Title
Estimating the size of unobserved populations in human rights: Problems in Syria and El Salvador

Author
Mejia, Robin Krieger

Publication Date
2016

Peer reviewed|Thesis/dissertation

Estimating the size of unobserved populations in human rights: Problems in Syria and El Salvador

by

Robin Krieger Mejia

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biostatistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Nicholas P. Jewell, Chair
Professor Mark van der Laan
Professor Barbara Abrams

Spring 2016

Estimating the size of unobserved populations in human rights: Problems in Syria and El Salvador

Abstract

Estimating the size of unobserved populations in human rights: Problems in Syria and El Salvador

by

Robin Krieger Mejia

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Nicholas P. Jewell, Chair

In this dissertation, I examine two human right estimation problems.

First, I assess data on child abductions from El Salvador's civil war. Between 1979 and 1992, El Salvador was wracked by conflict between leftist guerrilla groups and right-wing nationalist governments. One feature of the conflict was the abduction of children by government military forces, or the forced surrender of children to those same forces. Since 1994, La Asociación Pro Busqueda de Niñas y Niños Desaparecidos has investigated cases of these child abductions. To date, they have opened more than 950 cases and located nearly 400 abducted children (now, young adults). The organization remains active, and new cases come to light each year. In Chapter 2, I examine Pro Busqueda's data, assessing what can be said to date about the total as yet unknown number of abductions that occurred. I demonstrate that more abductions occurred than the number of currently known cases discuss capture-recapture estimates under a range of assumptions about the data available today. I then lay out a plan for updating estimates as new data becomes available.

Then, I examine current data on deaths from the ongoing conflict in Syria. Early in the conflict, the United Nations Office of the High Commissioner for Refugees (UNOHCR) contracted with statisticians at the Human Rights Data Analysis Group (HRDAG) to analyze data from multiple human rights groups that were documenting deaths from the conflict there. HRDAG produced three reports from the United Nations and has maintained ongoing relationships with the local human rights groups that are collecting the raw data. HRDAG is now in the unusual position of possessing a series of multiple "snapshots" of each group's data, collected at a number of points between 2012 and 2016. Using those snapshots, I examine how each group's data is changing over time, and discuss how those changes can impact resulting estimates of unreported deaths, showing that the changes can result in estimates for a single governorate that vary by nearly 100,000. In addition, I take advantage of the large number of processed cases to assess the performance of a variety of classification algorithms in determining whether two records refer to the same individual.

For Deborah Bogen

Contents

List of Figures

List of Tables

Acknowledgments

This work was only possible with the support of a large number of people.

I owe a huge debt to Nick Jewell, who supported my decision to pursue this degree and who patiently advised my work on a variety of projects as I worked my way slowly towards the three presented here. Mark van der Laan and Barbara Abrams both provided insight and feedback on this work. I've benefited from small courses with a truly incredible set of faculty: Sandrine Dudoit, Alan Hubbard, Jon McAuliffe, Maya Petersen, and Steve Selvin, many, many thanks. Marco Carone, I so wish you still had office hours within walking distance.

This work absolutely would not have been possible without the support of my collaborators, as well. Megan Price and Patrick Ball have shared generously both their time and data, and I have benefited from getting to know everyone at the Human Rights Data Analysis Group. The staff at La Asociación Pro Busqueda de Niñas y Niños Desaparecidos continue to amaze me, and I am grateful for the University of California Human Rights Fellowship that enabled me to visit Pro Busqueda's offices.

When I found out I was accepted to Berkeley for my MPH, one of the first things I did was contact Eric Stover at the Human Rights Center. Working for HRC shaped my thinking about the role of research in understanding – and responding to – conflict. Eric, Camille Crittenden, Alexa Koenig, Kim Thuy Seelinger, Julie Freccero, Stephen Cody, Cristian Orrego, Alexey Berlind, Melissa Carnay, Patrick Vinck and Phuong Pham – I hope our paths continue to cross.

On visting day for new admits to the Berkeley Biostatistics program, we always say that students learn as much from each other as from the faculty, and I've certainly found that to be true. A special thanks to the people who worked through problems sets or final projects with me: Alison Cohen, Rose Kagawa, Ellen Leucke, Jade Benjamin-Chung, Jeremy Coyle, Erin Ledell, Oleg Sofrygin, Alex Leudtke, Sam Lendle, Sara Moore, Lucia Petito, Luca Pozzi, and Kelly Street.

It's been a special pleasure to be able to call my sister, epidemiologist Wendy King, for advice, both on my research and career plans, and, at 42, to come home to care packages from my parents, Deborah and Jim Bogen. All of my family and friends have been incredibly supportive through this process. I couldn't have done it without you.

Chapter 1

Introduction

1.1 The need for evidence in human rights investigations

Wars (and civil conflicts and police actions) kill large numbers of individuals. However, it's often extremely difficult to determine the total number of killings. This comes as a surprise to researchers in many fields, which is understandable, as death seems like a very quantifiable event. After all, every death produces a body. However, war creates chaos. Professional militaries often maintain good records of the fates of their own soldiers, but deaths of civilians or irregular combatants may not be recorded.

Not all bodies are recovered from battles or bombings. In some cases mass graves are only uncovered years after an event. Even deaths that are given more traditional ceremonies may not make it into official records. Many countries only maintain full birth and death registration in capital cities; births and deaths in the countryside often receive no legal recognition. This is true in much of sub Saharan Africa, for example. (Lack of birth and death registration presents a number of challenges for both individuals and countries; here we only note its effects on attempts to understand the effects of conflict.) Just as important, in countries that do have full vital registration prior to the onset of conflict, such as Iraq, the systems tend to break down during wars. When cities are being bombed, other needs often trump the maintenance of death registration systems. Accounting for other violations of human rights that occur during wars and civil conflicts can be even more challenging. Many human rights violations are activities for which there are no official reporting practice in peacetime. Some horrific violations, such as rape or dismemberment, may make international news; however, it can be hard to determine whether reported incidents are isolated or part of a broader pattern. Other problems, such as forced displacement, may place a large burden on those affected yet receive less widespread attention.

That is not to say that there is no data on human rights violations in conflict. The resolution of civil conflicts now often involves a formal peace and reconciliation process. Such processes can involve documenting testimonials of large numbers of crimes in the immedi-

ate aftermath of a conflict. Truth and reconciliation commissions may collect individual testimonies from survivors. They may also attempt to investigate large scale atrocities or document patterns. There are organizations that collect data during conflicts as well, including international human rights organizations such as Amnesty International and Human Rights watch, and local non-governmental organizations. These organizations frequently have goals that involve minimizing harm to civilian populations and ending conflicts, and the collection and dissemination of data on deaths and other violations may be a part of efforts to engage the international community with those goals. Humanitarian organizations collect such data as well, in order to assess the needs of the population. News organizations also document killings and other violations. However, none of these groups is capable of capturing information on every death or other human rights violation that occurs, nor is that usually the goal of the work.

While none of these groups can record information on all violations that occur, a full accounting of atrocities would be useful. During a conflict, information on how many people are being killed, who they are, and where and when the killings are taking place can help shape responses. This is true for other crimes, as well. Humanitarian organizations can best target their work when they know who is being harmed, where, and in what way. After conflicts, such information can also help shape responses. Such responses can still be humanitarian, including targeting of resources for rebuilding, as well as historical, as part of establishing an accurate record of what happened. Additionally, establishing an evidence base is crucial for accountability. Individuals involved in perpetrating atrocities may be charged with crimes in national or, more likely, international criminal proceedings. In these situations, questions about killings and human rights abuses often revolve around determining magnitude and patterns. To address these questions, we often want to know how many people total were killed, and when and where those killings took place, or if one ethnic group or gender was more victimized than another.

The increasing availability of some information has led many to look for answers to these questions in the available data. Such work can be valuable, but it can also be misleading. On the one hand, publicizing available data can draw needed attention to a conflict and demonstrate that high levels of violence are occurring. The danger lies in attempts to document patterns using incomplete data. Most data collected by NGOs, news organizations and individuals during a conflict is not going to be representative of the total violations occurring. It can be tempting to generate maps or timelines that show differences in reported crimes. However, the true number of crimes is often different from the reported number in ways that are not proportional to time or geographic region. This means that such graphical representations of data can be misleading, as most readers will assume they represent trends in total crimes over time and/or space.

Data collected by means other than sample surveys or a comprehensive census will inevitably be a biased sample of the total population of violations. A local NGO may be be better known in the area in which it's based and therefore more likely to collection information in that region. Even large international organizations may receive more testimonies from individuals who live near their offices, simply because it's easier to for people close by

to reach them. Additionally, a group may be aligned (or be perceived to be aligned) with a certain faction, and individuals who sympathize with that faction may be more likely to talk with the group, while those who oppose that faction may avoid that group. At the beginning of a conflict, no one may yet be collecting data, so early crimes may not be recorded. Interest may peak as violence worsens, increasing interest in data collection, then interest may wane as a conflict continues and "news fatigue" sets in. Interestingly, even if interest increases as violence does, it is often harder to collect data as the situation worsens, so the percentage of crimes documented may decrease. Together, these issues explain why data collected during conflict in a non-probabilistic way will not tell the entire story of the violence.

Such biases don't mean the information collected by human rights and humanitarian organizations is bad. To the contrary, such information can be quite powerful, highlighting specific atrocities and drawing attention to the conflict and need for response. Problems arise when researchers fail to recognized the potential for bias in the data and infer patterns in the underlying population as if it were a probability sample. (For discussion of an example from Syria, see [25].)

These issues highlight the importance of bringing a statistical perspective to the analysis of conflict data. Statistical methods can be brought to bear on questions of the magnitude and patterns of human rights abuses in conflicts. Such methods can be descriptive, summarizing available data in meaningful ways that note the limitations of what conclusions can be drawn from available summaries. Additionally,the data can sometimes be modeled to enable statistical inference, providing estimates of total crimes committed, and patterns therein .

Examples of descriptive work include analyses state violence in Chad conducted by the Human Rights Data Analysis Group (HRDAG) [52], or the same group's work enumerating observed conflict deaths in Syria [45, 46, 44]. Such numbers can be important in demonstrating that a large number of crimes occurred, or in supporting a specific charge such as genocide or command responsibility. However, they don't tell the whole story of a conflict, a fact the reports note.

In any conflict, the number of deaths or other violations that occur is almost never known in real time. Large battles may be reported, but the number of civilian casualties is often a matter of contention. Smaller battles may receive less attention, and conflicts that don't feature western actors may receive less attention in the international press. Furthermore, war is a horrifically messy business. Professional militaries are generally good at keeping track of their own soldiers, but the affect of conflict on the enemy, and also civilian populations, is not necessarily tabulated in real time. Such a tabulation would be hard to do. Additionally, it may not be in the interest of warring parties to advertise civilian casualties. Furthermore, many modern conflicts feature non-state forces, which may have limited record-keeping ability. As a result, there have been debates over death tolls in many recent conflicts, including in the Democratic Republic of Congo, the Darfur, and even Iraq [55, 42]

1.2 Users of human rights data

In recent decades, many countries have established truth and reconciliation commissions, primarily after civil conflicts, with the goal that establishing the truth of what happened is instrumental to healing form the conflict, and that such a reckoning will hopefully forestall future aggression. Such commissions vary in composition and mission, but their work generally involves collecting data to determine what happened during the conflict. As a part of this work, truth and reconciliation commissions frequently solicit testimony from victims and perpetrators. Additionally, commissions in Peru, Sierra Leone, and Timore-Leste, have hired statisticians to assess the data that was collected, and sometimes to conduct analyses designed to estimate the total scope of violations and killings that occurred, even those for which they do not have direct testimonials.

Statistical analyses of human rights data can play also role in prosecutions. In recent years, international criminal tribunals such as International Criminal Tribunal for the former Yugoslavia and the International Criminal Court have brought charges against high ranking officials for atrocities committed during conflicts, primarily for genocide, crimes against humanity, and war crimes. More recently, domestic courts have begun to take on such efforts as well, such as the trial of Rios Montt in Guatemala. International war crimes, such as genocide and crimes against humanity are formally defined in international conventions [23]. Prosecuting such crimes frequently requires demonstrating the scope of crimes committed and also patterns of violence.

This is where statisticians come in. Statisticians can clearly summarize available data and determine what can be said from such descriptions. Additionally, statisticians have a number of tools at their disposal that can let them move beyond summarizing available data to creating a fuller (inferential) picture of events. And statisticians are uniquely qualified to understand and address questions about scope and pattern, which can be instrumental in demonstrating that specific groups were targeted. For example, in Guatemala, researchers at HRDAG compared rates of violence against indigenous populations vs others[6].

Generally, statistical analysis provide circumstantial or corroborating evidence. Normally, a statistical analysis will not by itself prove a crime occurred, but in combination with other evidence, such an analysis can determine whether the data is consistent with the crime being charged. At time the statistical analysis may disprove other explanations being offered by the accused.

1.3 What is conflict mortality?

Efforts to quantify the impact of war in ways that included civilian deaths only begin to appear in English language documentation about a century ago. Attempts to determine the number of deaths caused by a war can be found starting early in the 20th century, often using census data to determine how death rates changed during conflict [31]. In more modern

times, surveys have been used as well, including in the Central African Republic [59], the Democratic Republic of Congo [15], [49, 12, 1, 26], and other conflicts.

Both surveys and the census-based methods described above invoke a definition of war death toll now referred to as excess mortality. The basic idea of excess mortality is that war kills in many ways. For example, when a hospital is bombed, people are killed directly when the bomb explodes, but the next day, a woman may die giving birth at home because the hospital is not longer there. Similarly, disruption of sanitation and health services can lead to outbreaks of diseases such as cholera in a region that would not have experienced such an outbreak in the absence of war. By estimating rates of death before and after the onset of conflict, excess death estimates attempt to capture both deaths that result directly from a gunshot or bomb explosion and also deaths that occurred because of the disruption of infrastructure and health services caused by war.

Another way to conceptualize war death is to focus on violent deaths: individuals killed in battle, primarily by bullets and bombs. These are the kinds of deaths that are generally recorded as war deaths by observer groups, which means that frequently, more data is available on violent deaths than on excess deaths. In fact, which definition of war death a researcher focus on may be driven by the available data. It's important to notice however, that the two numbers will differ, and that the number excess deaths produced by war will always be higher than the number of violent deaths, as violent deaths are a subset of excess deaths. In this dissertation, two chapters focus on the analysis of violent death data from Syria. I believe the question of excess deaths is equally valid and important; however, the data available to date does not permit us to address that question.

1.4 Methods used in human rights data analyses

A full review of all methods of data analysis used in human rights investigations is beyond the scope of this dissertation. As in other fields, methods used vary according to the data that is available and the question being asked, and can range from straightforward tabulations to complex modelings. For a range of examples, see [4]

When the question is "Can we quantify the scope of killings or other violations?", statisticians generally rely on methods developed to estimate the prevalence of a condition or the size of a population.

To understand the prevalence of a condition in a population, an obvious method would be to conduct a survey. Surveys have many advantages. The theory behind representative sample surveys is well developed [38]. Surveys can be stratified to collect adequate information on subgroups of interest. Researchers can develop questionnaires to address exactly the questions that interest them. This ability to tailor a questionnaire to a specific research question is particularly appealing, as methods that rely on existing data sources, such as those discussed in this dissertation, can only answer questions about the data that already exists. For example, when we talk about the number of individuals killed by a conflict, we may be interested in both violent deaths and excess deaths. (As described above, excess

deaths are all deaths that occurred during the conflict that wouldn't have occurred if the conflict hadn't happened. In addition to those killed by bullets and bombs, that number can include mothers who died in childbirth because the hospital had been bombed and so they delivered at home without a doctor.) A survey can be designed to capture those kinds of deaths, whereas other methods are frequently limited to estimating the number of direct killings.

Surveys have been conducted during ongoing conflict or in immediate post-conflict settings, including during the war in Iraq and in several conflicts in Africa [24, 49, 12]. However, such surveys present significant challenges. Conducting a nationally representative survey requires having interviewers contact a sample of the population, generally through home visits around the country. This logistically challenging and difficult to implement during an ongoing conflict, and also expensive. For example, it can be challenging to obtain an accurate sample frame if a large portion of the population has been displaced from their homes. It can be dangerous or at time impossible for interviewers to access certain parts of a country. The point about cost is important, as well, as not all of groups that want information on a conflict have funding to implement a large scale activity that requires trained staff. Additionally, the nature of conflict violence itself can present issues. Even when death rates are extremely elevated, mortality is a rare event. In addition, conflict mortality is generally distributed quite unevenly. In recent years, researchers have raised concerns about whether smaller scale cluster surveys, which are common in epidemiological context to measure such things as vaccination rates and nutrition indicators, can provide precise enough estimates of death tolls [33] Additional issues include survivor bias; when entire household or settlements are decimated, no one is left to report those death. This problem can be an issue with other methods, as well.

A full discussion of the merits and challenges of surveys in conflict settings is beyond the scope of this paper. Suffice it to say that surveys are an excellent method for answering many questions, and can be a valuable tool in understanding the harms caused by conflict, but that they are not always feasible for addressing some of the questions that arise in human rights settings.

Hence, researchers studying conflicts frequently begin by summarizing data collected for other purposes. For example, if two human rights groups, one government office and one truth and reconciliation commission have compiled lists of known deaths, researchers can compare the lists, remove duplicates, and produce a master list of all known deaths. The United Nations Office of the High Commissioner for Human Rights (UNOHCR) used this methodology to estimate the number of known deaths in Syria during the first years of the war, through a contract with statisticians at the Human Rights Data Analysis Group, who did the actual work. It's important to note the use of the word estimate in that last sentence: even consolidating duplicate records within a list and matching records across lists is a statistical problem, as researcher must decide whether two similar but non-identical entries actually refer to the same person. This kind of work is referred to as record linkage. As described in Chapters 3 and 4, record linkage has been studied in both the fields of statistics and machine learning.

Additionally, researchers have adapted methods used in other fields to estimate the size of a population from samples from that population. In recent decades, much work has involved the use and development of capture-recapture methods [36] Capture-recapture methods, also called multiple systems estimation in this context (with the somewhat confusing acronym MSE), were developed in ecology at the turn of the twentieth century and later adapted for use in other fields, including public policy and epidemiology [28, 27, 29]. These methods use information in the overlap between multiple samples from a population (or lists developed for another purpose) to estimate the number of individuals that were never sampled. As described in Chapter 2, the initial capture-recapture estimator relies on strong assumptions, including independence of sampling occasions and homogeneity of individuals (with respect to probability of capture on a given occasion) . When there are three or more data sources, more modern methods can relax these assumptions.

When multiple groups are collecting data on casualties or other events and their lists overlap, it's possible to use record linkage techniques to deduplicate the lists, and capture recapture methods to model the data generating distribution that produced the lists. In this dissertation, we use these methods to examine questions relating to conflicts in El Salvador and Syria.

In the case of El Salvador, we examine data on child abductions from that country's civil war. Between 1979 and 1992, El Salvador was wracked by conflict between leftist guerrilla groups and right-wing nationalist governments. One feature of the conflict was the abduction of children by government military forces, or the forced surrender of children to those same forces. Since 1994, La Asociación Pro Busqueda de Niñas y Niños Desaparecidos has investigated cases of these child abductions. To date, they have opened more than 950 cases and located nearly 400 abducted children (now young adults). The organization remains active, and new cases come to light each year. In Chapter 2, we examine Pro Busqueda's data, assessing what can be said to date about the total as-yet unknown number of abductions that occurred, and we lay out a plan for updating estimates as new data becomes available.

In the case of Syria, we examine current data on deaths from the ongoing conflict there. Early in the conflict, the United Nations Office of the High Commissioner for Refugees (UNOHCR) contracted with statisticians at the Human Rights Data Analysis Group (HRDAG) to analyze data from multiple human rights groups that were attempting to document deaths from the conflict there. HRDAG produced three reports from the United Nations [45, 46, 44] and has maintained ongoing relationships with the local human rights groups that are collecting the raw data. HRDAG is now in the unusual position of possessing a series of multiple "snapshots" of each group's data, collected at a range of points between 2012 and 2016. We use some of those snapshots to assess how each group's data is changing over time, and conduct a preliminary analysis to assess whether those changes are likely to impact resulting estimates of unreported deaths. In addition, we take advantage of the large number of processed cases to assess the performance of a variety of classification algorithms in determining whether two records refer to the same individual.

Chapter 2

El Salvador Child Abductions: Estimates from Different Scenarios

2.1 Introduction

El Salvador suffered a brutal civil war between 1980 and 1992. A United Nations-sponsored truth and reconciliation commission formed after peace accords were signed in 1992 reports that 75,000 people were killed [7]. Rural areas, especially the departments north and east of the capital San Salvador, were often targeted by government forces, and sometimes entire villages were massacred in policies intended to reduce the range from which guerrilla forces could operate [17] To put these numbers in perspective, at the time, the population of El Salvador was roughly 5 million.

In addition to known deaths, may more individuals were disappeared [7]. "Disappeared" is a term used in the human rights literature to refer to individuals who were taken by government security forces and never seen alive again. Disappeared are not included in death counts, as their fates remain officially unknown. However, most frequently the disappeared were tortured by security forces and then killed. The lack of finality for families of the disappeared, who never receive a body or know for certain the fate of their loved ones, adds to their pain.

In this report, we examine a different type of disappearance. Military operations during the civil war at times included the forceful abduction of children. Sometimes these children were presumed to have been killed during the chaos of a massacre. Other times, a child was taken from their family by soldiers, who only knew that that was the last time they saw the child. In some cases, children were taken en masse from a village. In other cases, families surrendered children under duress, usually with the hope they could retrieve the child later when conditions were safer.

Jon de Cortina was a Jesuit priest from Spain who lived in El Salvador through the civil war [56]. An outspoken critic of the right wing government, he served a rural parish, and was also on the faculty of Universidad Centroamericana (UCA), a Jesuit university in San

Salvador (he had a doctorate in engineering). On November 16, 1989, six Jesuit priests at UCA were massacred, along with their housekeeper and her daughter [7]. Cortina had shared a residence with those priests and would have been killed had he not spent that night at his rural parish. When the the war officially ended in 1992, Cortina continued to serve those most affected by it. He heard reports of children who had gone missing during the conflict and who remained unaccounted for. Eventually, he located several abducted children at an orphanage. And in 1994, with parents of the disappeared youth, he formed La Asociación Pro Busqueda de Niñas y Niños Desaparecidos to search for the children. Since then, more than 948 cases have been opened, and in more than 389 of those, a child has been located.

However, the total scope of child disappearances during the war remains a mystery. The majority caseload of Pro Busqueda represents what executive director Eduardo Garcia Dobles refers to as "denunciations." That is, cases represent disappearances where parents have come forward to report an abduction. For this to happen, several conditions must be met: parents must suspect that their child is still alive, they must know that there is an organization that will investigate the disappearance, and they must trust that opening a case will not put them or the child in further jeopardy. It is clear that for some disappearances, not all three conditions are have been met.

This is evidenced by the fact that parents continue to come forward with new cases today. According to Pro Busqueda staff, between five and ten new cases are still being registered each year, more than two decades after the signing of peace accords. Pro Busqueda reports that fewer cases are opened annually now than in the years immediately following the war. This makes sense; one would expect parents and other family members to begin searching for their lost children as soon as possible. However, the fact that parents continue to come forward at all suggests that either they were unaware of the resources available to them earlier, or they were afraid of making a public statement about the event.

In addition, researchers estimate that more than 350,000 and possibly as many as 850,000 left El Salvador for the United States during the war [47], and very little is known about how many of those individuals lost children during the war. Pro Busqueda is a small organization with a modest profile; families living in the United States are unlikely to have heard of the group. To date, only one Salvadoran family in the United States has opened a case, and this occurred after a regional outreach campaign in San Francisco in 2013 (personal communication, Cristián Orrego). How many other Salvadorans may have lost a child before leaving is unknown.

And, tragically, in some cases, parents may be unaware that their child is still alive. In one case, Pro Busqueda investigators were searching for one child they knew was alive and ended up finding a second child from the same family. The parents had not reported that child missing because they believed he was dead (personal communication, Cristián Orrego).

All of these factors indicate that there were more child disappearances than have been registered with Pro Busqueda. However, the total remains unknown. Why does this matter? Even if the total number of child abductions is several times the number known to Pro Busequeda, it will still be small compared to the number of killings and adult disappearances. Also, the fact that large number atrocities were committed during the war has already been

documented. Though the report of the truth commission did not address child abductions specifically, it's unlikely that an estimate of the total number of such cases will, for example, lead to a sudden change in public perception of the war years in El Salvador. Furthermore, if prosecutions were to take place for cases of child abductions, they would likely focus on documenting specific cases for which details are known. (Prosecutions have not taken place to date; in 1994, El Salvador passed an amnesty bill covering crimes from the war. This spring, the government did agree to cooperate with Spanish courts in a case against Salvadoran military officers for the killing of the Jesuit priests [40] Still, truth is important, and a full understanding of El Salvador's war requires an understanding of this specific and specifically horrible crime.

Even more important, this is a crime for which remedial actions can still be taken. Many of the parents who lost children are still alive. The children of the civil war years are now young adults. There is still time for families to learn what happened to their babies, and for the young adults who were taken from their families to learn the truth of their origin. The psychological ramifications of learning this information – especially for people who grew up believing that their parents were military officers, or that they were a war orphan – are complex, and a subject that could likely justify an entire dissertation in another discipline. Pro Busqueda has wrestled with these issues and employed psychologists to help develop processes to support both families of origin and abductees. Yet, the right to know the truth of one's origins remains evidence. Determining the total number of cases that occurred can help justify action to identify and address those cases. This ability to still address a crime is unusual when examining human rights violations nearly 25 years after a conflict has ended.

We approach this question from multiple angles. None allow us to present a single estimate of the total number of child abductions. However, a careful consideration of the data and plausible models indicate that the total number of abductions is considerably larger than the 948 open cases known to date. We focus primarily on Pro Busqueda's case data. Pro Busqueda has cases opened by parents as well as by children – now young adults – who suspect they may have been abducted. We assess the feasibility of parsing that data into two lists and generating a capture-recapture estimate of the total number of child disappearances. In addition, we provide some description the population that fled El Salvador during the war using both US and Salvadoran sources, and provide suggestions for assessing the rate of child abductions suffered by that population.

2.2 Analysis: Estimating the total number of children abducted from families that still reside in El Salvador

We are, of course, interested in all children that were abducted or surrendered under duress during the war in El Salvador. However, when considering case data from Pro Busqueda, it is useful to subdivide cases into children abducted from families that still reside in El

Salvador, and children abducted from families that now reside in the US or elsewhere. This is because almost all of Pro Busqueda's caseload reflects abductions from families that still live in El Salvador, so we have no information about families that fled the country during the war. Therefore, the focus of this paper is to estimate the number of as yet unknown cases from parents still residing in El Salvador. In Section 3, we present some information that can help us understand the potential for additional cases among families in the diaspora, but we do not produce an estimate of that number.

Capture-Recapture Analysis

Capture-recapture is a technique originally developed to estimate the size of wildlife populations [43, 35]. In recent decades, capture-recapture methods been extended to estimate the size of hidden or difficult to measure populations in many fields, including traffic accidents [2], marijuana growers [9], sex workers [32], and even the dead in conflict settings [36], as well as the US Census undercount [18]. In these contexts, the available data are generally not probability samples. Rather, researchers are frequently presented with multiple lists of individuals.

To address the issues such data presents, including unequal capture probability of individuals and dependencies between lists, researchers have extended traditional capture-recapture using combinations of stratification and modeling There is a substantial literature on loglinear models [20, 16, 48], Bayesian approaches [54, 53, 61] and Grade-of-Membership models [41]. However, for the most part these techniques are not directly applicable to the problem at hand. Generally three or more lists are required to model the dependency structures, and for now we have, at most, two lists, as discussed and refined below.

Here, we provide a brief overview of two-list methods for capture-recapture analysis. Then, we discuss the issues that arise in modeling Pro Busqueda's data and provide a range of estimates that vary according to assumptions. We also consider the the potential value stratification could provide in the future when we are able to access appropriate covariates.

The method

Generally, capture-recapture methods are used to estimate the size of an entire population from multiple samples from that population. The technique was developed in ecology as a way to estimate animal populations from repeated samples, and it has been adopted in epidemiology and other fields to estimate difficult to reach populations, using different lists of individuals instead of random samples. In its most basic form, the technique uses two independent simple random samples. That is, each sample should be a random draw from the population, and the fact that an individual appears in one sample should have no effect on the chance that the individual will appear in the other sample. That is, the two sample schemes act independently

These usual assumptions required by the primary two list capture recapture estimator are:

1. Independence of capture occasions (or lists). That is, the probability that an individual shows up on one list is independent of the probability that they show up on a different list.

2. Equal capture probability of individuals. Each individual should have the same probability of showing up on a specific list. (The sampling probabilities do not need to be the same across lists.)

3. Perfect matching across lists. In the case of tagging experiments, this means that no tags are lost. In epidemiological contexts, it means that it is possible to correctly determine whether two separate list entries refer to the same person. (This may be necessary to deduplicate entries within a list as well as match across lists.)

In epidemiology and other contexts, the technique is frequently referred to as Dual Systems Estimation or Multiple Systems Estimation (MSE), when more than two lists are employed. In these contexts, researchers are not usually sampling from a population, but, rather, using existing data. For example, to estimate the number of cases of a reportable disease, one might use a dataset of reported cases, hospital records, and lab test results (i.e., three lists). To estimate war deaths, one might use lists from a truth and reconciliation commission, a government agency, and an NGO that tracked deaths (also three lists). These hypothetical examples illustrate the challenges of using data collected for another purpose: such lists are unlikely to be independent, and for some lists, individuals are unlikely to have an equal capture probability (that is, they are unlikely to have an equal probability of being listed). In the case of war deaths, for example, deaths in a capital city where observer groups are based may be more likely to be reported. Often researchers stratify on time and geography to account for varying capture probabilities. Additionally, dependencies between lists and heterogeneous capture probabilities can be modeled to some extent when there are three or more lists. However with only two lists, the assumptions of independence and equal capture probability are effectively required to develop the estimator.

The first capture-recapture estimator, developed in for a two-list case, is known as the Lincoln-Petersen estimator [43, 35]. In their early papers, Petersen and Lincoln implicitly relied on the assumptions above using somewhat ad hoc derivations of the estimator that implicitly followed the logic below. $\mathbf{1}$ is the indicator function. N is considered fixed (though unknown of course). A, B, and M are random variables.

$$
\begin{aligned}
N &= \text{total number of individuals} \\
A &= \mathbf{1}(\text{individual is on list A}) \\
B &= \mathbf{1}(\text{individual is on list B}) \\
M &= \mathbf{1}(\text{individual is on list B and list A})
\end{aligned}
$$

Then, assuming independence:

$$P(M) = P(A)P(B)$$

Assuming that the lists can be formally considered to be separate simple random samples from the N individuals, estimates for these probabilities are:

$$P(A) = \frac{\sum A}{N} \equiv \frac{N_A}{N}$$
$$P(B) = \frac{\sum B}{N} \equiv \frac{N_B}{N}$$
$$P(M) = \frac{\sum M}{N} \equiv \frac{N_{AB}}{N}$$

So, assuming independence, both $\frac{N_A B}{N}$ and $(\frac{N_A}{N})(\frac{N_B}{N})$ can be considered as estimates of $P(M)$ and thus should be "close" as long as N_A and N_B are not too "small."

Equating the estimates them immediately yields the Lincoln-Petersen estimator for N:

$$N = \frac{\sum A \sum B}{\sum M} \equiv \frac{N_A N_B}{N_{AB}}.$$

A common example used to describe the estimator is to consider estimating the total number of fish in a lake. On the first day, N_A fish are caught, tagged and released. One week later, N_B fish are caught. N_{AB} is the number of fish that are caught both days; that is, the number of fish in the second sample that have a tag. If all fish are equally likely to be caught on a given day and the fact that a fish is caught on the first occasion does not cause it to alter its behavior, then the conditions required by the Lincoln-Petersen estimator are met.

Initially, it might be challenging to see how such an estimator could be applied to the problem of estimating the number of child disappearances. Before we return to this question, we take a more straightforward human rights example: deaths during a war. For example, imagine that two different groups were trying to record deaths as they occurred: a government agency and a human rights group. If each group worked independently to try and record deaths and then compared lists, and each death was as likely to be recorded as any other death on a given list, the the estimator above would work as well for estimating a total number of human deaths as it does for fish.

Of course, it's easy to imagine that the necessary formal conditions could be violated. In fact, it's hard to imagine a scenario where the assumptions would hold. Even when different groups are collecting data, their lists may not be independent. For example, if the human rights group is seen as aligned with a rebel faction, then families aligned with that group

may be more likely to report the death to the human rights group than the government. In such a case, the fact that a death is recorded by the human rights group may indicate that it less likely to be reported by a government group. A more recent phenomenon is the possibility of copying between lists. This has arisen as more groups are making their data available online, and it is an issue we consider in Chapter 3. Additionally, individual deaths may not be equally likely to be recorded on a given list. Groups tracking deaths are usually more likely to hear about deaths closer to where they are based, and the real or perceived affiliation of an organization may make reporting more or less appealing to different groups in a conflict. Additionally, when violence skyrockets, certain areas may become harder to access, actually limiting reporting when death tolls are highest. Of course, it is not a requirement that lists be comprehensive. If they were, there would be no need for estimation. However, if the probability of being recorded on a list depends on the level of violence in a region, the individual deaths are not all equally likely to be captured. Deaths may be more or less likely to be recorded during different time periods for other reasons, as well. For example, if an NGO steps up (or decreases) its efforts as a conflict progresses this could also lead to heterogeneous sampling probabilities for deaths. Such heterogeneous selection probabilities can induce dependencies between lists.

To mitigate the issues of unequal capture probabilities and potential dependence between lists, investigators may stratify the data into smaller units (usually over time and space) where the assumptions are more likely to hold. In a situation where more than two lists are available, researchers may also attempt model the dependency structure.

As discussed above, considerable work has been done on capture-recapture estimation since the original Lincoln and Petersen papers. (For an overview, see [29]) A core methodological development came from Fienberg in 1972. If t is the number of lists or capture occasions, counts of the $2^t - 1$ observable possible capture histories (or patterns of appearances on lists) can be viewed as an incomplete contingency table, as the 0 capture history (individuals that are never captured or recorded on any list) is never observed. A loglinear model of these counts can then be parameterized, with abundance estimates derived from the parameters. Since that paper, the use of log linear models to estimate capture-recapture models has been extended, with packages such as Rcapture (Rivest 2007), providing parameterizations for a variety of capture-recapture models, as well as straightforward customization.

Other approaches, including Bayesian models, have seen recent developments. Additionally, in ecology, much effort has been focused on developing experimental designs that will increase the chances that the data collected meets the conditions required by the estimator; for example, using different trapping mechanisms, as well as developing estimators based on different sampling procedures.

However, here we are confronted with existing data, and a limited number of potential lists. Below, we discuss the issues the El Salvador data presents and proceed with two-list estimation.

Pro Busqueda's data

Could capture-recapture be used with Pro Busqueda's case data? At first, it would seem the answer is no: Pro Busqueda is a single organization, and capture-recapture requires multiple sources of data. However, cases at Pro Busqueda can be opened in different ways. The primary avenue is for parents or other family members to come forward and open a case. At times, a Pro Busqueda investigator may identify a new case in course of investigating one opened by a family. These cases are similar to those opened directly by a family, and for simplicity, we refer to both types of cases as "family cases". This comprises our first "list" of disappeared children, and there are 934 of these cases. On the other hand, some cases are initiated by potentially disappeared children themselves: either young people living in El Salvador who have suspicions about their origins or young adults from other countries who know they were adopted from El Salvador. (In general, families adopting stolen children were told they were receiving war orphans.) Pro Busqueda has gotten many contacts from young adults who saw a news story about the organization and realized their own life story sounded similar to the one they were reading about: they were adopted from or within El Salvador during the war but little or nothing was known about their biological family. These young adults then contact Pro Busqueda to find out if they have a living biological family looking for them. Based on documents and DNA samples from families that have opened cases, Pro Busqueda investigators attempt to answer that question. These cases represent a second "list" of disappeared children, the "Child list." There are 295 of these cases. (However, not all are usable in this analysis; we discuss the cases in much more detail below.)

As indicated in the above paragraph, we are proposing to consider the data maintained by Pro Busqueda as two distinct lists. This raises the question, if we consider cases reported by parents (or opened by Pro Busqueda investigators) as one list, and cases opened when children approached Pro Busqueda as a distinct list, could we use multiple systems estimation to estimate a total?

The first issue to discuss relates to "matching." Here, cases on the family list refer to a single child, even if multiple family members are involved in a search and have donated DNA, so that list is already deduplicated. Matching between lists can occur via DNA testing. Pro Busqueda maintains a database of DNA samples provided by families who are searching for their children: as of October 2014, Pro Busqueda had DNA samples associated with 469 cases of the 934 opened by families of (in 110 of those cases, a child has been found. The rest of the cases remain open). When a young person contacts Pro Busqueda, staff request that the adult child send a DNA sample. If the child sends one, and it matches with one of the DNA samples matches one in the database, the child has found a birth family. In 95 cases, children have provided DNA, and 13 of those samples have matched to a family looking for the child. When a family member opens a case, a child may be found through investigative efforts, an issue we discuss below. However, when a child contacts the organization, as far as we know all matches are what investigators call "cold hits" – a match to an existing family case file through DNA matching.

The existence of DNA samples addresses key issue in capture recapture analysis, the

matching of individuals across lists. When comparing two or more lists of individuals, in order to determine the overlap between the lists, one first must determine which entries refer to the same individual. This can be a challenge when identifying information is limited, as is frequently the case with human rights data. For example, groups tracking casualties are unlikely to have access to a unique identifier such as a social security number. Rather, they may record an individual's name and some details about the event itself. One list may refer to a person as "Rob Smith" while a different source may refer to a person as "Robert Smith." One group may include information about Smith's age, but another may not. Even when more information is available, most items won't provide foolproof identification. Addresses may also vary, as people move, as may other identifying information. At times the same individual may be repeated twice within a list as well. In an epidemiological context, this could happen if the person returned to a program after an absence and was added de novo with a new address. In a human rights setting, this could occur if the same death was reported through two different channels. The issue of deduplication (removing within list repeats) and matching across lists in such cases are both referred to as record linkage, a problem discussed in Chapters 3 and 4. Here, we rely on DNA analysis to tell us of the child that comes in is the same person as the child that was reported missing by the family. Additionally, while multiple family DNA samples may exist for a case (perhaps a mother and a sibling, for example) Pro Busqueda connects each sample to a case, so the numbers we report in Table 1here are already deduplicated: that is, they represent individual disappearances, not the total number of available DNA samples.

It's important to first note that the unit in each "list" is the case. If a parent opens a case, it's considered an event even if the child has not been found yet. That is, we are assuming that the family did indeed lose a child, as there is nothing to be gained by reporting a false case, and potentially some risk. We are also assuming the child is still alive, a fact that has been true in 86 percent of the cases resolved thus far. (There have been exhumations that have confirmed the death of 54 children.) Similarly, to use the child-approach data as a list in capture-recapture analysis, we must assume that each child that contacts the organization is a case, even if he or she is not immediately matched to a family. This second assumption is more problematic than the preceding, as some adoptees who contact Pro Busqueda are possibly actual war orphans, an issue we discuss further below. That is, when a family reports that a child was taken from them in a military operation, we treat that as a true case. However, when a child contacts Pro Busqueda with questions about his or her origin, most of the time that is a "possible" case, especially in cases of international adoptions.

The use of DNA to match children to open cases solves one the common problems of use of capture-recapture with existing data sources: the matching process. DNA matches provide more certainty than comparison of names and addresses, for example. That said, we are using the term match somewhat loosely here; in the majority of cases, Pro Busqueda's DNA analysis supports a match but does not reach a threshold of 99 percent certainty, especially if the only samples were donated by more distant relatives such as aunts, uncles or grandparents rather than parents. For now, for cases with DNA evidence, we are considering a case a match if Pro Busqueda investigators have labeled it as such. At times, such matches

are based on genetic evidence supported by documentary evidence.

However, while the use of DNA to match between lists solves one problem, it adds a complication as well. Not all families who have reported missing children to Pro Busqueda have provided DNA, nor have all children who have contacted Pro Busqueda provided DNA. According to data provided by Pro Busqueda in March 2014 and an audit report by Michelle Arevelo-Carpenter conducted for the UC Berkeley Human Rights Center and Pro Busqueda [3], 469 out of 934 family cases have a DNA sample associated with them, while only 95 of 295 child cases have a processed DNA sample. As we are relying on DNA to match cases across lists, we are limited to considering cases for which a DNA sample is available when conducting estimation. There are several challenges to this approach, both at the conceptual level and in terms of the data available from Pro Busqueda, which we discuss further below.

Data Access

Currently, Pro Busqueda maintains case data in multiple legacy systems and it has not yet been possible to create an analytic dataset of case level data. In addition to the audit report mentioned above, staff at Pro Busqueda have provided counts of different types of cases, and have answered many questions. Table 1 summarizes our current understanding of Pro Busqueda's data and provides an update to numbers presented in 2014 at the Latin American Studies Association annual meeting [47]. The substantial differences between these numbers and those presented at LASA highlight the need for access to case level data to verify counts. The numbers presented here therefore remain provisional. Pro Busqueda categorizes cases differently than we do in this paper. The current case categorization was developed as part of the organization's investigative process and includes categories based on who opened the case as well as categories based on the relationship of the case to the conflict.

SDM (Solicitud del Menor): Requested by a child. This means that the case was opened at the request of a child.

DCC (Desaparicion del Conflicto): Disappearance related to the conflict. This indicates that the family separation was related to the conflict.

IDI (Investigacion de Identidad): Identity investigation. This is an investigation that was opened by Pro Busqueda. These cases are based on information information uncovered or received by investigators but are not directly requested by a family or child.

DFN (Desparicion Forzada del Nino): A family separation caused by the government.

SFN (Separacion Forzada del Nino): A family separation caused by guerrillas.

As we can see, the categories do not appear to be comprised of mutually exclusive definitions: some describe how a case was opened and some describe the relationship of the case to the war. However, currently, cases can only have one label. Furthermore, the way a case is categorized can be changed as a case progresses. The audit report by Arevalo-Carpenter notes that staff at Pro Busqueda occasionally re-categorize cases as they progress. For example, a case that starts as an inquiry by a child could be reclassified as a case related to the conflict as an investigation progresses. This particular re-categorization would be a concern for our purposes. It is not clear whether this is a frequent occurrence, but it is troubling in that we rely on the categorizations from Pro Busqueda from 2014, not the categorization when the case was opened, in building our lists.

For this paper, Pro Busqueda provided totals of cases that were opened by a child vs cases opened by family or their own staff. However, it would be valuable to have that information permanently captured at the case level. Arevelo-Carpenter's report suggests Pro Busqueda adopt a two-level case categorization that separates the way in which the case was initiated from the findings about its relationship to the war, a recommendation we endorse. (At the moment, some information about case origin exists entirely as part of the organization's institutional memory in the person of Pro Busqueda's lead investigator)

As discussed below, Pro Busqueda has hired a consultant to integrate its existing data systems and that work is expected to be complete by late summer 2016, at which point we will be able to access case level data and provide updated estimates. However, it is still useful to assess what can be understood from the data as available to us today, and we proceed with that effort here, starting with a discussion of whether our data meets the conditions required by the methods we wish to apply.

Data Issues

Do both lists capture the same thing?

Cases opened by parents are, by definition, child disappearances from the war. However, children approach Pro Busqueda for a number of reasons. Some have been raised in a military family but have always been suspicious of their origin, perhaps because they look visibly different from their parents or because they were old enough at abduction to have some memory. Some are young adult adoptees living in other countries who see a news story about Pro Busqueda and realize that their life story sounds similar to a Pro Busqueda case. Some of these individuals have been matched with families that are searching for them, validating that, yes, they were disappeared. However, among the cases that have not been matched, it is not always possible to know for certain whether the individual was truly disappeared (i.e., their family was still alive at the time of the abduction) or if the person was an actual orphan.

Additionally, disappeared children may have been taken from families who fled the country during or after the war. This presents a challenge as all but one of Pro Busqueda's

parent-initiated cases come from within El Salvador. As we do not have information from the parent side on families who fled, we will focus here on constructing an estimate of children abducted from families that still live in El Salvador.

Below, we consider several possible scenarios, including some in which not all child-initiated cases reflect real disappearances, and demonstrate the impact our assumptions have on resulting estimates. It is clear that determining what percentage of adoptions are likely to have have been irregular – that is, to have involved a case where a child had living relatives who would have preferred to keep him or her – is an important step, as this may help us arrive at a better estimate of what percentage of child initiated cases are likely to be real disappearances and what percentage are likely real orphans. However, this will not address the question of the number of disappearances suffered by families in the diaspora, a separate question that we cannot immediately investigate from the data available to date.

Are the lists independent?

If we consider Pro Busqueda's data to be two separate lists as described above, the assumption that cases that were reported by families and opened by Pro Busqueda investigators are independent from cases initiated by children seems plausible; each group is acting without awareness of the other's action. However, it's possible that some types of cases may be more or less likely to be reported by each group. If, for example, parents from a certain village were more likely to open cases, and children who were taken from that village were placed with military families, and children who lived with military families were more or less likely to self-report, a correlation can be induced.

Does each individual have an equal capture probability

As mentioned above, we know that among families that fled El Salvador during the war, there is only one reported case, the result of a 2013 regional outreach effort in the San Francisco Bay Area. Effectively, families that fled during the war appear to have a near zero probability of reporting. It is believed that this is primarily due to a lack of awareness of Pro Busqueda as an avenue to finding a disappeared child. However, some believe that fewer families who fled suffered the loss of a child. It's plausible that families who lost children were more likely to stay, in hopes of locating their lost child. However, an argument could also be made for the converse: perhaps families who suffered such a tragedy would have been more likely to endure the hardship of fleeing in the hopes of getting their remaining children to safety. What seems clear is that the fact of losing a child seems almost certain to influence a family's decision to leave or stay, and there is currently no data that would allow an exploration of that relationship.

Ideally, we would like to stratify on location of the family, but that is not currently possible. For cases on the child list that have not yet resulted in a match, it is impossible to know where the parents are living or where the child was abducted. We discuss this issue below.

In addition, our data presents a challenge not seen in mortality counting: each time Pro Busqueda investigators locate a child and resolve a case, that child is removed from the population of cases that could be discovered another way. That is, if Pro Busqueda investigators locate an adoptee living in the US and connect that person to a family in El Salvador, that adoptee can no longer contact the organization for the first time with questions about their identity. We discuss strategies for dealing with this issue below.

The Data

	Family Cases	Child Cases
Total cases of this type	934	295
Cases for which there is DNA evidence	469	95
Number of matches where child died	54	0
Number of matches where child is living	334	13
Number of matches with DNA evidence	110	13

Table 2.1: Pro Busqueda's data as of May 2014 (Arevelo-Carpenter 2014 and email communication with Pro Busqueda staff. Note that the columns are not mutually exclusive. The 13 matches in the child column are also included in the matches in the Family column.)

Table 2.1 presents Pro Busqueda's data as investigators there consider it. The two columns are not exclusive; the 13 matches identified on the child list are also included in the "family case" list. (This is more intuitive than it first sounds: a child case can only match to a family case if it is already on the family list).

Let's first consider the data on parent cases. As of July 2014, 934 cases had been opened by parents or staff at Pro Busqueda (our "family cases"). In 334 those cases, a living child was located and matched to a family. In an additional 54 cases, exhumations confirmed a child sought by family was dead. Most of these matches were the result of investigative efforts by Pro Busqueda staff. In many of the cases, DNA evidence did not play a role, as the organization began using DNA evidence around 2005. In 110 of the 334 cases where a match was made, DNA was used in determining or confirming a match. In 13 of those cases, a child contacted Pro Busqueda and their DNA matched to a family with an open case. In the other 97 of the 110 cases where DNA was used in matching, it provided confirmation of a potential match unearthed by investigative efforts. That is, in those cases Pro Busqueda investigators identified a child and then confirmed his or her identity with DNA. Two of those were already deceased when found; the other 95 were alive. None of those 97 cases are included on the child list, as the children did not come forward by themselves, but, rather, were located by Pro Busqueda. Furthermore, there are additional cases where Pro Busqueda investigators located a child but no DNA analysis was conducted; most of these are matches completed prior to the introduction of DNA analysis at the organization.

Looking at the Child case column, we can see that 295 individuals have contact Pro Busqueda with questions, but only 95 of those queries have resulted in completed DNA

profiles. It is currently not clear if there are more profiles awaiting sequencing or if this is the total number that have been provided. According to conversations with staff, and also Arevelo-Carpenter's report, in some cases, individuals decide not to pursue the question about their origin after an initial contact. Others request a DNA kit but fail to mail it back. Either way, for our purposes, we can only formally consider cases with a DNA profile available for matching.

Table 2.2 presents one possible two way table that could be constructed from the data in Table 1. To construct the table, we made several decisions. First, we included only cases with an associated DNA sample, as child cases can only match to a family looking for them if a DNA sample is available. That would give us 469 parent cases. However, we subtract 95 from that number, as those 95 cases with located living children with DNA samples that were resolved through investigative efforts are presumably not available to be matched to a child who approaches the organization. This gives us 374 family cases, 13 of which are matched and 361 of which are not matched. For the child list, we have the 95 cases where a child contacted the organization and provided a DNA sample that has been processed. Thirteen of those are matched, and 82 are not.

It's worth noting that Pro Busqueda investigators have located living children beyond those matched by DNA. According to communication with staff, there are 226 located living children (334 - 108) for whom there has not been a DNA confirmation. (Two DNA matches were deceased; 108 out of 110 were for living children) We assume that those individuals were matched to families for whom there is no DNA profile. That means that those individuals are not represented in the 469 cases for which there is an associated DNA sample (our starting point for this analysis). We would expect our analysis to generate an estimate large enough to at least cover those known cases not used in estimation.

In the section on estimation, we consider other ways to handle cases where a Pro Busqueda investigation found a child, both with and without a DNA match.

		Family Cases		
		Known	Unknown	Total
Child Cases	Known	13	82	95
	Unknown	361	?	
	Total	374		

Table 2.2: Scenario 1: A possible two-way table

Estimates

To implement a capture-recapture estimation, we need three quantities: the number cases on each list and the number of cases on both lists. To begin, let's return to the numbers in Table 2. Ideally, we would want our parent list to include all cases opened by parents and our child list to include all inquiries initiated by adult children. However, as discussed above,

for a child-initiated cases to match a parent case, there must be a DNA sample associated with each case. So for the scenarios presented here, we restrict ourselves to cases that have a DNA sample associated with them, and we assume these cases are samples of the universe of parent- and child-initiated cases.

In this section, we begin with an analysis based on the numbers as shown in Table 2, and then we discuss a range of possible scenarios that address the issues described above, showing how different assumptions affect our estimates.

For example, if we consider the child list to only include cases that were actively initiated by a child, then we are, in effect, assuming that all cases that were opened by parents and resolved through Pro Busqueda's investigative efforts involve children who would never have self reported. We explore the effect of this assumption by relaxing it in several scenarios below. We must also consider the fact that we have no data on parent cases for parents who fled the country.

We explore the effect of this issue in the final scenarios below.

Scenario 1: This is the scenario described in the previous section. We treat the observed parent-initiated cases and child initiated cases that have DNA samples as random samples. We also must assume that all 95 self-reports with DNA samples are real abductions from families still living in El Salvador. We further assume none of the children identified through investigative efforts would have self-reported.

We have 374 parent cases, 95 child cases, and 13 matches.

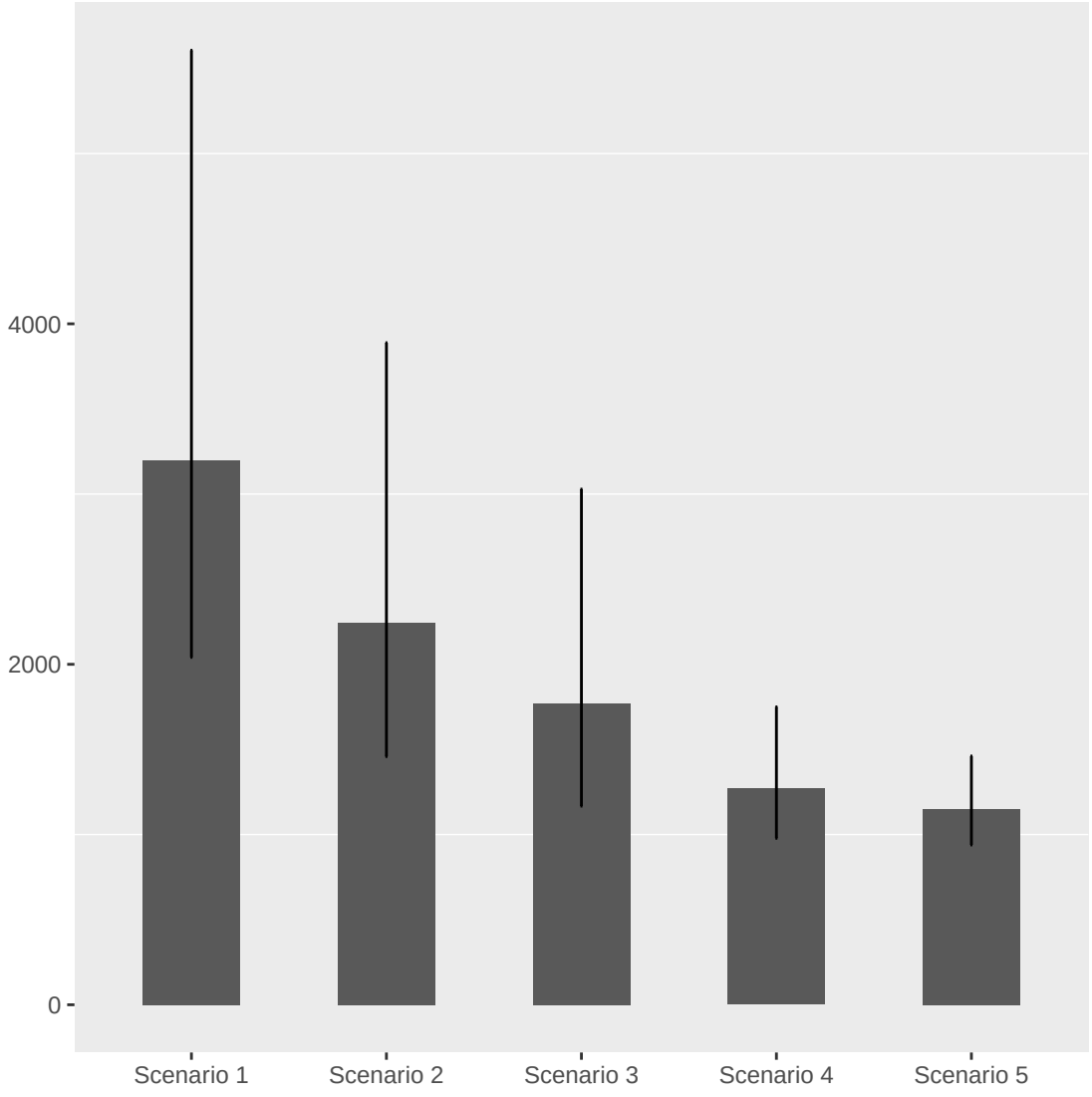
Scenario 2: Same as above, but we assume that 1/3 of the 95 self-reports were legitimate adoptions and exclude them from this analysis. (We choose 1/3 for convenience; this number is unknown.) We have 374 parent cases, 63 child cases, and 13 matches.

Scenario 3: Same as above, but we also now assume that 1/6 of the abductions are from families that now reside outside of El Salvador and we exclude them from this analysis. (This is approximately the percentage of Salvadorans who left the country during the war. However, this is still a guess; we cannot assume that the families who have left the country suffered the same pattern of child loss as those who stayed.) We have 374 parent cases, 47 child cases, and 13 matches.

Scenario 4: Same as scenario 3, but we also assume that 5 percent of the 321 children identified by investigators (334 matches - 13 who self presented) would have self-reported and matched to a family, and we add them to the child list. (Again, this is a guess. Four percent of the total found so far were self reports, so it doesn't seem implausible that another 5 percent might have been if they had the chance.) We have 374 parent cases, 63 child cases, and 29 matches.

Scenario 5: Same as scenario 3, but we assume that 10 percent of the 321 children identified by investigators would have self-reported and we add them to the child list. We have 374 parent cases, 79 child cases, and 45 matches.

Figure 2.1: Estimated disappearances under different scenarios



We estimated the total number of cases using the Rcapture package in R. Table 3 presents results from a loglinear model: the point estimates are identical to those given by the Lincoln Petersen estimator. In addition, we present results from a multinomial likelihood using profile likelihood confidence intervals, as the estimator is known to have a non-normal sampling distribution in finite samples. Those results are graphed with error bars in Figure 1.

	Total	SD
Scenario 1	3215	822
Scenario 2	2262	562
Scenario 3	1786	431
Scenario 4	1278	189
Scenario 5	1154	130

Table 2.3: Estimated Disappearances, Poisson Model

	Total	LCI	UCI
Scenario 1	3196	2041	5607
Scenario 2	2245	1458	3889
Scenario 3	1769	1167	3029
Scenario 4	1270	978	1750
Scenario 5	1149	939	1461

Table 2.4: Estimated Disappearances, Multinomial Likelihood

As mentioned above, in constructing these estimates, we ignore data on parent cases that do not have DNA associated with them. By assuming the cases with DNA are a sample, we are assuming that those known cases that we ignore are being estimated. Therefore, any estimate should be at least as large as the total number of known cases, including those without DNA samples. Scenario 5 here produces a lower confidence interval that is fairly close to the known total. It is possible to construct scenarios with assumptions that lead to an estimate (from the cases with DNA samples) is smaller than the total number of known cases.

(A similar issue exists for child cases without a DNA sample. However, we run into the same problem discussed above of not knowing definitively whether all child contacts are real cases.)

Next Steps

The previous sections suggest several ways it could be useful to stratify our data. At the moment, Pro Busqueda's data exists in multiple legacy systems, and Pro Busqueda staff have provided counts of different types of cases. However, Pro Busqueda is in the midst of a data transition project that will bring legacy data systems together and we expect to have access to case level data during summer 2016. This will allow us to stratify our data on several variables.

These are the variables we hope to access, and why they would be useful for stratification:

Time and location of disappearance. Just as the intensity of the war varied by region, it also varied over time. Additionally, as cases are solved, the population of individuals and families eligible to report a case is changing over time. This data should be available for

families. For child cases, at least the timing of the adoption should be known; it's not clear whether all cases will include the region from which the child came.

Location of family today (whether remained in El Salvador or migrated). The fact that all but one family-initiated case come from families within El Salvador means that we would like to stratify on location, as we do not have data on families who emigrated on the family list. To address this issue, it is theoretically possible to stratify on the location of family (at the time of reporting) and only attempt to estimate for families residing in El Salvador. However, for cases on the child list, it is not clear whether the child's biological parents still reside in El Salvador at the time the child contacts Pro Busqueda. This is because all information on the family is unknown at that point; it is when a child is matched to a family that the family location becomes evident. That is, this data is available for families, but not for the child list, as the reason adult children contact Pro Busqueda is to find out where their birth family is or was from.

Location of child (whether child was kept in El Salvador or adopted out). Families reporting disappearances generally do not know where the child is, so it seems plausible that in the case of family-reports, there is an equal chance of the case being reported if the child remained in El Salvador or was sent abroad. However, if we consider child self-reports to be a separate list, the probability of a case being reported by a young adult child will be related to where the child ended up. This data is available for the child list, but not necessarily for the family cases, as the reason parents open a case is because they don't know where their child is.

Time of case opening and closing (for resolved cases). This data should be available for all cases. This will allow us to see when a child was removed from the population of children eligible to self report.

Again, Pro Busqueda has provided some aggregate data that give a sense of what we can expect. Table 5 is adapted from a 2013 official Pro Busqueda slide presentation on cases resolved through 2012. As new cases have been resolved since 2012, the total number of cases does not match the numbers in our earlier tables, but it gives a sense of where the children that have been found so far ended up. We cannot infer from this data percentages of final locations for all abducted children, as other factors may influence whether a child is found. Particularly, it is logistically much easier to investigate cases in El Salvador, so that could be partly responsible for the high number of resolved cases with children in El Salvador. However, given that issue, it's even more unnerving to see the that 146 children were found outside the county – 44 percent of the total found so far. More than 60 cases of located children were children placed in adoptions with US couples. Given the more than 500 unresolved cases, it seems unavoidable that there are additional Salvadoran-American young adults who have grown up believing they were war orphans who in fact have living family searching for them right now.

Table 6 is also based on data through 2012, and shows where the cases came from. That is, where in El Salvador the known abductions occurred. Even when one accounts for the differing population sizes of departments, it's clear that the known child abductions were not evenly distributed across the country.

	Number
El Salvador	183
United States	61
Italy	37
France	19
Honduras	10
Guatemala	4
Belgium	3
Switzerland	3
Belize	2
Mexico	2
Holland	1
Spain	1
England	1
Monaco	1
Nicaragua	1

Table 2.5: Location of the child in resolved cases, up to 2012.

	Number
Ahuachapan	4
Cabanas	52
Chalatanango	168
Cuscatlan	50
LA Libertad	28
La Paz	19
La Union	5
Morazon	93
Santa Ana	40
San Miguel	21
San Salvador	174
San Vincente	169
Sonsonate	3
Usulutan	69

Table 2.6: Location where the case originate, all cases known up to 2012.

2.3 Data on emigration during the war

As illustrated in Table 7, research on emigration during the war suggests that between 350,000 and 850,000 Salvadorans left their own country for the US between 1980 and 1990 [47]. Lower estimates come from US census data, and are almost certainly conservative, as most Salvadorans entering the United States at that time were doing so without legal paperwork, and e had incentive to hide their status and existence in the Unites States from authorities. (Asylum cases were largely unsuccessful at that time. The United States was supportive of Salvadoran governments during this period and backing the government during the war through successive aid packages.) Higher estimates come primarily from the work of Salvadoran researcher Sergio Montes [47], who surveyed Salvadoran households to determine how many had relatives abroad in the United States, and also Salvadoran communities in the United States during the war years. An additional point raised by researchers is that the nature of migration changed during the conflict. Of particular interest to us, the percentage of migrants who were women of childbearing age increased. According to US Census data, which we believe to be conservative, 114,503 adult women entered the United States between 1980 and 1990 [47]. As we stated in earlier sections, the data we have to date does not allow us to estimate the rate of child loss within that population. However, given the existence of families among the migrants during the war (as opposed to men migrating alone for economic reasons in earlier years) and the fact that many were fleeing human rights violations, we believe it would be useful to explore the existence of child loss within this population.

In Table 7, we can see that the number of Salvadorans living in the United States increased by at least a factor of 5 and possibly a factor of 10 between 1980 and 1990, a decade that included much of the worst violence of the war. Additionally, the diaspora has continued to grow. The Salvadoran government estimate from 2007 would suggest that possibly more than 20 percent of the Salvadoran population now resides in the United States.

	1970	1980	1990	2007
In ES - ES Census	3,598,232	4,585,925	5,110,176	5,744,113
In US - US Census	15,717	94,447	465,430	1,104,390
In US - S. Montes	-	-	950,255	-
In US - ES Government Est	-	-	-	1,842,100

Table 2.7: Number of Salvadorans living in El Salvador and the United States, based on US Census data, El Salvador Census data, and survey work by Sergio Montes, as collected in [47]

2.4 Discussion

In this paper, we have explored using the data from the organization La Asociación Pro Busqueda de Niñas y Niños Desaparecidos to estimate the unknown total number of cases

of child abductions during El Salvador's civil war. All analyses presented here indicate that the cases of child abductions known to Pro Busqueda are only a subset of a larger universe of child abductions. However, at this point, the number of unknowns in our data mean that is not possible to pinpoint a solid estimate. Our work highlights the need for full access to case level data in order to confirm our categorizations of cases. As we can see in this analysis, the re-categorization of a modest number of cases has a large effect on the resulting estimate. This is easy to understand if one looks at the Lincoln-Petersen estimator described above. If the number of matches between the lists increases, this increases the denominator and decreases the estimate. If the number of cases on either or both lists is increased, this increases the numerator and therefore the size of the estimate. This is why, for example, it is important to understand what percentage of child-reported cases are true cases. If the child list decreases in size, the estimate decreases. Similarly, when we consider whether the cases that were located through Pro Busqueda investigators should be added to the match list, we are increasing the size of the denominator. Given these limitations, all results presented here should be considered provisional. We intend to produce revised capture-recapture estimates will when the data is available with appropriate covariates.

In addition to a case-level review of Pro Busqueda's data, it would be useful to explore additional avenues to address these substantive questions. For example, could experts in El Salvador give us a better sense of what percentage of adoptions may have been irregular.

It may also be helpful to explore other data sources that could shed light on the question of how many abductions occurred. For example could the United States Department of State provide data on the number of adoptions visas issued during this period. This could help us put a ceiling on the number of irregular adoptions that were possible in the United States. Access to records in El Salvador could be informative as well, if they could provide an improved understanding of the adoption process.

Despite all of the limitations discussed here, our analyses do show that even with fairly conservative assumptions, a number of unreported cases exist. It's especially notable that this is true even when we restrict ourselves to children abducted from families that still reside in El Salvador. How many children were abducted from families that fled in El Salvador remains an open question. Opinion varies on how likely it is that those who fled would have lost children. However, census data from that period do indicate that women of childbearing age were entering the United States, likely as part of a family unit, and it makes sense sense to explore whether child abductions occurred in that population. This, unfortunately, is a challenging question. There are large Salvadoran populations in many large US cities; however, reaching this population is more challenging for Pro Busqueda than reaching the Salvadoran population with in El Salvador. Additionally, even though the rate of child abduction in El Salvador appears to have been higher than seen in other Latin American conflicts, it is still a rare event. That is, one cannot expect to partner with a single local or regional agency and necessarily reach large numbers of Salvadorans who lost children. Additionally, human rights abuses are by their very nature, sensitive topics. Finding a way to reach the Salvadoran population in the United States such that people will feel safe enough to answer truthfully is an additional issue.

Despite these challenges, we look forward to continuing this line of research. All human rights abuses are, by definition, terrible, and a full accounting of such abuses is crucial to understanding conflict, and to both justice and reconciliation processes. However, in this instance, we are fortunate in that the research also poses the possibility of aiding people who lived through the war. The value to families who suffered the abduction of a child in knowing what occurred is clear. Additionally, there are likely young adult Salvadoran-Americans who believe they are orphans when in reality they have biological family searching for them. To date, more than 60 such cases have been confirmed out of 388 cases that have been solved. Given that Pro Busqueda has more than 500 open cases, there are almost certainly more Salvadoran-Americans waiting to be found just within the open case files.

Our work could help identify how many more biological families were torn apart who have not opened any formal case. Such an estimate could help prompt further investigative efforts. It will also help establish an accurate historical record for the period, and could prove useful if criminal trials are ever pursued. To date, this hasn't happened. In 1994, shortly after the UN-sponsored truth commission released it's report, the government of El Salvador passed an amnesty law covering crimes committed during the war. However, there remains a significant interest in accountability in El Salvador. This spring, the government agreed to cooperate with a case brought in Spain against 17 former Salvadoran military officers who killed six Jesuit priests in one of the most famous massacres of the war [40]. Two years ago, in Guatemala, one of El Salvador's neighbor's with a similar political history, that country's attorney general tried a former Guatemalan head of state for war crimes in a national trial, an event that experts had considered impossible up until a short time before it happened.

Chapter 3

Syria Casualty Data: reporting patterns over time

3.1 Introduction

Assessing the effects of ongoing war and civil conflict is challenging for researchers. War destroys people, infrastructure, and institutions. During active fighting, a seemingly simple question such as "How many people are dying?" may be impossible to answer. In many cases, wars and civil conflicts occur in regions in which birth and death registration systems are woefully incomplete, as is the case in many countries in Sub Saharan Africa. Even in countries with good record keeping infrastructure prior to conflict, such systems are often an additional casualty of violence. Other methods of assessing the effect of conflict on a population, such as surveys, a common method for assessing the health of populations, become logistically challenging to implement during active fighting.

Syria provides a case study of such a situation. In the spring of 2011, what began as protests in the south of the country over the arrest and torture of teenagers who'd painted revolutionary slogans on a school wall quickly morphed into nationwide protests demanding the resignation of Syrian president Bashar al-Assad. Each government response, including firing into crowds of protesters, fanned the protests, and by 2012, conflict had spread through the country [51, 8]

Fairly quickly, humanitarian aid was effectively shut out of the country, and the UN was unable to operate a mission that could provide independent assessment of the situation on the ground. Reports of killings, including on in the media and on YouTube indicated the death toll was rising. Rami Abdul Rahman, a Syrian exile in London, worked to focus attention on the conflict through a group he called the Syrian Observatory for Human Rights (SOHR), tracking data about deaths online and publishing it on his website [39]. Yet, no one knew how many people total were actually dying.

As an attempt to address that question, the United Nations High Commissioner for Human Rights contracted with the Human Rights Data Analysis Group (HRDAG), anon-

profit organization based in San Francisco, CA. The UN asked HRDAG to assess the data available on deaths from several different groups. Multiple groups, including the Syrian Observatory for Human Rights, were documenting casualties. Many of these groups had large observer networks on the ground in the country. Even the Syrian government provided the UN with data on casualties for the first few months of the conflict. UNOHCRR wished to know how many total unique deaths had been counted by the all the groups it was communicating with. Obviously, they could not simply add the totals from each, as many deaths would have been counted by more than one organization. HRDAG deduplicated the groups' individual lists and compared records across lists, determining when a person was included on multiple lists. By determining when multiple records referred to the same person, a process described in more detail below, they were able to estimate how many unique individuals were present in the full dataset (a compilation of records from all lists). Estimates of the total number of unique deaths were published in January 2013, June 2013 and August 2014 [45, 46, 44].

This was important work; however, it did not answer the question of how many people total were dying in the conflict; it only described how many deaths had been recorded. Inevitably, the total death toll will be higher than the number recorded by observer groups. During the chaos of conflict, human rights observer groups and media simply cannot be in all places violence is occurring at all times. Prior to this, researchers at HRDAG had analyzed data on violence in post-conflict settings, frequently using capture-recapture methods to estimate the total number of deaths that occurred, including undocumented deaths. (Capture-recapture methods are described in Chapter 2.) They have always found that the number of counted deaths is never the same as the total number of deaths that occurred.

In working with the UN on estimating the number of documented deaths in Syria, the researchers recognized that the project provided data that could be used to estimate the number of deaths that had not been reported, as well. Work on that estimate is ongoing at the time of this writing. One challenge the researchers have faced is that several of the observer groups continue to provide updates. Those updates are not simply new deaths that have occurred since the last provision of data. Rather, each group continually updates its records, so deaths that occurred earlier in the conflict may be added, and records may be updated. At times, a record may be removed when new information comes to light. This has meant frequent re-starts of the estimation process.

This scenario is a novel situation for HRDAG researchers. Traditionally, they have worked in post-conflict settings, for example, preparing reports for use by truth and reconciliation commissions or war crimes trials. In those situations, they collect the best available data that exists, and that's what they have to work with. By contrast, in the case of Syria, they are trying to estimate conflict deaths using similar methods, but while the civil war progresses. In 2015, we began a conversation about attempting to assess how the documentation of deaths is changing over time in Syria.

In the case of Syria, we have the unusual situation that estimates of known deaths were created at multiple time points during the conflict. On each occasion, groups monitoring deaths submitted a full snapshot of their data to HRDAG. For example, in May 2013, the

groups provided data to HRDAG from the start of the war until that time point. Then, in the June 2014, the same groups provided data from the start of the war until that time point. At each time point, researchers at HRDAG (under contract for the UN) used the most current version of the data to conduct a record linkage analysis to estimate the total number of observed deaths. On each occasion, the HRDAG analysis used the data as provided at that point in time.

On page 9 of the August 2014 report, the researchers note "Looking at the period of study from the previous report, this updated analysis finds that... three sources recorded 116,046 unique killings between March 2011 and April 2013. Notably, this is a higher count than the total enumeration previously reported (92,901). This is due to a combination of newly-documented deaths that occurred between March 2011 and April 2013, additional information discovered about previously-reported deaths, and refinements to the matching model."

Additional discussion with the researchers provided more details. The observer groups had alerted HRDAG that they updated their records continually, adjusting or adding records or even deleting records when new information came to light. In addition, the record linkage process was completed differently in 2013 and 2014. In 2013, the team used a classification algorithm, discussed in more detail below, after an an initial blocking of the data into potential record pairs. In 2014, all final match decisions were made by human review after the initial blocking. Using human reviewers to match record pairs is a common method for developing a test set to be used in training supervised classifiers. Human judgement of whether two records refer to the same person can be thought of as similar to a clinical diagnosis of depression. Both are subjective decisions, but such judgments are frequently considered a gold standard that more automated systems try to match. In the August 2014 report, all records were matched "by hand." This was a laudable effort, but not an especially sustainable or replicable one. Additionally, the fact that the deduplication was conducted differently in 2013 and 2014 means that comparing the deduplicated lists from 2013 and 2014 would not give an exact apples to apples comparison to determine how many new records were added.

In this paper, we attempt to assess how the data provided by the groups is changing over time, and whether and how those changes affect estimates of observed deaths and also of total deaths. To do this, we take advantage of this otherwise challenging situation, using the various snapshots of the data each group has provided.

3.2 Methods used and emerging data issues

Before we go any further, this is a good time to consider an overview of the steps needed to form an estimate of the total death toll from a conflict, starting with the raw data provided by observer groups:

1. Obtain raw data. How this is done varies by conflict and by group. In some cases,

paper records may need to be digitized, or information may need to be abstracted from narrative testimonies. In the case of Syria, observer groups generally provide raw data electronically in the form of Excel spreadsheets to HRDAG. HRDAG researchers have also scraped the websites of groups that publish data publicly.

2. Clean and standardize the datasets. Specifics will vary by conflict, but broadly, researchers decide on a minimum amount of information that must be recorded about a death for the record to be considered identifiable – that is, for it to be usable. Usually, this includes name, date, and location of death. This is the information that will be used to determine if multiple records refer to the same individual. Standard formats are applied across all lists. In the case of Syria, for example, in addition to including names in the native Arabic, researchers also included translations to English.

3. Deduplicate the datasets and match records across datasets. Together, these tasks are referred to as record linkage. This is a crucial, labor intensive step that we discuss in more detail below as well as in Chapter 4. Briefly, duplicates are removed from the individual lists as well as across the lists. This provides the estimated number of known deaths.

4. Model the data. Using the information about how many individuals appeared on multiple lists, model how many individuals are likely to have appeared on no lists. This is done using capture recapture methods, also called multiple systems estimation. Capture recapture is an umbrella term for a class of estimators that were originally developed in ecology and now used in many disciplines.

Normally, researchers using these methods collect the best available data after a conflict has subsided and proceed from there. Historically, datasets were collected from groups such as truth and reconciliation comissions and established non-governmental organzations, so researchers had some information about the methods used to create the lists and therefore expert knowledge about capture probabilities and dependency structures that could inform modeling decisions. For example, researchers might know that two groups were based in the same city and had similar political leanings, and so were likely to record the same deaths. Or they might know that one group was sympathetic to a certain faction and was more likely to record those deaths than other groups, or that a group was the only one based in the east of the country, etc. Further more, these groups often had missions that required a certain level of documentation and detail in incident recording.

In recent years, especially in the wars in Iraq and Syria, new types of groups have emerged that exist for the sole purpose of collecting names or numbers of the dead, and organizations such as Every Casualty to develop best practices and support this kind of work. Their work may involve a range of methods, from collecting first hand accounts to monitoring news reports. Such groups may or may not be in the same country as the conflict: Iraq Body Count was based in London, and the Syrian Observatory for Human Rights was as well. Such organizations may collate a great deal of data from disparate sources. However, the

data generating process for the lists they eventually produce may be harder for researchers to understand in an expert/a priori manner. Equally important, such groups may publish all or parts of their data online. Once such publication occurs, there's a possibility of groups using each other as sources, either copying records or updating their own based on the work of another group. Such publication and copying may fall exactly within the mandate of an individual group, if that group is attempting to collect as much information as possible about the war. But copying will generate dependencies between the lists.

Modern capture recapture methods allow for dependencies between data sources to be modeled. However, when the dependency structure is not known a priori, researchers rely on model selection procedures, using metrics such the Akaike Information Critiera (AIC) or Bayesian Information Criteria (BIC) to chose a model. However, it is a well known problem that sometimes models with similar values for a metric such as AIC can provide wildly different estimates of population size [20, 30]. One approach is to average across models, which has been used in [37]. However, in situations where models produce extremely different point estimates, one may still be concerned about averaging. In situations where this occurs, researchers frequently stratify data based on region and time to find more stable estimates.

Concerns about using automated model section procedures are not unique to casualty estimation or capture-recapture analysis; researchers in many fields are currently wrestling with these issues. For example, the field of causal inference is built around a philosophy for developing a model based on expert knowledge and determining a priori whether the effect of interest is identifiable from the data at hand based on the expected interactions between variables.

In this case, as well as in many similar situations where multiple groups are tracking some portion of the total death toll, the dependence structure of the lists is not know for certain. That is, the exact generating distribution of the data is unknown. In some conflicts, there may be significant subject matter knowledge available to statisticians, such as that Group A was based in City A and only took testimonials from individuals who came into their offices, or that Group B canvased communities in Region B, or that Group C was perceived as allied with the government and therefore individuals who are sympathetic to rebels may not cooperate with them. In the case of Syria, there is little such knowledge available to us. However, it is almost certain that our data sources are not independent.

A model could be developed to capture an expected dependency structure between lists, if such a structure were known. For example, one could fit a log-linear capture recapture model using appropriate interaction terms. (The data is considered as a multi-way contingency table and expected cell counts are modeled as a function of list inclusion [20, 13].

However, in our situation, as well as in many conflict settings, the dependency structure is unknown. One possibility in our case would be to fit a model with all possible interactions; that is, if one had k lists, this would mean all interactions except the k-way interactions. With moderately large numbers of lists and limited overlap between lists, one can quickly run into problems with small cell counts or even zero cells. Alternately, one could use model selection techniques to compare the fit of a range of models. Model selection is frequently employed and raises the concerns above. Additionally, we cannot be certain no k-way interaction

exists.

In this paper, we address a number of questions. First, considering a specific time frame, we examine two separate "snapshots" of data from each group. This allows us to examine how the reporting of deaths by different groups is changing over time, and explore whether there is an overarching pattern to those changes or if they are group-specific. This question is of particular interest, as researchers would like to know if there is a common time frame over which the data is likely to "settle" into its final form. We then examine how changes in the underlying data affect estimates of observed deaths. We also assess empirically whether the different versions of the lists leads to different estimates of unreported deaths. To do this, we use two different versions of the data to estimate a possible total death count using capture-recapture estimation, including model selection according to commonly used criteria.

If the model selection procedure is accurately finding the correct dependency structure of the lists – the true data generating distribution – we would expect different versions of the data to produce similar estimates (albeit possibly through different final models). However, estimates provided here are not meant to be definitive. At the moment, researchers at HRDAG are continuing to revise their record linkage model, and the final version of the linked dataset is expected to differ slightly from that presented here. However, the current version of the matching model allows us to assess whether the methods produce the same estimate when applied to the early and later versions of the data. For illustrative purposes, we focus on estimates for the governorate of Homs.

3.3 Data

To estimate known deaths in 2013, HRDAG relied on data from eight different sources. In 2015, after a meeting with several of the groups, they utilized five sources, three of which provided updated datasets. (The Syrian government only provided data covering the very beginning of the war and the Syrian Observatory for Human Rights chose to stop sharing its data after the 2013 report.) Here, I examine two "snapshots" of the data from each of the three groups that provided updates.

- Syrian Network for Human Rights (SNHR)
- The Syrian Center for Statistics and Research (SCSR)
- The Violations Documentation Centre (VDC)

HRDAG provided 2 datasets from each of the three groups. The first version of each list, which we refer to as List A for each group, was provided to HRDAG in May 2013. The second snapshot of the data, which we refer to as List B for each group, was provided to HRDAG in June 2014. For all groups, List B included updates to the data in the time period covered by List A. For this analysis, we truncated lists at April 31, 2013 to compare list coverage over the same time period.

In the versions provided to me, names were in Arabic and the governorate was already translated into English. I dropped entries with death locations outside Syria. Additionally, HRDAG provided a list of Arabic terms that included phrases such as "Unknown" or other phrases indicating that the name wasn't known, and I dropped entries where the name field matched an entry on that list. I selected complete cases based on the name, date of death and governorate of death fields, and deduplicated each list using exact matches those fields (and only those fields).

Table 3.1 shows the raw data from each group, the number of complete cases (those that included data on name, date, and location of death), and the number of unique individuals (based on exact matches of the name, date, and location fields). For these comparisons, I used the Arabic versions of the names. When translated to English, the number of unique names was slightly smaller, suggesting that sometimes more than one Arabic name maps to a single English translation. In addition, the table shows how many unique individuals appear in both versions of the data. Several things are immediately apparent. The data from SNHR contained nearly 7000 duplicates, much more than is seen in any other data source. Additionally, there are more than 20,000 additional records in List B that do not occur in List A. The data from SCSR also indicate the addition of new records, though far fewer, while the data from VDC is much more similar in each version of the list. We explore each organization's data in more detail below, using the deduplicated versions of the data.

For each organization, there were noticeable differences between the data provided in 2013 and 2014 for the data covering the same time period. Each organization showed a different type of change. Highlights include:

1. The Syrian Network for Human Rights showed approximately 20,000 more records in List B.

2. The Syrian Network for Statistics and Research showed approximately 4000 more records in List B.

3. The Violations Documentation Centre showed about 7000 records that only occurred on List A and a similar number that only appeared on List B. This is suggestive of an updating process, though we don't know for certain whether the records were actually updated or whether some were dropped and new ones added.

3.4 Within list comparisons

Table 3.1: All List: complete and unique cases; matches across versions of the group's data as provided in 2013 (A) and 2014 (B)

	List A	List B	Matched
SNHR Original	57422	81364	NA
SNHR Complete Case	57414	81364	NA
SNHR Unique	50855	75000	45143
SCSR Original	47789	52472	NA
SCSR Complete Case	47756	52402	NA
SCSR Unique	47102	51204	46897
VDC Original	70383	69621	NA
VDC Complete Case	70262	69561	NA
VDC Unique	68412	67924	60860

SNHR Data

The most striking differences between data provided in 2013 (List A) and data provided in 2014 (List B) was seen with the SNHR data: approximately 20,000 new unique records appeared in the data in 2014. Additionally, nearly 5000 records were dropped or updated between the two datasets. (If they were dropped, then the total added in 2014 is even higher.) The plots that follow show an overall comparison of lists followed by more detailed breakdown of the 2014 and 2013 data. The addition of death records varies by governorate and by month. There is not a strong pattern geographically, though governorates that showed the highest percent added tended to be those with lower total death tolls. Looking specifically at List A, we can see that while a moderate number of deaths was added to the early months of the war, those deaths constitute a high percentage of the deaths captured during that period. Looking at List B, it appears that a high number of the deaths that show up only on that list are also from the early months of the war, suggesting possible updating.

The overall addition of records between 2013 and 2014 raises questions. If the data were copied from another source, that knowledge should inform the use of the data in estimating unknown deaths. However, other explanations for the change are possible, and it would be useful to ask the organization about the differences.

Figure 3.1: SNHR Deaths through April 31: complete cases and unique cases on each list, and matches

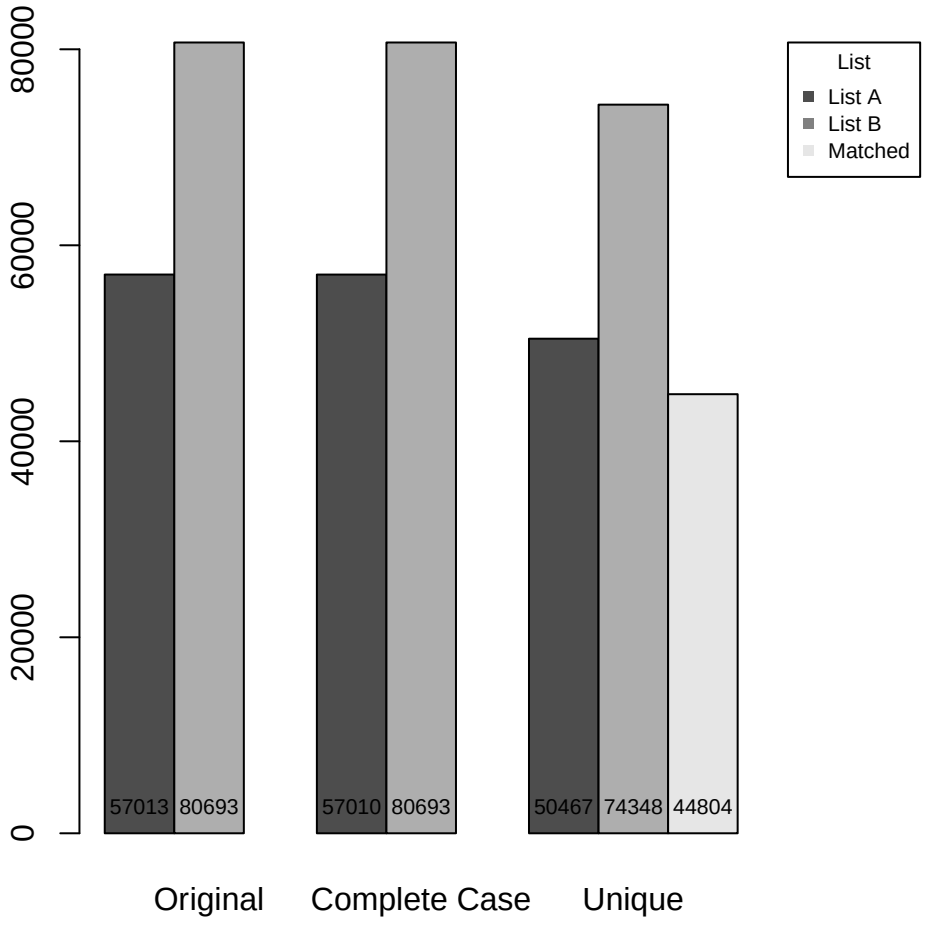


Figure 3.2: SNHR deaths by governorate. Light gray indicates records that only occur on that list

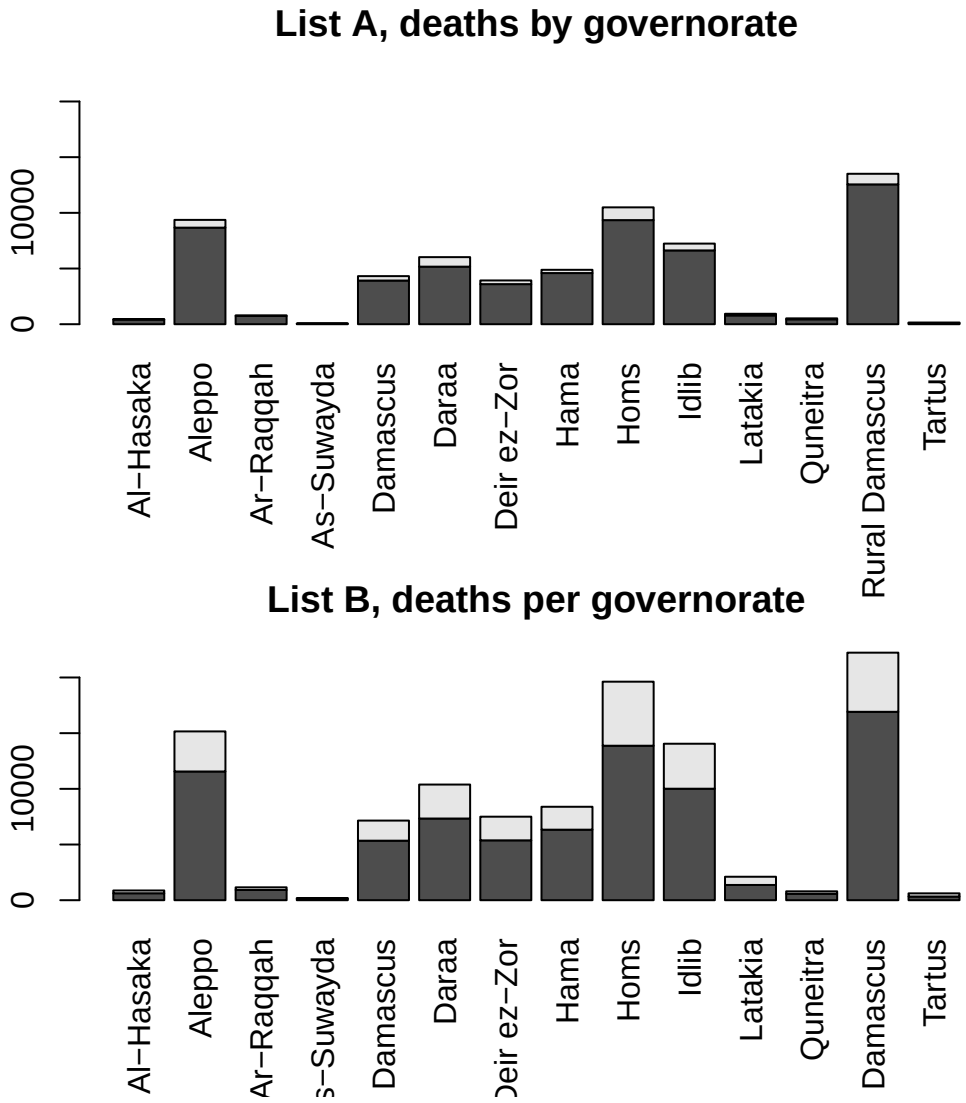
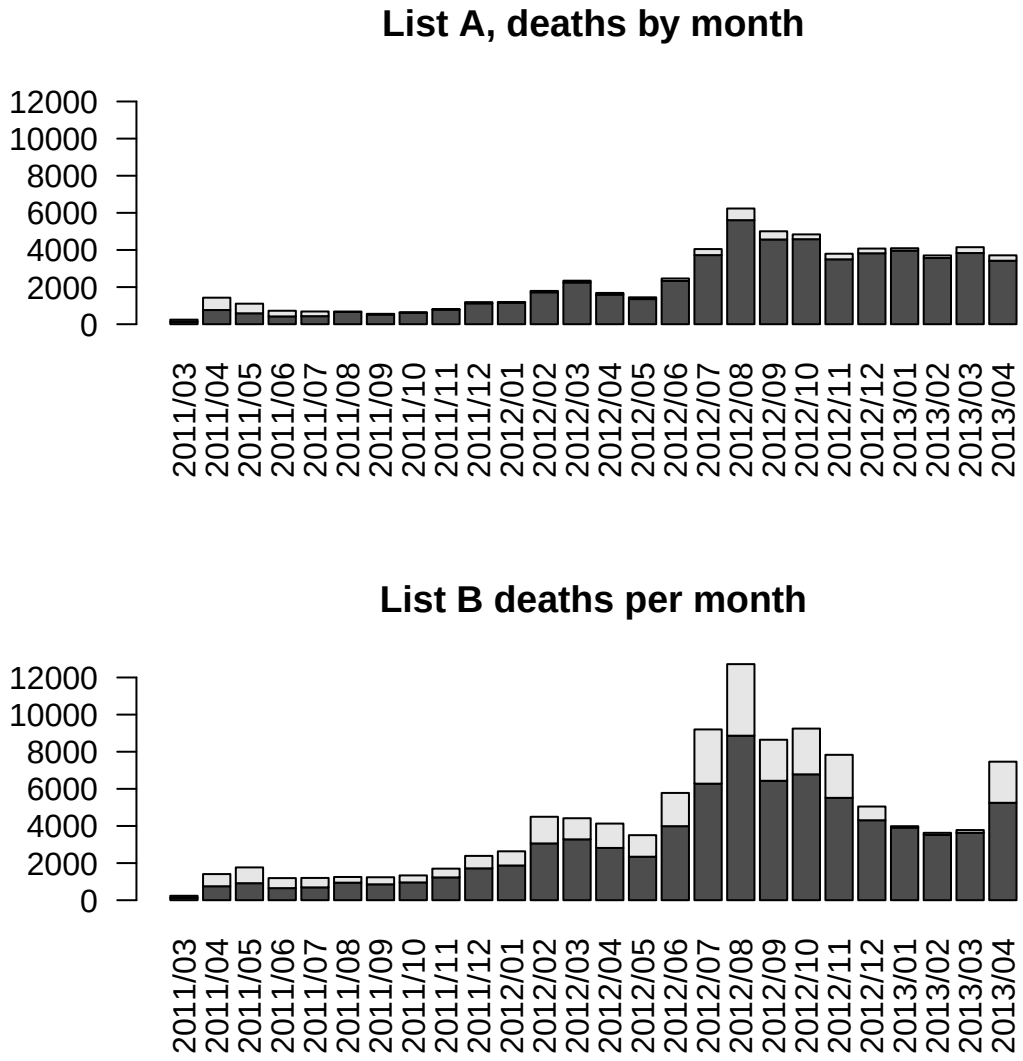


Figure 3.3: SNHR deaths by month. Light gray indicates records that only occur on that list



SCSR Lists

This is the only list where first and last names were considered in separate fields. For purposes of deduplication, both names had to match. There more approximately 4000 more records on the version from 2014 (List B), and only 205 records that appeared in only 2013 (List A). Looking at the 2014 data, a large number of the deaths were added in Aleppo, As-Suwayda, and Quneitra. The greatest percentage additions were in As-Suwayda and Quneitra. The greatest number of deaths were added between October 2011 and and August 2012; however, large percentage increases occurred in the early months of data collection.

Figure 3.4: SCSR Deaths through April 31: complete cases and unique cases on each list, and matches

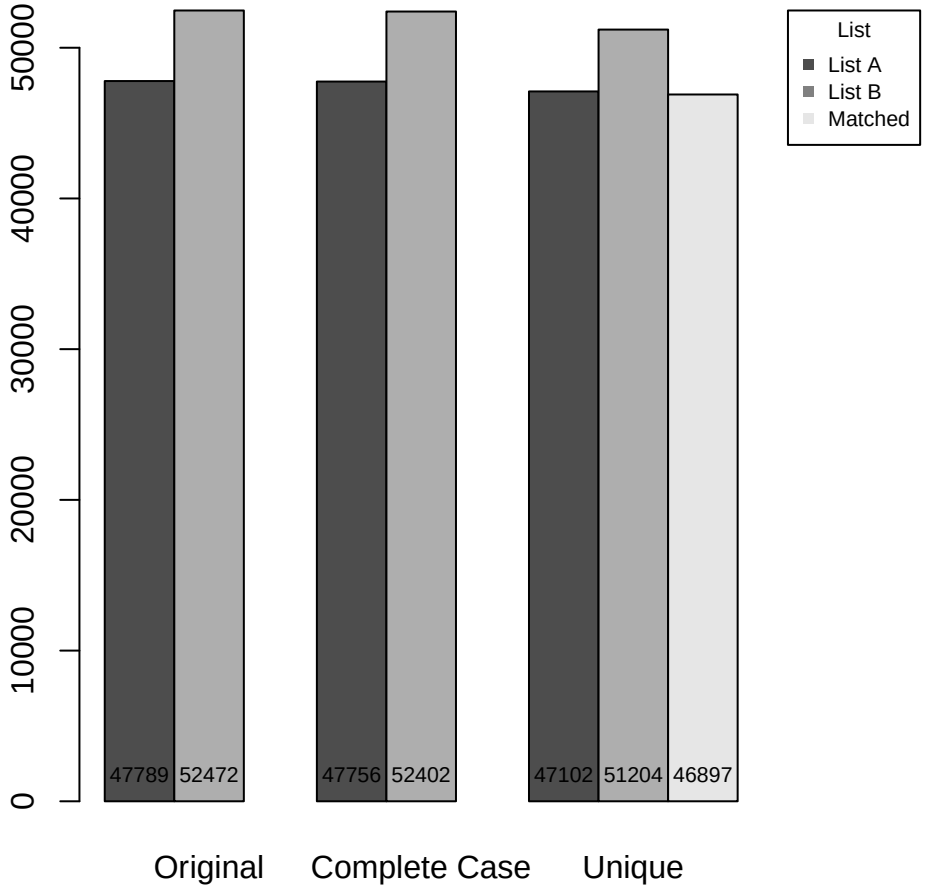


Figure 3.5: SCSR deaths by governorate. Light gray indicates records that only occur on that list

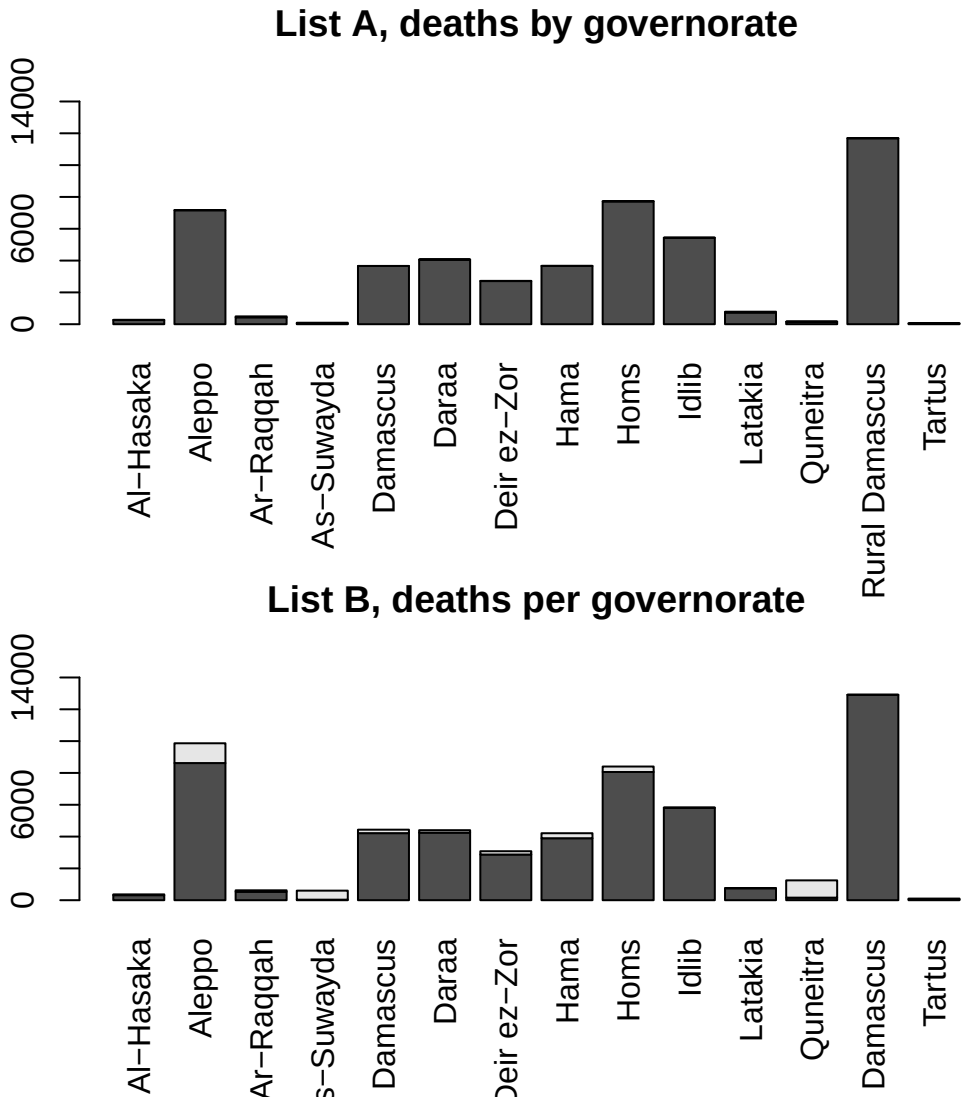
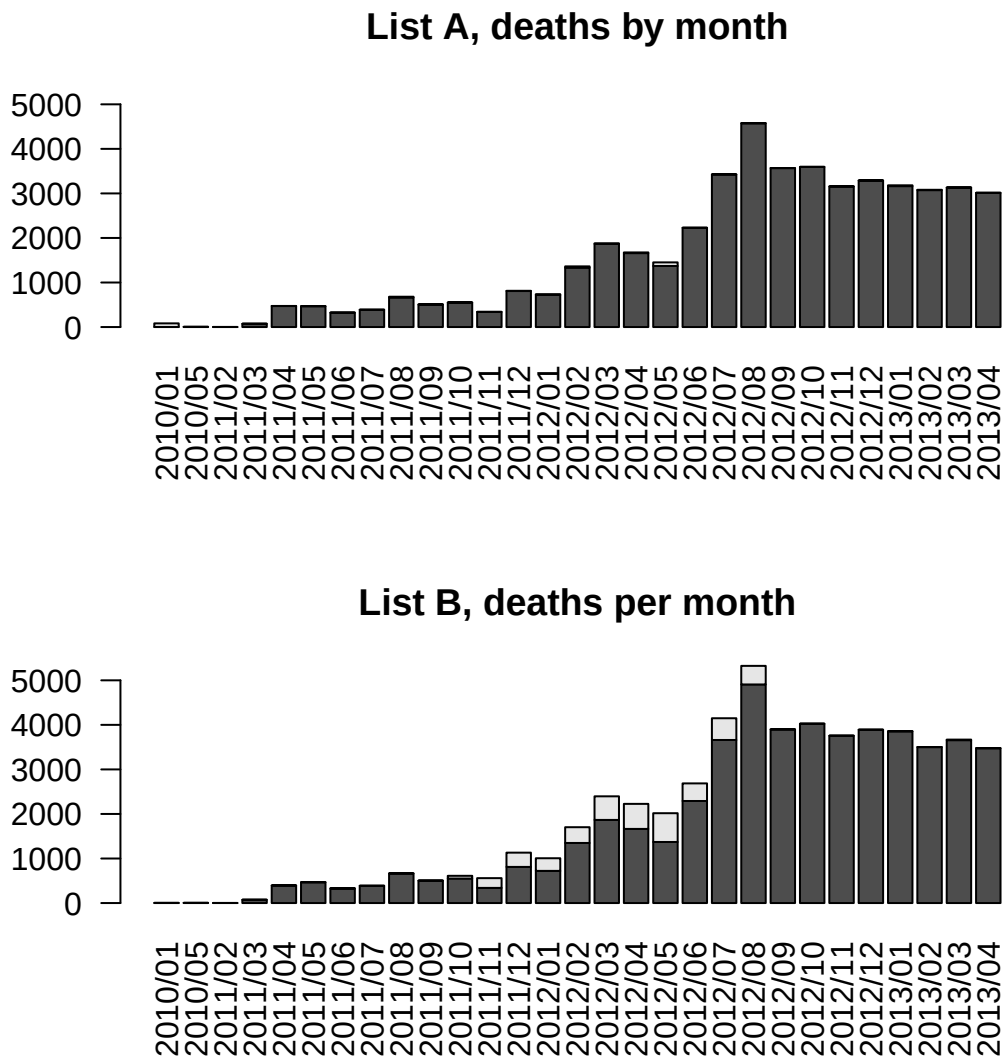


Figure 3.6: SCSR deaths by month. Light gray indicates records that only occur on that list



VDC Data

The overall numbers of records was very similar for both versions of the lists. The VDC was the only organization for whom the total count decreased slightly between the 2013 (List A) and 2014 (List B) versions of the list. Also, more than 7000 records differed between the lists when the lists were compared by exact matches on name, date, and governorate of death. This indicates that either the records were updated, or that 7000 were dropped from the list in 2013 and then another 7000 were added to in 2014, or perhaps a combination of updates, drops and adds occurred.

Figure 3.7: VDC Deaths through April 31: complete cases and unique cases on Each List, and matches

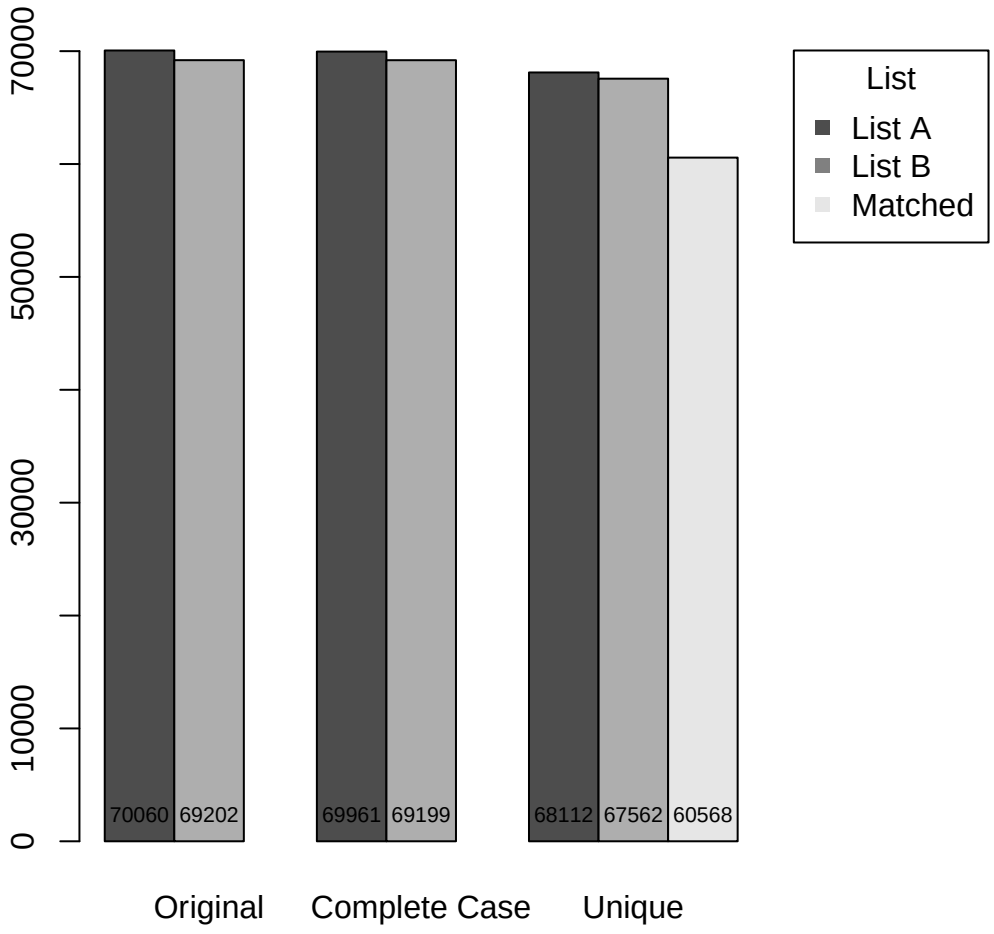


Figure 3.8: VDC List: deaths by governorate. Light gray indicates that deaths occur only on that list.

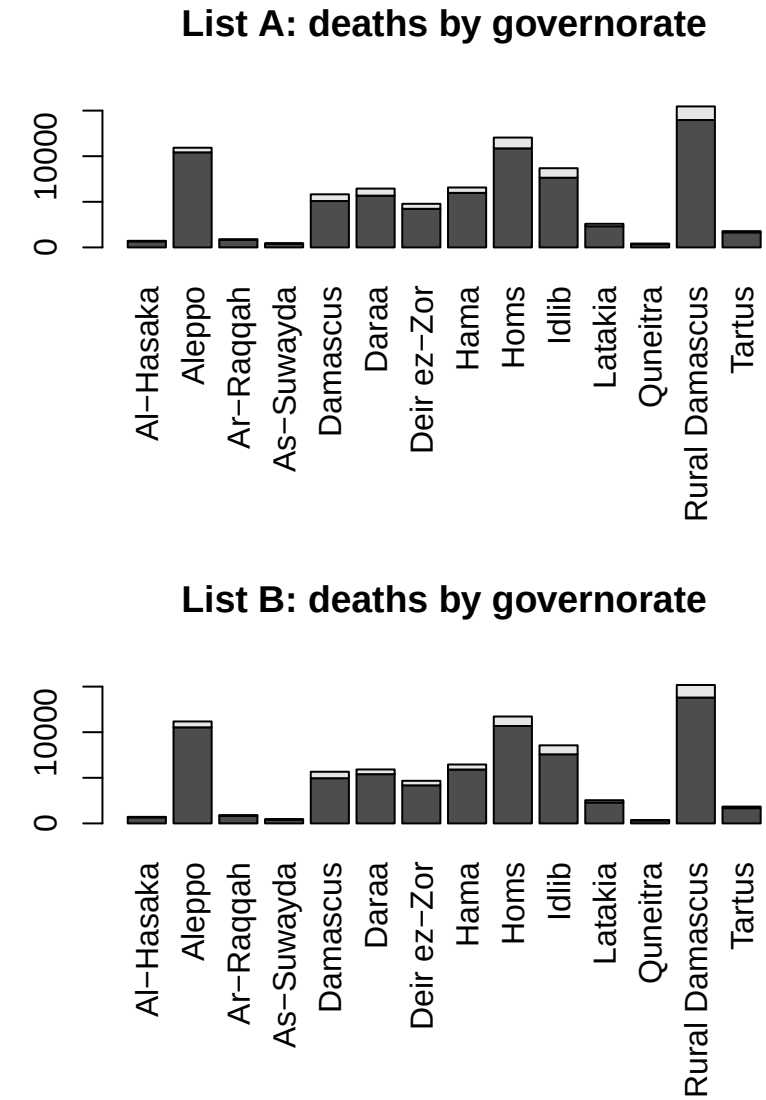
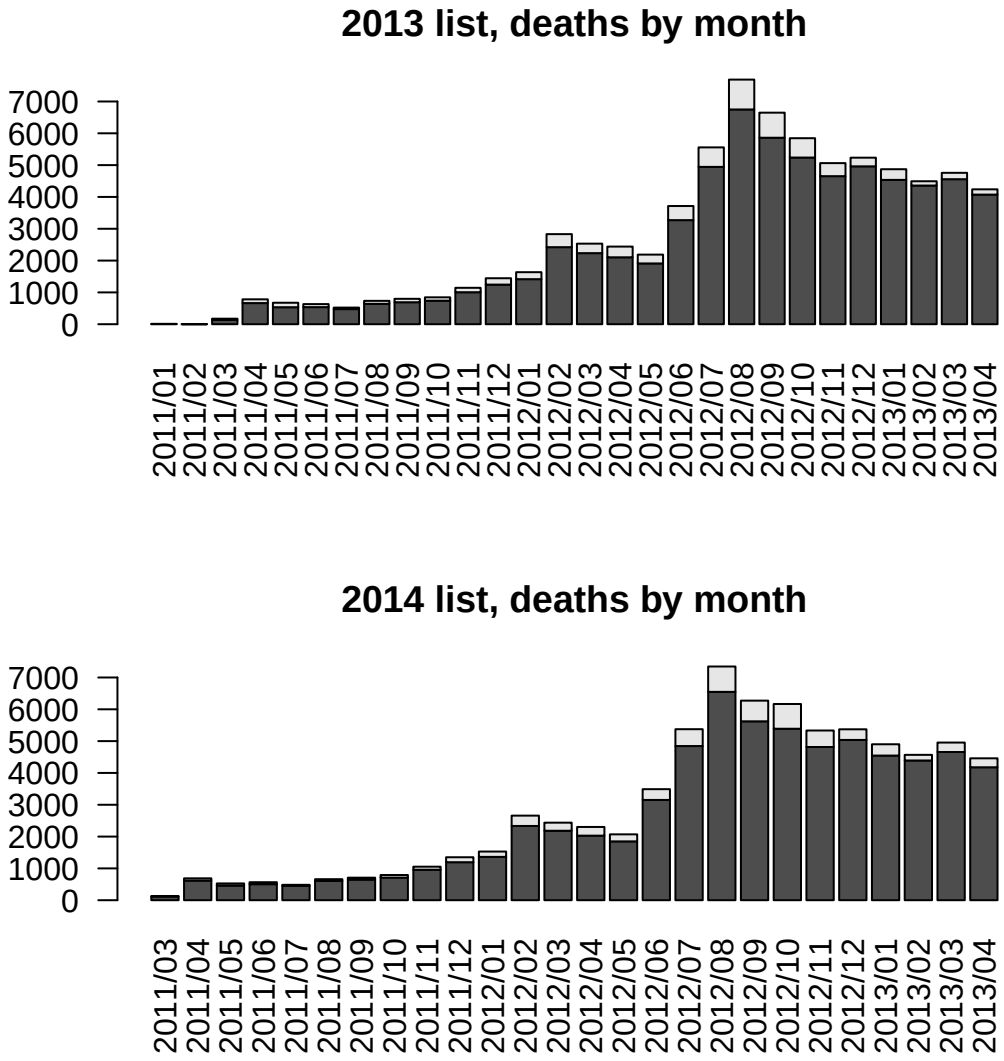


Figure 3.9: VDC Lists: Deaths by month. Light gray indicates that record occurs only in that list.



Missingness Tables

Table 3.2: VDC Missingness, 2013 list

	Missing	Percent
photo	46475	66.03
name	55	0.08
governorate	22	0.03
sex	22	0.03
age_group	22	0.03
age	51708	73.47
civilian_status	10250	14.56
mother.s_name	68443	97.24
family_status	61077	86.78
area	7052	10.02
occupation	67213	95.50
national_number	70263	99.83
id_card_number	70279	99.85
date_of_death	115	0.16
location_of_death	52729	74.92
cause_of_death	843	1.20
rank	48973	69.58
notes	31101	44.19
video_of_the_martyr	52743	74.94
funeral_procession	66773	94.87
facebook_page	66895	95.04
news_article	65328	92.82
generic_video	67535	95.95
list2013	0	0.00

Table 3.3: VDC Missingness, 2014 list

	Missing	Percent
name	1	0.00
governorate	57	0.08
sex	1	0.00
age_group	46904	67.37
family_status	56254	80.80
mother.s_name	65042	93.42
age	46904	67.37
area	5922	8.51
occupation	64893	93.21
national_number	69264	99.49
id_card_number	69288	99.52
date_of_death	4	0.01
location_of_death	50176	72.07
cause_of_death	1	0.00
rank	46411	66.66
notes	21171	30.41
civilian_status	10303	14.80
video_of_the_martyr	51495	73.96
funeral_procession	65288	93.78
facebook_page	65019	93.39
news_article	64305	92.36
generic_video	65979	94.77
photo	14510	20.84
list2014	0	0.00

Table 3.4: SCSR Missingness, 2013 list

	Missing	Percent
record_id	0	0.00
first_name	0	0.00
father_name	20846	43.62
last_name	1	0.00
sex	1	0.00
place_of_birth	0	0.00
age	36380	76.13
occupation	44663	93.46
marital_status	45901	96.05
num_children	0	0.00
health_status	47709	99.83
governorate	0	0.00
location_of_death	3455	7.23
date_of_death	32	0.07
trial	47789	100.00
trial_date	0	0.00
details	4861	10.17
note	40572	84.90
list2013	0	0.00

Table 3.5: SCSR Missingness, 2014 list

	Missing	Percent
surname	1	0.00
name	0	0.00
father_name	24924	47.50
governorate	0	0.00
date_of_death	69	0.13
sex	1	0.00
born_state	80	0.15
born_town	4776	9.10
age	40621	77.41
incident_town	3731	7.11
incident_details	4806	9.16
record_id	0	0.00
list2014	0	0.00
month_year_death	69	0.13
first_name	0	0.00

3.5 Initial estimates from different versions of the data

The above section demonstrates that the data maintained by at least three of the observer groups changed between the time points at which data was shared, even when we limit ourselves to comparing identical time periods. The next question is, do these changes affect resulting estimates of the total death toll, including unobserved deaths.

As described above, HRDAG used different record linkage processes in 2013 and 2013, which means that the two sets of deduplicated records are not an exact apples to apples comparisons. However, the previous section demonstrates that a great deal of the changes between those time points involved actual changes to data provided by the organizations, and it is informative to see whether using the different versions of matched data provide different estimates of total deaths.

As an example, we consider data from the governorate of Homs, including deaths that occurred between January 1, 2012 and before April 30, 2013. We can see in table 3.5 that the total number of recorded changes for each group, though the changes are not identical for each group. Both VDC and SCSR showed similar numbers of records on each list, with the later dataset have slightly fewer than the earlier on. SNHR added more than 2000 records.

Furthermore, as we can see in Table 3.5, the overlap patterns between the lists change substantially. For example, a number of deaths that appeared in SCSR's 2013 list appear to no longer show up in the 2014 version, leading to fewer than half has many deaths appearing on all three lists.

We used Rcapture [5] to fit all possible hierarchical loglinear models (that is, all hierarchical models except the one with a 3-way interaction term). We then ordered the top ten models by AIC. As we can see, this model selection procedure chooses the same choice top model with each version of the data – the model with all two-way interactions terms. As we look down the list, however, the models are ranked differently. This is to be expected, as the dependency structures between the lists have likely changed, especially if there is copying between lists. However, it is more concerning that the resulting point estimates are quite different, as well. Using the 2013 version of the data, the estimate suggests approximately 20,000 unreported deaths – more than twice as many as reported. Using the 2014 data, the model predicts more than 110,000 unreported deaths, more than seven times as many as were reported.

An important next step will be to redo the record linkage for each set of lists using the same procedure to ensure that these results are not in part an artifact of changes to the record linkage process.

	2013 Data	2014 Data
in_SNHR	6334	8880
in_VDC	8129	8016
in_SCSR	5647	5543

Table 3.6: Total count for each list. Data on deaths in Homs between January 1 2012 and April 30, 2013.

	in_SNHR	in_VDC	in_SCSR	2013 Data	2014 Data
111	1	1	1	3540	920
110	1	1	0	1058	5023
101	1	0	1	603	347
100	1	0	0	1133	2590
011	0	1	1	706	105
010	0	1	0	2825	1968
001	0	0	1	798	4171
Total				10663	15124

Table 3.7: Totals for each capture history. Data on deaths in Homs between January 1 2012 and April 30, 2013.

	abundance	stderr	AIC	BIC
12,13,23	30738	1660	77	128
13,23	13688	152	885	928
12,13	13856	184	1041	1085
13,2	12013	55	1386	1422
12,23	12162	100	2834	2878
23,1	11606	46	2920	2956
12,3	11488	44	3010	3046
1,2,3	11395	34	3024	3053

Table 3.8: Estimates of total deaths (observed and unobserved) from hierarchical models fit on data from data provided by SNHR, VCR and SCSR in 2013. Estimates are ordered by AIC. The lefthand column describes the top level interaction terms.

	abundance	stderr	AIC	BIC
12,13,23	121998	13017	76	130
12,13	93300	7928	97	142
12,3	44251	969	250	289
12,23	46256	1852	251	296
13,23	16139	46	2909	2954
13,2	17437	82	5989	6027
23,1	17695	85	6013	6051
1,2,3	19811	121	8068	8098

Table 3.9: Estimates of total deaths (observed and unobserved) from hierarchical models fit on data from data provided by SNHR, VCR and SCSR in 2014. Estimates are ordered by AIC. The lefthand column describes the top level interaction terms

3.6 Discussion

This initial analysis demonstrates that indeed, the lists of deaths recorded by human rights groups in Syria are dynamic entities, and that groups are not only adding new records as the conflict progresses but are also adding and/or changing records of deaths from earlier periods. It will be important to repeat this work with later snapshots of the data to assess whether the same level of change is seen in records from the first year of the war, or if the data does, in fact, settle over time. Already, we can see that there is no clear pattern of how groups' data change generally, but that the process seems to be unique to each group. This suggests that this kind of analysis should be considered not as indicating "how reporting changes during ongoing conflict", but rather an indicator that this issue needs to be assessed in each situation.

Additionally, this analysis suggests that concerns about the need to better understand the data generating process when using "real time" conflict data to estimate a total death toll are warranted. The analysis of data in Homs is not meant to produce a final number; researchers may stratify further on time as well as assess different models. However, this is a reasonable enough facsimile of what will be done to raise important questions. An important next step will be to repeat the analysis using two versions of the data that have been matched by the same model. The review of the raw data highlights that are clearly substantive differences to the raw inputs into matching and estimation. However, it will be important to assess whether any of the variation we see in list overlap and in estimates of unobserved deaths are due to differences in the matching model.

Chapter 4

Classifying Record Pairs

4.1 Introduction

Determining whether two records refer to the same individual is a problem that arises in many fields. In situations where records are indexed by unique identifier such as a social security number, the task is trivial. However, frequently researchers want to link records in cases where there is no such identifier that is common across datasets. In these cases, the determination of whether two records refer to the same person can be made using the information in the record. For example, name, birthday, and address fields can be used to help decide whether multiple records refer to a single individual.

The easiest way to determine whether multiple records refer to the same person is for a human to review the records and make a decision. Such human determinations are frequently used as a gold standard against which other classifications are measured. While imperfect, human record review is akin to using a doctor's diagnosis of a medical condition as a gold standard. In either case, the human (doctor or record reviewer) might make a decision that's different from the underlying truth, but the human decision making process is considered to be the closest we can get to that truth. Humans are able to take advantage of a great deal of external and contextual information. Doctors bring many years of training to bear on each individual diagnosis. Human record reviewers bring an intuitive understanding of the prohibitive value of many pieces of information (is the name Chris or Humberto more individualizing? Does the answer to that question vary based on location?)

However, when comparing large lists of individuals, it is not feasible to have humans review all possible record pairs to determine which refer to the same person. The question seems ripe for a statistical approach, and has in fact been studied in both the statistics and machine learning fields [14, 60, 50]. In both contexts, the problem is generally referred to as record linkage. Record linkage encompasses both the task of linking records across datasets (matching), and also linking records within a dataset (deduplication).

The simplest way to decide matches is a rule-based approach, where if certain conditions are met, the pair is declared a match. In 1969, Fellegi and Sunter formalized a probabilis-

tic approach to record linkage, determining an optimal rule for classifying record pairs as matches, possible matches, and non matches [19]. Record Linkage has also been approached as a binary classification problem, an approach we investigate here. Considering each record pair as a unit, we wish to determine whether that pair refers to a single individual or two distinct individuals. This is discussed in more detail below. Briefly, using training data where the match status is known, we train a classification algorithm, which can then be used to classify record pairs where the truth is not known. In this paper, we focus on evaluation of various algorithms.

Record linkage problems arise naturally in human rights investigations. It is common for more than one observer or advocacy groups to be attempting to document crimes. Frequently, officials or others wish to know how many total crimes occurred or how many distinct individuals were affected. A deduplicated compilation of the data from multiple groups can answer those questions. A deduplicated list can also serve as a floor (or minimum) for the total number of violations, though it's important to remember that additional crimes that were not recorded by any group. No observer group or news reporter can capture every individual crime committed during a conflict; the observed data never tell the entire story. Record linkage is also an important step in addressing the question of unrecorded crimes; it is needed to prepare the data for modeling when one wishes to estimate the number of unrecorded violations. As discussed in Chapter 3, using capture-recapture methods to model the relationships between data sources relies on the assumption that records in individual lists can be matched. Additionally, in these context individual data sources can contain duplicate records.

In this paper, we examine data from Syria, where multiple groups are documenting war deaths. Each group only sees a subset of the total deaths that occurred. Those subsets overlap; however, each subset contains some deaths that are unique to that group that created it. A natural question is, how many total unique deaths are documented by all the groups? In the early years of the conflict, the United Nations Office of the High Commissioner for Human Rights contracted with statisticians to answer that question. The Human Rights Data Analysis Group (HRDAG), a nonprofit statistical consultancy based on San Francisco, produced three reports with the United Nations [45, 46, 44]. The UN contract is now completed, but HRDAG has continued to collaborate with observer groups in Syria to update estimates. Currently, they are working with three lists that contain approximately 360,000 total records, describe in detail by Patrick Ball in a post on HRDAG's website. Each list contains some duplicated individuals, and there are many instances of the same individual being recorded by multiple groups.

We describe the data in more detail below. However, we can immediately see the problem of scale here. Checking all pairwise comparisons of records would be an n by n problem, and the overwhelming majority of those comparisons would be of records for different people. For example, the rates of within-list duplication are low, so most pairwise comparisons within a list are of distinct individuals. Researchers currently deal with this problem by developing blocking rules to determine which records to compare [14]. Often these are ad hoc rules about proximity of fields based on researcher's prior experience or expectations. The current

blocking rules developed at HRDAG for the Syria data were developed through an adaptive procedure described on the organization's web site, and bring the number of pairs requiring classification down to approximately 43,000,000. This is still a large number. Existing packages designed to handle deduplication such as RecordLinkage in R become unstable with datasets of fewer than 1 million. Additionally, they are not easily extensible beyond provided classification options.

HRDAG researchers have decades of experience dealing with large datasets; however, they are continually refining their methods, and each of the three reports produced for the UN followed a different methodology, including complete human review of matches in the third report. However, in general they employ an iterative process that works roughly as follows:

1. Develop training data. Determine an initial set of record pairs to classify and have human matchers make determinations of whether they refer to the same person. The training pairs are selected purposively to ensure a reasonable number of matching pairs, as matches are a rare occurrence if all n by n possible record pairs are considered. This selection can be done through blocking rules based on closeness of names and dates, for example. The training set will be refined as researchers iterate through this process.

2. Develop a set of comparison vectors for all candidate record pairs. This creates a dataset with one row for each record pair. Features include fields comparing names, location, and date of death. Comparisons include exact matches, partial matches, edit distances, phonetic matches, etc.

3. Determine which record pairs to compare. This can also be done initially through a set of ad hoc rules designed to create blocks of record pairs that include likely matches. Such rules can include closeness of names or dates, for example. Recently, HRDAG switched to an adaptive blocking procedure. This step produces what HRDAG refers to as candidate pairs, a subset of all possible pairs.

4. Using the training data, train a classification algorithm. A variety of algorithms can be used; currently, HRDAG is primarily using gradient boosting.

5. Using pairwise classifications, generate transitive match groups. Transitive match groups are networks where A matches B, and B matches C, and C matches D and E, etc. If groups are too large to plausibly refer to a single individual, employ a clustering algorithms to search for groups that likely represent unique individuals. Currently, HRDAG uses hierarchical agglomerative clustering for this step.

6. Examine results and iterate through the process again. If there are types of pair matches where the classification is borderline – or wrong – draw more similar record pairs for human review to enrich the training set.

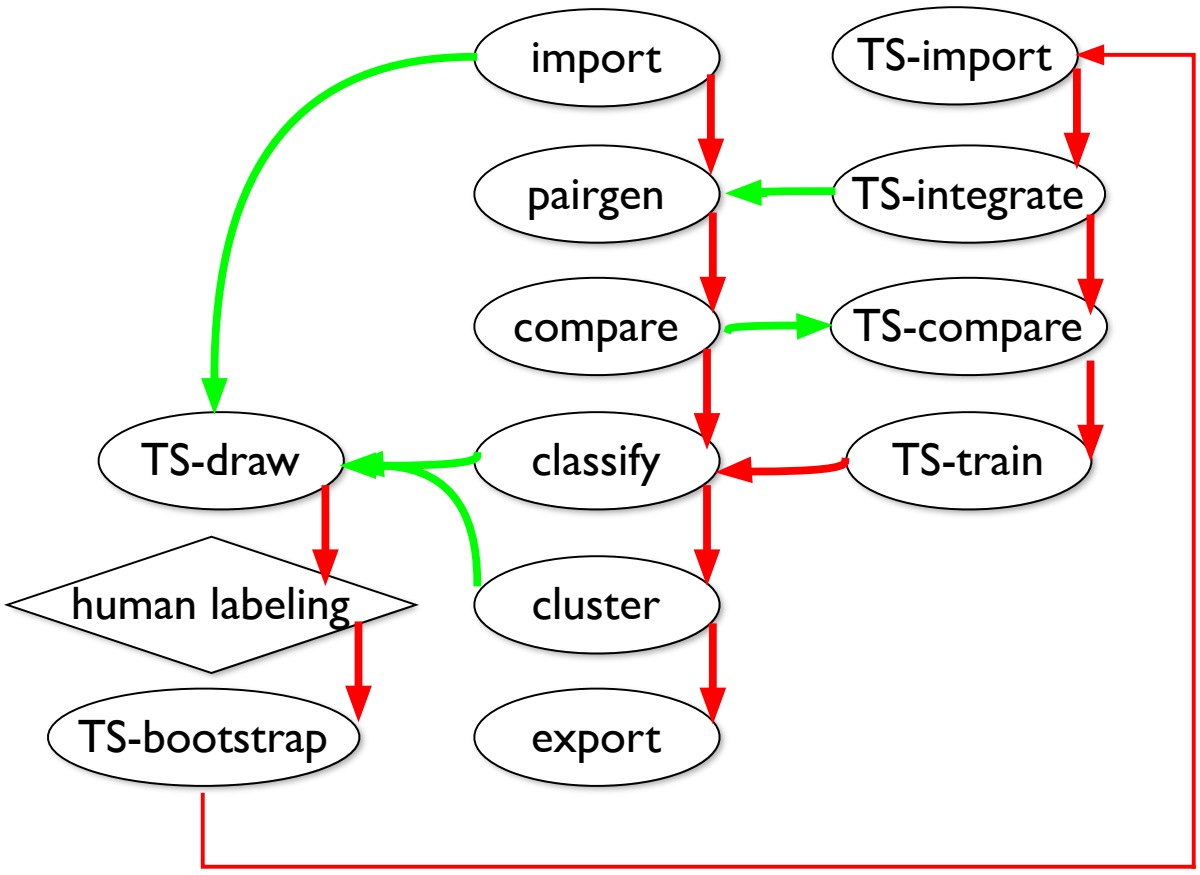


Figure 4.1: HRDAG's graphical representation of the record linkage process

Figure 4.1 is HRDAG's graphical representation of the process, which highlights it's iterative nature, particularly the refinement of the training data.

In this analysis, we focus on step 4 of this process, the classification of record pairs. We compare an implementation of the SuperLearner version of stacked regression to the component classifiers included in the ensemble, including gradient boosting and random forests. This analysis does not complete the process of record linkage. To determine the number of unique individuals, it's essential to determine transitive closures. However, generating classifications that are as correct as possible is crucial to generating transitive groupings that refer to one unique individual. Perfect classification would lead to perfect transitive groupings, and no further analysis would be required.

4.2 Classification

The problem of classifying record pairs as matches (two records that refer to a single individual) or non-matches (two records that refer to distinct individuals) can be framed as a binary classification problem where each pair represents a single record with a binary outcome (match/non-match). The data can be considered to be n observations of $O = (X_i, Y_i), i = 1, ..., n$ where X_i is a vector of comparison features and Y_i is an indicator of whether the two records used to form X_i refer to a single individual. Our observations O are not truly independent, given that conflict frequently kills in a clustered fashion.We wish to estimate a function that will map X_i into a prediction for Y_i.

Classification is a common problem, and many algorithms have been proposed, including regression models and non-parametric tree-based methods. More modern methods generate predictions from an ensemble of weaker models. Bagging (bootstrap aggregation) generates an average prediction from models fit on bootstrap samples of the data. Random Forests [10] is the classic example of a bagging algorithm, generating a prediction from an ensemble of classification and regression trees fit on bootstrap samples of the data, and is still widely used. Boosting methods build a series of models sequentially, iteratively reweighting the observations such that poorly classified observations are given more weight in subsequent fits [21]. Stacked regression provides a method to generate ensemble predictions from a diverse set of base algorithms. Presented by Breiman [11], the properties of stacked regression were studied by van der Laan [57], who proved the oracle properties of the algorithm, providing a theoretical foundation for stacking and introducing the term SuperLearner.

Simpler parametric models such as logistic regression are well understood in many disciplines and provide easily interpretable parameters; however they rely on strong assumptions about the data generating distribution, which are generally untrue. Non-parametric and semi-parametric methods generally provide superior prediction at the expense of easy model interpretation. Such predictions can be used to estimate components of interpretable parameters [58]. Frequently, however, accurate prediction is a problem in itself, as is the case here.

The SuperLearner Algorithm

The SuperLearner algorithm has been described in [57, 58]. Briefly, SuperLearner is a method of stacking models that weights the contribution of each algorithm according to its predictive value, using cross validated estimates risk. Original versions of stacking focused on ensembles of a similar base algorithm, varying the values of tuning parameters. Such ensembles are still useful; however, with SuperLearner, often the focus is on combining a diverse set of base algorithms, as different algorithms may model different aspects of the data better. Informally, this can be thought of as crowdsourcing a characterization of the underlying truth. Rather than attempting to generate the one best model to capture any data generating distribution, SuperLearner asks existing algorithms to do the best job they possibly can and then finds the best possible combination of those algorithms. As such,

SuperLearner relies on work done by a range of models developed by diverse groups over many years, and it can be easily expanded to include new algorithms as they are developed.

Briefly, for a set of n observations of $O_i = (X_i, Y_i)$, where Y_i is the outcome or response variable, and X_i is a vector of predictor variables, or features, the algorithm is:

1. Define the prediction problem and choose an appropriate set of L algorithms for the ensemble.
2. Fit each algorithm on the data O using $K - fold$ cross validation. Generate cross validated predictions.
3. Create a n x L matrix (Z) of predicted values, with each of the L columns a vector of predictions for each learner, and each of the n rows an observation.
4. Train a metalearner using the Z matrix to predict Y.
5. Then with each of the L baselearners, predict the outcome on the full dataset.
6. Use those predicted values to form a Z matrix and the use the metalearner fit determined in the previous steps to predict the outcome Y

For this analysis, we fit base models including a logistic regression, random forests, gradient boosting, and a deep learning neural net. For the metalearner, used a binomial glm with an elastic net penalty (to handle correlated predictor set). We conducted the analysis using the H2O implementations of all algorithms, with model parameters set to default values, specifically:

- Logistic Regression
- Random Forests
  - 50 trees
  - Sample rate of .632
  - Maximum tree depth of 20
  - Minimum of 1 row per terminal node
- Gradient Boosting
  - 50 trees
  - Sample rate of 1
  - Maximum tree depth of 5

  - Mimi mum of 10 row per terminal node

- Deep Learning neural net

  - 200 hidden layer size
  - 10 epochs

4.3 The Data

Researchers at HRDAG are involved in an ongoing analysis of data from four groups monitoring deaths in Syria. Since 2013, HRDAG has reviewed data from at least eight groups, and has used different subsets in previous reports. Over time, some have merged, others only provided data for a limited period of time, and, others have chosen to pursue advocacy efforts not focused on data collection. In their ongoing work, researchers at HRDAG have received repeated data updates from multiple groups, so the dataset has evolved over time. This process is described in more detail in Chapter 3. Briefly, the groups have provided repeated "snapshots" of their data. In February 2016, HRDAG provided us a snapshot of their most current processed data at that point. For the analysis, we focus on that one version of the data.

The dataset we analyze here was based on records from four groups: the three discussed in Chapter 3, along with data from an new organization, the Damascus Center for Human Rights in Syria. All groups maintain networks of observers and researchers inside Syria. The directors of all organizations, and the headquarters of most, are based outside of Syria for security reasons. The security concerns are not hypothetical. Razan Zaitouneh, a human rights lawyer and opposition leader with the Violations Documentation Centre has been missing since 2013, when she was dragged from the organization's offices with three colleagues [22]. Remarkably, VDC is one of the only organizations to still maintain an office in Damascus.

The organizations are:

- Syrian Network for Human Rights (SNHR)
- Syrian Center for Statistics and Research (SCSR)
- Violations Documentation Centre (VDC)
- Damascus Center for Human Rights in Syria (DCHRS)

The analytic dataset was comprised of feature vectors based on comparisons of records from all four organizations. That is, each row in the dataset is based on a comparison of two records. As provided to us, the dataset included training and testing data, with indicators of whether the pair had been determined to be a match or non-match based on human review, and a large set of record pairs for which no true status was known.

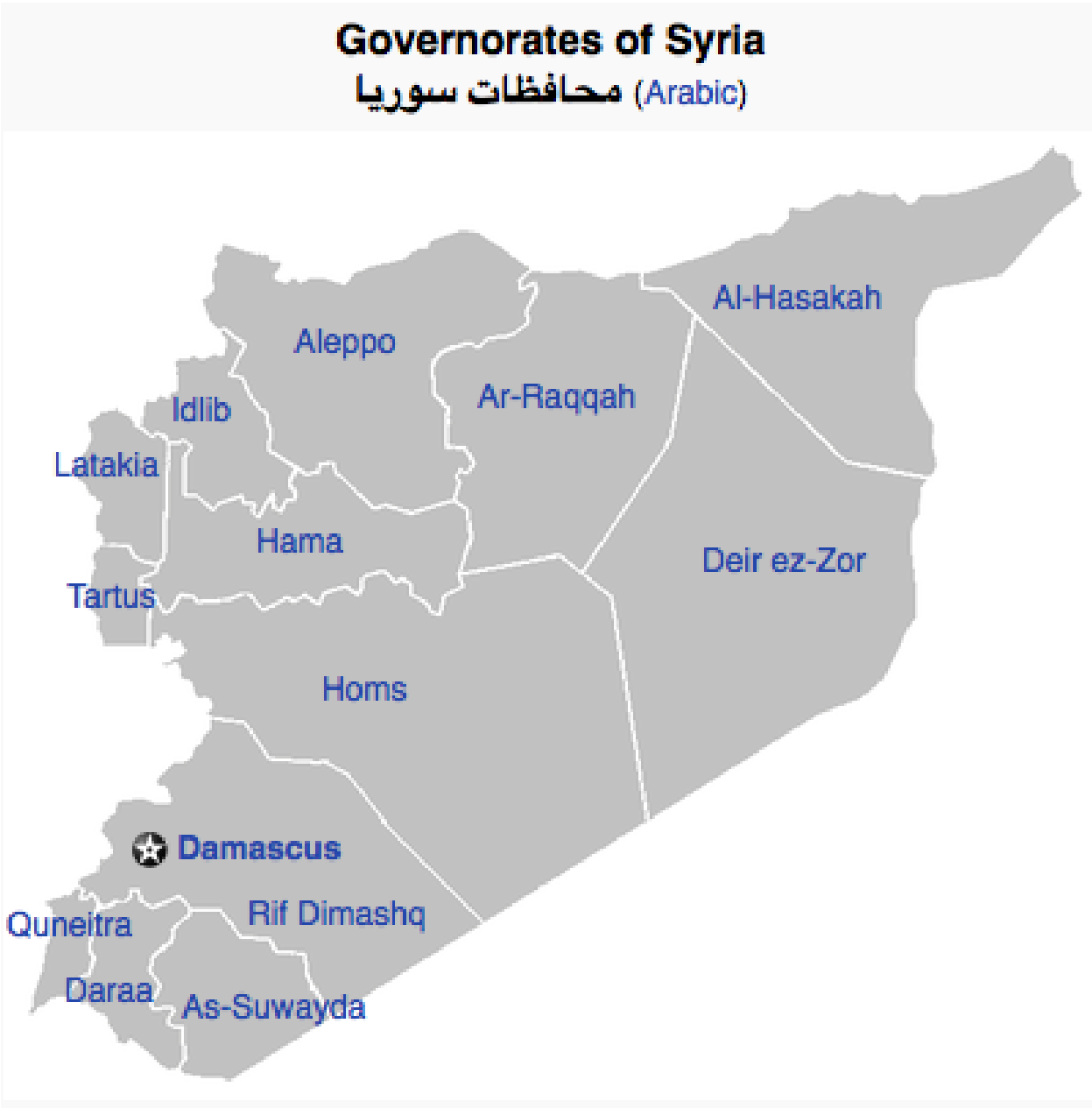


Figure 4.2: Map of Syria's governorates. Used with permission.

As the researchers have been working with versions of this data for more than three years, they have developed a remarkable set of human-classified data: 11,567,888 record pairs with 132,146 matches. (For the third UN report, all matches received human review, greatly enriching the total known match/non-match pairs) A training set was developed from the full set of human-classified match/not match pairs, and was enriched for matches and "hard to classify" record pairs. Hard to classify pairs are those that previous classifiers have failed to identify correctly as a match or non-match, or pairs where the predicted probability of being a match was close to the decision boundary, even if the match decision was correct. (These are also referred to as least informativity pairs) The training set included 120,407 total records with 12,930 matches. An additional 11,432,021 records with 114,815 matches are available for validation and testing.

The dataset also contained 43 million record pairs for which no true status was known. These 43 million records had been selected by an adaptive blocking scheme out of a possible 130,000,000,000 pairs, to capture as many potential true pair matches as possible while limiting the dataset to a manageable size. (The blocking procedure was developed using training data. For details, see the HRDAG website). Those record pairs were not used in this analysis, as we sought to test different classification algorithms on data for which there is a gold standard to test against. However, to determine the total number of deaths, as well as to prepare the data for a capture-recapture analysis, the next step would be to use the best algorithm to predict the classification of these 43 million record pairs as matches or not matches.

The analytic dataset is based on the original records of recorded deaths provided to HRDAG by observer groups, but it does not include those original records. We were provided with a dataframe of features based on comparisons of those record pairs, with each row representing a record pair. Each row included an ID number for the two original records on which it was based and a series of fields based on comparisons of the name, date of death, and governorate of death. However, the dataset did not contain the actual names, dates or locations. The majority of features were yes/no fields, and several were scores or distance measures. In addition, there were a number of features based on the fields of age and sex, which had very high levels of missingness, as those variables were only available in a small minority of records. For this analysis, we used only fields for which there was complete data: namely, those based on name, date of death and governorate of death. The fields based on name, date of death and location of death were complete for all records, as only records that included all three pieces of information were considered for analysis.

The digitized records provided to HRDAG by observer groups were generally provided as Excel spreadsheets, with text fields (name, location, etc) in Arabic. Arabic names were translated to English using Google Translate. This was considered an acceptable translation as the primary concern for translation in this context was consistency. Even if the Google translation were imperfect compared to using a professional translator, the Google version should be consistent. Additionally, most fields were short – generally names – so there was minimal concern about accuracy of grammar, for example. Both the original Arabic and English versions of names were used to generate comparison vectors, as described below.

Governorates were translated via a dictionary, as there were only 14, and the English version was used, as well as an adjacency matrix to generate features.

Additional steps were applied to the name fields. Where first and last names were recorded separately, the were combined into a single name field for consistency with other groups. A version of the names without the word "Mohammed" was created, as Mohammed is an extremely common name in the region, and this allowed for a comparison to focus on other parts of an individual's name. Additionally, a sorted version of the name field was created that ordered the names (first middle, last, etc) alphabetically. While most groups had a defined policy for recording names, it appeared that there were some inconsistencies in, for example, the order of first and last names. The sorted name field was designed to catch instances where the first and last name were identical but reversed in one record, for example. Fields were compared with a number of distance measures.

For the governorates, an adjacany matrix was created manually for use in building adjancy features.

The full list of features is describe in Table 4.1. Many of the features were built using these comparisons algorithms:

- Jaccard similarity is an index that measures similarity using the union and intersection of sets, which can be applied to the set of characters in a word.
- Jaro similarity is a measure of the edit distance between two words that uses the number of matching characters and transpositions.
- Metaphone is an algorithm that provides a phonetic representation of a word. This is useful in cases such as Kathryn and Catherine, where it's more likely that they could refer to the same person than one might guess by simply computing the edit distance between the two words.
- Locality sensitive hashing is a form of hashing designed to bin similar items.

	Description
date_of_death_exact	Exact match on date of death
dtdeath_dtdist	Number of days between the dates of death
governorate_exact	Exact match on governorate
governorate_govadj	Adjacent governorate
name_en_exact	Exact match of English translation of names
name_en_first4_exact	English names, first 4 letters match exactly
name_en_jacc	Jaccard similarity score for English names
name_en_jaro	Jaro similarity score for English names
name_en_lsh_exact	English names, locality sensitive hashing (LSH) matches exactly
name_en_meta_first_exact	English names, metaphone of first names match exactly
name_en_meta_last_exact	English names, metaphone of last names match exactly
name_en_no_mo_lsh_exact	English names, Mohammed dropped, LSH matches exactly
name_exact	Exact match on names
name_first5_exact	First five characters of name match exactly
name_jacc	Jaccard similarity score for names
name_last5_exact	Last five letters of name match exactly
name_meta_first_exact	First name, metaphone matches exactly
name_meta_last_exact	Last name, metaphone matches exactly
sortedname_en_exact	Sorted name in English
sortedname_en_first5_exact	Sorted name in English, first five characters match exactly
sortedname_en_jaro	Sorted name in English, Jaro score
sortedname_en_last5_exact	Sorted name in English, last five letters match exactly
sortedname_en_meta_first_exact	Sorted name in English, metaphone of first names match exactly
sortedname_en_meta_last_exact	Sorted name in English, metaphone of last names match exactly
sortedname_exact	Sorted names match exactly
sortedname_first5_exact	Sorted names, first five letters match exactly
sortedname_lsh_exact	Sorted names, LSH matches exactly
sortedname_meta_first_exact	Sorted names, metaphone of first names matches exactly
year_exact	Years match exactly
yearmo_exact	Year and month match exactly

Table 4.1: Comparison features used to determine if two records refer to the same individual.

4.4 The Analysis

For this analysis, we looked at two different versions of the data. The first was a subset of the data that included only the fields for exact matches on first name, last name, date of death and governorate where death occurred. The second was a subset of data that included those fields and also fields generated from the name, date, and governorate of death described above.

The training set was enriched for positive matches, and also for cases that had been determined in previous analysis to be "hard to classify." The testing data is believed to more closely mimic the full data.

For the dataset that included only exact match variables, we fit two ensembles. The first included logistic regression, random forests, gradient boosting machines and neural network. This was used with the exact match variables. The second ensemble omitted the logistic regression and included the other three algorithms (random forests, gradient boosting machines, and neural networks). These methods are better suited to a highly correlated feature set, and this ensemble was fit on the feature set including derived variables as well as that with only exact matches.

The analysis was conducted using H2OEnsemble, the H2O distributed implementation of the SuperLearner stacked regression algorithm [34]. H2O Ensemble relies on the H2O implementations of the base learners: generalized linear models, random forests, gradient boosting machines, and the deep learning neural network. The metalearner was a binomial GLM with an elastic net penalty with alpha = .05 and lambda= .00001.

To generate match/not match decisions from the predicted probabilities, cutoffs were chosen to maximize the F1 score when the model was fit on training data. F1 score is a measure that takes into account both the precision and recall (also known as sensitivity and positive predictive value) of a classification.

If you consider that a model can give four results, as shown in table 4.2, then:

Table 4.2: Confusion Matrix for a classifier

	Label Positive	Label Negative
True Condition Positive	True Positive	False Negative
True Condition Negative	False Positive	True Negative

The F1 score is the harmonic mean of precision and recall. As such, it's a useful measure in situations with a rare outcome. (Variations of the F score can put higher weight on either precision or recall). $F1 = 2 \times (\text{precision} \times \text{recall})/(\text{precision} + \text{recall})$, where precision = $(\sum \text{True Positive})/(\sum \text{Labeled Positive})$ and $recall = (\sum \text{True Positive})/(\sum \text{True Condition Positive})$.

Metrics for the ensemble learners were calculated on test data that was not seen when fitting the model. Because the outcome of a record pair being a match is rare in the data set, we evaluated models using the area under the ROC curve. The ROC curve is created by plotting the true positive rate vs the false positive rate for a classification as the cutoff

for classification is varied. Like the F1 score, the Area Under the Curve (AUC) values both positive and negative outcomes. Evaluating a model based on accuracy of predictions is problematic in situations with rare outcomes, as simply predicting that no record pairs match would provide a high level of accuracy.

For the random forests and gradient boosting models, we provide variable importance measures from the base models. For all models, we provide performance metrics based on test data not used to fit the model.

4.5 Results

Ensemble with Logistic Regression, Random Forests, Gradient Boosting, and Deep Learning

The base models and the ensemble all performed well, with AUCs over .9, as seen in Table 4.3. In fact, with the training data, the base models had indistinguishable AUCs and the ensemble performed nearly as well. Interestingly, the test metrics are significantly better than the training metrics. Normally one expects to see the reverse. The improved performance on test data is likely due to the fact that the training data was designed to include a high percentage of challenging classifications. That is, it was upsampled for cases that had been misclassified or near the decision boundary in previous classification attempts. In substantive terms, these may be cases where the name and location match but the date is slightly off, or where the names are similar but different enough to make the decision challenging. There are a limited number of such pairs, but they are the most important for training a classifier. Most models can accurately predict that pairs where all three fields disagree refer to two separate people. We can get a sense of how the training data differs from the testing data by examining the predicted values for both the training and test data, which are shown in Figure 4.5. The majority of both the training and testing data are grouped near zero, which is unsurprising, as most pairs refer to distinct individuals. However, in the training data, we can see a bump between .3 and .5. These are cases that were hard for the classifier to predict. About 13 percent of the training cases have predicted probabilities of between .3 and .7.

Table 4.6 shows the coefficients on the base learners, which indicate their importance in the metafit. Both the gradient boosting and random forest classifiers show large contributions to the ensemble prediction. The logistic regression (h20.glm.1) has a large coefficient as well; however it is negative. In the next section, we drop the logistic regression from the ensemble and focus on the less parametric classifiers. We repeat our analysis on the exact match field, and then add the derived comparison feature vectors to see how they affect classification.

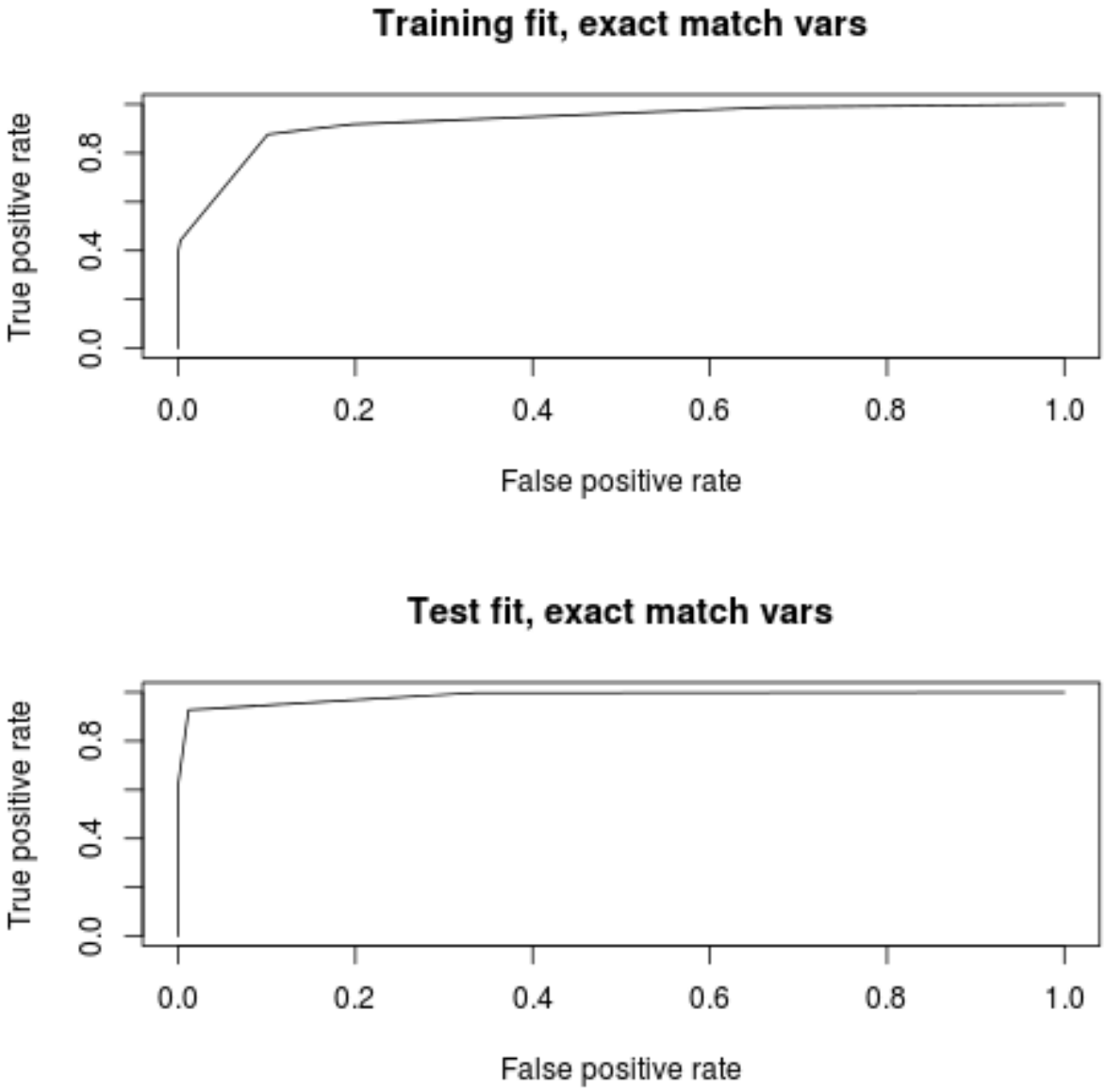


Figure 4.3: ROC Curves for model fit on exact match variables; ensemble of logistic regression, random forests, gradient boosting, and a neural net

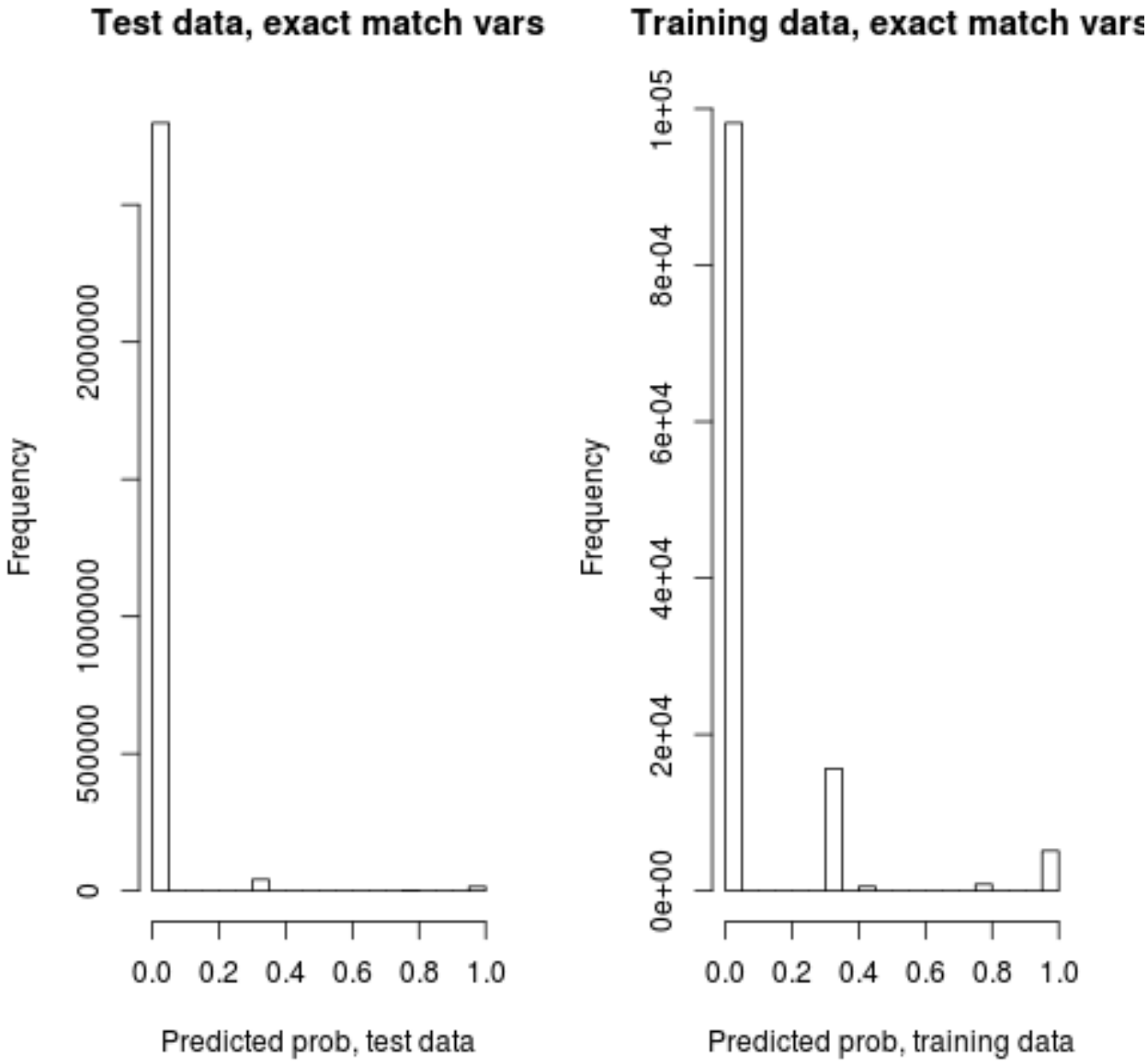


Figure 4.4: Predicted values for model fit on exact match variables; ensemble of logistic regression, random forests, gradient boosting, and a neural net.

	Training AUC	Testing AUC
Ensemble	0.93113057	0.98379735
Logistic Regression	0.93119342	0.98379859
GBM	0.93119342	0.98379859
Random Forest	0.93119342	0.98379859
Deep Learning	0.93119342	0.98379859

Table 4.3: AUC for predictions from ensemble of logistic regression, Random Forests, gradient boosting, and deep learning neural net

	0	1	Error
0	96606	10871	0.10
1	1590	11340	0.12
Totals	98196	22211	0.10

Table 4.4: Confusion Matrix for training data, ensemble of logistic regression, random forests, gradient boosting and deep learning fit on exact match features. Rows are true values and columns are classifications.

	0	1	Error
0	2796116	33027	0.01
1	2089	26701	0.07
Total	2829143	28790	0.01

Table 4.5: Confusion Matrix for testing data, ensemble of logistic regression, random forests, gradient boosting and deep learning fit on exact match features. Rows are true values and columns are classifications.

	coefficients	sign
h2o.glm.1	3.90	NEG
h2o.randomForest.wrapper	3.06	POS
h2o.gbm.wrapper	3.05	POS
h2o.deeplearning.wrapper	0.08	POS

Table 4.6: Standardized Coefficients for metafit, ensemble including logistic regression, gradient boosting, random forests, and deep learning, fit with exact match fields

Ensemble with Random Forests, Gradient Boosting, and Deep Learning

As noted above, we next generated classifications from an ensemble of random forests, gradient boosting and the deep learning neural network. In this section, we included features derived from the name, date, and locations fields, as well as the exact match fields. This gave us a much larger dataset with many highly correlated features, a reasonable process as we are not trying to provide interpretable coefficients for these parameters, but, rather, wish to maximize the predictive capability of the model. In addition, this allows us to explore the importance of different components of the name, date and location fields, as random forests and gradient boosting do provide variable importance measures. For comparison, we fit the same model on the data with only exact match features to ensure that no predictive value was lost in excluding the logistic regression.

Table 4.7 and Figure 4.5 show performance for the ensembles fit on the exact match data and the larger feature set. Again, all algorithms performed well by conventional standards, with AUCs (area under the curve) over .9. The fact that all models produce AUCs above .9 is likely due to the fact that the overwhelming majority of record pairs are easy to classify as matches or (primarily) non-matches. For example, if name, date of death and location of death match exactly, it's probably a match. If all three are quite different, the pair represents two distinct individuals. Records that are harder to classify – where date and location match, and first or last name matches but not both, for example, are the most important. However, they are also relatively rare.

For the model where only the exact match variables were included, all methods performed similarly to each other. In fact, the were indistinguishable to 8 decimal places, with the ensemble performing identically to the base learners. The performance was nearly identically to the ensemble that included the logistic regression. Again, we see a similar pattern where the test metrics are noticeably higher than the training metrics, and a similar pattern in predicted probabilities, with a greater number of hard to classify cases in the training data, as shown in Table 4.8 and Figure 4.5.

Adding features provided improved prediction. This was most noticeable in the training metrics; all AUCs were greater than .99. Test metrics were similar to the training metrics for this model, and slightly higher than the test metrics for the small model.

The difference in testing AUC between the smaller model and the one fit with the full feature set is small, but even very small improvements matter when large numbers of records are being analyzed. This is evident from the confusion matrices for the ensembles fit with only exact match variables or the larger feature sets. The larger model's AUC is only .015 higher than that of the smaller model. However, in the larger model, the number of false positives and false negatives in the test data are roughly the same, with 1402 false positives and 1748 false negatives. In the smaller model, there are 2089 false negatives and 33027 false positives. The dataset that had only 29790 total positives, so this model predicts more than twice as many matches as actually exist.

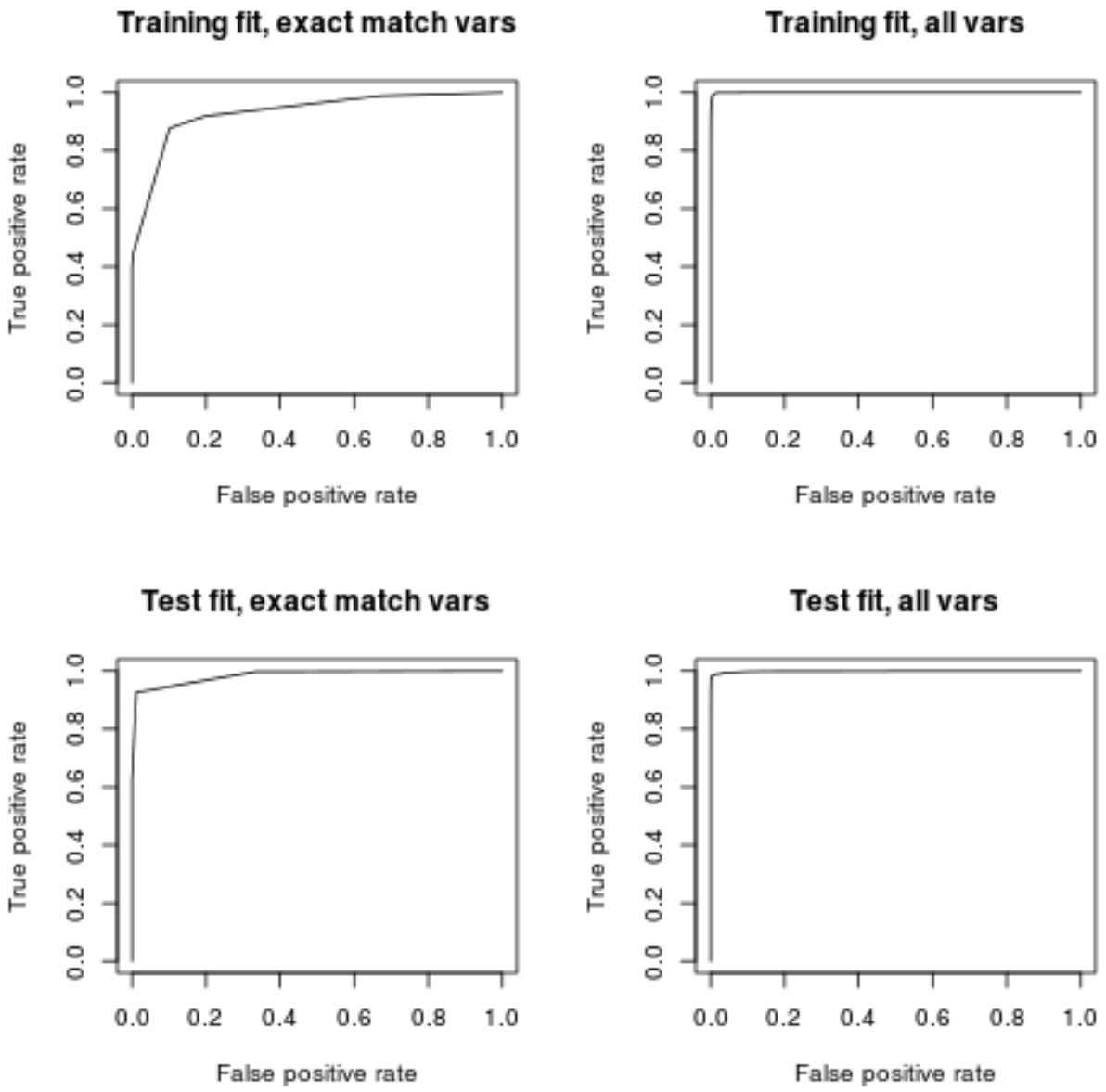


Figure 4.5: ROC curves for models fit on exact match variables and (separately) on all variables; ensemble of random forests, gradient boosting, and neural net

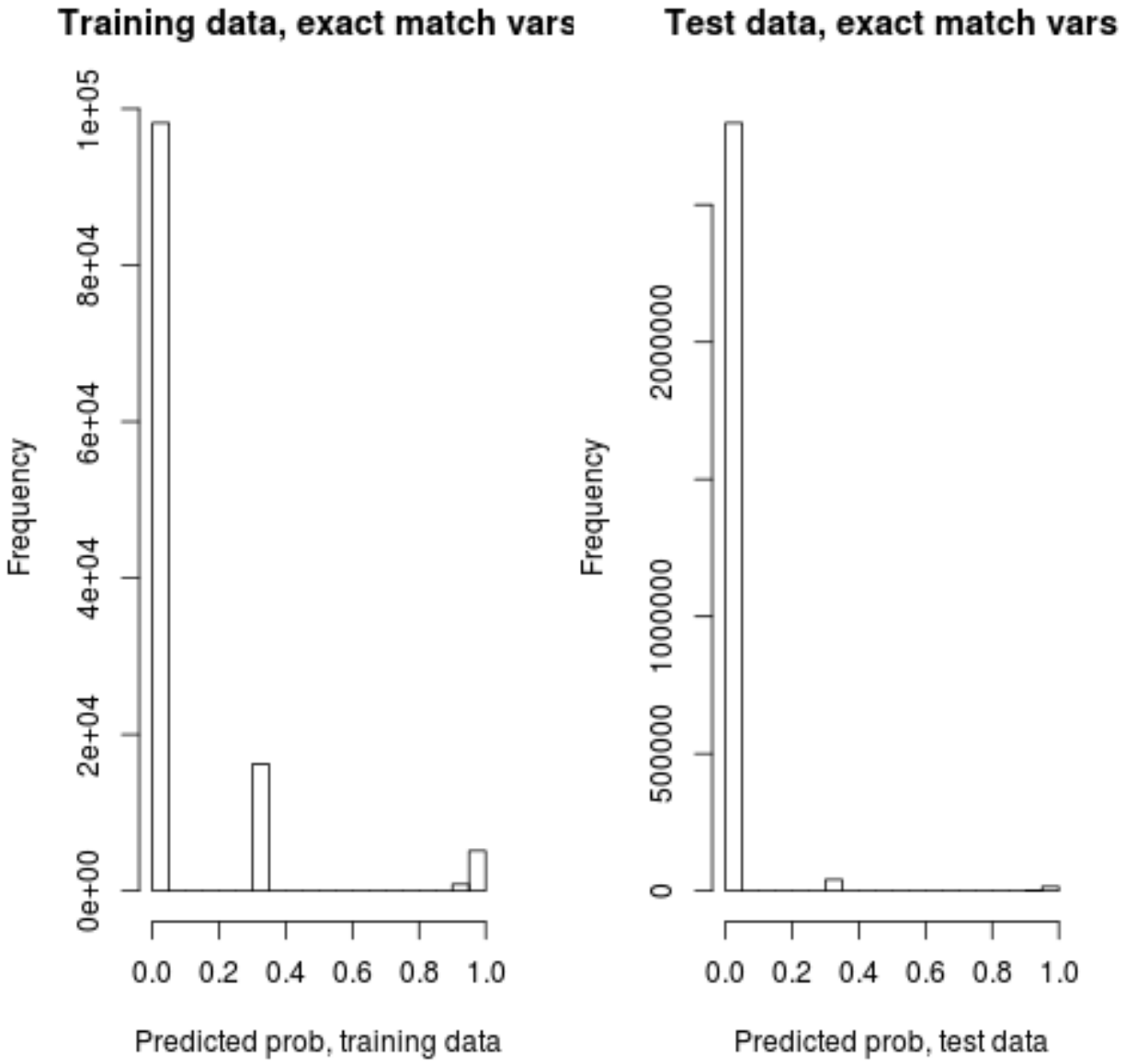


Figure 4.6: Predicted values for model fit on exact match variables; ensemble of random forests, gradient boosting, and neural net

	Small Train	Small Test	All Vars Train	All Vars Test
Ensemble	0.93119342	0.98379859	0.99963310	0.99879557
GBM	0.93119342	0.98379859	0.99718363	0.99913778
Random Forest	0.93119342	0.98379859	0.99980413	0.99596209
Deep Learning	0.93119342	0.98379859	0.99668417	0.99892723

Table 4.7: AUC for predictions from ensemble fit with only exact match fields and from ensemble fit with exact matches and derived features

	Small Model	All Vars
Training Data	13.00	0.55
Testing Data	1.46	0.03

Table 4.8: Percent of Predicted Probabilities that fell between .3 and .7

	0	1	Error
0	96606	10871	0.10
1	1590	11340	0.12
Totals	98196	22211	0.10

Table 4.9: Confusion Matrix for training data, model fit on exact features. Rows are true values and columns are classifications.

	0	1	Error
0	2796116	33027	0.01
1	2089	26701	0.07
Total	2829143	28790	0.01

Table 4.10: Confusion Matrix for ensemble fit only exact match features. Rows are true values and colums are classifications.

	0	1	Error
0	106694	783	0.01
1	1017	11913	0.08
Totals	107711	12696	0.01

Table 4.11: Confusion Matrix for training data, model fit on all features, training data. Rows are true values and columns are classifications.

	0	1	Error
0	2827981	1402	0.00
1	1748	26906	0.06
Total	2829383	28654	0.00

Table 4.12: Confusion Matrix for ensemble fit with all features, testing data. Rows are true values and columns are classifications.

	Small Model	All Vars
h2o.gbm.wrapper	1.09	1.90
h2o.randomForest.wrapper	1.07	0.53
h2o.deeplearning.wrapper	0.07	0.46

Table 4.13: Standardized Coefficients for metafit, ensemble fit with exact match fields and also with exact match and derived features

Variable Importance

Both the gradient boosting and random forests algorithms returns variable importance metrics. In the models that only used exact match comparisons, the name field was by far the most discriminating in determining whether two listings referred to the same person. (This was also true for logistic regression, not shown here) That the name was most important is unsurprising. Generally, we would expect more than one death to occur on many days during the conflict. And many, many deaths will occur in each governorate. In addition, war frequently kills in clusters, leading to instances where multiple distinct individuals are killed at the same place on the same day. Therefore, it's logical that names are crucial to locating individuals in these clusters.

A review of the importance scores for the similarity algorithms is more interesting. Again, names are important. The first four variables listed for gradient boosting, and the first six for random forests are derived from names. In each case, the top three are based on English translations of the name, as well as Jaccard and Jaro scores. Both of these scores address similarity in a generic fashion – a straightforward edit distance and the union and intersection of sets. Metaphone, which is designed to capture phonetic similarity, is also useful, but metaphone variables are noticeably less important for both gradient boosting and random forests. This highlights the importance of considering context when designing variables: Arabic names may have fewer instances of similar sounding names with dramatically different spellings. Each list also had a variable based on the sorted name in the top five, suggesting that transpositions of first and last names may still be an issue in these datasets.

	variable	relative_importance	scaled_importance	percentage
1	name_exact	136790.00	1.00	0.58
2	dtdeath_exact	79018.40	0.58	0.33
3	governorate_exact	20943.03	0.15	0.09

Table 4.14: Variable importance for Random Forests, exact features

	variable	relative_importance	scaled_importance	percentage
1	name_exact	21252.38	1.00	0.70
2	dtdeath_exact	5801.55	0.27	0.19
3	governorate_exact	3122.31	0.15	0.10

Table 4.15: Variable importance for gradient boosting model, fit with exact features

	relative_importance	scaled_importance	percentage
name_en_jacc	81829.5312	1.0000	0.1740
sortedname_en_jaro	81087.4062	0.9909	0.1724
name_en_jaro	79893.6406	0.9763	0.1699
name_jacc	42468.3086	0.5190	0.0903
sortedname_en_exact	29568.5801	0.3613	0.0629
sortedname_exact	27175.6113	0.3321	0.0578
dtdeath_dtdist	14926.9668	0.1824	0.0317
yearmo_exact	14838.0166	0.1813	0.0316
date_of_death_exact	14774.4443	0.1806	0.0314
name_en_exact	14210.7607	0.1737	0.0302
name_exact	13213.0166	0.1615	0.0281
name_en_meta_first_exact	12273.3047	0.1500	0.0261
name_en_first4_exact	6893.1318	0.0842	0.0147
sortedname_en_meta_last_exact	6608.1963	0.0808	0.0141
name_last5_exact	3958.9871	0.0484	0.0084
sortedname_lsh_exact	3880.3669	0.0474	0.0083
name_en_meta_last_exact	3863.9912	0.0472	0.0082
year_exact	3673.8909	0.0449	0.0078
name_first5_exact	2557.5437	0.0313	0.0054
governorate_exact	2219.7615	0.0271	0.0047
sortedname_en_last5_exact	1803.0952	0.0220	0.0038
name_en_no_mo_lsh_exact	1666.4011	0.0204	0.0035
sortedname_first5_exact	1615.3002	0.0197	0.0034
sortedname_en_first5_exact	1473.2816	0.0180	0.0031
governorate_govadj	1411.0375	0.0172	0.0030
name_en_lsh_exact	1200.9813	0.0147	0.0026
sortedname_en_meta_first_exact	1195.5269	0.0146	0.0025
sortedname_meta_first_exact	0.0055	0.0000	0.0000
name_meta_first_exact	0.0000	0.0000	0.0000
name_meta_last_exact	0.0000	0.0000	0.0000

Table 4.16: Variable importance for Random Forests, derived and exact features

	relative_importance	scaled_importance	percentage
name_en_jaro	19921.2363	1.0000	0.4007
name_en_jacc	12259.7666	0.6154	0.2466
sortedname_en_jaro	4058.5117	0.2037	0.0816
name_jacc	3874.5371	0.1945	0.0779
dtdeath_dtdist	3491.5491	0.1753	0.0702
yearmo_exact	3110.4541	0.1561	0.0626
date_of_death_exact	1460.9479	0.0733	0.0294
name_en_meta_first_exact	315.7815	0.0159	0.0064
sortedname_lsh_exact	209.6048	0.0105	0.0042
name_last5_exact	160.9481	0.0081	0.0032
governorate_exact	160.0280	0.0080	0.0032
sortedname_en_meta_last_exact	151.4111	0.0076	0.0030
name_en_meta_last_exact	144.3151	0.0072	0.0029
name_first5_exact	89.9987	0.0045	0.0018
name_en_first4_exact	84.9506	0.0043	0.0017
governorate_govadj	66.5080	0.0033	0.0013
name_en_no_mo_lsh_exact	52.7825	0.0026	0.0011
name_exact	19.4272	0.0010	0.0004
sortedname_en_meta_first_exact	17.2669	0.0009	0.0003
name_en_exact	16.1041	0.0008	0.0003
sortedname_en_last5_exact	12.7171	0.0006	0.0003
year_exact	10.3419	0.0005	0.0002
name_en_lsh_exact	9.6885	0.0005	0.0002
sortedname_first5_exact	7.0691	0.0004	0.0001
sortedname_en_first5_exact	6.9491	0.0003	0.0001
sortedname_en_exact	2.6825	0.0001	0.0001
sortedname_exact	2.4121	0.0001	0.0000
name_meta_first_exact	0.0000	0.0000	0.0000
name_meta_last_exact	0.0000	0.0000	0.0000
sortedname_meta_first_exact	0.0000	0.0000	0.0000

Table 4.17: Variable importance for gradient boosting model, fit with derived and exact features

4.6 Discussion

In this analysis we examined a key component of the record linkage process: the classification of potential record pairs as referring to a single individual (matches) or multiple individuals (non-matches). We compared a number of machine learning classification algorithms to themselves and to SuperLearner ensembles of the base algorithms. We fit the models using a training dataset that was enriched for matches as well as for "hard to classify" pairs – those pairs that had predicted probabilities close to the decision boundary in previous classification attempts. We tested the model on test data believed to mimic the overall composition of the records need to be classified.

In general, all machine learning approaches performed similarly, and the SuperLearner tended to perform equally well to its component algorithms. Further research should include a grid search over the tuning parameters of the base algorithms, though we do not expect to see large gains from this. The most noticeable improvements came from the creation of comparison vectors based on components of the original record fields, such as matches on the first five letters of a name, or an edit distance between names. Further work could include the creation of additional features. For example, the date fields could likely be further exploited through variables that allow for the same day and month but a different year, or for possible transposition of month and day. Additional string comparison approaches may prove useful, as well.

In addition, it will be important to explore the inclusion of features based on other fields. Fields aside from name, date and location tend to have high rates of missingness (see Chapter 3); however, some, such as gender, may be highly discriminating.

However, the results to date are quite promising and support the use of a machine learning based approach to classification of record pairs as matches or non matches as part of a scaleable record linkage strategy. It's important to remember, however, that such a strategy remains contingent on adequate training data for the models. In the case of human rights investigations, that likely means the development of training data for each investigation based on human review of records, as conflicts occur in a range of societies with different languages and name distributions. As we see here, the application of tools developed to match Western names, such as the Metaphone algorithm, should be studied when applied to other contexts.

In addition, it is important to develop relationships with the observer groups who are doing actual data collection. As we saw in Chapter 3, such groups may vary greatly in the quality of their record keeping, leading to noticeable differences in duplication rates, for example.

Bibliography

[1] Amir H Alkhuzai et al. "Violence-related mortality in Iraq from 2002 to 2006." In: *The New England Journal of Medicine* 358.5 (2008), pp. 484–493.

[2] Emmanuelle Amoros et al. "Actual Incidences of Road Casualties, and Their Injury Severity, Modelled from Police and Hospital Data, France". In: *European Journal of Public Health* 18 (2008), pp. 360–365.

[3] Michelle Arevalo-?Carpenter. *Successes and Lessons Learned in Finding and Reuniting Abducted Children of the Salvadoran Civil War - An Evaluation of Objectives and Processes*. A report by Michelle Arevalo-?Carpenter, Evaluator for the UC Berkeley Human Rights Center, Asociación Pro Busqueda, and the DNA Reunification Project Partnership. July 2014.

[4] Jana Asher, David Banks, and Fritz Scheuren, eds. New York: Springer, 2008.

[5] Sophie Baillargeon and Louis-Paul Rivest. "Rcapture: Loglinear Models for Capture-Recapture in R". In: *Journal of Statistical Software* 19.5 (Apr. 3, 2007), pp. 1–31. ISSN: 1548-7660. URL: `http://www.jstatsoft.org/v19/i05`.

[6] Patrick Ball. "Metodologica Intermuestra". In: *Guatemala: Memoria del Silencio, Vol. XII* (1999). URL: `http://shr.aaas.org/guatemala/ceh/ceh.htm`.

[7] Belisaric Betancur, Reinaldo Figueredo Planchar, and Thomas Buergenthal. *From Madness to Hope: Report of the Commission on the Truth for El Salvador (Original: Spanish)*. Report of the Commission on the Truth for El Salvador to the United Nations. Apr. 1993. URL: `http://www.derechos.org/nizkor/salvador/informes/truth.html`.

[8] Christopher M Blanchard, Carla E Humud, and Mary Beth D Nikitin. "Armed conflict in Syria: overview and US response". In: DTIC Document. 2014.

[9] Martin Bouchard. "A capture–recapture model to estimate the size of criminal populations and the risks of detection in a marijuana cultivation industry". In: *Journal of quantitative criminology* 23.3 (2007), pp. 221–241.

[10] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[11] Leo Breiman. "Stacked regressions". In: *Machine learning* 24.1 (1996), pp. 49–64.

[12] Gilbert Burnham et al. "Mortality after the 2003 invasion of Iraq: a cross-sectional cluster sample survey". In: *The Lancet* 368.9545 (2006), pp. 1421–1428.

[13] Anne Chao et al. "Tutorial in Biostatistics: The Applications of Capture-Recapture Models to Epidemiological Data". In: *Statistics in Medicine* 20 (2001), pp. 3123–3157.

[14] Peter Christen. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media, 2012.

[15] Benjamin Coghlan et al. "Mortality in the Democratic Republic of Congo: a nationwide survey". In: *The Lancet* 367.9504 (2006), pp. 44–51.

[16] R. M. Cormack. "Log-Linear Models for Capture-Recapture". English. In: *Biometrics* 45.2 (1989), pages. ISSN: 0006341X. URL: `http://www.jstor.org/stable/2531485`.

[17] Mark Danner. *The massacre at El Mozote: A parable of the Cold War*. Vintage, 1994.

[18] John Darroch et al. "A Three-Sample Multiple-Recapture Approach to Census Population Estimation with Heterogeneous Catchability". In: *Journal of the American Statistical Association* 88.423 (1993), pp. 1137–1148.

[19] Ivan P. Fellegi and Alan B. Sunter. "A Theory for Record Linkage". In: *Journal of the American Statistical Association* 64.328 (1969), pp. 1183–1210. URL: `http://www.jstor.org/stable/2286061`.

[20] S. Fienberg. "The Multiple recapture census for closed populations and incomplete contingency tables". In: *Biometrika* 59 (1972), 591fffdfffdfffd603.

[21] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)". In: *The annals of statistics* 28.2 (2000), pp. 337–407.

[22] Janine di Giovanni. "The Fate of Two Human Rights Lawyers Missing in Syria". In: *Newsweek* February 9 (2015).

[23] Dermot Groome. *The Handbook of Human Rights Investigation: A comprehensive guide to the investigation and documentation of violent human rights abuses*. Human Rights Press, 2001.

[24] Debarati Guha-Sapir and Willem Gijsbert Panhuis. "Conflict-related Mortality: An Analysis of 37 Datasets". In: *Disasters* 28.4 (2004), pp. 418–428.

[25] Debarati Guha-Sapir et al. "Civilian deaths from weapons used in the Syrian conflict". In: (2015).

[26] Amy Hagopian et al. "Mortality in Iraq associated with the 2003–2011 war and occupation: findings from a national cluster sample survey by the university collaborative Iraq Mortality Study". In: *PLoS Med* 10.10 (2013), e1001533.

[27] EB Hook and RR Regal. "Accuracy of alternative approaches to capture-recapture estimates of disease frequency: internal validity analysis of data from five sources." In: *Am J Epidemiol* 152.8 (2000), pp. 771–9.

[28] Ernest B Hook and Ronald R Regal. "Capture-recapture methods in epidemiology: methods and limitations". In: *Epidemiologic reviews* 17.2 (1995), pp. 243–264.

[29] International Working Group for Disease Monitoring and Forecasting. "Capture-Recapture and Multiple-Record Systems Estimation I: History and Theoretical Development". In: *Am. J. Epidemiol.* 142.10 (1995), pp. 1047–1058. URL: `http://aje.oxfordjournals.org/cgi/content/abstract/142/10/1047`.

[30] Nicholas P. Jewell. "MSE and Casualty Counts: Assumptions, Interpretation, and Challenges". In: *Counting Civilian Casualties: An Introduction to Recording and Estimating Nonmilitary Deaths in Conflict*. Ed. by Taylor B Seybolt, Jay D Aronson, and Baruch Fischhoff. Oxford University Press, 2013. Chap. 10, p. 185.

[31] NP Jewell, M Spagat, and B Jewell. "Accounting for Civilian Casualties: From the Past to the Future". In: *In revision, Social Science History* (2016).

[32] Natalie Kruse et al. "Participatory mapping of sex trade and enumeration of sex workers using capture–recapture methodology in Diego-Suarez, Madagascar". In: *Sexually transmitted diseases* 30.8 (2003), pp. 664–670.

[33] J van der Laan Mark et al. "Confidence Intervals for the Population Mean Tailored to Small Sample Sizes, with Applications to Survey Sampling". In: *The International Journal of Biostatistics* 5.1 (2009), pp. 1–46.

[34] Erin E LeDell. "Scalable Ensemble Learning and Computationally Efficient Variance Estimation". PhD thesis. University of California, Berkeley, 2015.

[35] F.C. Lincoln. *Calculating Waterfowl Abundance on the Basis of Banding Returns*. 1930.

[36] Kristian Lum, Megan Emily Price, and David Banks. "Applications of multiple systems estimation in human rights research". In: *The American Statistician* 67.4 (2013), pp. 191–200.

[37] K. Lum et al. "Measuring Elusive Populations with Bayesian Model Averaging for Multiple Systems Estimation: A Case Study on Lethal Violations in Casanare, 1998-2007". In: *Statistics, Politics, and Policy* 1.1 (2010), p. 2.

[38] Thomas Lumley. *Complex surveys: a guide to analysis using R*. Vol. 565. John Wiley & Sons, 2011.

[39] Neil MacFarquhar. "A Very Busy Man Behind the Syrian Civil War's Casualty Count". In: *The New York Times* April 9 (2013).

[40] Elisabeth Malkin. "El Salvador Arrests Ex-Military Officers in 1989 Jesuit Killings". In: *The New York Times* February 6 (2016).

[41] Daniel Manrique-Vallier and Stephen E. Fienberg. "Population Size Estimation Using Individual Level Mixture Models". In: *Biometrical Journal* 50 (2008), pp. 1051–1063.

[42] Robin Mejia. "The Iraq Math War". In: *Mother Jones* (Nov. 2008).

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

[44] Megan Price, Anita Gohdes, and Patrick Ball. *Updated Statistical Analysis of Documentation of Killings in the Syrian Arab Republic*. Report by the Human Rights Data Analysis Group to the United Nations Office of the High Commissioner for Human Rights (OHCHR). Palo Alto, CA, Aug. 2014.

[45] Megan Price, Jeff Klingner, and Patrick Ball. *Preliminary Statistical Analysis of Documentation of Killings in the Syrian Arab Republic*. Report by the Benetech Human Rights Data Analysis Group to the United Nations Office of the High Commissioner for Human Rights (OHCHR). Palo Alto, CA, Jan. 2013.

[46] Megan Price et al. *Full Updated Statistical Analysis of Documentation of Killings in the Syrian Arab Republic*. Report by the Human Rights Data Analysis Group to the United Nations Office of the High Commissioner for Human Rights (OHCHR). Palo Alto, CA, June 2013.

[47] Leila Juzam Pucheu et al. *The Salvadoran Diaspora in North America and the Disappeared Children of El Salvador's Civil War: How Many Don't We Know About Yet?* Paper presented at the Latin American Studies Association International Congress, 2014. Chicago, IL, May 2014.

[48] Louis-Paul Rivest and Sophie Baillargeon. "Applications and extensions of Chao's moment estimator for the size of a closed population". In: *Biometrics* 63 (2007), pp. 999–1006. DOI: 10.1111/j.1541-0420.2007.00779.x. URL: http://www3.interscience.wiley.com/journal/118538514/abstract.

[49] Les Roberts et al. "Mortality before and after the 2003 invasion of Iraq: cluster sample survey". In: *The Lancet* 364.9448 (2004), pp. 1857–1864.

[50] Fritz Scheuren and William E. Winkler. "Regression Analysis of Data Files that are Computer Matched - part II". In: *Survey Methodology* 23 (1997), pp. 157–165.

[51] Jeremy M Sharp and Christopher M Blanchard. "Syria: Unrest and US Policy". In: *Congressional Research Service (CRS) Report for the Congress*. 2012.

[52] Romesh Silva, Jeff Klingner, and Scott Weikart. *State Coordinated Violence in Chad under Hissène Habré, A Statistical Analysis of Reported Prison Mortality in Chad's DDS Prisons and Command Responsibility of Hissène Habré, 1982-1990,* Report by the Benetech Human Rights Data Analysis Group to Human Rights Watch and the Chadian Association of Victims of Political Repression and Crimes. Palo Alto, CA, Feb. 2010.

[53] Philip J Smith. "Bayesian analyses for a multiple capture-recapture model". In: *Biometrika* 78.2 (1991), pp. 399–407.

[54] Philip J Smith. "Bayesian methods for multiple capture-recapture surveys". In: *Biometrics* (1988), pp. 1177–1189.

[55] Christine Tapp et al. "Iraq War mortality estimates: a systematic review". In: *Conflict and health* 2.1 (2008), p. 1.

[56] Ginger Thompson. "The Rev. Jon de Cortina, 71, Is Dead; Saved Salvadoran Children". In: *The New York Times* January 9 (2006).

[57] Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. "Super learner". In: *Statistical applications in genetics and molecular biology* 6.1 (2007).

[58] Mark J Van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

[59] Patrick Vinck and Phoung Pham. "The Fastest to Die". In: *Foreign Policy* August (2010).

[60] William E. Winkler. *The State of Record Linkage and Current Research Problems*. Tech. rep. RR1999/04. Statistical Research Division, U.S. Census Bureau, 1999. URL: `http://www.census.gov/srd/papers/pdf/rr99-04.pdf`.

[61] Jeremy C. York and David Madigan. *Bayesian Methods for Estimating the Size of a Closed Population*. Tech. rep. 234. University of Washington, July 1992. URL: `http://www.stat.washington.edu/research/reports/1992/tr234.pdf`.