# An exploration of multiple systems estimation for empirical research with conflict-related deaths

Jule Krüger and Kristian Lum*

**Abstract**

Conflict and/or violence are often measured by counting the casualties in a specific area and period. An unbiased account of conflict lethality is a prerequisite for testing key hypotheses with empirical data. Visibility and security issues however inhibit the collection of complete enumerations or random samples required for reliable statistical inference. In this paper, we demonstrate that the statistical method of *multiple systems estimation* (MSE) can be used to estimate the total number of conflict-related fatalities during an episode of lethal violence. We introduce this technique and apply it to the case of lethal violence in Kosovo (March-June 1999). We estimate the total number of victims from three incomplete lists of casualties and missing persons. We compare our estimates to the observed data, as well as a recently completed census of all war victims during our period of observation. With this - to our knowledge - first test of multiple systems estimation in the context of conflict casualties, we show that MSE addresses problems of incomplete and biased registration common in data on observed lethal violence.

## 1 Introduction

Empirical research on conflict and violence is currently flourishing thanks to an ever-increasing availability of many types of data sources from which information on observed violent events and/or conflict-related casualties can be obtained (cf. the widely used data by Eck and Hultman 2007; Lacina and Gleditsch 2005; Raleigh et al. 2010; Sundberg and Melander 2013). The new influx of 'microlevel data' has significantly expanded the range of research topics. For example, scholars examine whether the level of civilian victimization

*Paper prepared for presentation at the Visions in Methodology Conference at the University of Kentucky, May 13-16, 2015.

influences bargaining between the regime and insurgents during a civil war (Wood and Kathman 2014), whether different forms of third-party intervention affect civilian death tolls (DeMeritt 2015; Hultman et al. 2013), whether actors engaged in armed conflict are punished for inflicting collateral damage on civilians (Condra and Shapiro 2012), or whether armed actors are more likely to abuse the civilian co-ethnics of their enemy (Fjelde and Hultman 2014).

What these exemplary research questions have in common is that, to be answered accurately, they require knowledge of the 'true number of civilian casualties' that occurred within the area and period under study. For the purpose of this paper, we refer to this number as '*the ground truth*,' i.e, the total number of individuals who are killed or disappeared due to the use of armed force within a given spatial area and period.[1]

The above-mentioned research projects further have in common that, in measuring violence, data on civilian death tolls was obtained via *direct observation* with information being primarily coded from media reports (cf. Ulfelder and Schrodt 2009; Sundberg and Melander 2013; Iraq Body Count n.d.). In the statistical sense, data from observable information is a convenience sample of 'the ground truth' as it neither constitutes a random sample, nor a full enumeration of the full population of civilian casualties that occurred.

Our empirical exploration in this paper is motivated by our concerns regarding current empirical practice in using conflict-related deaths to answer research questions such as the above. While scholars require knowledge of the ground truth to provide valid and reliable answers for the theoretical and empirical puzzles of interest to the field, research designs yet rely on indicators of *observed* conflict-related casualties to measure theoretical concepts such as 'conflict-years' and 'violent events' (Gleditsch et al. 2002; Raleigh et al. 2010, 2012; Sundberg et al. 2012; Sundberg and Melander 2013). We posit that the existing observed data are being treated *as if* they represent ground truth, which risks to produce erroneous findings.

With our paper, we join an emerging debate on the representativeness of observed casualties when used for conclusions about the ground truth (cf. Davenport and Ball 2002; Andreas and Greenhill 2010; Carpenter et al. 2013; Chojnacki et al. 2012; Eck 2012; Gohdes and Price 2013; Krüger et al. 2013; Landman and Gohdes 2013; Siegler et al.

[1] We posit that there is only 'one true number' of the total of individuals who perish within an area that experiences an episode of lethal violence and that this true number can hardly be contested. We acknowledge that the establishment of a 'ground truth' is less clear-cut when it comes to establishing the circumstances that led to each individual death, i.e., whether it was (un)related to the use of armed force.

2008). These scholars argue that a multitude of issues inhibits our ability to observe and register both the entirety and a representative sample of the ground truth. Only random samples or full enumerations warrant statistical inference about magnitudes and patterns of violence, unless under-registration and selection bias in convenience samples are explicitly controlled for in a given empirical approach.

For macro- and microlevel research on conflict and violence to be accurate and valid, empirical methods are required that control for the 'visibility problem' in violence data. For example, we could mispecify the link between civilian victimization and bargaining between armed parties to a conflict, if not all civilian victims are observed, if there is spatiotemporal variation in the observation of victims, and if that variation is also systematically correlated with where, when, how, and/or by who these casualties were caused. Because we do not know whether these various *if's* are indeed an empirical issue, our conclusions about the link between deaths and bargaining remain highly uncertain.

In this paper, we demonstrate that the statistical method of multiple systems estimation (MSE) addresses the visibility and uncertainty problems inherent in violence data. Originating from the disciplines of biology and population ecology where this technique is commonly known as 'capture-recapture', multiple systems estimation has been developed to estimate the size of a population whenever it is impossible to fully observe that population. MSE has already been applied by human rights advocates to estimate conflict-related deaths in the context of truth commissions and criminal tribunals (e.g., Ball et al. 1999, 2002*a*; Ball and Asher 2002; Ball et al. 2003; Silva and Ball 2006). However, it yet remains to be adopted in political-science work on conflict-related deaths.

It is our goal to advocate for MSE as a promising tool to advance existing empirical practice in political-science research on conflict and violence. At the example of lethal violence in Kosovo between March and June 1999, we examine the reporting of killings and disappearances by three available data sources: the *American Bar Association* (ABA), *Human Rights Watch* (HRW), as well as the *Organization for Security and Cooperation in Europe* (OSCE). We compare these data sources to each other to identify commonalities and differences. We also compare the observed patterns to 'the ground truth', a recently completed census of war victims by the Humanitarian Law Centre in Belgrade, Serbia, and the Humanitarian Law Centre Kosovo (HLC) who have produced the 'Kosovo Memory Book' (Humanitarian Law Centre 2015; Krüger and Ball 2014; Spagat 2014). In comparison to the HLC data, we show that each of the three data sources under study

misrepresents magnitudes and patterns of violence by providing only an incomplete and biased snapshot of what happened in Kosovo between March and June 1999.

In a second step, we use the HLC data as a benchmark to evaluate the estimates we obtain from combining the ABA, HRW and OSCE data. This constitutes to our knowledge the first test of MSE in the context of estimating lethal violence. Usually, the performance of MSE can only be assessed theoretically or in experimental settings precisely because our knowledge of the underlying population suffers from non-visibility issues. The availibility of the HLC census provides us with a unique research opportunity to assess the performance of MSE on a real case.

In the remainder, we briefly review the origins of multiple systems estimation in other disciplines, contrasting it with current practice in empirical scholarship on conflict and violence. In Section 3, we introduce the basic statistical logic for two- and multiple-systems approaches. In Section 4, we present our empirical strategy - the data sources, the record linkage and estimation approaches we adopted, as well as the HLC data as our empirical benchmark. This is followed by descriptive analysis and a discussion of our estimates in Section 5. We conclude with suggestions for future research.

## 2 State of the art

Some populations of interest to scholars are not readily accessible, or even entirely visible, to obtain a full enumeration or apply some principled sampling procedure for valid and reliable inference. Originally an issue in the study of animal populations, scholars in the fields of biology and population ecology developed the method of capture-recapture, or *multiple systems estimation* (MSE), to address restrictions to fully observing a population of interest.

MSE comprises a class of statistical methods that estimate the total size of a population given several partial samples from the population and the overlaps among them. The earliest use of MSE used only two samples and was applied to fish populations in the fjords of Denmark (Petersen 1896). Since then, MSE has been utilized in many ecological applications, including estimation of the number of *** (@cite further animal studies here). Amstrup et al. (2010) offer a great overview of this literature.

MSE has been applied to human populations for correcting the US Census (Sekar and

Deming 1949; Darroch et al. 1993), and in epidemiology (Madigan and York 1997; Robles et al. 1988; Hook and Regal 1995). Indeed, many different types of human populations are understood as being 'hard-to-reach', 'hidden', or 'elusive.' Examples are the use of MSE to estimate the number of drug users (Larson et al. 1994; Buster et al. 2001; Comiskey and Barry 2001; Hope et al. 2005) and marijuana growers (Bouchard 2007), individuals with HIV infection (Abeni et al. 1994; Mastro et al. 1994) or diabetes (Gill et al. 2003; Haynes et al. 2004), homeless people (Fisher et al. 1994), lesbians (Aaron et al. 2003), sex workers (Kruse et al. 2003; Khan et al. 2004; Paz-Bailey et al. 2011), and recently, deaths during process of arrest in the US (Banks et al. 2015).

Similar to other disciplines in which the partial and non-random visibility of units in the population of interest is recognized, it is crucial that deaths in any population of conflict-related casualties (i.e., a certain episode of conflict-related violence) are also understood as remaining partially hidden, hard-to-reach, or elusive. To date scholars in the field acknowledge, for example, that perpetrators have plausible motives and may hence take practical measures to conceal their violent deeds, that conflict sites may become inaccessible to witnesses or information collectors for political restrictions, inaccessible terrain, or security reasons, or that the full extent of lethal violence may exceed available registration capacity (@cite). There is further reason to assume that such visibility-inhibiting factors may be systematically linked to characteristics of the units of interest (i.e., the events, the victims, the perpetrators, the observers, the context), which could introduce systematic measurement error, i.e., *selection bias* (@cite). For example, it is suggested that violence in urban areas may be more visible than in rural areas.

The characterization of any population of conflict casualties as elusive would dictate that political-science research designs incorporate such capture complexity of the units of analysis in the empirical work. We argue that multiple systems estimation provides an appropriate tool to address this empirical challenge as evidenced in other disciplines that deal with similar population characteristics of elusiveness this way.

In the sphere outside of political-science research on conflict and violence, MSE has now been used multiple times to estimate conflict-related deaths in support of historical verification and truth commission processes, national human rights campaigns by non-governmental organizations, as well as criminal justice tribunals. More precisely, it was used to provide an estimate of conflict-related deaths in Guatemala (Ball et al. 1999), Kosovo (Ball 2000; Ball et al. 2002*a*; Ball and Asher 2002; Ball et al. 2002*b*), Bosnia

(Brunborg et al. 2003; Zwierzchowski and Tabeau 2010), Perú (Ball et al. 2003), Timor Leste (Silva and Ball 2006), and Colombia (Guzmán et al. 2007; Guberek et al. 2010; Lum et al. 2010; Guzmán et al. 2012). In the recent statistical literature, specifics of MSE applications have been discussed with regard to estimating conflict-related deaths (Lum et al. 2013; Mitchell et al. 2013, 2015). In all the mentioned applications, samples are typically lists of casualties collected by various agencies.

Political-science work on conflict and violence has to yet incorporate multiple systems estimation into empirical research designs. A few works mention, propose, describe or discuss MSE as a method suited to address problems of under-registration and selection bias in violence data (Landman 2006; Landman and Carvalho 2010; Jewell et al. 2013; Manrique-Vallier et al. 2013; Otto 2013; Salehyan 2015). To our knowledge, however, only a few works have relied on some version of system-estimation logic to obtain an estimate of ground truth (Hoover Green 2011; Birnir and Gohdes 2014; Gohdes 2014; Hendrix and Salehyan forthcoming). To help advance current practice in the field, we demonstrate the usefulness of the MSE technique on the context of conflict lethality at the example of a real case for which we also have the unique opportunity to assess our findings against the ground truth.

## 3 Theory of multiple systems estimation

The logic of multiple systems estimation has been described elsewhere many times and it is not our goal to duplicate those efforts. Rather, we seek to provide the most basic ideas and refer the interested reader to more in-depth explanations provided in Manrique-Vallier et al. (2013) and Lum et al. (2013).

To start, let us briefly clarify the two major concepts involved in MSE theory. A 'system' describes some type of sampling mechanism that produces a set of records. Ideally, this would be some kind of random sampling process. In our particular case, a system would more likely be a list of casualties coded from news reports, or a database gathered by a military institution, a police station, a non-governmental group, a truth commission, and so forth, which all represent convenience samples. The 'population', in turn, circumscribes the entirety of research subjects for whom the total size is unknown and hence to be estimated. A system produces a subset of that population, either by simple random sampling or some other, less principled sampling procedure – in our case the latter.

### 3.1 Estimation with two systems

The intuitive logic behind multiple systems estimation is best explained at the example of a two-systems model and then expanded to the case of more than two systems. The two-systems case builds on probability theory, assuming two random-sample systems $A$ and $B$ that each draw a set of units from finite population $N$.

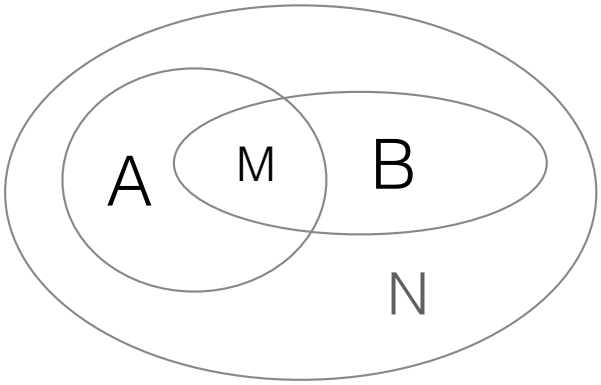



Figure 1: Drawing two systems A and B from population N.

The probability of a unit to be sampled into either $A$ or $B$ is $A/N$ or $B/N$, respectively. We refer to the conjunction of $A \cap B$ as $M$. The probability of a unit to be sampled into $M$ is $M/N$. We can calculate $M/N$ by multiplying the two individual probabilities: $M/N = A/N * B/N$. Solving this equation for $N$, the – to us unknown – size of the population, we obtain our main equation from which $N$ can be calculated using the observed counts:

$$N = AB/M \tag{1}$$

For each unit in $N$ ($i_n$), there are four possible *inclusion patterns* with regard to $A$ and $B$ ($i_{AB}$) in this two-systems selection process, which can be summarized in a 2x2-contingency table:

| Selection | $B$ | $\neg B$ |
|---|---|---|
| $A$ | M = $i_{11}$ | A = $i_{10}$ |
| $\neg A$ | B = $i_{01}$ | N = $i_{00}$ |

Table 1: Contingency table for two-system selection.

To estimate a population of conflict-related deaths with two random samples of casualties, for example, we would estimate the unknown size of $N$ in the fourth cell in Table 1 by applying our main two-systems equation (1) above: multiply the counts of individuals captured by only $A$ and $B$, respectively, and divide that product by the number of individuals -observed in both $A \cap B$, or $M$.

While the two-systems logic to calculate $N$ is very intuitive, it only holds when four crucial modeling assumptions are met:

[1] *Closed system:* Each system must refer to the same population $N$, i.e., none of the units can enter or leave the population within the space-time window of interest.

[2] *Perfect matching:* Every unit has to be uniquely and accurately identifiable to correctly determine the size of each system $A$ and $B$, as well we as their conjunction $M$.

[3] *Homogeneity:* Every unit in the population has equal probability of capture in a given system, here $A$ and $B$.

[4] *Independence:* The systems are independent of each other, i.e., a unit's probability of (non-)capture in $A$ does not impact its probability of (non-)capture in $B$, and vice versa.

The first assumption of a *closed system* is met when estimating conflict-related deaths. Different to animals, for example, human deaths – once they occurred – cannot migrate into or out of the area and time period for which we seek to estimate all deaths that occurred.

The second assumption of *perfect matching* requires that the process of determining whether a given pair of captured individuals (within one or across two systems) refers to the same or different people is accurate. In Section 4.2 below, we discuss how we address this requirement.

The [3] homogeneity assumption is usually violated as lists of conflict-related deaths rarely or ever constitute random samples of the true number of deaths. Instead, we observe 'unit heterogeneity' of capture (or, *capture heterogeneity*) as some deaths are more likely to be registered than others. For example, existing literature suspects that deaths are more visible in urban areas (vs. rural), in central, easily accessible areas (vs. remote), with more victims (vs. events with less casualties), or due to the identity of the target (highly visible figures of public interest such as politicians, journalists or peacekeepers vs. individuals outside of public interest) (@cite). As a consequence, the population of deaths results in unequal capture probabilities for different 'strata', i.e., subsets of the population. Such 'strata' are, for example, individuals of one sex, within a given age group, or deaths occurring within a certain area or time period.

The [4] independence assumption is also often violated in the case of systems capturing conflict-related deaths. The various actors who are producing lists of deaths (e.g., human rights activists, the police, the military, a city council, a truth commission) are not operating in a vacuum but either simultaneously or successively. They may be drawing from the same sources (e.g., the news, witnesses, a public database), collect data from separate but overlapping populations, or even draw information from each other (e.g., refering cases to, copying from, exchanging with or consulting the other). List dependence may also be related to the issue of capture heterogeneity. As a consequence of list dependence, the capture probability of a given death into one system (e.g., a truth commission) is higher if that death was also captured by another system (e.g., a non-governmental organization) because the two systems are somehow related (e.g., the NGO supports the truth commission with information about lethal violence during the conflict or they both have their headquarters in the same district).

Capture heterogeneity and list dependence in violation of assumptions [3] and [4] cannot be addressed in two-systems estimation. Instead, three or more systems of registered deaths, as well as more complex statistical techniques are required to estimate the size of the unknown ground truth from multiple systems.

### 3.2 Estimation with multiple systems

When we overlap multiple systems, we make more information available from which the ground population can be estimated because we gain more complex inclusion patterns. For example, three systems yield seven different combinations ($i_{100}, i_{010}, i_{001}, i_{110}, i_{101}, i_{011}, i_{111}$) to estimate $i_{000}$, four systems result in 15 different inclusion patterns to obtain $i_{0000}$, five systems 30, etc.

With a more complex overlap structure, we require more sophisticated statistical solutions that model for every death the probability distribution of belonging to one particular inclusion pattern while being excluded from every other. The MSE literature has established a variety of statistical models to address specific characteristics of the populations that are to be estimated, such as log-linear models (@cite), Bayesian modifications (@cite), discrete mixture models (@cite), and many more. Each of these models relies on a different set of assumptions which are usually laxing those of the two-systems estimator.

The main advantage of using more than two lists is that capture heterogeneity and

list dependence can be addressed. In the context of estimating conflict-related deaths, practitioners use non-parametric 'stratification' to model capture heterogeneity directly. 'Stratifying' means that the entire pool of observed records is grouped into plausible latent classes, or 'strata' (i.e., subsets), for which class-specific capture probabilities are expected. The size of the unknown subset populations is then estimated separately based on the given overlap structure within a subset (@cite). Stratification therefore models selection bias directly. Below, we illustrate our reasoning in creating different strata to estimate conflict casualties in Kosovo between March and June 1999.

When more than two systems are available, we can also model potential list dependence. Log-linear models offer one solution to the list dependence problem by exploring different dependence models (Bishop et al. 2007). Typically in the log-linear regression framework for MSE, a Poisson likelihood is assumed, though over dispersed models with a negative binomial likelihood have been proposed (@cite). In this framework, list dependence is represented by pairwise or multi-way interaction terms. These interaction terms allow the probability of capture on each combination of lists to vary from what would be expected if the lists were all assumed to be independent. This adjusts the estimates to account for list dependence.

However, this also raises the additional question of which interaction terms to include in the model. If one places no restrictions on the structure of interaction terms to be included in the log-linear regression (each possible combination of interaction terms is called a model), the number of possible models grows super-exponentially in the number of lists. Standard model fit statistics have been employed to select from among the set of possible models (@ cite to AIC and BIC approaches).

Alternatively, rather than selecting one single model to represent the dependence between lists, others have used model averaging to propagate model uncertainty all the way through to the estimates (Madigan and York 1997; Lum et al. 2010). Calculating the necessary weights to perform Bayesian model averaging is aided by moving away from the log-linear modeling setup, where marginal likelihoods are not available in closed form. By using a multinomial likelihood with a hyper-Dirichlet prior on the capture pattern probabilities, closed form model averaging weights are available. This is the method we use in the analysis presented in this work.

Other models have focused more on mitigating list dependence by directly modeling

capture heterogeneity through stratification. Within each strata, lists are assumed to be independent. Marginal of latent class, however, this structure induces list dependence across all lists (Manrique-Vallier and Fienberg 2008; Manrique-Vallier 2014). Models that apply specific parametric distributional assumptions to the underlying individual-level catchability can be found in Darroch et al. (1993); Rivest and Baillargeon (2007); Coull and Agresti (1999). Lastly, given the unidentifiability of the form of the catchability distribution, others have focused instead on estimating a lower bound on the population size, regardless of the form of the distribution of catchability (Chao 1987; Rivest 2011).

## 4 Empirical strategy

Kosovo provides us with a unique research opportunity to assess the promise of using MSE in empirical research on conflict and violence. In the first half of 1999, the Kosovo region in the south of Serbia saw an intense period of massive violence against ethnic Kosovar Albanians. The Serbian government was staging a counter-insurgency campaign to stifle growing mobilization into and radicalization of the Kosovo Liberation Army (Hayden 1999). Increasingly, the Serbian leadership was accused of conducting an ethnic cleansing campaign given its massacres and mass expulsions of the local Albanian population. Accusations of violence ultimately triggered NATO intervention in the form of air strikes between March and June 1999.

Knowledge about the killing of Albanians on the ground in Kosovo was desired both during and after the Kosovo conflict. Prior to NATO intervention, policy-makers assessed the decision to intervene, while during the intervention military leaders sought to update their bombing strategy (Clark 2001). Next to governmental intiatives (Organisation for Security and Co-operation in Europe 1999), multiple agencys collected information on killings and disappearances, for example to support criminal tribunals at the ICTY (cf. Human Rights Watch 2001; Ball and Asher 2002; Ball et al. 2007).

In this section, we describe the three systems we use to estimate conflict-related deaths in Kosovo between March and June 1999 that became available shortly after conflict in Kosovo ended. We further discuss how we match victims across these three lists and what statistical estimation model we select. We also introduce our benchmark data to which we compare reported and estimated deaths.

## 4.1 Three data sources on lethal violence in Kosovo

Three data sources on victims of lethal violence became available soon after conflict in Kosovo ended. The *American Bar Association/Central and East European Law Initiative* (ABA) interviewed ethnic Albanian refugees in camps or private homes in Kosovo, Albania, Macedonia, Yugoslavia, Poland and the United States. *Human Rights Watch* (HRW) equally conducted interviews with ethnic Albanian refugees at Kosovar border crossings as individuals were fleeing into Albania, Macedonia, or Montenegro in the period of March and June 1999. Additionally, HRW collected interviews in various Kosovar regions throughout Kosovo in the second half of 1999. The *Organization for Security and Cooperation in Europe* (OSCE) was engaged in a 'Kosovo Verification Mission' since October 1998, in addition to an OSCE Mission in Kosovo since June 1999. Between March and June 1999, OSCE interviewed ethnic Albanians refugees in camps, private homes or communal places in Albania and Macedonia, but had no access to Kosovar territory. Observers only resumed collecting testimonies within Kosovo after the OSCE Mission was established in June 1999.[2]

The ABA, HRW and OSCE data provide individually identifying information on victims who were either reported killed or disappeared due to armed conflict in Kosovo for the period of March and June 1999. For every victim, each list provides information on their name, date of birth, sex, location and date of violence. Such personally-identifying information of every death provides the first step in satisfying MSE assumption [2] (cf. Section 3.1).

These victim lists are ideal for a three-system estimation of deaths in Kosovo between March and June 1999. We assume that every reported victim did, in fact, occur (i.e., there is no false reporting). The main advantage of using the ABA, HRW, and OSCE data is that we expect them to be incomplete but suitable to estimate the true number of all deaths during said episode of lethal violence.

## 4.2 Matching

To identify the inclusion patterns for every victim reported in the ABA, HRW, and OSCE lists, personally identifying information of victims was evaluated within, as well as across the three data sources. 'Record linkage' or 'matching' was performed to determine whether

[2] A more detailed description of these sources can be found in Ball and Asher (2002).

a given pair of two records designated the same, a 'match,' or different people, 'a non-match.' Positive matches help eliminate duplicates within a given list, but most importantly determine whether a given victim, reported by at least one of the sources, was also reported by one or even both of the other two list.

To ensure that all records were perfectly matched, a human coder reviewed name, age, sex, location and date of each reported death. It was thus determined whether reported victims matched each other within a given data source (duplicates), as well as across the ABA, HRW, and OSCE data sources (overlap). Because date and location information seemed to be volatile even within matching record clusters, name information was deemed most decisive in matching decisions. Clusters of records identified to belong to the same match group were merged into one single record by preserving an arbitrary selection of location and date values within the cluster.[3]

An inter-coder reliability review of a random sample of records was performed by both authors in which no false positives (i.e, records that were marked as matches when they did not seem to match), or false negatives (records that were kept as non-matches when they seemed to match) could be identified. This was deemed sufficient evidence that perfect matching (cf. Section 3.1) was achieved.[4]

From record linkage, we obtained binary inclusion variables for every data source given the total pool of unique records reported jointly by the ABA, HRW, and OSCE data. For each data source, the inclusion variable denotes whether a given record was captured (1) by this source or not (0). The three inclusion variables for ABA, HRW, and OSCE jointly provide the seven observable overlap patterns for every record in our uniquely-identified record pool – $i_{100}, i_{010}, i_{001}, i_{110}, i_{101}, i_{011}, i_{111}$. Note that the eighth inclusion pattern $i_{000}$, i.e., the number of records not captured by either of the three data sources, is unknown and to be estimated.

Information on the location and date of a violation which remained missing even after merging match clusters was imputed. Simply dropping records with missing information from the analysis is ill-advised, as this changes the population of all deaths from which the data is drawn to the population of all deaths for which the location and date of death or disappearance is known. We therefore impute missing data conditional on all other known variables, including list inclusion variables. We describe the details of our imputation

[3] A future iteration of this work will explore different merge decisions in comparison.

[4] We are grateful to Michelle Dukich for matching the Kosovo data.

procedure in Appendix A.

### 4.3 Statistical model: hyper-Dirichlet approach

In this paper, we use the model developed in Madigan and York (1997) to perform MSE. In their model, the log-linear representation of expected list overlap counts (cf. Section 3.2) is replaced with a hyper-Dirichlet prior distribution on the probabilities and a Multinomial likelihood. This substitution is not unfounded -– the Multinomial distribution arises in the case of Poisson sampling with a fixed total. Here, we place a prior on $N$, the total population size, and average over this parameter. The standard priors used all require the analyst to specify a range of possible values of the total population size, thus constraining the estimates to a plausible range. This constraint has a distinct advantage relative to other approaches. It prevents the estimates from becoming implausibly large (many orders of magnitude larger than the total number of recorded individuals), which is common in other estimation techniques, particularly when many list intersections contain no records. Whereas in the Poisson log-linear framework, list dependence is modeled via interaction terms on the list effects, here list dependence is represented by graphical models with list intersection probabilities modeled as products of marginal probability vectors.

Often, we do not know *a priori* which list dependence structure best represents our data, so we must resort either to variable selection (i.e., selecting interaction terms for inclusion or exclusion in the log-linear framework) or model averaging (i.e., averaging over all possible list dependence structures according to weights that are based on how well each structure fits the given data). Model averaging is preferable to variable selection in this case, as model averaging propagates the uncertainty about the correct list dependence structure all the way through to the final estimates and intervals Draper (1995); Hoeting et al. (1999). In our approach, there is a closed form expression for the posterior probability of each model of list dependence. That is, the weight of evidence to be apportioned to each list dependence structure given the observed data can very easily be calculated. The computational ease with which sensible weights on each model are obtained is one of the main advantages of this approach.

### 4.4 Benchmark data

We face an unusual research opportunity in having a benchmark to evaluate the ABA, HRW, and OSCE data, as well as our estimates. The Humanitarian Law Centre in Belgrade, Serbia, and the Humanitarian Law Centre-Kosovo (here co-jointly referred to as 'HLC') have produced a census of human losses in connection with armed conflict in Kosovo between 1998-2000 (Humanitarian Law Centre 2015). HLC defines as a 'war victim' any individual who was killed or disappeared due to the use of armed force (cf. Krüger and Ball 2014, 5,9).

With the magnitudes and patterns it documents, HLC provides an accurate and reliable data source for statistical analysis of the conflict (Krüger and Ball 2014; Spagat 2014). Full enumerations of ground victims such as the HLC database are rarely, if ever, available to scholars or practitioners.

We extracted a total of 9,945 war victims from the HLC database. These are all human losses HLC reports for the territory of Kosovo for the four-month period of observation between March and June 1999.

## 5 Discussion of results

### 5.1 Descriptive evidence of underregistration and selection bias

In a first step, we compare the reporting of the three data sources ABA, HRW, and OSCE to each other, and to the HLC data. For the three data sources to be considered representative of the ground truth of war victims in Kosovo, they need to report either a complete count of victims with regard to the given space-time dimensions. Or, to be at least representative in terms of spatiotemporal patterns (i.e., unbiased), we had to observe the same spatial and temporal patterns as documented in the HLC database regardless of potential under-registration. If either of the two conditions was met, each of the ABA, HRW, or OSCE data may be used for statistical analysis of lethal violence patterns in Kosovo between March and June 1999. However, an in-depth cross-examination of the three data sources in comparison to the HLC provides strong evidence of underregistration and selection bias with regard to both the spatial and temporal dimension.

With regard to the quantity of victim registration, the three data sources can by far not be considered complete. ABA reports a total of 561 victims, HRW documents 677

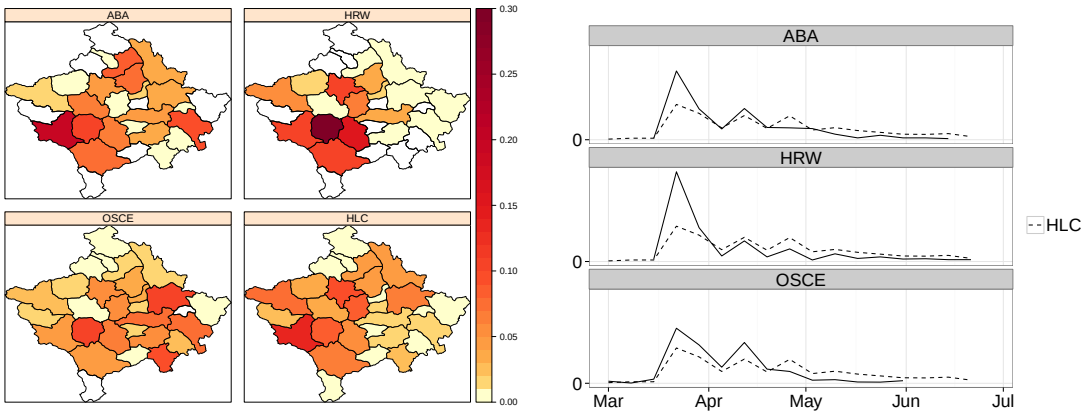
victims, and OSCE 1,834. Even combined (2,676 unique pooled victims), these reported victim totals are very short of the total of 9,945 war victims documented by the HLC within our chosen space-time window.

Even though we find the three data sources to suffer from significant under-registration, someone may argue that they could still be representative of spatial and temporal patterns on the ground. In Figure 2, we compare the percentage distributions of the four data sources by municipality and by week. Comparing the maps in Figure 2(a), we can see that each of the four data sources reports a rather different spatial distribution of violence across Kosovo's 29 municipalities. While violence is reported to spread between the northeast and southwest regions in ABA, HRW reports the center of violence mostly for the southwest. In both the OSCE and HLC data, violence is reported more widespread throughout Kosovo, albeit with differing municipal centers, i.e., OSCE east vs. HLC west. In general, all three data sources diverge signifantly from the true municipal pattern reported in HLC.

In Figure 2(b), we observe the percentage distribution of the temporal trend of lethal violence for each of the three data sources against the HLC trend. All three data sources are found to overrepresent violence during the second half of March, while OSCE also overrepresents violence during the first half of April in comparison to HLC. All the data sources underrepresent the temporal distribution of violence from mid-April onwards. In particular, ABA, HRW, and OSCE seem to miss a violent episode during the second half of April.

The sources' inability to capture violence from mid-April onwards matches historical accounts of political violence in Kosovo. After mid-April, Serbian authorities closed border crossings. This political measure made it significantly less likely that witness-statements of casualties traveled with refugees to interviewers outside of Kosovo – the main method of data collection for the ABA, HRW, and OSCE data generators. It is noteworthy that the OSCE data also underrepresent violence during this second half of the war. The organization resumed its statement-taking mission within Kosovo after the ceasefire agreement came into force in early June. However, it does not seem to recover sufficient information on lethal violence ex-post.

To obtain a better understanding of biased reporting patterns compared to the HLC census, further bargraphs are provided in Figure 3. Both ABA and HRW agree with HLC that the majority of lethal violence occurred in the western region (Figure 3(a)).




(a) By municipality

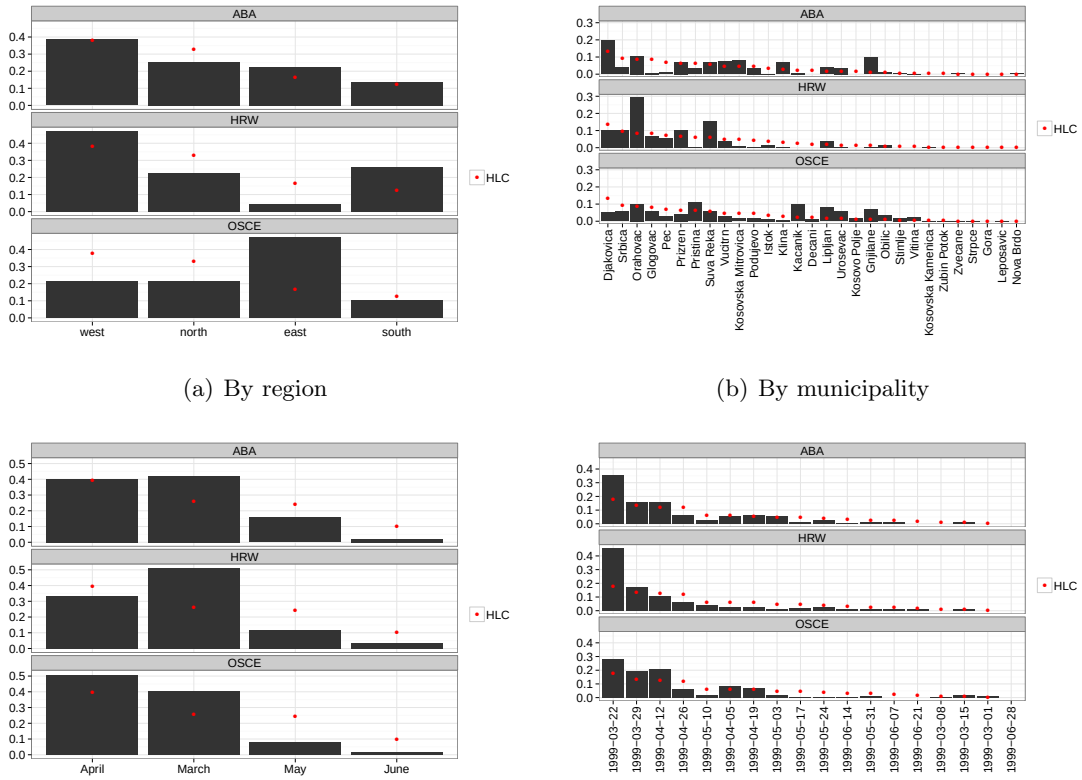
(b) By week

Figure 2: Reported patterns (%) of lethal violence, by data source.

ABA fits the regional pattern best, underrepresenting the north while overrepresenting the east. HRW significantly overrepresents violence in the south, OSCE in the east. It is notable that even at this most aggregate spatial level, at least two of the data sources do not capture a rather representative regional distribution of violence (HRW and OSCE). If a scholar had had access to all three data a few years after the conflict when the HLC data was not yet available, she may not have been tempted to trust the ABA distribution given this data source is the smallest overall. Looking at the percentage distribution of reported victims across municipalities, the differences in spatial patterns become even more pronounced (Figure 3(b)).

With regard to temporal trend, OSCE is closest to the monthly HLC pattern (Figure 3(c)). Both ABA and HRW overrepresent March compared to April. All three data sources underrepresent violence in May and June. Disaggregating to the week-level, allows us to inspect differences in reported temporal patterns even further (Figure 3(d)). In both graphs, we reconfirm the observation that all three data sources underrepresent violence from mid-April onwards. In particular, lethal violence in the weeks of April 26 and May 10 remains considerably underregistered.

Visual inspections of the reported patterns over space and time against the HLC distributions show us that all three data sources suffer from significant spatial and temporal selection biases. With the exception of the regional distribution reported in ABA, scholars or policymakers would risk drawing inaccurate conclusions about the spatiotemporal spread of violence in Kosovo between March and June 1999 were they only to rely on observed information reported in the ABA, HRW, or OSCE data.




(a) By region

(b) By municipality

(c) By month

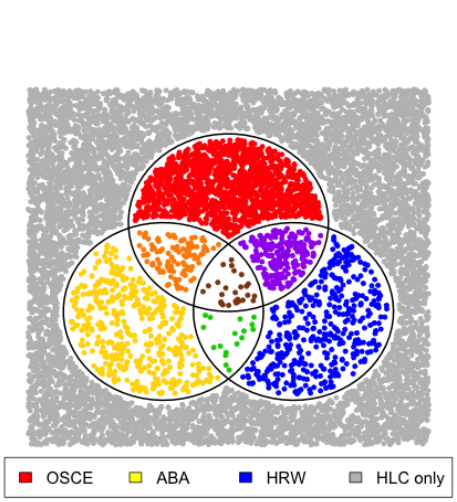
(d) By week

Figure 3: Reported spatial and temporal patterns of violence vs. order of magnitude in HLC, by data source and strata.

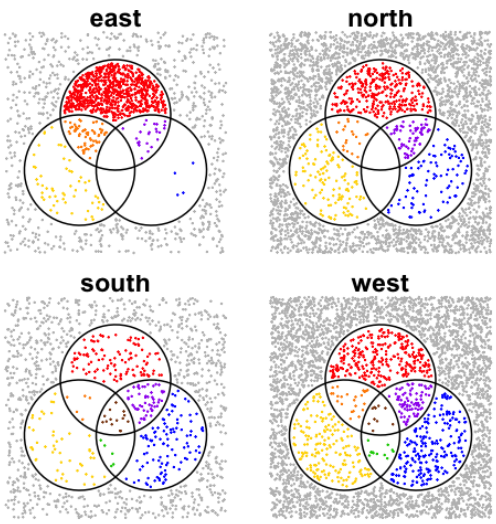
A further opportunity for a visual comparison of reported patterns is inspecting the overlap patterns of the inclusion variables with the help of venn diagrams. In Figure 4, the victims reported by each data source at a given aggregate level are color-coded within a circle. The amount of the 'unknown' non-captured records ('000') that is documented in the HLC data for a given aggregation is depicted by gray dots outside of the capture circles. Note that usually we do not have benchmark data such as the HLC available, therefore having to estimate the amount of the gray dots. The venn diagrams show that there is some pair-wise overlap between OSCE and HRW, especially in March 1999. In general however, we observe that pair- and three-source overlaps are rather rare at the regional and monthly levels.[5]

The overlap patterns provide further evidence that the stories of lethal violence told by ABA, HRW, and OSCE have little in common. Rather, every data source seems to capture a distinct snapshot of victims of lethal violence during this episode of the Kosovo conflict. Information between the three data sources is complimentary, not confirmatory. Using these three data to inform policy-making or scientific inquiry would risk incorrect

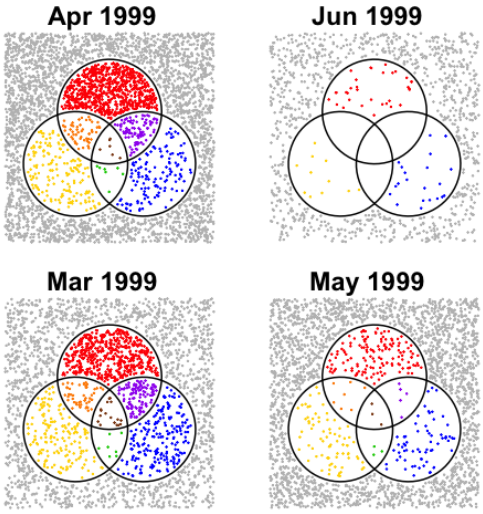
[5]Venn diagrams of record overlap by municipality and by week are provided in Appendix C.




(a) Aggregate list overlap




(b) By region




(c) By month

Figure 4: Overlap and additional HLC records, by counts and stratifications.

conclusions.

### 5.2 Estimates of lethal violence in Kosovo, March-June 1999

We now investigate whether multiple systems estimation corrects for underregistration by the three data sources, as well as for selection bias in the reported patterns that we uncovered relative to HLC's census of war victims.

In Figure 5, we present the estimated war victim totals we obtain by stratifying the underlying data in different ways. These estimated totals are derived from summing the point estimates of all the strata within a given stratification model. The dashed red line denotes the 'ground truth,' i.e., the total of 9,945 war victims documented by the HLC.

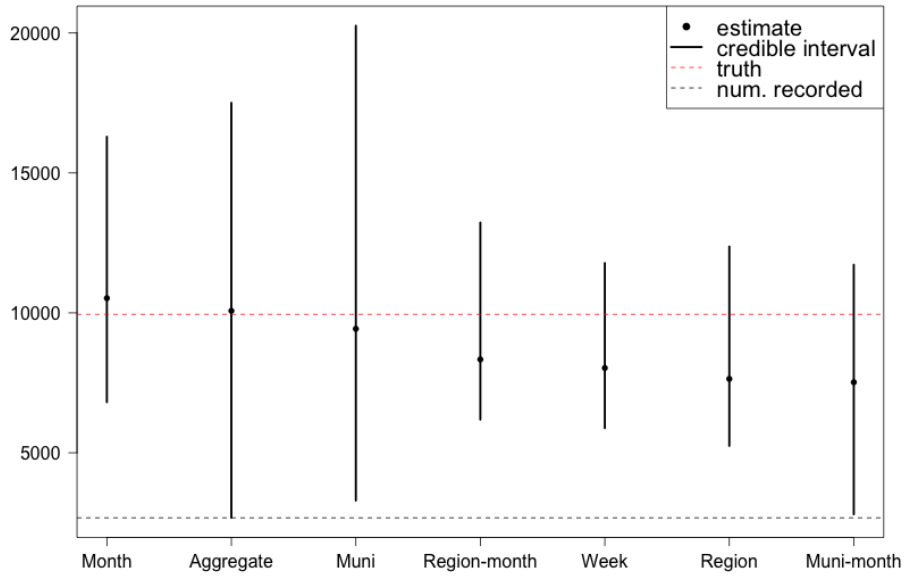



Figure 5: Estimates of total war victims, by stratification.

The dashed black line denotes the total of 2,676 unique victims jointly reported by ABA, HRW, and OSCE.

As can be seen, all seven models' credible confidence intervals contain the true count of war victims documented by the HLC. The aggregate model is the most simple MSE model of only one stratum which we obtain by estimating from the basic overlap structure of the three data sources (cf. Figure 4(a)). Note that this plain model without any stratification can not correct for spatial or temporal reporting differences.

To produce different strata for which to estimate, we agreed on two criteria: (1) a chosen stratum had to contain a total of at least 200 victims, and (2) there had to be at least 2 non-zero overlap cells out of the four possible ones to reliably estimate a stratum. Whenever these criteria were not met by a suggested stratum, these latent classes were combined into an 'Other' stratum. To divide the observed victims into different latent classes we considered information on the violation region, month, municipality and week, as well as the region-month, and the municipality month.

Additional to the aggregate model, Figure 6 shows the estimates resulting from six different spatiotemporal stratifications of the Kosovo data. The white areas in the stacked bars denote the total of uniquely recorded victims within a given strata when the uniquely identified individuals are pooled across the three data sources. The black areas denote the 'black figure', i.e., the number of undocumented victims we estimate from the recorded

data with associated uncertainty. The red dots represent the 'ground truth' represented by the HLC.

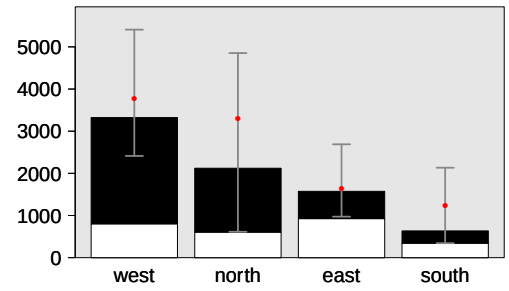
Stratifying by region (Figure 6(a)), it becomes clear that we are largely underestimating violence in the north of Kosovo. This is not surprising however because all three data sources – ABA, HRW, and OSCE – underreport violence for this region given our knowledge of the HLC census (cf. Figure 3(a)). In the most simple terms, MSE is unable to magically correct for information that is entirely missing from all available systems.

More formally, this issue results from a special case of *capture heterogeneity*. As mentioned earlier, it is being assumed that deaths have different likelihoods of being captured by one, several, all, but also *none* of the systems considered. If some records are systematically different from others by being unlikely to be captured in even one of the systems, we are going to underestimate this type of completely undocumented individuals altogether.

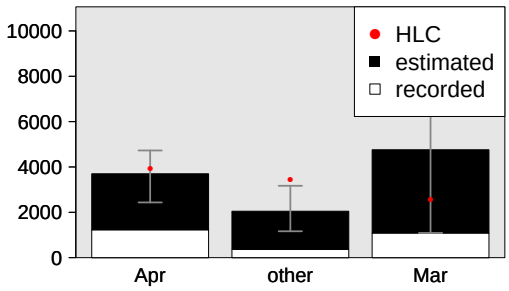
At this point, we can only speculate why the three data sources underreport violence in the north. One plausible explanation could be that a significant proportion of violence in that region occurred in the time period after mid-April for which we found the three data sources to underrepresent victims. An exploration of regional HLC timelines (cf. Figure 7 in the appendix) supports this suspicion as it turns out that a major episode of violence is documented in the north for the end of April/beginning of May 1999 that we earlier found to have been missed by all three data (cf. Figure 2(b)).

Even more important with regard to the regional stratification model is our finding that MSE corrects the regional misrepresentation of violence that we obtain from the HRW and OSCE data, respectively, as well as from the three data sources pooled. The pooled regional pattern of reported victims suggests that most violence occurred in the east, followed by the west and north. This finding, contrasted with the HLC's true count, provides strong evidence against the assumption that pooling various data sources cancels out existing biases. In the Kosovo case, pooling the ABA, HRW, and OSCE data perpetuates regional selection bias but MSE provides a correction.

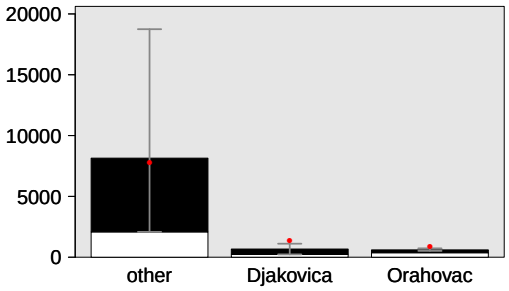
To stratify estimates by month (Figure 6(b)), we had to combine May and June into one stratum. As can be seen in Figure 4(c), there is no pair-wise or three-source overlap between the ABA, HRW, and OSCE data in June, which is why we cannot estimate this month separately. We clearly overestimate lethal violence in March and April, while underestimating May and June. This, again, is driven by the fact that all three data sources



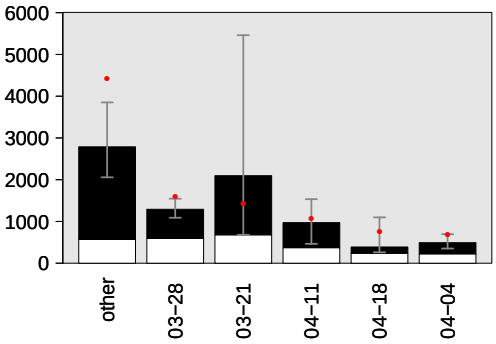

(a) By region



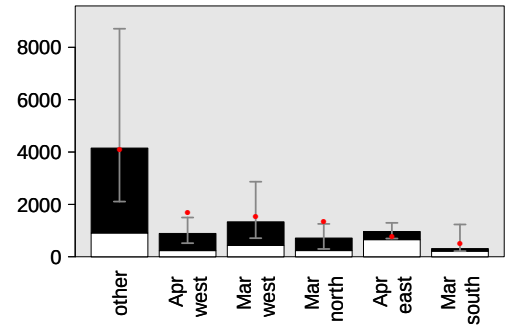

(b) By month



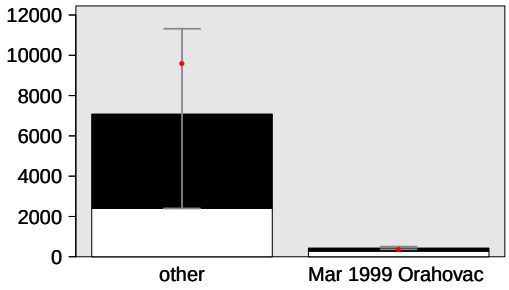

(c) By municipality




(d) By week




(e) By region-month




(f) By muni-month

Figure 6: Comparison of data source counts and estimates to HLC counts, for different stratification models.

overreport for the month of March, while deaths in May and June are significantly underreported. As discussed earlier, HLC can correct for a biased representation in the observed data unless all of the available sets suffer from the same type of catpure heterogeneity in the most severe sense (i.e., no chance of being captured by either of the available systems).

A by-municipality estimate is not very different from the aggregate model. Only two municipalities satisfy our stratification requirements (Figure 6(c)). Similarly, a by-week estimate only estimates five strata out of 18 possible ones (Figure 6(d)). Because the data sources underreport victims from the end of April onwards, strata for late April or even later disappear in the 'other' category. Stratifying by region-month or muni-month (Figures 6(e) and 6(f)), does not provide any additional insight either.

In our case, stratifying documented victims by either month or region appears most plausible. Our data is too sparse and heterogeneous in capture to warrant more fine-grained stratification strategies. Even without the HLC census available, our suspicion of capture heterogeneity during the second half of the conflict, as well as across different Kosovar regions can be informed by our contextual knowledge of this conflict episode and the data sources that generated the victim lists.

Both the by-region and the by-month estimates' credible intervals contain the true HLC count. Without the victim census, we would expect the population of conflict-related deaths to have ranged between at least 5,248 (by region) and at maximum 16,293 victims (by month). The lower bound of the region estimate would almost double the number of victims documented by the ABA, HRW, and OSCE data combined (2,676).

### 5.3 Sensitivity analysis

In future iterations of this paper, we will conduct different sensitivity analysis to evaluate the robustness of our results (log-linear models, lower-bound models).

## 6 Conclusion

Does MSE present an opportunity for the empirical study of conflict and violence in current political-science scholarship? Practically, one was able to obtain an accurate conflict-related death estimate more than ten years prior to the publication of the HLC census using three very incomplete and non-random victim lists. In most episodes of armed violence, we

are most likely never going to have a casualty census available. As we demonstrated in this paper, MSE provides a promising tool for obtaining a credible estimate of the lethality of episodes of armed violence for which the true size and distributions are usually unknown.

It became clear in the descriptive summary of the three data systems and their respective overlap structures, that even for a small area of the size of Kosovo and a very short period of four months on the European continent, three data sources were unable to capture a true representation of ground violence. We posit that someone had to present very strong arguments to claim that data sources covering conflict-related deaths in periods of armed conflict outside of Europe, of possibly much higher lethality, spanning significantly larger territories and time periods often involving years would provide a better capture of conflict-related casualties to reliably represent magnitudes and trends than in the case of Kosovo. More likely, we suspect a link between the visibility of lethal violence in Kosovo and there being a census of war victims today.

While a general notion in the field may currently be that "bad data is better than no data," we would like to suggest that there is no such thing as 'bad data' as long as the available data is of sufficient usable quality. The ABA, HRW, and OSCE data used here each constitute reputable data projects that undertook a recommendable effort to document what was documentable at their time. It is bad practice however to use good data in ways that are not fitted to the specific characteristics of the research population for which statistical inference is sought. Researching true magnitudes and patterns of lethal violence with convenience sample data without correcting for very likely underregistration and selection bias is bad practice. MSE provides a good practice solution to use good convenience sample data.

[Provide a step-by-step how-to guide – what is needed (data, software), what to look out for. Statistical packages available in R: Rcapture(Baillargeon and Rivest 2007), dga (Johndrow et al. 2015).]

We see the following opportunities for future research:

1. context-informed discussion of severe capture heterogeneity: our knowledge of the data-generating processes in ABA, HRW, OSCE; vs. cases documented in HLC.

2. additional estimates with data that became available at a later stage, e.g., ICMP, ICRC, OMPF in 2007, – what further knowledge do these data provide and how does that change our estimates?

## References


Aaron, D. J., Chang, Y.-F., Markovic, N. and LaPorte, R. E. (2003), 'Estimating the lesbian population: a capture-recapture approach', *Journal of Epidemiology and Community Health* **57**(3), 207–209.

Abeni, D. D., Brancato, G. and Perucci, C. A. (1994), 'Capture-Recapture to Estimate the Size of the Population with Human Immunodeficiency Virus Type 1 Infection', *Epidemiology* **5**(4), 410–414.

Amstrup, S. C., McDonald, T. L. and Manly, B. F. (2010), *Handbook of capture-recapture analysis*, Princeton University Press.

Andreas, P. and Greenhill, K. M., eds (2010), *Sex, Drugs, and Body Counts: The Politics of Numbers in Global Crime and Conflict*, Corn, Ithaca and London.

Baillargeon, S. and Rivest, L.-P. (2007), 'Rcapture: Loglinear Models for Capture-Recapture in R', *Journal of Statistical Software* **19**(5), 1–31.

Ball, P. (2000), *Policy or Panic? The Flight of Ethnic Albanians from Kosovo, March–May 1999*, American Association for the Advancement of Science, Washington D.C. Last accessed May 29, 2014.
**URL:** *https://hrdag.org/wp-content/uploads/2013/07/kosovo-Policy_or_panic-2000.pdf*

Ball, P. and Asher, J. (2002), 'Statistics and Slobodan: Using Data Analysis and Statistics in the War Crimes Trial of Former President Milosevic', *CHANCE* **15**(4), 17–24.

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'.
**URL:** *http://shr.aaas.org/peru/aaas_peru_5.pdf*

Ball, P., Betts, W., Scheuren, F., Dudukovich, J. and Asher, J. (2002*a*), 'Killings and Refugee Flow in Kosovo March - June 1999: A Report to the International Criminal Tribunal for the Former Yugoslavia', Washington, DC: AAAS and ABA/CEELI. Last accessed November 15, 2011.
**URL:** *http://shr.aaas.org/kosovo/icty_report.pdf*

Ball, P., Betts, W., Scheuren, F., Dudukovich, J. and Asher, J. (2002*b*), 'Killings and Refugee Flow in Kosovo March - June 1999: A Report to the International Criminal Tribunal for the Former Yugoslavia, Corrigendum', Washington, DC: AAAS and ABA/CEELI.
**URL:** *http://shr.aaas.org/kosovo/corrigendum/corrigendum-021115.pdf*

Ball, P., Kobrak, P. and Spirer, H. F. (1999), 'State Violence in Guatemala, 1960-1996: A Quantitative Reflection', American Association for the Advancement of Science. Last accessed Oct 20, 2011.
**URL:** *http://shr.aaas.org/guatemala/ciidh/qr/english/en_qr.pdf*

Ball, P., Lynch, M. and Hoover, A. (2007), 'Revisiting "Killings and Migration in Kosovo": responses to additional data and analysis', Benetech: Human Rights Data Analysis Group.
**URL:** *http://www.icty.org/x/file/About/OTP/War_Demographics/en/milutinovic_kosovo_070128.pdf*

Banks, D., Couzens, L., Blanton, C. and Cribb, D. (2015), 'Arrest-Related Deaths Program Assessment', Technical Report by RTI International. Last accessed April 7, 2015.
**URL:** *http://www.bjs.gov/content/pub/pdf/ardpatr.pdf*

Birnir, J. K. and Gohdes, A. (2014), Voting in the Shadow of Violence: Electoral Politics and Conflict. Unpublished manuscript.

Bishop, Y. M., Fienberg, S. E. and Holland, P. W. (2007), *Discrete Multivariate Analysis: Theory and Practice*, Springer.

Bouchard, M. (2007), 'A Capture–Recapture Model to Estimate the Size of Criminal Populations and the Risks of Detection in a Marijuana Cultivation Industry', *Journal of Quantitative Criminology* **23**(3), 221–241.

Brunborg, H., Lyngstad, T. H. and Urdal, H. (2003), 'Accounting for Genocide: How Many Were Killed in Srebrenica?', *European Journal of Population / Revue Européenne de Démographie* **19**(3), 229–248.

Buster, M., van Brussel, G. and van den Brink, W. (2001), 'Estimating the number of opiate users in Amsterdam by capture–recapture: The importance of case definition', *European Journal of Epidemiology* **17**(10), 935–942.

Carpenter, D., Fuller, T. and Roberts, L. (2013), 'WikiLeaks and Iraq Body Count: The Sum of Parts May Not Add Up to the Whole - A Comparison of Two Tallies of Iraqi Civilian Deaths', *Prehospital and Disaster Medicine* **28**(3), 223–229.

Chao, A. (1987), 'Estimating the Population Size for Capture-Recapture Data with Unequal Catchability', *Biometrics* **43**(4), 783–791.

Chojnacki, S., Ickler, C., Spies, M. and Wiesel, J. (2012), 'Event Data on Armed Conflict and Security: New Perspectives, Old Challenges, and Some Solutions', *International Interactions* **38**(4), 382–401.

Clark, W. K. (2001), *Waging Modern War: Bosnia, Kosovo and the Future of Combat*, PublicAffairs, New York.

Comiskey, C. M. and Barry, J. M. (2001), 'A capture recapture study of the prevalence and implications of opiate use in Dublin', *The European Journal of Public Health* **11**(2), 198–200.

Condra, L. N. and Shapiro, J. N. (2012), 'Who Takes the Blame? The Strategic Effects of Collateral Damage', *American Journal of Political Science* **56**(1), 167–187.

Coull, B. A. and Agresti, A. (1999), 'The use of mixed logit models to reflect heterogeneity in capture-recapture studies', *Biometrics* **55**(1), 294–301.

Darroch, J. N., Fienberg, S. E., Glonek, G. F. and Junker, B. W. (1993), 'A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability', *Journal of the American Statistical Association* **88**(423), 1137–1148.

Davenport, C. and Ball, P. (2002), 'Views to a Kill: Exploring the Implications of Source Selection in the Case of Guatemalan State Terror, 1977–1995', *Journal of Conflict Resolution* **46**(3), 427–450.

DeMeritt, J. H. R. (2015), 'Delegating Death: Military Intervention and Government Killing', *Journal of Conflict Resolution* **59**(3), 428–454.

Draper, D. (1995), 'Assessment and propagation of model uncertainty', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 45–97.

Eck, K. (2012), 'In Data We Trust? A Comparison of UCDP GED and ACLED Conflict Events Datasets', *Cooperation and Conflict* **41**(1), 124–141.

Eck, K. and Hultman, L. (2007), 'One-Sided Violence Against Civilians in War: Insights from New Fatality Data', *Journal of Peace Research* **44**(2), 233–246.

Fisher, N., Turner, S. W., Pugh, R. and Taylor, C. (1994), 'Estimated numbers of homeless and homeless mentally ill people in north east Westminster by using capture-recapture analysis', *BMJ* **308**(6920), 27–30.

Fjelde, H. and Hultman, L. (2014), 'Weakening the Enemy: A Disaggregated Study of Violence against Civilians in Africa', *Journal of Conflict Resolution* **58**(7), 1230–1257.

Gill, G. V., Ismail, A. A., Beeching, N. J., Macfarlane, S. B. J. and Bellis, M. A. (2003), 'Hidden diabetes in the UK: use of capture–recapture methods to estimate total prevalence of diabetes mellitus in an urban population', *Journal of the Royal Society of Medicine* **96**(7), 328–332.

Gleditsch, N. P., Wallensteen, P., Eriksson, M., Sollenberg, M. and Strand, H. (2002), 'Armed Conflict 1946-2001: A New Dataset', *Journal of Peace Research* **39**(5), 615–637.

Gohdes, A. and Price, M. (2013), 'First Things First: Assessing Data Quality before Model Quality', *Journal of Conflict Resolution* **57**(6), 1090–1108.

Gohdes, A. R. (2014), Repression in the Digital Age: Communication Technology and the Politics of State Violence, PhD thesis, University of Mannheim.

Guberek, T., Guzmán, D., Price, M., Lum, K. and Ball, P. (2010), 'To Count the Uncounted: An Estimation of Lethal Violence in Casanare', Benetech Human Rights Program. Last accessed October 24, 2011.
**URL:** *http://www.hrdag.org/resources/publications/results-paper.pdf*

Guzmán, D., Guberek, T. G. and Price, M. (2012), 'Unobserved Union Violence: Statistical Estimates of the Total Number of Trade Unionists Killed in Colombia, 1999-2008', The Benetech Human Rights Program. Last accessed October 30, 2012.
**URL:** *https://www.hrdag.org/resources/publications/uv-estimates-paper.pdf*

Guzmán, D., Guberek, T., Hoover, A. and Ball, P. (2007), 'Missing People in Casanare', Benetech.
**URL:** *https://hrdag.org/wp-content/uploads/2013/02/casanare-missing-report.pdf*

Hayden, W. (1999), 'The Kosovo conflict: The strategic use of displacement and the obstacles to international protection', *Civil Wars* **2**(1), 35–68.

Haynes, A., Bower, C., Bulsara, M., Jones, T. and Davis, E. (2004), 'Continued increase in the incidence of childhood Type 1 diabetes in a population-based Australian sample (1985–2002)', *Diabetologia* **47**(5), 866–870.

Hendrix, C. S. and Salehyan, I. (forthcoming), 'No News is Good News?: Mark and Recapture for Event Data When Reporting Probabilities are Less than One', *International Interactions* .

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999), 'Bayesian Model Averaging: A Tutorial', *Statistical Science* **14**(4), 382–401.

Hook, E. B. and Regal, R. R. (1995), 'Capture-recapture methods in epidemiology: methods and limitations', *Epidemiologic reviews* **17**(2), 243–264.

Hoover Green, A. (2011), Repertoires of Violence Against Noncombatants: The Role of Armed Group Institutions and Ideologies, PhD thesis, Yale University.

Hope, V. D., Hickman, M. and Tilling, K. (2005), 'Capturing crack cocaine use: estimating the prevalence of crack cocaine use in London using capture–recapture with covariates', *Addiction* **100**(11), 1701–1708.

Hultman, L., Kathman, J. and Shannon, M. (2013), 'United Nations Peacekeeping and Civilian Protection in Civil War', *American Journal of Political Science* **57**(4), 875–891.

Human Rights Watch (2001), 'Under Orders: War Crimes in Kosovo', New York, Washington, London, Brussels.

Humanitarian Law Centre (2015), 'Kosovo Memory Book, 1998-2000'. Last accessed February 17, 2015.
**URL:** *http://www.kosovskaknjigapamcenja.org*

Iraq Body Count (n.d.).
**URL:** *https://www.iraqbodycount.org/*

Jewell, N. P., Spagat, M. and Jewell, B. L. (2013), MSE and Casualty Counts: Assumptions, Interpretation, and Challenges, *in* T. Seybolt, J. D. Aronson and B. Fischhoff, eds, 'Counting Civilian Casualties: An Introduction to Recording and Estimating Nonmilitary Deaths in Conflict', Oxford University Press, New York, chapter 10, pp. 185–211.

Johndrow, J., Lum, K. and Ball, P. (2015), 'dga: Capture-Recapture Estimation using Bayesian Model Averaging'.
**URL:** *http://cran.r-project.org/web/packages/dga/index.html*

Khan, S. I., Bhuiya, A. and Uddin, A. J. (2004), 'Application of the Capture-Recapture Method for Estimating Number of Mobile Male Sex Workers in a Port City of Bangladesh', *Journal of Health, Population and Nutrition* **22**(1), 19–26.

Krüger, J. and Ball, P. (2014), 'Evaluation of the Database of the Kosovo Memory Book', Human Rights Data Analysis Group. Last accessed February 5, 2015.
**URL:** *https://hrdag.org/wp-content/uploads/2013/01/Evaluation_of_the_Database_-KMB-2014.pdf*

Krüger, J., Ball, P., Price, M. and Hoover Green, A. (2013), It Doesn't Add Up: Methodological and Policy Implications of Conflicting Casualty Data, *in* T. Seybolt, J. D. Aronson and B. Fischhoff, eds, 'Counting Civilian Casualties: An Introduction to Recording and Estimating Nonmilitary Deaths in Conflict', Oxford University Press, New York, chapter 12, pp. 247–264.

Kruse, N., Behets, F. M.-T. F., Vaovola, G., Burkhardt, G., Barivelo, T., Amida, X. and Dallabetta, G. (2003), 'Participatory Mapping of Sex Trade and Enumeration of Sex Workers Using Capture–Recapture Methodology in Diego-Suarez, Madagascar', *Sexually Transmitted Diseases* **30**(8), 664–670.

Lacina, B. and Gleditsch, N. (2005), 'Monitoring Trends in Global Combat: A New Dataset of Battle Deaths', *European Journal of Population/Revue européenne de Démographie* **21**(2-3), 145–166.

Landman, T. (2006), *Studying Human Rights*, Routledge, London and New York.

Landman, T. and Carvalho, E. (2010), *Measuring Human Righs*, Routledge, London and New York.

Landman, T. and Gohdes, A. (2013), A Matter of Convenience: Challenges of Non-Random Data in Analyzing Human Rights Violations during Conflicts in Peru and Sierra Leone, *in* T. Seybolt, J. D. Aronson and B. Fischhoff, eds, 'Counting Civilian Casualties: An Introduction to Recording and Estimating Nonmilitary Deaths in Conflict', Oxford University Press, New York, chapter 5, pp. 77–94.

Larson, A., Stevens, A. and Wardlaw, G. (1994), 'Indirect estimates of 'hidden' populations: Capture-recapture methods to estimate the numbers of heroin users in the Australian capital territory', *Social Science & Medicine* **39**(6), 823–831.

Lum, K., Price, M. E. and Banks, D. (2013), 'Applications of Multiple Systems Estimation in Human Rights Research', *The American Statistician* **67**(4), 191–200.

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**(1), 1–26.

Madigan, D. and York, J. C. (1997), 'Bayesian methods for estimation of the size of a closed population', *Biometrika* **84**(1), 19–31.

Manrique-Vallier, D. (2014), Bayesian population size estimation using dirichlet process mixtures, Technical report, University of Indiana.

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

Manrique-Vallier, D., Price, M. and Gohdes, A. (2013), Multiple Systems Estimation Techniques for Estimating Casualties in Armed Conflicts, *in* T. Seybolt, J. D. Aronson and B. Fischhoff, eds, 'Counting Civilian Casualties: An Introduction to Recording and Estimating Nonmilitary Deaths in Conflict', Oxford University Press, New York, chapter 9, pp. 165–182.

Mastro, T. D., Kitayaporn, D., Weniger, B. G., Vanichseni, S., Laosunthorn, V., Uneklabh, T., Uneklabh, C., Choopanya, K. and Limpakarnjanarat, K. (1994), 'Estimating the number of HIV-infected injection drug users in Bangkok: a capture–recapture method', *American Journal of Public Health* **84**(7), 1094–1099.

Mitchell, S., Ozonoff, A., Lum, K., Zaslavsky, A. M. and Coull, B. A. (2015), Population Size Estimation with Inactive Lists: Multilevel Mixture Models and Missing Data with Applictation to Armed Conflict Data.

Mitchell, S., Ozonoff, A., Zaslavsky, A. M., Hedt-Gauthier, B., Lum, K. and Coull, B. A. (2013), 'A Comparison of Marginal and Conditional Models for Capture–Recapture Data with Application to Human Rights Violations Data', *Biometrics* **69**(4), 1022–1032.

Organisation for Security and Co-operation in Europe (1999), 'KOSOVO/KOSOVA: As Seen, As Told', An analysis of the human rights findings of the OSCE Kosovo Verification Mission, October 1998 to June 1999. Last accessed November 28, 2012.
**URL:** *http://www.osce.org/odihr/17772*

Otto, S. (2013), 'Coding one-sided violence from media reports', *Cooperation and Conflict* **48**(4), 556–566.

Paz-Bailey, G., Jacobson, J. O., Guardado, M. E., Hernandez, F. M., Nieto, A. I., Estrada, M. and Creswell, J. (2011), 'How many men who have sex with men and female sex workers live in El Salvador? Using respondent-driven sampling and capture–recapture to estimate population sizes', *Sexually Transmitted Infections* **87**(4), 279–282.

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.

Raleigh, C., Linke, A. and Dowd, C. (2012), 'Armed Conflict Location and Event Dataset (ACLED): Codebook, Version 2', Trinity College Dublin, University of Colorado, Boulder, and Centre for the Study of Civil War, International Peace Research Institute, Oslo (PRIO). Last accessed April 26, 2012.
**URL:** *http://www.acleddata.com/wp-content/uploads/2012/ACLED_Codebook_2012.pdf*

Raleigh, C., Linke, A., Hegre, H. and Karlsen, J. (2010), 'Introducing ACLED: An Armed Conflict Location and Event Dataset', *Journal of Peace Research* **47**(5), 651–660.

Rivest, L.-P. (2011), 'A lower bound model for multiple record systems estimation with heterogeneous catchability', *The International Journal of Biostatistics* **7**(1), 1–21.

Rivest, L.-P. and Baillargeon, S. (2007), 'Applications and extensions of chao's moment estimator for the size of a closed population', *Biometrics* **63**(4), 999–1006.

Robles, S. C., Marrett, L. D., Clarke, E. A. and Risch, H. A. (1988), 'An application of capture-recapture methods to the estimation of completeness of cancer registration', *Journal of clinical epidemiology* **41**(5), 495–501.

Salehyan, I. (2015), 'Best practices in the collection of conflict data', *Journal of Peace Research* **52**(1), 105–109.

Sekar, C. C. and Deming, W. E. (1949), 'On a method of estimating birth and death rates and the extent of registration', *Journal of the American Statistical Association* **44**(245), 101–115.

Siegler, A., Roberts, L., Balch, E., Bargues, E., Bhalla, A., Bills, C., Dzeng, E., Epelboym, Y., Foster, T., Fulton, L., Gallagher, M., Gastolomendo, J. D., Giorgi, G., Habtehans, S., Kim, J., McGee, B., McMahan, A., Riese, S., Santamaria-Schwartz, R., Walsh, F., Wahlstrom, J. and Wedeles, J. (2008), 'Media Coverage of Violent Deaths in Iraq: An Opportunistic Capture-Recapture Assessment', *Prehospital and Disaster Medicine* **23**(4), 369–371.

Silva, R. and Ball, P. (2006), 'The Profile of Human Rights Violations in Timor-Leste, 1974-1999', A Report by the Benetech Human Rights Data Analysis Group to the Commission on Reception, Truth and Reconciliation. Last accessed May 14, 2014.
**URL:** *https://hrdag.org/wp-content/uploads/2013/02/Benetech-Report-to-CAVR.pdf*

Spagat, M. (2014), 'A Triumph of Remembering: Kosovo Memory Book'. Last accessed February 17, 2015.
**URL:** *http://www.kosovomemorybook.org/wp-content/uploads/2015/02/Michael-_Spagat_Evaluation_of_the_Database_KMB_December_10_2014.pdf*

Sundberg, R., Lindgren, M. and Padskocimaite, A. (2012), 'UCDP Geo-referenced Event Dataset (GED) Codebook Version 1.5-2012', Department of Peace and Conflict Research, Uppsala University. Last accessed December 13, 2012.

Sundberg, R. and Melander, E. (2013), 'Introducing the UCDP Georeferenced Event Dataset', *Journal of Peace Research* **50**(4), 523–532.

Ulfelder, J. and Schrodt, P. A. (2009), 'Political Instability Task Force Worldwide Atrocities Event Data Collection Codebook, Version 1.0B2'. Last accessed April 6, 2015.
**URL:** *http://eventdata.parusanalytics.com/data.dir/Atrocities.codebook.1.0B2.pdf*

Wood, R. M. and Kathman, J. D. (2014), 'Too Much of a Bad Thing? Civilian Victimization and Bargaining in Civil War', *British Journal of Political Science* **44**, 685–706.

Zwierzchowski, J. and Tabeau, E. (2010), The 1992-95 War in Bosnia and Herzegovina: Census-Based Multiple Systems Estimation of Casualties' Undercount, Conference Paper for the International Research Workshop on 'The Global Costs of Conflict', Berlin, p. 25.

# A Procedure for imputing missing information

Even after matching, there are some victims for whom information on the location and/or date of the reported death or disappearance remains missing. Simply dropping these records from the analysis is ill-advised, as this changes the population of all deaths from which the data is drawn to the population of all deaths for which the location and date of death or disappearance is known. If the deaths with unknown date or location are not missing at random, e.g. some regions are more likely to contain deaths with unknown location than others, removing these records from the analysis may bias the estimates inconsistently relative to one another. That is, our ability to discern patterns in the numbers of deaths may be reduced.

Instead, we suggest imputing the missing data conditional on all other known variables, including list inclusion variables. In this analysis, we perform imputation non-parametrically by taking a random sample from the empirical joint distribution of all missing variables conditional on the known variables. In particular, we divide the records into three types: (a) records for which only the date of death is missing, (b) records for which only the location of death is missing, and (c) records for which both the date of death and the location is missing.

For each record that is to have its missing fields imputed, we first identify all records that have non-missing values for the the fields we will impute and share the same list overlap pattern with the record in question. Consider an example victim that appears on the first two lists but not the third and whose location of death is known but the date of death is missing. For this record, we first identify all other records with non-missing date of death that were also reported to the first two lists but not the third. This set of records constitutes the initial pool of records from which we will draw a representative record to "donate" its values to the missing fields of the record we are imputing.

For a record of type (a), from the initial pool of records, we select only the records that also share the same location of death as the record to be imputed. If no such record exists, we revert back to the initial pool. For records of type (b), we follow a similar procedure. Here, we sub-select records that have date of death within one week of the date of death in either direction of the record whose location we are imputing. Again, if there is no record that shares both list overlap structure and date of death to within one week, we revert back to the initial pool described above. For records of type (c), we have no further data on which to reduce the pool of potential matches and select only from the initial pool.

In all cases, having identified the pool of potential "donor" records, we select one record as a match and impute all of the fields that are missing in the record to be imputed with the fields given in the donor record. Table 2 shows the number of records with each class of missingness used in our analysis. Ideally, one would multiply impute, repeating this imputation procedure many times and averaging across the estimates obtained from each imputed dataset to test for the sensitivity of the estimates. Here, we perform imputation only once and leave multiple imputation for a future iteration of our work.

| | date present | date missing |
|---|---|---|
| location present | 2244 | 356 |
| location missing | 122 | 50 |

Table 2: The number of records exhibiting missing date and/or location.

# B Additional descriptives

# C Overlap by municipality and week

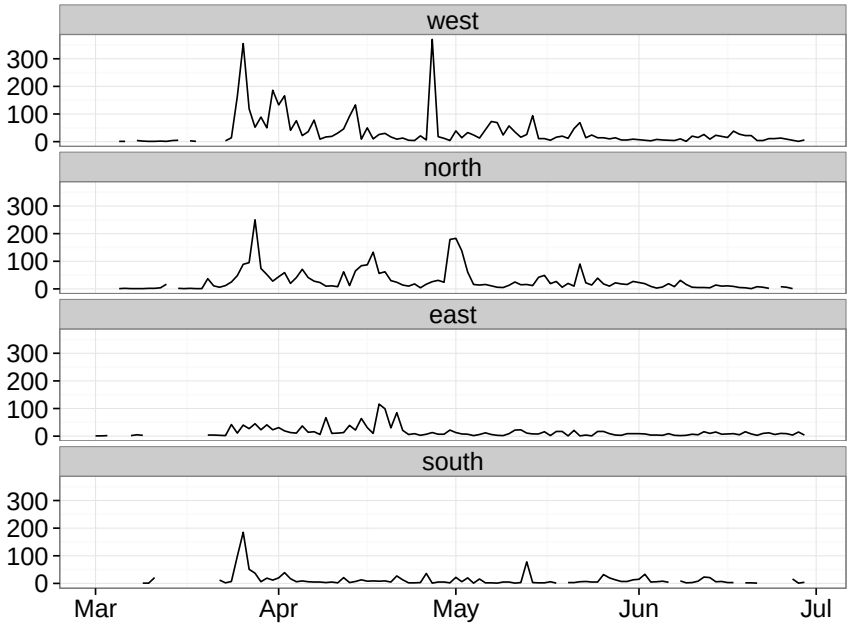
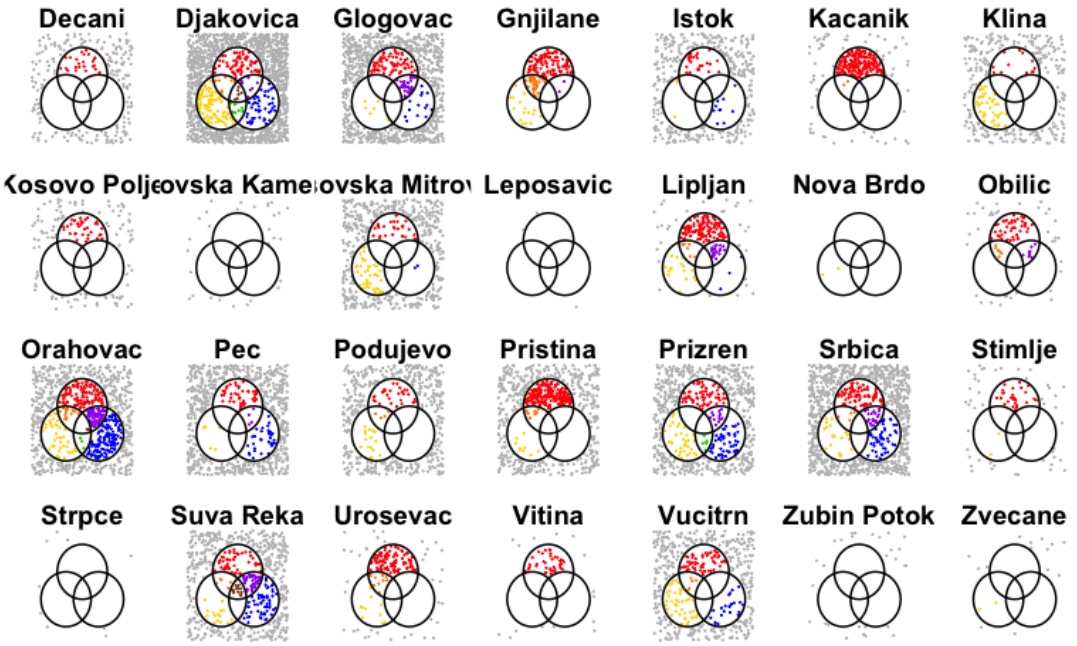
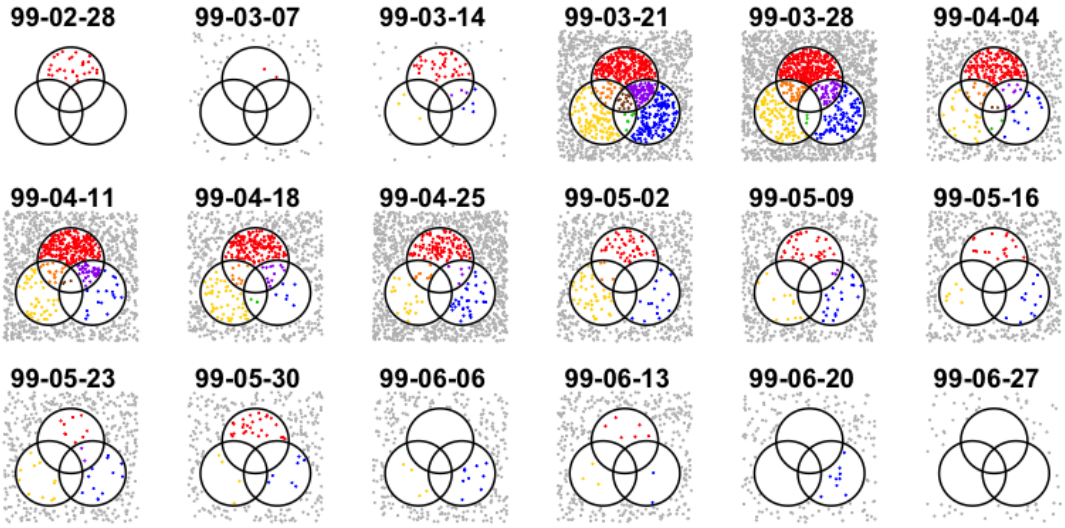



Figure 7: War victims (HLC) over time, by region and day.




(a) By municipality




(b) By week

Figure 8: Overlap and additional HLC records, by counts and stratifications.