

Killings and Refugee Flow in Kosovo March - June 1999

A Report to the International Criminal
Tribunal for the Former Yugoslavia

3 January 2002

Patrick Ball, Wendy Betts, Fritz Scheuren,
Jana Dudukovich, and Jana Asher



AMERICAN ASSOCIATION FOR THE
ADVANCEMENT OF SCIENCE



American Bar Association
Central and East European Law Initiative

The materials contained herein represent the opinions of the authors and editors and should not be construed to be the view of the American Association for the Advancement of Science (AAAS) Science and Human Rights Program or the American Bar Association Central and East European Law Initiative (ABA/CEELI), or any of the contributing organizations. The AAAS Committee on Scientific Freedom and Responsibility (CSFR), in accordance with its mandate and Association policy, supports publication of this report as a scientific contribution to human rights. The interpretations and conclusions are those of the authors and do not purport to represent the views of the AAAS Board, Council, the CSFR, or the members of the Association. Likewise, the views expressed herein have not been approved by the House of Delegates or the Board of Governors of the ABA and, accordingly should not be construed as representing the policy of the ABA.

Nothing contained in this publication is to be considered as the rendering of legal advice for specific cases, and readers are responsible for obtaining such advice from their own legal counsel. This publication and any forms and agreements herein are intended for educational and informational purposes only.

Copyright 2002 by the
American Association for the Advancement of Science
1200 New York Avenue, NW, Washington, DC 20005

Contact Information:

AAAS Science and Human Rights Program
1200 New York Avenue, NW
Washington, DC 20005

Tel: 202 326 6600
Fax: 202 289 4950
Email: shrp@aaas.org
URL: <http://shr.aaas.org>

Contents

Executive Summary	1
Killings and Refugee Flow in Kosovo, March–June 1999: Analysis and Conclusions	2
1 Purpose of report	2
1.1 Hypotheses	2
1.2 Data and analysis	3
1.3 Principal findings	3
2 Identifying a pattern	4
3 Statistical analysis of refugee flow	4
4 Statistical analysis of killings	5
4.1 Estimated total number of killings	5
4.2 Killing patterns over time	7
4.3 Refugee flow and killings by geographic location	8
5 Examination of proposed hypotheses	8
5.1 Kosovo Liberation Army activity	11
5.2 NATO airstrikes	12
5.3 Effect of KLA activity and NATO airstrikes	13
5.4 Yugoslav forces	15
6 Summary of conclusions	15
Appendix 1: Data and Matching	17
1 Introduction	17
2 Data sources	17
2.1 American Bar Association Central and East European Law Initiative (ABA/CEELI)	18
2.2 Exhumations (EXH)	19
2.3 Human Rights Watch (HRW)	20
2.4 Organization for Security and Cooperation in Europe (OSCE)	21
3 Initial data editing	21
3.1 Geographic coding	22
3.2 Name and gender editing	22
3.3 Date of death formatting	23
4 Initial data matching	23
4.1 Variables used for intra-system matching of individual records	24
4.2 Basic approach of intra-system matching of individual records	25

4.3	Inter-system matching of individual records	27
4.4	Intra- and inter-system handling of anonymous group records	28
4.5	Merging anonymous group killings across systems	29
5	Refinements in data editing and matching	29
5.1	Inconsistent matches	29
5.2	Choosing the “best” dates	30
5.3	Exhumation data	31
5.4	Other edits of the final matches	31
6	Final summary of data results	32
6.1	Data handling by source for individual records	32
6.2	Data handling across sources for named, individual records	32
6.3	Evaluating the completeness of the individual data	34
Appendix 2: Statistical Methodology and Analysis		35
1	Introduction	35
1.1	Limitations of direct observations	36
2	Methodological background	37
2.1	Dual systems estimation	37
2.2	Triple systems estimation	38
2.3	Multiple systems estimation	41
2.4	Model selection	41
3	Methodology	42
3.1	Exploratory data analysis	42
3.2	Fitting and selection of a model for the total number of killings	45
3.3	Aggregation of the cross-classification tables to account for sparseness	45
3.4	Global model fitting across all temporal and spatial points	49
3.5	Piecewise modeling across temporal and spatial points	49
3.6	Projection of 2-day time point series to 6-day time point series for each region	50
3.7	Comparison of results of global and piecewise modeling	51
3.8	Analysis of relationship between original lists, complexity of models selected by the selection rule, and time and space	52
4	Analysis of relationship between multiple systems estimation modeling results and KLA/NATO activity	56
5	Discussion	59
5.1	Sensitivity analysis of date of death reporting	59
5.2	Summary of modeling conclusions	62

Appendix 3: Additional Sources on KLA and NATO Activity	63
References	65
Acknowledgments	68
Authors and Editors	70
Scholarly Review Panel	71
Authoring Organizations	72
About the Authors	73

Executive Summary

This study presents the results of analyses of the statistical patterns of refugee flow and killings in Kosovo during the period March–June 1999. The data were drawn from the Albanian border guard registries of people entering Albania through the village of Morina; interviews conducted by the American Bar Association Central and East European Law Initiative (ABA/CEELI) and its partners; interviews conducted by Human Rights Watch (HRW); interviews conducted by the Organization for Security and Cooperation in Europe (OSCE); and records of exhumations conducted by international teams on behalf of the International Criminal Tribunal for the Former Yugoslavia (ICTY). These analyses describe the estimated total number of killings and estimated number of refugees leaving their homes over time and location.

This report finds that killings and refugee flow occurred in a regular pattern characterized by three phases. In each phase, a high volume of killing and refugee flow was followed by a much lower level of killing and refugee flow. These findings are then used to evaluate three possible explanations for the pattern.

- Action by the Kosovo Liberation Army (KLA) motivated Kosovars to leave their homes.
- Air attacks by the North Atlantic Treaty Organization (NATO) created local conditions that led to Kosovars being killed and leaving their homes.
- A systematic campaign by Yugoslav forces expelled Kosovar Albanians from their homes.

This study concludes the following:

- The patterns of both refugee flow and killings exhibit characteristics consistent with the existence of an external cause.
- Refugee flow and killings occurred in the same places at the same times, implying a common cause of both phenomena.
- Two of the hypotheses proposed to explain the patterns in killing and migration, KLA and NATO activity, are inconsistent with the observed patterns of refugee flow and killings.
- The statistical evidence is consistent with the hypothesis that Yugoslav forces conducted a systematic campaign of killings and expulsions.

Killings and Refugee Flow in Kosovo, March–June 1999: Analysis and Conclusions

1. Purpose of report

This study presents the results of analyses of the statistical patterns of refugee flow and killings in Kosovo during the period March–June 1999. This data analysis describes the estimated total number of deaths and estimated number of refugees leaving their homes over time and location. The objective of the analysis is to compare three hypotheses about what may have caused killings and refugee flow in order to conclude which hypotheses are contradicted, and which supported, by the analysis.

1.1. Hypotheses

The study first examines whether there was a regular structure in killings and refugee flow. Thus our first hypothesis is

- *Hypothesis 1:* Killings and refugee flow occurred in distinct patterns indicating the existence of a common cause of both phenomena.

If the data analysis supports Hypothesis 1, there are three possible explanations for the pattern.

- *Hypothesis 2.1:* Action by the Kosovo Liberation Army (KLA) motivated Kosovars to leave their homes, either directly because the KLA ordered people to leave, or indirectly because Kosovars fled fighting between KLA and Yugoslav forces.
- *Hypothesis 2.2:* Air attacks by the North Atlantic Treaty Organization (NATO) created local conditions that led to Kosovars being killed and leaving their homes. The NATO influence could either have been direct, because people were killed in airstrikes and others fled, or indirect, because *local* Yugoslav authorities responded to the airstrikes by killing Kosovars and forcing them from their homes.
- *Hypothesis 2.3:* A systematic campaign by Yugoslav forces drove Kosovar Albanians from their homes. Killings were used either to motivate the departures, or the killings were a result of the campaign.

Although there may be other explanations for regular patterns in killings and refugee movement, we consider these three to be the most likely. The hypotheses are distinct. Although they are not necessarily mutually exclusive, each of the hypotheses in 2.1–2.3 implies differing responsibility. It is beyond the capacity of statistical analysis to *prove* that any of these hypotheses is the definitive cause of the patterns seen in the two forms of violence. However, as will be seen, the data can be found to contradict some hypotheses while being consistent with other hypotheses.

1.2. Data and analysis

The data for this project came from several sources.

- *Refugee flow*: The analysis of refugee flow uses the Albanian border guard registries of people entering Albania through the village of Morina. Additional sources were used to transform the statistical patterns of people entering Albania into an analysis of people leaving their homes and becoming refugees.¹
- *Killings*: The data on killings were drawn from four sources: interviews conducted by the American Bar Association Central and East European Law Initiative (ABA/CEELI) and its partners; interviews conducted by Human Rights Watch (HRW); interviews conducted by the Organization for Security and Cooperation in Europe (OSCE); and records of exhumations conducted by international teams on behalf of the International Criminal Tribunal for the Former Yugoslavia (ICTY).

The statistical analysis of killings aggregates information from more than 15 000 interviews and exhumation reports.² The analysis includes a statistical estimate of the killings that were not reported to any of the four sources.³

1.3. Principal findings

This report finds:

- Killings and refugee flow occurred in a regular pattern characterized by three phases. In each phase, a high volume of killing and refugee flow was followed by a much lower level of killing and refugee flow. Killing and refugee flow tend to occur at the same times and places. We conclude that this pattern is consistent with Hypothesis 1;
- An estimated 10 356 Kosovar Albanians were killed;⁴

¹The refugee flow data are based primarily on the records maintained by Albanian government border guards. Additional administrative records from the United Nations High Commission for Refugees and the Albanian government, and survey data from several human rights organizations augmented the analysis of the border records. Note that this analysis does not include data from internally displaced persons who never crossed the border. Thus, the estimates do not represent overall totals of people leaving their homes. See Ball (2000).

²The direct results are presented in Appendix 2.

³In an effort to assure quality, all the data coding involving comparisons between data sources was done independently by two different people; their results were compared, and all differences were reviewed and reconciled by an author of the study.

⁴All of the statistical programming connected to the estimation of the results was done independently by two analysts using separate computers and different software, and their results were identical.

- Observed and estimated patterns are inconsistent with Hypotheses 2.1 and 2.2, KLA activity or NATO airstrikes. Patterns are consistent with Hypothesis 2.3, activities of Yugoslav forces.

Each of these findings is explained in the sections that follow.

2. Identifying a pattern

The structure of the patterns in both refugee flow and killings over the time period in question is the key component for the findings of the present study. In this context, a pattern means a series of distinctive, clearly non-random movements, trending upward and downward, in the volume of refugee flow and the number of people killed. Two or more patterns are considered to be similar if they exhibit similar high points and low points at the same (or nearly the same) times.

Statistically, it is implausible that patterns such as those indicated by the findings would result simply from ad hoc decision-making or random external causes. The correlated, nearly simultaneous variations in the social phenomena being measured (killings and refugee flow) in time and location strongly suggest a common, systematic cause of which the patterns are results.

The identification of a pattern does not by itself support or contradict Hypotheses 2.1, 2.2, or 2.3. It does, however, weigh against the claim that the killings and refugee flow were random. That is, the existence of a pattern strongly suggests that there was a common cause, and that the killings and refugee flow did not occur independently.

3. Statistical analysis of refugee flow

This section describes the departure of ethnic Albanians from Kosovo from late March to May 1999.⁵ Approximately 95% of the Kosovar Albanian refugees who entered Albania did so between 24 March and 11 May (Ball 2000, p.5). Analysis of the flow of these refugees during this period shows a pattern of surges followed by steep descents.

An earlier analysis of refugees' departures from their homes showed that from late March through late May 1999, ethnic Albanians left their homes in Kosovo in three distinct time periods, or phases (see Figure 1). These phases were: 24 March to 6 April; 7 April to 23 April; and 24 April to 11 May.⁶

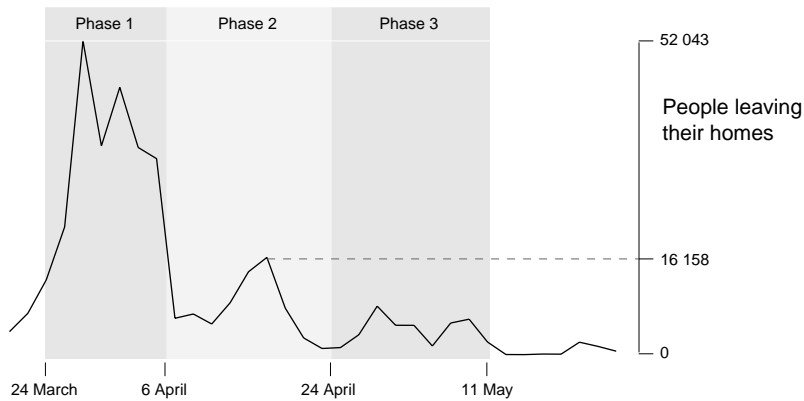
The essential characteristic of this phase structure is the presence of low points in the number of refugees leaving their homes on 6-8 April and 23-25 April, the phase transition dates.⁷ These low points are significant because they do not last for extended periods of time and are preceded and followed

⁵Although the analysis of killings covers the period 20 March - 22 June, the analysis of refugee movement ends in late May, for two reasons. First, the registries maintained by the Albanian border guards ended at that time. Second, anecdotal reports indicated that there was very little movement over the border after that time; this was later confirmed by surveys taken among residents in refugee camps in mid to late June.

⁶See Ball (2000). The three phases reflect the patterns of refugees departing their homes, not the patterns of refugees crossing the border. On any given day, slightly more than half of the refugees who crossed the border had left their homes earlier that same day. However, the remaining refugees crossing the border that day had been in transit for longer times. The transit delay was accounted for in the analysis of the data.

⁷The March - June period was aggregated into two-day intervals for this analysis. Aggregating the time to this level provided enough data at each time for the statistical analysis to stabilize; see

Figure 1: Estimated total refugee flow over time



by distinct peaks. In other words, during these two transition intervals, the incidence of people leaving their homes nearly ceases, compared to the high numbers observed during the phases.

As Figure 1 shows, during the 6-8 April phase transition, refugee flow falls to approximately 6 000 people, down from the phase one peak of slightly more than 52 000. During the 23-25 April phase transition, refugee flow falls to approximately 1 000 from the phase two peak of more than 16 000. The third phase sees refugee flow rising to two peaks of approximately 8 000 and 6 000 in early May, representing the last surges. Refugee flow declines to fewer than 100 people per two-day period after 11 May.

The extreme fluctuation between high and low points constitutes the pattern in the refugee flow. Migration that resulted from dispersed, decentralized causes would not show distinct separations between moments of high flow and moments of low flow. If the incidence of people leaving their homes occurred at random, there would be a more uniform distribution of numbers over time, with occasional small peaks. The extreme, well-defined surges observed in this analysis would not occur by chance. The mass exodus of Kosovar Albanians on this scale and in this pattern could only have been driven by a common cause.

4. Statistical analysis of killings

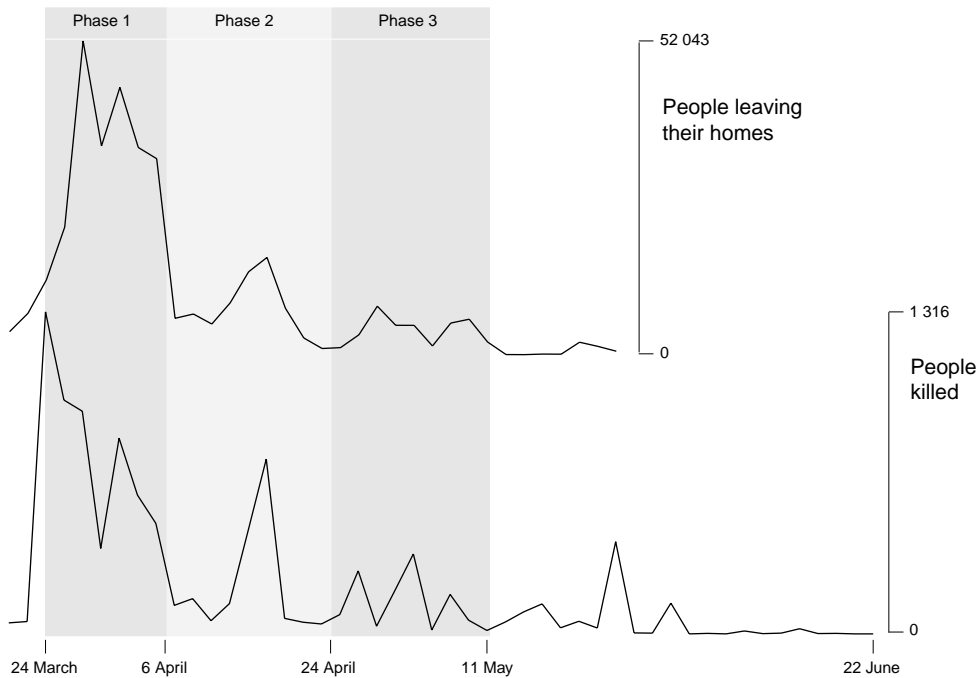
This section describes the number and pattern of killings that occurred in Kosovo from late March to mid-June 1999. Analysis of the data on killings finds that an estimated 10 356 Kosovo Albanian civilians were killed, and that the patterns of killing are similar to the pattern of refugee flow. As with refugee flow, we conclude that the statistical patterns of killings indicate that they resulted from a common cause.

4.1. Estimated total number of killings

Before studying when and where killings took place, it is necessary to first estimate the number of total killings that occurred during this time period. To make

Appendix 2. The value of the estimated number of killings or refugees plotted for a given time on the horizontal axis of the graph represents the number for the related two-day period.

Figure 2: Estimated total refugee migration and killings over time



this estimate, a series of steps was taken. First, the total number of individual victims, documented by name, was tabulated. All victims identified by name in one or more of the data sources were listed; descriptive information on the victims was compared in order to eliminate duplicates; and the total number of unique individuals was tallied.⁸ From approximately 10 000 victims reported by name, this process identified 4 400 unique individuals. The number 4 400 is not an estimate; it is the actual count of uniquely reported victims.

Second, because the victims identified in the data sources were not the only victims of killings, it was necessary to estimate the number of undocumented victims to determine the overall estimate of total number killed. This figure, 10 356, was generated by means of a widely-used demographic statistical technique known as “multiple systems estimation,” which depends on samples of the population.⁹ Because the overall estimate of 10 356 killed was generated from samples — and not from the (unknowable) perfect list of deaths — a margin of error must be computed. We estimate this interval to be 9 002 to 12 122. Note that the estimate and margin of error are consistent with estimates of killing victims in Kosovo in previous work by ABA and AAAS, as well as those in other, independent studies.¹⁰

⁸See Appendix 1 for a complete description of this process.

⁹See Appendix 2 for a complete description of this procedure.

¹⁰See ABA/AAAS (2000), PHR (1999), Spiegel and Salama (2000).

Figure 3: Regions of Kosovo



4.2. Killing patterns over time

When the estimated number of people killed is considered over time, using the same two-day intervals employed with the refugee flow data, the observed pattern of killings closely resembles the pattern of refugee flow. The analysis is shown in Figure 2.

The data show a peak in the number of killings in late March, and another peak in mid-April. Most noteworthy is that, similar to the refugee flow data, the incidence of killings fell to nearly zero on 6-7 April and again on 22-24 April. Thus, not only does the number of killings exhibit the same extreme contrasts between the high and low points as observed in refugee flow, these high and low points occur at the nearly the same times as those in refugee flow. These surges would not occur by chance, and we conclude that they are the result of a common cause.

4.3. Refugee flow and killings by geographic location

In addition to examining when refugee flow and killings happened, it is important to study where the events occurred. An analysis of the locations where the refugee flow originated, and the killings occurred, shows widespread patterns consistent with acts of violence associated with displacements.

When the number of people leaving their homes and the number of people killed are analyzed on a regional level, one can identify the extreme contrasts in high and low points following a phase pattern similar to that described above for the overall analysis (see Figures 4–7). Their relative patterns over time and space are similar. In all regions, the 6-7 and 22-24 April dates mark low points in both the flow of refugees and the number of people killed.

An earlier analysis of refugee flow observed that more than three-quarters of the refugees crossing into Albania during Phase 1 originated in the southern and western areas of Kosovo, while during Phase 2, more than three-quarters of refugees originated in the northern and eastern areas of Kosovo (Ball 2000). Figures 4–7 show that killings follow a similar pattern. Killings in the western and southern regions occur primarily during Phase 1; during later phases, there are relatively fewer killings in these two regions. In the northern and eastern regions, killings also occur during Phase 1. However, in these regions and unlike in the southern and western regions, there are also substantial numbers of people killed during Phase 2.¹¹

As these graphs show, not only do the patterns of refugee flow and killings share similar characteristics over time, the patterns are similar in different regions. Although when viewed in isolation local refugee movement and killings may look like a local response to a local cause, seen in the aggregate, statistical analysis reveals a pattern implying a common cause. In other words, the killings and the exodus of refugees occurred in the same places at roughly the same times. The analysis shows that these events occurred in similar patterns in each of the four regions. The analysis does not prove what caused either pattern, nor that one of the patterns caused the other. The analysis does show that acts of violence — killings — were associated in time and space with the refugee departures from their homes.

5. Examination of proposed hypotheses

As noted above, statistics do not prove that any particular process caused either refugee flow or mass killing patterns. However, analysis can show whether hypotheses are consistent with or contradicted by the statistical evidence. There have been three hypotheses about the causes of the patterns in refugee flow and killings. These three hypotheses are KLA activity, NATO airstrikes, or a systematic campaign conducted by Yugoslav forces.

It is possible to use statistical methods to examine the relationship between KLA activity or NATO airstrikes and the patterns described above. If KLA activity or airstrikes occur immediately before or during periods of high levels of killing and migration, these events may plausibly be the cause of the rise and fall

¹¹There is an anomalous point in the southern region (Figure 5) in late May. This estimate of more than 200 people killed in one two-day period results from fewer than 20 documented killings. Appropriately, this point also has a relatively high level of error associated with it, as shown in Appendix 2, Figure 12. As is clear in that figure, most points have modest errors which do not weaken the interpretation of the pattern. This point, however, has a sufficiently wide margin of error that the point may not be significantly different from zero.

Figure 4: Estimated total refugee migration and killings over time, northern region

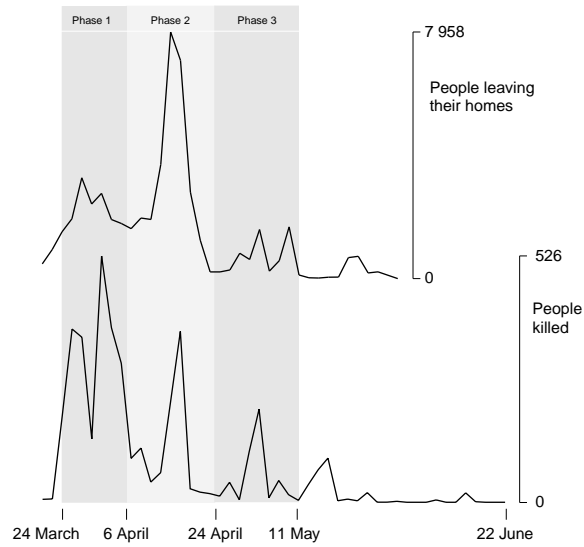


Figure 5: Estimated total refugee migration and killings over time, southern region

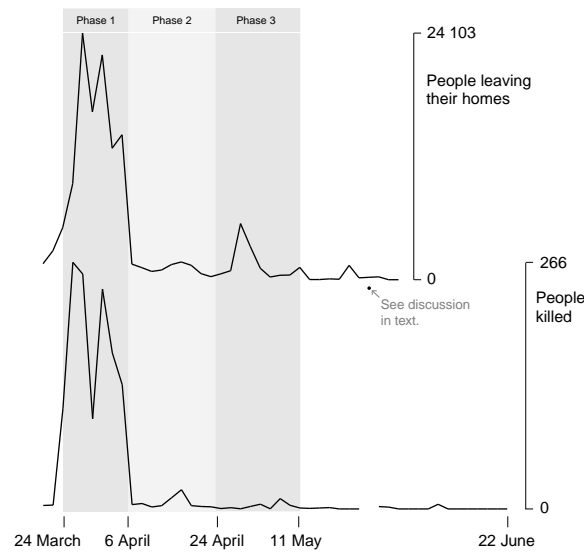


Figure 6: Estimated total refugee migration and killings over time, eastern region

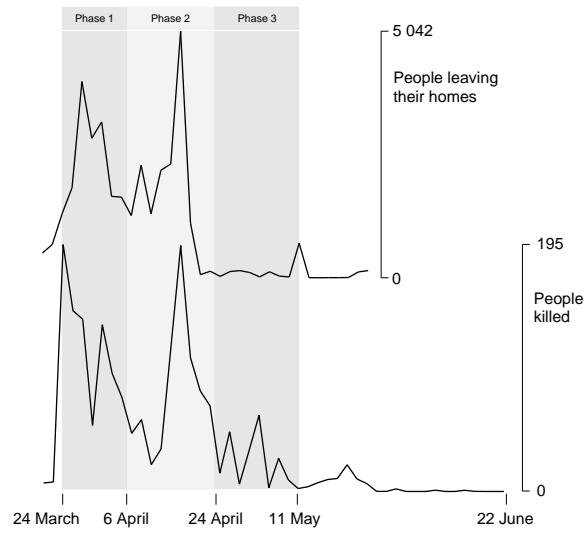


Figure 7: Estimated total refugee migration and killings over time, western region

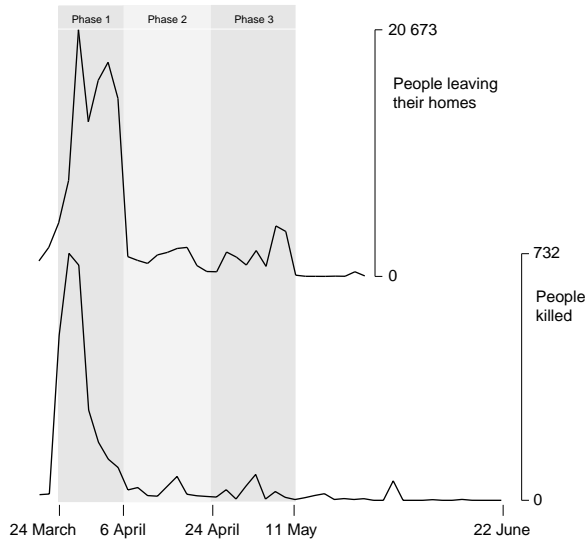


Figure 8: Timing of KLA attacks with killings and refugee flow

Timing	Killings	Percent	Refugee Flow	Percent
Preceded or coincided with peak	11	38%	10	34%
Followed peak	12	41%	11	38%
Inconclusive	6	21%	8	28%

pattern. However, if airstrikes and KLA activity do not precede the peaks in the number of killings and refugee flow, then the causal relationship should be questioned or rejected. An analysis of KLA activity and NATO airstrikes over time and place shows that neither occurred at the times and places necessary to be the primary cause of the refugee flow and killings.

To analyze the occurrence of KLA or NATO activity in relation to the pattern of killings and refugee flow, we used the following procedure. For each municipality in Kosovo, we listed chronologically, by two-day period, the numbers of refugees departing their homes, the number of reported killings, and the incidence of KLA and NATO activity.¹² For this analysis, KLA activity included both battles and isolated killings of Serbs. The two-day periods marking the peak for refugee flow and killings, respectively, were identified. If an incidence of KLA or NATO activity fell within the same period or in the two-day period preceding the peak, we concluded that the two events coincided. If there was no record of KLA or NATO activity at any point prior to the peak, we concluded that KLA or NATO activity occurred only after the peak. If an incidence of KLA or NATO activity occurred earlier than two days prior to the peak period, the municipality was counted as having an inconclusive pattern.

To test the conclusions drawn by this method, we used another statistical method to consider the joint correlations of KLA and NATO activity with refugee flow and killing patterns. The point is to use the second statistical technique to control for the correlation of KLA activity and NATO airstrikes with the quantity of killings and refugee flow, over time and space.

5.1. Kosovo Liberation Army activity

Information on KLA activity was obtained from interview accounts and a variety of non-governmental reports summarized and provided to this project by the ICTY.¹³ Using that information, the present study counted the number of reported battles between the KLA and Yugoslav forces occurring in each municipality over time. No effort was made to quantify the intensity of individual battles, but distinct engagements were counted separately. Isolated KLA attacks that resulted in the injury, disappearance, or deaths of ethnic Serbs were also tabulated by the number of casualties. These are counts of reported Serb casualties, not estimates. The data were insufficient for estimating the missing totals.

¹²Note that for this analysis, we used only the number of reported killings, not the estimated total number. The data are inadequate to make estimates at the municipality level. See Section 5.3 for an analysis using the estimated number of killings at the regional level.

¹³A summary of sources is provided in Appendix 3.

Figure 9: Timing of NATO airstrikes with killings and refugee flow

Timing	Killings	Percent	Refugee Flow	Percent
Preceded or coincided with peak	3	10%	9	31%
Followed peak	20	69%	13	45%
Inconclusive	6	21%	7	24%

As testing the hypotheses necessitates, reported KLA activity was plotted against killings and refugee flow for each of the 29 municipalities in Kosovo. The results of the analysis of timing are in Figure 8, which shows that in 11 of the 29 municipalities, 38%, KLA activity coincided with the overall peak in the number of killings, or it occurred within the two-day interval prior to the peak. In 12 of the municipalities, 41%, KLA activity either occurred only after the peak in number of killings or did not occur at all. In 6 municipalities, 21%, there is an inconclusive pattern.

Refugee flow has a similar pattern. In 10 municipalities, 34%, KLA activity coincided with the overall peak in number of refugee flow or occurred within the two-day interval prior to the peak. In 11 municipalities, 38%, KLA activity either occurred only after the peak in refugee flow or did not occur at all. In the remaining 8 municipalities, 28%, KLA activity occurred at points in time coinciding with other high points, low points, or interim points in the numbers of killings and refugee flow.

For KLA activity to have caused the pattern observed in killings and refugee flow, the instances of activity would have to precede the high points. However, this analysis shows that KLA activity followed the peaks in the killing and refugee numbers in more places than it preceded them. Thus, there is no clear cause and effect relationship between KLA activity and the pattern described here.

5.2. NATO airstrikes

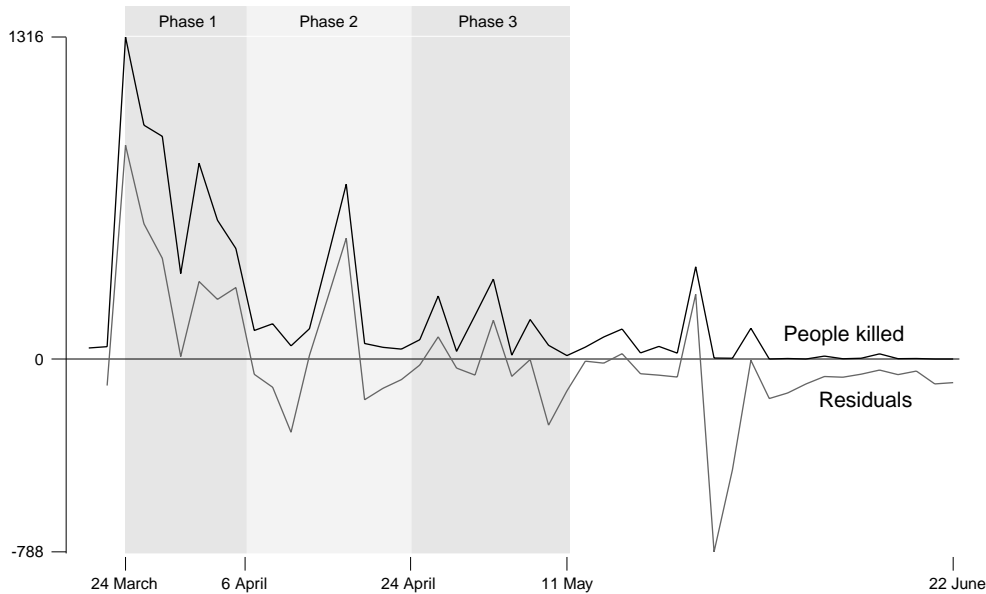
This analysis considers the number of NATO airstrikes, as reported by Yugoslav government sources.¹⁴ No effort was made to quantify the severity of each airstrike, but reports of different airstrikes were counted separately. Similar to KLA activity, reported airstrikes were plotted against killings and refugee flow for each of the 29 municipalities in Kosovo.

In only 3 of the 29 municipalities, 10%, did NATO airstrikes coincide with the overall peak in the number of killings, or occur within the two-day interval prior to the peak. In 20 municipalities, 69%, NATO airstrikes either occurred only after the peak in the number of killings or did not occur at all, and in 6 municipalities, 21%, the pattern was inconclusive.

The refugee flow pattern is not as lopsided, but it leads to the same conclusions. In 9 municipalities, 31%, NATO airstrikes coincided with the overall

¹⁴The Yugoslav government was the primary proponent of the claim that NATO airstrikes were responsible for the killings and refugee flow in Kosovo. Therefore, the strongest test of this hypothesis is to use the Yugoslav government's own information concerning when and where airstrikes occurred.

Figure 10: Estimated total killings and residuals over time



peak in number of refugee flow or occurred within the two-day interval prior to the peak. In 13 municipalities, 45%, NATO airstrikes either occurred only after the peak in the refugee flow or did not occur at all. In the remaining 7 municipalities, 24%, NATO airstrikes occurred at other times, coinciding with other high points, low points, or interim points in the killings and refugee flow.

One other noteworthy fact regarding NATO airstrikes was that during 2-4 April, attacks were greatly reduced due to bad weather.¹⁵ Yet this period, during which there were relatively few NATO airstrikes, includes substantial peaks in Kosovo-wide killings and refugee flow. As with the findings regarding data on the KLA, the analysis of data on NATO shows that the airstrikes more often followed the peaks in the killings and refugee numbers than preceded them. Therefore, the hypothesis that NATO airstrikes directly or indirectly caused the patterns in killing and refugee flow should be rejected.

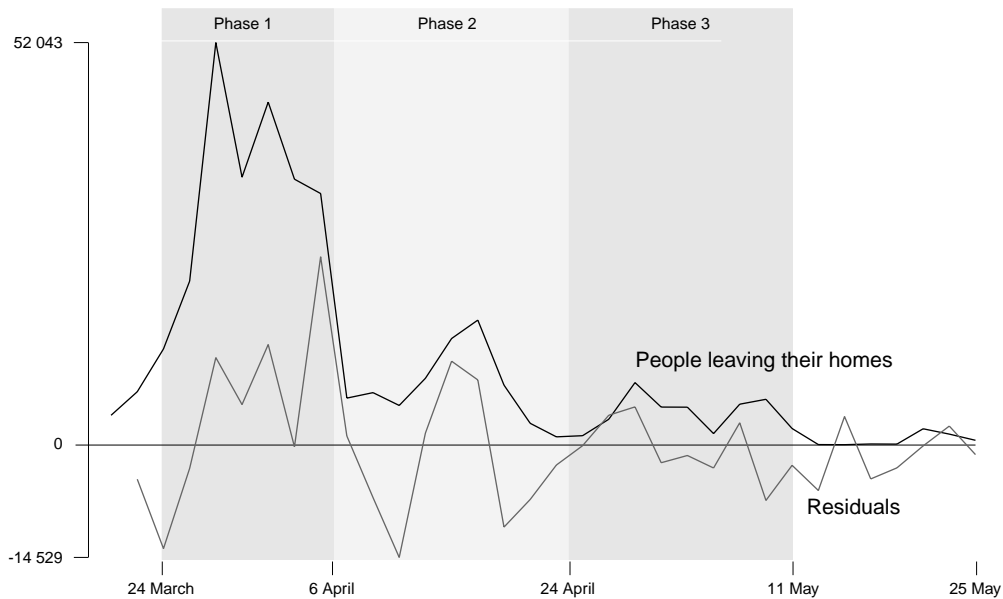
5.3. Effect of KLA activity and NATO airstrikes

In addition to the preceding analysis, the data were also aggregated to regional levels, and patterns over time in each of the four regions were analyzed jointly with the patterns of killings and refugee flow. The objective was to examine the pattern of killing net of the statistical correlation with KLA activity and NATO airstrikes.

In other words, this analysis looks at the joint effect of KLA and NATO activity by estimating the numbers of killings predicted by the statistical interaction of the KLA and NATO data, and subtracting that estimate from the original pattern. The result of the subtraction is called the “residual,” and it describes the

¹⁵UK Ministry of Defense Briefing, Deputy Chief of the Defense Staff, Sir John Day; available at <http://www.kosovo.mod.uk/brief040499.htm> as of 3 January 2001.

Figure 11: Estimated total refugee flow and residuals over times



pattern in killings and refugee flow that remains after the effect of the control variables (KLA and NATO activity) has been removed. The result of this analysis is shown in Figure 10.

In Figure 10, the upper line reproduces the total estimated number of deaths over time as seen in Figure 2. The lower line in Figure 10 is the same pattern controlling for the statistical influence of the KLA and NATO patterns.¹⁶ With the influence of the correlations with NATO airstrikes and KLA activity removed, the pattern of killings over time remains essentially the same. All of the peaks are the same, although some of the troughs are slightly exaggerated in the lower line.

The same analysis can be performed for refugee flow. The results are shown in Figure 11. As with killings, the pattern of the refugee flow, controlling for the correlations with the NATO and KLA patterns, is strongly similar to the original pattern. However, the statistical measures suggest that the KLA activity (but not NATO airstrikes) has a weak but noticeable relationship with the refugee flow pattern.¹⁷ The relationship is particularly evident at two points in time: during Phase 1 in the northern region, and during the Phase 1–Phase 2 transition in the eastern region. In these two regions at these two times, the pattern in the residual diverges from the pattern in the estimated refugee flow. Other than these exceptions, NATO and KLA activity have little influence on the pattern of refugee flow.

The analysis of patterns of killing and refugee flow while controlling for the influence of KLA activity and NATO airstrikes shows that while there may be

¹⁶For a more detailed discussion, including the underlying regression analysis, see Appendix 2.

¹⁷See Appendix 2, Figure 21 for a detailed analysis. Other points at which the estimates and residuals diverge occur when flow is near zero, and therefore are not meaningful.

occasional coincidences, the overall effect of KLA activity and NATO airstrikes does not much change the killing and refugee flow patterns. This provides further evidence to reject the hypotheses that KLA activity or NATO airstrikes caused the killings or refugee flow.

5.4. Yugoslav forces

Turning to the third hypothesis – that Yugoslav forces organized and implemented a systematic campaign of violence resulting in killings and refugee flow: the statistical analysis of correlations cannot prove that the Yugoslav forces were the external influence responsible for the observed patterns. However, the findings of this study are consistent with the hypothesis that action by Yugoslav forces was the cause of the killings and refugee flow.

In particular, one of the findings of this study shows a circumstantial link between Yugoslav army activities and the observed pattern in killings and refugee flow. The extreme decline in the number of killings and refugee flow observed during the period 6-7 April coincides with the unilateral ceasefire declared by the Yugoslav authorities in recognition of Orthodox Easter.¹⁸ During the period when Yugoslav forces ceased hostilities, the number of killings and refugee departures fell drastically. Further links could be drawn if Yugoslav troop movements could be shown to have occurred in the same patterns observed in killings and refugee flow. However, such analysis lay outside the scope of this study.

6. Summary of conclusions

Consistent with earlier analyses, the findings of this study show that killings and refugee flow occurred in distinct surges. Over time, the flow of refugees departing their homes originated from different regions of Kosovo, and the flow occurred in peak periods, separated by periods of much lower level flow. As Figure 2 shows, killing patterns over time track the refugee flow. Thus, the patterns of both refugee flow and killings exhibit characteristics consistent with the existence of an external cause. The observation that the two processes move together strengthens this finding.

This study has also analyzed the patterns of these two series over time and by region. When the overall estimates are compared at the regional level, a clear relationship remains between the patterns of refugee flow and killings. Thus, refugee flow and killings occurred in the same places at the same times, implying a common cause of both phenomena.

The analysis also shows that two of the hypotheses proposed to explain the patterns in killing and migration, KLA and NATO activity, are inconsistent with the observed patterns of refugee flow and killings. Both KLA and NATO activity occurred more frequently after the largest number of killings and highest levels of refugee flow than it did before the peaks. When controlling for the statistical effect of KLA activity and NATO airstrikes, the patterns of killing and refugee flow over time are substantially unchanged.

¹⁸ABC News reported a Yugoslav government statement that “[t]o honor the biggest Christian holiday, Easter, all actions of the army and police will stop in Kosovo against the terrorist organization KLA [Kosovo Liberation Army] starting April 6 at 8 p.m. [3 p.m. ET].” Orthodox Easter fell on Sunday 11 April 1999. See <http://abcnews.go.com/sections/world/DailyNews/kosovo.bombing990406.html> as of 3 January 2002.

The analysis is consistent with the hypothesis that Yugoslav authorities conducted a campaign of killings and expulsions. The Yugoslav government's Orthodox Easter ceasefire coincides exactly with a drastic reduction in killings and refugee movement, and this observation reinforces the agreement of the analysis with this hypothesis.

Each of these findings is consistent with the narrative accounts of the situation in Kosovo during this time period, reported by numerous nongovernmental organizations. The coherence of the phases, the close relationship between estimated number of killings and refugee flow, and their occurrence across broad regions of Kosovo each support the claim that there was a coordinated cause of violence against ethnic Albanians during the period March–June 1999.

Appendix 1: Data and Matching

1. Introduction

The present study is based on the collection of more than 62 000 reported deaths, of which approximately 52 000 were anonymous.¹⁹ The names of 9 569 people were reported to one or more organizations that collected information about killings in Kosovo during the period March–June 1999. Appendix 1 describes how we managed both anonymous and named reports of death. As will be seen, these two types of data represented quite different challenges.

This Appendix is divided into sections, beginning with the introduction (as Section 1). In Section 2, we describe the data collection procedures that generated the basic inputs for our work. Section 3 details the initial data editing steps to clean the data and prepare them for analysis. The next section (Section 4) describes our initial attempts to identify multiple reports of the same death. In Section 5 we describe how we reviewed the matching in a second round. The final data are summarized in the last section of Appendix 1 (Section 6).

2. Data sources

The data analyzed in this study were assembled from four sources: interviews conducted by the American Bar Association/Central and East European Law Initiative (ABA/CEELI), Human Rights Watch (HRW), and the Organization for Security and Cooperation in Europe (OSCE), as well as exhumation reports produced by a number of international teams on behalf of the International Criminal Tribunal for Former Yugoslavia (EXH). Overall project summaries are shown in Figure 1.

The first row summarizes collection efforts by ABA/CEELI. They conducted 1 674 interviews in which 5 089 incidents of killing were reported. They did their data collection in five countries. The final column in the figure indicates by the “yes” entry that the ABA/CEELI data gatherers all employed a standardized questionnaire.

More generally, for each source an “incident” could involve information on deaths of more than one person. In an interview, the witness might describe one or several such incidents. Thus, an incident was a report of a single person identified by name, or of an anonymous person or group of people who were not specifically identified.

Among the reported killings from a single data source, different witnesses often reported the deaths of the same victims. Some witnesses identified victims

¹⁹Data on an additional 18 000 anonymous deaths were available but were not included because of lack of time.

Figure 1: Summary of data sources

Project	Interviews	Incidents	Where	When	Qstn.
ABA/CEELI	1 674	5 089	Albania	May–Jun 1999	Yes
			Macedonia	May–Jun 1999	Yes
			USA	May–Jun 1999	Yes
			Poland	May–Jun 1999	Yes
			Yugoslavia	Aug 99–Aug 00	Yes
Exhumations	n/a	1 767	Kosovo	Jun 1999–Apr 2001	n/a
HRW	337	1 717	Albania	Mar–Jun 1999	No
			Macedonia	Mar–Jun 1999	No
			Kosovo	Jun–Dec 1999	No
OCSE	1 837	6 686	Albania	Mar–Jun 1999	Yes
			Macedonia	Mar–Jun 1999	Yes

specifically, listing each victim by his or her full first and last name,²⁰ age, and gender, as well as date and place of death.

Other victims were identified only anonymously. Some individual victims were reported, but without a specific name (“I saw the body of an old man”). Other victims were identified as members of groups (“I saw ten people dead in a pile by the side of the road”). Bodies that were exhumed but never identified are also included in this category. These victims are referred to as *groups* (even if there is only one victim in the “group”).

2.1. American Bar Association Central and East European Law Initiative (ABA/CEELI)

The sources of the 1 674 interviews which comprise the ABA/CEELI data varied by country of collection, with different partners in each. The countries where interviewing was done included Albania, Macedonia, the United States, Poland, and Kosovo, Yugoslavia.

Albania: ABA/CEELI conducted 35% of its interviews in Albania, where they partnered with a coalition of local Albanian non-governmental organizations (NGOs) called the Center for Peace through Justice. With the Center, ABA/CEELI conducted interviews in the refugee camps and among refugees in private homes throughout Albania. Data collection in Albania began in May 1999 and ended in August 1999. In the camps, interviewers sought interviewees tent by tent.

Macedonia: About 16% of the ABA/CEELI interviews were collected in Macedonia. The interviewees included Kosovars residing with host families throughout Macedonia, but the interviews were primarily collected in refugee camps. ABA/CEELI worked with a team of ethnic Albanian citizens of Macedonia to conduct these interviews. ABA/CEELI secured interviewees through referrals from humanitarian organizations, word of mouth, and advertising in local newspapers. In the camps, interviewers sought interviewees tent by tent. The Macedonia data collection began in May 1999 and ended in August 1999.

United States and Poland: American attorneys, working through interpreters, collected interviews from refugees housed on the military base in Fort Dix, New Jersey. ABA/CEELI recruited 10% of its interviews in Fort Dix, and interviewees

²⁰The terms “last name” and “surname” will be used as synonyms.

were found through advertising and word of mouth within the camp. U.S. data collection began in May 1999 and ended in July 1999. CEELI also collected a small number of interviews (4) from a refugee camp in Poland and received a small amount of interview information collected by the Kosovo Diplomatic Observer Mission in Poland.

Yugoslavia: ABA/CEELI partnered with two organizations in Kosovo to collect information after the Yugoslav withdrawal in June 1999; interviews taken in Kosovo account for 38% of the ABA/CEELI total. Data collection by the Center for Peace Through Justice began in August 1999 and ended in November 1999 and was undertaken in the following municipalities: Djakovica, Glogovac, Klina, Mitrovica, Pec, Podujevo, Pristina, Prizren, Orahovac, Suva Reka, Vucitrn, and a small number elsewhere in Kosovo. Additional data were collected by the Council for Defense of Human Rights and Freedoms. Their interviews began in July 2000 and ended in August 2000. Interviews were conducted by opening general collection points in the centers of the following towns: Gnjilane, Vucitrn, Kacanik, Urosevac, and Stimlje.

All interviews were conducted using a standardized questionnaire that allowed for a narrative description of events. The information on the questionnaire was then keyed into a database. The coding team paid particular attention to the precision of the dates expressed by the interviewees. Some dates were identified exactly, while other dates were identified relatively (“two weeks before we left our homes”), or approximately (“some time before the Serbs came”). The date precision coding was used later for the analysis of sensitivity of the findings to date reporting errors.²¹

For the statistical purposes of the present study, all of the data were re-categorized from the original database into new data structures. All data were recoded from their original formats into standard geographic classifications and date precision codes.

The ABA/CEELI data were processed in two parts: The first portion of ABA/CEELI data included the 634 interviews taken in Kosovo. These data were compared to and completely merged independently with the HRW, OSCE, and exhumation data as described in Section 4. The second set of ABA/CEELI data (comprised by the interviews conducted outside Kosovo) had been used in an earlier publication by ABA/CEELI and AAAS (2000). These 1 040 interviews were self-matched, then integrated with the entire dataset (which included data from OSCE, HRW, the exhumations, and the ABA/CEELI data in the first set). The second phase of ABA/CEELI work was done at the end of the inter-system matching process (see also Section 4 and Section 5).

2.2. Exhumations (EXH)

Exhumations were conducted in locations thought to contain graves of Kosovars killed during the months leading up to the Yugoslav withdrawal. Although exhumations were not evenly spread across Kosovo, exhumations were conducted in 24 of Kosovo’s 29 municipalities. The total number of bodies exhumed and the number identified for each municipality are presented in Figure 2.

The exhumation data did not identify the date on which the victims had been killed, and so these data only have date identification when they match

²¹The results of the sensitivity analysis are covered in Appendix 2. We found that the substantive interpretation of the results is robust to the residual imprecision in dates due to reporting error or missing data.

Figure 2: Total number of bodies exhumed and percent identified, by municipality

Municipality	Total Exhumed	Percent Identified
Missing place	4	0.0%
Decani	54	9.3%
Dakovica	388	33.5%
Glogovac	421	39.4%
Gnjilane	54	83.3%
Dragas	1	0.0%
Istok	208	40.4%
Kacanik	142	69.0%
Klina	24	41.7%
Kosovo Polje	11	54.5%
Kamenica	8	100.0%
Mitrovica	149	50.3%
Lipljan	91	92.3%
Obilic	5	100.0%
Orahovac	368	40.8%
Pec	312	62.5%
Podujevo	90	72.2%
Pristina	357	21.0%
Prizren	510	21.6%
Srbica	343	64.1%
Stimlje	24	75.0%
Suva Reka	371	50.1%
Urosevac	22	77.3%
Vitina	8	100.0%
Vucitrn	246	61.0%
Total	4 211	45.4%

to victims in another data system (see Section 5). The place of the exhumation may or may not have been the place in which the victim was killed. Identifications were made carefully, and so the exhumation data were an especially important source to check for repetition of the same name. Many people in Kosovo have similar surnames, and it can be difficult to distinguish between people by last name alone.²²

2.3. Human Rights Watch (HRW)

From March to June 1999, HRW interviewed refugees as they left Kosovo. Of all the interviewees who gave statements to HRW, 25% were interviewed as they

²²The exhumation data provided a basic early reference for the matching issues we would encounter later. We thought initially that the names in the exhumation data were unique. In the end, while this did not prove true, the exhumation data still had the best record of identification of victims by name.

crossed the border into Albania or when they had settled in refugee camps or private homes; 11% were interviewed in Macedonia, and 3% in Montenegro.²³

From June through December 1999, HRW conducted interviews in Kosovo; 60% of the interviews given to HRW were conducted in Kosovo. The geographic regions within Kosovo were selected based on refugee reports of mass human rights violations and on reports of mass violations from sources other than refugees. Interviewees were selected for their knowledge of specific abuses inside the province. Interviews were conducted in the municipalities of Decani, Djakovica, Gllogovac, Gnjilane, Istok, Kacanik, Kamenica, Klina, Kosovo Polje, Lipljan, Mitrovica, Orahovac, Pec, Podujevo, Pristina, Prizren, Orahovac, Suva Reka, Srbica, and Vucitrn.

All interviews were conducted to elicit open narratives of what the interviewee had seen. Standardized questionnaires were not used (HRW 2001). Despite not having used a standardized questionnaire, the interviews were rich sources of information about killings. They were coded and entered into a database. Coding for the present study was independent of the original HRW database and the statistical work presented earlier in HRW (2001).

2.4. Organization for Security and Cooperation in Europe (OSCE)

The OSCE Kosovo Verification Mission (OSCE-KVM) collected 1 837 interviews which mention one or more killings. The statements were taken from March through June 1999. The interviews were conducted in more than 90 distinct locations in Albania (37% of the interviews) and at least six locations in Macedonia (61% of the interviews). There was a small number of interviews (22) for which the place of interview was not noted. No information was gathered in Kosovo itself. OSCE-KVM interviewers opened offices in central locations near refugee gathering points (mostly camps), and interviewees came to give statements. The OSCE informed potential interviewees of the project through local non-governmental organizations, announcements in the press, and contact with local clinics or hospitals. Most of the interviews (over 80%) were conducted in refugee camps; the remainder of the interviews were collected in public gathering spaces or private homes.

OSCE-KVM used standardized interview forms similar to those used by ABA/CEELI. The information was then entered into a database, also similar to that used by the ABA. For our study, the data were independently recoded, as we did for HRW. The semi-structured OSCE interviews were reformatted to be compatible with the format we developed for use with the ABA/CEELI data.

3. Initial data editing

Although the data sets were all carefully compiled by each of the collecting organizations, considerable effort was required to standardize the data to formats that permitted us to determine which records identified the same victims. Two rounds of data editing were done. In the first round, we prepared the data to be matched. In the second round, additional edits refined the data and finalized the matches. This section describes the initial editing. Section 5 describes the final edits.

²³The percentages do not sum to 100% due to rounding.

3.1. Geographic coding

All places identified by interviewees were coded to specific geographic locations. Before matching, all the several geographic systems were made to agree with a single coding scheme. A coding scheme uses a *list* of place names. Since many places have the same name, a place list is not uniquely identified by names. Instead, a *code* is assigned to each distinct place. The codes were mapped to the latitude and longitude of the place to which they referred.

We began with the geographic structure described in Ball (2000) and ABA/AAAS (2000), using 29 municipalities. These structures omitted many places introduced in the new data acquired for the present study. The place list available at the online Humanitarian Community Information Centre (HCIC) linked place names to grid positions on a detailed atlas. The HCIC list was used to standardize place names.²⁴

All place codes were coded for latitude and longitude. A first pass used the U.S. National Imaging and Mapping Agency's (NIMA) Populated Place Locations list.²⁵ The NIMA list includes latitude and longitude. The NIMA list was linked to the HCIC list using place names. When names were ambiguous, we hand-linked the codes using municipality names and checking places on the HCIC map and a commercial map.²⁶ Using the HCIC list and map, as well as a commercial map, we developed computer routines that confirmed every place code's latitude and longitude against the grid coordinates in the HCIC map. Locations which did not fall in their grid coordinate were hand-plotted and rechecked.

Several cities and villages have the same names as municipalities. Given one of these names, it was not always possible to determine whether the municipality or the village was being described. Sometimes, too, the same place name occurred in more than one municipality (e.g., Drenovac is a city or village name in four municipalities: Orahovac, Decani, Pristina, and Klina). Finally, there were cases where no place coding could be assigned (e.g., "in the mountains").

Distances between locations were calculated using their latitude and longitude.²⁷ These distances were used to determine whether witnesses' conflicting reports of locations plausibly referred to the same place. Locations less than 10 kilometers distant from each other were routinely treated as the same location.²⁸

3.2. Name and gender editing

We consulted with native Albanians and several Internet-accessible Albanian name indices in order to help interpret the names reported in the data sources.

²⁴See www.reliefweb.int/hcic/ as of 10 October 2001. Note that the HCIC list includes the municipality of Malisevo which did not exist during the first two quarters of 1999. During the time of the conflict, Malisevo was part of four other municipalities.

²⁵See the NIMA GEONet Names Server (GNS), found at <http://gnpswww.nima.mil/geonames/GNS/index.jsp> as of 3 January 2002

²⁶Interestingly, we found more than 50 locations in the NIMA list for which the latitude and longitude were 25 or more km away from their plotted positions on several maps. When this occurred, the NIMA coordinates were rejected by our grid-square check which compared latitude/longitude positions against grid coordinates in the HCIC maps

²⁷We used Haversine's Formula to calculate distances; see, e.g., <http://mathforum.org/dr.math/problems/longandlat.html>.

²⁸More distant locations were occasionally treated as the same; this occurred when far-apart places had the same name and might have been confused either by the witnesses or by the data coders. See Section 5.4 for examples.

Common misspellings of first and last names were corrected. First and last name reversals were detected and corrected either before or during matching. Some of the spellings were phonetic, having been recorded by individuals who did not speak Albanian, and others were obvious data entry errors.

Gender was given directly by the interviewee or it was coded from the first name (when a first name had been given). First names were cross-checked to be sure that the same first name was always assigned the same gender.²⁹

A search was made for identifiable Serb victims. Some of these were obvious, e.g., a designation such as “Serb Commander.” Others were identified by checking against a list of common first and last names for Serbs.³⁰ A total of 30 Serbs, identified by reference to the name lists, were dropped from the estimates.³¹

3.3. Date of death formatting

Data edits were performed to correct confusion caused by differences in the order in which day and month conventionally are entered by Europeans (day, month) and Americans (month, day). Other records had dates of death with out-of-range year values (e.g., “1990” and “2999” which were both reset to 1999). After editing, all dates were standardized to the ISO YYYY-MM-DD format.

Although labeled “date of death,” interview reports usually were a mixture of actual remembered incident dates and dates when bodies were seen. Therefore, the date given could have been later than the day that the death actually occurred.

When the original interviews were entered into the databases, the precision of the date information was coded as “exact,” “approximate,” “imprecise,” or “unknown.” The precision coding corresponds roughly to the degree of precision defined to the day, week, or month of the event, or no precision. As a result of matching records to other records, multiple dates were sometimes available for each record. The precision coding was used to select the “best” date, as described in subsection 5.2.³²

4. Initial data matching

It was our working hypothesis that each data system (except the exhumation data) could contain many duplicated reports of the same victim’s death. These duplicates, of course, had to be found prior to doing any analysis. The problem of duplicate records was divided into four subtasks: intra- and inter-system matching for individuals, and intra- and inter-system matching for groups.

Each identifiable individual record was first compared to the rest of the identifiable individual records in its dataset of origin. This process is called

²⁹A reviewer noted that some first names may be used by people of different genders. Since gender played a relatively small part of our matching logic, this editing rule cannot have significantly affected our matching.

³⁰For some of the resources we employed, see <http://www.kabalarians.com/male/serb-m.htm> and <http://toybox.flickr.com/onomastikon/Europe-Eastern/Former-Yugoslavia/Serbia/Surnames.htm>.

³¹Logically, an equally tiny fraction of the anonymous deaths would also be Serb victims. These numbers would be too small to affect the interpretation of our estimates. We have therefore ignored their effect.

³²As already noted, Appendix 2 examines the robustness that exists relative to known weaknesses in date reporting.

“intra-system matching” or “self-matching” because it matches a single data source to itself.³³

After each data source’s individual records were self-matched, the reports from each source were matched to those of every other source. We called the process “inter-system matching.” Named individuals in each system were compared to named individuals in every other system and matches were recorded. The primary variables employed to check for duplicates were name (first, last) and geographic location. Other information, such as date of death, age, or gender, were also considered in order to confirm or reject possible matches.

The same process of intra- and inter-system matching was then repeated for anonymous group data. Location and time were the two key variables used to bring potentially duplicative reports of anonymous killings together. Conceptually, reports of anonymous deaths could contain individuals who were identified by name in other reports. We found several ways to combine the individual and group data as described below. The approaches we considered provide a credible set of lower bounds on the total number of killings, which are described in Section 6. In Appendix 2, modeling approaches are described to improve on these lower bounds.

Although the matching process was computer-assisted, the decisions were made by people. Matching was done by a small team of carefully-trained coders supervised by one of the authors. During the second round of matching described in Section 5, other steps were taken to measure the quality of the matching.

4.1. Variables used for intra-system matching of individual records

The primary keys for identifying duplicate reports of the same individuals were last and first names. There were many common misspellings (or erroneous transcriptions) of certain names, including the following: Hysen, Hyseni, Iseni; Ymeraj, Imeraj; Krasniqi, Krasnici, Krasniki; Kuci, Kuki, Quki, Quci; Cake, Caka, Cakaj; and Loki, Loku. In general, “H” as the first letter was often omitted in the transcriptions of the interviews.

Last names beginning with certain letters often occurred in different combinations: K could be C or Q; Y and I were often confused; and Xh could be Gj or Sh. All of these combinations were routinely compared. In addition to the routine rules, less regular but obvious misspellings were sought involving similar-sounding letters in the middle of names. Over time, the coders became familiar with the variety of possible spellings for different names.

When records contained similar names, they were considered to be matches unless other information clearly distinguished them from each other. Information that might weigh against matching two records included the age and sex of the victims, and the dates and places of death. Ages were rarely useful since they were frequently approximate. However, when the ages differed by 20 or more years, it was assumed that the records were different; we theorized that two identically named victims of different ages were likely to be relatives. Information on the sex of the victim rarely differed, unless the first names did as well, because in earlier data editing stages, first names had been used to differentiate between the sexes.

³³The exhumation data were not checked for duplicates in this round because the records were assumed to be unique. In Section 5, we describe how this assumption was later changed, and some names were found to be duplicated.

Place of death was the most important additional information used to evaluate name matches. Records for which the names matched and the ages differed by less than 20 years were considered to be matched when the places were identical. In particular, if the places were in the same municipality or in adjoining municipalities, they were treated as a match.

4.2. Basic approach of intra-system matching of individual records

The record linkage literature offers many approaches for matching individuals in lists.³⁴ The complexity of the data described here, however, required us to use manual methods. As we learned from and edited the data, we were able to increase the automation of the process. Due to resource limitations, we were not able to quantify all the errors in the matching process such as was done in Belin and Rubin (1995). We relied instead on repeated rounds of computer-assisted matching that concluded with very few matches being found in the final passes of the last round. To minimize the impact of residual matching errors, we tried in all cases to err in the direction of too many matches, which would tend to decrease the estimates.

Within each dataset, records were matched by printing paper lists in a spreadsheet format. The lists were sorted on several variables: by place, then in a subsequent pass by last name–first name, then in another pass by first name–last name.³⁵ Although matchable records may sort to positions quite distant in one sort, they would appear close together in at least one of the other sort orders. For example, two records with identical first and last names would appear together on both name sorts. Two records with differently spelled last names but identical first names would appear together when sorted by first name. When found together the varying last names would be readily identifiable.

Using the multiple-sorting (or multiple-blocking) technique, coders identified blocks of records that were the same. The record with the best data (the best name spelling and most precise date and place location) was chosen. All the record numbers were grouped in a “circuit” and preserved for later analysis of the most likely date (as described in Section 5) for the final record.³⁶ This record was called the “key” record.³⁷

The technique changed as confidence in the coders increased. In the first part, three coders independently self-matched all the HRW data and the portion of the ABA data that consisted of the interviews conducted in Kosovo. After the HRW and the ABA self-matches were completed, the approach seemed sufficiently routine so that one coder could achieve results of almost as high a reliability as two or three. Therefore, only one coder self-matched the OSCE data and the second portion of the ABA data.

Although the process was routine, the match results were complex because they involved collapsing an indefinite and unpredictable number of records into one “key” record. Although errors were not always obvious, the internal consistency of the collapsed records could be assessed (see below).

For the HRW and ABA self-matches, the individual victims who were matched differently by the two coders were identified. The coders discussed the differ-

³⁴Two linkage conferences (1985 and 1997) were co-organized by one of the authors of the present study and provide access to this rich literature

³⁵See Scheuren (e.g., 1985) for more on the statistical properties of multiple blocking methods.

³⁶The use of the term “circuit” in Asher and Ball (2001) is different, but the analytic issues are similar.

³⁷See Appendix 2 for the use made of this information in sensitivity analysis.

ences and jointly developed a consensus list of matches.³⁸ Inter-coder agreement varied from between 75% and 90%. All coders' differences were reviewed and resolved by an author of this report.

For all self-matches, the structure of the match decisions was evaluated for its plausibility. In particular, for each system, each circuit containing two or more records was evaluated, comparing the key match fields for agreement. For each source, for each circuit containing more than one record, all the records were compared to the key record. For each of three fields (surname, date, and place), a count of how many records agreed with the key record was made.

Surnames The proportion of records within circuits which matched within the first 3 and 7 characters was tabulated. This comparison was done only for records with names at least 3 or 7 characters long. The last-name matches tended to be very close: considering the three datasets, between 85% and 95% of the matched records shared at minimum the first three letters. The records that do not share the first three letters of the last name are not necessarily mismatched. Arguably, the name-similarity index measures the rate of spelling or transcription variation among the original interviewers. We interpret the high rate of agreement as an indication that the names were most often matched to other similar names.

Dates The proximity in time of the self-matches was also considered. Between 79% and 84% of the precise dates on records matched in self matching were within one week of one another. Dates coded as approximate were not compared.

Locations With precise location codes, individuals in the self-matches had identical codes between 66% (OSCE) and 99% (HRW) of the time; by expanding the place comparison to places within 25 km of each other, an additional 28% of the OSCE matched records agreed on the place coding, raising the place-agreement rate to 94%.

Similar names, dates and places that do not match exactly may reflect differences in the witnesses' recollections — they are not necessarily coding errors. However, having high agreement on these measures suggests that records that had similar names, and that dates and places were appropriately matched when they were close in time or space. Records with dissimilarities were less often matched.

Our results from these initial intra-system matches of named individuals reduced by about one-third the number of records that went on to later steps. The duplicates found afforded us a way of improving the reporting of dates and learning more about Albanian name variations. Although this round did not find all the duplicates, it provided a foundation on which to begin the inter-system comparisons described in subsection 4.3.

As we will discuss in Section 5, later steps in the process made it possible to detect most of the remaining duplicates. There may have been a small amount of "overmatching," however, which would not be detectable in later steps and, hence might be a cause for concern. Overmatching can occur when records for two distinct individuals are linked in the self-matching step. There is no way to fix overmatching errors later in the process since the record for one of the individuals is not available. Overmatching has the effect of reducing the total

³⁸The individual coders' decisions were preserved for use in sensitivity analysis.

number of killings, and hence it tends to lead to underestimates of the total deaths.³⁹

4.3. Inter-system matching of individual records

Inter-system matching consisted of comparing each individual record in one data source (the “source”) with all of the possible matches in another data source (the “target”). As with the intra-system matching the work was all done manually. The possible records in the target database included all the individuals whose names began with the same letter (or one of the sound-alike letters described above). The spreadsheet approach used in the self-matching was replaced by custom software designed to facilitate matching decisions.⁴⁰

Each source dataset was divided into subsets. Each subset (called a “slice”) was a proportionally stratified (on date of death and region) random sample of the whole. Slices were designed to represent approximately one half-day’s work. Whenever a pair of coders finished the same slice, a supervisor compared their results and reviewed all disagreements with both coders. In this way, different coding styles were brought together, and subtle differences in coding practice were detected and eliminated. Coders were given one or more training slices, and their work was not accepted until they reached at least 90% agreement with the standard answers for the training slices. Coders whose work had low rates of agreement with the training slices were identified and they received additional training.

The unduplicated “key” records remaining after the self-matching step were the inputs to the inter-system matching. The match comparisons were made using the same rules as in subsection 4.1. Each record in each source was exposed to every record in each of the other three sources. The comparisons were not symmetrical. That is, if source A were compared to source B, source B was not then compared to source A (although for additional redundancy, occasional symmetric comparisons were made). Since at least two (and often more) coders made match decisions for each comparison, there are more than twice as many decisions as possible comparisons. Altogether there were 18 462 match decisions made, and the raw proportion of agreement overall for the decisions was 94%.

Even though the proportion of agreement among coders was quite high, there were some disagreements. An author of this study reviewed every disagreement among the coders and made the final decision.

As previously mentioned, the interviews taken by the ABA outside Kosovo were handled differently than the other sources. While the self-matched steps for this portion of the data were identical to those described in subsection 4.2, the inter-system matching was done after the other systems had been matched and merged. In all other respects, though, the steps were the same. The matches were performed independently by two different coders, and their decisions were compared. The inter-rater proportion of agreement was relatively lower for this portion of the matches (approximately 80%). However, as with all the matches, all disagreements were reviewed and resolved.

³⁹As noted elsewhere, when errors were unavoidable, we have elected to err in a direction that would lead to understating the number of killings.

⁴⁰The matching application was an HTML client which the coders accessed using web browsers. The application itself was written in PHP and MySQL running on a local intranet. Data were processed using Python and SQL, and the statistics and graphs were generated using Stata. The graphs were edited in Adobe Illustrator, and the text typeset using $\text{\LaTeX} 2_{\epsilon}$.

The initial inter-system matching noted areas in the data where additional checking was needed, notably for the exhumation data, but in other places as well. Interestingly, as with the self-matching round, the number of deaths was again reduced by about one third. In Section 5, we describe how we used what we learned to make the final match decisions for individual deaths.

4.4. Intra- and inter-system handling of anonymous group records

All of the matching described so far has compared named individual victims to other named victims. The sources, however, describe approximately seven times more victims anonymously in groups than individually by name. By its nature, anonymous group reporting has less information available for determining whether group reports uniquely match each other. After substantial analysis, we found that it is not possible to match individuals to groups within data systems, nor to match groups to other groups across data systems with sufficient reliability to model the missing information, as described in Appendix 2.

Nonetheless, there are several benefits from matching groups to each other and to individuals. First, matching records of all kinds provides additional information about the precision (or possible imprecision) of date and place identifications. If the most likely match for a given record (based on qualitative information in the notes fields, or a match by place while the date is different) is another record that is distant in time or space, this implies that one of the two records has imprecise date or place information. Second, by matching groups and individuals within systems, a basic lower bound for the number of killings missing from the identified list can be estimated.

The matching process for groups was as follows.

First, we self-matched group reports within each data source. Anonymous groups were collapsed within specific places and dates. That is, groups who were identified as killed in a particular location within approximately 10 days of each other were considered duplicates; wider date ranges were collapsed when the more distant dates were imprecise, or when narrative information available in the interview notes suggested that the incidents were the same. In nearly all village-level locations, there were very few groups, and they clustered at particular dates. After the group matches were made, all groups with more than one matching record were evaluated. Only 15 had dates spread more widely than 14 days apart. The “best” date was chosen using the same logic used for the individuals who were self-matched.

Second, individuals were matched to groups within each data system. The matching was done primarily by location and date, although in many cases additional information about the group aided the matching. Some otherwise unidentified records were noted “brother of victim 27,” or a group might be documented as “the X family.” Matching individuals to groups increased the information about the individual killings in some cases. For example, if the report of the individual killing did not have a specific date on which the killing occurred, an anonymous (group) report could provide a date. Even if the individual report contained date information, the group report was used to confirm the original.⁴¹

Third, individuals were matched to groups in other data sources. This process provided the same information-leveraging benefits described above.

⁴¹The additional dates were used in the evaluation of date precision. See Appendix 2.

Fourth, the group counts provide a method for examining the number of deaths that were unreported as individual records. With a count of the unique groups, we can evaluate whether the pattern of killings that were not reported as individuals is distributed uniformly or non-uniformly over space and time. This question is examined in detail in Appendix 2.

4.5. Merging anonymous group killings across systems

Approximately five times more victims were reported to the interviewing projects as members of anonymous groups than were reported by name.⁴² In this section, we describe how we use this information to explore the number of victims who were not identified by name.

After self-matching the groups, we determined that the level of duplication in the group data was high: the total reduction from the reported data to the unduplicated data was more than a factor of five. The unduplicated group records resulting from the match process were composed of one or several group records. Thus each record contained a distribution of group sizes that could be used to estimate its “best” size. Three sizes were estimated for each group. For an estimated minimum size, we took the smallest reported group size that was greater than the number of individuals matched to the group (within that data system). The median size was the median of all the sizes greater than the number of individuals matched to the group. The maximum size was the greatest reported size of any group in the circuit.

Sums over the group records by time and space are unduplicated within each data system. However, group counts cannot be directly summed across data systems because the group data were not matched across systems. We unduplicated the group counts by comparing the sum of group counts (using the minimum, median, and maximum counts) across the three systems. We chose the maximum of the three systems at each point, thereby assuming that the other two systems were completely matched to the largest one. This is the most conservative possible merging rule.

The resulting data are the estimated total anonymous killings over time and space, and they are used to evaluate the completeness of the unique individual records (see Section 6.3).

5. Refinements in data editing and matching

Once all the initial intra- and inter-system matching and editing decisions had been applied to the raw data, the combined dataset could be reviewed using information accumulated in the previous rounds. The review focused on match inconsistencies, on choosing the best variable to represent the entire set of matched records, on imputing for missing dates, on reassessing the exhumation data, on a general cleanup of spelling inconsistencies, and on other errors that were found in our initial work.

5.1. Inconsistent matches

Because of the way we matched targets to sources in the initial match, it is possible for records to match in inconsistent patterns across datasets. For example,

⁴²As previously mentioned, records on an additional 18 000 group deaths were available, but were not processed due to lack of resources.

consider three pairwise match decisions: $A1 \rightarrow B1$, $C1 \rightarrow A1$, and $B1 \rightarrow C2$. In this example, record $A1$ (from dataset A) was compared to dataset B and matched to record $B1$. In a separate match, record $C1$ was compared to dataset A and matched to record $A1$. In a third match, record $B1$ was compared to dataset C and matched to record $C2$.

When merged, records $A1$ and $B1$ seem to match both $C1$ and $C2$. If dataset C was properly self-matched, this is a contradiction. In total, there were 298 records like $C1$ and $C2$ were found. The solution was clear: in each of the pairs of overmatches, the matching pattern of one of the records had to be modified.

All of these contradictions were reviewed and resolved. There are two ways to resolve them: One possibility was that $C1$ and $C2$ should have been matched in the intra-system matching. In this case, one of the two records can simply be dropped from the analysis.⁴³ Eighty records were resolved this way. The second possibility is that either $C1$ or $C2$ was matched in error and should be unmatched; the remaining 218 overmatched records were resolved by separating them. In some cases, if the records were matched in error, there may have been a true match which was obscured by the erroneous match. These potentially missed matches were sought in the final editing step (see Section 5.4).

5.2. Choosing the “best” dates

Each record in the final dataset was a combination of all the records that matched to it. This combined record potentially has many dates to choose from. The selection of the “best” date proceeded as follows. First, we aggregated the matched records by date. Usually in a set of matched records, one date is much more common than other dates. We chose this date first, if possible. Among the remaining dates, we chose the date with the highest level of precision (defined by the most precise record within that circuit). If there was more than one date with that level of precision, we chose the date with the largest number of constituent records in the circuit. When there were ties (dates with the same precision and number of constituent records), we chose the earlier date. We reasoned that later dates were more likely the result of people having seen the bodies after the killing, rather than having seen the killing itself.

For 204 records with no date information, a “hot deck” procedure was employed to assign a date at random from a “donor” record that was geographically closest to the location of the record with the missing date.⁴⁴ Three dates were randomly selected from the potential donors, and copies of the original record were created with each of the sampled dates. The new records were each assigned a weight of 0.33.⁴⁵

⁴³Dropping a record means adding it to the circuit of matching records in the self-match. In this way, information in the “dropped” record is still available to the “kept” record.

⁴⁴“Hot decking” (e.g., Ford 1983) imputes missing information to a record by finding another record—a “donor” record—with non-missing information which is identical, or nearly so, in all other respects. Here we used geographic proximity to select the donor. To reduce the Monte Carlo error introduced by the imputations we first created potential “donor” groups of 60 records each, (in 85% of the 457 villages identified as locations of one or more killings we were able to find 60 or more valid records available within 10 km).

⁴⁵Again, as part of dampening the imputation error, dates were imputed to records three times, with a weight of 0.33 assigned to each resulting record. The motivation for this use of multiple imputation is set out in Oh and Scheuren (1983). We are not using multiple imputation in the sense described by Rubin (1987). In particular, our goal is not to try to calculate variances. The residual uncertainty arising from the imputation process is almost certainly small (see, e.g., Converse and

Some of the hot-decked dates were outside the date range of interest to this study (20 March–22 June). Those records (and their partial weights) were therefore excluded from the analysis.

The sensitivity of the estimates to the “best” date choice and the imputation are analyzed in Appendix 2. As Appendix 2 shows, the main statistical results are robust against uncertainties caused either by date inconsistencies or because in a small fraction of the cases, an imputation had to be made.

5.3. Exhumation data

In the initial editing and matching steps described in Sections 2 through 4 we had assumed that the exhumation data were unique by definition. That is, we assumed that there were no cases in which different remains were identified as the same victim. At this stage, we examined that assumption critically. Records with identical names, in which the place of death was the same village, and the ages were identical (or missing) were matched. However, even if the names were identical, but the ages were recorded and distinct, the records were not merged.

In all, 232 named exhumation records were self-matched in this way. This resulted in a net decrease of the total named, individual exhumation records by this number. We increased the number of anonymous deaths by the same amount with the reasoning that the exhumation reports documented two bodies, but the bodies had other forms of indirect identification that led to duplicate registration of the same individual.

5.4. Other edits of the final matches

When the entire dataset had been compiled, one of the authors spent five days reviewing it. In this process, she identified 329 match modifications, as well as 400 corrections to names, dates and places. The process was simple: sort the list by a key field (last name, first name, or place), and then scan the list looking for repetitions.⁴⁶

The 400 corrections were primarily corrections to variable name spellings. For example, one victim’s surname was “Pashi,” a clear misspelling of “Gashi.” Special attention was given to the first names and surnames that occurred only once, on the theory that these were likely misspellings.

At this stage, names, dates, and places that were clearly inconsistent across data systems were reconciled. For example, there were cases in which an entire family had been identified in one location in the exhumation data, but were reported by another source in an entirely different municipality. Checking the locations, we determined that these were not simply case of miscoding. We believe that they represent cases in which the bodies were buried in locations distant from the killings.

One of the most striking examples of this sort included the bodies of seven members of one family who were exhumed in Donja Sipasnica. Five of them were reported to the ABA/CEELI project as having been killed in Susica, 70

Scheuren 2001). Nonetheless, there is some possibility for residual bias, so the step of conducting sensitivity analyses (as in Appendix 2) seemed warranted. For this reason, we have also described all the significance testing and confidence intervals calculated in Appendix 2 and included in the body of the report as “nominal.”

⁴⁶In the cases of commonly misspelled first and last names, the list was arranged so that records with the common variant spellings appeared next to each other.

km distant. The remaining two family members were not identified in any of the interview-based datasets. All seven people were assigned to Susica. When records disagreed in the exhumations and the other datasets disagreed, the location from the interviews was used in preference to the exhumation location.

Some of the other checks were more complex. For example, we discovered additional matches which had been missed because they were coded to locations quite distant from one another. The coders did not originally match the records because the distance measure indicated that the places of death were far apart. When we reviewed the list, we discovered that the villages were in different municipalities but that they had the same names. For example, the villages of Pograde (in Gnjilane) and Pograde (in Klina) were confused, and different reports about the same victim were coded variously to the two locations. The coders rejected the match because the two villages are 83 km distant from each other. We found 5 cases of this sort.

Some records required more significant modifications. Once the complete name corrections were made, we found 153 records that should have been matched in the self-matching stages. Finally, 176 new matches were found after the editing. Each of these records was merged with its new links, increasing the number of sources in which they were found.

Although undetected duplicate records may still exist, we believe that there are now very few. After we finished the final review, we checked how many times we had missed a match that we should have caught using our initial matching rules. We found 97 new matches that theoretically could have been found under the original matching rules. This represents an error rate in our initial process of less than 2%. This low error is the reason we believe that after the final matching round described in this section, any remaining errors are negligible.

6. Final summary of data results

To summarize the results of the data management steps, we will look at the results achieved source-by-source for individual records. Second, we examine the results obtained by linkages across sources, again for individual killings. Third, we examine ways to combine anonymous killings with individual deaths.

6.1. Data handling by source for individual records

Earlier, in Figure 1 we provided counts of the inputs we obtained from each source. We now can summarize the number of killings of named individuals that result after both the initial and final intra-system matching steps. These results are shown in Figure 3.

The interview sources were reduced by 29%–51%, while the exhumation data were reduced by 11%. What Figure 3 does not show is the extent to which the matching and editing process improved the quality of information in all the data sources by leveraging information across matching records.

6.2. Data handling across sources for named, individual records

Another way we can summarize the results of our data handling is to look at the total number of individual deaths from all sources, after double counting has been eliminated. The number of unique individual records found in each

Figure 3: Individual counts from basic sources, gross total and net unduplicated counts

Dataset	Gross Individuals	Unduplicated Individuals
ABA/CEELI	2 800	1 528
EXH	2 155	1 910
HRW	966	685
OSCE	3 648	1 786

Figure 4: Number of individual victims of killing, by documentation status (including victims with imputed dates of death)

		ABA					
		yes	yes	no	no		
		yes	no	yes	no		
HRW	OSCE					Total	
yes	yes	27	32	42	123		
yes	no	18	31	106	306		
no	yes	181	217	228	936		
no	no	177	845	1 131	n.a.		
Total						4 400	

combination of matches is presented in Figure 4. It includes only records that had valid date information in the range 20 March to 22 June. Victims whose deaths occurred on dates before 20 March or after 22 June were not included.

The table indicates that of the total number of 4 400 individual deaths, relatively few victims were documented by three or four projects, as shown by the cells in the upper left. For example, only 27 victims were documented in all four data sources. This cell is at the upper left of the table, at the intersection of the “yes-yes” row and the “yes-yes” column.

Moving down and to the right, the cells show the values for progressively less frequently documented victims. For example, 1 131 victims were documented only in the exhumation data.⁴⁷ Figure 4 does not include an estimate for the “no-no-no-no” cell (shown as n.a.), that is, the number of people who were not individually documented by any of the four projects.

⁴⁷Victims with imputed dates are disproportionately in the cells with fewer matches. Records that had more matches had more opportunities to acquire date information, while more sparsely matched records had fewer opportunities to get date information. The way the imputations were done left some records with fractional values that are summed in the figure. The results were rounded to the nearest integer. For example, the “yes-yes-no-no” cell was rounded in this way from 176.66 to 177. The total reflects the rounded sum.

6.3. Evaluating the completeness of the individual data

The individual deaths must be underestimates. The “no-no-no-no” cell in Figure 4 above cannot be empty. There are several possible methods by which the number of killings that were not individually identified may be estimated. The anonymous group data provide the first indication that there is a substantial number of victims who were not documented among the named individuals.

The group estimate was compared to the individual estimate at each time and space point, as described in subsection 4.5. Again, taking the most conservative possible merging rule, we subtracted the sum of the individuals at each time-space point from the group sum at that point. This assumes that every documented individual was also documented as a member of an anonymous group. The result of the subtraction is a “net” group count, that is, the number of victims identified in anonymous groups that remain after all the fully identified individuals are removed. This can be interpreted as a minimum lower bound of the number of victims who were not documented as individuals. The minimum, median, and maximum net group counts are 2 755, 2 889, and 5 859, respectively.

When summed with the 4 400 individual victims, the group counts produce overall estimates ranging from 7 155 to 10 259. These values estimate the total number of documented victims. These estimates exclude victims who were not documented as groups or as individuals, and so these numbers still underestimate the total deaths. Other methods of estimating the total number of victims of killing are presented in Appendix 2.

Appendix 2: Statistical Methodology and Analysis

1. Introduction

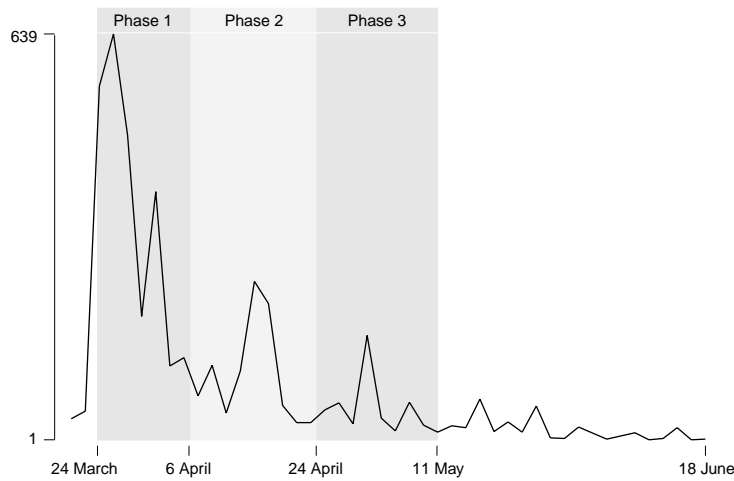
Appendix 1 of this study details the process by which a cross-classification table is created of counts of individual victims of killing in Kosovo during the period of March–June 1999. The total number of identified killings given in Appendix 1 is 4 400, and a range of estimates of the total number of killings based on documented killings of unidentified victims is given as 7 155 to 10 259. It is improbable that all killings were captured by the data collection process underlying these raw counts. For this reason, a suite of statistical methodologies is required to use the data described in Appendix 1 to estimate the most likely number of killings in Kosovo in March–June 1999. Appendix 2 describes, in detail, all of the statistical methodology required to produce the estimates used in the main body of this study and the logic used to choose the methodology.

The organizational structure of Appendix 2 is as follows. In Section 1.1, the limitations of the direct counts as a measure of total killings and the trends of killings over time and space are discussed. These limitations motivate the use of multiple systems estimation techniques to model the counts of killings over space and time. Sections 2.1 through 2.3 introduce several methods of multiple systems estimation modeling, and Section 2.4 discusses model selection procedures. Section 3 begins with an exploration into the validity for these data of the assumptions underlying the models described in Section 2. It continues into a collection of modeling procedures that both accounts for these assumptions and also results in a series of internally consistent estimates for different levels of temporal and spatial aggregation.⁴⁸

While this appendix mainly documents our thinking on ways to estimate deaths that were not reported, we also look at other modeling issues important to the hypotheses evaluated in the body of the study. Section 4 presents an analysis of the relationship between the killing counts and the NATO and KLA data. This appendix concludes, in Section 5, with a brief sensitivity analysis of the date of death reporting described in Appendix 1.

⁴⁸As a quality check on the estimate production process required for this study, software routines for all the estimation procedures discussed in Appendix 2 were independently created by Jana Asher using Splus 2000 and SAS Version 8 and Patrick Ball using Stata 7. The results were compared, and when there were differences, the routines were debugged until the results matched.

Figure 1: Documented killings over time



1.1. Limitations of direct observations

A direct estimate of killings documented in the four data sources used in this study is 4 400, and a direct interval estimate of the total number of killings is given as 7 155 to 10 259.⁴⁹ There is good reason to believe that these numbers do not represent an accurate count of the killings in Kosovo during this time period. These data were compiled from a collection of interview and exhumation data. Believing that all killings were documented by these sources assumes that all relevant bodies were exhumed and identified, or that all killings were witnessed and reported in the survivors’ interviews, or that all killings were captured by at least one of these processes. This scenario is implausible.

Given that the data used to develop this study are incomplete, the question of how accurately they reflect the true patterns of killings over space and time arises. Figure 1 plots the 4 400 documented killings given in Figure 4 of Appendix 1 over time by two-day periods. Note that the characteristics of the estimated count time series presented in the main body of this study are clearly apparent in the time series of the raw counts given in Figure 1. The largest number of killings occur in Phase 1; the quantity declines sharply on 7 April, then rises again to a mid-April peak. After a decline to near zero 23-25 April, the series rises to several small peaks in early May.

This raw count series might be sufficient to substantiate the analysis made in the body of the report if it could be shown that the pattern over space and time of the “true” counts of killings in Kosovo during March–June 1999 is accurately reflected by the documented killings cross-classified in Figure 4 of Appendix 1. There is no way to directly compare the time series of the raw counts of killings to these “true” counts. We can, however, compare the group data counts described in Section 6.3 of Appendix 1 to the individual count data. As

⁴⁹The direct intervals are lower than they would have been had we had time to process 18 000 anonymous deaths reported on the ABA interviews done outside of Kosovo that we were unable to use. If we had been able to integrate these additional records, the directly observable lower bounds would have increased, bringing them even closer to the model estimate we chose.

the analysis of the net group data in Appendix 1 showed, there were many victims who were not documented as named individuals. If the distribution of the individual count data over space and time varies considerably from the distribution of the group data, it is likely that it could vary considerably from the distribution of the true number of killings as well.

A measure of the similarity of the distributions of the individual and group data is given by the absolute relative difference between the group and individual data counts:

$$\text{absolute relative difference} = \frac{|\text{group count} - \text{individual count}|}{\text{individual count}} \quad (1)$$

If the distributions over space and time for the individual and group counts are similar, we would expect the absolute relative differences defined in (1) to be constant, and the distribution of these absolute relative differences have a small standard deviation. In fact, if the absolute relative differences of the two-day regional counts are created, 114 of 192 are zero, indicating perfect agreement between the group and individual data.⁵⁰ The 78 remaining absolute relative differences, however, range from .01 to 48, with a median of .83, mean of 2.67, and standard deviation of 6.48. We conclude that the absolute relative differences are not constant, suggesting that the individual and group data counts by two-day period and region do not follow the same distribution. As a result, we must attempt to estimate the “true” counts of killings in order to support or contradict the hypotheses given in the main body of this study about the distribution of killings in Kosovo in March-June 1999. Section 2 presents the statistical technology required to do just that.

2. Methodological background

Multiple systems estimation, or multiple recapture estimation, has a long history that originates in the estimation of counts in wildlife populations. Basic capture-recapture modeling goes back to at least Peterson (1896) and has been used in a diverse set of fields, including epidemiology (see International Working Group, 1995a, 1995b), general population counts (see Sekar and Deming, 1949; also Hogan, 1993, and Anderson and Fienberg, 2001a), and, in the area of human rights, for estimation of the number of killings during the violence in Guatemala between 1960 and 1996 (Ball, 2000b).

2.1. Dual systems estimation

The simplest version of this methodology, dual systems estimation, occurs when two separately collected but incomplete lists of the members of a population are available. Dual systems estimation relies on three statistical assumptions. The first assumption is independence of the lists, which is described statistically as follows:

$$\Pr(\text{record } i \text{ on list } L_1 \mid \text{record } i \text{ on list } L_2) = \Pr(\text{record } i \text{ on list } L_1).$$

⁵⁰When the total of the group data was smaller than the total of the individual data, the relative difference was set to zero.

In other words, if the lists are independent, the presence of a person on one list does not predict the presence or absence of that person on the second list. The second assumption is homogeneity of the population being captured; in other words, that each member of the population has an identical capture probability for a given list. The final assumption is error-free matching across the lists.

If the three conditions described above are met, then dual systems estimation is a viable estimation technique for a total population count. Let $x_{ij}, i, j \in \{0, 1\}$, represent a count in a two-way cross-classification table of population counts for two lists, as follows:

		List 2		Total
		In	Out	
List 1	In	x_{11}	x_{10}	x_{1+}
	Out	x_{01}	x_{00}	x_{0+}
	Total	x_{+1}	x_{+0}	$x_{++} = N$

Here x_{00} represents the count of members of the population that are not captured by either list, and “+” represents the summation of the counts over the lists (e.g., $x_{1+} = x_{11} + x_{10}$; $x_{+1} = x_{01} + x_{11}$). The goal is to estimate N , the total count of members of the population, and the traditional estimator for N is simply

$$\hat{N} = x_{10} + x_{01} + x_{11} + \lfloor \frac{x_{10}x_{01}}{x_{11}} \rfloor \quad (2)$$

where $\lfloor \frac{x_{10}x_{01}}{x_{11}} \rfloor$ is the integer produced by rounding $\frac{x_{10}x_{01}}{x_{11}}$. It can be shown that the estimator of \hat{N} given by (2) is derived by assuming the following identity holds:

$$\frac{x_{10}}{x_{00}} = \frac{x_{11}}{x_{01}} \quad (3)$$

A problem arises in dual systems estimation if the underlying assumptions of independence of lists and homogeneity of capture probabilities are not valid. In that case, several alternative methods of estimation have been developed, but all rely on the addition of at least one more list to the system.

2.2. Triple systems estimation

While there is basically only one method of estimation in dual systems estimation, the addition of a list allows greater flexibility of modeling for triple systems estimation. In this subsection, several methods of triple systems estimation are explored. The first, due to Marks, Seltzer, and Krótki (1974), uses a combination of dual system estimators to determine \hat{N} . The second, taken from Bishop, Fienberg, and Holland (1975), is based on loglinear models. For completeness, triple systems estimation via full and quasi-symmetry models is briefly discussed. The underlying data structure is the same for all of these methods; triple systems estimation relies on a three-way cross-classification table of population counts formed as follows:

An alternative triple systems approach is estimation of \hat{N} through loglinear modeling (e.g., Bishop, Fienberg, and Holland, 1975). In some data collection settings, loglinear modeling can better account for dependency than (11), as well as allow for reasonable sample standard error calculations and fit statistics. By creating a loglinear representation for the expected counts, $m_{ijk} = E(x_{ijk})$, a model for the observable cells is formed that is then projected to the unobserved cell. The form of this model is as follows:

$$\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} \quad (12)$$

with constraints on the u -terms (e.g., that they add to zero across any subscript). (12) is the standard no-second order interaction model, or, in other words, the model that allows for dependency between pairs of lists but not three-way list dependency. Because there are only seven potentially observable cell counts available, this is the saturated triple system model that fits the data perfectly, i.e., the maximum likelihood estimates for the expected counts are $\hat{m}_{ijk} = x_{ijk}$.

Within this framework, reduced models can be fit to the data by removing parameters from (12). Typically, these parameters are removed carefully to ensure that the resulting loglinear model is hierarchical. In other words, higher order terms may only be included if the related lower order terms are also included, so that higher order parameters reflect only the higher order relationships between the lists (see Fienberg, 1978).

For any hierarchical loglinear model chosen, the expected cell values under the model are estimated and the resulting model is projected to the missing $(0, 0, 0)$ cell. For both the saturated and the reduced models, the estimate of N , \hat{N} , is:

$$\hat{N} = n + \frac{\hat{m}_{111}\hat{m}_{100}\hat{m}_{010}\hat{m}_{001}}{\hat{m}_{110}\hat{m}_{101}\hat{m}_{011}}. \quad (13)$$

Bishop, Fienberg, and Holland (1975) give asymptotic variance equations,⁵¹ derived via the δ -method, for each \hat{N} derived via triple systems hierarchical loglinear models. These equations are used within this document to form approximate nominal⁵² 95 % confidence intervals for estimates derived from loglinear models for 3-way cross-classification tables.

Other models

Both the Marks, Seltzer, and Krótki model and the loglinear models account for dependencies across lists. They do not, however, account for heterogeneity

⁵¹In loglinear modeling, estimation of the standard errors assumes no missing data, no clustering of reports, and no matching error. However, the relative confidence interval lengths from alternative loglinear model estimates are expected to be robust to the small disturbances caused by these data blemishes. Nonetheless, the confidence intervals themselves, as calculated under the model, are too short. In our view, this limitation is not sufficient to be misleading.

⁵²The word "nominal" is used here because the confidence coefficient should be corrected when multiple comparisons are being made. Bonferroni adjustments, albeit generally conservative, would be one approach. Furthermore, we often visually and verbally compared two estimates or two series of estimates without remarking about unmeasured covariances which may exist.

of capture probabilities. One simple method for modeling this heterogeneity is the stratification of the target population by demographic characteristics (see Hogan, 1993). Another, model-based approach is the use of quasi-symmetry models (see Cressie and Holland, 1983, Fienberg and Meyer, 1983, Holland, 1990, Darroch et al., 1993, and Fienberg, Johnson, and Junker, 1999). A detailed technical explanation of these models for triple systems estimation can be found in Asher and Fienberg (2001). For the purposes of this document, it is sufficient to state that for triple systems estimation the partial quasi-symmetry models produce identical results to the six parameter loglinear models described above. As such, several quasi-symmetric models are used in the modeling procedure outlined in Section 3 of this appendix. Full quasi-symmetry models are not explored in this study, as capture heterogeneity is not believed to be identical across lists.

2.3. Multiple systems estimation

The hierarchical loglinear approach extends naturally to allow for the modeling of more intricate dependencies among 4 lists. If n is the sum of all records observed in all 4 lists combined, then:

$$\hat{N} = n + \frac{\hat{m}_{odd}}{\hat{m}_{even}}, \quad (14)$$

where \hat{m}_{odd} is a product of estimated expected cell values over all cells whose subscripts sum to an odd value, and \hat{m}_{even} is a product of estimated expected cell values over all cells whose subscripts sum to an even value. Formula (14) is just the generalization of formula (13), and as such is still the maximum likelihood estimate of the population total (see Fienberg, 1972).

For the 4-way multiple systems estimation models fitted in this document, interval estimates for N are computed using the profile likelihood methods of Cormack (1992) via a program developed by Matthew Johnson of the Educational Testing Service. The profile likelihood estimate of the $1 - \alpha$ confidence set for N is defined to be

$$\{N : G^2(N - n) - G^2(\hat{N} - n) < \chi_{(1),1-\alpha}^2\}, \quad (15)$$

where G^2 is the model deviance, and $\chi_{(1),1-\alpha}^2$ is the $1 - \alpha$ quantile of a $\chi_{(1)}^2$ distribution. Because Splus's `glm()` function estimates the multinomial capture-recapture model using a Poisson likelihood, we must approximate the multinomial deviance, \hat{G}^2 , from the Poisson fit. We use an approximation suggested by Cormack (1992):

$$\hat{G}^2(z) = D(z) - \log \left\{ \frac{z^z (n+z)!}{(n+z)^{n+z} z!} \right\}, \quad (16)$$

where $D(z)$ is the model deviance for a loglinear Poisson model fit to the 2^J contingency table with z in the missing cell.

2.4. Model selection

With the exception of the Marks, Seltzer, and Krótki model, the fit of all models described in this section is typically assessed using one of the following two statistics:

$$X^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}, \text{ or}$$

$$G^2 = 2 \sum (\text{Observed}) \log\left(\frac{\text{Observed}}{\text{Expected}}\right).$$

Both of these statistics, the Pearson chi-square (X^2) and the deviance (G^2), have approximate χ^2 distributions on q degrees of freedom, where q is the number of cells in the cross-classification table minus the number of parameters fitted in the model.⁵³ Both statistics produce similar results; the Pearson chi-square statistic, however, is better known. Therefore, within this document, the Pearson chi-square will be used to assess the fit of the models attempted, and the deviance will be used in the development of some confidence intervals via profile likelihood methods.

In order to assess the fit of a loglinear model using the Pearson chi-square statistic, a balance needs to be struck between neither underfitting nor overfitting the data. This is done by only accepting models whose Pearson chi-square statistic, when compared to a χ^2 distribution of the appropriate number of degrees of freedom, yields a p-value within a set range. A standard lower cutoff for the p-value for the Pearson chi-square statistic is 0.05; models with p-values below this do not fit the data well and are abandoned. An appropriate upper cutoff, required to prevent overfitting, must also be determined within the context of the models available.

3. Methodology

The goals for the statistical analyses undertaken during this study are as follows:

1. Development of a global estimate of the number of killings.
2. Estimation of the number of killings within each of four regions for every 2-day time period between 20–21 March and 22–23 June.
3. Analysis of the relationship between number of killings for every two-day time period and KLA/NATO activity.

Developing the models and analyses required to fulfill these goals is complicated. The remainder of this section outlines, in order, the methodological steps followed to complete the creation of the estimates required to meet goals 1 and 2. Section 4 will address goal 3.

3.1. Exploratory data analysis

Exploratory data analysis in the context of multiple systems estimation takes two forms. The first is an exploration of possible list dependence and heterogeneity through direct analysis of characteristics of the lists. For example, by analyzing patterns of data collection over list, time, and space, we can hope

⁵³In the case of multiple systems estimation, the number of cells in the cross-classification table is 2^{J-1} where J is the number of lists; the x_{000} cell is considered to be a “structural zero” and therefore is not included in the calculation of degrees of freedom.

Figure 2: Percentage of documented killings, by data source and municipality

Region	ABA	EXH	HRW	OSCE
1	24.6	32.9	24.6	22.5
2	33.4	22.1	3.5	45.6
3	11.8	15.1	28.2	11.9
4	30.3	30.1	43.5	20.2

to gain some insight into the complexity of the modeling procedure required for estimation. The second is the comparison of several low-level saturated model multiple systems estimation results. We must be careful to clarify that these models, due to the fact that they are saturated, are not candidates for our estimation procedure. At this point, we are not interested in the value of the estimates produced during our analysis; we only wish to note the relationship of these estimates to each other. In this way the models fitted during this procedure are explanatory only and not confirmatory. Once this exploratory data analysis is complete, we will begin our estimation procedure.

Direct list analysis

In order to understand spatial and temporal heterogeneity for the four lists, we will analyze data collection patterns. Figure (2) presents patterns of data collection for each of the four lists over region.⁵⁴ The percentages represent the proportion of documented killings for a given list within each municipality or region. Note that the lists have distinctly different patterns of data collection; for example, HRW covers proportionately less of region 2 and proportionately more of region 4 than the other lists. This indicates that there is heterogeneity of the lists that may be addressed by stratifying by region.

A similar analysis can be performed in order to determine the patterns of data collection by list over time. Figure (3) presents patterns of data collection for each of the four lists over 2-day time periods, where the percentages represent the proportion of documented killings for a given list within each time period. The periods in the table represent breaks in time that are of interest to the main body of this study. Again, HRW appears to follow a different data collection distribution over time than the other lists, indicating heterogeneity that may be addressed by stratifying over time.

Exploratory dual and triple systems estimation

Dependence and heterogeneity of the lists can also be explored directly through the statistical machinery of multiple systems estimation. The $\binom{4}{2}$ pairs of lists can be used to form six dual systems estimates of the global number of killings by observing how closely they match each other. These six estimates are listed in Figure (4).

⁵⁴The structure of the data is important here, not its content. Therefore, the regions are referred to by number. The northern region is region 1; the eastern region is region 2; the southern region is 3; and the western region is 4.

Figure 3: Percentage of documented killings, by data source and 2-day time period

Time Period	ABA	EXH	HRW	OSCE
20 – 23 March	2.1	1.8	1.6	2.9
24 March – 5 April	59.4	57.8	67.1	54.5
7 – 23 April	21.0	23.0	11.1	27.7
25 April – 9 May	11.0	9.2	10.6	10.3
11 May – 18 June	6.6	8.5	9.4	4.4

Figure 4: Dual system estimates

	EXH	HRW	OSCE
ABA	7 245	9 689	5 970
EXH		6 777	7 135
HRW			5 461

Note that all but one of these estimates falls at or below $n = 4\,400$, the total number of killings in the overall 4-way cross-classification table. This suggests a great deal of positive dependence between the lists, forcing the number of killings that are recorded in both lists (in the x_{11} cell) higher, and therefore the overall estimate lower. The exception is the dual systems estimate produced using the two lists ABA and HRW. Additionally, there appears to be some variability between the estimates (they range from 5 461 to 9 689), suggesting heterogeneity of the underlying capture probabilities.

Looking at the $\binom{4}{3}$ saturated triple systems estimates may yield greater insight into the higher level dependencies between the lists. Figure 5 lists these estimates.

These results are somewhat more promising; the positive list dependencies evident in the dual systems estimation chart have disappeared, suggesting that the list dependencies are modeled well by two-way interaction terms. Note, however, that the ABA, EXH, and OSCE estimate is overly large compared to the rest of the estimates, suggesting some higher order negative dependencies

Figure 5: Triple and 4-way system estimates (saturated)

Lists	\hat{N}
ABA, EXH, HRW	11 818
ABA, EXH, OSCE	22 331
ABA, HRW, OSCE	12 252
EXH, HRW, OSCE	8 014
ABA, EXH, HRW, OSCE	12 565

for this set of systems, while the EXH, HRW, and OSCE estimate appears a little low, suggesting higher order positive dependencies among these lists.

Both the direct analysis of the patterns of data collection and the exploration of the patterns within the dual and triple system estimates suggest that there is a great deal of dependence and heterogeneity between the lists. It is therefore appropriate to explore complicated 4-way multiple systems estimation for the full cross-classification table presented in Figure 4 of Appendix 1. Section 3.2 describes the selection procedure used to determine the top level model for the global⁵⁵ count of killings.

3.2. Fitting and selection of a model for the total number of killings

There are 113 possible hierarchical loglinear models for the four-way cross classification table presented in Figure 4 of Appendix 1, but only nine of these yield a Pearson chi-square statistic with a p-value greater than 0.05. For completeness, these nine models are presented in Figure (6). The notation used to represent the models is as follows: interaction terms are presented as lists multiplied together; e.g., ABA*EXH*HRW represents a three-way interaction term of these lists. Because the models presented are hierarchical, each interaction term presented also represents its lower order terms; e.g., ABA*EXH*HRW represents the set of terms {ABA, EXH, HRW, ABA*EXH, ABA*HRW, EXH*HRW, and ABA*EXH*HRW}.

Five of the models (1-4 and 7) have p-values above .30, while the remaining four have p-values between .06 and .08. Choosing an upper cutoff for the p-value of .3, for the purpose of avoiding overfitting, seems logical given this fact. The task then remains to pick the best of the four remaining models (5, 6, 8, and 9). It is tempting to simply pick the model with the lowest Pearson chi-square statistic. Doing so, however, ignores another good measure of a model - parsimony, measured by the minimization of the number of parameters, or conversely the maximization of the degrees of freedom. Minimizing the Pearson X^2 will tend toward the tightest fitting, and therefore most complicated, models. As a compromise between the desire to pick the model that fits best and the desire for the simplest model possible, we choose as our "best" model the model with the minimal adjusted Pearson chi-square statistic, X^2/d , where d is the degrees of freedom. Using this statistic, the model chosen is (9), leading to a global estimate of 10 356 killings, with a 95% confidence interval of (9 002, 12 122). This model produces a conservative estimate, in that only one other model in Figure 6 produces a lower \hat{N} .

3.3. Aggregation of the cross-classification tables to account for sparseness

The next goal for this analysis is the estimation of the number of killings for each of 192 space/time points representing 48 two-day time periods and four geographical regions. For this to occur, the cells in Figure 4 of Appendix 1 must be disaggregated into 192 cross-classification tables. Attempting to perform this disaggregation, clearly proves to be problematic. Column 4 of Figure 7 lists the frequency of 2-day cross-classification tables with 0-15 zero cells. Note, as a rule of thumb, that the maximum number of zero cells that still allows for meaningful loglinear modeling for a four-way multiple systems estimate is 10;

⁵⁵The term "global" will be used for the remainder of this Appendix to refer to the total number of killings within Kosovo for the March-June 1999 time period.

Figure 6: Results for models of global count of killings

	Model	\hat{N}	Fit Statistics			Profile Likelihood	
			X^2	d	Pr.	Dev.	95% C.I.
1	ABA*EXH*HRW+ABA*EXH*OSCE+EXH*HRW*OSCE	13 760	0.6	1	0.434	0.603	(9 695, 20 752)
2	ABA*EXH*HRW+ABA*HRW*OSCE+EXH*HRW*OSCE	22 923	0.8	1	0.386	0.755	(18 122, 29 394)
3	ABA*EXH*OSCE+ABA*HRW*OSCE+EXH*HRW*OSCE	13 467	0.9	1	0.337	0.917	(9 030, 21 419)
4	ABA*EXH*OSCE+EXH*HRW*OSCE+ABA*HRW	12 845	1	2	0.603	1.014	(9 700, 17 979)
5	ABA*HRW*OSCE+EXH*HRW*OSCE+ABA*EXH	20 734	4.9	2	0.085	4.964	(16 813, 25 889)
6	ABA*EXH*HRW+EXH*HRW*OSCE+ABA*OSCE	20 550	5.4	2	0.068	5.269	(16 708, 25 585)
7	ABA*EXH*OSCE+EXH*HRW*OSCE	12 741	1.0	3	0.796	1.021	(10 202, 16 742)
8	ABA*OSCE*EXH+HRW*EXH+HRW*ABA+HRW*OSCE	9 824	7.2	3	0.065	7.063	(8 449, 11 632)
9	ABA*OSCE*EXH+HRW*OSCE+HRW*EXH	10 356	8.9	4	0.063	9.333	(9 002, 12 122)

X^2 = Pearson chi-square statistic, d = degrees of freedom, Pr. = p-value,
Dev. = Residual Deviance, C.I. = confidence interval.

Figure 7: Counts of zero cells for the 4-way tables

Count of Zero Cells	by Six-Day Period	by Four-Day Period	by Two-Day Period
0	3	1	0
1	0	2	2
2	5	5	5
3	1	2	3
4	2	2	2
5	5	5	3
6	7	4	4
7	4	7	11
8	4	9	11
9	4	7	16
10	3	4	13
11	3	5	11
12	2	5	11
13	10	10	14
14	5	14	35
15	6	14	51

this allows one non-zero count for each of the parameters of an independence model. Of the 192 2-day tables, 122 (64%) contain more than 10 zero cells. Additionally, the sparseness of the tables that allow multiple systems estimation but contain a large number of zeros could lead to distorted estimation.

Collapsing to the 24 four-day periods over the four regions yields 96 cross-classification tables; of these, 48 (50%) contain more than 10 zeros. Collapsing to the 16 six-day periods over the four regions yields 64 cross-classification tables; of these, 26 (41%) contain more than 10 zeros. Collapsing further will impede the analysis desired.

Another option to collapsing across time points is given by collapsing across lists. There are $\binom{4}{3}$ possible 3-way cross classification tables for each four-way cross-classification table. This yields three-way cross classification tables, each representing a 2-day interval, region, and “system” of lists. In this case, more than three zeros will impede loglinear modeling for triple system estimation. In Figure 8, the systems range between 117 and 142 tables with more than three zeros, yielding at least 70% of the total tables between the four systems as too sparse for triple systems estimation. This percentage is a little misleading, in that each space and time point is represented by four cross-classification tables (one for each list). The actual coverage of space/time points by these 768 tables may be significantly higher than 30%.

It appears that triple systems estimation at the 2-day by region level may be difficult, and a combination of reducing the cross-classification tables (from four-way to three-way) and collapsing across 2-day periods is necessary. Figures 9 and 10 give the zero counts for the cross-classification tables for four-day and six-day periods, respectively.

Collapsing to six-day periods within three-way cross-classification tables appears to be an acceptable solution to the sparseness of the data. Although there are still a large number of sparse cross-classification tables at this level of ag-

Figure 8: Counts of zero cells for 3-way tables (two day period)

Count of Zero Cells	ABA, EXH and HRW	ABA, EXH, and OSCE	ABA, HRW and OSCE	EXH, HRW and OCSE
0	10	29	7	12
1	4	14	9	6
2	16	15	13	13
3	20	17	22	16
4	25	17	28	27
5	26	14	20	17
6	35	34	39	33
7	56	52	54	64

Figure 9: Counts of zero cells for 3-way tables (four day period)

Count of Zero Cells	ABA, EXH and HRW	ABA, EXH, and OSCE	ABA, HRW and OSCE	EXH, HRW and OCSE
0	10	28	9	10
1	3	7	5	8
2	14	6	13	8
3	15	11	14	16
4	10	6	13	6
5	13	9	9	10
6	16	15	17	15
7	15	14	16	23

Figure 10: Counts of zero cells for 3-way tables (six day period)

Count of Zero Cells	ABA, EXH and HRW	ABA, EXH, and OSCE	ABA, HRW and OSCE	EXH, HRW and OCSE
0	10	23	9	11
1	3	7	6	8
2	13	4	10	7
3	10	7	11	10
4	4	2	5	2
5	11	9	7	6
6	6	6	9	11
7	7	6	7	9

gregation, the redundancy of the four three-way systems allows for most of the six-day time and region points to be estimable. The inestimable time and space points tend to occur later in the 96 day time range, where there is less interest in understanding the trends in the data. Additionally, collapsing the 2-day tables across region within three-way cross-classification tables will allow modeling at a finer level of time, thereby allowing a better understanding of the general temporal trends in the data. Two sets of estimates will therefore be created in Sections 3.4 and 3.5 of this appendix, then compared in Section 3.6.

3.4. Global model fitting across all temporal and spatial points

This analysis commenced with the production a global estimate of the number of killings, estimated via a four-way multiple systems estimation model fit to Figure 4 of Appendix 1. One option for modeling the counts at individual space/time points is to build a generalized linear model, in which parameters representing each unique space/time point, as well as the parameters associated with the chosen global multiple systems estimation model 9 from Figure 6, are estimated. This overall model will project the global model down to the disaggregated tables, allowing for an complete modeling procedure for the entire system of estimates.

To adjust for sparseness, the data are collapsed to six-day time points, and only the first 10 of these points are included.⁵⁶ The result is a 71 parameter model fit to the counts of killings, with a 70 column matrix of indicator variables serving as predictors. Although the results suggest interesting trends, the p-value for the Pearson X^2 statistic is insignificant. This lack of fit can be explained as follows: although the global model describes the disaggregated table well, it doesn't describe each space/time point well. The overall generalized linear model allows for one model of list interactions to describe the relationships between the lists for each space/time point; heterogeneity of space/time points causes this model to fail at a local level.

A solution is to allow the overall generalized linear model to contain list parameters for each space/time region; starting with a fully saturated matrix of 14 list parameters * 40 space/time points + 40 space/time indicators, a stepwise procedure can be applied to select the 40 models that best describe the 40 space/time points. This is equivalent to running 40 separate generalized linear models, except the fit statistic for the overall model measures the fit of the entire system, while the fit statistics for the 40 local models measure each fit individually. Clearly it is simpler to just run separate loglinear models for each space/time point. This piecewise multiple systems estimation procedure will be discussed in more detail in Section 3.5.

3.5. Piecewise modeling across temporal and spatial points

Due to the sparseness of the data, implementation of four-way multiple systems estimation models, even with data aggregated to six-day periods, will yield very few acceptable models in terms of fit. Collapsing to dual system estimates yields models for which there are no measures of fit and for which we know the assumptions of independence and homogeneity do not hold. The remaining solution - triple system estimation - contains its own complexities. There are

⁵⁶If all 16 6-day points are included in the modeling, the resulting estimates are "flat" (identical) for the last six time points.

$\binom{4}{3} = 4$ possible three-way systems, and within each system there are 8 possible models. The result is up to 32 models for each of 64 time and space points.

A choice rule must be developed by which a “best” model can be picked. The following sets of rules - one for two-day table models, and one for six-day by region table models - mirror the model selection procedure for the overall model, with one important difference: the selection of an upper cutoff for the p-value of the Pearson chi-square statistic. Moving the upper cutoff too high will result in models that overfit, but moving the cutoff lower eventually removes all the models for a particular time and/or spatial point. The upper cutoff for the p-values for each of the two sets of models is therefore chosen to be as small as possible while maximizing the number of space/time points that are estimated. For the six-day by region models, the p-value is 0.7, and for the two-day models, the p-value is 0.5.

The model choice rules for the two-day estimates are as follows:

- Remove all models for which $\hat{N} < x_{++++}$ (1132 out of 1408 models retained).
- Remove all models for which $\hat{N} > 10\ 356$ (974 out of 1132 models retained).⁵⁷
- Remove all models for which $p < .05$ (657 out of 974 models retained).
- Remove all models for which $p > .5$ (247 out of 657 models retained).
- Choose the model with the lowest adjusted Pearson chi-square statistic.
- If no such model exists, then $\hat{N} = x_{++++}$.

The model choice rules for the six-day by region estimates are as follows:

- Remove all models for which $\hat{N} < x_{++++}$ (1455 out of 1856 models retained).
- Remove all models for which $\hat{N} > 10\ 356$ (1 235 out of 1 455 models retained).
- Remove all models for which $p < .05$ (876 out of 1235 models retained).
- Remove all models for which $p > .7$ (381 out of 856 models retained).
- Choose the model with the lowest adjusted Pearson chi-square statistic.
- If no such model exists, then $\hat{N} = x_{++++}$.

3.6. Projection of 2-day time point series to 6-day time point series for each region

The goals outlined at the beginning of this section included the creation of estimates for each 2-day time period between 20-21 March and 22-23 June within each region. To this point, estimates for six-day periods within region have been created, and estimates for two-day periods aggregated over region have been created as well. A series of two day estimates for each region can be created from these two separate sets of estimates as follows:

⁵⁷ 10 356 is the global estimate for the number of killings; it is illogical to believe that any estimate of a single space-time point will be greater than this number.

- Each six-day estimate at the regional level maps to three two-day estimates at the global level. Let $t \in (1, 16)$ designate a six-day interval, and $t_1, t_2,$ and t_3 designate the two-day intervals associated with t . Let \widehat{N}_{tr} designate the estimate for six-day interval t and region r . Finally, let $\widehat{N}_{t_j r}^*$ designate the estimate for the two-day interval t_j and region r .
- For the six-day estimate \widehat{N}_{tr} , create a proportion for each of the three two-day estimates as follows:

$$\widehat{p}_{t_i r} = \frac{\widehat{N}_{t_i r}^*}{\sum_{j=1}^3 \widehat{N}_{t_j r}^*}. \quad (17)$$

- Form a two-day estimate for time t_i , region r as follows:

$$\widetilde{N}_{t_i r} = \widehat{N}_{tr} \widehat{p}_{t_i r}. \quad (18)$$

The resulting two-day estimates for region r represent a blend of information about the regional trend and the global trend in the data.

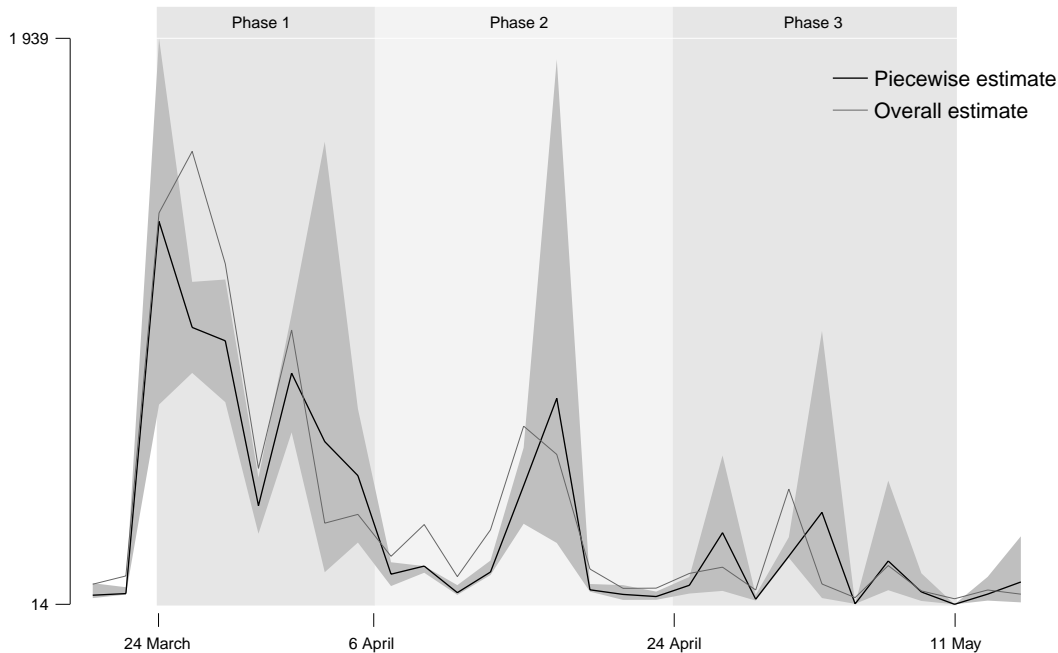
3.7. Comparison of results of global and piecewise modeling

In Figures 11 and 12, the piecewise and overall model estimates of killing counts, along with the confidence intervals for the piecewise estimates, are plotted together. There are two reassuring characteristics of these plots. The first is that the estimates derived from the overall model are quite similar to the estimates derived from the piecewise models; both suggest the same temporal trends in the form of waves of killings. The second is that the shape of the confidence bands, formed around the piecewise model estimates, maintain the shape of the curves formed by the estimates; taking any point in the confidence intervals as the true count of killings will not remove these trends.

Another reassuring characteristic of the results of the multiple forms of modeling procedures is how well the estimates track each other. Figure 13 shows several versions of aggregated estimates from the four different modeling procedures performed. The overall estimates of the total number of killings all fall within the confidence interval of the global model's estimate; some aggregated estimates fall quite close to 10 356. General trends across time appear similar for both the six-day period estimates aggregated across region and also the two-day period estimates aggregated across time. The system of models appears to work well.

As a final check of the quality of the modeling procedure developed for this study, the estimates produced within each modeling system are compared to the estimates produced by taking the average of the four triple-system Marks, Seltzer, and Krótki estimates given by equation (11). A count of the number of times the Marks, Seltzer, and Krótki estimate falls within the confidence interval of the piecewise estimate is given in Figure (14). The agreement between the two estimates appears to be better at lower levels of aggregation across space and time. This is not surprising; while the Marks, Seltzer, and Krótki model does account for some dependence between lists, it is not as flexible a tool as the hierarchical loglinear models. As such, it will perform better where heterogeneity is mild or not present.

Figure 11: Estimated killings over time, with nominal confidence interval



3.8. Analysis of relationship between original lists, complexity of models selected by the selection rule, and time and space

To this point, the focus of this appendix has been on the development of estimates of counts of killings. An interesting side analysis of the relationships between the lists is possible, however, because of the nature of the model selection procedure for the piecewise estimates. By performing this analysis and comparing it to the patterns given in Section 3.1, we can assess how well the piecewise modeling procedure adjusts to patterns of spatial and temporal heterogeneity.

Figure 15 lists the source systems for each of the six day by region and two-day estimates. The first column of counts displays the number of estimates derived from each of the three systems of lists; the second column of counts displays the number of estimates to which that list contributed. Although each list appears to contribute to a roughly equivalent number of estimates, the ABA, EXH, OSCE system appears to yield the most estimates overall. This makes sense given the structure of the underlying lists. HRW employed a different data collection strategy, leading to a different across both time and space, adding heterogeneity to the list systems containing it. Also, Human Rights Watch relied on an investigative data collection strategy which creates a different set of individual capture probabilities than the enumerative strategy employed by the other organizations.⁵⁸

⁵⁸For more information on the effects of enumerative versus investigative data collection strategies, see Asher and Ball (2001).

Figure 12: Estimated killings with nominal confidence interval, by region over time (6-day periods plotted to the middle day of the period)

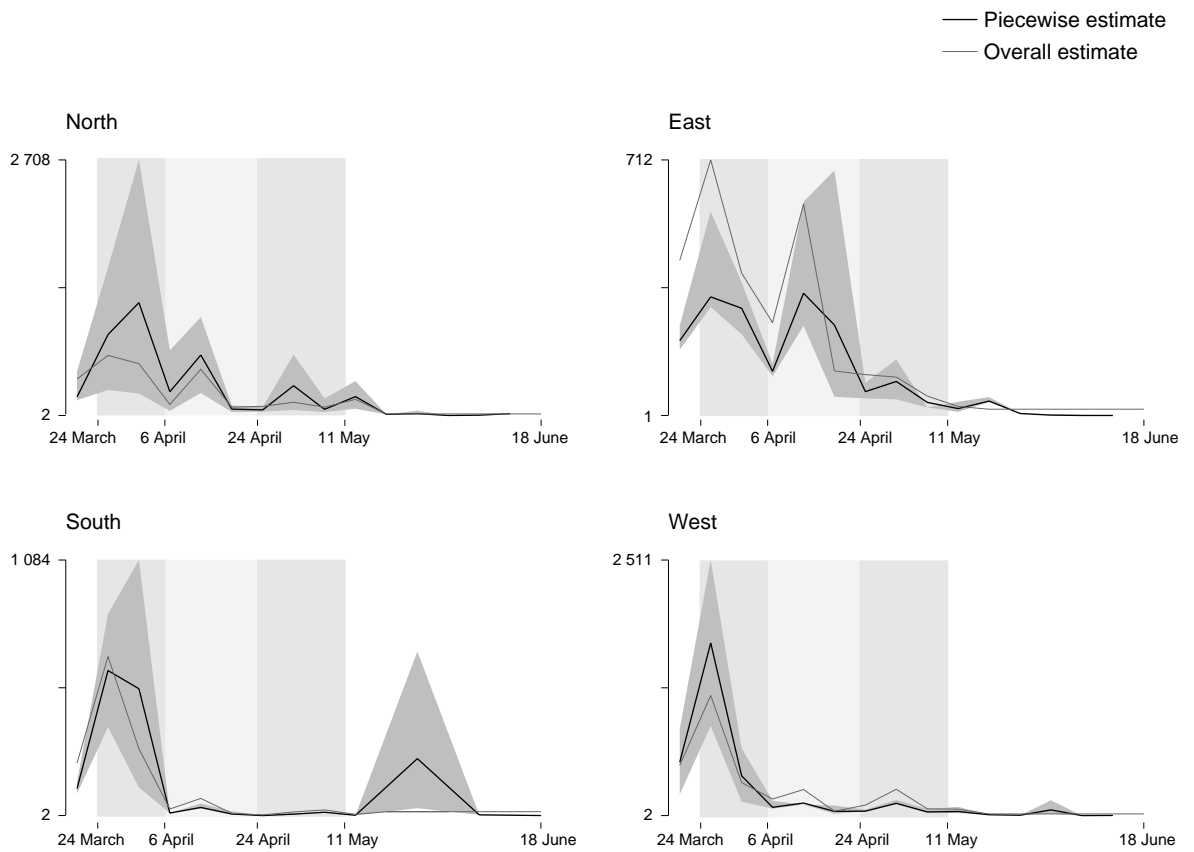


Figure 13: Comparison of estimates from different modeling procedures

Area	Piecewise Models			Overall Model
	Six Day Period Within Region	Two Day Period	Global	Direct GLM
Global	10 548	9 375	10 356	10 004
Region 1	3 925			2 748
Region 2	1 827			2 863
Region 3	1 608			1 393
Region 4	3 188			3 000
20 March - 25 March	1 048	1 372		1 538
26 March - 31 March	3 502	2 322		3 203
1 April - 6 April	2 426	1 735		1 557
7 April - 12 April	472	280		571
13 April - 18 April	1 144	1 312		1 411
19 April - 24 April	373	227		271
25 April - 30 April	175	246		322
1 May - 6 May	542	479		526
7 May - 12 May	157	216		238
13 May - 18 May	266	538		178
19 May - 24 May	64	124		61
25 May - 30 May	275	357		62
31 May - 5 June	62	128		23
6 June - 11 June	13	13		13
12 June - 17 June	25	25		25
18 June - 23 June	2	2		2

Figure 14: Comparison of Marks, Seltzer, and Krótki estimates to estimates from different modeling procedures

Status of Marks, Seltzer, and Krótki Estimates	Piecewise Models		
	Six Day Period Within Region	Two Day Period	Global
Below 95% C.I.	6	3	1
Within 95% C.I.	30	22	
Above 95% C.I.	10	10	
Missing/Inestimable	12	9	

Figure 15: Relationship between model selection criteria and lists

System 6-Day X Region	Models Selected	List	Models Selected
ABA,EXH,HRW	10	ABA	33
ABA,EXH,OSCE	14	EXH	37
ABA,HRW,OSCE	9	HRW	32
EXH,HRW,OSCE	13	OSCE	36
Total	46	Total out of 46	

System 2-Day	Models Selected	List	Models Selected
ABA,EXH,HRW	7	ABA	28
ABA,EXH,OSCE	13	EXH	27
ABA,HRW,OSCE	7	HRW	22
EXH,HRW,OSCE	8	OSCE	28
Total	35	Total out of 35	

Figure 16: Relationship between model selection criteria and lists by region

System (6-Day X Region)	Regions				Total	List	Regions				Total
	1	2	3	4			1	2	3	4	
ABA,EXH,HRW	2	2	3	3	10	ABA	8	10	8	7	33
ABA,EXH,OSCE	4	5	3	2	14	EXH	10	9	7	11	37
ABA,HRW,OSCE	2	3	2	2	9	HRW	8	7	6	11	32
EXH,HRW,OSCE	4	2	1	6	13	OSCE	10	10	6	10	36
Total	12	12	9	13	46	Total	36	36	27	39	138

Spatial and temporal dependencies of the list might be noticeable in the patterns of systems selected for the estimates within particular regions or time intervals. The following two tables, in order, show the counts of systems contributing to estimates by region and then time. Again, the counts of the number of estimates contributed to by each list is shown, as well as the number of estimates to which each list contributes. The 6 estimates to which EXH, HRW, OSCE contributes in Region 4 are especially interesting. Nearly half of this system's models are in Region 4, and half of the models in Region 4 are from this system. Also of interest is the pattern of the two HRW, OSCE systems over time; while the other systems roughly contribute more earlier in time and less over time, ABA, HRW, OSCE contribute more towards the middle of the time frame. These patterns reflect some of the spatial and temporal dependencies between the lists and further confirm the need for a flexible and complex modeling system.

A final measure of spatial and temporal dependencies between the lists is given in Figure 18. This set of counts by the number of parameters in the model indicates the complexity of the models for the estimates. While the 6-day by region estimates are evenly split between the most complicated model type and

Figure 17: Relationship between model selection criteria and lists by time point

System (2-Day)	3/20 to 4/6	4/7 to 4/24	4/25 to 5/12	5/13 to 5/30	5/31 to 6/17	Total
ABA,EXH,HRW	3	2	1	0	1	7
ABA,EXH,OSCE	5	2	3	3	0	13
ABA,HRW,OSCE	1	3	1	2	0	7
EXH,HRW,OSCE	1	2	3	2	0	8
Total	10	9	8	7	1	35

Figure 18: Complexity of models selected compared to aggregation level

Parameters in Model	6-Day X Region Models Selected	2-Day Models Selected
4	6	0
5	16	7
6	24	28
Total	46	35

the simpler model types, four-fifths of the 2-day estimates are derived from the most complicated model type.

The interpretation of this observation is relatively straightforward. The six-day by region estimates reflect a lower level of geography; within the smaller geographical units the relationships between the lists are less complicated than for the larger geographical units. In other words, by geographically “stratifying” the estimation areas, dependencies and heterogeneity are reduced. In the case of the 2-day estimates, however, list dependencies have not been “stratified” out, and the more complex models fit better.

4. Analysis of relationship between multiple systems estimation modeling results and KLA/NATO activity

To this point, all the statistical methodology and analyses described by this Appendix have directly related to the estimation of counts of killings. Only one additional statistical analysis is done using the estimates once they are created; this analysis will now be discussed.

In the main body of this study, the relationship between NATO air strikes on Kosovo, KLA activity within Kosovo, and patterns of killings and refugee migration is discussed. A statistical analysis technique by which these relationships can be understood better is simple linear regression using estimates of killings or migration flow within a particular spatial and temporal region as the dependent variable. Potential explanatory variables for the model include number of KLA battles within the spatial and temporal region, number of KLA killings within the spatial and temporal region, number of NATO air strikes within the spatial and temporal region, number of KLA battles within the previous spatial

and temporal region, number of KLA killings within the previous spatial and temporal region, and number of NATO air strikes within the previous spatial and temporal region. Dummy variables for regional effects can be used as independent variables as well, in order to control for the possibility of the activities in one region dominating the analysis.

Through a regression analysis using the variables described above, the association between the NATO and KLA activities and the Albanian migrations and killings can be assessed via the significance levels of the model and each individual parameter in the model. Figure (19) displays the results for four regression models; in the first and third the variables represent two-day periods for the whole country, in the second and fourth the variables represent six-day periods within regions. The significance level of the parameters is indicated by $\star = 0.05$, $\star\star = 0.01$, and $\star\star\star = 0.001$. For the regression models using killings as the dependent variable, the only significant parameters are for regional effects. In other words, the association between the KLA and NATO variables and the counts of killings is weak. This is confirmed by the low R^2 's for these two models of 0.253 and 0.147.

The regression models using migrations as the dependent variable, however, yield a different interpretation. In this case, there appears to be an association between KLA activity and migration; specifically, the association between KLA battles within the previous time period and migrations in the current time period appears to be significant. The R^2 's for both of these models are high, further confirming an association.

At this point, the regression results suggest that the pattern of NATO bombings in Kosovo over time is not significantly associated with the pattern of killings or migrations in Kosovo over time. The pattern of KLA activity, however, appears to be associated with the pattern of migration.

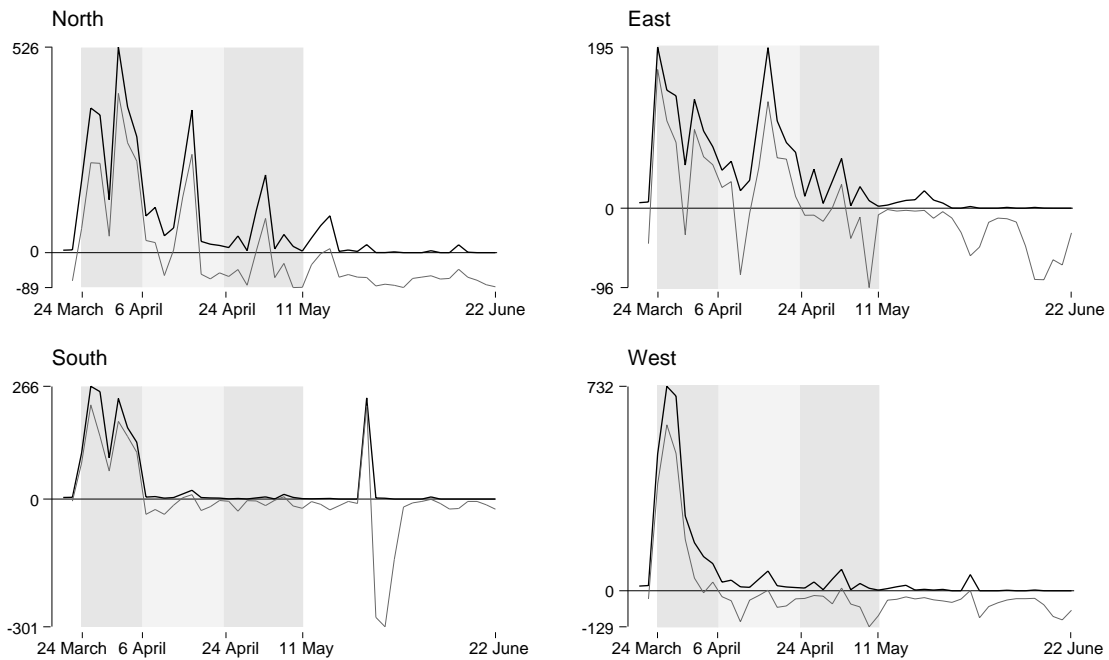
Further evidence of the lack of association of the KLA activity, NATO bombings, and killing patterns in Kosovo is given by a comparison of the residuals from each of the regressions described above to its dependent variable. If a regression model describes its dependent variable well, then the pattern of the residuals for that regression model will be random. If, however, the regression model does not describe its dependent variable well, then the residuals will follow the same pattern over time as the original dependent variable. Figures 20 and 21 display comparisons of the residuals for the models for which the dependent variables are killings and migrations within six-day period and region. In Figure 20, the residuals very closely track the estimated counts of killings closely, picking up clear trends in the data.

In Figure 21, the relationship between the residuals and migration flow is evident, but in but not as strong as the relationship between the residuals and killings for the previous model. This is not surprising given the better fit of the regression model for migration flows. However, the divergences between the series occur at only a few points. The residuals track the estimated values closely in the southern and western regions. In these regions, when the series diverge it is only because the residual is exaggerating a trend clearly present in the refugee flow series; this pattern is seen in the western region during the early part of Phase 2. In the northern and eastern regions, the series differ more strongly. But even in these regions, the mid-April peaks during Phase 2 are clearly similar in the refugee flow and residuals series. In the northern and

Figure 19: Regression coefficients

Explanatory Variables	Response Variables			
	Killings over Time	Killings over Time and Region	Refugee Flow over Time	Refugee Flow over Time and Region
Region 2		* -52.3 (20.4)		-721.4 (665.3)
Region 3		** -57.0 (21.3)		**3 017.8 (1 048.9)
Region 4		-34.4 (25.2)		-193.9 (862.5)
KLA (kill)	-1.1 (5.2)	1.6 (4.1)	*-634.9 (318.9)	-184.2 (130.8)
KLA (battle)	34.7 (32.1)	13.3 (12.2)	2 728.6 (1030.5)	**1 879.3 (583.2)
Lag-KLA (kill)	0.2 (4.4)	3.3 (3.4)	491.9 (384.4)	277.9 (167.1)
Lag-KLA (battle)	21.2 (17.7)	11.6 (11.9)	**2 794.3 (827.7)	***2 138.4 (633.7)
NATO	10.9 (11.1)	11.4 (6.7)	327.6 (390.3)	565.8 (379.7)
Lag-NATO	-4.8 (6.9)	-2.5 (4.1)	-28.0 (325.5)	29.2 (234.0)
Constant	83.9 (51.1)	**62.7 (19.9)	122.3 (3 933.2)	-484.8 (608.8)
R^2	0.3	0.1	0.7	0.5

Figure 20: Estimated total killings and residuals by region over time



eastern regions during Phase 1 and into the transition to Phase 2, the series contradict each other.

Our conclusions are as follows:

- Based on our analysis of these data, there is no evidence to support the theory that NATO bombings or KLA activity is associated with patterns of killings in Kosovo.
- There is some evidence that there is an association between KLA activity and migration patterns in the northern and eastern regions, especially during Phase 1.
- The association between KLA activity and migration flows does not fully explain the pattern of migration, especially in the western and southern regions.

5. Discussion

This appendix has presented the main modeling methods we employed. In this final section, a method of sensitivity analysis of the date of death reporting is explained, and summary conclusions for this appendix are given.

5.1. Sensitivity analysis of date of death reporting

As individual records were matched to other individuals and to groups, they accumulated dates. The choice of the "best" date is described in Appendix 1.

Figure 21: Estimated refugee flow and residuals by region over time

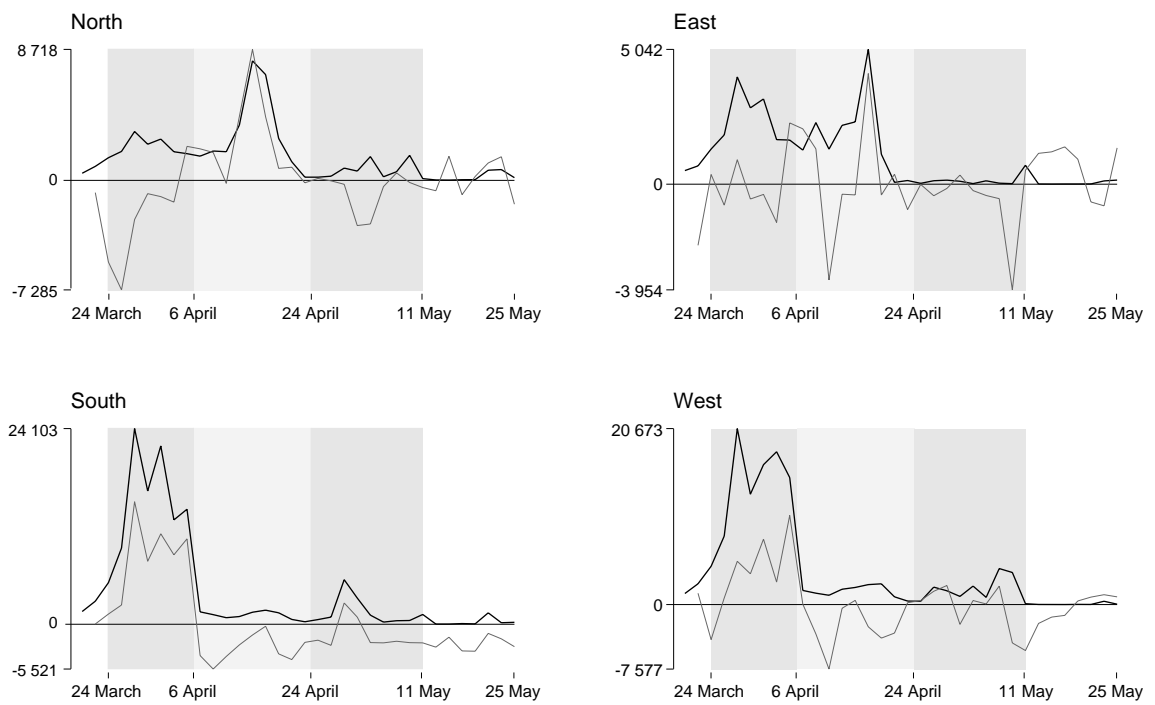
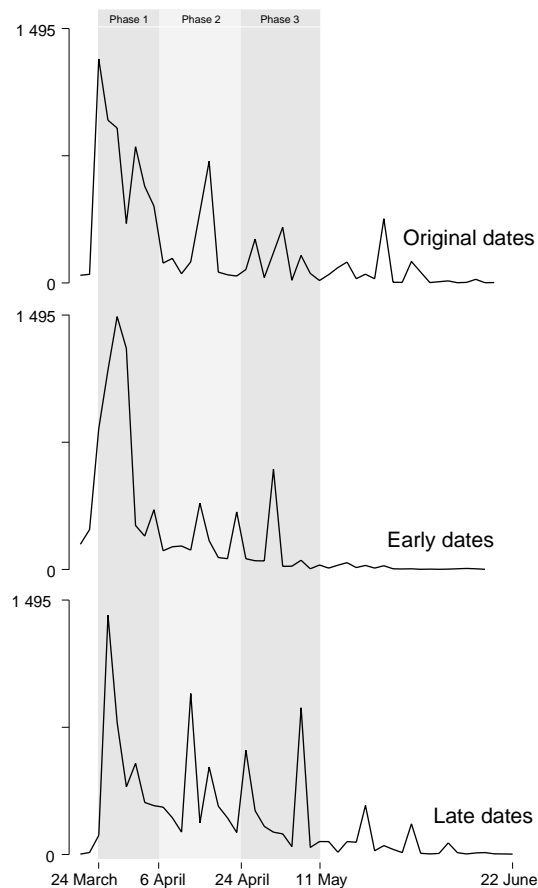


Figure 22: Estimated killings over time with alternative date assignments



Even with an appropriate choice of the best date, the “best” date might still occasionally be in error. Time is a central dimension of the hypotheses being tested. It is, therefore, useful to consider whether the substantive interpretation of the results is robust to different assumptions about the quality of the date information.

For the sensitivity analysis, plausible “early” and “late” dates were chosen as alternatives for each record. Dates were accumulated from all the group and individual records that matched each individual record, both in the self-matching, and the inter-system matching. Records that had 3 or more dates in their distribution took the dates at the 25th and 75th percentiles as the “early” and “late” dates. Records with 2 dates took those two dates as the early and late dates. The difference between the early and late dates defined a range.

Records with 1 or 0 dates were assigned a range by hotdecking; as before, the hotdecked records were matched by geographic location. The late and early dates for these records were plus and minus half the hotdecked range; the values were rounded up to the next integer. In this way, all records were perturbed by at least one day as we tried to maximize the impact of this test.

The resulting distribution of killings over time is shown in Figure 22. Perturbing the dates does affect how the curves are shaped. Shifting killings to earlier dates fills the late March and early April dates while taking quantity from the mid-April peak. Shifting killings to later dates moves them from the Phase 1 peaks to peaks in Phase 2 and Phase 3.

The most important finding from this analysis is that even this significant restructuring of how the dates are handled does not change the fundamental characteristics of the pattern over time. Both of the perturbed series have high peaks during the early or middle part of Phase 1. Both series show substantial declines during the 5-8 April phase transition, and each has a peak in the middle of Phase 2. The perturbed series disagree about exactly when the transition from Phase 2 to Phase 3 occurs: two days earlier (in the late and random series), or two days later in the early series.

If the peaks and troughs in the pattern of killing over time had been created by a particular date handling technique, then one or more of the perturbations would have shown a random pattern. If the reported dates had been widely dispersed, the ranges would have been wide enough that the perturbations would have smoothed over the “start and stop” pattern that characterized killings and refugee flow in Kosovo during March–June 1999. The observation that the smoothing did not occur is evidence that the estimation procedure is robust to imprecise reporting of the date of killings.

5.2. Summary of modeling conclusions

We began this appendix by posing the dilemma of how to estimate unobserved (or at least unreported) killings in order to estimate total deaths. The modeling presented in Appendix 2 convinces us that there were probably a little over 10 000 killings of Kosovar Albanians in the period 20 March to 22 June 1999. The largest direct estimate is comparable to this “best” model estimate, and different models produce similar estimates. We believe that we have made our case about the overall total number of killings and for the pattern of killings during the period in question. It is on this basis that we made the conclusions presented in the body of the report.

Appendix 3: Additional Sources on KLA and NATO Activity

Albanian Human Rights Group

Albanian Media

Belgrade Center for Human Rights

Center for Peace and Tolerance

Daily Telegraph

Danas

Egyptian National Community in Kosovo

European Community Monitor Mission

European Roma Rights Center

Federal Republic of Yugoslavia (FRY) Ministry of Defense

FRY Civil Defense Authorities

FRY Ministry of Foreign Affairs

FRY Ministry of Information

FRY, Aide-Memoire on the Use of Inhumane Weapons in the Aggression of the North Atlantic Treaty Organization Against the Federal Republic of Yugoslavia. Belgrade, 15 May 1999

Fund for the Humanitarian Right

The Guardian

Human Rights Board of Sandzak

Information Service of Church and National Assembly (Kosovo)

International Committee of the Red Cross

International Criminal Tribunal for the Former Yugoslavia

Koha Ditore

Kosovapress

Kosovar Media
Kosovo Verification Mission
Local Church Councils (Kosovo)
Los Angeles Times
NATO Operation Allied Force Update
Open Society Institute
Organization for Security and Cooperation in Europe
Organization of Families of Disappeared
Orthodoxy Press
Politika
Report by Bishop Artemije "List of Killed and Kidnapped Serbs." Republic
of Serbia Ministry of Internal Affairs
RTS TV Belgrade
Serbian Media
Serbian Orthodox Church
Serbian Unity Congress NewsBits
SVEDOK-Belgrade weekly
Tanjug
United Nations High Commission for Refugees
VI.P Daily News Report

References

- American Bar Association Central and East European Law Initiative and the American Association for the Advancement of Science. 2000. *Political Killings in Kosova/Kosovo, March-June 1999*. Washington, DC: American Bar Association Central and East European Law Initiative.
- Anderson, Margo and Stephen E. Fienberg. 2001a. *Who Counts? Census-Taking in Contemporary America*. Revised Paperback Edition. New York: Russell Sage Foundation.
- Anderson, Margo and Stephen E. Fienberg. 2001b. *Counting and estimation: Methodology for Improving the Quality of Censuses. The U.S. 2000 Census Adjustment Decision*. Paper presented at the International Conference on Quality in Official Statistics, Stockholm, Sweden, May 14-15, 2001.
- Asher, Jana and Patrick Ball. 2001. Understanding Human Rights Violation Data through the Analysis of Circuits. To appear in the *Proceedings of the American Statistical Association* (Social Statistics Section).
- Asher, Jana and Stephen E. Fienberg. 2001. Statistical Variations on an Administrative Records Census. To appear in the *Proceedings of the American Statistical Association* (Government Statistics Section).
- Ball, Patrick. 2000a. *Policy or Panic: The Flight of Ethnic Albanians from Kosovo, March-May 1999*. Washington D.C.: American Association for the Advancement of Science.
- Ball, Patrick. 2000b. The Guatemalan Commission for Historical Clarification: Intersample Analysis. Chapter 11 in *Making the Case: Investigating Large Scale Human Rights Violations using Information Systems and Data Analysis*, edited by Patrick Ball, Herbert Spierer, and Louise Spierer. Washington, DC: American Association for the Advancement of Science.
- Belin, Thomas R. and Donald B. Rubin. 1995. A Method for Calibrating False-Match Rates in Record Linkage. *Journal of the American Statistical Association*.
- Bishop, Yvonne M. M., Stephen E. Fienberg, and Paul H. Holland. 1975. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.
- Converse, N. and F. Scheuren. 2001. Workarounds in Survey Data Handling. Submitted to the new *Journal of Data*.
- Cressie, Noel and Paul W. Holland. 1983. Characterizing the Manifest Probabilities of Latent Trait Models. *Psychometrika* 48: 129-141.
- Cormack, R. 1992. Interval Estimates for Mark-Recapture Studies of Closed Populations. *Biometrics* 48: 567-576.

- Cowan, Charles Douglas. 1984. The effects of misclassifications on estimates from capture-recapture studies. Ph.D. diss., George Washington University.
- Darroch, John N., Stephen E. Fienberg, Gary Glonek, F.V. Gary, and Brian W. Junker. 1993. A Three-Sample Multiple-Recapture Approach to Census Population Estimation with Heterogeneous Catchability. *Journal of the American Statistical Association* 88: 1137–1148.
- Fienberg, Stephen E. 1972. The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika* 59: 591–603.
- Fienberg, Stephen E. 1980. *The Analysis of Cross-Classified Categorical Data*. Second Edition. Cambridge, MA: MIT Press.
- Fienberg, Stephen E., Matthew S. Johnson, and Brian W. Junker. 1999. Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society, Series A* 162: 383–405.
- Fienberg, Stephen E., and Michael M. Meyer. 1983. Loglinear models and categorical data analysis with psychometric and econometric applications. *Journal of Econometrics* 22: 191–214.
- Ford, B. 1983. Hot Deck Imputation. Ch. 14 in vol. 2, part 4 of *Incomplete Data in Sample Surveys*, edited by William G. Madow, Harold Nisselson, and Ingram Olkin. New York: Academic Press.
- Hogan, Howard. 1993. The 1990 Post-Enumeration Survey: Operations and Results. *Journal of the American Statistical Association* 88: 1047–1060.
- Holland, Paul W. 1990. On the sampling theory foundations of item response theory models. *Psychometrika* 55: 577–601.
- Human Rights Watch. 2001. *Under Orders: War Crimes in Kosovo*. New York: Human Rights Watch.
- Independent International Commission on Kosovo. 2000. *The Kosovo Report: Conflict*International Response*Lessons Learned*. New York: Oxford University Press.
- International Working Group for Disease Monitoring and Forecasting. 1995a. Capture-recapture and multiple-record systems estimation, I: History and theoretical development. *American Journal of Epidemiology* 141: 1047–1058.
- International Working Group for Disease Monitoring and Forecasting. 1995b. Capture-recapture and multiple-record systems estimation, II: Applications in human diseases. *American Journal of Epidemiology* 141: 1059–1088.
- Marks, E.S., W. Seltzer, and K. J. Krótki. 1974. *Population Growth Estimation: A Handbook of Vital Statistics Measurement*. New York: The Population Council.
- Oh, H. and F. Scheuren. 1983. Weighting Adjustment for Unit Nonresponse. Chap. 13 in vol. 2, part 4 of *Incomplete Data in Sample Surveys*, edited by William G. Madow, Harold Nisselson, and Ingram Olkin. New York: Academic Press.

- Organization for Security and Cooperation in Europe. 1999. *Kosovo/Kosova As Seen As Told: An Analysis of the Human Rights Findings of the OSCE Kosovo Verification Mission October 1998 to June 1999*. Warsaw, Poland: OSCE Office for Democratic Institutions and Human Rights.
- Peterson, C. G. J. 1896. The yearly immigration of young plaice into the Limfjord from the German Sea. *Report of the Danish Biological Station to the Ministry of Fisheries* 6: 1–48.
- Physicians for Human Rights. 1999. *War Crimes in Kosovo: A Population-Based Assessment of Human Rights Violations of Kosovar Albanians by Serb Forces*. Boston: Physicians for Human Rights.
- Record Linkage Techniques. 1985. *Record Linkage Techniques – 1985 Proceedings of the Workshop on Exact Matching Methodologies*. Washington, DC: U.S. Internal Revenue Service, Statistics of Income Division.
- Record Linkage Techniques. 1997. *Record Linkage Techniques – 1997 Proceedings of An International Workshop and Exposition*. Washington, DC: Ernst and Young, LLP
- Rubin, Donald B. 1987. *Multiple imputation for nonresponse in surveys*. New York: Wiley.
- Scheuren, F. 1985. Methodologic issues in linkage of multiple data bases. *Record Linkage Techniques – 1985 Proceedings of the Workshop on Exact Matching Methodologies*. Washington, DC: U.S. Internal Revenue Service, Statistics of Income Division.
- Sekar, C.C. and Deming, W.E. 1949. On a Method of Estimating Birth and Death Rates and the Extent of Registration. *Journal of the American Statistical Association*. 44:101-115.
- Spiegel, Paul B. and Peter Salama. 2000. War and Mortality in Kosovo, 1998–1999: An Epidemiological Testimony. *Lancet* 355: 2206–2211.
- Splus, Insightful Corporation. 2001. “Generalizing the Linear Model.” Ch. 12 in *S-PLUS 6 for Windows Guide to Statistics, Volume 1*. Seattle, WA: Insightful Corp.
- Stata Corporation. 2001. Section on generalized linear models in *Stata 7 Reference Manual*. Vol 1 A-G. College Station, TX: Stata Corporation.

Acknowledgments

This report would not exist without the collaboration of many individuals and organizations. It is based primarily on data provided by the American Bar Association Central and East European Law Initiative (ABA/CEELI), the American Association for the Advancement of Science (AAAS), The Center for Peace Through Justice, the Council for Defense of Human Rights and Freedoms, Human Rights Watch, the Organization for Security and Cooperation in Europe, and the International Criminal Tribunal for the Former Yugoslavia. We are very thankful for the willingness of these organizations to share the results of their work with each other. Without this cooperative spirit, this project would not have been possible.

In 1999, Scott Carlson, Director of Central and East European Programs at ABA/CEELI, brought human rights organizations together to share their information regarding Kosovo and to benefit from their past experience in documenting human rights abuses. ABA/CEELI and AAAS published a report entitled *Political Killings in Kosovo/Kosova, March-June 1999* in 2000.

In April 1999, Patrick Ball and Fritz Scheuren of AAAS, with Fron Nazi of the East-West Management Institute and the Institute for Policy and Legal Studies began a study of the statistical patterns of refugee flows out of Kosovo. This work was published as *Policy or Panic: The Flight of Ethnic Albanians from Kosovo, March-May 1999*. Organizations and individuals who contributed time, data, and other assistance to these earlier projects include Physicians for Human Rights, the the Human Rights Center and the Department of Demography of the University of California-Berkeley, Fred Abrahams, Vasian Cepa, Blerina Kashari, Julia Belanger, Andrea Lako, Eric Stover, Dr. Sandra Eyster, Ilir Gocaj, and many others.

In addition, a number of individuals generously contributed their time, energy, and expertise to this project. Matt Zimmerman wrote the software application used to perform the inter-system matching. He also designed the layout and provided crucial assistance with graphic design and technical editing. Patricia Hawkins provided early programming assistance. Sarah Churchill and Maya Goldstein oversaw the data coding of the ABA and OSCE data. Jason Sanders helped with coding and administrative issues. Jeff Henigson recoded the HRW data from the original interviews.

Support for the project was provided to ABA/CEELI from the Bureau of Democracy, Human Rights, and Labor at the U.S. Department of State and the U.S. Agency for International Development. While U.S. Government support was essential to the project, it should also be emphasized that at no time did U.S. Government personnel seek to infringe upon our independent management of the project or influence our substantive reporting. ABA/CEELI struc-

tured its relations with the U.S. Government as a “cooperative agreement” to ensure its independence in this respect. Consequently, this report was not submitted for U.S. Government review, and any convergence with the views of the U.S. Government is purely coincidental.

In addition to support via a subcontract with ABA/CEELI, AAAS received support from the Institute for Civil Society and the John D. and Catherine T. MacArthur Foundation.

Authors and Editors

Patrick Ball programmed the database, managed the quality control, and created statistical software, wrote portions of each section of the report, and provided overall direction. Wendy Betts wrote the body of the report and coordinated the coding of the ABA/CEELI and OSCE data. Fritz Scheuren provided statistical guidance and wrote Appendix 1. Jana Dudukovic managed the data clerks, oversaw the matching process, coded the KLA data, and contributed to Appendix 1. Jana Asher programmed statistical routines and wrote Appendix 2. Patrick Ball and Jana Asher developed the modeling procedures. All of the authors jointly edited the report.

Scholarly Review Panel

A number of people reviewed the report. An international review team was chaired by Dr. Helge Brunborg (Senior Research Fellow, Statistics Norway), and consisted of Dr. Ronald Lee (Professor of Economics and Demography, University of California-Berkeley); Dr. Françoise Seillier-Moiseiwitsch (Associate Professor of Statistics and Director of the Bioinformatics Research Center, University of Maryland-Baltimore County, and Chair, Human Rights Committee, American Statistical Association), Dr. Jean-Louis Bodin (Past President, International Statistics Institute); Dr. Carlo Malaguerra (Director General of the Swiss Federal Statistical Office–SFSO); Dr. Philippe Eichenberger (Head of the Department of Statistical Methods, SFSO); Dr. Beat Hulliger (Deputy Head of the Department of Statistical Methods, SFSO). The reviewers provided extensive comments on two preliminary drafts of the report.

A number of additional reviewers worked with us less formally. These included Dr. David Banks (U.S. Department of Transportation), Herbert F. Spierer (Adjunct Professor, Columbia University School of International and Public Affairs Human Rights Program), Louise Spierer (independent scholar), and Dr. Denise Albanese (Department of English, George Mason University).

This report is much stronger as a result of the frank critical assessments made by the reviewers, and we are very grateful to them. Of course, the authors bear sole responsibility for the analysis and opinions expressed in this study.

Authoring Organizations

AAAS Science and Human Rights Program

The Science and Human Rights Program of the American Association for the Advancement of Science (AAAS) seeks to protect the human rights of scientists and to bring the methods of science to human rights work. The Program develops and advances methods for human rights documentation and monitoring, fosters support for human rights among scientists, and conducts research on a variety of related issues. The Program's work is based on the premise that respect for human rights is essential to the conduct of science. For more information about the Program and its activities, visit <http://shr.aaas.org>.

ABA Central and East European Law Initiative

The Central and East European Law Initiative (CEELI) is a public service project of the American Bar Association (ABA). The project is designed to advance the rule of law by supporting the law reform process underway in Central and Eastern Europe and the New Independent States of the former Soviet Union (NIS). Through various programs, CEELI makes available the legal expertise of American and European volunteers to assist emerging democracies in modifying or restructuring laws and legal systems.

The ABA/CEELI War Crimes Documentation Project (WCDP) began in May 1999 with funding from the U.S. Agency for International Development and the U.S. State Department. The WCDP has two objectives: 1) to assist efforts to investigate war crimes and prosecute perpetrators, and 2) to increase public awareness of war crimes, their prosecution, and the role of the International Criminal Tribunal for the former Yugoslavia (ICTY) in the process. On war crimes issues, ABA/CEELI has worked closely with several other nongovernmental organizations, including the Coalition for International Justice (CIJ), Chicago-Kent College of Law, and The Center for Peace Through Justice. For more information about ABA/CEELI and its activities, visit <http://www.abanet.org/ceeli/>.

About the Authors

Patrick Ball, Ph.D., is Deputy Director of the AAAS Science and Human Rights Program. Since 1991, he has designed information management systems and conducted quantitative analysis for large-scale human rights data projects for truth commissions, non-governmental organizations, tribunals and United Nations missions in El Salvador, Ethiopia, Guatemala, Haiti, South Africa, Kosovo, and Sri Lanka.

Wendy Betts, M.A., is Co-Director of the ABA/CEELI War Crimes Documentation Project. She has contributed to a number of publications on international and internal conflict, and post-conflict transition.

Fritz Scheuren, Ph.D., is Vice President, Statistics, at the National Opinion Research Center, a research arm of the University of Chicago. He has extensive experience in record linkage both in survey and administrative settings. Currently, he is working full time on Native American issues.

Jana Dudukovic is an independent scholar studying under the tutelage of Louise Spierer.

Jana Asher, M.S., has wide research experience in small area estimation, administrative records, record linkage, and multiple systems estimation. She is currently pursuing a Ph.D. in statistics at Carnegie Mellon University under the guidance of Professor Stephen E. Fienberg.