Kosovo Data Analysis

Richard Blankenhorn

5/1/2017

1: Objective and Introduction to the Data

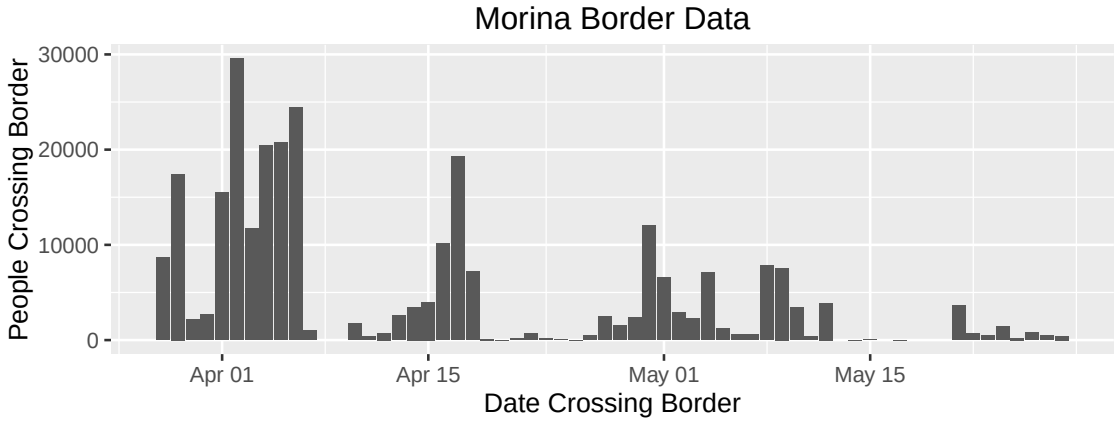
The objective of this analysis is to provide evidence for or against the argument that the NATO bombings and NATO involvement in the Kosovo conflict caused an increase in the number of individuals fleeing their homes and crossing the border. This analysis consists of several data sets from the HRDAT site which includes five migration data sets, a NATO airstrike data set and a Kosovo Liberation Army (KLA) activity data set. With the migration data sets, the variables of interest include the number of individuals being reported or interviewed, the date they crossed the border and the municipality they are from. With regard to the airstrikes and KLA activity data sets, the variables of interest are the number of strikes, attacks or casualties, the date of these events and the municipalities impacted for an event.

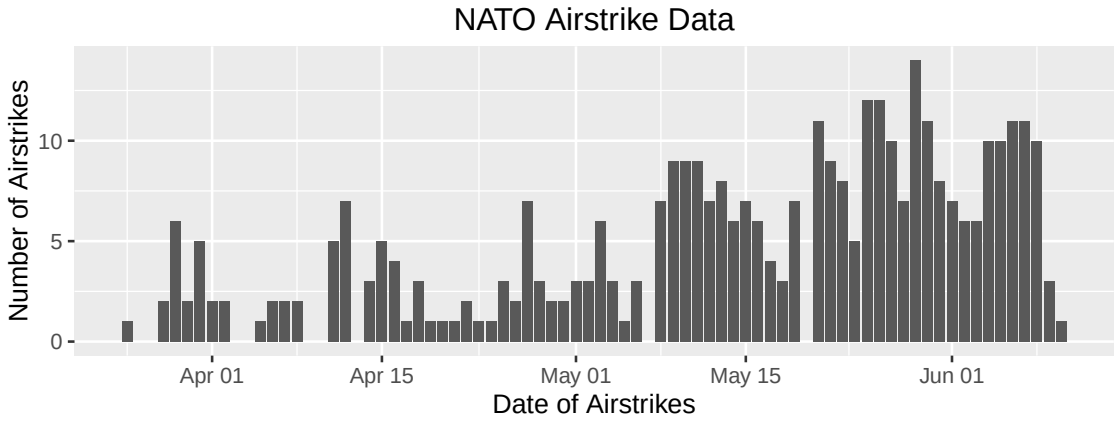
2: Exploratory Analysis

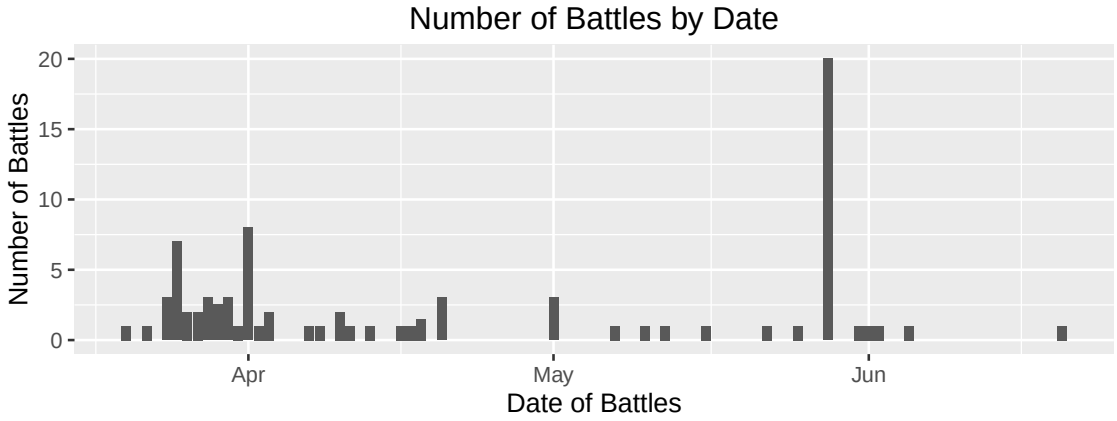
We begin by looking at the different migration data sets through frequency plots. The first relates to the Morina Border data set. This data set contains 19,126 records of which two observations were dropped since there is no date listed when crossing the border. This data set runs from March 28th of 1999 through May 28th of 1999 and contains the border ID, pcode (village code), date crossing the border and the number of people in the party. The border ID variable was dropped along with the pcode variable after a new *muni* variable was created to show the municipality from which the refugees fled. There were also a number of missing values for the pcode and these entries received a value of 0 for the *muni* variable. The distribution of the border crossings were plotted by the date of the migration. As can be seen by the Morina Border plot on the following page, there is a surge in migrations from late March through early to mid April. This happens again in late April and early to mid May but the trend is a steady decline.

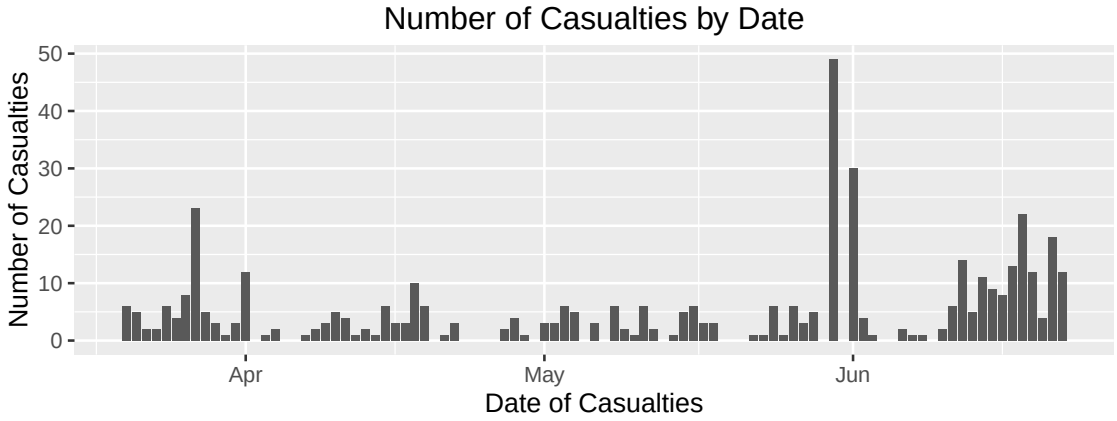
We also look at the NATO airstrike activity and the KLA activity. The NATO bombing data set contains information reported to individual airstrikes and there are 364 of them in total. These airstrikes only occur in Kosovo and the data set includes the bombing date, the municipality impacted and a description of the bombing. The KLA activity date set references the Kosovo Liberation Army activity maintained by the Kosovo investigation team. Each record in this data set represents an instance of either Yugoslav government casualties due to confrontation between the KLA and Yugoslav authorities or an exchange of fire between the Yugoslav forces and the KLA. The variables in this data set include the event date, type of event (battle or casualty), the municipality impacted and the number of such events. The variable *details* was dropped and the data was split in to KLA Battles and KLA Casualties. In addition, there were no missing value for these data sets.

The NATO airstrikes show someowhat of a different trend in that the majority of the bomings occur after mid May. There appears to be an inverse relationship between the number of individuals migrating and the number of airstrikes as time goes on. On the other hand, the KLA activity plot shows more activity in the months of late March and April. Here, there are two plots, one showing the number of casualties and another showing the number of conflicts or battles. Apart from a dramatic spike in battles and casualties in June, the majority of the battles appear to occur in late March and through April. To break down the impact of the airstrikes, battles and casualties even further, these figures are plotted by each municipality and displayed on page 3. Please note that the top 10 municipalities are reflected as these municipalities account for approximately 75% of the migrations in this data set. Also, the figures for airstrikes, battles and casualties are divisible by 500 (100 for KLA Casualties). This was done in order to reflect these events on the same plot as the migrations.

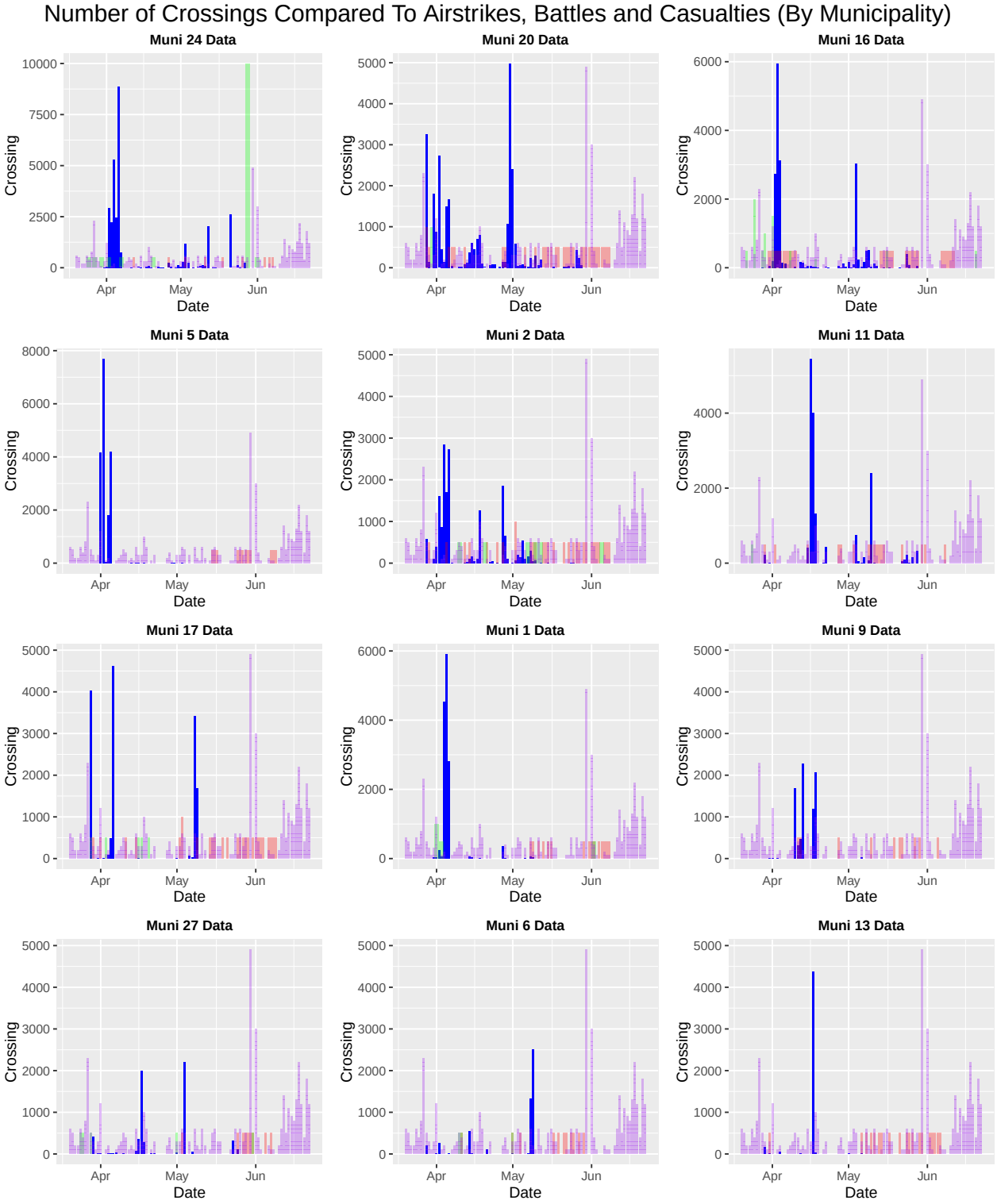

Morina Border Data
People Crossing Border
30000
20000
10000
0
Apr 01
Apr 15
May 01
May 15
Date Crossing Border


NATO Airstrike Data
Number of Airstrikes
10
5
0
Apr 01
Apr 15
May 01
May 15
Jun 01
Date of Airstrikes


Number of Battles by Date
Number of Battles
20
15
10
5
0
Apr
May
Jun
Date of Battles


Number of Casualties by Date
Number of Casualties
50
40
30
20
10
0
Apr
May
Jun
Date of Casualties

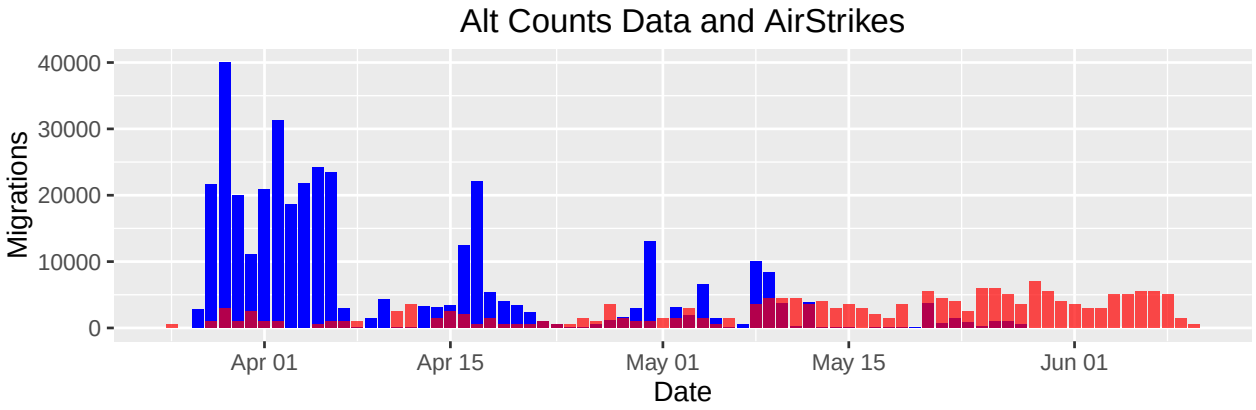
Migrations = Blue, Airstrike = Red, Battle = Green, Casualties = Purple

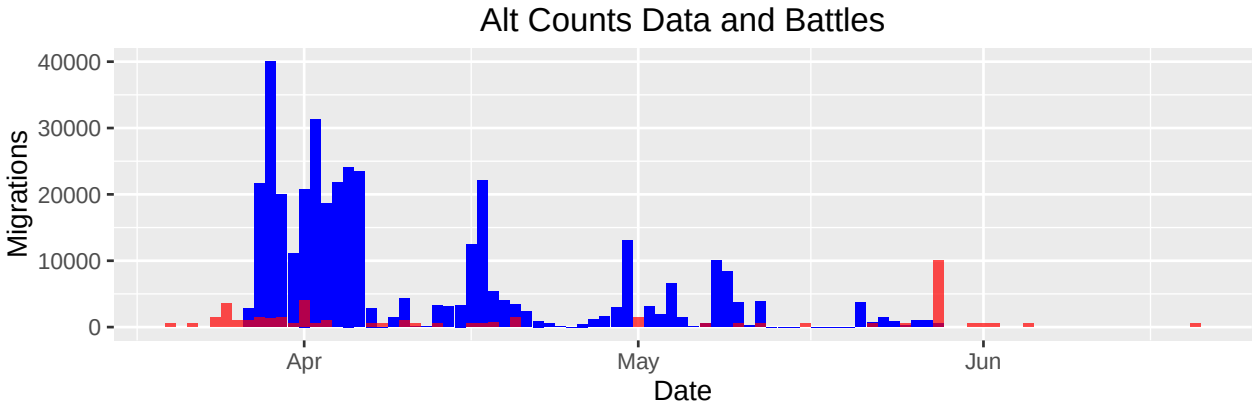


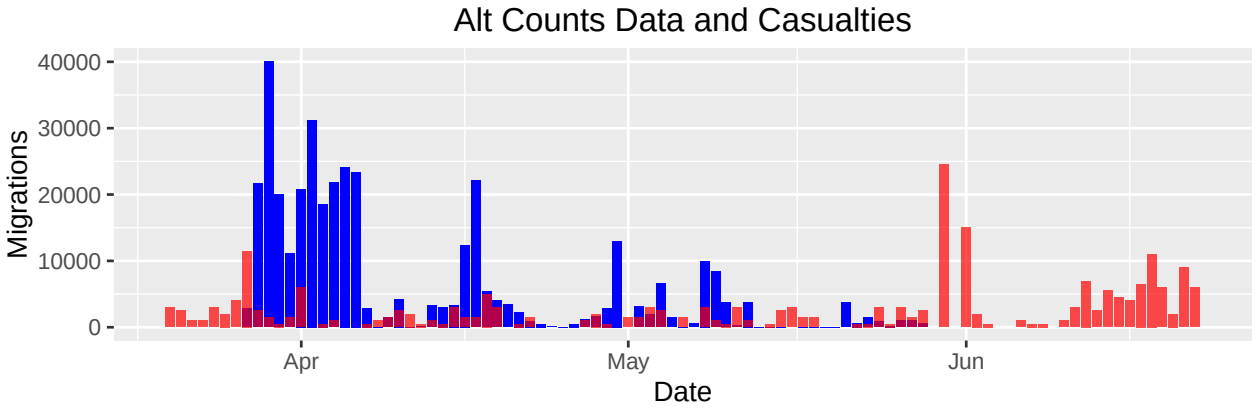
Looking at these plots, there appears to be more KLA activity around or before migrations compared to NATO activity. In fact, for the majority of these plots, most of the NATO activity is occuring after the report of migrations. We will now look at the other migration data sets to see if the same discoveries are present.

The next migration data set is the Alternative Border Counts data. Starting on March 27th of 1999, the UN High Commission for Refugees issued daily reports for individuals crossing the border. In addition, an Albanian government group began reporting similar figures on April 13th of 1999. The variables included in this data set are the date crossing the border, the number of people crossing reported by UNHCR and the number of people crossing reported by EMG. These figures do not always match each other and there are some missing values. The rows in which data is missing for both UNHCR and EMG were removed. If a row contained numbers from UNHCR but not for EMG, the former number was kept and vice versa. If there were numbers reported from both groups, the average was taken. Lastly, the number of airstrikes is divisible by 500 and this was done in order to display airstrike activity on this plot (a value of 500 = 1 airstrike).

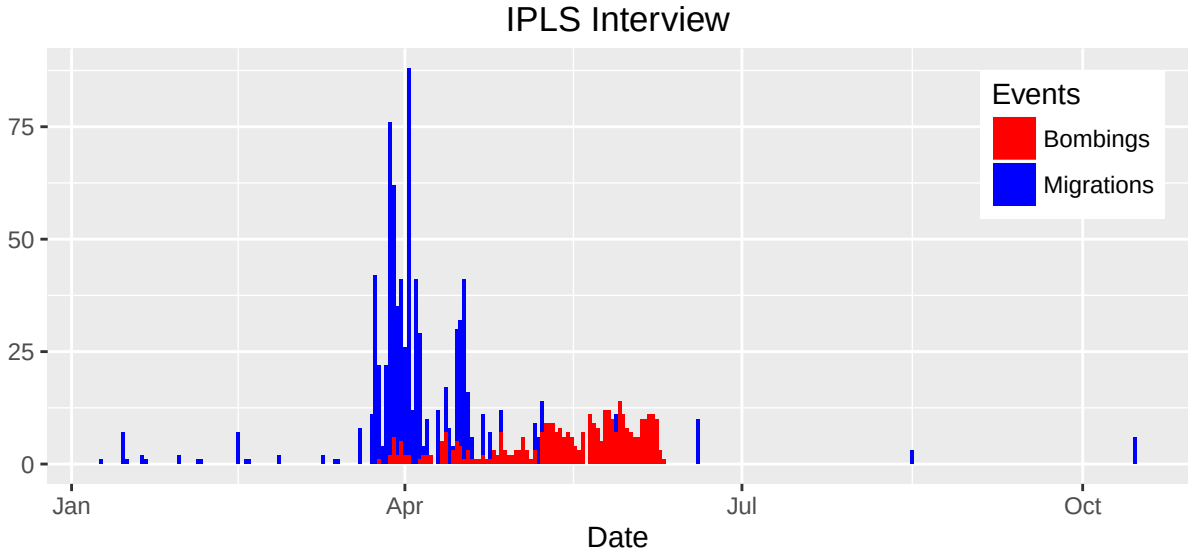
The below plot appears to show the same results from the Morina Border count data in that the majority of NATO activity occurs after a bulk of the migrations. There is an inverse relationship between the number of people migrating and the number/frequency of NATO airstrikes as time goes on. Additionally, there appears to be more KLA activity prior to and during the migrations. (Migrations in Blue, Other Event in Red)

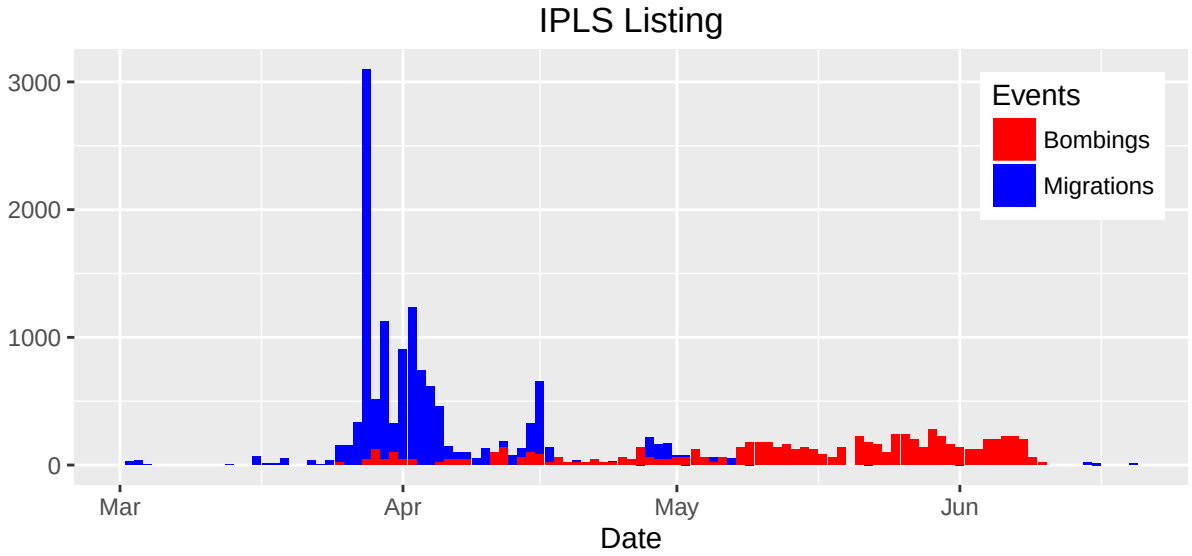


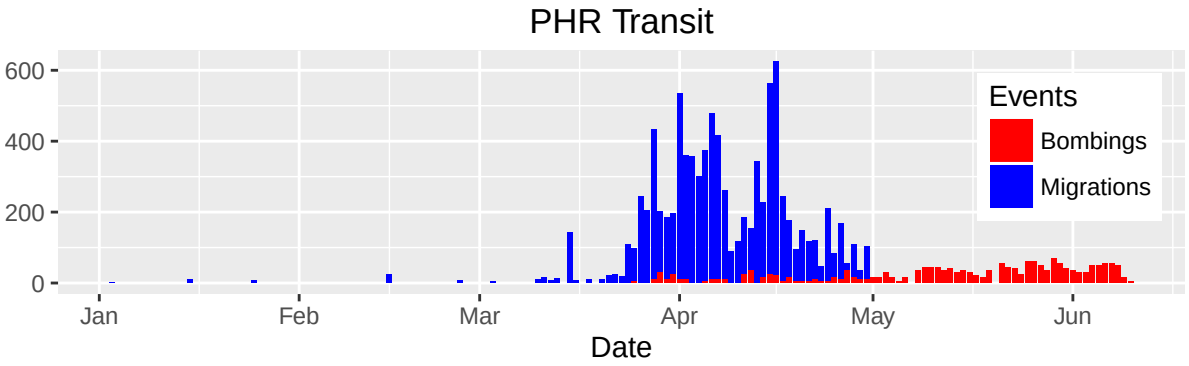


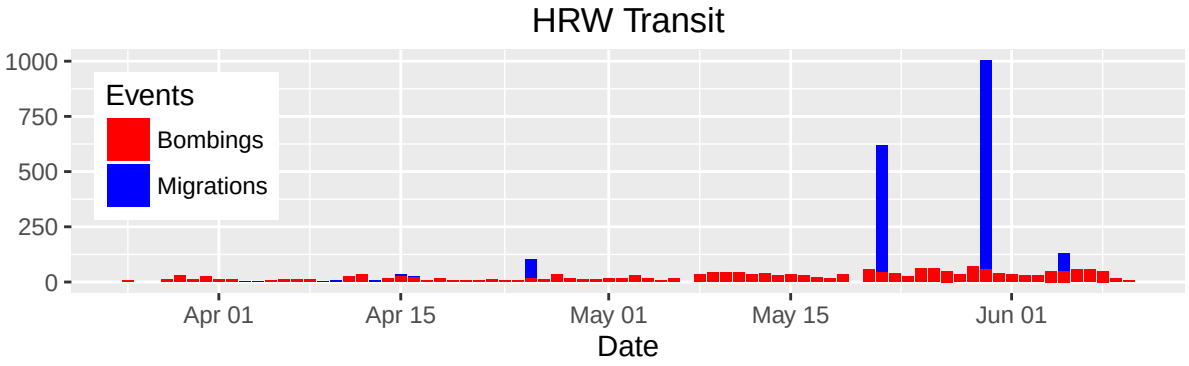


We will now go over the remainder of the migration data sets and will just provide summation plots showing the number of migrations and the dates of the migrations. The next two data sets come from IPLS/AAAS Listings and Interviews. For the listings data set, a sample of 1837 Kosovar Albanian families were taken and the information collected included the number of people in the household, the date crossing the border and the code for the village or town where they lived in Kosovo. For the interview data set, a total of 265 Kosovar Albanian households were interviewed and a code for their village or town was recorded along with the date they left their homes, the date crossing the border, the number of people in the group and whether they were interviewed in Bosnia or Albania. The Physicians for Human Rights data set contains records for 671 households that crossed the border into Albania and 509 records for households crossing in to Macedonia. This data set includes the number of people in the interviewed household, their leaving date, the date they crossed the border, where they were interviewed (Albania or Macedonia) and the code of the village or municipality from which they came. Finally, the Human Rights Watch data set consists of 123 people interviewed in Albania. This set includes the number of people in the interviewed household, their leaving date, the date they crossed the border and the code for the municipality from which they came.

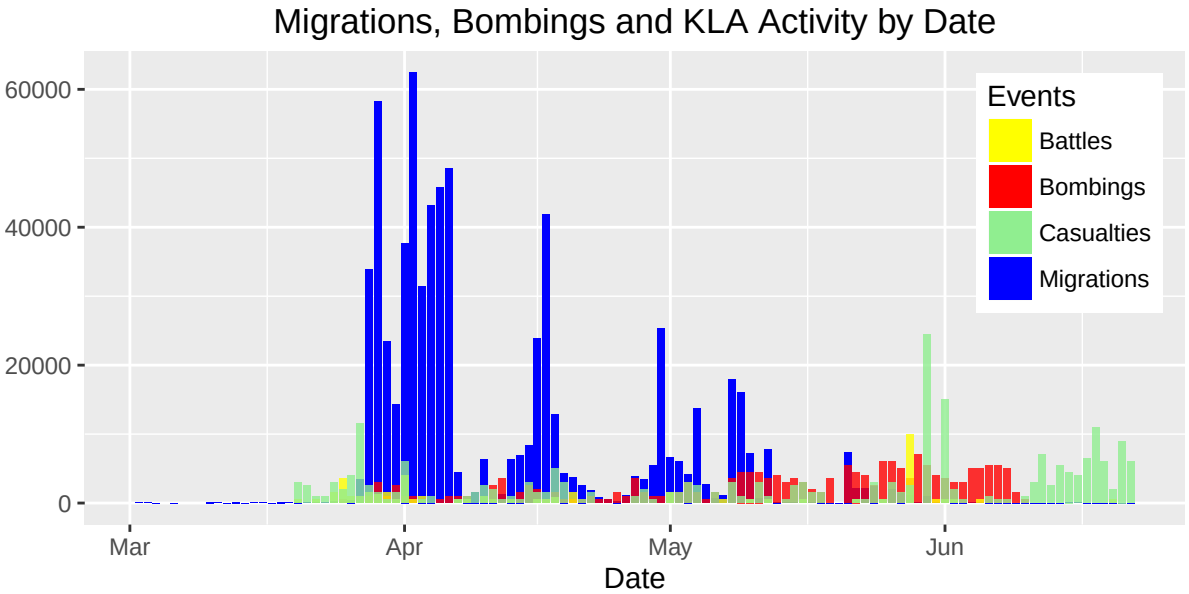








Apart from the HRW Transit plot, these plots show again that the bulk of the NATO bombing activity comes after the surge in migrations. There still appears to be a negative relationship with migration activity and bombing activity. To pull this all in to perspective, all of the mentioned migration data sets were combined to produce one single data set. The airstrike and KLA activity data sets were then merged with these migration data sets to produce a final data set. Below is a plot showing all of the combined migrations over time. As shown before, overall there appears to be a surge in migrations in late March to early and mid April and these migrations slowed as time went on. This plot also shows the airstrikes occurring over this time period along with the KLA activity. Please note that the plot only includes data points after 03/01/1999 as the activity prior to this date is minimal. In addition, airstrikes are divisible by 500. This was done in order to reflect airstrikes on the plot.



3: Model Building

Now that we have our final data set, we can begin to build models to determine if there is a significant relationship between the NATO airstrikes and the migrations in Kosovo or to discover other factors that may contribute to migrations. From the explanatory analysis, there does not appear to be a strong relationship between the NATO involvement and the migrations. Instead, the KLA activity appears to be more influential in the migrations, particularly the reported casualties. To explore the relationships, I will begin with linear regression and then also add some different higher polynomial terms.

A: Multiple Regression & Polynomial Regression

Table 1: Full Linear Model

	Estimate	Std..Error	t.value	Pr. . . t..
(Intercept)	-59488	31144	-1.91	0.05679
exdate	5.738	2.932	1.957	0.05097
strike	-60.2	78.55	-0.7664	0.4439
battle	118.7	96.45	1.23	0.2192
casualty	-30.42	46.77	-0.6504	0.5158

After fitting a linear regression model to the full data set, there is not a significant relationship between airstrikes and the number of individuals migrating out of Kosovo. However, as suspected, there appears to be a negative relationship between the number of people migrating and the number of airstrikes due to a negative coefficient of -60.202 for the variable strike. In addition, while the relationship between migrations and battles is not significant, it does appear to be positive due to the coefficient. Now, I will explore the impact of adding some higher order polynomials for airstrikes, casualties and battles individually.

Table 2: 2nd Degree for Strike

	Estimate	Std..Error	t.value	Pr. . . t..
(Intercept)	1571	233	6.743	5.036e-11
poly(strike, 2)1	856.9	4837	0.1772	0.8595
poly(strike, 2)2	-14599	4837	-3.018	0.002696

Table 3: 2nd Degree for Battle

	Estimate	Std..Error	t.value	Pr. . . t..
(Intercept)	1571	231.5	6.786	3.847e-11
poly(battle, 2)1	6728	4806	1.4	0.1623
poly(battle, 2)2	-17193	4806	-3.577	0.0003871

Table 4: 2nd Degree for Casualty

	Estimate	Std..Error	t.value	Pr. . . t..
(Intercept)	1571	235.2	6.679	7.5e-11
poly(casualty, 2)1	-41.12	4884	-0.008419	0.9933
poly(casualty, 2)2	-4501	4884	-0.9216	0.3572

After adding a 2nd degree polynomials for strike, battle and casualty, there appears to be more significance in the variables strike and battle, but more importantly there is a strong inverse relationship with the number of individuals migrating and airstrikes and battles. As airstrikes or battles increase, we would expect a sharp decrease in the number of individuals migrating. These results make sense in light of the previous plots. We will now incorporate battles and airstrikes in to the same model and review the results.

Table 5: 2nd Degree Terms for Strike & Battle

	Estimate	Std..Error	t.value	Pr...t..
(Intercept)	1571	230.1	6.827	3.001e-11
poly(strike, 2)1	2274	4823	0.4714	0.6376
poly(strike, 2)2	-12750	4850	-2.629	0.008877
poly(battle, 2)1	4536	4855	0.9342	0.3507
poly(battle, 2)2	-16565	4818	-3.438	0.0006431

Again, the higher order terms are significant and show a negative relationship between the number of airstrikes and battles to the number of migrations. We will now fit additional polynomails along with interaction terms and examine the results.

Table 6: 5th Degree for Strike

	Estimate	Std..Error	t.value	Pr...t..
(Intercept)	1571	230.9	6.804	3.464e-11
poly(strike, 5)1	856.9	4794	0.1788	0.8582
poly(strike, 5)2	-14599	4794	-3.045	0.002468
poly(strike, 5)3	12030	4794	2.51	0.01246
poly(strike, 5)4	-7989	4794	-1.666	0.09636
poly(strike, 5)5	6302	4794	1.315	0.1893

Table 7: 5th Degree for Battle

	Estimate	Std..Error	t.value	Pr...t..
(Intercept)	1571	230.8	6.808	3.379e-11
poly(battle, 5)1	6728	4791	1.404	0.1609
poly(battle, 5)2	-17193	4791	-3.589	0.0003711
poly(battle, 5)3	4766	4791	0.9947	0.3204
poly(battle, 5)4	-4170	4791	-0.8703	0.3846
poly(battle, 5)5	-9618	4791	-2.007	0.04533

Table 8: 5th Degree for Strike & Battle

	Estimate	Std..Error	t.value	Pr...t..
(Intercept)	1571	228.1	6.889	2.063e-11
poly(strike, 5)1	1521	4875	0.312	0.7552
poly(strike, 5)2	-12665	4813	-2.632	0.008812
poly(strike, 5)3	9866	4870	2.026	0.04339
poly(strike, 5)4	-7272	4829	-1.506	0.1328
poly(strike, 5)5	7347	4829	1.521	0.1289
poly(battle, 5)1	3813	4856	0.7852	0.4328

	Estimate	Std..Error	t.value	Pr. . . t..
poly(battle, 5)2	-13236	4950	-2.674	0.007796
poly(battle, 5)3	6292	4816	1.306	0.1921
poly(battle, 5)4	-4927	4826	-1.021	0.3078
poly(battle, 5)5	-10369	4765	-2.176	0.03011

Table 9: Interaction Strike & Battle

	Estimate	Std..Error	t.value	Pr. . . t..
(Intercept)	1152	335.2	3.435	0.0006496
strike	55.58	73.12	0.7601	0.4476
battle	609.9	224.2	2.721	0.006774
strike:battle	-87.7	37.16	-2.36	0.0187

Table 10: Interaction Strike & Casualty

	Estimate	Std..Error	t.value	Pr. . . t..
(Intercept)	1490	347	4.295	2.16e-05
strike	24.17	76.7	0.3152	0.7528
casualty	18.09	67.38	0.2686	0.7884
strike:casualty	-3.661	9.251	-0.3958	0.6924

Table 11: Interaction Battle & Casualty

	Estimate	Std..Error	t.value	Pr. . . t..
(Intercept)	1529	285.7	5.352	1.42e-07
battle	23.01	201.9	0.1139	0.9093
casualty	-25.97	51.32	-0.506	0.6131
battle:casualty	20.31	32.29	0.629	0.5297

These results again seem in line with the previous results in that the 2nd degree term is the most significant and reflects a negative relationship between airstrikes/battles with migrations. There also appears to be a significant relationship between battles and airstrikes on migrations. For all of these models, we will now perform 10 fold cross validation to assess the predictive power of these models on the response *numpers*. To begin, the data is mixed up using a random sample of all observations. After performing 10 fold cross validation, the results are listed below. The model that performed the best was model 5, which consisted of the formula (numpers ~ poly(strike, 2) + poly(battle, 2)).

Table 12: 10 Fold CV MSE Averages by Model (continued below)

Model 1	Model 2	Model 3	Model 4	Model 5	Model6	Model 7	Model 8
23753095	23557020	23314264	23894961	23068197	23485515	23573554	23501545

Model 9	Model 10	Model 11
23662385	23945133	24299430

B: Subset Selection

To see if we can come up with even better predictions on this data, we will use the best subset selection method to choose a set of variables and also make predictions. Below are the results of the best subset selection using the formula (numpers ~ exdate + poly(strike, 2) + battle + casualty). The best model chosen by BIC was the model with only one predictor, strikeˆ2. As can be seen from the coefficient, there is a negative relationship between airstrikes and the number of migrations. However, this model is not performing as well as the linear regression model. Lastly, 10 fold cross validation was performed for each different model and their average cross validation MSEs are listed below. From this table, the model that includes all 5 variables is performing the best. If we then apply this model to the data, we still see there is a negative relationship between the number of migrations and the number of airstrikes.

	Exdate	Strike	Strikeˆ2	Battle	Casualty
1 (1)			*		
2 (1)	*		*		
3 (1)	*		*	*	
4 (1)	*		*	*	*
5 (1)	*	*	*	*	*

Table 15: Coefficients for Best Model (By BIC)

	coef.subset.selection..best.
(Intercept)	1571
poly(strike, 2)2	-14599

Table 16: Average MSE For Each Model

Model 1	Model 2	Model 3	Model 4	Model 5
25195413	24954065	25050942	25021168	24919922

Table 17: Coefficents From Full Model

	coef.subset.selection..5.
(Intercept)	-39004
exdate	3.803
poly(strike, 2)1	-2482
poly(strike, 2)2	-12169
battle	86.04
casualty	-22.79

C: Generalized Additive Models

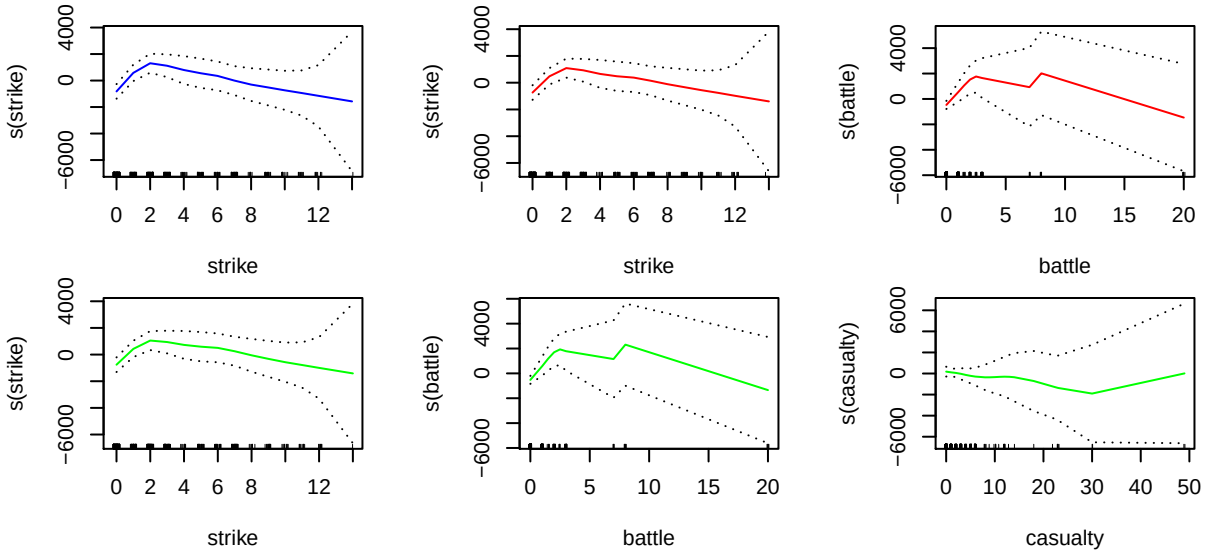
We will now use generalized additive models to fit the final data set. The first model only includes a smoothing spline for strike, the second model includes smoothing splines for strike and battle and the third model includes smoothing splines for strike, battle and casualty. These three models are initially compared by using ANOVA and the best result appears to be the second model (strike and battle). This is similar to the results obtained using regression above. In addition, cross validation was used to examine the prediction power of these models and the results are listed below. Again, the model with smoothing splines for strike and battle performed best. This indicates that there is a strong non-linear relationship between the airstrikes, battles and the migrations. Even more so, these results show that the relationship between battles and migrations is stronger than the relationship between airstrikes and migrations.

Table 18: Anova Test

Resid. Df	Resid. Dev	Df	Deviance	F	Pr(>F)
426	9.774e+09	NA	NA	NA	NA
422	9.496e+09	4	2.78e+08	3.072	0.01636
418	9.458e+09	4	37888391	0.4186	0.7952

Table 19: 10 Fold CV MSE Averages by Model

Model 1	Model 2	Model 3
23426935	23062871	23114182



D: Conclusion

After fitting various models and reviewing the results, there does appear to be some evidence that NATO was not directly responsible for the surge in migrations in Kosovo. The graphs provided point to a decrease in migrations as airstrikes accelerated. This trend was seen after examining the various migration data sets along with the data on airstrikes. After fitting various regression models, there is a non-linear relationship between the NATO airstrikes and the surge in migrations. In addition, this relationship appears to be an inverse relationship due to the strength of the regression coefficients in the models with 2nd degree polynomials. Overall, there is evidence to argue that NATO was not solely responsible for the migrations.