Appendix A: Data and Methods

This project is based on records maintained by the Albanian officials at the Morina border crossing from Kosovo. These data are the major component of the statistics computed to estimate the population parameter B_{vd}, the number of people crossing the border on day d from city or village of origin v.

These registries are partial: the records list the place of origin only for some people, and other people crossed the border without being registered at all. The number of people crossing the border on each day d from each village v will be estimated by imputing the missing information about border-crossers' places of origin, and this estimate is denoted $\hat{b}_{vd}$ where the hat indicates that this is an estimate. The data used for the estimation of $\hat{b}_{vd}$ are described in Section A1.

To impute the two components of $\hat{b}_{vd}$ for which some data are missing, data collected from refugees sampled from camps in Albania and Macedonia are used to allocate counts of people with missing origin data to municipalities. Section A2 describes how the imputation was done.

The border registries record when people left Kosovo, not when they left their homes. The analytic goal of this project is to map the flow of refugees out of Kosovo according to their locations of origin and the time at which they left that origin. Formally, the number of people leaving their homes is denoted G_{vd}, the number of people who left each city or village v on day d. G_{vd} will be estimated by transforming $\hat{b}_{vd}$ using additional data sources described in Section A1. Most Kosovar Albanian refugees left their homes and exited from Kosovo on the same day. However, many other refugees were in transit for varying periods. The transformation of the estimated number of people exiting Kosovo, $\hat{b}_{vd}$, into an estimated number of people departing from their homes, denoted $\hat{g}_{vd}$, is described in the Section A3.

Both the imputation and projection processes just described use sample data, and so both introduce sampling error into the estimate. Section A4 uses a statistical technique called jackknifing to estimate the error of $\hat{g}_{vd}$ due to sampling.

The estimate $\hat{g}_{vd}$ depends on the imputation to municipalities of approximately half of $\hat{b}_{vd}$. In Section A5, the sensitivity of $\hat{g}_{vd}$ to the imputation process is tested by considering the effect of ratio bias. That section concludes that the substantive interpretation of $\hat{g}_{vd}$ is not affected by ratio bias.

The imputation of people with missing municipalities of origin involved several assumptions. Sensitivity analyses of several of these assumptions are presented in Section A6. The section concludes that the estimates $\hat{g}_{vd}$ are robust to different means of conducting the imputation.

The estimate $\hat{g}_{vd}$ also depends on the method by which the estimated number of people exiting Kosovo, $\hat{b}_{vd}$, are projected back to the estimated time that they left their homes some days prior to crossing the border. The Section A7 examines the sensitivity of $\hat{g}_{vd}$ to the estimated transit time, and it concludes that $\hat{g}_{vd}$ is unaffected by varying transit time distributions that fit the observed data.

The substantive analysis in this report has used the patterns of refugees' entering Albania as an indicator of all Kosovar Albanian refugees' departure from their homes, including those who left Kosovo but went to countries other than Albania. Section A8 examines how well refugees who entered Albania may have represented the entire universe of Kosovar Albanians who left their homes, and it concludes that including data from other countries to which refugees fled would not substantially alter the analysis based on the Albanian data alone.

A1 Data: Sources and notation

The core data for this project are the records maintained by the Albanian border guards at Morina. During the period from 28 March to 28 May 1999, the guards registered 19,126 households or groups crossing the border. Most records included the name of the head of the household, his (or less frequently, her) age or year of birth, the household head's home village, and the date on which the group entered Albania. The number of people registered with their place of origin for each day is denoted as b^R_{vd}, where the superscript upper case "R" indicates that these people were fully registered. This number is obtained directly from the data and is therefore denoted with a lower-case "b" without a hat. The total number of people is 191,693 (see Equation 1), obtained by summing over all villages v and days d during the period 28 March – 28 May.

Equation 1

$$b^R = \sum_d \sum_v b^R_{vd} = 191{,}693$$

Records in the border registry represented either families – as described in the previous section – or sometimes larger groups listed only by an estimated total in the group: "95 people on foot." These daily counts are denoted $b^a_{\bullet d}$ where the superscript lower-case "a" indicates that these people were registered only in the aggregate, and the subscript "dot-d" indicates that the data are available for each day d but that the places of origin v are unknown and represented only by their sum. The total number of people registered in the aggregate is 85,678 (see Eq. 2), where d is summed over the period 28 March – 28 May. The number of people registered in the border records is 276,461 (see Eq 3). The totals $b^a_{\bullet d} + b^R_{\bullet d}$ are aggregated to two-day periods and presented in Graph 1.1 (in Part I).

Equation 2

$$b^a = \sum_d b^a_{\bullet d} = 85{,}678$$

Equation 3

$$b^R_{\bullet d} = \sum_v b^R_{va}$$

Estimating G_{vd} required that the border data be transformed in several ways. To begin with, many people crossed the border but were not registered at all. The number of people who were not registered at all is denoted $b^o_{\bullet d}$ with the superscript lower-case "o" indicating that these are the people who overflowed the registry process. The number $b^o_{\bullet d}$ is derived directly from the data and so it is lower-case and has no hat. Also, among those people who were registered, approximately one-third of them do not have their home village recorded, and these people are denoted $b^a_{\bullet d}$. In order to correct these various problems, the border data were augmented with several additional sources.

- UNHCR reports were issued at the daily press conferences in Tirana (24 March – 28 April). These reports estimated the number of Kosovar Albanians entering Albania in the previous 24 hours.
- The Albanian Government's Emergency Management Group (EMG) reported daily (or every 12 hours) the number of Kosovar Albanians entering Albania (14 April – 28 May).
- The IPLS/AAAS team registered a sample of 1,837 Kosovar Albanian families (including some 12,092 individuals) residing in 18 camps in Albania. The sample was collected by in May-June 1999 as listings from which samples were to be drawn. Some camps were listed completely, but in mid-June when it became clear that refugees would soon return to Kosovo, non-probabilistic samples were taken in other camps so that data could be obtained from as many camps as possible.
- As a pilot project, 83 interviews were conducted by the IPLS/AAAS team in May-June 1999, probabilistically sampled from listings in four camps in Albania.[1]
- 136 interviews were conducted by the University of California (Berkeley) among Kosovar Albanians sampled probabilistically from camps and private residences in Bosnia in July, 1999, as part of the IPLS/AAAS study.
- 123 interviews conducted by Human Rights Watch with Kosovar Albanians in various parts of Albania in March-June, 1999. The respondents were chosen according to HRW's judgement and brief screening conversations as HRW sought information on human rights violations.[2]
- 1,180 interviews conducted by PHR/Columbia in their survey of Kosovar Albanians, April-May 1999.[3]

The following sections explain how these additional data sources were used to resolve the problems with missing data in the border registry and transform the border-crossing counts into estimated numbers of people leaving their homes.

A2 Imputing missing data

The objective of this section is to estimate b_{vd}, the number of people from each village who cross the border each day. Ignoring places of origin for a moment, the number of people crossing the border on each day d is estimated as $b_{\bullet d} = b^{R}_{\bullet d} + b^{a}_{\bullet d} + b^{o}_{\bullet d}$, where $b^{o}_{\bullet d}$ must still be estimated. Subsection A2.1 below constructs an estimate for $b^{o}_{\bullet d}$.

In Subsection A2.2, an intermediate statistic is defined as the estimated number of people who crossed the border for whom place of origin was unknown: $\hat{b}^{A}_{\bullet d} = b^{a}_{\bullet d} + b^{o}_{\bullet d}$, where the superscript upper-case "A" means that this

is the sum of the place-unknown border crossers for day d, and the subscript "dot-d" for each term indicates that the villages v are unknown and represented by their sum. $\hat{b}^A_{\bullet d}$ must be transformed into $\hat{b}^A_{vd}$ through imputation using interview data (described below) and the data from the fully-registered border crossers b^R_{vd}.

The estimated number of people from each village who crossed the border each day is then the sum of the people who were fully registered plus the people imputed: $\hat{b}_{vd} = \hat{b}^A_{vd} + b^R_{vd}$.

A2.1 Unregistered border crossers

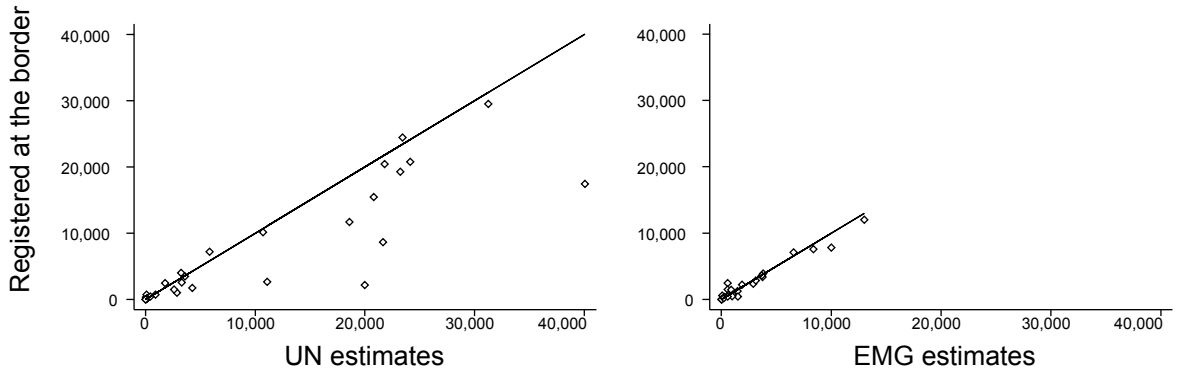
The border at Morina was a chaotic place during March-May 1999. Often thousands or tens of thousands of people would surge across the narrow road from Kosovo toward the border point. The border guards' job was to register the name and home city or village of the head of household of each group crossing into Albania. When flow increased to very high levels, the border guards would be unable to register all the people entering without creating huge delays, backing people up onto the Serbian side. The guards decided that it was more important to let people cross into safety in Albania than to delay people in insecure conditions in order to fulfill bureaucratic requirements.[4] When this happened, the guards would simply permit people to cross without being registered.

As described in the list of data sources in the previous section, there were other organizations keeping track of the overall number of entrants. UNHCR had people counting Kosovars as they came through the border point, and later the Albanian government Emergency Management Group (EMG) coordinated a variety of organizations that continued this work. Crucially, the UNHCR and EMG estimates were independent of the work of the border guards. The border guards' counts, the UNHCR counts, and the EMG counts provide independent means for assessing the number of Kosovars Albanians entering Albania day-by-day.

In Graph A1, the daily counts are presented for the border registry, the UN estimates, and the EMG estimates. On the left, the border counts and UN estimates are compared for the period 24 March – 28 April; on the right, the border counts and EMG estimates are compared for the period 27 April – 28 May. The 45-degree line through both graphs shows the points at which the vertical and horizontal data are equal; note that most of the dots are on the line, but the dots that are not on the line are below it. As expected from the overflow analysis described above, this indicates that when the border counts are not equal to the UN or EMG estimates, the border counts are smaller.

The series disagree on the high-flow days, such as 28 and 29 March when the UN counted more than 21,000 and 40,000 refugees each day, respectively, and the border guards registered one-third to one-half of the entrants. Note that

Graph A1: number of Kosovar Albanians crossing the Albanian border at Morina, by day, comparing border registration counts with UNHCR estimates (left), and with EMG estimates (right)



most of the days on which the border count and the UNHCR count disagree, there were more than 20,000 people crossing the border, and the border registration process clearly collapsed. The EMG counts show a similar pattern for days of more than 10,000 entrants where there is overflow, although there are several days in which the border counts are very slightly higher than the EMG estimates.[5]

On low flow days, the three series tend to agree closely. On both graphs, most of the dots toward the lower left corner are within a few dozen or a few hundred people of the line indicating that the border count and the independent estimate are equal. For example, on 14 April, when the UNHCR estimated that 3,600 people came into Albania, the border guards registered 3,475. These differences confirm the explanation for border underregistration: on high-flow days, people simply streamed around the registration process.

The motivation for using the UNHCR and EMG estimates to supplement the border registration counts is that the border data underreport the number of border crossers for each day due to overflow. The UNHCR and EMG counts (denoted as b_d^U, with the upper-case "U" indicating that these data come primarily from the UNHCR) were used to estimate the overflow as described below. The total estimated number of people who crossed the border between 24 March and 28 May is approximately 404,000 (see Eq. 4), where *d* is summed over the period 28 March – 28 May.

Using the b_d^U from the UNHCR and EMG counts, the daily overflow is computed as $b_{\bullet d}^O = b_{\bullet d}^U - b_{\bullet d}^R - b_{\bullet d}^a$. The total daily number of people $b_{\bullet d}^A$ that crossed the border and had to be imputed to villages is computed[6] as $b_{\bullet d}^A = b_{\bullet d}^O + b_{\bullet d}^a$.

Equation 4

$$b^U = \sum_d b_d^U \approx 404{,}000$$

A2.2 Missing data: border crossers with missing place identification

Of the original 276,000 people registered in the border data, approximately 69% are identified by their home village. When the overflow from the UNHCR and EMG counts are added, only about 49% of the people estimated to have crossed at Morina came from known home villages. The estimated number of these people for each day is denoted $b^A_{\bullet d}$ and computed following the method described above. The objective of this section is to disaggregate $b^A_{\bullet d}$ into estimated numbers of people leaving each village on each day, denoted $\hat{b}^A_{vd}$.

There are several reasons why the refugees who were registered at the border (b^R_{vd}) and those who were registered only in the aggregate or not registered at all ($b^A_{\bullet d}$) might be quite different. For example, the border guards told this project that people in cars were always registered, but that people on foot were sometimes missed. If people in cars originated in locations systematically different from the people on foot, then the distribution of people across origin villages in b^R_{vd} could be different from the distribution in $b^A_{\bullet d}$.

Relative to the distribution in b^R_{vd} of people among municipalities and villages, people in refugee camps in Albania might be a more representative sample of the unregistered people with respect to their distribution among villages of origin, especially if the list of people from camps could exclude people who were registered at the border. In other words, imputation using surveys in camps as the donor data might be less biased than imputation using just the registered border crossers as donors.

A list of people sampled from camps was constructed by combining two samples: respondents in the PHR interview project (including 598 interviews), and respondents to the IPLS camp listing (1,837 interviews). Only respondents who crossed the border at Morina were included in these lists. The two lists were matched to each other by respondent name, and duplicate entries were eliminated. The lists were appended together to form a unified list from both sources; this list of 2,409 respondents is called the survey list. The survey list was then matched to the border registries, and 154 respondents who appeared in the survey list and the border registry were deleted from the survey list. Also deleted were respondents who entered Albania before 24 March or after 11 May. The resulting list is called the "reduced survey list."[7]

In the reduced survey list, each respondent was identified by the date (d) he or she crossed the border at Morina and his or her municipality of origin (u). The list is denoted as S^L_{ud}, where the superscript upper case "L" indicates that this is the listing part of the sample, and the subscript d and u indicate the day the respondent crossed the border and the municipality of origin of the respondent. The reduced survey list included 2,078 families made up of 14,864 individuals.[8]

Equation 5

$$w_{ud} = \frac{s_{ud}}{s_{\bullet d}}$$

The counts were converted to proportions[9] of each municipality for each d as shown in Eq. 5. Since these are proportions, summing w_{ud} over d is equal to 1 for each u.

Using w_{ud}, $b^A_{\bullet d}$ can be disaggregated: $\hat{b}^A_{ud} = w_{ud} \cdot b^A_d$, across all u and all d, linking on d. This estimate, $\hat{b}^A_{vd}$, is the number of people estimated to have crossed the border on each d from each municipality u without being registered by place of origin.

Computing w_{ud} by using S^L_{ud} introduces sampling error which will be discussed later. However, this method is believed to largely avoid the bias that would have been introduced by estimating $\hat{b}^A_{vd}$ solely from distributions calculated from the registered border crossers, b^R_{vd} (note that every v is contained in a u). Nevertheless, the data in S^L_{ud} did not have enough information to permit the imputation to city and village v, as implied by the notation (S^L_{ud} does not have a v subscript). To disaggregate $\hat{b}^A_{vd}$ to villages, the distributions in the registered data within municipality and time period were used as the donor data. Although this does introduce the bias avoided by using in S^L_{ud} in the prior step, the effect of the bias is limited to distributions among cities and villages within each municipality. Since the substantive interpretation in this report is at the inter-municipality level, the potential intra-municipality bias introduced here does not affect the main analyses.

Equation 6

$$w_{vud} = \frac{b^R_{vud}}{b^R_{\bullet ud}}$$

As with the prior step, the distribution was computed as shown in Eq. 6. The next step is the imputation, computed[10] as $b^A_{vud} = w_{vud} \cdot b^A_{ud}$, for all v and all d, linking on u and d.

This estimate (with the municipality subscript u dropped) is $\hat{b}^A_{vd}$. This is the piece needed to compute $\hat{b}_{vd} = \hat{b}^A_{vd} + b^R_{vd}$, the number of people crossing the border from each village on each day. The next step is to convert the counts of people crossing the border into counts of people leaving their homes. This transformation is the subject of the following section.

A3 Projection from exit time to leaving time

On any given day, approximately half of all the refugees crossing the border had left their homes on that same day, but the other half of the refugees had left their homes some days prior to crossing the border. This section describes how estimated counts of people crossing the border, $\hat{b}_{vd}$, were projected to estimated counts of people exiting their homes, $\hat{g}_{vd}$.

A sample list was composed of interviews conducted by HRW, PHR, and IPLS among Kosovar Albanian refugees in camps in Albania. Each of the three lists was matched to the other two and duplicates dropped to create a single, unified dataset of 753 interviews.[11] The dataset is denoted S^I_{udt}, where the superscript upper case "I" indicates that these are interviews. The information sought from these interviews was how long each respondent had been in transit between leaving his or her home in Kosovo and crossing the border at Morina. Since the IPLS camp listings did not include this information, there are fewer cases available for S^I_{udt} than there were in S^L_{ud}.

In addition to the lower case subscripts u and d (indicating the municipality of origin and the day the respondent crossed the border at Morina), S^I_{udt} includes the length of time (in days) between when they left their homes and when they crossed the border, denoted t, where $0 \leq t \leq 70$. Refugees for whom t > 70 are out of range because although they crossed the border during the period of analysis (24 March $\leq d \leq$ 11 May), if t > 70, then no matter when they crossed the border, they left their homes before 24 March and so fall outside the scope of this analysis.

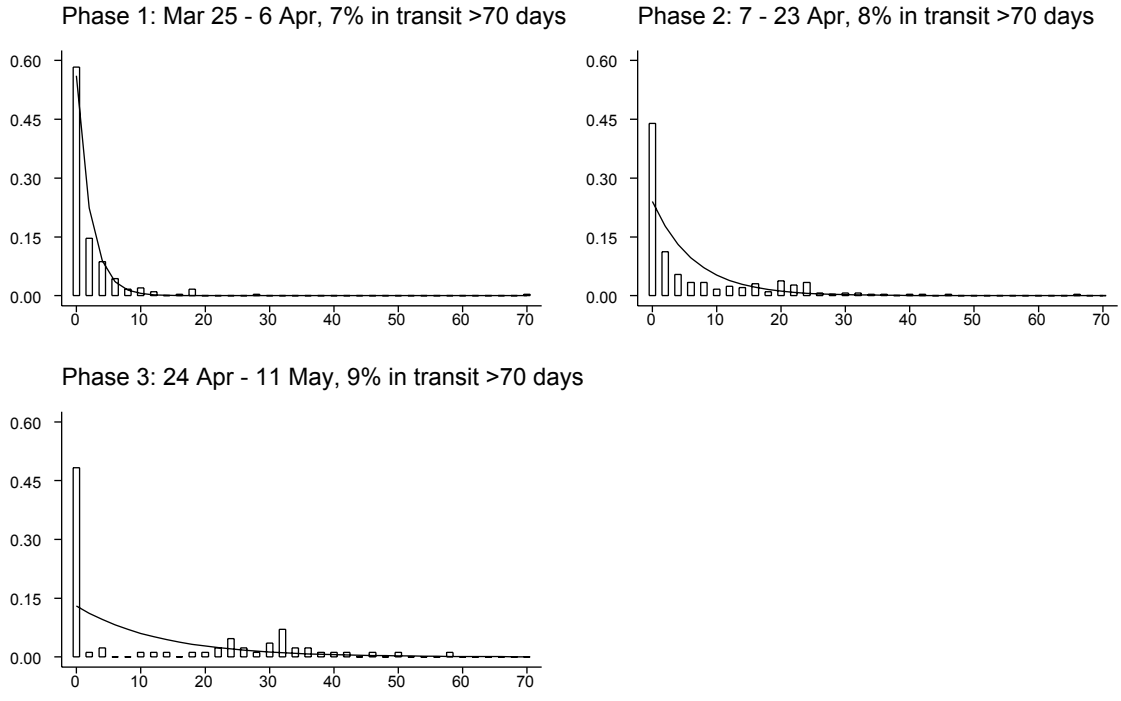
As with the imputation of missing places, the list S^I_{udt} was converted to proportions. First the list was summed to counts of families departing each day who experienced different transit times,[12] denoted s_{dt}. The distribution of transit times among municipalities was also considered, comparing different regions of Kosovo, but only very small differences in the distribution of transit times across regions were found. We judged these differences not to be significant.[13] Therefore, we have ignored region in this analysis.[14]

Equation 7

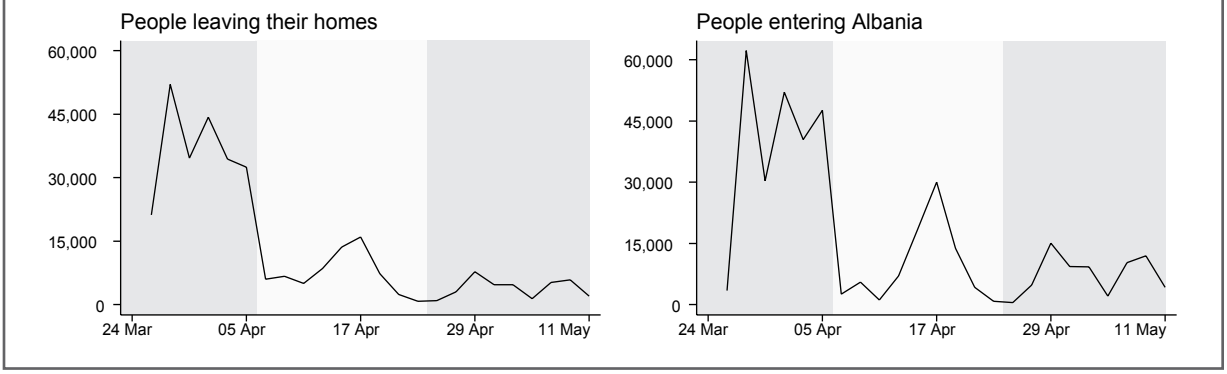
$$w_{dt} = \frac{s_{dt}}{s_{d\bullet}}$$

The counts in s_{dt} were converted to proportions w_{dt} (as shown in Eq. 7), the distribution among transit times t for each d. The distributions of w_{dt} for each of the three phases is shown in Graph A2.

Graph A2: Proportion of people entering Albania by length of time in transit (in 2-day periods), with exponential distribution (plotted line)



Graph A3: Number of Kosovar Albanians departing home and crossing the border, by two-day period



In Graph A2 is it clear that during all three phases, most refugees exited Kosovo on the same day that they left their homes; expressed differently, they were in transit for zero or one days. Across all the periods, 50% of all refugees groups that exited Kosovo left their homes earlier the same day, and 71% of all refugees left Kosovo the same week as they left their homes. Some refugees had been in transit for very long times, but they are a small minority – less than 8% of all migrant groups were in transit for more than 70 days. Although the people for very long times are not shown in Graph A2, the proportion of people in transit for longer than 70 days are shown in the titles. Since these distributions may at first glance appear to be exponential, the exponential distributions are also plotted to show that the best fit exponential distribution differs considerably from the empirical distribution among transit times (see Section A7 for a more thorough examination of the transit time distributions).

Phase 1: 24 March – 6 April

Phase 2: 7 – 23 April

Phase 3: 24 April – 11 May

Multiplying the computed number of people crossing the border from each village by the transit time distribution spreads each estimate $\hat{b}_{vd}$ to a range of transit times t. The actual computation is $\hat{b}_{vdt} = \hat{b}_{vd} \cdot w_{dt}$, linking on d.

The time at which each $\hat{b}_{vdt}$ left their home in village v is denoted q and computed as $q = d\text{-}t$. Each $\hat{b}_{vdt}$ can be mapped to q by $\hat{b}_{vdtq} = \hat{b}_{vdt(\,d-t)}$. Summing across all d and t for each q transforms border-crossing counts b into home-departing counts as shown in Eq. 8. Both d and q are dates; q expresses that a particular group of people from village v who crossed the border on day d left their homes on day q. The total (summed across villages v) daily number of people leaving their homes, $\hat{g}_d$ and crossing the border b_d are presented below in Graph A3.

Equation 8

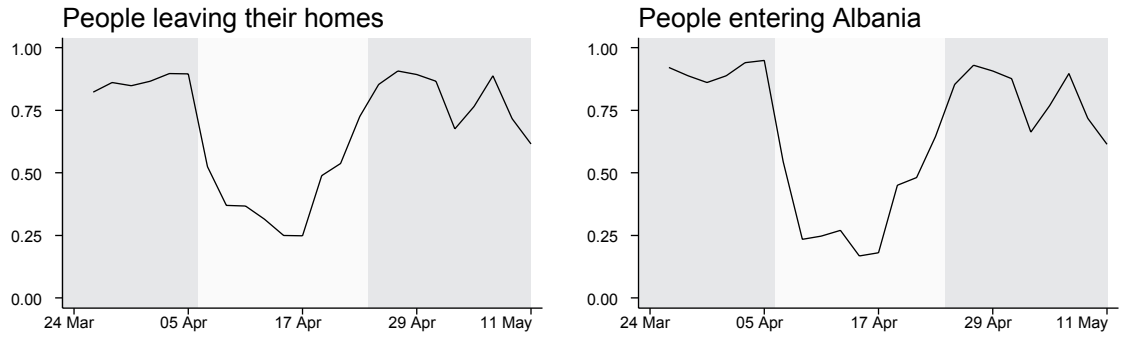
$$\hat{g}_{vd} = \sum_{d} \sum_{t} \hat{b}_{vdtq}$$

Graph A3 shows the number of people leaving their homes and the number of people entering Albania, both graphed across time in two-day periods. The pattern in time is very similar in the two graphs, although the home leaving pattern is smoother than the pattern of people entering Albania: the lows are a bit higher and the highs are a bit lower. Furthermore, there are fewer people represented in the home leaving graph $\hat{g}_d$ than in the entering Albania graph b_d. This difference is a result of the periods shown on these graphs. Some of the

people who entered Albania (b_d) between 24 March and 11 May departed from their homes before 24 March, and therefore are omitted from the graph of $\hat{g}_d$. Between 24 March – 11 May, slightly more than 350,000 people are estimated to have left their homes and crossed the border at Morina; slightly more than 50,000 people left their homes before 24 March or entered Albania after 11 May. The municipality-specific graphs used in the analysis of Phases (i.e., Graphs 2.1, 3.1, and 4.1) are all versions of the graph on the left of A3.

The greater smoothness of the home leaving pattern relative to the entering Kosovo pattern may reflect the cumulating effect that transit may have had: it is possible that people were detained in transit, and then at particular moments, large number of people were permitted to exit Kosovo. For example, as was noted in the narrative section analyzing Phase 2, substantial numbers of people from Srbica left their homes and were in transit during early April but then exited during a concentrated period in mid-April. It may also be that the distribution of transit times w_{dt} does not fully capture the "lumpiness" of how people exited Kosovo, and so using w_{dt} may have created more smoothness in the estimated number of people leaving their homes than in fact occurred. However, both of these possible effects would act to reduce the difference between the high points during the middle of each phase and the low points that define the phase boundaries. That is, the smoothness resulting from the computation

Graph A4: Proportion of people who originated in southern and western Kosovo, by 2-day period, as a fraction of the people leaving their homes and as a fraction of people entering Albania



is conservative because it reduces the fit of the estimates to the hypothesized phases. Even in the smoother graph on the left of Graph A3, the phase structure remains clear.

The interpretation of the changing concentration of people from western and southern Kosovo is identical when considering either people entering Albania or people leaving homes in Kosovo. In both graphs in A4, the proportion from the south and west is high (greater than 75%) during Phase 1, low (less than 50%) during Phase 2, and high again (greater than 75%) during Phase 3.

Furthermore, in both graphs the shifts from high to low and low to high proportions of people originating in the south and west occur at the phase boundaries on 7 April and 24 April. However, the proportions of people from the south and west during Phase 2 vary less and are consistently lower in the home-departure data than in the exiting-Kosovo data. As in Graph A3, the relatively smoother patterns in the pattern of people leaving their homes relative to people entering Albania results from the smoothing effect of w_{dt}. Also as in Graph A3, the smoothing effect does not affect the substantive interpretation of Graph A4.

The patterns in g_{vd} provide a basis for empirical analysis of the patterns of people leaving their homes during the period 24 March – 11 May. The estimation of $\hat{g}_{vd}$ included the use of samples for both the imputation of missing place information and for the projection of numbers of people crossing the border, and therefore $\hat{g}_{vd}$ is subject to sampling error. In addition, a number of assumptions were made in the transformation of $\hat{b}_{vd}$ to $\hat{g}_{vd}$. The next sections investigate the effect on the substantive interpretations of the sampling error and sensitivity to assumptions. Section A4 estimates the sampling error of $\hat{g}_{vd}$. Sections A5 and A6 analyze the sensitivity of $\hat{g}_{vd}$ to different assumptions in the place imputation and time projections from home departure to entering Albania. Section A7 examines to what extent conclusions drawn from analysis of refugees entering Albania through Morina can be interpreted as an analysis of refugee movements out of Kosovo more generally.

A4 Sampling error of $\hat{g}_{vd}$

For each estimated number of people leaving their homes in village v on day d, denoted $\hat{g}_{vd}$, there is an associated error. This section provides the means by which these errors are calculated.

Given the complexity of directly estimating the errors associated with an estimator based on two different samples, the jackknife method was used.[15] In this method, each of the two samples is divided into k distinct groups. The entire estimation for each $\hat{g}_{vd}$ is then calculated k times, omitting the *k-th* group from each of the samples each time the estimation is rerun, producing k distinct versions of $\hat{g}_{vd}$ denoted $\hat{g}_{vd(k)}$. The values $\hat{g}_{vd(k)}$ are transformed into k pseudovalues as shown in Eq. 9. The mean of the pseudovalues, $\hat{\bar{g}}_{vd}$, is shown in Eq. 10. The standard error for $\hat{g}_{vd}$ is then estimated as shown in Eq. 11. This estimates the error for the estimated number of people leaving every village v at each d.

Equation 9

$$\hat{g}_{vd\alpha} = k \cdot \hat{g}_{vd} - (k-1) \cdot \hat{g}_{vd(k)}$$

Equation 10

$$\hat{\bar{g}}_{vd} = \frac{1}{k} \sum_{\alpha=1}^{k} \hat{g}_{vd\alpha}$$

Equation 11

$$SE(\hat{g}_{vd}) = \sqrt{\frac{\sum_{\alpha=1}^{k} (\hat{g}_{vd\alpha} - \hat{\bar{g}}_{vd})^2}{k \cdot (k-1)}}$$

It would be convenient to be able to make a single statement about the overall error in this entire matrix of 122 places mapped across 37 times. By plotting the errors estimated by the jackknife process against the estimated number of people leaving from that place at that time, the relationship of the error to the estimates in $\hat{g}_{vd}$ can be shown visually. Furthermore, the computed regres-

sion coefficient is the coefficient of variation, that is, the mean ratio of the error to the estimate. These results are in Graph A5.

Lines are plotted through the error-estimate points. There are different rates of error for smaller estimates and for larger estimates: larger estimates tend to have proportionally less error. The coefficients are presented below in Table A1.

These coefficients represent the standard error as the mean proportion of the estimated counts, and they can be interpreted as relative standard errors. For example, estimates of fewer than 5,000 people[17] leaving home during any given 2-day period fall on average inside a 95% confidence interval of ±(1.96 * 0.109) = ±22% of the estimate. However, estimates of 5,000 or more people leaving home fall on average inside a 95% confidence interval of ±(1.96 * 0.056) = ±11% of the estimate. Errors of this size are unlikely to affect the overall interpretation of any given phase, especially since most of the substantive interpretation is driven by the largest flows (which are subject to the smaller errors). However, all interpretations should be cautious, given these margins of error.

Graph A5: Error plotted against the estimates of people leaving their homes[16]

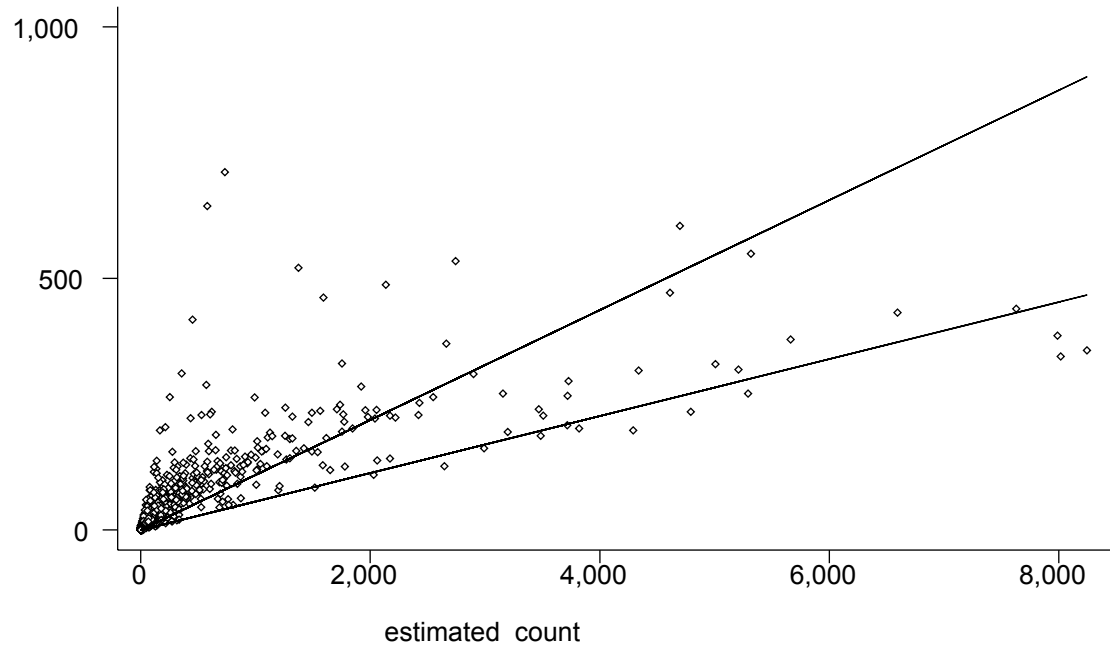


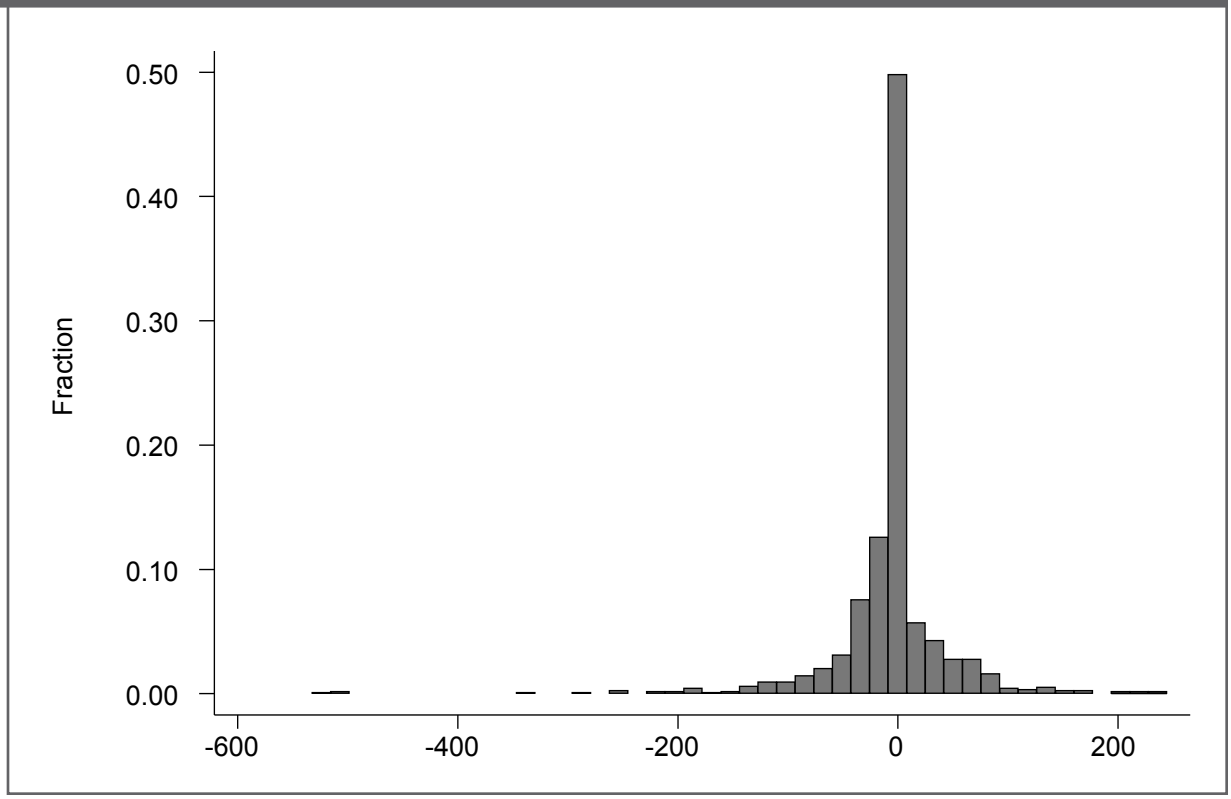
Table A1: Regression coefficients of jackknifed standard errors against estimated counts $\hat{g}_{vd}$ (all coefficients are significant at the 0.01 level)

	<5,000 people	>5,000 people
Slope coefficient	0.109	0.056
Measure of fit (R^2)	0.65	0.93
Number of points	1181	10

A5 Ratio bias

In the computations described above, there are several times a vector of proportions is multiplied by a scalar count to produce a vector of estimated counts. First, the imputation of data on people crossing the border with missing place information ($b^a_{\bullet d}$) into estimated numbers of people leaving each village on each day ($\hat{b}^A_{vd}$) uses a vector of observed sample proportions among municipalities (w_{ud}) computed from a sample of camp residents who were not matched to the border records (S^L_{ud}). And second, the projection of numbers of people crossing the border ($\hat{b}_{vd}$) to estimated counts of people leaving their homes ($\hat{g}_{vd}$) uses a vector of observed sample proportions among transit times

Graph A6: Histogram of values of estimated biases of $\hat{g}_{vd}$



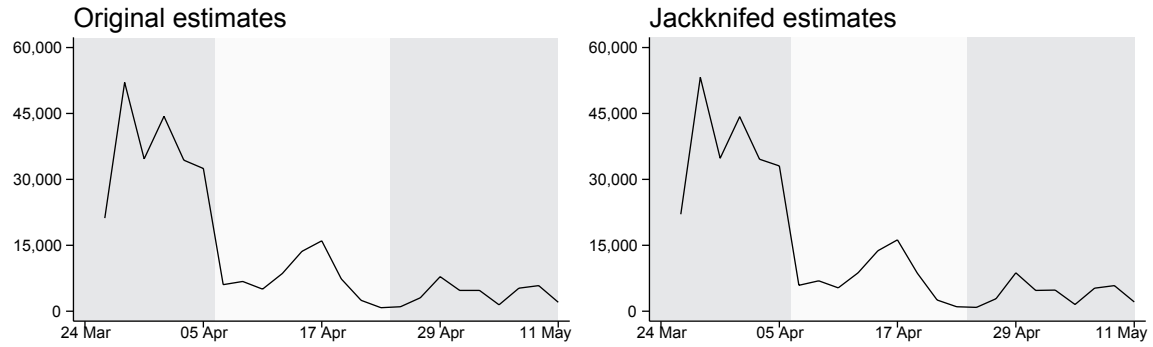
(w_{dt}) calculated from a sample of camp residents who were interviewed about the process through which they left their homes and came to Albania (S^{I}_{udt}).

The use of proportions in these two calculations admits the possibility that ratio bias could affect the estimations.[18] Ratio bias results from covariance between the values of the numerator and the denominator: if outflow from a block of municipalities tended to move consistently across time, then the totals from these municipalities might dominate the total number of people exiting Kosovo. This influence would create a correlation between the municipalities' total and the overall total, thereby creating bias when the municipality proportions are applied across time. Bias can be estimated as the difference between the original estimates of number of people leaving their homes ($\hat{g}_{vd}$) and the jackknifed estimates ($\hat{\bar{g}}_{vd}$) calculated in the previous section. Jackknifed estimators are free of first-order bias,[19] and therefore the difference between $\hat{g}_{vd}$ and $\hat{\bar{g}}_{vd}$ can be taken as approximating any ratio bias. The magnitude of this bias could be sufficient to affect the interpretation, and so it is examined below.

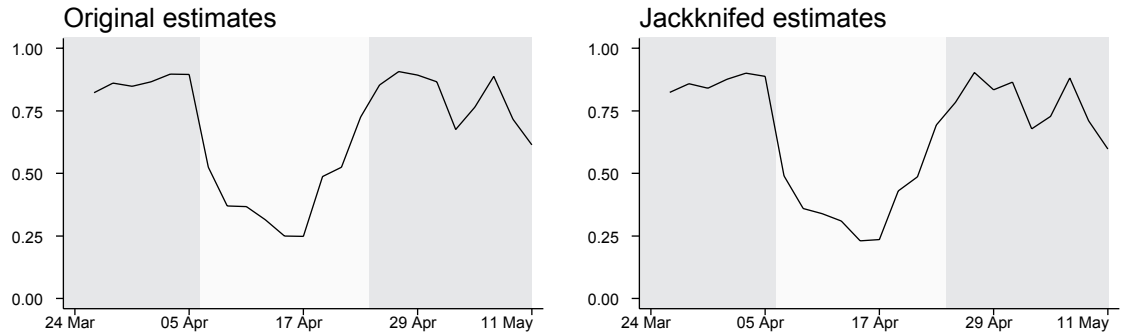
Although occasionally the estimated bias is large, the mean bias is –6 and the median is 0. This is a relatively small value considering that the scale of the standard error of the estimates is about five times larger (the mean of the standard errors of $\hat{g}_{vd}$ is 30) and the mean estimated number of people leaving any given village v on any given two-day period d is about fifty times larger (294). A histogram of the biases is presented in Graph A6.

Although the biases shown in Graph A6 seem to be distributed around zero such that they would have little effect in the overall interpretation, a more thorough look at the impact of ratio bias on the interpretation is to compare graphs from the original estimates ($\hat{g}_{vd}$) with graphs from the jackknifed estimates ($\hat{\bar{g}}_{vd}$). Because jackknifed estimators are free of first-order bias, this comparison tests the impact of the bias. If bias were significant, then the graphs from the jackknifed estimates would lead to different substantive interpretations of the

Graph A7: Number of people leaving home by 2-day period, original estimates and jackknifed estimates



Graph A8: Proportion of people who originated in southern and western Kosovo, by two-day period, as a fraction of the people leaving their homes, original estimates and jackknifed estimates



data. In fact, the graphs using the jackknifed estimates are barely distinguishable from the original graphs.

The differences between the left and right versions presented in Graphs A7 and A8 are nearly invisible. Apparently the individually estimated biases tend to cancel each other in the aggregate, and so the impact of ratio bias on these estimates is negligible.

A6 Stability of missing place imputations relative to concentrations of migrants from the southwest

If the camp residents' data were biased in some way, or if the data had substantial error unaccounted for above, using the camp residents' data to impute the missing place data would transfer the bias and error onto the border data. In this section, the potential bias and error in the camp residents' data are addressed in three ways.

A6.1. Inter-project consistency

Since the camp resident data were collected by two independent projects (PHR/Columbia and IPLS/AAAS), they can be compared to each other and to the border data to determine if in broad terms they agree. The collection processes of the three sources were very different: the border data were collected by an administrative system; PHR conducted systematic sampling in some camps; and IPLS undertook listing in camps that overlap but are not the same as the camps in which PHR worked. Because the collection processes were so different, is unlikely that the three sources share the same biases, so

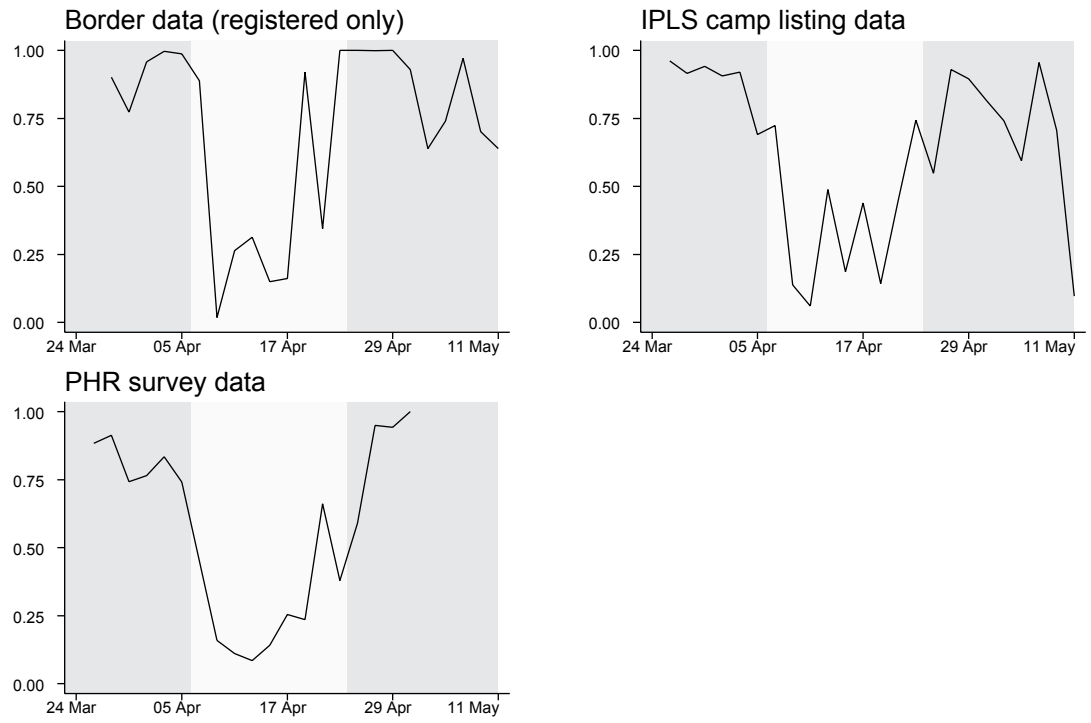
agreement between the sources would imply that the sources are relatively unbiased. Agreement would not constitute proof that the sources are unbiased, of course, but the data could fail this test if the sources had turned out to be very different from one another.

The proportions of Kosovars from the southwest municipalities, by two-day period, are presented below in Graph A9. The municipalities are listed in Appendix C with the regions (south and west, or north and east) to which they were assigned.

The three series are similar, but not identical. All three begin with a high (greater than 75%) concentration of people from the south and west; after 7 April, all three show a decline of the concentration of people from this region (mostly less than 40%), and then after 21 April show a return to a high concentration from the south and west. This broad similarity is also reflected in correlation coefficients presented in Table A2, below.

The hypothesis for this section was that if the three sources either share the same biases or are all relatively unbiased, they should be correlate with each other, and the finding is that the three sources do track each other closely. Table A2 shows that the proportion of refugees from southern and western municipalities calculated from three sources correlate strongly with each other,

Graph A9: Proportions of Kosovar migrants from the south and west in b_{vd}^{R}, IPLS camp listings, and PHR survey



although the sources based on surveys in refugee camps correlate more strongly with each other than they do with the border data.

Both camp sources were derived from samples of camp residents,[21] and so the next question is what effect sampling error may have had on the derivation of the distribution of origin places over time.

Table A2: Correlation coefficients of proportions of Kosovar migrants from the SW in three series (based on two-day counts between 24 March – 11 May 1999)[20]

Series	PHR	IPLS/AAAS
Border b_{vd}^{R}	0.72 (n=18)	0.70 (n=23)
IPLS/AAAS	0.86 (n=19)	

A6.2 Confidence intervals

This section considers the effect of sampling error on the imputation. Because the estimate of the proportion of migrants from the southwest is derived in part from the allocation process based on the camp sources, the estimated proportions could be treated as sample estimates, and nominal 99% confidence intervals could be drawn around each estimate. If the results are consistent even in the high or low scenarios indicated by the confidence interval, this would suggest that the findings are stable with regard to sampling error in the camp sources.

In order to derive confidence intervals around the estimated proportion of border crossers who were from the southwest, several assumptions must be made: about the nature of the samples, and about the applicability of the sampling error to the overall estimate. There are 2,078 interviews in S_{ud}^{L}, the list of camp residents composed by joining the PHR survey with the IPLS camp listing. Respondents who appeared in both lists were included only once, and respondents who were matched to the border registry were excluded. The reduced list S_{ud}^{L} has been taken to be reasonably representative of Kosovar Albanian refugees who were not registered at the border.

For simplicity in this sensitivity analysis, the jackknifed variances were not employed. Instead we acted as if the camp data were simple random samples of Kosovars crossing the border at Morina. There are three possible problems with this assumption. First, neither sample was balanced across camps by either a sampling strategy or a proportional distribution, and so there may be bias resulting from overrepresenting or underrepresenting some camp residents relative to others. Second, although the PHR sample was taken systematically from some camps,[22] the IPLS/AAAS data were gathered as complete listings in some camps and as small samples in other camps as the project strat-

egy changed, thus adding another component to the inter-camp non-sampling error. Third, and possibly of most importance, not all Kosovars who entered Albania were in camps: according to the UNHCR and as of late April, approximately two-thirds of Kosovars in Albania were in private accommodations.

The second assumption is that the sampling error affects the entire estimate of $\hat{b}_{vd}$. As discussed in the previous section, the sampling error affects only that portion of the data for which the place of origin information is missing and was therefore imputed ($\hat{b}^{A}_{vd}$). This missing-place portion of the data, $\hat{b}^{A}_{vd}$, is approximately half of the total data ($\hat{b}_{vd}$). The binomial calculation of the sampling error (defined below) will be applied to the entire estimate, and therefore it will tend to overstate the level of real sampling error by approximately a factor of two. The overestimate of the error in the binomial calculation partially offsets the hidden non-sampling error discussed in the previous paragraph. The sampling errors shown in Graph A10 below are calculated by the binomial formula shown in Eq. 12. In the notation developed earlier, $\hat{p}$ is the proportion w defined as shown in Eq. 13, and u is summed over the municipalities of the south and west (see Appendix C). The time d is aggregated by the standard two-day periods. The sample size n for each $w_{(SW)d}$ was the number of interviews (out of the 2,078 interviews in S^{L}_{ud}) for which the respondent entered Albania during d. Multiplying each $SE(w_{(SW)d})$ by 2.33 (the tabled z-value for a 99% confidence interval) and adding and subtracting the result to $w_{(SW)d}$ gives the confidence intervals shown in Graph A10.

Equation 12

$$SE(\hat{p}) = \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{(n-1)}}$$

Equation 13

$$w_{(SW)d} = \frac{\sum_{u} s_{ud}}{s_{\bullet d}}$$

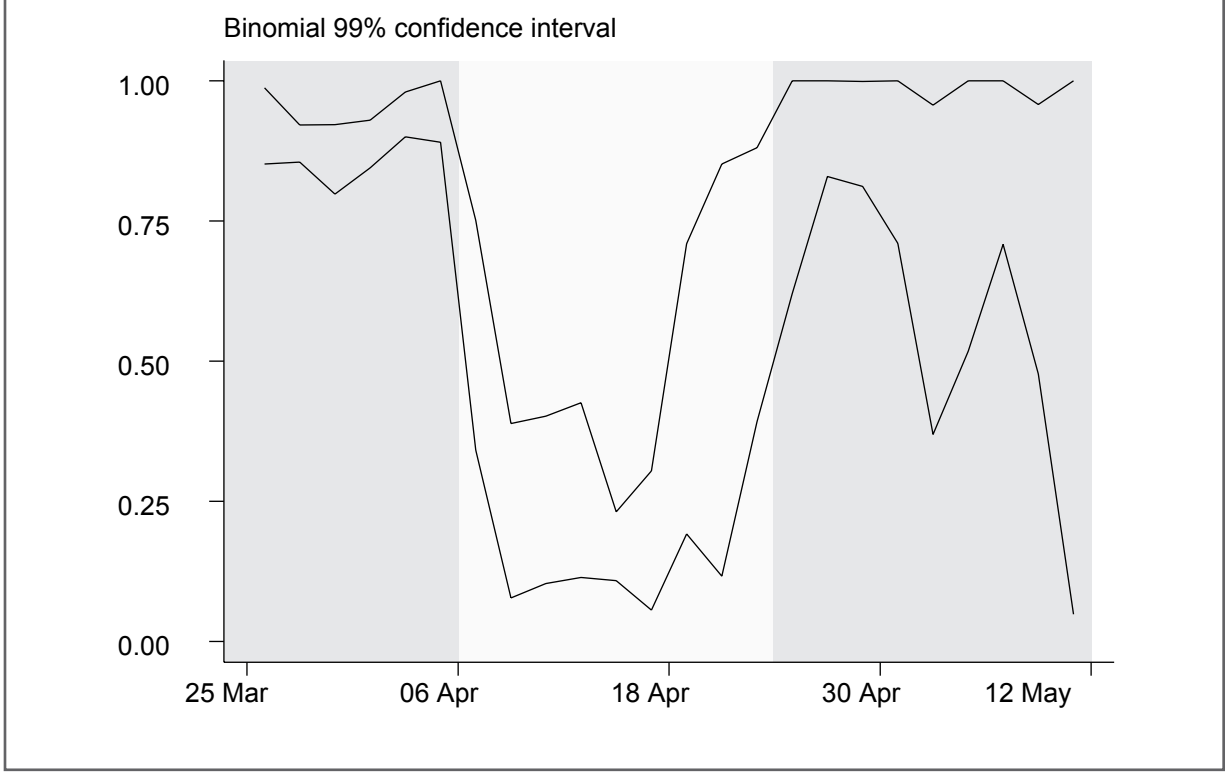
The results shown in Graph A10 make clear that sampling error – taking into account the assumptions described above – does not significantly affect the interpretation. During the period of substantial interest (late March to early May), the high and low bounds of the confidence interval are consistent with the interpretation given in the text. The upper and lower bounds of the nominal 99% confidence interval occasionally show patterns slightly different from the trend described by the point estimate. However, the overall pattern of the proportion of refugees from the south and west across time (high, low, high) is the same for any combination of the higher and lower bounds of the interval.

During the later period there are relatively fewer interviews because many fewer Kosovars were exiting after Phase 3 ended on 11 May. Fewer interviews were conducted with people exiting during the later period because there were fewer people exiting from whom to sample. The smaller number of interviews result in much higher levels of sampling error for estimates during this period.

A6.3 Robustness to noise

This section examines the sensitivity of the findings to random noise. By introducing noise into the imputation, we can consider if the estimated proportion of migrants from the southwest would change substantially if the imputa-

Graph A10: Estimated proportion of Kosovar Albanians entering Albania from southwest municipalities, by two-day periods, with nominal 99% confidence interval for sampling error



tion process were done in some other way.

If the results of the analysis were an artificial product of the process by which the missing data were imputed to municipalities using the camp sources, then imputing the missing data to municipalities by some other process could result in different findings. If, however, other reasonable processes produce substantially the same findings as the camp source imputation, then the conclusion is that the results are not the product of the imputation process. Indeed, if the results do not change much, the imputed results may be taken as robust.

The simplest way to look at this conjecture is to assign some fraction of the missing data border crossers to municipalities at random. This process essentially adds random noise and considers the robustness of the findings to this noise. Recall that the number of people from each village v crossing the border on day d was estimated as $\hat{b}_{vd} = \hat{b}^{A}_{vd} + b^{R}_{vd}$, where b^{R}_{vd} were the people registered by the border officials and $\hat{b}^{A}_{vd}$ were the people imputed to village v by a two-step process (from raw counts to municipalities, then within each municipality to villages) described in a previous section. The first step – imputing raw counts to municipalities – is the point at which this sensitivity analysis is applied.

The distribution of people crossing the border across municipalities on a given time d is calculated as $\hat{b}^A_{ud} = w_{ud} \cdot b^A_d$, where w_{ud} defines the proportion of sample data for each municipality u with respect to each day d as shown in Eq. 14. The noise to be added is a random distribution c_{ud} where, for each d, c_{ud} summed over $u = 1$ (see Eq. 15). The amount of noise to be added is α where $0 < \alpha < 1$. The distribution of people crossing the border across municipalities with noise added is estimated as shown in Eq. 16. The results are added to b^R_{vd} to get an estimate of $\hat{b}_{vd}$ with noise added. The results of setting α = 0%, 20%, 50%, and 80% are shown in Graph A11.

Equation 14

$$w_{ud} = \frac{s_{ud}}{s_{\bullet d}}$$

Equation 15

$$\sum_u c_{ud} = 1$$

Equation 16

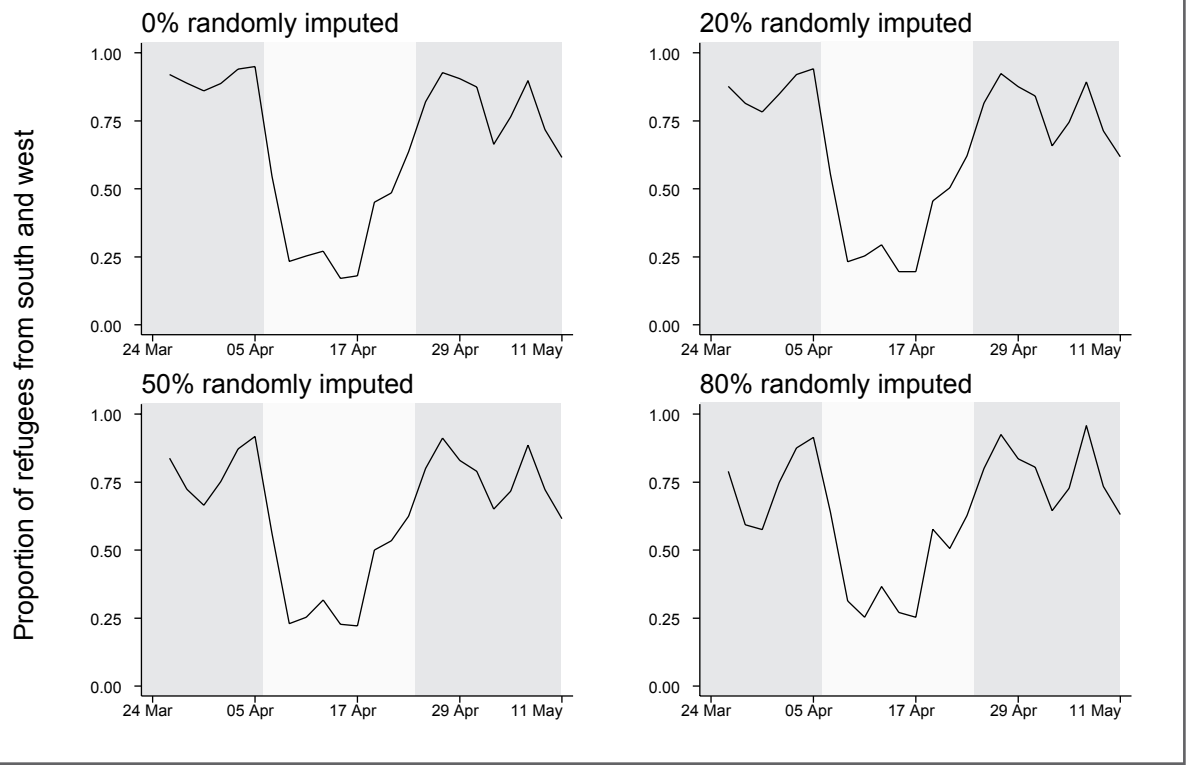
$$\hat{b}'^A_{ud} = (1-\alpha)\cdot(w_{ud}\cdot b^A_d) + \alpha\cdot(c_{ud}\cdot b^A_d)$$

Equation 17

$$0.52 < \frac{b^A_d}{b_d} < 0.91$$

The alternative formulations shown in Graph A11 are barely different from the original estimate, largely because the noise affects only that part of the data for which the place of origin is missing ($\hat{b}^A_{vd}$). As more noise is added to the allocation of the missing data, the only substantial change comes in a weakening of the finding for the very first days of the conflict (27 March - 1 April). During this week, the proportion of border crossers with unknown places of origin was relatively high (see Eq. 17). But soon thereafter, as the border guards improved their registration process in order to manage the massive flow of refugees and as the flow diminished, the proportion of unregistered refugees dropped to less than 30% of the total border crossers.

Graph A11: Proportion of Kosovar Albanian border crossers originating from municipalities in the south and west of Kosovo, with random noise added



The early period of disorganized registration left the data less stable than at other periods. As random noise is added to the data in this early period (in the graph above), the proportion of migrants from the south and west decreases relative to the interpretation made with data imputed without added noise.

What is remarkable about these graphs is the stability of all the estimates outside this early period. For example, the crucial early-April period – the period of the heaviest refugee flow of the entire conflict – the data are very robust to adding random noise of as much as 80% of the data for which origin places are missing. Even at this level, there are no substantial changes in the interpretation of this graph.

A7 Response of $\hat{g}_{vd}$ to changes in the distribution among transit times

For each time d, the distribution w_{dt} defines what proportion of people crossing the border left their homes on each of t days prior to crossing the border (because the model is estimated separately for each period d, the d subscript will not be used in this section). The subscript t denotes a range of transit times $0 \le t \le 70$. Refugees for whom $t > 70$ are out of range because although they crossed the border during the period of analysis, they left their homes before the beginning of the first phase on 24 March.

When the distributions w_{dt} were shown above in Section A2 on the transformation of $\hat{b}_{vd}$ to $\hat{g}_{vd}$ (see Graph A2), the exponential distribution was plotted by a line across the observed distributions. The point of showing the exponential distribution, as stated earlier, was to demonstrate that the exponential does not fit the observed distribution. The best fit for the observed distribution has both an exponential component and a linear component. The distribution f_t models w_t and is defined as shown in Eq. 18, where t is the number of transit days for which a proportion is being predicted; a is the average transit time for people leaving in six or fewer days as shown in Eq. 19; w is the cumulative proportion of the people leaving in six or fewer days as shown in Eq. 20; and p is the point k in w defined by Eq. 21. A fourth parameter, r defines the proportion of the total refugees whose transit time was greater than 70 days and should therefore be dropped from the modeling. People with very long transit times were kept in the analysis, but they do not appear in considerations of people leaving their home because people with long transit times left their homes before 24 March. Each distribution f_t is then normalized to sum to 1.00. Each of these statistics was computed for all three phases, and the results are presented below in Table A3.

Given the statistics a, w, p, and r, the parameter α was computed for each phase by minimizing the summed error between the estimated and observed distributions. The observed and modeled distributions, denoted w_{dt}^{O} and w_{dt}^{M} respectively, are shown in Graph A12.

Equation 18

$$f_t = (1-\alpha)\left(\frac{1}{a}\right)\exp\left(\frac{-t}{a}\right)+\alpha\left(\frac{1-w\ -0.01}{p\ -6}\right)$$

Equation 19

$$a\ = \frac{1}{6}\sum_{t=0}^{6} t \cdot s_t$$

Equation 20

$$w\ = \sum_{t=0}^{6} w_t$$

Equation 21

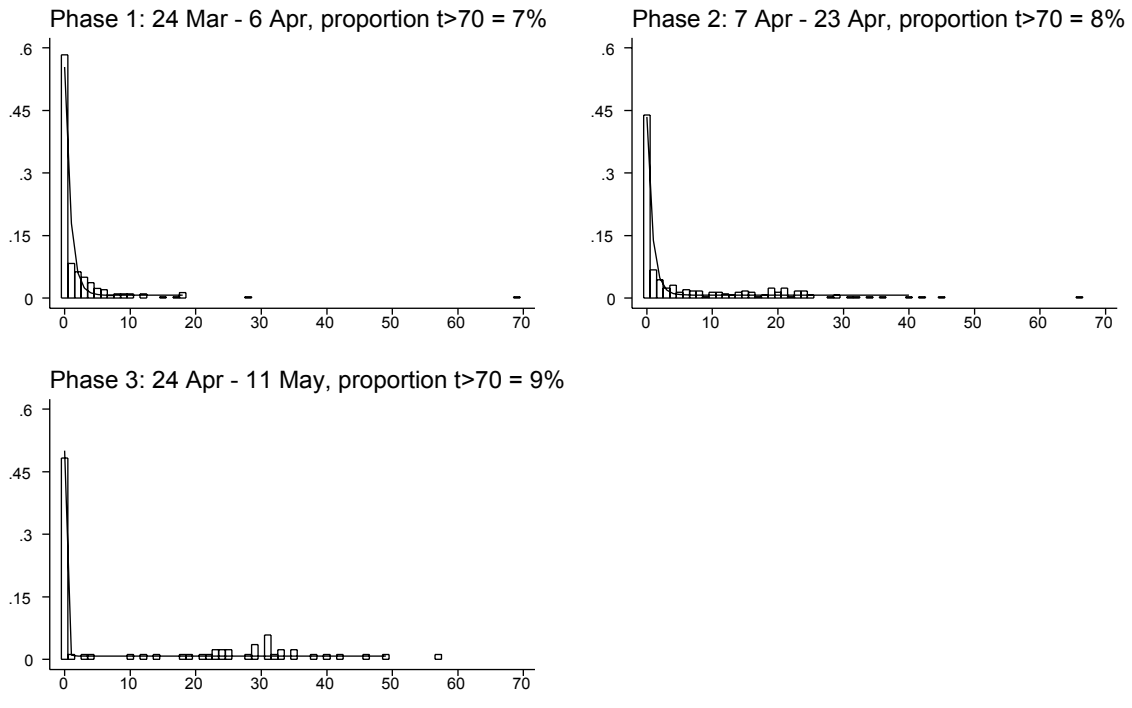
$$p\ = \sum_{t=0}^{k} w_t \ge 0.9$$

Table A3: Estimated parameters for f_t, with standard errors in parentheses[23]

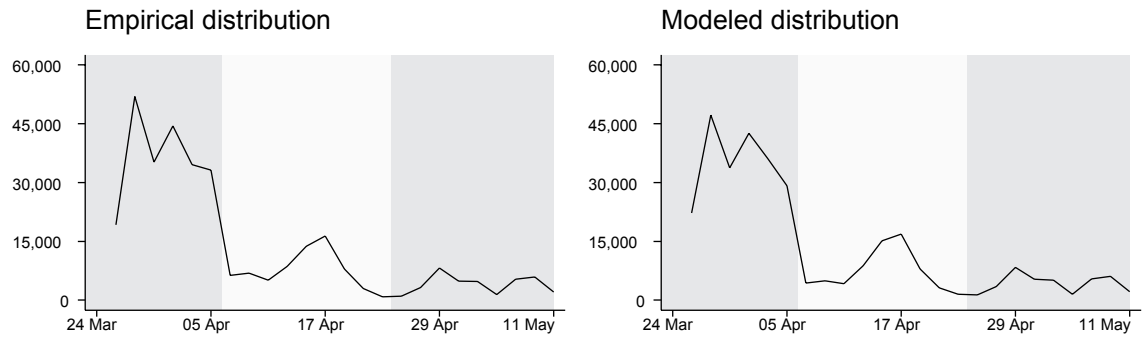
Phase	*a*	*w*	*p*	*r*	α	*n*
1	0.864 (0.098)	0.921 (0.016)	18 (0.987)	20 (4.291)	0.69	280
2	0.846 (0.112)	0.696 (0.026)	40 (7.731)	24 (4.692)	0.69	270
3	0.182 (0.115)	0.571 (0.054)	49 (4.158)	8 (2.655)	0.90	77

The modeled distributions w_{dt}^{M} are quite similar to the observed distributions w_{dt}^{O}. For both the observed and the modeled distributions, the largest number of refugees are in transit for zero days. Rapidly dropping proportions of refugees are in transit for 1-6 days. A small proportion of refugees are in transit for longer periods. The two versions of the distribution among transit times can be combined with $\hat{b}_{vd}$ to generate alternative versions of $\hat{g}_{vd}$. The distributions among w_{dt} are defined without regard to refugees' region of origin (there is no v subscript), and therefore changes in w_{dt} would not affect the distribution of refugees among places of origin. However, changing w_{dt} might change the pattern of people leaving their homes summed across origin points ($\hat{g}_{\bullet d}$). The two

Graph A12: Proportional distribution of number of groups entering Albania, by length of time in transit and phase during which they entered Albania, observed (bars) and modeled (line)



Graph 13: Number of people leaving their homes, by 2-day period, estimated by observed and modeled distribution of transit times



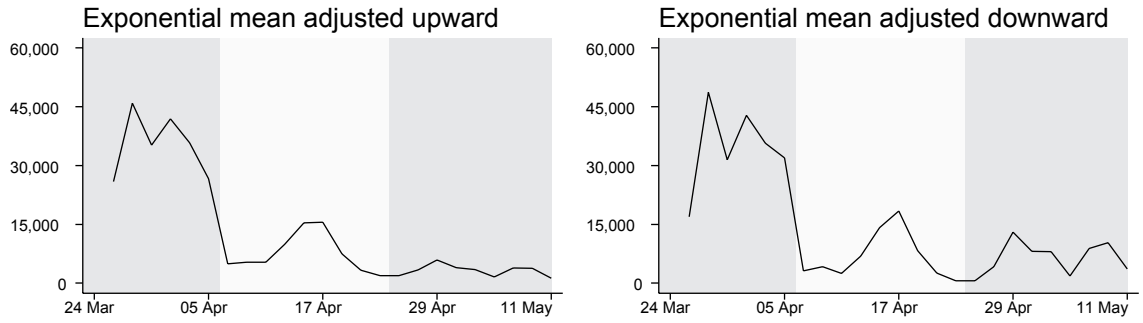
versions of $\hat{g}_{\bullet d}$ that result from using the empirical and modeled distributions are shown in Graph A13.

Graph A13 indicates that the estimated $\hat{g}_{\bullet d}$ differs only very slightly when different versions of w_{dt} are used. Relative to the $\hat{g}_{\bullet d}$ estimated with the observed distribution (w_{dt}^{O}) shown on the left of Graph A13, in the version of $\hat{g}_{\bullet d}$ estimated using the modeled distribution (w_{dt}^{M}) the peaks in Phase 1 are slightly lower, and the peaks in Phase 3 are slightly more pronounced. The phase structure defined by the sharp drops on 6 and 24 April is nearly identical between the two versions of $\hat{g}_{\bullet d}$. The similarity of the two versions means that the model *f* can be used to test the sensitivity of $\hat{g}_{\bullet d}$ to changes in w_{dt}.

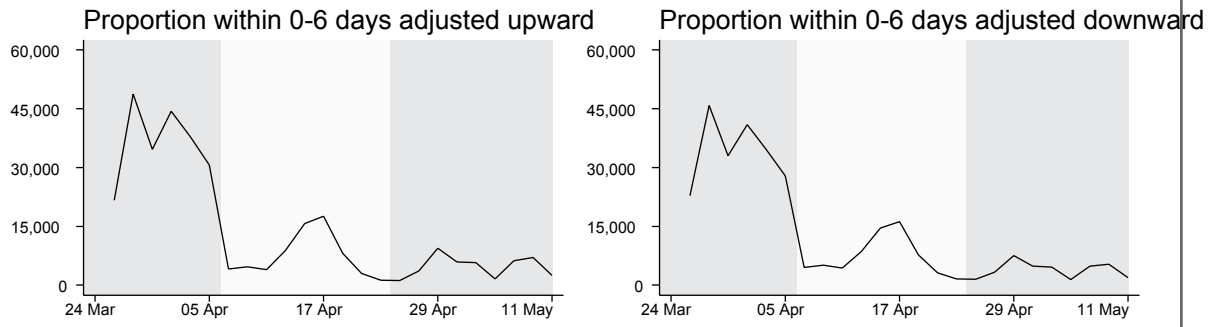
Using nominal 99% confidence intervals and the standard errors in Table 1, reasonable ranges of possible alternative transit time distributions w_{dt} can be spanned with an upper- and lower-bound for each of the four statistics in Table A3. In Graphs A14, A15, A16, and A17 below, $\hat{g}_{\bullet d}$ is shown using eight alternative versions of w_{dt}. With each of the four parameters in Table A3, two alternative w_{dt} were estimated by multiplying the standard error by 2.33 and adding or subtracting the result to the estimated parameter.

None of the alternative versions of the departure patterns shown in Graphs A14-A17 differs in substance from the departure pattern derived from the unadjusted w_{dt}. In fact, Graphs A14-A17 are barely distinguishable from the unadjusted $\hat{g}_{\bullet d}$. In Graph A14, the mean *a* is adjusted upward and downward, changing the concentration of people who left in 0-6 days at zero days. On the left graph in A14, *a* was shifted upward (by adding 2.33 times its standard error). Increasing *a* models the condition that fewer people were projected to have left in zero days relative to people leaving in 1-6 days; the right graph in A14 tests the opposite, that more people left in zero days. The effect is most obvious in the right tail of each graph: Phase 3 in the left graph in A14 is much smoother than Phase 3 in the right graph. But even in the left graph of A14,

Graph A14: Number of people leaving their homes, estimated with alternative *a*



Graph A15: Number of people leaving their homes, estimated with alternative *w*



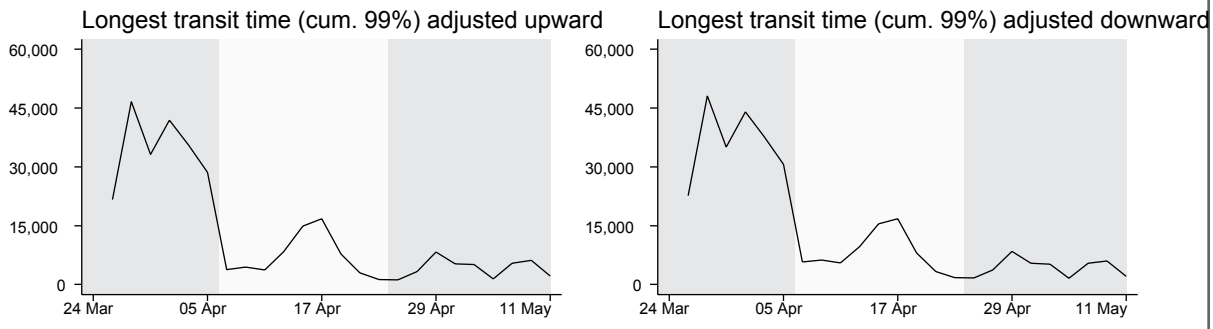
Phases 1 and 2 remain clearly distinguishable, although Phase 3 becomes less clear.

The differences between the right and left graphs in A15, A16, and A17 are nearly invisible. After combining the upward effects into a single estimate, and the downward effects into a single estimate, the resulting versions of unadjusted $\hat{g}_{\bullet d}$ look no more different from each other than those in A14 (this graph is not presented). In conclusion, the estimated number of people leaving on each day d, unadjusted $\hat{g}_{\bullet d}$ is not sensitive to changes in w_{dt} that span the reasonable range of possible distributions.

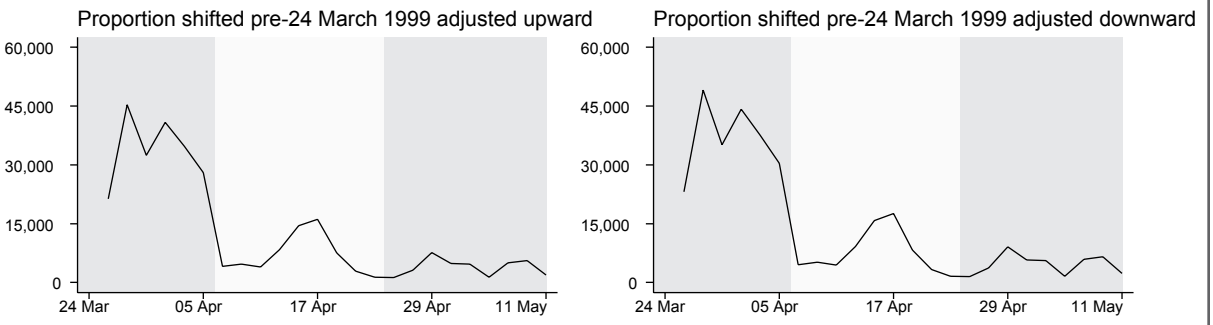
A8 Stability of findings relative to refugees exiting to countries other than Albania

Throughout this report, refugee movements in Kosovo have been analyzed using data only from the Albanian border point at Morina. We have asserted, however, that these conclusions would be unchanged if we added data from

Graph A16: Number of people leaving their homes, estimated with alternative *p*



Graph A17: Number of people leaving their homes, estimated with alternative *r*



Kosovar migrants who entered other countries. In this section we show that the patterns of Kosovar migrants who entered the other countries were substantially the same as the pattern of migrants entering Albania. It is our view that the conclusions presented in this report would be unchanged even if data from, for example, the Former Yugoslav Republic of Macedonia were introduced into the analysis.

This analysis should not be overemphasized because nearly half of all Kosovar refugees passed through Morina, and substantially more than half of all of the refugees were in Albania more generally. In order for the differences in patterns of Kosovar Albanian refugees exiting to other countries to affect the findings in this report in some drastic fashion, the differences would have to contradict the patterns in the Albania data.

There are two levels on which data from other countries might be similar (or dissimilar) to data from Albania: in terms of the relative magnitude over time, and in terms of the relative proportions of migrants from various regions of Kosovo over time. Both considerations will be considered below.

A8.1 Comparisons of relative magnitude of Kosovar Albanian entrants over time for four countries

The 6 Dec. 1999 OSCE report presented data on the entry of Kosovar Albanian refugees into Albania, Macedonia, Bosnia-Herzegovina, and Montenegro. The data are aggregated across some periods (e.g., 24-30 March, 15-21 April) and presented by day for other periods. This presentation style makes it difficult to compare the relative magnitude of numbers of migrants across periods directly. The AAAS/IPLS estimate of the total number of border crossers is presented in the first column; the OSCE figures are presented in columns 2-5 of Table A4.

From these data it is difficult to establish the flow patterns for Bosnia-Herzegovina, Macedonia, and Montenegro because the UNHCR/OSCE data presented in this table are imprecise. Furthermore, the numbers presented in the OSCE report, though attributed to the UNHCR, are not consistent with UNHCR's contemporaneous reporting in daily press briefings during March and April. It is clear that the analysts who constructed the table for the UNHCR/OSCE took total numbers at particular dates and then averaged the differences across intervening dates. This process smooths (and obscures) the variation which would indicate whether or not the flows into these countries actually track each other.

The OSCE estimates are higher than the estimated entry through Morina for 31 March, and for 2-4 April. The IPLS/AAAS estimates in Table A3 include only people crossing through Morina, whereas the OSCE estimates apparently include people who crossed through the border points at Trepoye and Krume. This can only be assumed because the OSCE table is not explicit on this point.

All countries show heavy flows in the early period before 3-5 April, and then lower flow levels after that. In late April, flows seem to increase again, but since the data are aggregated by week, it is hard to determine if these peaks and valleys track the phased structure suggested in the text. The data in Table A4 are suggestive of rough agreement with the Albania data, but they may be judged inadequate for stronger conclusions.

A8.2 Pattern of origin municipalities over time: Macedonia and Bosnia

The second manner in which the Albania data might be compared to other countries' patterns is by analyzing the proportion of refugees from the southwest who entered Macedonia over time, and comparing that proportion to the proportions found in the three Phases of migration into Kosovo. If the full data from the Macedonian border guards were available, this analysis could parallel that done for Albania. However, the only data from Macedonia available for this analysis are 540 interviews done by PHR as part of their survey. The proportions are calculated by the same methods described earlier for Graph A4,

Table A4: Daily number of refugees entering four countries, according to AAAS/IPLS and OSCE estimates[24]

Date	AAAS / IPLS Albania (Morina only)	OSCE Abania	OSCE Montengro	OSCE Bosnia-Herzogovina	OSCE FYR Macedonia
Phase 1					
24–30 Mar	84,500	60,000	7,500	3,000	4,500
31 March	11,100	25,000	2,500	1,000	7,500
1 April	20,800	0	5,000	1,000	15,000
2 April	31,251	34,500	10,000	1,000	0
3 April	18,577	62,000	2,000	1,000	65,000
4 April	21,808	47,000	2,000	1,000	1,000
5 April	24,111	7,813	1,000	2,000	1,125
6 April	24,448	7,812	3,000	1,000	1,125
Total Phase 1	*236,201	244,125	33,000	11,000	95,251
Phase 2					
7 April	2,877	7,813	2,000	1,000	1,125
8 April	16	7,812	1,000	1,000	1,125
9 April	1,459	7,813	1,000	1,000	1,125
10 April	4,265	7,812	1,000	1,000	1,125
11 April	371	7,813	1,000	1,000	1,125
12 April	742	7,812	0	1,300	1,125
13 April	3,309	4,800	2,500	100	0
14 April	3,600	0	1,700	300	0
15–21 April	67,956	42,700	2,800	4,600	14,000
Total Phase 2	*83,461	94,375	13,000	11,300	20,750
Phase 3					
22–28 April	7,473	10,200	0	0	12,650
29 April-5 May	36,427	37,000	0	2,600	68,690
6–12 May	27,829	22,800	2,700	900	19,860
Total Phase 3	*67,511	70,000	2,700	3,500	101,200
Total Phases 1-3	387,183	408,500	48,701	25,801	217,201

* The phase totals for the IPLS/AAAS estimates for Albania are slightly different from the sum of the daily counts because the phase totals are calculated as the sum of the days that are *precisely* within the phase. To be consistent with the OSCE data (some of which is presented only by week), the phase timing is somewhat approximate. In the AAAS/IPLS column, exact phase totals are reported in order to be consistent with earlier presentations and with Table A5.

Note: at some points the number of refugees entering a country declined as the humanitarian evacuation program helped some refugees leave a particular country (especially Macedonia) in order to settle in countries outside the region. This process results in negative flows for those periods; the negative periods are represented here by zero.

and the top line in Graph A18 below is the same data as shown in the PHR graph in Graph A4. The proportions of survey respondents from the municipalities of the southwest for people who exited into Albania and Macedonia, respectively, are presented below in Graph A18.

Graph A18 shows the proportion in terms of the dates that refugees left their homes. The graph on the left shows the proportions for the two countries. The graph on the right shows the same line for Macedonia as on the left, but the line for Albania has been shifted down by subtracting 0.6 from each of the original values. It is clear from the graph on the left that for all periods (with the exception of 10-11 April), a higher proportion of people entering Albania are from the south and west than those who enter Macedonia. This is logical since Albania borders Kosovo's southwest and Macedonia borders Kosovo's southeast, and refugees would in most cases take the shortest route out of Kosovo.

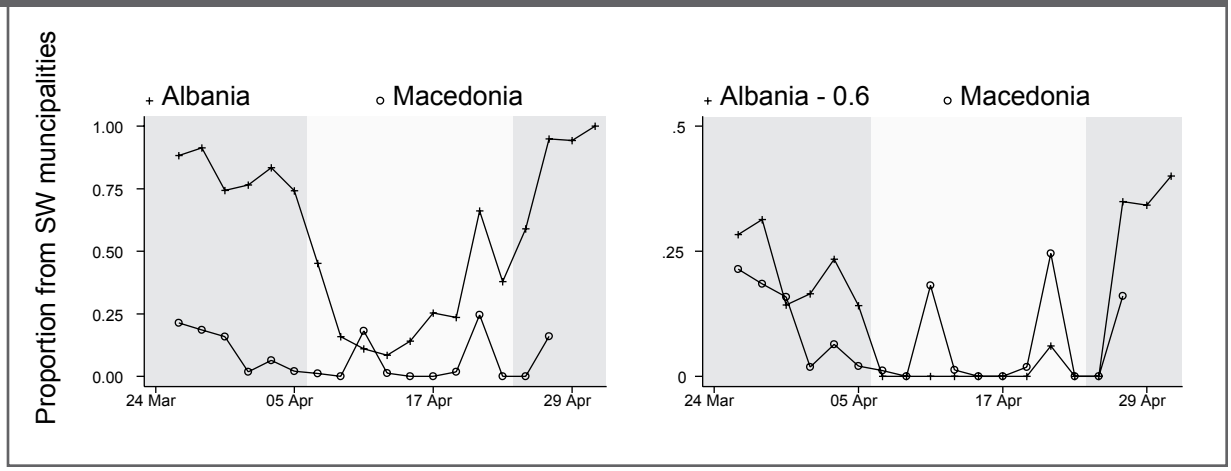
The pattern of ethnic Albanian refugees from southern and western municipalities of Kosovo entering Macedonia is roughly the same as the analogous pattern for Albania. From 24-30 March, the proportion from the south and west is at relative high points for both series, with the proportion for Kosovar Albanians entering Macedonia is greater than 16% for all six days. The proportion drops after 30 March for those entering Macedonia, though it remains high for those entering Albania for about another week. Both series show a minor local peak on 2-3 April.

There is an anomalous peak for the Macedonia data on 10-11 April. During this period, the proportion for those entering Macedonia is greater than for those entering Albania. The two countries' trends agree that the proportion of refugees from the south and west begins to rise on 20-21 April, but then both series decrease. The Macedonia percentage goes to zero during the transition to Phase 3 while the Albania data dip below 50% before rising toward 100%. The proportion rises on the last point of the Macedonia series, 26-27 April, as the series end.

There is only one two-day period (10-11 April) on which the two series move in contradictory directions. Over most of the period, the series do agree, as is clear from the comparison between the Macedonia and shifted Albania data in the right graph of Graph A18.[25] With the exception of 10-11 April, the Macedonia series shows the same high-low-high pattern as was observed for the Albania data. However, the anomalous point on 10-11 April and the consistently lower proportion for Macedonia leave open the question of whether adding the Macedonia data would fundamentally change the analysis.

There is one remaining dataset available with which to test this question, a survey of Kosovars in Bosnia. These data are very sparse: only 136 interviews, of which the respondents for nearly half left Kosovo much earlier than this analysis covers and so were excluded from the analysis.[26] The few remaining interviews provide a limited but interesting perspective. There is inadequate data for a graph, so the pattern will be discussed only in the text; the proportions were calculated by the same methods discussed for Graph A4.

Graph A18: Proportion of Kosovars from the SW municipalities exiting to Albania and Macedonia, by date they left their homes (PHR survey data)



Among Kosovar Albanians who went to Bosnia, the period of the highest proportion of people from the southwest is at the beginning of Phase 1, 25-30 March, when 56-80% of the people entering Bosnia were from southern and western Kosovo. In the survey data, no Kosovar Albanians from the south and west enter Bosnia again until 8-9 April, when 12% of the entrants are from the south and west. The data end at this point when routes from Kosovo to Bosnia apparently became too difficult for refugees to travel. The Bosnia data do not include sufficient respondents to draw firm conclusions, but they are consistent with the previous analysis of Phases that during Phase 1 a high proportion of refugees originated in the south and west, but that during Phase 2 a smaller proportion of refugees came from those areas. This pattern suggests agreement with the data from Albania.

A8.3 Pattern of origin municipalities over time: consistent and inconsistent scenarios

If data from Kosovar Albanian refugees entering Macedonia, Montenegro, and Bosnia were added to the Albania data, would the findings change? To evaluated this conjecture, plausible scenarios about the overall pattern can be generated using the data that is available on the timing and magnitude of migration out of Kosovo into neighboring countries and survey data from Kosovar Albanians in Bosnia and Macedonia.

Table A5 combines the analysis of Sections A8.1 and A8.2 to create the scenarios suggested in the previous paragraph. Using information about the relative magnitude of refugee flow over time from Bosnia, Macedonia, and Montenegro, multiplied by estimates of the proportion of refugees originating

in municipalities of the south and west, a combined estimate of the proportions of all refugees from the municipalities of the south and west can be produced for each of the three phases. All of the estimates are in part or in whole based on samples, and so nominal 99% confidence intervals are used to span the reasonable range of the combined estimates.

The total column on the right of Table A5 gives the result of the weighted sum of each of the country proportions. The totals are summarized graphically in Graph A19.

Table A5: Total number of Kosovar Albanians entering neighboring countrieswith two scenarios of estimates of proportion of migrants from the south and west, by period and destination country

High and low estimates of the proportion of Kosovar Albanians from the south and west, by Phase		Albania: IPLS/AAAS Project		OSCE/UNHCR estimates			Total
		Morina [note 1]	Trepoje/Krume [note 2]	Monte-negro [note 3]	Bosnia-Herzegovina [note 4]	FRY Macedonia [note 5]	
Phase 1	Total	236,201	85,000	33,000	11,000	95,250	460,845
	SW High	92.5%	92.5%	68.5%	44.5%	9.3%	72.4%
	SW Low	88.7%	88.7%	50.3%	11.8%	2.3%	66.2%
Phase 2	Total	83,461	0	13,000	11,300	20,750	129,645
	SW High	31.4%	31.4%	29.6%	27.7%	6.9%	26.9%
	SW Low	21.8%	21.8%	10.9%	0.0%	1.2%	15.5%
Phase 3	Total	67,511	0	2,700	3,500	101,200	179,129
	SW High	87.2%	87.2%	68.6%	50.0%	15.5%	44.7%
	SW Low	70.2%	70.2%	35.1%	0.0%	0.0%	27.6%

Note 1: The high and low estimates are nominal 99% confidence intervals from the binomial standard error resulting from the samples used for the imputation of missing data; see the section “Stability of missing place imputations: confidence intervals,” above.

Note 2: The high and low estimates use the same standard errors as the rest of the Albanian data. These estimates probably underestimate the proportion of people from Kosovo’s south and west because Trepoje and Krume would be inconvenient exit points for Kosovar Albanians from the north, east, or central regions. The total figures were given to this project by Albanian border officials in Krume in June, 1999, though we have reduced the number by 10,000 to account for people who entered before 24 March.

Note 3: There is no survey data from Montenegro. The high and low proportions were chosen by taking the mean of the data from Albania and from Bosnia-Herzegovina. Since the primary border point between Kosovo and Montenegro is west of Pec, it is logical that most of the people entering Montenegro would be from the west or northwest.

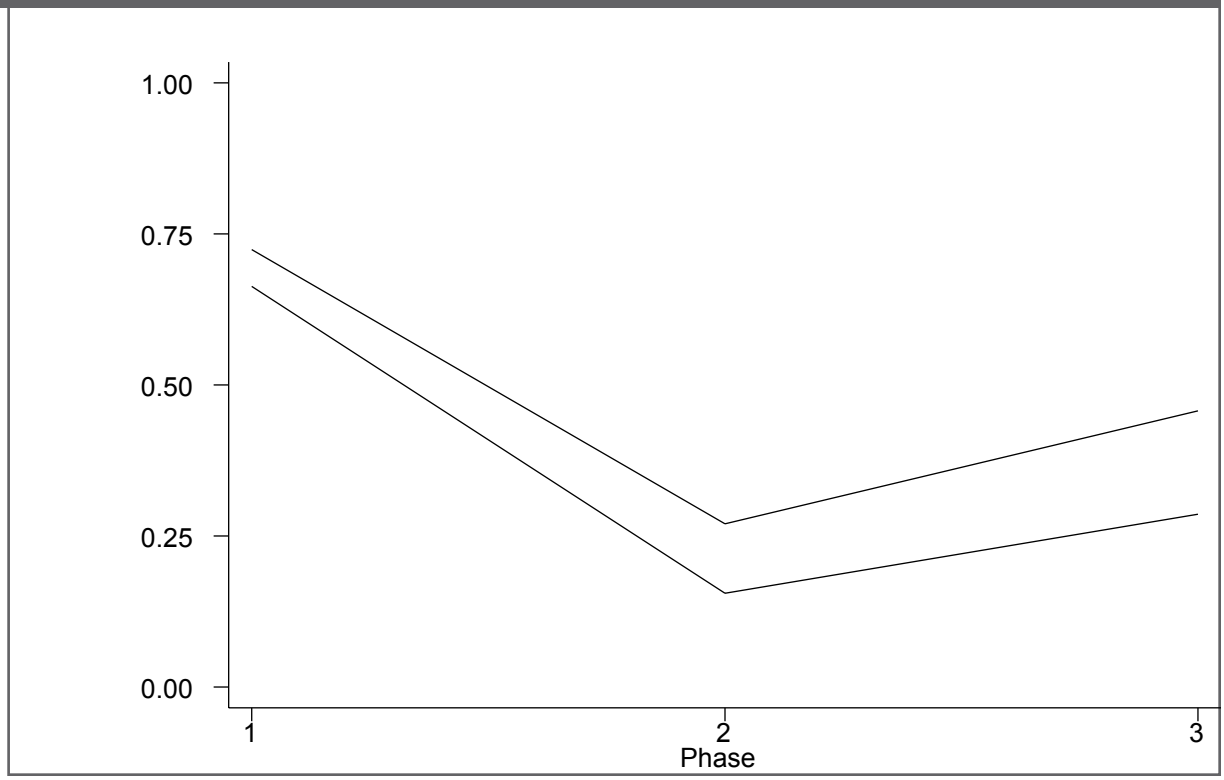
Note 4: The high and low estimates are the nominal 99% confidence interval from the binomial standard error calculated from the sample of Kosovar Albanians interviewed in Bosnia in June-July 1999 by the UCB/IPLS/AAAS project. There is no survey data for Phase 3, so the proportions are set very widely, but there were relatively few refugees entering Bosnia during this period.

Note 5: The high and low estimates are the nominal 99% confidence interval from the binomial standard error calculated from the sample of Kosovar Albanians interviewed in Macedonia by PHR in April-May 1999.

Adding data from Bosnia, Montenegro, and Macedonia lowers the overall effect, principally because fewer refugees from the south and west entered Macedonia at all points. Refugees entering Bosnia and Montenegro have only a small effect on the overall pattern because there were relatively few of them. The characteristic shape of the distribution (high – low – high) is retained, more so in the high scenario than in the low.

If the estimates of the proportions of Kosovar Albanians from the south and west were placed at the most inconsistent points (allowed by nominal measures of sampling error), the interpretation would only have to be weakened slightly. In particular, the proportion of Kosovar Albanians from the south and west exiting in Phase 3, taking into account all Kosovar Albanians exiting, may not be substantially higher than in Phase 2. For the maximum weakening of the explanation, however, the sampling errors would have to work to depress the estimated proportion for Phases 1 and 3 while working in the opposite direction to inflate the estimated proportion for Phase 2. The analysis of the Phase 1 – Phase 2 transition in the estimate of all Kosovar Albanian refugees remains consistent with the observation from the Albanian border estimates from Morina.

Graph A19: Estimated proportion of all Kosovar Albanians leaving Kosovo who are from southern and western municipalities, by phase and scenario, upper and lower bounds defined by nominal 99% confidence intervals



A8.4 Conclusion about generalizing from Albanian data to all Kosovar Albanian refugees during this period

In the previous three sections, the basic conclusions from the Albanian data have been evaluated using several partial sources on Kosovar migration patterns to other countries. The conclusion is based on the following empirical observations.

a) The data available for Macedonia yield findings suggestive of agreement with the findings for Albania.
b) Preliminary evidence from refugees in Bosnia and Montenegro also tends to support the Albanian analysis.
c) An overall analysis that combines data from Bosnia, Montenegro, and Macedonia with data from Albania shows that the combined analysis is not substantially different from the analysis using the Albania data alone.

Based on these observations, we conclude that the unequivocally strong and clear findings from Albania are generally applicable to the universe of all Kosovar Albanian refugees, within tolerances discussed in Section A4.

Notes for Appendix A

[1] These interviews were part of the pilot project for a much larger project. However, when the refugees returned to Kosovo in mid-June, the interview component of the project was cancelled.

[2] Used by permission of Human Rights Watch (HRW).

[3] See Physicians for Human Rights (PHR) 1999; these data used by permission of PHR.

[4] During discussions in early June, the border guards expressed this position quite vehemently to our project.

[5] There are also three days (14, 15, and 17 May) when most of the border records were lost and so the only available data is from the EMG; for 16 May, the data were lost for both the border and for the EMG. These days appear on the graph as three points along the bottom axis, near zero for the border count and between 4,000-8,000 on the EMG estimates. These days are not within our three Phases and so no additional analysis has been done for them.

[6] The computation of b^A is presented in this way to make obvious the various data components used. More directly, $b^A_{\bullet d} = b^U_{\bullet d} - b^R_{\bullet d}$.

[7] The matching process from camp residents to people registered at the border underrepresented the real match rate because the average group size at the border was larger than the families in the camps. Therefore several families may have been registered at the border with only one name, yet they would have disaggregated themselves in the camps. This means that some people in our camp data used to impute missing place data were actually registered at the border but cannot be traced to a particular border record. We did not resolve this problem but we do not believe that it is serious. The bias that might be introduced by including a few respondents who were actually registered is still far less than would be created by using b^R_{vd} directly. Furthermore, as discussed in Section A6.1, the differences between the camp data and b^R_{vd} are small.

[8] The mean number of people per family according to the IPLS interview data is about 7.15. Since people crossing the border and registered by name were sometimes registered in groups larger than just one family (with a mean of 10.5 and a standard deviation of 19), the mean group size for border crossers is larger than the average family size in camps.

[9] S^L_{ud} is sparse with respect to d, and so distributions among municipalities by d would be unstable because a small number of additional respondents in S^L_{ud} might substantially change the proportion of respondents in each municipality. To reduce the instability, the list was aggregated into counts of respondents for each 6-day period ($d[6]$) and municipality, creating counts of respondents for all u and $d[6]$. For simplicity, the notation will remain as d, but it should be

Equation 22

$$w_{vud[2]} = \frac{b^R_{vud[2]}}{b^R_{\bullet ud[2]}}$$

understood that any d can be mapped into a $d[6]$, and aggregation over changing time periods can be managed via this kind of mapping.

[10] The proportions were calculated with data aggregated to two-day periods as shown in Eq. 22. For simplicity, the time notation has been maintained as d.

[11] This analysis includes 753 interviews cumulated from the sources described in the text: 572 from PHR, 61 from the IPLS interviews, and 123 from the HRW interviews. Because the sources were matched to each other and the duplicates dropped, the individual project totals are greater than the final dataset used. Note that the unit of analysis for the distribution is groups, not individuals, because in the interviews it became clear that groups traveled together. The correlation between group size and transit time in the original data is 0.02 (n = 753), not significant at the 0.05 level, suggesting that group size is unrelated to travel time. This finding about group size is different from the border data, where group size did vary depending on the size of the daily flows.

[12] As in prior steps, the aggregation across time periods was the minimum necessary to smooth the distributions.

[13] We conducted Kolmogorov-Smirnov (KS) tests of equality of distribution functions across transit times on all pairs of regions, and two of ten were nominally significant at the 0.01 level (two more were nominally significant at 0.05). However, all three pairs of periods were nominally significant at the 0.01 level. The word "nominally" has been used here when speaking of the KS tests because the data handling problems, especially those due to missingness, that arose in the project made formal testing difficult. The use of nominal P-values is meant to convey not that we have a confirmatory outcome but that there is considerable evidence that the distributions do not vary meaningfully among regions.

[14] The author is grateful to the participants of a special seminar in the Department of Demography at the University of California, Berkeley (9 Nov. 1999) for pointing out that since the transit time distribution does not vary across regions there is no reason to include regional factors in the projection.

[15] See Kirk M. Wolter, *Introduction to Variance Estimation*, New York: Springer-Verlag (1986), pp. 154-156. Issues exist when using jackknife variance estimation in the presence of missing data; see, for example, J. N. K. Rao and J. Shao (1992) "Jackknife variance estimation with survey data under hot deck imputation," *Biometrika* **79**(4):811-22. This is one of the reasons nominal confidence intervals have been used. While the issues have not been fully addressed, we do not believe they impact the basic conclusions in this report.

[16] Note that place-times for which there are zero people leaving are omitted from this error analysis.

[17] The value of 5,000 as the cut-off was arbitrarily chosen as about the location where a block of points with an apparently smaller slope departs from the

cloud of points with lower values.

[18] On ratio bias, see Leslie Kish (1965) *Survey Sampling*, New York: Wiley, pp. 207-208.

[19] See Wolter, op. cit., p. 155.

[20] PHR's survey was completed in early May, and so there are no data points to compare from their series after 1 May. The border data only become available as of 26 March, and so the 24 March data point is only compared for the IPLS-PHR pair. The missing data were dropped from the pairwise correlations shown in Table A1.

[21] The samples aggregated for this analysis were mostly taken by probabilistic means, but some were complete camp listings and others were judgement samples taken quickly before camp residents returned to Kosovo.

[22] PHR used a "systematic" sampling scheme in which families in every *n*-th tent were chosen for interviews. See PHR (1999: 35).

[23] The errors were calculated using the jackknife method as described previously, but with $k = 75$. Note that the estimates for each phase were computed separately. That is, jackknifing was done for each phase randomly assigning cases from the sample data for that phase into *k* groups. Phase 3 has only 77 cases, so *k* could not be 100 as it was for the estimation of the overall error of $\hat{g}_{vd}$.

[24] The OSCE estimates are derived from a table in Part III, Chapter 14, of the 6 December 1999 report; the OSCE cites the UNHCR as the source of the figures.

[25] As a further test of the agreement between the proportions of refugees from the south and west entering Albania and Macedonia, the points in Graph A18 from 24 March to 30 April, excluding 9-10 April, were fitted in a regression equation: $\hat{w}_d^A = 0.43 + 2.08 \cdot w_d^M + e_d$, where the superscript "A" and "M" indicate the proportions for Albania and Macedonia respectively. The slope and intercept coefficients are significant at the 0.01 level, and the goodness-of-fit R^2 is 0.38. This result (including the moderate goodness-of-fit and the significant coefficients) is consistent with the hypothesis that the proportions of refugees from the south and west entering Albania and Macedonia tend to track each other across this period.

[26] There are some respondents who exited as late as 27 April, but the data are very sparse after 9 April.