Dealing with Missing Data in Group-Level Studies of Terrorism

Bryan Arva and John Beieler
Pennsylvania State University

September 30, 2014

Abstract

One prominent area of research in terrorism studies focuses on the analysis of terrorist groups. Studies in this area examine group attributes, such as ideology, size, and state sponsorship, in order to determine the impact these factors have on phenomenon such as the number of attacks conducted or the targets of attacks. One significant issue for this type of research, however, is that a large number of attacks are not attributed to a specific group. Many researchers simply remove these missing cases under the assumption that the data is missing at random. In light of this, we examine whether the missingness of group information can be predicted with a high degree of accuracy. We find that, when making use of machine-learning algorithms and out-of-sample testing with a limited number of covariates, we are able to predict whether group information will be missing with an accuracy of greater than eighty percent. This finding suggests that the group-level data is not missing at random, which further implies that existing statistical studies that remove this data have results that are biased in some manner. Given this, we provide discussion of the implications this has for existing and future works within the field.

1 Introduction

The field of terrorism studies is a relatively new one in comparison to many other areas of international relations. Before the September 11 terrorist attacks, few scholars studied terrorism as their main field. Since then, however, it has become a hotbed for research and a burgeoning field in IR. When terrorism studies first began to take shape as a discipline, any data was considered good data. While this was true at the beginning, due to lack of resources and the small number of people involved in terrorism research, it is no longer the case. We believe that one of the problems with terrorism research is that people became comfortable with the idea that terrorism data just was not as good as other data. Because of this, the bar has been set lower. This can be seen in the way people deal with the problem of missing data in group level studies. Although many group level studies have close to, or more than, 50% missingness in their data, very few scholars acknowledge that this is a problem. By missingness we mean the number of terrorist attacks, in the dataset, that are attributed to unknown groups. In this paper, we argue that there is a large problem with the way past scholars have dealt with the problem of missingness by showing that the data is not missing at random. We use machine learning techniques to predict whether an attack was attributed by a known or unknown group. Our ability to predict at a much higher level than chance supports our hypothesis that the data are not missing at random.

2 Studies of Terrorist Groups

While most of the early terrorism research was theoretical and based on case studies, a strong wave of quantitative research has been going on for over a decade. At first, most people conducted cross-national studies on international terrorism because that was the only data available to them. In more recent years, people have begun focusing on group level studies of terrorism. There are a great number of things that studying terrorism, at the group level, can tell us. Some of the more theoretical work can help us construct terrorism typologies in order

to deal with numerous definitions for terrorism that are prevalent in our field (Ganor, 2008). We can learn about what aspects of terrorist groups make them more lethal: do certain things make them more likely to commit a higher number of attacks or commit attacks that lead to a higher number of casualties (Asal and Rethemeyer, 2008). Additionally, are there certain characteristics of terrorist groups that make them more or less likely to use certain weapons or have certain targets? Other studies have been carried in order to assess what factors affect the longevity of terrorist groups (Carter, 2012). These are very important topics and, the progress the field of terrorism has undergone, is a very encouraging sign. Still, the problem of missingness in terrorist group studies is very troubling. Not all terrorist group studies suffer from the problem of missingness though. Most of the theoretical work and case studies can be continued without a problem. Additionally, if you are studying outbidding, like Findley and Young (2012), then your results should be robust. However, anytime one uses data where the attacks attributed to unknown groups are dropped, this is a problem.

2.1 Missing Data in Group Studies

The problem of missing data in terrorist group studies is real and, to our knowledge, no one has offered a solution for it. Two of the most popular sources for group level data, MIPT's Terrorism Knowledge Base (TKB)[1] and the Global Terrorism Database (GTD),[2] attribute roughly 72% and 40% of the attacks in their database to "unknown" groups. Even more troubling, few studies acknowledge that their data suffers from a problem of missingness. Most scholars, including ourselves in the past, have simply listed the data source that they were using and left it at that. Some, like Asal and Rethemeyer (2008), acknowledge that there is a serious problem with the number of missing observations in the data. Still, most of these scholars drop the missing observations and analyze their results like they have complete data. Little, if any, time is spent in the robustness checks sections talking about the potential

[1] http://www.tkb.org/Home.jsp
[2] http://www.start.umd.edu/gtd/

effect of the missing data. People can drop data because the data is assumed to be missing at random but this seems highly unlikely. If data is not missing at random, and is list wise deleted, it could lead to biased predictions, give the findings poor external validity, and leave us with a low degree of confidence in our results.

3 Theory and Hypotheses

There seem to be two primary reasons for missing group-level information: media reporting and group reporting. In the first case, the media simply fails to report the name and information of a group claiming an attack. This might have been more plausible in the past but, with the technology we have today and the number of news sources that provide global coverage, it seems highly unlikely now. However, this is something that we could study in the data: were there a higher percentage of attacks attributed to unknown groups in the pre-internet era? In order to answer this question, we split up the dataset to see if better technology has led to better media reporting and, in turn, less unknown group attacks. The results are actually quite interesting. From 1970-1999, 34.8% of terrorist attacks were attributed to unknown groups. During the 21st century, from 2000-2011, that number climbs to 57%. While it is true that there were many more attacks reported, in total, over the latter time period, these figures provide evidence that the number of attacks attributed to unknown groups was not higher in the past because of poor media reporting. Still, poor reporting could be a byproduct of the political regime in a country so we cannot discount this as an explanation for why many attacks are attributed to unknown groups.

The second possibility is that, for some reason, a group fails to lay claim to an attack. There are many possible explanations for this. For example, Juergensmeyer (2003) argues that religious terrorist groups may be less likely to report attacks because they do not want that level of attention. Additionally, there could be circumstances where the group feels that their goals would be better served if they remain anonymous. This could be the case with

splinter groups or extremist members of groups, working on their own, without the consent of their leaders. Kydd and Walter (2006) would lead us to believe that one example where we might find this is during civil conflicts where these aforementioned groups are attempting to act as spoilers. If they do not like the agreements their groups are about to sign with the governments, they might commit unclaimed attacks in order cast a negative shadow on the groups and get the governments to walk away from the negotiation table. Finally, groups may want to remain anonymous for fear of reprisal. If they are a weak group, or fear that the government will react harshly to the attack, they may not claim responsibility.

The fact there are theoretical reasons for the large number of attacks attributed to unknown terrorist groups leads us to our one, and only hypothesis: *The data are not missing at random which means we will be able to predict whether a group observation is missing, with a higher degree of accuracy, than chance.*

4 Is the data missing at random?

The question of whether data is missing at random is an empirical one. We propose that if it is possible to predict whether a group observation is missing with a high degree of accuracy, then the data is not missing completely at random. Missing completely at random implies that the observed data carries no information regarding missingness, while the ability to predict implies that the data *does* carry such information. Thus prediction means that the data is not missing completely at random. The following sections explore this question in greater depth and outline how we go about answering this empirical question.

4.1 Analysis

4.1.1 How Much Is Missing?

Before conducting an in-depth statistical analysis of the data, it is useful to question whether it really matters if the data is missing at random. While it is well known that

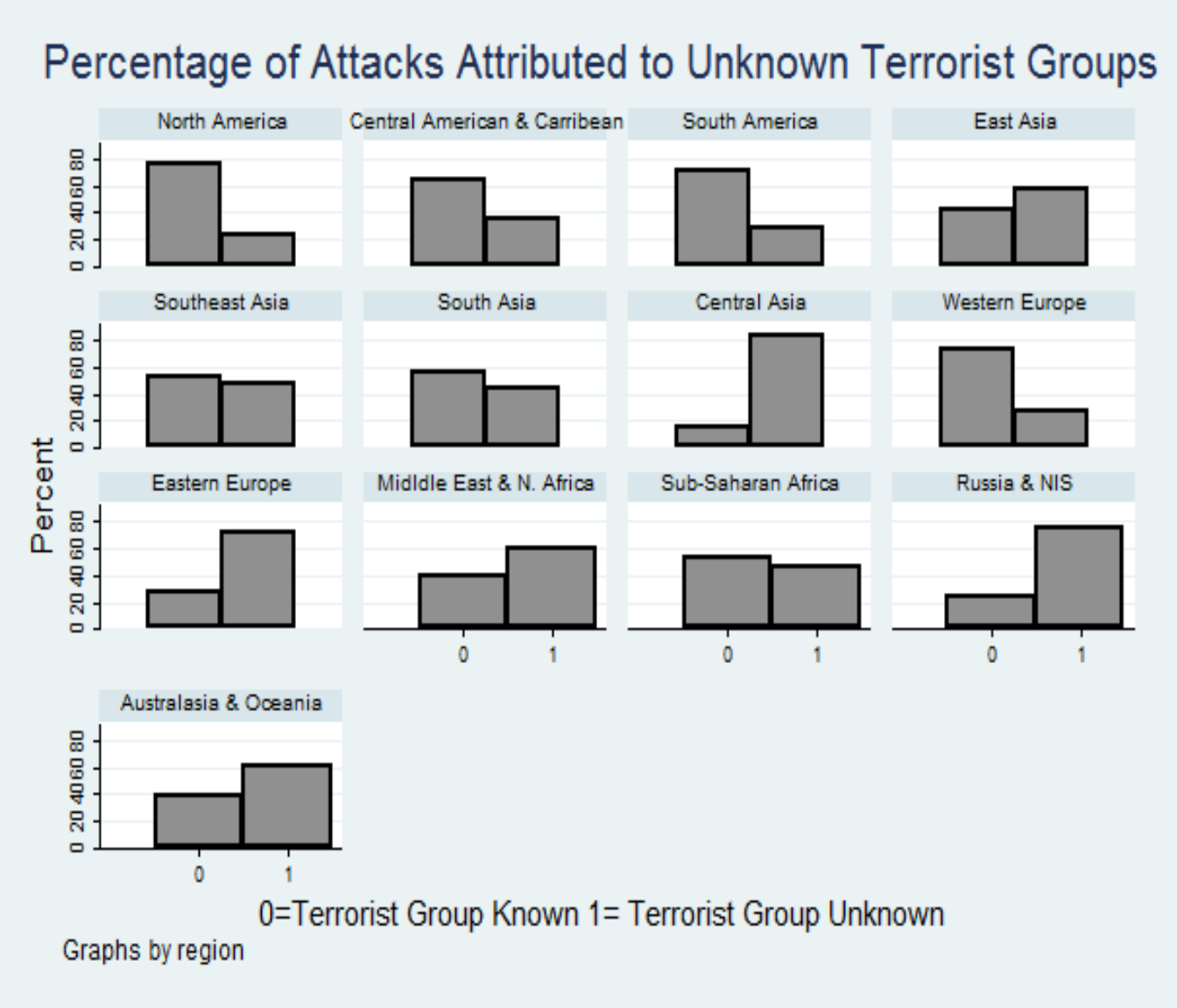
list-wise deletion of data that is not missing at random can introduce bias in statistical estimation, if only a handful of cases are missing within the dataset it might be enough to "shrug off." A quick examination of the START GTD data shows that 44,201 observations have a group that is recorded as "Unknown." The total size of the dataset is 104,689 observations, which means that roughly 40% of the observations would be considered missing in a group-level study of terrorism. This point bears repeating; nearly half of the data would be removed from the dataset when performing list-wise deletion of missing data. Thus, at the very least, this is a problem that deserves some consideration. The next question this brief analysis presents, then, is the degree to which this large amount of missing data poses an issue for statistical analysis. As stated previously, systematic bias in the data is problematic, while missing group-level data that is randomly distributed does not introduce problems for statistical analysis. The next section attempts to determine whether the group-level missingness in the GTD is random or non-random in nature.

4.2 Which Data are more likely to be Missing?

This section serves to give the reader a better understanding of which types of terrorist attacks are more likely to be attributed to unknown groups. In the following paragraphs, we will present some graphs, and descriptive statistics, to help the reader better understand whether where the attack occurred, type of weapon used in the attack, the target of the attack, and/or the type of attack affects the likelihood that we will know the perpetrator of the attack. As can be seen, from the figure below, the region of the world where the attack occurred does seem to have an affect on whether or not the attack is attributed to an unknown group. As stated above, we are missing data on roughly 40% of the attacks. Some regions of the world, such as the Americas and Western Europe, are better able to identify the perpetrators of terrorist attacks. Others, most notably Central Asia, Eastern Asia, and Russia, have a very high number of attacks that are attributed to unknown groups. The fact that these three regions have the highest percentage of missingness is quite telling. All three

of these regions are very close to one another geographically, so it may indicate that this is one place in the world where identifying terrorist groups is more difficult. Going forward, we would like to try and determine if there are certain similarities amongst these regions, which are specific to those areas, that leads them to have a higher degree of missingness. If we can figure of the problem, then we can try to come up with a solution and create better datasets.

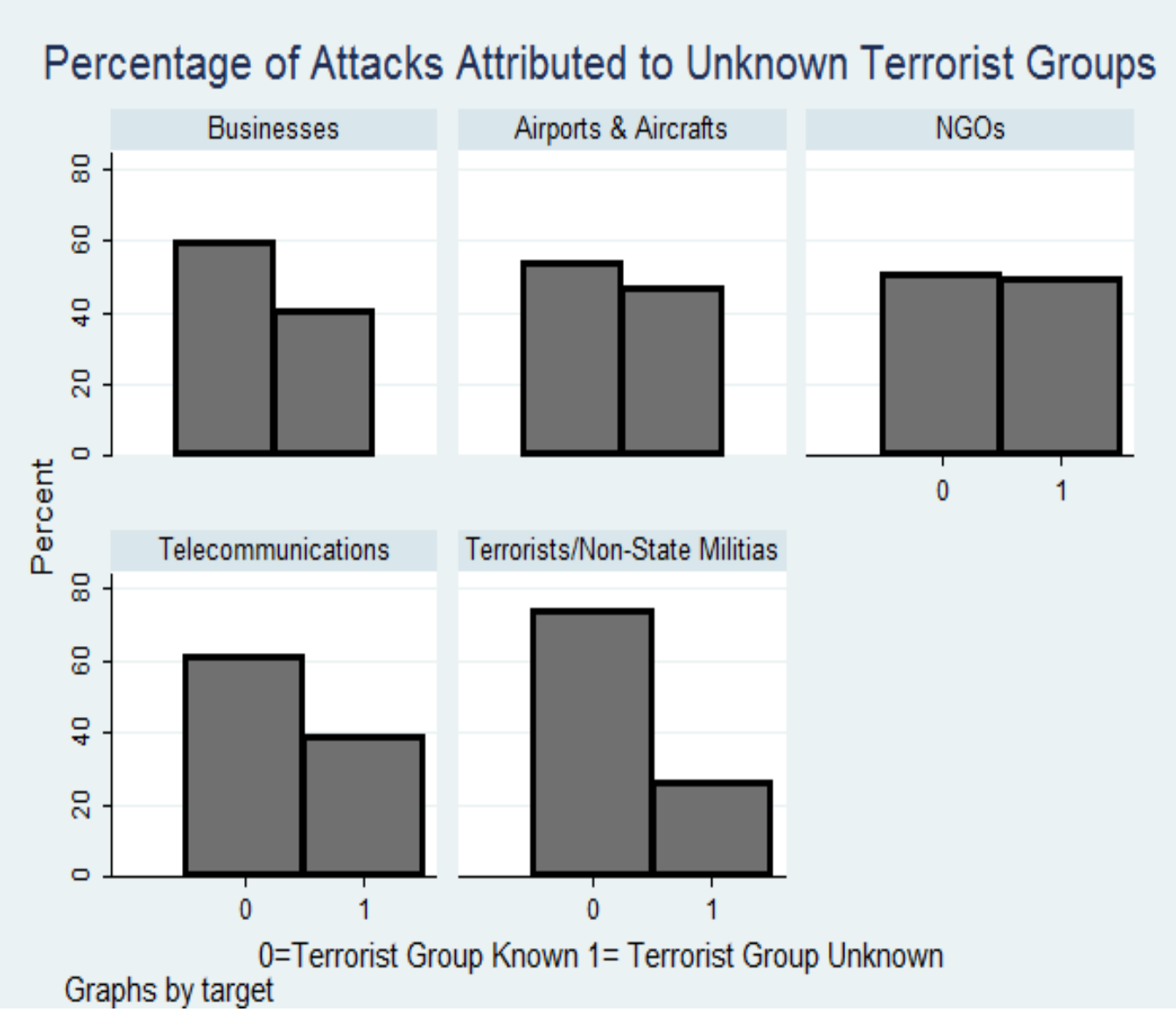
Figure 1



The figure below is a graphical representation of the relationship between the target of

the terrorist attack and whether or not the attack was attributed to an unknown group. There are two interesting things that can be taken from these data. First, when terrorists are targeting other terrorists/non-state militias, it seems that they are either more likely to claim responsibility for their attacks or less careful at trying to hide their identify. This is an interesting finding and one that we hope to delve into more in further iterations of this paper. Additionally, of all the target types, National Governmental Organizations have the highest level of missingness.

Figure 2



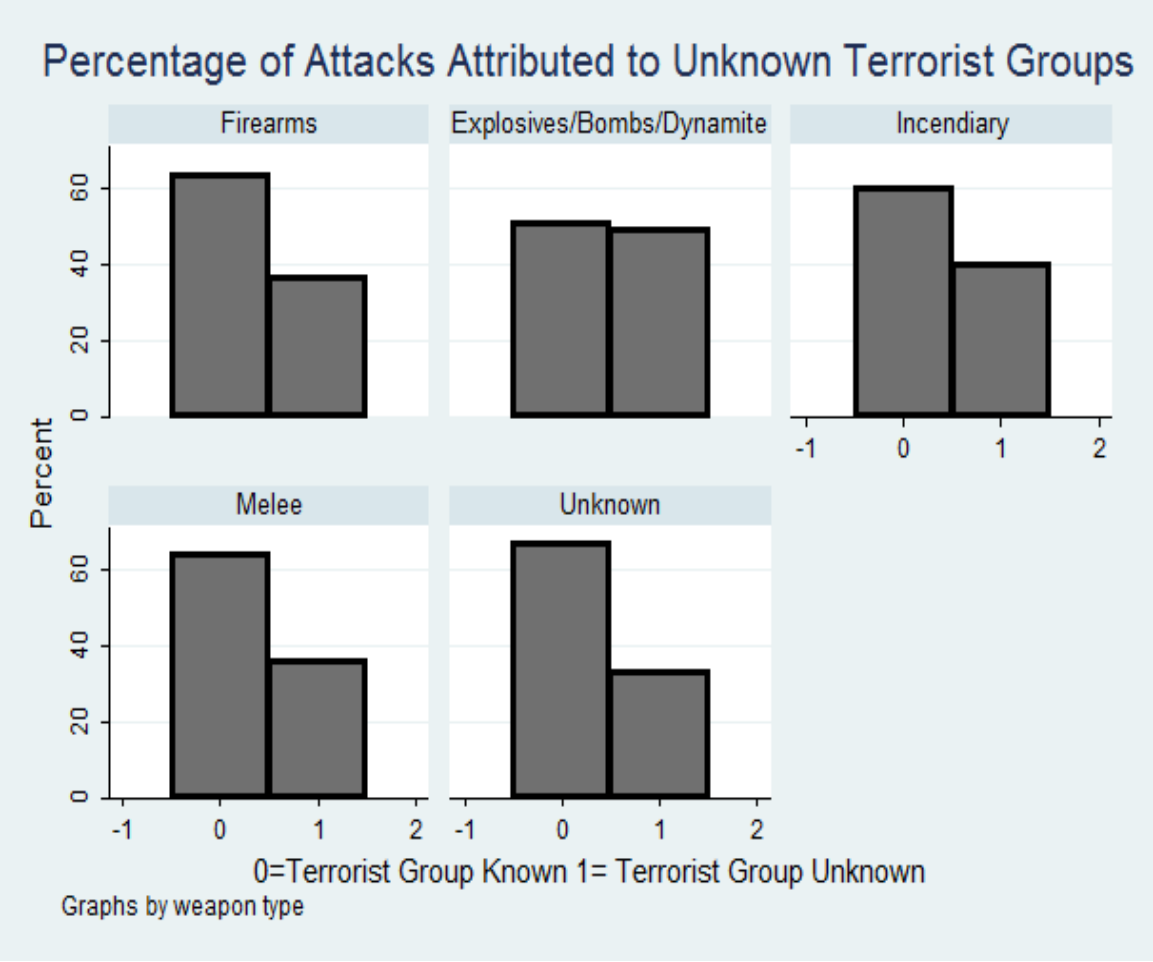
We also thought it would be interesting to see if the type of weapon used in an attack had an effect on the level of missingness. This is one of the variables that seems to have the least fluctuation. Most of the weapon types have roughly the same 60/40 split as the overall data set, with regard to the ratio of attacks attributed to known v. unknown groups. Because of this, explosives/bombs/dynamite does stand out to a degree: there is roughly a 50/50 split here. It is possible that because one can be far away, or use themselves as a human bomb, it is less likely that the identity of the perpetrator will never be known. In following, it would most likely be harder to figure out the group responsible for the attack as well. It is also very interesting that the attacks committed with unknown weapons are one of the most likely to be attributed to a known terrorist group. This finding does not seem to make much logical sense and must be looked into further.

4.2.1 Predicting Missingness

In order to determine whether or not the data within the GTD is missing at random or not we approach the modeling as a simple prediction problem. In other words, can we build a model that is able to accurately predict whether an observation will be missing or not? If the data *is* missing at random it should not be possible to construct a model that is able to accurately predict missingness. Our models should do no better than a coin flip in this scenario.

Towards this end, we construct one of the most simple datasets available to us. We mainly make use of variables in the GTD data itself. Specifically, we use the date of the attack, the country, region, latitude and longitude, whether it was a suicide attack, the attack type, the target type, and the nationality of the target. We also include the Polity scores for each country in which the attack occurs. This variable is included since the regime type of a country may impact whether or not a group claims a terrorist attack. We then encode the group name variable as 1 if the group is "Unknown" and 0 otherwise. This dataset is then split 75%-25% into training and test sets. This is standard practice in machine learning

Figure 3



approaches. The algorithm is trained, or fit, on the 75% sample and the models accuracy is then tested on the 25% "hold-out" data. This is done to avoid problems of overfitting.[3] Using this split, we train two models. The first is a decision tree, while the second is a random forest. A decision tree is different from the linear methods usually seen in political science. Instead of fitting a line to a set of data, a decision tree finds splits within the dataset that increases the purity of the data at each split. Put differently, the algorithm finds potential divisions in the data, such as variable X less than 15, that best reduces the amount of mixing within the outcome variable. The second algorithm, a random forest, is a

[3]Overfitting can be thought of as the algorithm memorizing the answer to a problem, rather than truly learning what characteristics truly distinguish a positive observation from a negative observation.

form of ensemble learning that combines many decision trees[4] together and tallies the votes of each individual decision tree. One of the distinguishing factors of a random forest is that a bootstrap sample is constructed for each individual decision tree along with a random sample of the variables. Random forests have many favorable properties that make them an easy favorite for a "first cut" for many modeling exercises. These properties are beyond the scope of this paper, however, and it can suffice to say that random forests often perform well under a variety of situations.

We fit both of these algorithms to the 75% test set and then evaluate the performance using the 25% test set. The results for these algorithms are below. Before turning to these results, a brief discussion of model evaluation is necessary. The most straightforward measure of algorithm performance is accuracy. This simple measures how many observations an algorithm got right. At first glance this is useful. A deeper look shows that, using our data as an example, guessing all 0 would lead to an accuracy of about 60%. This is the case since we have roughly 60% negative observations in our dataset and by guessing 0 we would classify each of these correctly, but miss the remaining 40%. Given this, we use three metrics: precision, recall, and receiver-operating characteristic (ROC) curves. Precision is defined as:

$$precision = \frac{truepositives}{truepositives + falsepositives} \tag{1}$$

This equation can be interpreted as the percentage of positive classifications that were correct. The equation for recall is:

$$recall = \frac{truepositives}{truepositives + falsenegatives} \tag{2}$$

In contrast to precision, this measure reports how many of the positive observations our model was able to correctly identify. Finally, a ROC curve plots the true positive rate and

[4]A forest is made up of trees.

false positive rate for various probability cutoff points for classification. The area under the ROC curve (AUC) is a commonly-used metric to asses model performance. The AUC is functionally equivalent to a rank-sum test, which means that the AUC shows how well an algorithm is able to discriminate between positive and negative observations (Flach, Hernandez-Orallo and Ferri 2011). An AUC of 50% is equivalent to a random guess. Given this, Table 1 shows the results for each of these three metrics for both of the algorithms examined in this paper.

Table 1: Held-out Data Metrics

	Precision	Recall	AUC
Decision Tree	71%	71%	75%
Random Forest	79%	75%	89%

We also present the ROC plot and classification table for each of these models to provide further insight into the results.

Table 2: Classification Table - Random Forest

	Predicted	
Actual	0	1
0	11,553	2,031
1	2,587	7,825

In addition to these scores, the random forest allows us to calculate variable importances in order to determine which variables play the most important role in the model. Figure 6 shows these scores. Overall, the most important features for predicting whether or not

Figure 4

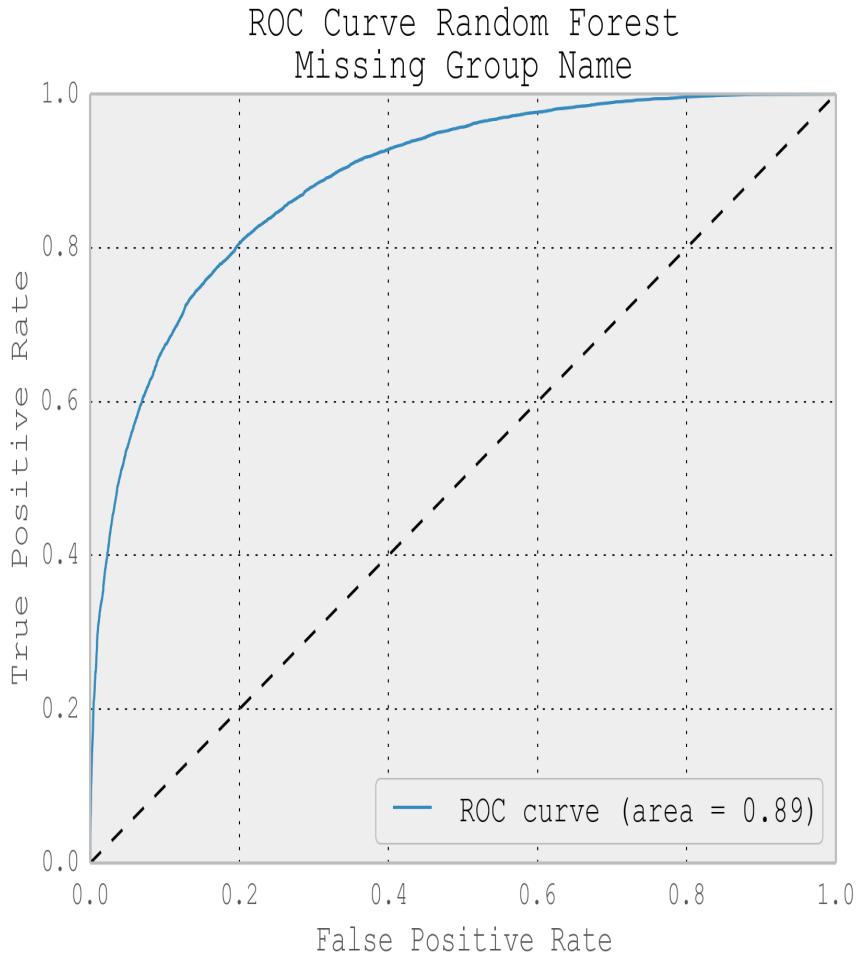


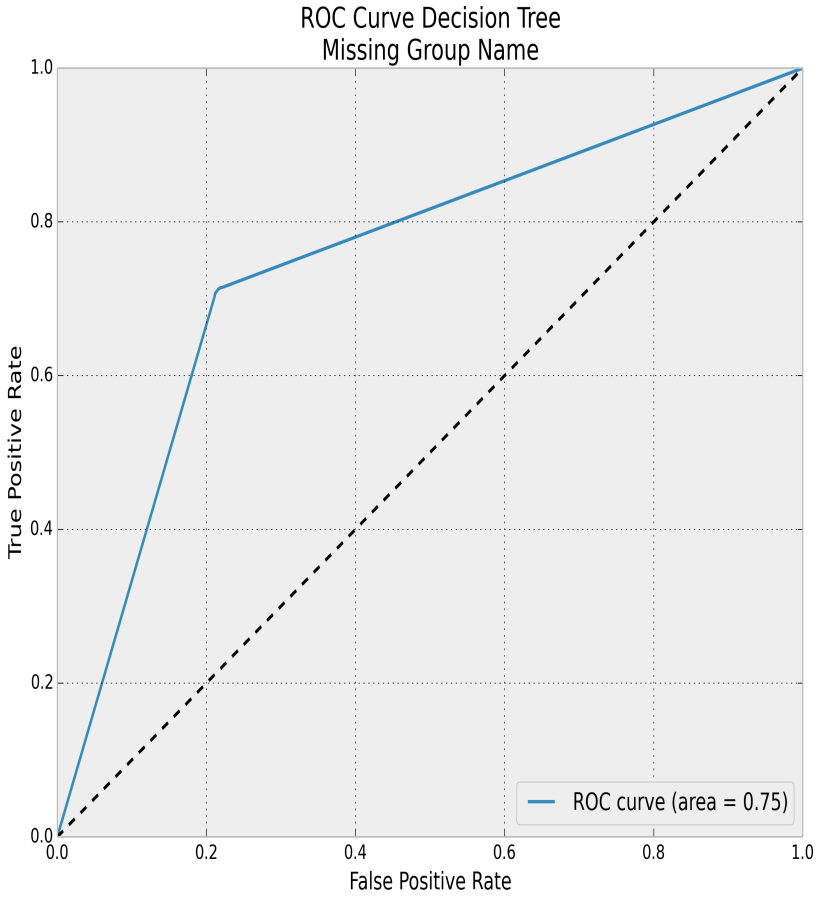
Table 3: Classification Table - Decision Tree

	Predicted	
Actual	0	1
0	11,891	3,194
1	3,251	7,837

a group will be coded as missing are the date information, the target type, the country in which the attack occurs, the nationality of the target, and the country's Polity score.

The overall conclusion these models present is that the group-level data in the Global Terrorism Database is not, in fact, missing at random. This is shown by the ability of our models to gain high scores on all of the metrics presented. Perhaps even more surprising is

Figure 5

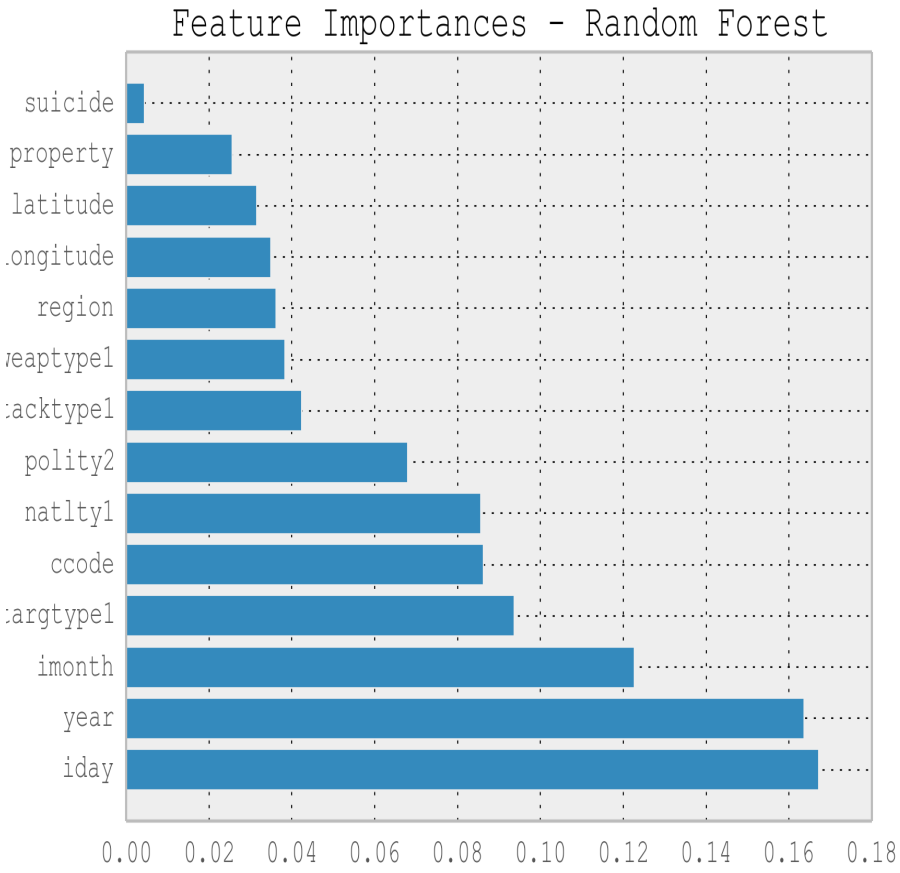


the models' ability to correctly predict many of observations correctly with a fairly small, and limited, set of covariates. Just using information about the attack itself it is possible to create fairly accurate models. If one so desired, a fairly comprehensive story regarding the missing observations could be constructed if things such as Polity scores or other country-level variables were included. We address this issue a bit more in the next section.

5 Next Steps and Conclusion

Given our findings, the obvious question is: what next? There are a few possible options. The first, and easiest, is for papers to draw the appropriate conclusions given the limited

Figure 6


Feature Importances - Random Forest
suicide
property
latitude
ongitude
region
weaptype1
attacktype1
polity2
natlty1
ccode
targtype1
imonth
year
iday
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18

data available. The findings are not global in nature. Instead, the results are derived from a limited, and biased, sampling of group-level terror events. While this may be unsatisfying from an external validity, or generalizability, standpoint it is much less worrisome than presenting results as indicative of broad, global trends in terrorist events.

The second option is to somehow impute the missing data using a technique such as multiple imputation (van Buuren and Groothuis-Oudshoorn 2011; King, Honaker, Joseph and Scheve 2001). Before using some form of multiple imputation, however, we must determine *in what way* the GTD data is missing. King, et al (King et al. 2001, 50-51) outline three possibilities: missing completely at random (MCAR), missing at random (MAR), and non-ignorable (NI). Since we are able to predict missingness, it seems clear that the data is not MCAR. The main difference between MAR and NI data is whether the $P(M)$ is dependent on the data. While King, et al (King et al. 2001, 51) note that "the presence or absence of NI can never be demonstrated using only the observed data" it seems that a bit of thought regarding why the data is missing in the GTD is called for. As we stated above, there are two primary reasons for missing group-level information: media reporting and group reporting. We could likely include more information in our model, such as additional country-level variables, to build a more complete model of the missingness, but that exercise was beyond the scope of this paper. Given both of these possible explanations, it seems to us that NI should not be written off.

In addition to the issue of how the data is missing, there's the practical considerations of imputing missing data for group-level variables. The core of this problem is a prediction task. The end goal is, at its heart, to build a model that predicts which group carried out an attack. We believe that it is possible to create such a model. Intuitively, it is possible for an analyst to conclude that if an attack takes place in country X, within region Y, and between a certain set of years, the attack was most likely carried out by group Z. Our modest proposal is to construct a formalized and quantitative algorithm to encode this knowledge. Given the results presented in this paper, we believe that it is possible to make some progress

on this front.

In conclusion, we believe that there is a severe problem with missingness in the area of terrorist group studies. Furthermore, we are worried by the fact that most scholars seem to dismiss it and not give it the attention it deserves. Assuming that these data are not missing at random, which we believe we have demonstrated throughout this paper, we can have very little confidence in the results of past group level research, including our own. here are too many great things scholars can learn from conducting group level studies in terrorism to simply dismiss this as a small data problem. We hope that this paper will open up people's eyes to this issue, that it can spur conversation, and that we can collectively come up with a solution.

References

Asal, Victor and R. Karl Rethemeyer. 2008. "The Nature of the Beast: Organizational Structures and the Lethality of Terrorist Attacks." *The Journal of Politics* 70(2):437–449.

Carter, David B. 2012. "A Blessing or a Curse? State Support for Terrorist Groups." *International Organization* 66(1):129–151.

Findley, Michael G. and Joseph K. Young. 2012. "More Combatant Groups, More Terror?: Empirical Tests of an Outbidding Logic." *Terrorism and Political Violence* 24(5):706–721.

Flach, P.A., J. Hernandez-Orallo and C Ferri. 2011. A coherent interpretation of AUC as a measure of aggregated classification performance. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. pp. 657–664.

Ganor, Boaz. 2008. "Terrorist Organization Typologies and the Probability of a Boomerang Effect." *Studies in Conflict & Terrorism* 31(4):269–283.

Juergensmeyer, Mark. 2003. *Terror in the Mind of God: The Global Rise of Religious Violence*. Berkley, California: University of California Press chapter Chapter 7.

King, Gary, James Honaker, Anne Joseph and Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review* 95(1).

Kydd, Andrew H. and Barbara F. Walter. 2006. "The Strategies of Terrorism." *International Security* 31(1):49–80.

van Buuren, Stef and Karin Groothuis-Oudshoorn. 2011. "mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software* 45(3).