

3 Case Studies

3.1 Kosovo

The tragic events that unfolded in Kosovo between March and June 1999 present a case where the application of MSE methods significantly improved the knowledge of the extent of violence that was exercised against Kosovars. The research was conducted by Ball et al. (2002), who used four data sources to conduct MSE analyses of the patterns of refugee flow and killings in Kosovo for said time period. The data sources comprised of interviews conducted by the American Bar

[18] see also Bishop et al. (1975) Ch. 6 for a discussion about essentially the same extreme situation

Association Central and East European Law Initiative (ABA/CEELI), interviews conducted by Human Rights Watch (HRW), interviews conducted by the Organization for Security and Cooperation in Europe (OSCE), and records of exhumations conducted on behalf of the International Criminal Tribunal for the Former Yugoslavia (ICTY). It is important to note that while excellent sources of information, and exemplary data collection efforts, none of these sources are uniformly representative of the entire underlying population of individuals killed in Kosovo between March and June 1999. For example, ABA/CEELI conducted interviews in Macedonia (among other locations) relying on referrals from humanitarian organizations, word of mouth, advertising in local newspapers, and moving tent by tent through refugee camps (Ball et al. 2002). These were reasonable methods to locate individuals with crucial information on killings in Kosovo. Similar methods are employed in a variety of conflict and post-conflict regions and are often necessary to obtain information where there is no hope of a complete census or a sampling frame from which to build a random sample. But it would be unreasonable to assume that these methods result in a representative sample. Therefore we must rely on statistical methods, such as MSE, which are suitable for calculating population-level estimates from available data.

A total of 4,400 individual victims were identified across the four data sources. Many of these were listed in more than one source. Based on the inclusion patterns of these 4,400 identified individuals, an estimated total of 10,356 victims was calculated (95% confidence interval: [9,002, 12,122]). This number was surprising, as it implies that more victims went undocumented—namely, 5,956—than were jointly recorded in the four available data sources.

Prior to building and selecting the models necessary to calculate these estimates, exploratory data analysis was conducted to evaluate the plausibility of the classic MSE assumptions outlined in Section
2.1. This analysis indicated potential sources of heterogeneity, and led researchers to stratify MSE calculations by space (geographic region) and time. Two-systems estimates (as described in Section
2.1) were also calculated for each pair of sources to identify possible dependences between lists; numerous positive dependences were identified. An extension of this method, using hierarchical log-linear models, was used to examine the relationships between three of the data sources at a time.[19] These results indicated that the pairwise dependencies, identified by the two-systems estimates, were likely well modeled by including two-way interaction terms. Such direct analysis of the data patterns (over time and space) and exploratory two- and three-systems MSE calculations indicated the need for careful stratification and complex modeling to account for the intricate heterogeneity and dependence structure. This procedure illustrates how, as mentioned in Section 2.2, assumptions must be checked so that the most appropriate MSE method is chosen for a given situation.

There were many possible complex models to describe the observed inclusion patterns. Traditional model selection techniques (i.e., best fit statistics) were used to identify the model used to calculate 10,356 killings with a 95% confidence interval of [9,002, 12,122][20].

[19] See Bishop et al. (1975) for details on hierarchical log-linear models.

[20] See Ball et al. 2002 for full analytical methods and results, including model results calculated for space and time strata.

It is important to note a few things. First, without MSE calculations, at that time, we would have lacked an estimate of the nearly 6,000 undocumented killings in Kosovo between March and June 1999. Second, if we relied solely on the observable, available data from the four sources, we would be unable to choose between the contradictory conclusions regarding the pattern of violence over time and space provided by each data source (see Kruger et al. in this volume for an example comparing geographic regions). And lastly, thanks to the work of the the Humanitarian Law Center[21], this estimate has recently been largely corroborated. Their attempt to generate an exhaustive list of victims has documented 9,030 murders and 1,200 missing from this time period.

3.2 Peru

Between 1980 and 2000, Peru witnessed a bloody armed internal conflict that was primarily carried out between the state forces and the insurgent Communist Party of Peru-Sendero Luminoso (PCP-SLU) movement. This fighting received only limited attention from the international community.

Prior to the establishment of the Truth and Reconciliation Commission (Comisión de la Verdad y Reconciliaci´ón, CVR), conventional wisdom had situated the number of victims claimed by the conflict to be approximately 25,000. Using three different lists enumerating deaths and disappearances in Peru between 1980 and 2000, researchers at the CVR and the American Association for the Advancement of Science (AAAS) were able to conduct MSE analyses, which revealed that the total number of victims was in the vicinity of 70,000 (see Ball et al. (2003)).

The CVR collected reports documenting deaths and disappearances of approximately 24,000 people, of which 18,397 could be identified sufficiently to be matched with two further lists.[22] Importantly, the addition of the second and third list only amounted to approximately 6,000 more documented cases. This exemplifies the fact that the size of the additional lists used for MSE is less important than the pattern of overlap between the lists. Despite almost 75% of all reported cases having been recorded by the CVR, the two lists added for MSE delivered the missing information that was required to calculate an estimate.

Local area experts expected incidences to be reported with varying probability in different regions (i.e., violations of the homogeneity assumption), so data were first stratified by geographic location of the death or disappearance. Depending on the amount of information available for each region, the data were stratified by departments, provinces and—where possible—even districts. For example, in the department of Ayacucho, the data could be stratified right down to the district-level, as all three lists had recorded a disproportionate number of incidences in this department. Besides the assumption that different regions would produce heterogeneous capture probabilities, it was assumed that the perpetrator by whom an individual was killed or disappeared would have an influence on whether the incident was reported or not. As demonstrated in Landman and Gohdes (this volume), the three lists offered a very different

[21] www.hlc-rdc.org/index.php?lid=en&show=kosovo&action=search&str_stanje=1

[22] See chapter XX, page X by Landman and Gohdes for further details on the other lists.

answer to the question of which perpetrator should be held responsible for the majority of atrocities committed. Of all cases attributed to PCP-SLU, 80% were exclusively recorded by the CVR database. For each geographical stratum, the researchers thus attempted to calculate individual estimates for the different perpetrators.

The log-linear models used for the estimation allowed for the modeling of interactions between the different lists, enabling the researchers to select the best fitting model out of seven possible models for each stratum.[23] The models that were selected with the greatest frequency were those in which there was at least one interaction between the two smaller lists, with the Truth Commission's list being independent. Accordingly, not once was a model selected that assumed an interaction between the Truth Commission's data and *both* of the other lists.

With the help of this method, it could not only be revealed that the majority of atrocities were actually committed by the PCP-SLU (31,331, 95% confidence interval:[24,823; 37,840]) and not by the state (20,458, 95% ci:[17,023; 23,893]), as had been assumed by many human rights groups (the data they had collected supported their claim), but that the conflict had primarily affected the rural, poor areas of Peru, furthest away from the urban agglomeration of Lima. Of the estimated 69,280 (95% ci:[61,007; 77,552]), deaths and disappearances, 26,259 could be attributed to the region of Ayacucho alone, which is located in the south-central Andes.

It is important to take into consideration in this case that, despite its horrifying magnitude, the total death toll directly attributable to the conflict represents a minute fraction of the total Peruvian population at the time—approximately 27 million in 1993. This means that other, more traditional, techniques of assessing conflict-related mortality rates, such as survey sampling, might have been unfeasible due to the low prevalence of the effects we would need to detect (see (BallSilva2008)).

4 Conclusion

Multiple Systems Estimation methods encompass a broad variety of techniques that offer promising solutions to some of the challenges that researchers and practitioners face in casualty estimation in conflict and post-conflict situations. In the demanding circumstances of wartorn regions, obtaining reliable estimates of killed and disappeared persons poses a difficult, and sometimes seemingly impossible task. Documentation of violent events is often rare and biased, and the lack of infrastructure and resources presents a challenging situation for the conduct of surveys. MSE techniques offer a way to use existing, sometimes unrepresentative, information on casualties to produce a less biased and more complete number of atrocities committed. While certainly not a 'foolproof' class of methods, our case studies of Kosovo and Peru illustrate that in certain situations, it can considerably improve our knowledge of conflict trajectories.

In this paper, we have attempted to present the general intuition that lies behind MSE methods. Instead of focusing on one particular technique, we introduced a general framework for MSE analysis in order to explore more deeply the assumptions and the subtle interplay between

[23] The seven models included one model that assumes independence between the lists, three that assume one interaction (i.e. two lists are dependent), and three that assumed two interactions (i.e. one list interacts with the other two lists, but the other two lists are independent of each other).

heterogeneity and dependence. As with any other statistical technique, the most basic forms of MSE rely on strong assumptions which, in real life applications such as casualty estimation, are almost never met. Fortunately, the availability of more than two lists makes it possible to apply methods that replace those assumptions with more appropriate ones. Furthermore, multiple lists make it possible to test for violations of many of the assumptions implied by different models.

As illustrated in our case studies, researchers must examine the data and choose appropriate methods. This requires both statistical and local area expertise, as contextual knowledge about the data can guide researchers in terms of which assumptions are likely to be violated and which tests to check. At the same time, the appropriate method can not only reveal which assumptions are *not* met, it can also help account for these violations.

The Kosovarian and Peruvian cases presented here exemplify the significance of data analysis methods that correct for recording biases in casualty estimation. In both situations, the number of casualties additionally ‘uncovered’ through MSE were larger than the number of killed and disappeared people recorded by Truth Commissions, NGOs and international organizations together. The political relevance of such results is evident, and illustrates the importance of comprehending the assumptions and possible pitfalls of such estimation techniques.

The many advantages of MSE come at a price of high technical complexity. Understanding the assumptions and limitations, and correctly applying the methods require considerable statistical expertise. MSE techniques differ substantially from many other better known statistical techniques and this can easily induce misunderstandings, even in technically sophisticated audiences. A common source of misunderstandings is the extrapolation of assumptions and limitations from the two-list case to the multilist case, e.g. the need of strict homogeneity and independence assumptions. Other common misunderstandings, such as believing that MSE require representative samples from the population, sometimes arise from faulty analogies with more standard statistical techniques.

The complexity of MSE methods constitutes a communication challenge that risks the clear dissemination and discussion of results. Opaque presentations, coupled with the potential misunderstanding of the methods’ assumptions and subsequent analysis decisions, run the risk of undermining the credibility of otherwise sound conclusions. This can be particularly problematic in a politically charged debate, as can almost always be found in the casualty-estimation context.

Since their first development for the estimation of wildlife populations over a century ago, recapture methods have significantly progressed. The recent evolvement of techniques that address ’real-life’ problems, such as the estimation of casualties, presents an important step in the continuing development of this class of methods. Although they do present challenges and limitations, we believe that MSE methods are a versatile tool that enables the principled use of data frequently found in practice, and as such should be considered part of a standard ‘casualty-estimation toolbox’.

References

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," AAAS. Report to the Peruvian Truth and Reconciliation Commission (CVR). Also published as Anexo 2 (Anexo Estadístico) of CVR Report.

Ball, P., Betts, W., Scheuren, F., Dudukovich, J., and Asher, J. (2002), "Killings and Refugee Flow in Kosovo March-June 1999," *Report to the Intl. Criminal Tribunal for the former Yugoslavia*.

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

Bishop, Y., Fienberg, S., and Holland, P. (1975)*, Discrete Multivariate Analysis: Theory and Practice,* Cambridge, MA: MIT Press, reprinted in 2007 by Springer-Verlag, New York.

Darroch, J., Fienberg, S., Glonek, G., and Junker, B. (1993), "A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability," *Journal of the American Statistical Assocation*, **88**, 1137–1148.

Fienberg, S. (1972), "The Multiple recapture census for closed populations and incomplete 2^k contingency tables," *Biometrika*, **59**, 591–603.

Fienberg, S., Johnson, M., and Junker, B. (1999), "Classical multilevel and Bayesian approaches to population size estimation using multiple lists," *Journal of the Royal Statistical Society. Series A,* **162***,* 383–406.

George, E. and Robert, C. (1992), "Capture-Recapture Estimation Via Gibbs Sampling," *Biometrika*, 677–683.

Goudie, I. and Goudie, M. (2007), "Who captures the marks for the Petersen estimator? " *Journal of the Royal Statistical Society. Series A*, **170**, 825–839.

Hook, E. and Regal, R. (1999), "Recommendations for presentation and evaluation of capture-recapture estimates in epidemiology." *Journal of clinical epidemiology,* 52, 917.

International Working Group for Disease Monitoring and Forecasting (1995a), "Capture-recapture and multiple-record systems estimation I: History and theoretical development," *American Journal of Epidemiology*, **142**, 1047–1058.

— (1995b), "Capture-recapture and multiple-record systems estimation II: Applications in human diseases," *American Journal of Epidemiology,* **142**, 1059–1068.

Lum, K., Price, M., Guberek, T., and Ball, P. (2010), "Measuring Elusive Populations with Bayesian Model Averaging for Multiple Systems Estimation: A Case Study on Lethal Violations in Casanare, 1998-2007," *Statistics, Politics, and Policy*, 1, 2.

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

Petersen, C. (1896), "The yearly immigration of young plaice into the Limfjord from the German Sea," *Report of the Danish Biological Station*, **6**, 1–48.

Sekar, C. and Deming, W. (1949), "On a method of estimating birth and death rates and the extent of registration," *Journal of the American Statistical Association*, 44, 101–115..

.