Matching Migration Records: Record Linkage and Capture Recapture Analysis of Kosovo Border-Crossing Records in 1999

Stephanie Eaneff
Data Matching Methods and Their Uses

October 18, 2013

Abstract

Between March and May of 1999, hundreds of thousands of refugees fled from Kosovo, most of them ethnic Albanians. While it is unknown for certain what prompted these mass migrations across Kosovo's borders, there is no doubt that the massive border-crossings were heavily impacted by ongoing conflict between Yugoslav troops and the Kosovo Liberation Army, NATO air strikes, and reputed efforts towards ethnic cleansing of Albanians by the Serbian government.

Beginning in late March of 1999, the Albanian government began to collect records documenting the passage of hundreds of thousands of refugees across the border from Kosovo into Albania through a small border post near the village of Morina. One month later, the Institute for Policy and Legal Studies (IPLS) in Washington DC, in partnership with the American Association for the Advancement of Science (AAAS) conducted a series of interviews with Kosovan refugees in camps located throughout Albania and Bosnia, documenting the details and date of their passage across Kosovo's border.

Drawing upon these studies documenting the migration of individuals across the Kosovo border between March and May of 1999, record-linkage methods were used as a means of identifying groups of individuals who were documented as having crossed the Kosovo border at Morina and who were also surveyed as part of the later interview efforts in refugee camps of surrounding countries. Based on these linked records, capture-recapture methods were then used to generate an overall estimate of the number of individuals who crossed the Kosovo border between March and May of 1999, specifically, between March 28 and May 28.

The results of these analyses suggest that between 711,000 and 1,088,000 individuals crossed the Kosovo border between March and May of 1999, an estimate which is comparable to those reported by the United Nations High Commissioner for Refugees, the New York Times, and the Human Rights Data Analysis Group.[1, 10] Although the methods and applications discussed in this paper are not intended to provide absolute estimates, nor estimates associated with statistically interpretable confidence intervals, the computationally conservative methods discussed in the pages that follow can be applied to provide first-pass estimates of counts related to human migration or human-

rights violations based on record-linkage and capture-recapture methods which draw from multiple sources of records.

Introduction

Between March and May of 1999, a large number of individuals living in Kosovo crossed the country's surrounding borders to seek refuge in other countries, most commonly Albania, Macedonia, and Bosnia (although seeking asylum in Bosnia required these refugees to also cross through Montenegro).[1] Institutional efforts to document these mass migrations were incredibly challenging and time consuming, and it was impossible to capture records related to the total number of individuals who crossed the border at each possible location. However, an estimate of this final count of refugees would be incredibly valuable in helping to understand the social, political, and economic impacts of these migrations, both for historical purposes and also for use in investigating potential war crimes and human rights violations.

This paper considers two border-crossing records taken in early 1999 in order to generate an estimate of the total number of refugees who crossed the Kosovan borders between March and May of 1999. By combining record-linkage methods with capture-recapture analyses, this research draws from publicly-available refugee documentation without personally identifiable information in order to generate an estimate of the magnitude of refugee migrations out of Kosovo in early 1999.

Using a series of historically-informed thresholding criteria for matching records, these analyses use a relatively computationally conservative method for record-linkage of sparse observations related to human migration during periods of political and social conflict. These methods are both intuitive and could be replicated within hours, given the selection of proper thresholding criteria. Thus, while not intended to calculate absolute estimates of human migration, the methods and applications discussed in the pages that follow have the potential for use in providing rapid and first-past estimated counts of human migration using sampling from multiple sources of records.

Data

Given the significant influx of refugees who crossed the borders of Kosovo in early 1999, it is nearly impossible to generate complete records tracking each individual who crossed the border out of Kosovo. In part, this is due to a lack of the necessary institutional support to track and maintain these types of data. In order to document border crossings of each individual for over three months, a huge investment of physical infrastructure, personnel, and time would be required. Given the potential for border crossing at at number of specific border posts, as well as other geographic points along Kosovo's shared borders, generating a complete record of these border crossings would be nearly impossible, particularly because the migrations in question occurred at all hours of the day and night over a period of over three months.

However, time- and personnel-intensive efforts were still made to document the number and characteristics of the individuals who crossed the Kosovo border. These efforts represent the work of foreign governments, international research groups, and non-profit organizations, and the records collected by these groups, although far from comprehensive, represent the best-available data to consider the characteristics, timing, and magnitude of migrations across the Kosovo border in 1999.

Beginning in March of 1999, the Albanian government tasked a group of officials with collecting records documenting the passage of hundreds of thousands of individuals across the border from Kosovo into Albania through a small border post near the village of Morina. It is estimated that approximately half of all refugees who left Kosovo crossed the border at Morina, so documenting the massive influx of refugees across this border was no small task. [1] The records maintained by the Morina border officials do not contain any personally identifiable information, they simply identify the date at which a group of individuals crossed the border, the number of individuals in each group, and the home province (in Kosovo) of the individuals in the group. [2] The number of individuals known to have passed through Morina between March and May of 1999 can be considered below in Figure 1, which illustrates the change in refugee count over time for the months considered by the following analyses.

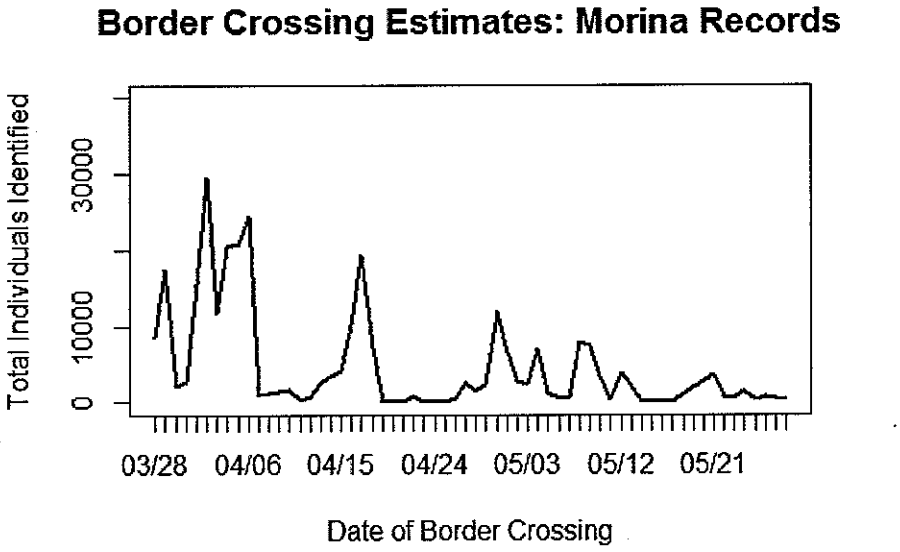

Figure 1

In June of 1999, an additional source of border-crossing records was generated by the Institute for Policy and Legal Studies (IPLS) in Washington DC, in partnership with the American Association for the Advancement of Science (AAAS), who together conducted a series of interviews with Kosovan refugees in camps located throughout Albania and Bosnia, documenting the date on which they crossed the border out Kosovo, the number of individuals in their group, and their home province in Kosovo.[2] These interviews took place at eight refugee camps in Albania and several refugee camps in Bosnia. Specifically, interviews were conducted in four camps in Tirana (Pool camptents, Greek camp-houses, Sports palace, and Mullet), two camps in Korça (Pojska and Qatrom), and two camps in Kukes (Kukes I-Arcobaleno and the United Arab Emirates Camp).[3] Publically-available documentation is not available regarding the specific site at which interviews took place in Bosnia. Once more, the date of border crossing of each individual considered in the record is plotted over time. Although the magnitude of observed individuals is significantly lower for the interview records than for the Morina border records, many of the same peaks and patterns of border-crossing can be observed in both sets of records, as can be seen in Figure 2.

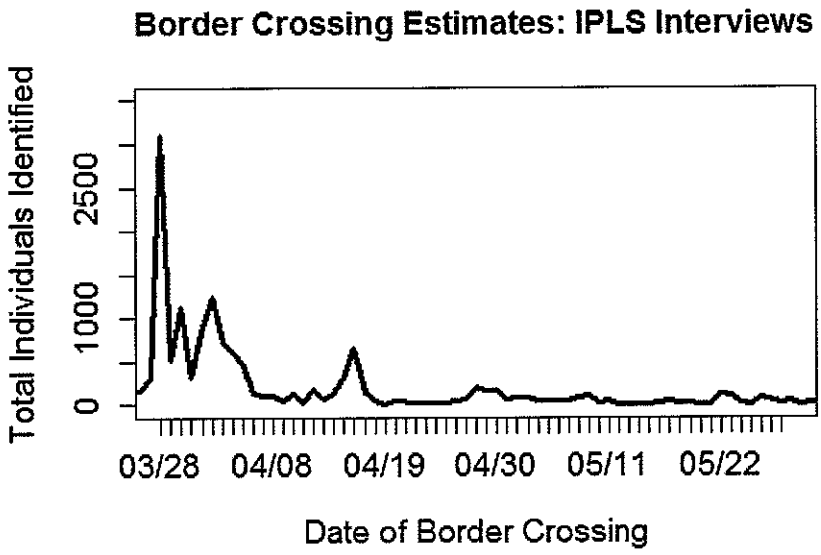

Figure 2

For the purpose of these analyses, 1,681 unique interview records were considered based on publicly-available documentation available from AAAS. These interview records are associated with over 14,430 unique individuals who were living in refugee camps in June of 1999. A total of 19,126 unique records generated by Albanian government officials at the Morina border were also included in these analyses, corresponding to a total of over 126,461 individuals known to have left Kosovo through the Morina border post. For both lists, records were maintained at the per-group level as opposed to at an individual level. Since both sets of records documented a number of the same observations (home province of the group, date of border crossing, and number in group), it was possible to use record-linkage methods to associate group-level observations from one record with group-level observations from the other record.

Methods

In order to generate an estimated range of border-crossing counts, data were aggregated and processed using record-linkage methods and capture-recapture estimation, which together yielded an estimate of the total number of groups who crossed the Kosovo border between March and May of 1999. A constant multiplier was then applied to this group-level estimate in order to identify the estimated number of individuals who crossed the Kosovan border during this time period. While the specifics of each of these steps will be discussed at length below, it is important to highlight the fact that, through this sequential processing, error is compounded at a variety of steps. There are errors associated with each level of this analysis, as data are further and further processed, there is a very real possibility that these errors are compounded in a way that meaningfully impacts results. During record-linkage, errors are associated with both false matches (instances in which truly different groups are classified as matching between the two lists) and false non-matches (instances in which truly identical groups are not classified as matches). During capture-recapture analyses, these errors are considered to be truth, and are processed as such. Since capture-recapture analyses are reliant on random sampling mechanisms with their own associated error, additional error is introduced into the estimates during this phase of the analysis. Similarly, the multiplier applied

to the capture-recapture count estimate has its own associated error since it is based on a sample statistic with associated sampling uncertainty.

As an attempt to capture as much variation as possible within these sequential analyses, estimates were processed as upper and lower bounds to generate a pseudo-confidence interval. Although this interval is not associated with a statistically meaningful level of confidence, it represents bounds of upper and lower estimates based on an understanding of the data and its underlying structure. Such confidence intervals could perhaps be made more rigorous through the use of non-parametric methods such as bootstrapping, although such methods are outside the scope of this report. One standard of confidence proposed by Seybolt et al. in the text *Counting Civilian Causalities*, albeit related to observation of casualties instead of border-crossings, is that:

> *"in some cases it may be preferable to give credible lower, and perhaps upper, bounds for war deaths based on the lists' data. These bounds can supplement, or even replace, estimates based on uncertain and unverified assumptions; often, they will be sufficient for practical purposes while directly conveying a sense of necessary imprecision."* [5]

Processing data as upper and lower confidence bounds, then, is a functional way of considering these multiple levels of error aggregation throughout each step of the analysis.

Record Linkage

In order to classify identical records between the two lists, comparisons were made between each level of observation of the interview records and each level of observation of the Morina border station records. Each comparison was made multiple times and with varied thresholding criteria. Thresholding criteria biased towards both false negatives (as in the case of "strict matching") and towards false negatives (as in the case of "flexible matches") were considered for the purposes of these analyses. These thresholding criteria were selected based on the assumption that errors associated with misinformation occurred much more frequently than errors associated with typographical errors. Thus, thresholding criteria were selected in ways that accounted for misreported or miscounted data. Flexible threshold criteria were informed by historical context provided by the work of Patrick Ball of the Human Rights Data Analysis Group, who found that, for multiple records regarding Kosovo border-crossing estimates, the estimated numbers of total individuals tended to agree when averaged over a two day period. [1].

The algorithm comparing interview observations to Morina observations compared each interview observation to each Morina observation, and, for each comparison, determined whether the two observations were matches or non-matches based on the specified thresholding criteria. Since each list contained multiple identical records (two truly different groups of the same size and from the same province who crossed the border on the same day), for some instances, multiple matches were found within the Morina list for a single observation of the interview list. In these instances, only one match was counted since the lists were considered to be deduplicated. This algorithm, while computationally minimalist and comparatively easy to implement, is slightly biased towards overmatching, since under the structure of the algorithm, it is possible for a single Morina observation to be matched to multiple interview observations. However, given the magnitude difference between the Morina data and the interview data (19,126 records vs. 1,837 records), this was determined to be an acceptable introduction to error given the reduction in computational demands allowed for by using this algorithm. A more computationally demanding alternative to the structure of

this algorithm would be to remove each Morina observation from consideration once it has been classified as a match; however, it is not expected that implementation of such an algorithm would dramatically change the results of this analyses given the relatively small count of Morina records which were matched to multiple interview records.

Capture-Recapture

Capture-recapture analyses have been used in a number of instances to consider lists of human-rights related data, including cases in Guatemala, Peru, East Timon, Columbia, and Bosnia, as well as in Kosovo [5]. When capture-recapture methods rely on overlapping incomplete lists, such as the method proposed in this paper, they are also sometimes referred to as dual-record estimation, or multiple systems estimation when more than two lists are used. Such methods are particularly useful in cases such as these, in which it is impossible or nearly impossible to observe a full population.

Capture-recapture methods are heavily reliant on the assumptions of capture-recapture analysis, including the existence of a closed population from which random samples were taken, accurate classification of matches and nonmatches, independence of samples.

The first assumption, the assumption of a closed population from which random samples are taken, requires that the population of refugees is closed, or that each individual who crossed the border at Morina survived and traveled to a refugee camp. Statistically speaking, this assumption requires that every refugee has a non-zero chance of discovery by at least one of the lists considered. This assumption is relatively well-met by the data considered, since the NATO and Albanian national presence near the Kosovo border helped to ensure that most individuals who crossed the border at Morina survived and received help in entering one of several refugee camps. While it is nearly certain that a proportion of individuals who crossed the border did not enter refugee camps and instead continued their travels or stopped to seek lodging with friends, relatives, or acquaintances, this proportion was assumed to be relatively small. The most likely violation to this assumption is related to the potential for certain sub-populations of refugees who were less likely to cross through a border post and less likely to seek refuge in an international camp (for instance, individuals who feared surveillance or discovery). These individuals are not accounted for by the methods considered in this analysis, and the presence of significantly sized unobserved groups such as these could negatively impact the accuracy of capture-recapture analyses applied to the entire population of refugees.

The second assumption of accurate classification of matches and non-matches is almost certainly not met through the use of point estimates, which even in the best cases nearly always include error due to misclassification. Violation of this assumption was addressed by considering upper and lower bounds of the counts of matching and non-matching through the use of both "strict" and "flexible" thresholding criteria. Although this method does not perfectly account for misclassification error, it gives a range of the probable number of matches in a way that can meaningfully account for this violated assumption. Matching criteria with bias towards over-matching, in this case, flexible thresholds, will generate capture-recapture estimates that are overly low. Conversely, matching criteria with bias towards under-matching, in this case, strict thresholds, will generate capture-recapture estimates that are inflated or overly high.

The final assumption made regarding capture-recapture analysis is that the samples considered in each part of the study are independent. In context, this assumption requires that individuals

identified in the border-crossing records at Morina are not more or less likely to be interviewed at a refugee camp. It is possible that, the two lists considered in these analyses are not independent, given the methods of data collection employed to generate these lists. Since the Morina border station lies on the border of Kosovo and Albania, and since individuals in Albanian refuge camps are heavily represented in the interview sample, there exists the potential that it is systematically more likely that each individual interviewed crossed at the Morina border station than at another border station, for instance, a border station near at the border of Kosovo and Macedonia. However, it is worth nothing that subject matter experts believe that the majority of individuals who left Kosovo in early 1999 did so by crossing the border between Kosovo and Albania, so despite the potential violation of independent samples, the sampling mechanisms were at least moderately representative of the full population of Kosovan refuges who left Kosovo between March and May of 1999.[1] However, potential violation of this assumption is clearly worth paying close attention to, since a lack of independence could dramatically impact the validity of capture-recapture estimates. For a case such as this, where it is probable that an observation is more likely to occur on a second list (the interview list) given that it appeared on the first list (the Morina border list), capture-recapture estimates are likely to underestimate the total population in question, since the number of individuals found on both lists may be artificially inflated by lack of independence.

Despite these potential violations of assumptions, capture-recapture was deemed to be a reasonable method given that the results of such analysis were caveated by the discussion of assumptions included above. The estimated total number of border-crossings was calculated based on the equation $n_{total} = {}^{n_{interview} \times n_{Morina}}/_{n_{matches}}$, where $n_{interview}$ is the total number of interview records, n_{Morina} is the total number of Morina border-crossing records, and $n_{matches}$ is the estimated number of matches based on the results of record-linkage analysis.

Extrapolation from Group-Level to Individual-Level Observations

Once capture-recapture methods were used to estimate the total number of groups who crossed the border, it was necessary to apply a multiplier estimating the number of individuals in each group. For these data, the mean number of individuals based on Morina border counts was used since it accounted for a presumably accurate outlier which fairly dramatically impacted the results, and since the mean provides an unbiased estimator of group size.

Results

Based on the matching thresholds and multiplier selected, the estimated number of individuals who crossed the Kosovo border between March and May of 1999 falls within a range of 711,000 to 1,088,000 refugees. While this number does not represent a statistically meaningful confidence interval, it accounts for uncertainty regarding the accuracy of classification of matches and non-matches between the interview list and the Morina border list.

As shown in Table 1, the thresholding criteria selected were found to dramatically impact the estimated number of matches and non-matches. For the sake of generating an estimated range which accounted for the uncertainty associated with the choice of these matching thresholds, one "strict" threshold and one "flexible" threshold were chosen: specifically, exact matching and a second threshold which used exact matching by province but allowed for variation of 1 day and up to 10% variation of group number. As shown in table 1, these two thresholds were the least and most

restrictive, respectively, with regard to categorizing matches.

It is worth noting that, in this instance, province was a particularly important matching criteria. While identical matching identified only 426 matches between the two lists, this number shot up to 1542 when province was removed as an exact matching criteria. Given the type and structure of the data considered, this is not particularly surprising. The dates considered came from a relatively narrow time period, and the group numbers of each party were low and not particularly variable (with the exception of a few notable outliers). With these characteristics in mind, the high level of matching observed following the removal of the province criteria is fairly intuitive, since there is a great deal of expected overlap between both date and group size. While this choice of thresholding criteria was observed to be successful in addressing the data sets considered in these analyses, the choice of thresholding parameters can and probably should vary for different applications and types of data.

Thresholding Criteria		**Matches (*% of IPLS records*)**
"Strict" Criteria: Exact Matching	date, number, province	426 (*25.3%*)
	date, province	592 (*35.2%*)
	date, number	1542 (*91.7%*)
"Flexible" Matching	date, province, $\pm$ 2 numbers	458 (*27.2%*)
	province, number, $\pm$ 2 days	484 (*28.8%*)
	province, number, $\pm$ 3 days	515 (*30.6%*)
	province, day $\pm$ 10% number	533 (*31.7%*)
	province, $\pm$ 1 day, $\pm$ 1 number	600 (*35.7%*)
	province, $\pm$ 1 day, $\pm$ 10% number	650 (*38.7%*)

Table 1: **Consideration of Strict and Flexible Thresholding Criteria**

Following consideration of the above matching criteria, a lower threshold using exact matching and an upper threshold which required exact matching by province but allowed variation of 1 day and up to 10% variation of group number were considered using capture-recapture analysis. This application of capture-recapture analysis can be visualized by a two-way table with a single missing cell, as shown below by Tables 2 and 3. The available cells shown in these tables are used to calculate an estimate for the total population. [6]

	Recorded in Morina data	Not Recorded in Morina data	
Recorded in interview data	426	1255	1681
Not Recorded in interview data	18700		
	19126		

Table 2: **Capture-Recapture at Lower Threshold (Exact Matching)**

	Recorded in Morina data	Not Recorded in Morina data	
Recorded in interview data	650	1031	1681
Not Recorded in interview data	18476		
	19126		

Table 3: **Capture-Recapture at Upper Threshold ("Flexible" Matching)**

Capture-recapture analyses considering the above described matching estimates yielded an estimated range of between 49,462 and 75,471 border-crossing groups. As previously discussed, potential violations to the assumptions required for capture-recapture analysis could cause these numbers to be artificially deflated due to overmatching influenced by non-independent samples. Consideration of upper and lower thresholds represents an attempt to address the uncertainty associated with matching estimates; however, it cannot address lack of independence of samples, which may influence the findings of these analysis.

Since capture-recapture provides only an estimate of the total number of records (at the party-level), a multiplier must be applied to these results in order to estimate the number of individuals. As shown in Figure 3, the distribution of group size is strongly positively skewed, and is heavily impacted by the presence of 3-5 large outliers, one corresponding to a group of 1,961 individuals who crossed the border as a single group on April 1, 1999. This specific observation, as well as other observed outliers, are referred to by Ball as an example of "outflow," related to the observation that less than 2% of the total records correspond to over 20% of people who crossed the border in total since these groups contained such high numbers of individuals.[1] The large group observed on April 1st was documented by the interview process in which testimony was obtained from refugees. One refugee reported that:

> *"The Serbs made people leave [Pec] on Thursday [April 1] at 10:00 AM. They said: "Go to Albania." They went from neighborhood to neighborhood. They burned the houses as people left. There were trucks and buses waiting for us. We were treated like cows. Men and women together. They didn't allow us any baggage, clothes. Just as we were. They slapped the men around a bit. They stole everything from the houses."* [1]

Based on the documentation of such outliers, they were included in the calculation of the standard multiplier, in this case, the mean. In other instances for which there may be less thorough documentation related to such outliers, it would be possible to use the median, with the understanding that the median could potentially be a biased estimator in cases where the observed outliers were not errors, but correct observations.

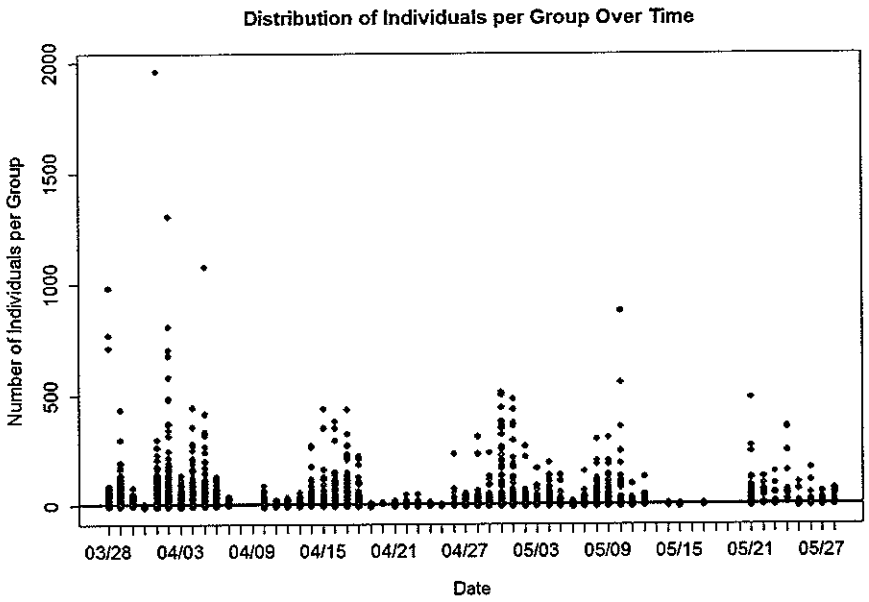

Figure 3

In this case, the sample mean group size was 14.46 individuals, while the observed sample median was far lower, 7 individuals, again highlighting the highly skewed nature of the data. When the standard multiplier was applied to the upper and lower bounds of the group size estimates, a final estimate of between 711,000 to 1,088,000 refugees was identified. Thus, based on the methods described above, the number of refugees who fled Kosovo between March 28 and May 28 of 1999 is estimated to be between 711,000 and 1,088,000 individuals.

Discussion

Although it is impossible to directly assess the estimate obtained above using standard methods, the estimate obtained using the methods proposed in this paper is consistent the results of other more computationally demanding analyses, including those identified below in Table 4. A brief review of the literature and news sources reveals that many sources cite one another without referring the original analyst or the methods used to generate the estimated number of refugees. As these types of international human-rights analysis can be highly politically controversial, one potential explanation for this phenomenon is a hesitance to release estimates publicly before they have been fully validated.

Source	Estimated Count	General Method, details
Eaneff (691)	711,000 to 1,088,000	record linkage, capture-recapture
AAAS (Ball)[1]	over 850,000	multiple-system estimation (capture-recapture)
UC Davis [7]	600,000	uncited estimate, unknown method
Doctors Without Borders[8]	over 500,000	estimate for April of 1999
The Guardian [9]	over 86,000	uncited estimate, unknown method
New York Times [10]	over 86,000	cites UNHCR (source not identified)

Table 4: **Estimates of Refugee Count by Source**

Despite the fact that these analyses present some methodological shortcomings with respect to quantification of error, the methods considered are computationally undemanding and may be useful in other applications as a first-pass method for estimating counts related to human migration or human-rights violations using sampling from multiple independent sources of records. While additional work is necessary to address a more complete consideration of error and thus, confidence levels, methods such as these may be useful as comparatively simple and easy to generate early and ongoing count estimates in scenarios such as the Kosovo refugee crisis in 1999.

Specific shortcomings of this type of analysis include the requirement of assumptions which may not realistically be met by the data being considered, a lack of consideration of geographic proximity as a matching criteria (except through exact matching), an inability of to quantify error in a statistically meaningful way, an inability to compare records between three or more lists (based on the current algorithm), and an inability to assess the results of such data.

For the case of migration records related to refugee crisis in Kosovo, it is unlikely that additional independent data will become available (and if such data does become available, it is likely to be heavily impacted by recall bias), and almost certainly impossible that an entirely accurate count of refugees will ever be known. It is worth nothing that future work in such areas should seek, whenever possible, to conduct interviews and collect data in ways that allow for independence of data. However, such a requirement is improbable at best and impossible at worst, given the unique challenges of collecting data during times of intense political and social conflict and violence. In situations that allow for independent samples to be taken, independence of the data collected should prioritized as much as reasonably possible.

Unlike the challenge of collecting data from truly independent samples or the challenge of identifying an accurate final count of refugees, it is quite possible that further methodological development for this type of analysis could allow for consideration of geographic proximity as a matching criteria. Further methodological development would also be necessary in order to quantify error in a statistically meaningful way. More accurate and more precise quantification of error, in particular, would be especially valuable for these types of analysis, although such methodological work would admittedly be more challenging that the consideration of geographic proximity as a matching criterion, which would be relatively simple to implement based on the existing algorithm.

Geographic proximity could be considered using the Euclidean distance between pairs of latitude and longitude, distances which could be stored in a look-up table to speed computation of each comparison. If such latitude and longitude measures were not available at the province level, or if such comparisons were not computationally feasible, distance could also be considered by means of a categorical location variable (e.g. Northern Kosovo vs. Southern Kosovo), although such criteria would be less sensitive than a consideration of Euclidean distance. Alternatively, for cases where computational power was not a concern, geographic boundaries (such as mountains or large rivers) and transportation routes (such as paved roads or other well-traveled routes) could also be considered alongside consideration of Euclidean distance, since such obstacles and aids to travel likely impact the true "similarity" of any two geographic locations.

Quantification of error using the methods described could draw upon recent work addressing the generation of of association rules, which are used to identify structures of correlation between patterns of binary events. [11, 12] Such rules could be used to identify clustered patterns of matching

that exist above a defined confidence threshold. If an algorithm was developed which generated probabilities of matching related to the thresholds described above, other non-parametric methods including kernel density estimation could also be useful for characterizing matches vs. non-matches. However, given the high level of similarity between true matches and true non-matches in this context, such methods may not be practically useful for identifying matches and non-matches. The data considered in the above analyses were taken from a period of only three months, so there is a great deal of overlap between the dates (which were limited to a three month period), the group size (which was generally fairly low), and the home-province of the group (since patterns of refugees were observed to cross the border based on the geographic location of military action, NATO bombings, and other events which systematically forced migration across the border [1]). While such methods would certainly be valuable, their ability to accurately characterize matches and non-matches seems to be far from guaranteed when applied this type of data.

While the estimates and methods presented in this paper are far from perfect, they identify a computationally undemanding and intuitive method for generating preliminary, first-pass estimates of human migration based on sparse, overlapping data which lack unique identifiers. Given the challenges of working with these types of data, which are often collected in attempts to quantify human migration or patterns of human rights violations, these methods may be useful for "quick and dirty" field calculations which could be used by non-profit organizations or other types of relief agencies who would benefit from order-of-magnitude count estimates related to observations which are inherently difficult to measure.

Acknowledgements

This project would not have been possible without the help and support of:

- Sam Ventura (for helping code this project into reality)
- Rebecca Nugent and Steve Fienberg (for help with project ideas)

References

[1] Ball, Patrick. "Policy or Panic." Prepared for the American Association for the Advancement of Science. Accessed online at: http://shr.aaas.org/projects/human$_r$*ights/kosovo/policypanic.pdf*

[2] American Association for the Advancement of Science, Scientific Responsibility, Human Rights and Law Program. "Data on Migration." Accessed online at: http://srhrl.aaas.org/projects/human$_r$*ights/kosovo/migration/index.html*

[3] IPLS/AAAS Survey of Kosovar Refugees Project. "Survey of Kosovar Refugees Project Survey of Refugees' Attitudes About Return." American Association for the Advancement of Science (14 June, 1999). Accessed online at: http://srhrl.aaas.org/projects/human$_r$*ights/kosovo/migration/survey.pdf*

[4] Asher, Jana; Banks, David; and Scheuren, Fritz. J. "Statisical Methods for Human Rights." New York: Springer, 2008. Print.

[5] Seybolt, Taylor B.; Aronson, Jay D.; and Fischhoff, Baruch. "Counting Civilian Casualties." New York: Oxford Univeristy Press, 2013. Print.

[6] Fienberg, Stephen E. "The multiple recapture census for closed populations and incomplete 2k contingency tables". Biometrika (1972) 59 (3): 591-603.

[7] "Kosovar Refugees". University of California, Davis. Migration News (1999) 6(5). Accessed online at: http://migration.ucdavis.edu/

[8] Perea, William A. "Rapid Needs Assessment Among Kosovar Refugees Hosted by Albanian Families and Assessment of Human Rights Violations Committed in Kosovo." Doctors Without Borders (1999). Accessed online at: http://www.doctorswithoutborders.org/publications/

[9] Hooper, John. "860,000 refugees present daunting task." The Guardian (11 June 1999). Accessed online at: http://www.theguardian.com/world/1999/jun/12/johnhooper.

[10] Giussan, Bruno, "New Technologies Employed to Trace Kosovar Refugees." New York Times, Technology (8 June, 1999). Accessed online at: http://partners.nytimes.com/library/tech/99/06/cyber/eurobytes/08eurobytes.html

[11] Kotsiantis, Sotiris and Kanellopoulous, Dimitris. "Association Rules Mining: A Recent Overview". GESTS International Transactions of Computer Science adn Engineering. (2006) 32 (1): 71-82.

[12] Agrawal, Rakesh; Imielinksi, Tomasz and Swami, Arun. "Mining Association Rules between Sets of Items in Large Databases". ACM SIGMOD Conference, Washington DC, USA. (2003).