Capture-Recapture for Casualty Estimation and Beyond: Recent Advances and Research Directions

Daniel Manrique-Vallier[*], Patrick Ball[†] and Mauricio Sadinle[‡]

November 11, 2019

Abstract

The most basic quantitative question about the consequences of armed conflicts is perhaps how many people were killed. During and after conflicts, it is common to attempt to create tallies of victims. However, destroyed infrastructure and institutions, danger to field workers, and a reasonable suspicion of data collection by victim communities limit the result of these efforts to incomplete and non-representative lists. Capture-Recapture (CR) estimation, also known as Multiple Systems Estimation (MSE) in the context of human populations, is a family of methods for estimating the size of closed populations based on matched incomplete samples. CR methods vary in details and complexity, but they all ultimately rely on analyzing the patterns of inclusion of individuals across samples to estimate the probability of not being observed and then the number of unobserved individuals. In this discussion, we describe the versions MSE with which analysts have estimated the total number casualties in armed conflicts. We explore the advances of the last fifteen years, and we describe outstanding statistical challenges.

1 Introduction

"How many people were killed?" This is perhaps the most basic quantitative question about the consequences of armed conflicts. While many groups attempt to create tallies of victims, these lists are usually subject to incomplete and non-representative registration. Difficulties faced by data-collection efforts include destroyed infrastructure, danger to field workers, suspicion of data collection by victim communities, among others. Tempting as it is, simply pooling existing registries together and eliminating duplicates is unlikely to produce a full enumeration. The result of such an approach can only be taken as an incomplete, nonrepresentative sample with unknown biases, and it can only lead to a lower bound on the total number of casualties.

[*]Department of Statistics, Indiana University, Bloomington, IN
[†]Human Rights Data Analysis Group, San Francisco, CA
[‡]Department of Biostatistics, University of Washington, Seattle, WA

Capture-Recapture (CR) estimation takes advantage of the multiple registries of victims that are often generated during or after a conflict. CR, also known as Multiple Systems Estimation (MSE) in the context of human populations, is a family of methods for estimating the size of closed populations based on matched incomplete samples. CR methods vary in details and complexity, but they all rely on analyzing the patterns of inclusion of individuals in the samples to estimate the probability of not being observed, and then the number of unobserved individuals. CR methods are best known for their application to animal abundance estimation, where they have developed considerably. In the context of conflict casualty estimation, they were first used in Guatemala by Ball (2000), where researchers used three incomplete sources of information, which jointly documented more than 54,000 unique killings, and used CR to estimate the total to be more than 132,000. This analysis and several follow-on projects helped support the case that the Guatemalan Army committed acts of genocide against the indigenous Mayan population (see Ball and Price, 2018).

Several challenges arise when using CR methodologies for casualty estimation. Many of these problems are common to other applications, and there are readily available methodologies to address them. For example, dependence between lists can be addressed using a log-linear CR approach (Bishop et al., 1975). Other problems in casualty estimation differ substantially from other contexts. For example, while several models for controlling individual heterogeneity in capture have been proposed in the ecology literature (Otis et al., 1978), most of them assume a symmetrical form of heterogeneity which is not realistic in our context as lists are not exchangeable.

In this article we discuss the challenges of applying CR methods to the problem of estimating the total number of deaths in armed conflicts, and explore the advances of the last fifteen years in the area. We also describe outstanding challenges and speculate possible solutions.

2 Capture-Recapture in Casualty Estimation: Challenges and Developments

2.1 The Capture-Recapture Approach

Consider a closed population of N individuals, and $J \geq 2$ incomplete lists taken from that population. In this context N will be the unknown number of victims of the armed conflict, and J the number of partial lists available. Let $x_{ij} = 1$ if individual i is recorded in list j and $x_{ij} = 0$ otherwise. We arrange all these indicators into individual-level vectors to form N individual *capture patterns* $\mathbf{x}_i = (x_{i1}, ...x_{iJ})$, one for each element of the population. For example a pattern $\mathbf{x}_i = (0, 0, 1, 1)$ indicates that individual i was recorded by lists 3 and 4, but missed by the first two. We note that even though each individual has a capture pattern, any individual with pattern $\mathbf{0} = (0, ..., 0)$ is by definition unobserved. Our objective is to produce an estimate of how many individuals in the population belong to that class.

Capture-Recapture estimation of N is based on estimating the probability mass function $f(\mathbf{x}|\theta)$ for the capture patterns $\mathbf{x} \in \{0, 1\}^J$ from a sample truncated at $\mathbf{x} =$

0. We then use that model to predict $f(\mathbf{0}|\theta)$, and then N. In order for this to be possible it is necessary, at the very least, that whatever the model $f(\mathbf{x}|\theta)$ is, it can be estimated from the data, which, by definition, will never include the capture patterns **0**. Conversely, even though we cannot observe **0**, the model should make it possible to evaluate $f(\mathbf{0}|\theta)$. The model's other assumptions are mostly related to the specific form of the data generation process, and these will be encoded as specific parametric assumptions in $f(\cdot|\theta)$.

Two assumptions are commonly associated with CR estimation. The first one, *independence*, states that the probability of appearing in one list is not affected by having appeared in another list. The second one, *homogeneity*, requires that this probability distribution is the same for each individual in the population. These two can be expressed as the so-called *independence model*:

$$\mathbf{x}_i \overset{iid}{\sim} \prod_{j=1}^{J} p_j^{x_{ij}} (1-p_j)^{1-x_{ij}}, \quad i = 1, ..., N, \tag{1}$$

where p_j is the probability of appearing in list j. The independence model lies behind the earliest and most famous CR techniques, for example the Petersen estimator,

$$\hat{N} = \frac{n_A \cdot n_B}{n_{AB}}, \tag{2}$$

where n_A and n_B are respectively the number of observed individuals in lists A and B, and n_{AB} is the number of individuals in common between the two.

Independence estimators like Petersen's are still occasionally useful—for instance, when both lists are independent simple random samples from the population—but their assumptions are unrealistic in the casualty estimation setting. Specifically, probabilities of capture tend to vary, sometimes greatly, from individual to individual. From qualitative conversations with victim communities and grassroots human rights activists documenting abuses, we have learned that the two primary factors that affect the probability that an event will be observed are trust and logistics. Interviewers are asking survivors to relate events that are among the most traumatic situations that can happen to anyone. The survivors' willingness to report these events requires them to trust the interviewers. Conversely, if survivors perceive the interviewers as from rival political positions, they may choose not to disclose information to protect themselves. The second major influence on documentation dynamics is the logistical capability of each organization. Can the groups conducting documentation access the affected communities? Much mass violence occurs in remote areas. Groups that have interviewers willing to make arduous journeys may be better able to capture information in those locations. High-resource groups may be able to afford more and more adaptable vehicles, or in the case of the UN, helicopters.

2.2 List Dependence

Violations of the assumptions underlying (1) in the form of dependence between lists are common in casualty estimation. In the original Guatemala analysis (Ball, 2000), researchers observed a form of negative dependence between two of the lists. The

first of these dependent lists was the result of a qualitative investigation that took testimonies among Catholic religious communities conducted in the mid-1990s; the second was gathered by a coalition of NGOs mostly associated with the political left which took testimonies in the early 1990s among communities which had been part of the guerrillas' civilian base. Researchers noted that people in the religious communities that trusted the Catholic researchers would be less likely to report to the NGOs, and vice-versa. In this scenario, individual witnesses prefer one documentation project to another, leading to negative list dependence.

Fienberg (1972) proposed to account for list dependencies through their explicit modeling as list-by-list interactions in log-linear models (see also Bishop et al., 1975). For example, using Bishop et al. (1975) notation, a model accounting for dependence between lists 1 and 2 when three lists are available would be

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)},$$

where $m_{ijk} = E[n_{ijk}]$ and n_{ijk} is the number of individuals with capture pattern $(i, j, k) \in \{0, 1\}^3$. A conditional maximum likelihood estimator for the undercount is given by the formula

$$\hat{n}_{000} = \frac{\hat{m}_{001}\hat{m}_{010}\hat{m}_{100}\hat{m}_{111}}{\hat{m}_{011}\hat{m}_{101}\hat{m}_{110}}. \tag{3}$$

This formula results from a no-second-order interaction assumption in the log-linear model, that is, $u_{123(ijk)} = 0$, necessary to ensure that the model is identifiable and that $f(\mathbf{0}|\theta)$ can be calculated.

Log-linear CR is a mature technology that has been used in several casualty estimation projects; e.g. Kosovo 2002 (Ball et al., 2002), Peru 1980–2000 (Ball et al., 2003), Guatemala 1982–1983 (Ball and Price, 2018), and Bosnia in 1992–1995 (Zwierzchowski and Tabeau, 2010). Nevertheless, this approach has several important limitations, and we will discuss some of them later in this article.

2.3 Heterogeneity

Differences in the probabilities of being listed due to individual traits are referred in the CR context as *heterogeneity of capture probabilities* (or "heterogeneity", for short). As discussed in the introduction, our experience has led us to believe that heterogeneity is the primary problem in CR applied to casualty estimation. Victims and witnesses of violence are subject to individual attributes that affect the listability of victims. We mentioned the problem of the degree of trust that witnesses put in different projects as one reason. Another important aspect is the social visibility of victims. Adults tend to be better known by their communities than children; authorities and famous people tend to be reported more than regular people; victims in remote rural locations tend to be less frequently reported than people in cities. All of these, and other more locally-specific or less describable factors, contribute to the violation of the "equal distribution" assumption in model (1) and need special treatment. We now describe two approaches to deal with the effects of heterogeneity in CR.

The first approach to deal with heterogeneity is stratification (Sekar and Deming, 1949). The idea is to use a discrete covariate that is known or suspected to be related

to the source of heterogeneity to segment the population into homogeneous (or at least more homogeneous) sub-populations, and estimate within them separately. A common choice in this context is place of death. For example in Ball et al. (2003) researchers resorted to expert knowledge for dividing the Peruvian territory into 59 geographic strata roughly corresponding to known insurgent-counterinsurgent conflict dynamics, which were then treated separately. Other typical choices in casualty estimation are perpetrator agent (Ball et al., 2003) and period (Ball et al., 2002).

When properly executed, stratification can greatly help reduce the impact of heterogeneity; however, it also has important limitations. The most obvious one is the reduction of within-strata sample sizes. This reduces inferential power and can lead to identifiability problems. Another one is the need of data for the stratification, and of specialized knowledge about the relationship between heterogeneity and the available covariates.

A second, complementary approach to dealing with heterogeneity is through modeling. Sometimes heterogeneity manifests itself as list dependency and can be dealt with using log-linear and related methods. For example, even though we presented the case of Guatemala as an illustration of list dependency, a closer look reveals that the driver of said dependency were differences of listability due to individual traits, i.e. the level of trust each individual had on each documentation project.

A more direct modeling strategy is to directly represent the individual traits that lead to the differential capturability. This approach was first introduced by Sanathanan (1973), and was greatly developed in the context of animal abundance estimation as the model M_h and its variants (Otis et al., 1978). All these approaches have in common to introduce some form of individual-level random effect ω_i driving capturability:

$$\mathbf{x}_i|\theta, \omega_i \overset{ind}{\sim} f(\cdot|\theta, \omega_i), \qquad \omega_i \overset{iid}{\sim} H.$$

Most of the models developed in the animal estimation literature assume symmetric heterogeneity effects, that is, ω_i affects all lists in the same direction. An example of this structure is the Rasch model (Agresti, 1994; Fienberg et al., 1999). This makes sense in ecology applications: if animals possess characteristics that make them difficult (or easy) to capture in general, they should be so for any trapping occasion.

Symmetric heterogeneity does not hold in casualty estimation. Different documentation projects often have different objectives, capabilities, and sympathies, resulting in different access to different types of victims. This means that the same individual traits ω_i may have different effects on different lists, sometimes in opposite directions. A dramatic case was observed in Peru (Ball et al., 2003). There victims of the Shining Path tended not to be captured by NGOs or the Ombudsman office, while victims of the armed forces tended to be favored by NGOs. In these cases symmetric effect models, like Rasch models, would be inadequate.

Models that allow for less constrained forms of heterogeneity have been proposed for casualty CR estimation. Manrique-Vallier (2016) proposed the use of Dirichlet process mixtures of independence models (NPLCM model). These models have been successfully used for re-analyzing heterogeneous data previously analyzed with log-linear models in Peru (Manrique-Vallier et al., 2019) and Kosovo (Manrique-Vallier, 2016). It has also been used to estimate the total number of people who disappeared in

the final three days of the Sri Lankan civil war (Ball and Harrison, 2018); the number of women held as sexual slaves by Japanese authorities during World War 2 in Palembang, Indonesia (Ball et al., 2018); the number of people killed in drug-related violence in the National Capital Region of the Philippines (Ball et al., 2019a); and the number of social movement leaders killed in Colombia in 2016–2017 (Ball et al., 2019b). We discuss more about them in Section 2.4.

We believe that the use of flexible models that directly address heterogeneity, like NPLCM, is preferable to techniques that address the induced list dependency, like log-linear models. In Manrique-Vallier et al. (2019) we re-analyzed the Peruvian data from Ball et al. (2003) (plus an extra data source) using LCMCR and log-linear models. We noted that in cases in which results from the two approaches diverged, log-linear models were complex and required many interaction terms. We attribute this behavior to the fact that log-linear models can only address heterogeneity through its approximation using interaction terms and the no-highest-order interaction assumption. While in some cases a simple log-linear representation exists (as in the Guatemala and Kosovo cases), in others the necessary models will be highly complex and not identifiable. For a study on the relationship between LCMCR-type mixtures and log-linear models see Johndrow et al. (2017).

2.4 Model Selection

Even after selecting a family of models for CR estimation (e.g. log-linear), it is usually necessary to choose among many competing models. As in any other statistical problem, model selection can be performed using both knowledge about the problem and by formal model selection techniques.

An example of the use of substantive knowledge to guide model selection is presented in (Zwierzchowski and Tabeau, 2010) for Bosnia and Herzegovina. There analysts for the International Criminal Tribunal for the Former Yugoslavia used twelve data sources (including eight enumerations of the names of people reported as dead) in a log-linear model to estimate the total of war casualties. They started by making dual-system estimates between pairs of systems. They noted which pairs seemed to produce plausible estimates, and which lead to substantially greater or lower than the plausible middle estimates. They decided that the pairs of lists that produced greater or lower estimates were those with substantial interactions. In the discussion, they describe how specific pairs of lists might be positively or negatively interacting. For example, they noted two projects that were both based in Sarajevo and both sampled deaths from Sarajevo with greater probability than deaths elsewhere. In the log-linear model, they included all the pairwise log-linear terms for the lists that they argued had substantial interactions. Naturally, such an approach is difficult to justify from a formal statistical point of view.

Formal model selection procedures have been prominently featured in casualty estimation studies using log-linear CR. The earliest of these (e.g. Ball et al., 2002, 2003) relied on exhaustive searches within the space of hierarchical log-linear models, and were conducted based on the minimization of indexes that balanced parsimony with fit, like the BIC or the χ^2/df statistic. Although this is common practice in applied

statistics and in CR in particular, the approach presents some important limitations. First, even though the model search is data-based (and therefore subject to sampling variability), estimation is performed conditioning on the selected model. This neglects the uncertainty associated with the model selection process itself. Second, in many cases equally plausible models can produce substantially different results with no clear way of choosing one over the other. Finally, some families of models (like log-linear with a large number of lists) can be too large to fully evaluate.

Bayesian model averaging (BMA) avoids the model selection issue altogether. Instead of selecting one single "best" model, we average the posterior distributions of interest (in this case over the population size) over all models of a family, weighting by the posterior probability of the models themselves. Lum et al. (2010) used a BMA approach proposed by Madigan and York (1997) to estimate the number of fatal human rights violations in the department of Casanare, Colombia, in the period 1998-2007. The method of Madigan and York (1997) uses BMA in the space of decomposable graphical models, which is itself a sub-family of hierarchical log-linear models. Madigan and York (1997)'s method works in practice because discrete decomposable graphical models allow posterior estimation in closed form, and the number of models is not too large to evaluate provided the number of lists is small. However, BMA can become computationally challenging with large numbers of lists. Furthermore, as decomposable graphical models are a sub-family of hierarchical log-linear models, they also share some of their limitations; in particular, they might not be sufficiently flexible for modeling dependence induced by heterogeneity.

A different approach was taken by Manrique-Vallier (2016), who proposed the use of Dirichlet process mixtures of product-Bernoulli (independence) models. In this case the model is theoretically infinite-dimensional, but has a structure that modulates the complexity of the mixture to what is needed to adjust well to the data. Similarly to BMA, this approach has the advantage of avoiding the model selection problem, but avoids having to deal with several models to begin with. It also has the advantage of being computationally tractable, scaling easily to very large numbers of lists.

Recent advances notwithstanding, the problem of model selection in CR estimation presents a unique challenge. Any formal model selection procedure can only ensure that the models under consideration fit the observed data well enough to some criterion. However, since capture pattern $\mathbf{0}$ is unobservable by definition, there is no way of ensuring that a model that fits the observed part of the data well enough will lead to recovery of the true value of $f_T(\mathbf{0}|\theta_T)$ under the true model $f_T(\cdot|\theta_T)$. This is a well-known problem (see e.g. the discussion section in Bishop et al., 1975, Ch. 6), and the ultimate reason why the non-parametric CR problem is unidentifiable.

2.5 Practical Invisibility: α-Observability

An important assumption for CR estimation is that every individual in the population of interest must have a positive probability of being listed. With perhaps the exception of projects actively refusing to register particular victims or types of victims, this condition is not difficult to meet in the context of casualty estimation: it is implausible that people can die or disappear without anybody at all noticing.

A related but less explored problem is when some individuals' probability of being listed is indeed positive, but very small. In these cases, even though the classical assumption of positive probability is satisfied, some individuals might be practically invisible to the sampling efforts. This phenomenon is specially problematic in heterogeneous populations, where it might be the case that we have several lists with plenty of data from individuals from easily observable sub-populations, but very few or none from less observable groups.

This problem was studied by Johndrow et al. (2019). They noted that this is an intrinsic problem in CR estimation under heterogeneity, and that an important consequence is that the estimation risk of the population size can, in many cases, be unbounded. As a compromise solution they proposed to abandon the objective of estimating the true population size, and re-define the problem as estimating the number of individuals with a probability greater than an arbitrary threshold α of being observed ("α-observable").

3 Some Open Problems and Research Directions

3.1 Models and Extrapolation Assumptions

CR estimation is in its essence an extrapolation problem: use data from capture patterns in $\{0,1\}^J \setminus \{\mathbf{0}\}$ to estimate $f(\mathbf{0}|\theta)$. Since $\mathbf{0}$ is unobservable by definition, the way to project to this probability will be completely determined by the model $f(\cdot|\theta)$; this also means that truly non-parametric CR estimation is essentially impossible; see discussion in Manrique-Vallier (2016). Conversely, the way in which the probability $f(\mathbf{0}|\theta)$ relates to the rest of $f(\mathbf{x}|\theta)$, $\mathbf{x} \neq \mathbf{0}$, can neither be learned from data nor tested. The projection of the joint model to the unobservable part is related to the concept of *extrapolation distribution* in the missing data literature (Hogan and Daniels, 2008).

Since statistical inference on the way in which observable patterns relate to the unobserved is impossible, selecting an appropriate model or family of models should be done in a way that best resembles the actual data generation process and with understanding of the implied extrapolation assumption. An important example are hierarchical log-linear models. As discussed in Section 2.2, log-linear models for J lists require an assumption of no $(J-1)$th-order interaction in order to be identifiable from data with capture patterns $\{0,1\}^J \setminus \{\mathbf{0}\}$. This condition itself defines the extrapolation assumption (from which the estimator in (3) is derived). The question then becomes: is this particular way of extrapolating reasonable for casualty estimation?

As explained in Sections 2.3, we believe that in most casualty estimation problems, heterogeneity is the main driver of departures from the independence model and so, with some exceptions, log-linear models are just an approximation to the true joint distribution of data. Therefore, even if the models fit the observed data well, the extrapolation assumption might not be appropriate for this problem. On the other hand, models that directly represent plausible heterogeneity structures, like NPLCM, might be more appropriate. Which models and extrapolation assumptions are better

for different scenarios in casualty estimation is an open question that would benefit from additional research.

3.2 Data-Based Stratification

Stratification is often used as a first approach to tackle heterogeneity (see Section 2.3). The usual practice consists in using qualitative expert knowledge to find a partition of the population that could result in homogeneous sub-populations, and estimate within each of them separately. Oftentimes, after trying a stratification scheme, some strata will still exhibit signs of residual heterogeneity. In these cases researchers sometimes revise the stratification scheme, adjust, and try again. For example, in Peru (Ball et al., 2003) researchers determined regional conflict dynamics and stratified accordingly. Then, after noting that model fitting in some of the regions was poor, they sub-divided them into smaller pieces forming a finer stratification scheme.

This iterative procedure seems natural and intuitive, but is statistically problematic. Specifically, the process of looking at the results obtained under a stratification scheme to modify it, is itself a data-based decision that is likely to alter the validity of inferences—similar to the so-called p-hacking problem (Gelman and Loken, 2013). Manrique-Vallier et al. (2019) noted this problem in their re-analysis of the Peruvian data. They addressed it using a partial blinding procedure: two of the authors performed the calculations without sharing the results with the third, while the latter proposed sub-stratification schemes only based on external qualitative knowledge. This procedure partially addressed the risk of cherrypicking results based on what the researchers would want to see. However, the selection of which regions to sub-divide was still based on data-based evaluations of model fitting.

A possible alternative is to formally incorporate the stratification process into the modeling and estimation procedures. Let $\mathcal{Y} = \{y_1, y_2, ..., y_M\}$ be the finest partition of the population we are willing to consider, determined from subject matter knowledge. Let us call these partitions *atomic strata*. Taking $\mathcal{Y}$ as the stratification scheme is equivalent to fitting M models $f(\cdot|\theta_{y_1}), ..., f(\cdot|\theta_{y_M})$ to each atomic stratum. On the other extreme, we can think of unstratified estimation as making the parameters of all M models equal, i.e. $\theta_{y_1} = \theta_{y_1} = ... = \theta$. In between, we can represent different stratification schemes as different agglomerations of atomic strata, where parameters are equal. For example, if we wanted to create a stratum that combines strata 1, 2, and 3, we would represent it by enforcing the restriction $\theta_{y_1} = \theta_{y_2} = \theta_{y_3}$. Using this idea we can think of performing simultaneous estimation of the stratification scheme and CR parameters (including the population size) by specifying prior distributions that put positive mass into relevant groupings of atomic strata by enforcing equality on their parameters. This idea is similar to the method of Price et al. (2019) for the automatic combination of categories in logistic regression. This construction can also allow enforcing meaningful structures, like geographic or temporal contiguity by appropriately allowing equality among neighboring atomic strata.

3.3 Missing Data

As noted by Fienberg and Manrique-Vallier (2009), CR can be seen itself as a missing data problem. Indeed, many estimation methods are based on data- or sample-augmentation schemes that represent unobserved individuals as missing records—see e.g. Manrique-Vallier (2016). This makes it natural to combine CR with other forms of missing data problems and methods.

A frequent scenario in casualty estimation is when data for stratification is missing for some individuals. For example, in the study of the Peruvian conflict (Ball et al., 2003; Manrique-Vallier et al., 2019), about 10% of the records missed perpetrator attribution. As noted by Zwane and van der Heijden (2007)—who studied the problem for the special case of variables completely missing in some of the lists—, in these cases the common practices of ignoring incomplete covariates, creating a special category out of them, or imputing "reasonable" values can be a source of either biases or too optimistic precision.

Manrique-Vallier et al. (2019) proposed a framework for Bayesian stratified CR estimation with incomplete stratification information in one covariate. They combined it with the model from Manrique-Vallier (2016) and used the resulting method to estimate deaths in the Peruvian conflict. The method is based on using a data-augmentation representation for both the unobserved individuals and the missing stratification which is then estimated using Markov Chain Monte Carlo simulation. At its core this method is based on a Missing at Random assumption (Little and Rubin, 2002) whereby the information used to infer the missing stratification is obtained from records with similar capture patterns. A natural extension of this idea is to complement the information from the capture patterns with other variables. For example, in Manrique-Vallier et al. (2019) researchers had access to covariates that were not used in stratification (like age) which might be related to the missing stratification labels and could be used to better estimate them.

3.4 Data Copying Between Lists

An important exception to our belief of heterogeneity being the main driver of dependence between lists in casualty estimation is the case of sharing or copying of records between documentation projects. In these cases, in addition to gathering first-hand information about casualties of a conflict, some projects directly incorporate data obtained by other projects into their listings. In our experience this is not a prevalent problem across the casualty estimation projects in which we have been involved. However, when it happens, its effect is noticeable. One example occurs in the conflict in Syria, where the Human Rights Data Analysis Group (HRDAG) has longitudinal access to lists put together by different projects. The databases are shared multiple times over time, as they are updated when new deaths are known and when additional information about previously reported deaths becomes available. In some cases, the overlap between two lists increases substantially between updates, where the newly-overlapping records are found not to be present in one of the databases in the previous iteration, and the new records match exactly records in the other database. HRDAG, in conversation with one of the groups, learned that they copy published records from

the others. This is a reasonable strategy for a group trying to maintain a comprehensive list, but it creates a strong positive dependence between the lists.

Copying of records between lists that also directly gather first-hand information is problematic in CR because it superimposes and confounds two data generation processes: the capture of individuals by documentation projects, and the relationship between those projects. From these, only the former process is useful for inferring the population size. Thus we need to somehow disentangle them. An ideal situation is that projects record the source(s) of each record so that we can identify which records have only been copied and remove them prior to statistical analysis. In the absence of such information we may try to model the copying process. This strategy will likely require external sources of information and/or strong and untestable assumptions to overcome unidentifiability. One of such possible additional sources of information can result from integrating the CR estimation and the record-linkage process.

3.5 Internal Duplication

Typical multi-list CR methods (like all the ones that have been used for casualty estimation studies so far) only work with information about presence or absence in lists, in the form of vectors in $\{0,1\}^J \setminus \{\mathbf{0}\}$. These vectors are usually the result of J-way record linkage among J lists, where individual lists are assumed to be free of duplicated records. In fact, in most projects an important part of the data preparation is making sure that the internal duplication within lists has been eliminated.

Internal duplication within lists carries plenty of useful information that can be lost during the "cleanup" process. In the same way in which the presence of an individual in more than one list is usually interpreted as an indication of a higher probability of being observed, repeated presence *in the same list* (or "duplication") can also contribute to the same conclusion. To take advantage of these data we need to create methods in which the multivariate capture patterns are not simply strings of zeros or ones, but of natural numbers, $\mathbf{x} \in \{0,1,2...\}^J$. A simple version of such a model, assuming independence between two lists, has been proposed by Lerdsuwansri and Böhning (2018). However the casualty estimation context is likely to require much more sophisticated multi-list models that represent plausible data generation scenarios, and that can be integrated with other sources of information. An additional level of complication comes from the fact that, in practice, there will be uncertainty on which records are duplicates within a single dataset, that is, the counts $\mathbf{x} \in \{0,1,2...\}^J$ will be known with error (see, e.g., Sadinle, 2014; Steorts et al., 2016).

3.6 Record Linkage Errors

The capture patterns $\mathbf{x}_i \in \{0,1\}^J \setminus \{\mathbf{0}\}$ are the essential input for all CR techniques. To obtain these we need to identify individuals that appear in multiple lists by linking their corresponding records. In the context of armed conflicts, witnesses or victims of violence may report an event to different organizations at different points in time and with different degrees of detail. Unfortunately, reporting or collecting unique identifiers, such as national identification numbers, is rare in this context. This means

that even the more basic question of how many unique casualties were reported to any one group cannot be easily answered, as it is often difficult to determine which records belong to the same individuals.

Probabilistic record linkage techniques (see, e.g., Fellegi and Sunter, 1969; Winkler, 1988; Jaro, 1989; Larsen and Rubin, 2001; Sadinle and Fienberg, 2013; Steorts et al., 2016; Sadinle, 2017) take advantage of imperfect partial identifiers collected on the individuals, such as names and demographic information, dates and locations of the events. These pieces of information are usually subject to typographical and other types of errors, which lead to uncertainty in the correct way of linking the records. The result of the record linkage process will typically contain errors termed *false links* and *false non-links*, that is, records that were incorrectly linked and records that were incorrectly left unlinked, respectively. A false non-link can for example lead to a true capture pattern $(0, 0, 1, 1)$ being incorrectly registered as two capture patterns $(0, 0, 0, 1)$ and $(0, 0, 1, 0)$; conversely, a false link can for example lead to two capture patterns $(0, 0, 0, 1)$ and $(0, 0, 1, 0)$ being incorrectly counted as $(0, 0, 1, 1)$. Similar errors appear when record linkage techniques are used for duplicate detection within each list.

The effect of linkage errors is clearly seen in the Petersen estimator (2) in the case of two lists. Between the lists, false links will lead to higher n_{AB} and thereby lower population size estimates, whereas false non-links will lead to lower n_{AB} and higher population size estimates. Within each list, false links will lead to lower n_A and n_B and therefore lower population size estimates, whereas false non-links will lead to higher n_A and n_B and therefore higher population size estimates. For multiple lists, the specific impact of linkage errors will depend on the models being used.

Broadly speaking, the output of a linkage procedure can be seen as an estimator for the underlying correct way of linking the records. As every estimation procedure, the linkage is subject to sampling variability, and we are interested in "transferring" this "linkage uncertainty" into the population size estimation, with the goal of having final estimates that reflect the fact that the linkage is subject to error. Two strategies come to mind: a joint modeling strategy for both the linkage and the population size estimation, and a two-stage strategy where the output of probabilistic linkage is fed into the population size estimation. The first approach has been undertaken by Liseo and Tancredi (2011) and Tancredi and Liseo (2011), who created a joint Bayesian modeling strategy that combines a model for record linkage with a model for population size estimation; although their work focuses on the case of two lists, their strategy could in principle be extended to more general models. The second approach was undertaken by Sadinle (2018), who proposed a procedure called *linkage-averaging*, where the linkage and the population size estimation can be carried out in two separate stages, while still leading to proper Bayesian inferences under some conditions.

A characteristic of the joint modeling strategy is that the analyst will have to run the record linkage and the CR model jointly for each different CR model, which can be computationally intensive, whereas in the two-stage strategy the results from record linkage can be reused with different CR models. Another characteristic of both of these approaches is that their success is determined by the success of their record linkage and CR components. For example, if the record linkage model over-links or under-links, then the population size estimates will be lower or higher, respectively, with respect to

what we would obtain under the correct linkage, regardless of whether one uses a joint model or a two-stage approach. Similarly, if the model for population size estimation is wrong, our estimates will be deficient regardless of whether one uses that model in a joint model or in a two-stage approach. Further research should be devoted to better understanding the properties of these strategies and to develop alternatives.

4 Final Comments

Our goal in this discussion was not to be exhaustive but rather to present some of the challenges, approaches, and directions we are most familiar with. CR for casualty estimation could benefit from developments in many other areas of statistics, such as model selection in regression problems, post-selection inference, small area estimation, and spatio-temporal modeling, just to name a few. Furthermore, CR techniques that are developed for estimating animal abundance in ecology, for corrections to census enumerations, or for disease prevalence estimation in epidemiology will also continue to be potentially useful for casualty estimation.

CR for casualty estimation is an area of research posing several technical challenges that have traditionally been bypassed in applications via ad-hoc solutions. More adequate solutions should account for the uncertainty in the the correct ways of modeling, extrapolating, stratifying, handling missing data, and deduplicating and linking records. Unfortunately, the flip side of better handling of these issues is that we will necessarily obtain casualty estimates with much broader uncertainty intervals. This can potentially mean that in certain situations the intervals will become too large to be practically useful. Nevertheless, it is desirable to have estimation methodologies that provide us with honest assessments of uncertainty and thereby avoid misleading and overconfident results.

References

Agresti, A. (1994), "Simple capture-recapture models permitting unequal catchability and variable sampling effort," *Biometrics*, 50, 494–500.

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

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., Dudukovic, J., and Asher, J. (2002), "Killings and Refugee Flow in Kosovo, March–June, 1999," Report to ICTY.

Ball, P., Coronel, S., Padilla, M., and Mora, D. (2019a), "Drug-Related Killings in

the Philippines," Tech. rep., Human Rights Data Analysis Group and the Stabile Center for Investigative Journalism.

Ball, P. and Harrison, F. (2018), "How many people disappeared on 1719 May 2009 in Sri Lanka?" Tech. rep., Human Rights Data Analysis Group and the International Truth and Justice Project.

Ball, P. and Price, M. (2018), "The statistics of genocide," *CHANCE*, 31, 38–45.

Ball, P., Rodrguez, C., and Rozo, V. (2019b), "Asesinatos de líderes sociales en Colombia en 20162017: una estimación del universo," Tech. rep., Human Rights Data Analysis Group and Dejusticia.

Ball, P., Shin, E. H.-S., and Yang, H. (2018), "There may have been 14 undocumented Korean comfort women in Palembang, Indonesia," Tech. rep., Human Rights Data Analysis Group and Transitional Justice Working Group.

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.

Fellegi, I. P. and Sunter, A. B. (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, 64, 1183–1210.

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.

Fienberg, S. E. and Manrique-Vallier, D. (2009), "Integrated methodology for multiple systems estimation and record linkage using a missing data formulation," *AStA-Advances in Statistical Analysis*, 93, 49–60.

Gelman, A. and Loken, E. (2013), "The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time," Unpublished paper.

Hogan, J. W. and Daniels, M. J. (2008), *Missing Data in Longitudinal Studies*, Boca Raton: Chapman and Hall.

Jaro, M. A. (1989), "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, 84, 414–420.

Johndrow, J. E., Bhattacharya, A., and Dunson, D. B. (2017), "Tensor decompositions and sparse log-linear models," *Annals of statistics*, 45, 1.

Johndrow, J. E., Lum, K., and Manrique-Vallier, D. (2019), "Low-risk population size estimates in the presence of capture heterogeneity," *Biometrika*, 106, 197–210.

Larsen, M. D. and Rubin, D. B. (2001), "Iterative Automated Record Linkage Using Mixture Models," *Journal of the American Statistical Association*, 96, 32–41.

Lerdsuwansri, R. and Böhning, D. (2018), "Extending the Lincoln-Petersen Estimator when Both Sources are Counts," in *Capture-Recapture Methods for the Social and Medical Sciences*, eds. Böhning, D., Van Der Heijden, P. G., and Bunge, J., Boca Raton, FL: Chapman & Hall/CRC, chap. 23, pp. 341–360.

Liseo, B. and Tancredi, A. (2011), "Bayesian Estimation of Population Size via Linkage of Multivariate Normal Data Sets," *Journal of Official Statistics*, 27, 491–505.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data: Second Edition*, New York: John Wiley & Sons.

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.

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

Manrique-Vallier, D. (2016), "Bayesian Population Size Estimation Using Dirichlet Process Mixtures," *Biometrics*, 72, 1246–1254.

Manrique-Vallier, D., Ball, P., and Sulmont, D. (2019), "Estimating the Number of Fatal Victims of the Peruvian Internal Armed Conflict, 1980-2000: an application of modern multi-list Capture-Recapture techniques," .

Otis, D. L., Burnham, K. P., White, G. C., and Anderson, D. R. (1978), "Statistical inference from capture data on closed animal populations," *Wildlife monographs*, 3–135.

Price, B. S., Geyer, C. J., and Rothman, A. J. (2019), "Automatic Response Category Combination in Multinomial Logistic Regression," *Journal of Computational and Graphical Statistics*, 28, 758–766.

Sadinle, M. (2014), "Detecting Duplicates in a Homicide Registry Using a Bayesian Partitioning Approach," *Annals of Applied Statistics*, 8, 2404–2434.

— (2017), "Bayesian Estimation of Bipartite Matchings for Record Linkage," *Journal of the American Statistical Association*, 112, 600–612.

— (2018), "Bayesian propagation of record linkage uncertainty into population size estimation of human rights violations," *Annals of Applied Statistics*, 12, 1013–1038.

Sadinle, M. and Fienberg, S. E. (2013), "A Generalized Fellegi-Sunter Framework for Multiple Record Linkage With Application to Homicide Record Systems," *Journal of the American Statistical Association*, 108, 385–397.

Sanathanan, L. (1973), "A comparison of some models in visual scanning experiments," *Technometrics*, 15, 67–78.

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, 101–115.

Steorts, R. C., Hall, R., and Fienberg, S. E. (2016), "A Bayesian Approach to Graphical Record Linkage and Deduplication," *Journal of the American Statistical Association*, 111, 1660–1672.

Tancredi, A. and Liseo, B. (2011), "A Hierarchical Bayesian Approach to Record Linkage and Size Population Problems," *Annals of Applied Statistics*, 5, 1553–1585.

Winkler, W. E. (1988), "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage," in *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 667–671.

Zwane, E. and van der Heijden, P. (2007), "Analysing capture–recapture data when some variables of heterogeneous catchability are not collected or asked in all registrations," *Statistics in Medicine*, 26, 1069–89.

Zwierzchowski, J. and Tabeau, E. (2010), "The 1992-95 War in Bosnia and Herzegovina: Census-Based Multiple System Estimation of Casualties' Undercount," *Berlin: Households in Conflict Network and Institute for Economic Research*, 539.