Comparing classification trees to discern patterns of terrorism*

Saiya, Nilay; Scime, Anthony

2019

Saiya, N., & Scime, A. (2019). Comparing classification trees to discern patterns of terrorism*. Social Science Quarterly, 100(4), 1420-1444. doi:10.1111/ssqu.12629

<https://hdl.handle.net/10356/145738>

<https://doi.org/10.1111/ssqu.12629>

This is the accepted version of the following article: Saiya, N., & Scime, A. (2019). Comparing classification trees to discern patterns of terrorism*. Social Science Quarterly, 100(4), 1420-1444. doi:10.1111/ssqu.12629, which has been published in final form at <https://doi.org/10.1111/ssqu.12629>. This article may be used for non-commercial purposes in accordance with the Wiley Self-Archiving Policy [<https://authorservices.wiley.com/authorresources/Journal-Authors/licensing/self-archiving.html>].

Downloaded on 26 Apr 2023 08:45:07 SGT

Comparing Classification Trees to Discern Patterns of Terrorism

Objective

Though applied widely in the fields of medicine, finance, ecology, psychology and computer science, machine learning algorithmic-based methods are a relatively novel approach to social scientific analysis that have yet to be extensively applied. Yet as we argue in this article, a specific form of algorithmic analysis known as C4.5 classification trees has much to offer social analysis and, specifically, the study of social and political violence.

Method

This article describes four novel classification model comparison techniques for the C4.5 classification method and applies them to the study of terrorism.

Results

Our state-level analysis suggests that there is something fundamentally different in the targeting choices of religious and secular terrorists.

Conclusion

This analysis highlights the ability of classification trees to heighten our understanding of terrorism and even provide recommendations to policymakers for avoiding future attacks.

Comparing Classification Trees to Discern Patterns of Terrorism

1. Introduction

The social sciences study human relationships in society. This is accomplished by comparing one social phenomenon to another, with the intent of identifying the similarities and differences between them. Often this analysis requires either a qualitative or statistical approach. Qualitative analysis, however, can fall prey to the biases of the analysis, while statistical analysis deal with averages and provides a single estimated value of the data (Spielman and Thill, 2008). In this article, we present four techniques, based on machine learning and applied to social data, which allows comparison of similar social phenomena using a non-biased, data-centric approach. These techniques not only compare social phenomena, but also allow for the segmentation of the phenomenon into component parts.

There are major differences between conventional statistical methods, such as regression analysis, and classification modeling in data analysis. According to Spielman and Thill (2008: 111), "In contrast to more conventional multivariate statistical methods such as factor analysis, principal component analysis, and multidimensional scaling, they [machine learning techniques] tend to be less bound by a priori assumptions." Watt et al. (2009) and Chang (2006) conducted studies comparing classification mining and logistic regression. Both studies found classification superior to regression, more easily understood, more accurate in describing the domain and more precise in predicting future actions in the domain. Interpretation of classification results are more "reader-friendly" compared to the odds ratios of logistic regression (Watt et al., 2009). Classification rules directly provide the independent variables

and their values that co-occur with a specific dependent variable's values, which define relationships in the data (Scime et al., 2008). It is the visibility of the relationships in the tree structure and the individual classification rules, not available from non-tree regression, that are used in this study to compare two datasets.

While conceptually informative, probability statements with confidence intervals and statistical hypothesis testing are not sufficient for making individual decisions, whereas the outcomes from classification can be actionable, practical and immediately applied to real-time data (Chang, 2006). As noted by Basuchoudhary and Bang (2019), "Machine learning is not sensitive to the sometimes heroic statistical assumptions [of] parametric econometrics...This increases the reliability of explanations while adding contextual nuance that captures the flavor of individualized case analysis." While both classification and logistic regression analyses can lead to the same results, classification provides more specific insight into the characteristics that are most relevant.

Though applied widely in the fields of medicine, finance, ecology, psychology and computer science, machine learning algorithmic-based methods are a relatively novel approach to social scientific analysis that are only beginning to be applied to the study of political violence. We argue in this article, a specific form of algorithmic analysis known as C4.5 classification trees (Quinlan, 1993) has much to offer social analysis and, specifically, the study of social and political violence. This article describes the C4.5 classification method, applies it to the study of terrorism and uses four novel classification model comparison techniques to compare and contrast two different forms of terrorism.

Briefly summarized, in tree-producing algorithms, data are iteratively portioned into smaller subsets at “break points” in independent variables, thus producing new tree “branches.” The process is repeated with the goal of creating similar class outcomes. Classification algorithms are used to create models that describe existing data and relationships within that data and predict the behavior of out-of-sample data. The resulting model is expressed as a classification tree that can be easily converted into if-then type rules. Additionally and importantly, when investigating similar phenomena, classification trees can be compared to find similarities and differences between those phenomena. In this article we present four methods for comparing similar classification trees.

This article progresses in six sections. The next describes how classification trees work. Section 3 discusses the potential merits of classification trees in the study of conflict. Sections 4 and 5 discuss the data and comparison methods used to analyze our test case, while Section 6 presents the results. Section 7 concludes.

2. How Classification Trees Work

Rather than achieving a single estimate of a model coefficient as in statistical modeling, machine learning algorithmic approaches leverage computational power to specify or train models through iterative processes to both construct tree representations of data and test out-of-sample records. The C4.5 classification algorithm uses pattern matching, where each variable-value pair is a distinct pattern. The dependent variable is categorical and may have many possible, unordered values, often referred to as classes. In regression analysis calculations are mathematical, where each variable value is represented by a number. The

dependent variable is numeric. The independent variables in classification analysis may be unordered (as in this study), ordered categorical variables or numeric variables, whereas regression requires categorical independent variables to be converted into a number of Boolean (0 or 1) dummy variables. While both classification and regression use probability to make decisions, classification matches the patterns of variable-value pairs between records, while regression looks to fit the variable values to a function or curve in the dimensional space.

Classification can be used to provide knowledge about the structure and interrelationships among data (Osei-Bryson, 2004). The tree structure produced by classification algorithms is a representation of the domain (the phenomenon in the real world under study). In the tree, each branch is composed of items (variable-value pairs), which lead to a leaf item. The leaf represents the dependent variable. The tree has two basic capabilities: (1) it can be used (as we do in this article) to characterize the class (dependent variable) in terms of a branch's items with respect to the dependent variable's value, and (2) it can be used to classify new, unseen records where the class is unknown to predict the record's class by following the branch with the items that correspond to the record's variables and their values.

The tree structure itself is critical to the analysis of comparing two datasets as we are doing in this article. Other methods, such as regression analysis (Schrodt, Yonamine and Bagozzi, 2013), neural networks (Schmidhuber, 2015), random forests (Breiman, 2001) and similar prediction models may provide better predictions, but that is not the goal of this analysis. Regression analysis, neural networks, random forests, and others do not provide the characterization of the dependent variable values in terms of the independent variables and their values necessary for the comparison analyses demonstrated in this article.

Classification algorithms construct classification trees by looking at the past performance of input variables (i.e., independent variables) with respect to an outcome variable (i.e., a dependent variable, known as the class) (Quinlan, 1993). The tree is constructed inductively from records with known values for the outcome variable. Input variables are selected from the dataset to construct the classification tree using a divide-and-conquer algorithm that is driven by an evaluation criterion. Using this selection process, the data are sub-divided until the set of records is indivisible. This repeated division creates the classification tree. As outlined below, the process is completed in four steps: firstly the classification tree is created, secondly the tree is pruned, thirdly the tree is validated. The fourth process is part of the first process as a separate algorithm to determine when a tree branch is initiated. In algorithmic form, the process used in this study proceeds as follows (adapted from Han and Kamber, 2001):

Inputs:

Dataset, D_t

Independent Variable List, IVL

Dependent Variable, V_d

Pruning Factor, PF

Output: Classification Tree

Process:

1. Divide D_t into 10 equal parts D_1 to D_{10}
2. Generate Tree (D , IVL)
 - 2.1. Create Node N

2.2. If the records D_1 all have the same V_d value, C_v

Then N becomes a leaf node with value C_v

endif

2.3. If IVL is empty

Then

For each C_i value of V_d

$|C_i| = \text{count}(C_i)$

endFor

$C_v = \max(|C_i|)$

N becomes a leaf node with value C_v

endif

2.4. Apply $\text{InformationGainAlgorithm}(D_i, IVL)$ to find the “best” splitting criteria, SC

2.5. N becomes SC

2.6. If SC is discrete valued

Then

$IVL = IVL - SC$

endif

2.7. For each value of SC , SC_v

$D_v = \text{set of data where } SC \text{ value is } SC_v$

If D_v is empty

Then

For each V_d value C_i of D_v

$|C_i| = \text{count}(C_i)$

endFor

$C_v = \max(|C_i|)$

Attach a leaf node to N with value C_v

Else

Attach a node to N returned by Generate Tree (D_v , IVL)

endIf

endFor

2.8. Return Tree, CT

3. Prune Tree (CT)

3.1. Select leaf node, LN_i from set of Leaf Nodes

3.2. Find LN parent, LNP

3.3. For each LN_i of LNP

$|C_v| = \text{count } V_d \text{ value, } C_v$

$|C_{v_i}| = |C_v|$

endFor

3.4. $|LNP|C_{v_i}|| = |C_{v_i}|$, LNP count $|C_{v_i}|$

3.5. $LNP_{\text{tot}} = \sum |LNP|C_{v_i}||$

3.6. $\text{Error}_{\text{beforeSplit}} = (LNP_{\text{tot}} - \max(|LNP|C_{v_i}||) + PF) / LNP_{\text{tot}}$

3.7. For each LN_i of LNP

If $C_{v_i} = \text{node } C_i$

Then

$$|C_{c_i}| = |C_{v_i}|$$

endif

$$|LN_i|C_{v_i}|| = |C_{v_i}|, LN_i \text{ count } |C_{v_i}|$$

$$LN_{tot} = \sum |LN_i|C_{v_i}||$$

$$|LNErr_{i}| = LN_{tot} - |C_{c_i}|$$

endFor

3.8. |Splits| = Count LNP LNs

$$3.9. Error_{aftersplit} = (\sum |LNErr_{i}| + |Splits| * PF) / LNP_{tot}$$

3.10. If $Error_{aftersplit} > Error_{beforeSplit}$

Then

For each LN_i

$$CT = CT - LN_i$$

$$\text{Set of Leaf Nodes} = \text{Set of Leaf Nodes} - LN_i$$

endFor

For each V_d value C_i of LNP

$$|C_i| = \text{count}(C_i)$$

endFor

$$C_v = \max(C_i)$$

Set LNP with value C_v

endif

3.11. Return CP_p

3.12. Prune Tree (CT_p)

4. Validate Tree, CT_p

4.1. For each D_1 to D_{10}

For each SC_j

Split D_i on SC_j

If SC_j does not exist

Then

For each D_i

$|C_{D_i}| = \text{Count } D_i\text{'s } C_k \text{ value of } V_d$

endFor

$|C_{\text{tot}_{D_i}}| = \sum |C_{D_i}|$

If $D_i\text{'s } C_i \text{ value of } V_d = D_1\text{'s } C_i \text{ value of } V_d$

Then

$|C_{c_{D_i}}| = \text{count}(C_i)$

endif

endif

endFor

$$|C_{\text{tot}}| = \sum |C_{D_i}|$$

$$|C_c| = \sum |C_{c_{D_i}}|$$

$$\text{Success}_{D_i} = |C_c|/|C_{\text{tot}}|$$

endFor

4.2. Tree Accuracy = Average(Success_{D_i})

4.3. Tree Error = 1 – Tree Accuracy

4.4. Return Tree Error

4.5. Return Tree Accuracy

InformationGainAlgorithm (D, IVL) is called by the Generate Tree algorithm to find the “best” split point for branching the tree. We present this algorithm with explanation. The independent variable selected at any given node in the classification tree construction is the independent variable that will divide the instances into subsets providing the greatest likelihood of reaching the point of indivisibility the quickest. That is, it generates the node, N, having the fewest number of future splits necessary to reach indivisibility. The most desirable attribute at a given point in the tree is the one with the greatest information gain.

Information gain is the change in information entropy from the current state of the set of instances to the proposed state of the set of instances. The entropy is a measure of the randomness of the distribution of the instances in a subset (s) of instances with respect to the dependent variable, D.

$$H(s) = - \sum_{i=1}^n P(v_i) \ln P(v_i)$$

where $H(s)$ is the entropy of a set, s

$P(v_i)$ is the probability that v_i is a value of independent variable i .

In the case of known instances, as in classification tree construction, the probability that an independent variable, v , contains value, v_i , is computed for discrete attributes as

$$P(v_i) = freq(v_i, s) / cnt(s)$$

where $cnt(s)$ is the count of occurrences of s

$freq(v_i, s)$ is the frequency of v_i in s

and where in the classification tree

$$v_i = d_i$$

where d_i is a value of the dependent variable.

When the independent variable contains a range of continuous values, say from 0 to 100, the probability is computed after finding a split point for the values, which divides the instances into two subsets. Instances with independent variable values below the split point are assigned to one subset, and the other instances are assigned to the other subset. The split point is calculated based on the known independent variable values in the instance dataset.

The independent variable's continuous values are sorted as an ordered set $\{c_1, c_2, \dots, c_{n-1}, c_n\}$ and the frequency is determined as the number of instances falling at or below the split point or above the split point.

$$freq(c_i, s) = freq(cnt(c_i \leq SplitPt(c)), s) \text{ for one subset}$$

where $cnt(c_i \leq SplitPt(c))$ is the count of instances c_i less than or equal to the value of $SplitPt(c)$

and

$$freq(c_i, s) = freq(cnt(c_i > SplitPt(c)), s) \text{ for the other subset}$$

where $cnt(c_i > SplitPt(c))$ is the count of instances c_i greater than the value of $SplitPt(c)$

where

$$SplitPt(c) = (c_1 + c_{n-1}) / 2$$

where c_1 is the first value in the ordered set of attribute values

c_{n-1} is the second last value in the ordered set of attribute values.

The entropy is calculated for an independent variable, $H(s)$, and for each resulting subset of instances created, $H(s_i/s)$. The entropy of the subsets is weighted by the ratio of dependent variable values in the subset.

$$H(s_i | s) = - \sum_{i=1}^n (cnt(d_i) / cnt(d)) * H(s_i)$$

where $cnt(d_i)$ is the count of instances with the dependent variable having value i in the subset.

$cnt(d)$ is the count of instances in the parent set.

The information gain, $Gain(s, s_i)$, is the change in entropy from the proposed division.

$$Gain(s, s_i) = H(s) - H(s_i | s)$$

The attribute with the greatest information gain is selected as the node for dividing the data in the decision tree at each node. The information gain with the overall highest gain with

respect to the dependent attribute is the root of the tree. A sample of a part of a classification tree we use in the present analysis is shown in Figure 1. This tree shows five decision nodes (points at which decisions are made). A record would be classified by this tree as having or not having a religious terrorist attack — the outcomes or values of the dependent variable at the leaf nodes. The records in the domain that apply to the partial tree are three branches, which continue to the dependent variable, religious terrorist attacks. All three branches start with the countries' population being medium, Muslim-majority, large geographical area, and eventually reaching the next node, foreign military occupation. At this point, the country-years that have foreign occupation and are before 2001 do not have a religious terrorist attack. Those foreign occupied country-years 2001 or after do have at least one religious terrorist attack. Other country-years follow the same branch up until foreign occupation, at which point the value is no and the period variable is not applicable.

Figure 1

About Here

After the tree is constructed, branches of the tree are converted into rules, shown in table form (Table 1):

Table 1

About Here

The rules provide insight into how the outcome variable's value is dependent on the input variables, and each rule constitutes a selection of records. A complete classification tree provides for all possible combinations of the input variables and their allowable values reaching a single, allowable outcome.

During the tree-construction process, branches may be pruned to increase classification performance. Pruning simplifies the tree by replacing or removing branches (Han and Kamber, 2001; Witten and Frank, 2005). Clearly, pruning introduces errors, but pruning also generalizes the tree, so that new datasets may be evaluated with a known accuracy. Pruning a tree can increase its accuracy.

When the tree is being constructed, it is possible that a given record will follow a branch, but when the leaf node is reached the value of the dependent variable is not the value at the end of the branch. This is reported as an error. Overall the tree has an accuracy rate, the percent of correctly classified records. To test the validity of a tree, a commonly used technique in classification analysis is to construct the tree and then test it with other datasets. With 10-fold cross validation the dataset is randomly divided into 10 equal size parts or folds; using one fold the tree is constructed and the remaining nine folds validate the tree. The accuracy of the validation folds is averaged to determine the tree's overall accuracy.

The goal is to find a classification tree that best represents the dataset being mined. To achieve this goal, a number of classification trees are created at varying levels of pruning. Selection of the tree for further analysis is based on Occam's razor; the simplest tree with the highest accuracy (or the lowest error rate, where error rate equals $1 - \text{accuracy}$) is selected (Osei-Bryson, 2004). The classification tree can be analyzed to determine the likelihood of a similar dataset to follow the same classification criteria.

There are limits to the suitability of C4.5 classification trees (Quinlan, 1993). First, the dependent variable must be categorical. If it is not, definite classes cannot be defined. If there is an overwhelming number of a particular value, that value will cause tree construction to ignore the other values. Likewise, if there is a dependent variable that is underwhelming it can result in an increase in the tree error rate. Second, with respect to the independent variables, if an independent variable has many values, the tree may have many branches with small numbers of records. This can result in overfitting the data. When such overfitting occurs the tree has many branches and very few records follow each branch. This creates a high level of specificity

in the characterization of the dependent variable values with respect to the independent variables values. Such overfitting is mitigated by pruning or by reducing the number of values by binning or grouping them, both of which we have done here.

3. Using Classification Analysis to Study Terrorism

Our example comes from the field of terrorism studies. Until recently, standard approaches to the study of terrorism changed little. With the rapid expansion of information processing capability, computational power, readily accessible software and wealth of machine readable, large-scale terrorism databases, new approaches to understanding terrorism are slowly beginning to be utilized in socio-political analyses of conflict.

Most of the empirical work on conflict (including terrorism) has utilized traditional statistical modeling, which focuses on interpreting coefficient estimates and significance testing, but has little usefulness for predictive modeling (Ward et al., 2010). More recently, scholars have attempted prediction more seriously using classical time series, vector autoregression, event history, survival and rare-events models, but the predictive utility of such models remains questionable.

We propose that machine learning algorithmic approaches have much to offer terrorism studies beyond conventional statistical techniques. Though underutilized in social science, computer scientists have developed various tools to forecast terrorist group activities and macro-level causes (Martinez et al., 2008; Silva et al., 2009; Scime, Murray and Hunter, 2010). In addition, machine learning techniques have already been used to generate forecasts of violent crime rates in the United States (Berk, 2009), civil war onset (Muchlinski et al., 2015)

and state repression (Hill and Jones, 2014). With respect to terrorism, Basuchoudhary and Bang (2019) found that machine learning, such as classification modeling, can help explain terrorism and provide nuanced explanations of terrorism. Bassetti, Caruso, and Schneider (2018) used regression trees to analyze Islamist (religious) terrorist attacks after 2001. Bang et al. (2017) evaluated a number of prediction determining machine learning techniques in their analysis of terrorism, finding that single regression trees, random forests, and bagging predictors are more accurate in determining the likelihood of a terrorist attack. It therefore stands to reason that it would behoove scholars attempting to forecast international conflict and terrorism to continue to experiment with algorithm-based techniques which have proven useful in other disciplines (Beck, King and Zeng, 2000).

Classification analysis, in particular, is an especially useful technique for studying the complex phenomenon of terrorism. Terrorism results from a complex interplay of different factors that co-occur where and when terrorist attacks transpire. Tree-based approaches allow the researcher to unearth those co-occurrences in a way not possible with statistical analyses. In contrast to regression analysis, tree-based approaches not only allow the researcher to determine which variables and their values are important in predicting terrorism's onset but also how those variables are related to each other. Our analysis attempts to determine the relative importance of various political, social and religious variables in determining terrorist incidents and to find and compare classification tree models that best characterize different forms of terrorism using these variables.

Schrodt et al. (2013, 145) identify four specific ways in which algorithmic approaches may be better suited than statistical models for analyzing terrorism. First, because algorithmic

techniques are not sensitive to degrees of freedom, they are generally better at handling “big data” with large quantities of independent variables, including many conflict and terrorism databases used by social scientists. Second, machine learning algorithms are less dependent on rigid assumptions about the distribution of data or how they were generated than more restrictive statistical methods. For instance, much of the data used in conflict studies are class imbalanced, contain complex nonlinear interactions among covariates or are linear in the parameters, which violate fundamental assumptions of standard regression methods and consequently lead to biased estimates. In these cases, flexible algorithmic approaches—which are still capable of modeling data structures common in social science like clustered hierarchical data and time-series cross-sectional data—might produce more reliable findings. Third, in contrast to many statistical methods used in empirical conflict studies, algorithmic approaches are intended to generate accurate predictions—a trait that would seemingly make them attractive to those seeking to understand terrorist patterns and trajectories and offer proposals to policymakers. Finally, algorithmic methodologies approximate the “case-based reasoning” found in qualitative approaches which match patterns of past performance to events observed in a current situation, using the best historical fit to predict the current outcome.

In what follows we demonstrate how classification trees can help researchers understand patterns of different types of terrorism. The literature, to date, has unearthed a number of correlates of terrorism. Yet these studies have not examined which ones apply to specific types of terrorist groups. As a step toward resolving this, we employ an approach that disaggregates religious and secular terrorist groups using a state-level analysis. To illustrate the applicability of classification trees to the study of terrorism we consider an important but

unexamined question that arises from the terrorism literature: whether religious and nonreligious terrorists attack different kinds of countries.

Because religious militants often understand their endeavors as part of a larger-than-life spiritual struggle, the constraints that normally limit secular terrorist groups seem to have little bearing on groups and individuals who commit violence in God's name. It stands to reason, then, that the structural, state-level conditions that enable religious terrorists might not have the same effect on non-religious terrorists and vice versa. Yet the quantitative literature on the state-level causes of terrorism tends to treat terrorism as a homogeneous occurrence and generally does not distinguish between attacks carried out by religious and secular actors (for exceptions see Saiya, 2014; Saiya, 2015a; Saiya, 2015b; Saiya and Scime, 2015; Saiya, 2016a; Saiya, 2016b; Saiya 2018). Thus a disaggregated approach is more promising for understanding the structural conditions that encourage particular forms of terrorist violence. Using tree-based classification analysis on two unique terrorism datasets, we analyze if religious and secular terrorists do indeed target different kinds of states.

4. Data

The dependent variable in this study is whether or not a country experienced any identifiably religious or secular terrorist attacks in each year between 1991 and 2013, inclusive, as this timeframe corresponds to years for which many of the data for our independent

variables are available.¹ Data for the dependent (class) variable, terrorist attacks, are derived from the Global Terrorism Database (GTD), an open-source database of terrorist attacks, hosted at the National Consortium for the Study of Terrorism and Responses to Terrorism (START). The GTD, widely considered the “gold standard” in empirical terrorism research, covers terrorist incidents from 1970-2016 and includes both domestic and international terrorist attacks, making it the most thorough publically-available terrorism dataset. From the GTD data, this study coded every terrorist attack a country experienced in the timeframe under investigation as either “religious” or “secular” if the perpetrator was known. An attack was coded as “religious” if it was carried out by a group that (1) conceives its identity and mission in religious terms (by reference to sacred texts or divine commands), (2) holds a discernible religious ideology or motivation that animates its views towards more temporal/political objectives, and (3) is comprised of a majority of members who belong to that particular religious tradition. If an attack does not clearly fall within these boundaries, it was coded as “secular.”² Cases in which the identity of the perpetrator was unclear were not included in the analysis (roughly 50 percent of total terrorist incidents). This dependent variable is represented in the dataset as two different variables, religious attacks and secular attacks.

¹ We believe using a binary class variable is appropriate for the following reasons. First, our figures show that in 91 percent of all country-years in our dataset, states either experienced “0” or “1” religious terrorist attacks. In 97 percent of our observations, the number of attacks was less than 10. Second, using the present/absent dichotomy ensures that the results are not being driven by single countries that have experienced disproportionately more religious terrorism in recent times (i.e., Iraq and Afghanistan).

² We consulted the “Terrorist Organization Profile” list found at the START website to aid in the coding of attacks. See http://www.start.umd.edu/start/data_collections/tops/.

The study also incorporates a number of standard control variables found in the terrorism literature. While the possible predictors of terrorism are myriad, we have selected state-level explanatory variables that past scholarship has linked to the onset of both forms of terrorism we examine here.

1. *Period.* First, some scholars have noted that the attacks of September 11, 2001 ushered in (or at least epitomized) a new era of terrorism qualitatively different from previous periods of terrorist activity (Rapaport, 2004). Accordingly, the number of terrorist attacks explicitly rooted in religion increased exponentially since 2001. To test the “recentness” aspect of the “new terrorism” thesis, we use a country-year format and then divide those years into two groups: country-years before 2001 (“*B-2001*”) and country-years 2001 and later (“*2001-A*”).
2. *Polity.* Our second independent variable is a measure of procedural democracy taken from the Polity IV database (*Polity*) to assess the effect of electoral politics on terrorism (Marshall and Jaggers, 2011). Some find that democracy inhibits terrorism (Eyerman, 1998; Gurr and Moore, 1997; Regan and Norton, 2005; Abrahms, 2007; Walsh and Piazza, 2010), while others find that democracy encourages it (Schmid, 1992; Eubank and Weinberg, 1994; Eubank and Weinberg, 2001; Li, 2005). The Polity project codes the authority characteristics of states using a scale that ranges from -10 (strongly autocratic) to +10 (strongly democratic). To force the C4.5 algorithm to create more than two branches, *Polity* was binned into three categories coded as “Autocracies” (for values -10 to -6, inclusive), “Anocracies” (-5 to 5) and “Democracies” (6 to 10). Binning variables removes variances in the results based on small variations in the variable values by grouping similar values together.

3. *Foreign Military Occupation.* Our third independent variable is a measure of *foreign occupation* to account for the finding that terrorism is commonly a reaction to the occupation forces of an external actor (Pape, 2003; Collard-Wexler et al., 2014). Data for this variable are drawn from the Collard-Wexler et al. (2014) list of foreign occupations. A country-year under foreign occupation was coded as “YES” or “NO” otherwise.
4. *Poverty.* Fourth, we test for the effect of poverty, which some have found to be related to the onset of terrorism (Ross, 1993; Li and Schaub, 2004; Burgoon, 2006, Krueger and Laitin, 2008; Azam and Thelen, 2008; Basuchoudhary and Shughart, 2010). We include a measure of logged per capita GDP (*Log_GDP*) taken from the United Nations’ statistical database (2013). These values we placed in one of three equal-depth bins, that is, bins with approximately an equal number of values to maintain granularity greater than just the two branches where the tree splits on *Poverty*. The bins were labeled LOW, MEDIUM, and HIGH to ensure creation of three branches for these variables.
5. *Religious Tradition/World Region.* We include two measures, one for predominant religious tradition and the other for world region. Predominant religious traditions are *Protestant, Catholic, Muslim, Eastern, Orthodox, and Other*. Regional values are *Africa, Asia, North America, South America, Europe* and the *Middle East*.
6. Finally, we include measures for a state’s logged population (*Log_Pop*) and area (*Log_Area*) to test the effect of population and geographic size on terrorism. Data are taken from the World Bank Development Indicators (2013). Both variables were grouped into three equal-depth bins (LOW, MEDIUM, HIGH) to ensure creation of three branches for these variables.

Data were arranged into a cross-national, time-series panel format, which is ideally suited for teasing out the potential effects of structural changes on terrorism. We employ data for 175 countries during the timeframe under investigation. Values were collected for the variables for each country and year creating 3648 records in the dataset, each record representing a country-year. From these data, two datasets were derived, one for religious terrorist attacks (where the dependent variable indicates whether or not a country experienced a religiously motivated attack), and the other for secular attacks (where the dependent variable indicates whether or not a country experienced a secular attack). The other variables in both datasets are the same, and the variables have the same values. We deal with the issue of reverse causality by entering the dependent variables with a one-year lag.

5. Methods

This article compares datasets in four ways to investigate the state-level correlates of religious and secular terrorism.³ First, we use proxy classification analysis to see if the complex interactions of structural determinants of both kinds of terrorism are the same. Second, we conduct diversity analysis to determine if the rule-sets derived from the representative religious and secular attack trees differ significantly. Third, we use gain ratio analysis to test the effect of individual independent variables with respect to religious and secular terrorist attacks. Finally, we use variable-branch analysis to compare the depiction of the independent variables between the trees. In this section, we explain how all four forms of analysis work.

Comparing Datasets – Proxy Analysis

³ Although these methods have been published elsewhere, they are assembled here together for the first time.

The first type of analysis used is called proxy classification analysis. As used here, proxy analysis determines if the data representation of the religious attacks dataset can be used to represent the secular attacks dataset, and, if the secular dataset can be used to represent the religious dataset.

There are country-years that have only secular attacks, country-years that have only religious attacks, country-years that have both types of attacks, and country-years that have no attacks. Dividing the data into two datasets allows clear delineation of all the possibilities; and permits the analysis to determine not only if an attack occurred, but also the attack's type and the independent variable values that co-occur with an attack and no attack of that type.

The classification tree segments the dataset based on the variables and their values as characterizations of the data records individually and as a dataset. As discussed in Scime et al. (2015), using a classification tree from one dataset as a proxy for another dataset informs about the similarity of the characteristics of the two datasets. A classification tree created and tested with one dataset (the model dataset) can be used as a proxy for another target dataset if the tree and the target dataset are shown to be sufficiently similar. If so, it can be concluded that the two datasets have the same characteristics and essentially will behave the same in similar circumstances. On the other hand, should the performance of the target dataset be significantly different from the performance of the tree, then it can be concluded that the datasets' characteristics are different and, consequently, the domains are different and will not behave in the same way.

There are three possible outcomes when evaluating for proxy trees. A *two-way* result of tree proxy analysis is when trees created from either dataset can act as a proxy for the other

dataset. A *one-way* result occurs when one dataset creates a tree that can act as a proxy for the other dataset, but the reverse is not true. When neither dataset creates a tree capable of being a proxy for the other; this is referred to as a *zero-way* result.

To determine if there is a significant difference in the performance of the tree model (m) on the target domain's dataset (t), the error of the model and the target domain dataset applied to the model at an established statistical confidence level is calculated to find the confidence interval. The domains are statistically similar when the confidence interval includes zero (Tan et al., 2006). To determine if two datasets are statistically different, the difference and the variance between the errors are calculated (Scime et al., 2015).

$$d = e_m - e_t$$

where d is the difference between the errors

e_m is the error of the tree model

e_t is the error of the target domain dataset

$$\sigma_d^2 \cong \frac{e_m(1-e_m)}{C_m} + \frac{e_t(1-e_t)}{C_t}$$

where C_m and C_t are the number of records in

the model building dataset and the

target domain dataset respectively

At a statistical confidence level, the true difference in the error rates is determined by the standard deviation of the errors, the error differences, and the critical value for the confidence level.

$$\sigma_d = \sqrt{\sigma_d^2}$$

$$d_i = d \pm z_\alpha \sigma_d$$

where z_α is the critical value for confidence level α

At a confidence level of 95%, $z_\alpha = 1.96$. The d_i values define the confidence interval. When the errors are not statistically different, d_i values cross zero, and the tree model can be used as a proxy to evaluate the target domain. The model statistically characterizes the target domain as well as the original domain. Alternatively, when the values for d_i are both positive or both negative, then the tree model does not statistically represent the target domain. In this case, it is concluded that the domains are different with different characteristics and behaviors.

Compare and Contrast Classification Trees – Diversity Analysis

Two classification trees derived from similar, but different datasets, may have characteristics, either items (variable-value pairs) or rules that are the same or completely different. Looking at these comparisons and discrepancies can inform the analysis on the similarity of the classification trees, the datasets, and ultimately the domains.

Classification trees partition a dataset completely. This partitioning allows the analysis to understand the domain and the characteristics of a particular segment. Each root-to-leaf segment that comes from the classification analysis is expressed as a rule. The branches

represent segments of the domain, and analysis may shift from the tree as a whole to individual rules. The rules are composed of one or more variable-value pairs in the premise, and the dependent variable and one of its values in the conclusion. The concluding item, or leaf, is the dependent variable and one of its values.

Individually the rules characterize a value of the dependent variable in terms of independent variables and their values in the dataset. Selection of the interesting rules from the domain's rule-set provides insight into the domain itself. The diversity interestingness measure quantifies the amount of difference between rules in the rule-set (Scime, 2018).

The diverse rule is the rule with variable-value pair (an item) frequencies that occur in very few rules. Items, however, may occur in multiple rules. Any item in the classification tree that branches to two or more items is in two or more rules.

Rules derived from the same branch of the tree have the most in common, but as the branch extends the distance from the root increases. The rule diversity measure (D_R) is the sum of the inverse of a rule's item's commonness in the rule-set (O_i) (Scime, 2018). The inverse of the commonness accounts for the abundance of occurrences of the item. Summing this inverse considers the distance from the root; that is, more terms in the sum for rules with more different items. Diversity is a measure of the distinctness of the combination of variable-value pairs in a rule with respect to the other rules in the rule-set mitigated by the number of items in the rule.

$$C_i = \begin{cases} 1/O_i & \text{if the item exists in a branch of the tree} \\ 0 & \text{otherwise} \end{cases}$$

$$D_R = \sum C_i$$

Where:

O_i = Count of the occurrences of item i in the rule-set

C_i = Inverse of commonness of item i in the rule

Likewise, the rule set diversity is the inverse of commonness of rules within the collection of rule sets where the number of rules, the less pruned rule-set, is the measure of distance (Scime, 2018).

$$C_r = \begin{cases} 1/O_r & \text{if the rule exists in the rule set} \\ 0 & \text{otherwise} \end{cases}$$

$$D_S = \sum C_r$$

Where:

O_r = Count of the occurrences of rule r in the collection of rule-sets

C_r = Inverse of commonness of rule r in the collection of rule-sets

The diversity of a rule is a measure of difference. It is a relative measure with respect to the other rules in the same rule-set. A diverse rule is a rule in which the items are significantly different from each other.

Comparing Datasets – Gain Ratio Analysis

The information gain algorithm used to decide the data division (splits) during classification tree construction can be extended to the gain ratio, which can be used to determine the relative importance of the variables in determining the value of the dependent variables. The results of this analysis can then be compared internally to determine the relative importance of the variables in a dataset and externally to see how the order of importance changes between datasets.

The gain ratio is based on the change in information gain algorithm, discussed previously, as applied to the dataset to find the initial (root) node of the tree. The information gain is adjusted by the number and size of the subsets to create the gain ratio. The independent variable with the greatest gain ratio has the most influence on the dependent variable's value.

The gain ratio is calculated as follows (Han & Kamber, 2001):

$$GainRatio(s) = \frac{Gain(s, s_i)}{SplitInfo(s)}$$

Where $GainRatio(s)$ is the gain ratio of a dataset

$Gain(s, s_i)$ is the change in entropy for a split on variable i , discussed above.

$SplitInfo(s)$ is the mitigating factor for subset number and size.

The $SplitInfo(s)$ is the sum of the subset adjustments for number and size into which the variable divides the dataset without considering the values of the dependent variable.

$$SplitInfo(s) = -\sum_{j=1}^k \frac{n_j}{n} \log \frac{n_j}{n}$$

where n is the number of records in the parent dataset

n_j is the number of records in the subset

k is the number of subsets

When the set of independent variables and their values is the same in two datasets with different dependent variables (or different dependent variable values) the gain ratio order of the independent variables may explain differences between the datasets.

Considering Critical Variables – Independent Variable-Branch Analysis

The independent variable-branch analysis is an extension of critical attribute analysis (Scime and Rainer, 2012) where each independent variable is considered critical. The branches of the classification tree provide insight into how the domain is organized with respect to the co-occurrence of the independent variables with each other and with respect to the dependent variable.

To accomplish this analysis, each branch is “cut-off” at the nodes for the independent variable being considered. The number of records that follow the branch through each sub-branch to the leaf nodes, the dependent variable, are counted. When the number of records reaching a leaf is substantial, a note is made of the independent variable nodes through which the record passed to reach the leaf node. Additionally, if the number of records that follow the branch, but disperse to a lot of leaf nodes, is substantial, a note may be made of the branch’s previous independent variable nodes and values.

For example, the records in the domain that apply to the partial tree in Figure 1 are three branches, which continue to the dependent variable. All three branches start with the

population being medium, the predominant religion as Muslim, the geographical area as large, and ultimately reaching foreign occupation. At this point, the country-years that have foreign occupation and are before 2001 do not have religious terrorist attacks. Those occupied country-years after 2001 do have a religious terrorist attack. Furthermore, there are seven country-years that create each of these two branches. Other country-years follow the same branch up until foreign occupation, at which point the value is no and the period variable is not applicable. There are 111 country-years that follow this branch.

Evaluation of the Period variable would lead to the seven country-years being counted among those having a religious terrorist attack and seven not having a religious terrorist attack. Evaluation of the foreign occupation variable provides seven country-years having a religious terrorist attack and 118 country-years as not having a religious terrorist attack. Notice that country-years are counted multiple times—once for each variable being considered, in this case, period and foreign occupation.

This analysis is conducted on each independent variable separately. Such an analysis gives an indication of the importance and effect of the independent variable and independent variable combinations on the dependent variable.

6. Results

To determine if the environment in a country is the same or different when secular or religious terrorist attacks occur we took a data-centric approach, which resulted in two classification trees, one tree from the religious attack dataset and one tree from the secular attack dataset. In selecting the trees to represent each dataset, a set of classification trees was

created by changing the pruning parameter from 0.05 to 1.0. This resulted in 20 trees for each dataset. The representative tree for the dataset to be used for further analysis is the tree with the lowest overall error rate, where error rate equals $1 - \text{accuracy rate}$. Overall error rate is simply the count of the number of records incorrectly classified versus the total number of records. In the event of a tie in error rate the simplest tree, the one with the fewest branches, is selected.

It is useful to look at the two classification trees individually before using our techniques to compare them. Each tree can be analyzed on its ability to predict attacks with respect to the macro-level characteristics of countries where attacks occur.

For purposes of future prediction of secular terrorist attacks, the most likely countries to be attacked (with a 75% probability) are highly populated, Western European and Catholic. There are 45 records that lead to secular attacks. Overall the secular attack classification tree has an accuracy rate of 84.32%. A multinomial logistic regression analysis on the same dataset yields a 79.33% accuracy rate, and the regression model cannot as easily determine the likelihood and, importantly, the characteristics under which secular attacks may occur.

The religious terrorist attack classification tree shows with 83.5% accuracy that a foreign occupied, low population Middle Eastern country will have religious terrorist attacks. There are 22 records that lead to such attacks. The religious terrorist attack classification tree's overall accuracy is 92.87%, compared to 89.20% for the logistic regression model.

While the accuracy rates of the classification and regression models are close in both cases, and the regression model may be sufficient for prediction, the goal here is not prediction,

but rather to determine if the datasets (secular and religious) represented by the classification trees are similar or different. In other words, do religious and secular terrorist attacks occur under the same conditions? We now look at a comparison of the classification trees to each other using our four analysis techniques: proxy analysis, diversity, gain ratio, and independent variable-branch analysis.

Comparing Models by Proxy Analysis

To complete the proxy analysis, classification tree models were created from each dataset and the most representative model for each dataset was selected. The other dataset was processed against the selected tree to test the similarity between the tree model and the other domain dataset.

The secular terrorist attacks dataset was used to create a series of trees using 10-fold cross validation. The tree with the greatest accuracy (84.32%) was selected as the representative tree for characterizing secular terrorist attacks. The analysis shows that this tree is not a proxy tree for the religious dataset. The religious attacks dataset was passed through this tree and resulted in an accuracy of 79.88%. The confidence interval at the 95% confidence level across the secular tree and the religious terrorist dataset is -0.06197 to -0.02684, which does not cross over zero. This secular tree and the religious dataset are statistically different.

Testing to check if the characteristics of religious terrorist attacks can represent secular attacks, a series of trees were built from the religious terrorist attacks dataset. The selected tree proved to have an accuracy of 92.87%. The religious attacks tree processing the secular attacks dataset results in an accuracy of 76.89%. Evaluation of the statistical significance of this

tree on this dataset also finds statistical difference (confidence interval -0.17584 to -0.14379) with 95% confidence.

From these results, it is clear that the explanatory variables and their values, while the same in the records having come from the same data source, mostly have a different cumulative result when predicting different kinds of terrorist attacks. While this analysis has shown that secular and religious attacks are different, additional analysis can show the *ways* they are different. There are a number of techniques that can be used to compare and contrast the classification models.

Comparing Models using Diversity

Calculating the diversity measure for the rule-sets derived from the representative (the same trees selected for the proxy analysis) religious attack and secular attack trees finds that the rule-sets are significantly different from each other. The comparative diversity measures for the religious rule-set and secular rule-set are 84.5 and 179.5, respectively. Furthermore, there are only three rules in common between the rule-sets:

```
IF Pop_BIN = HIGH
    AND Region = Asia
    AND Religion = Protestant
THEN Attack = NO

IF Pop_BIN = HIGH
```

AND Region = North America

AND Religion = Catholic

THEN Attack = NO

IF Pop_BIN = HIGH

AND Region = Asia

AND Religion = Catholic

THEN Attack = YES

Using diversity analysis on each rule-set identifies the rules that are the most different and most similar to other rules in the same rule-set. Multiple occurrences and uniqueness of factors in the model's rules can be used to identify situations where attacks are likely or not.

In the religious attacks model, the two most diverse rules—those rules whose variables and their values are most different from the other rules—are:

IF Area_BIN = LARGE

AND BIN-Polity = Anocracies

AND Foreign_OccupationYN = NO

AND GDP_BIN = MEDIUM

AND Period = B-2001

AND Pop_BIN = HIGH

AND Region = Middle East

AND Religion = Muslim

THEN Attack = NO

IF Area_BIN = LARGE

AND BIN-Polity = Anocracies

AND Foreign_OccupationYN = NO

AND GDP_BIN = MEDIUM

AND Period = 2001-A

AND Pop_BIN = HIGH

AND Region = Middle East

AND Religion = Muslim

THEN Attack = YES

Notice that the only difference in these rules is the period; the attacks occurred during or after 2001, not before. The most common rule in the religious attack rule-set is

IF GDP_BIN = LOW

AND Pop_BIN = MEDIUM

THEN Attack = NO

In contrast, the secular model's most unique characteristics are found in the rule

IF Foreign_OccupationYN = YES

AND GDP_BIN = MEDIUM

AND Pop_BIN = LOW

THEN Attack = YES

The factors that coincide the most often with secular attacks are found in the rule

IF Period = 2001-A

AND Pop_BIN = MEDIUM

AND Region = Asia

THEN Attack = NO

In short, comparing models using diversity presents yet another indication that the structural roots of religious and secular terrorism are entirely different.

Gain Ratio Analysis

Having established that the general complex interactions that invite religious terrorism are different than those that breed secular terrorism, Table 2 shows the gain ratio analysis results for religious attacks. Table 3 does the same for secular attacks.

Table 2

About Here

Table 3

About Here

Gain ratio analysis tells us that the most important characteristic of a country-year with respect to whether or not it experiences terrorist attacks is the country's population in that year. The third most important characteristic in both cases (religious and secular) is geographic area, and the fourth is region of the world. Poverty (GDP_BIN) and regime type (BIN-Polity) are two of the last three considerations for both datasets and domains.

For the religious dataset, foreign occupation is the second most important characteristic; while for the secular dataset, the period when the terrorist attack took place, before or after 2001, takes second place. Period is the least important characteristic in the religious attacks dataset. Religion is in the middle of the list for the religious attacks dataset, whereas it is the penultimate characteristic in the secular attacks dataset. This is a clear indication of the relative importance of religion in religious terrorist attacks. The varying placements of religion and period in the gain ratio analysis supports the idea that the datasets—and therefore domain—are indeed different. Religion is a more important consideration in religious terrorist attacks, period more important in secular terrorist attacks.

Comparing Models with Independent Variable-Branch Analysis

Finally, we look at the independent variables and how they create branches on the trees. The number of records that followed each branch is considered as well as interesting

observations preceding and following branches of the branch under consideration. Recall that the trees are not 100% accurate. The religious tree is accurate 92.87% of the time. The secular tree is 84.32% accurate. Also, individual records are counted multiple times in this analysis. A record is counted once for each branch of which it is a part.

Period: For the religious tree period does not appear in all the branches of the tree; when it does, it is the deciding split at the end of the branch. For example, see Figure 1. When the period node is reached, two branches are created, one for the value *B-2001* and the other for value *2001-A*. There are four times that branches occur like this. Fifty-four percent of the religious attacks occur after 2001. In the secular tree; period creates splits at different depths on different branches. Overall, period splits occur 11 times at depths ranging from the second to the last split on a branch. Sometimes the branches with period conclude with a secular attack and other times not, depending on how the branch after the Period split continues to split. Thirty percent of the secular attacks occurred after 2001.

Polity: For the religious tree, *Polity* splits occur at or near the end of five branches. Eventually, at the end of the branches that have a *Polity* split, there are 114 religious terrorist attacks and 183 no attacks. In the secular tree, *Polity* causes the tree to split 13 times, at varying tree depths. Of the 705 records that follow these branches 258 result in a secular terrorist attack. For the *Polity* variable, the percent of records that follow a Polity branch and result in an attack is about the same – 38% for religious attacks and 36% for secular attacks.

Foreign Occupation: Considering foreign occupation, religious terrorist attacks occurred 57% of the time there were foreign occupations. It splits this tree four times. The foreign

occupation variable splits the secular tree five times at various points. Thirty-two percent of the secular attacks occurred in country-years where there was foreign occupation.

Poverty: Poverty, represented as the log of the county-year's GDP, has three values and causes five splits into 15 branches in the religious tree. These splits occur at the second to the last level in the tree. Medium and low poverty levels split again in three of the fifteen poverty branches. The high levels of poverty always result in leaf nodes and account for 23.2% of the country-years experiencing religious terrorist attacks. Medium levels of poverty account for 47.8% of country-years having attacks, and the remaining 28.9% of attacks are where the poverty level is low. In the secular tree, there are nine splits on poverty, which lead to 27 branches. Poverty occurs well down the tree. Fifteen of the branches result in leaf nodes. Only 6.9% of the country-year records that have a secular terrorist attack on one of these branches are on a high poverty branch. Medium levels of poverty have 51.5% of the poverty related secular terrorist attacks, and low poverty country-years account for 41.5% of the poverty-related secular terrorist attacks. The evidence from the poverty variable leads to the possibility that high poverty is not as strongly related to terrorist attacks as believed.

Religious Tradition: In the religious tree, religion is the second split of the high and medium population branches (population is the most significant variable and the first split for this tree). Interestingly religion is not a factor in the low population country-years. Ninety percent of religious terrorist attacks occur in the more populated country-years, and 56.5% occur in country-years where the predominant religion is Muslim. Eastern religion country-years account for 17.7% of the religiously based terrorist attacks. The remaining religion branches have less than 10% of the religious terrorist attacks each. In the secular tree, religion

creates splits 14 times in the tree. This results in 84 religion branches, of which 72 lead directly to leaf nodes. This is a clear indication that religious tradition is not a significant factor in secular terrorist attacks. Of the country-years that follow a religion branch, 20.8% of them have secular terrorist attacks. The religion, again, with the most terrorist attacks is Muslim, with 32.9% of the secular terrorist attacks that follow a religion branch. However, other religions have a greater number of secular terrorist attacks than they do religious terrorist attacks. Catholic countries experienced 19.1% of the secular terrorist attacks, Eastern religions 18.5%, and Protestant country-years 16.0%. The remaining secular terrorist attacks are in Orthodox (5.4%) and other religion (8.2%) country-years.

World Regions: The Middle East has the most (112) religious terrorist attacks followed by Asia (72), Africa (54), and North America (9). The majority of the religious terrorist attacks in the Middle East, Asia and Africa occurred in the medium sized, Muslim country-years where the population was high. Asia, Africa, and South America have the most country-years containing secular attacks with 143, 142, and 131, respectively. These locations have the majority of the attacks coming from the branch that begins with high population. A third of the attacks in the Middle East come from the same high population branch. Another third of the Middle East secular terrorist attacks come from a branch of low population, foreign occupied, Muslim country-years with a medium poverty level.

Population: Both the religious and secular trees begin branching on the population variable. This means that population is the most significant indicator of terrorist attacks, either religious or secular. Religious terrorist attacks occur mostly in high population country-years, 80.9%; 59.6% of those in Muslim country-years. Medium population country-years account for

8.2% of the religious terrorist attacks. In small population country-years, all the religious attacks occurred in country-years where the religion is Muslim. Secular terrorist attacks also occur mostly in high population country-years, 64.8%. Medium population country-years have 29.3% of the secular attacks. These attacks are spread out through the secular tree across many branches. The low population country-year secular terrorist attacks primarily occur where there is foreign occupation and the religion is Muslim.

Area: In the religious tree, area causes the tree to branch in the middle of the tree. Medium population country-years only have terrorist attacks in Muslim countries. By far, most religious terrorist attacks, 80.3%, occur in the branch of geographically large countries, and 69% of those attacks are in large population Muslim country branches from there. Secular terrorist attacks also branch in the middle of the tree. The secular attacks do not consistently come from any other variable branches. Most of the secular terrorist attacks occur in large population country-years, 63.3%.

In summary, our state-level analysis suggests that there is something fundamentally different in the targeting choices of religious and secular terrorists. By proxy analysis of the terror attack models, we found that there is a statistically significant difference between the religious and secular terrorist datasets and therefore the domains. This shows that the conditions present in any given country-year that predict religious terrorist attacks cannot predict secular terrorist strikes. The two types of terrorism are distinct. We then looked at the characteristics—the variables and their values—which occurred most and least frequently in the models in the diversity analysis. Here we found only three combinations of characteristics common to both models. The information gain analysis provided insight at the dataset level of

the significant independent variables for each model. With the exception of population, the importance of the variables varied significantly between the models. Using the independent variable branch analysis, we broke the trees down and compared and contrasted the models. Branch by branch, we again found that the factors that co-occur with an attack vary significantly with the type of attack. Combined, the proxy, diversity, gain ratio and independent variable branch analyses support the hypothesis that religious and secular attacks do not occur under the same structural conditions. Our disaggregated approach also cautions against making blanket judgments about the characteristics of states where “terrorism” in general occurs, as much of the existing literature does.

7. Conclusion

The potential of tree-based methodologies has not been fully realized by scholars of socio-political violence but have been used successfully in other disciplines. This article has made the case that classification tree analysis can be used to evaluate the similarities and differences between the datasets and hence the occurrences of secular vs. religious terrorist attacks. The methods shown here offer flexible ways to understand state-level patterns of terrorism. The trees used in this analysis shed light on the different patterns of secular and religious terrorism that traditional regression analysis would be unlikely to capture. While it should not be considered a replacement for conventional statistical methods for understanding terrorism, classification analysis provides a level of information that allows easy and direct interpretation of the data and the domain it represents. We have demonstrated how two datasets can be compared using classification tree representations to determine if the datasets

and domains are similar, in other words, if the datasets and trees may be proxies for one another.

Given major recent trends in the study of conflict such as the information revolution and the increased use of machine learning software, we can expect algorithmic approaches beyond the tree-based approach used here to become a welcome addition to the political analyst's toolkit in the coming years. The utilization of association mining, neural networks, k-means approaches, Hidden Markov models, random forests, bagging and boosting of trees, outlier analysis, and sequence analysis all have the untapped potential of yielding improvements in predictive accuracy and domain understanding in the study of political violence and terrorism. Beyond prediction, the comparison methodologies presented here can assist in understanding the differences in similar domains.

References

Abrahms, Max. 2007. "Why Democracies Make Superior Counterterrorists." *Security Studies* 16(2): 223-253.

Azam, Jean-Paul and Veronique Thelen. 2008. "The Roles of Foreign Aid and Education in the War on Terror." *Public Choice* 135(3/4): 375-397.

Bang, James T., Atin Basuchoudhary, John A. David, and Aniruddha Mitra 2017. "Predicting Terrorism: A Machine Learning Approach" working paper (PredictingTerrorIncidents(WEAI2016).pdf), retrieved from: <https://www.researchgate.net/project/Applying-Machine-Learning-Techniques-to-predict-conflict-terrorism-and-growth>.

Bassetti, Thomas, Raul Caruso, and Friedrich Schneider. 2018. "The Tree of Political Violence: A GMERT Analysis" *Empirical Economics* 54(2): 839-850.

Basuchoudhary Atim and James T. Bang. 2019. "Predicting Terrorism with Machine Learning: Lessons from "Predicting Terrorism: A Machine Learning Approach." *Peace Economics, Peace Science and Public Policy* 24(4): DOI: <https://doi.org/10.1515/peps-2018-0040>.

Basuchoudhary, Atim, and William F. Shughart II. 2010. "On Ethnic Conflict and the Origins of Terrorism." *Defence and Peace Economics* 21(1): 65-87.

Beck, Nathaniel, Gary King, and Langche Zeng. 2000. "Improving Quantitative Studies of International Conflict: A Conjecture." *American Political Science Review* 94(1): 21-35.

Berk, Richard. 2009. "The Role of Race in Forecasts of Violent Crime." *Race and Social Problems* 1: 231-242.

Breiman, Leo. 2001. "Random Forests" *Machine Learning* 45(1): 5-32.

Burgoon, Brian. 2006. "On Welfare and Terror." *Journal of Conflict Resolution* 50(2): 176–203.

Chang, Lin. 2006. "Applying Data Mining to Predict College Admissions Yield: A Case Study." *New Directions for Institutional Research* 131: 53–68.

Collard-Wexler, Simon, Constantino Pischedda, and Michael G. Smith. 2014. "Do Foreign Occupations Cause Suicide Attacks?." *Journal of Conflict Resolution* 58(4): 625-657.

Eubank William and Leonard Weinberg. 1994. "Does Democracy Encourage Terrorism?." *Terrorism and Political Violence* 6(4): 417-443.

Eubank, William and Leonard Weinberg. 2001. "Terrorism and Democracy: Perpetrators and Victims." *Terrorism and Political Violence* 13(1): 155-164.

Eyerman, Joe. 1998. "Terrorism and Democratic States: Soft Targets or Accessible Systems." *International Interactions* 24(2): 151-170.

"Global Terrorism Database. 2014. National Consortium for the Study of Terrorism and Responses to Terrorism (START)", www.start.umd.edu/gtd, 2014.

Gurr, Ted R and William Moore. 1997. "Ethnopolitical Rebellion: A Cross Sectional Analysis of the 1980s with Risk Assessments for the 1990s." *American Journal of Political Science* 41(4): 1079-1103.

- Han, Jiawei and Micheline Kamber. 2001. *Data Mining: Concepts and Technologies*.
Massachusetts: Morgan Kaufmann.
- Hill, Daniel W. and Zachary M. Jones. 2014. "An Empirical Evaluation of Explanations for State Repression." *American Political Science Review* 108(3): 661-687.
- Krueger, Alan B, and David D Laitin. 2008. "Kto Kogo? A Cross-country Study of the Origins and the Targets of Terrorism. In: Phillip Keefer & Norman Loayza (eds). *Terrorism, Economic Development and Political Openness*. Cambridge: Cambridge University Press, 148–173.
- Li, Quan & Drew Schaub. 2004. Economic Globalization and Transnational Terrorism. *Journal of Conflict Resolution* 48(2): 230–258.
- Li, Quan. 2005. "Does Democracy Promote or Reduce Transnational Terrorist Incidents?." *Journal of Conflict Resolution* 49 (2): 278-297.
- Marshall Monty G, and Keith Jagers. "Polity IV Project: Political Regime Characteristics and Transitions, 1800-2010", <http://www.systemicpeace.org/polity/polity4.htm>, 2011.
- Martinez, Vanina, Gerardo Simari G, Amy Silva, Subrahmanian V. 2008. "The Soma Terror Organization Portal: Social Network and Analytic Tools for the Real-Time Analysis of Terror Groups." In Liu H, Salerno JJ, and Young (eds). *Social Computing, Behavioral Modeling and Prediction* (Springer/Wien, Mordestedt), 9-18.
- Muchlinski, David, David Siroky, Jingrui He, and Matthew Kocher. 2015. "Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data." *Political Analysis* 24(1): 87-103. DOI: 10.1093/pan/mpv024.

- Murray, Gregg R., Chris Riley, and Anthony Scime. 2009. "Pre-Election Polling: Identifying Likely Voters Using Iterative Expert Data Mining." *Public Opinion Quarterly* 73(1): 159-171.
- Murray Gregg R., and Anthony Scime. 2010. "Microtargeting and Electorate Segmentation: Data Mining the American National Election Studies." *Journal of Political Marketing* 9(3): 143-166.
- Osei-Bryson, Kweku-Muata. 2004. "Evaluation of Decision Trees: A Multicriteria Approach." *Computers and Operations Research* 31(11): 1933-1945.
- Pape, Robert A. 2003. "The Strategic Logic of Suicide Terrorism." *American Political Science Review* 97(3): 343-361
- Quinlan, J. Ross. 1993. *C4.5: Programs for Machine Learning*. San Francisco: Morgan Kaufmann.
- Rapoport, David C. "The Four Waves of Modern Terrorism." in Audrey Cronin and J. Ludes, eds., *Attacking Terrorism: Elements of a Grand Strategy*. Washington, DC: Georgetown University Press, 2004, 46-73.
- Regan Patrick M. and Daniel Norton. 2005. "Greed, Grievance, and Mobilization in Civil Wars." *Journal of Conflict Resolution* 49(3): 319-336.
- Ross, Jeffery Ian. 1993. "Structural Causes of Oppositional Political Terrorism: A Causal Model." *Journal of Peace Research* 30(3): 317–329.
- Saiya, Nilay. 2014. "Explaining Religious Violence Across Countries: An Institutional Perspective." in Kevin R. den Dulk and Elizabeth Oldmixon (eds.), *Mediating Religion and Government: Political Institutions and the Policy Process*. New York: Palgrave: 209-240.

- Saiya, Nilay. 2015a. "The Religious Freedom Peace." *International Journal of Human Rights* 19 (June 2015): 369-382.
- Saiya, Nilay. 2015b. "Religion, Democracy and Terrorism." *Perspectives on Terrorism* 9 (2015): 51-59.
- Saiya, Nilay and Anthony Scime. 2015. "Explaining Religious Terrorism: A Data-Mined Analysis." *Conflict Management and Peace Science* 32 : 487-512.
- Saiya, Nilay. 2016a. "Religion, State and Terrorism: A Global Analysis," *Terrorism and Political Violence*, DOI: 10.1080/09546553.2016.1211525
- Saiya, Nilay. 2016b. "Blasphemy and Terrorism in the Muslim World." *Terrorism and Political Violence*, DOI: 10.1080/09546553.2015.1115759.
- Saiya, Nilay. 2018. *Weapon of Peace: How Religious Liberty Combats Terrorism*. Cambridge: Cambridge University Press.
- Schmid, Alex P. 1992. "Terrorism and Democracy." *Terrorism and Political Violence* 4(4): 14-25.
- Schmidhuber, Jürgen 2015. "Deep Learning in Neural Networks: An Overview." *Neural Networks* 61: 85-117.
- Schrodt, Philip A., James Yonamine, and Benjamin E. Bagozzi. 2013. "Data-based Computational Approaches to Forecasting Political Violence." In V.S. Subrahmanian (ed.). *Handbook of Computational Approaches to Counterterrorism*. New York: Springer, 129-162

- Scime, Anthony. 2018. "Classification Diversity Measurement." *International Journal of Data Science*, 3(2):107-125.
- Scime, Anthony, Gregg R Murray, and Lance Y. Hunter. 2010. "Testing Terrorism Theory with Data Mining." *International Journal of Data Analysis Techniques and Strategies*. 2 (2): 122-139.
- Scime, Anthony, Gregg R. Murray, Wan Huang and Carol Brownstein-Evans. 2008. "Data Mining in the Social Sciences and Iterative Attribute Elimination." In Taniar, David. Ed. *Data Mining and Knowledge Discovery Technologies*. Hershey, PA: IGI Publishing, 308-332.
- Scime, Anthony and Summer Rainer. 2012. "Finding Interesting Classification Rules: An Application from Education." *Proceedings of the 2012 International Conference on Data Mining (DMIN 2012)*, Las Vegas, NV, July, CSREA Press, 37-43.
- Scime, Anthony, Nilay Saiya, Gregg R. Murray, Steven J. Jurek. 2015. "Classification Trees as Proxies." *International Journal of Business Analytics*, 2(2):31-44.
- Silva, Amy, Subrahmanian V., Vanina Martinez, Gerardo Simari. 2009. "Cape: Automatically Predicting Changes in Group Behavior." In Memon, Nasrullah, Jonathan David Farley, David L. Hicks, and Torben Rosenorn, eds. *Mathematical Methods in Counterterrorism* (Springer/Wien, Mordestedt), 253-269.
- Spielman, Seth. E., and Thill, Jean-Claude. 200). "Social Area Analysis, Data Mining, and GIS." *Computers, Environment and Urban Systems*. 32(2): 110–122.

Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. 2006. *Introduction to Data Mining*.
Massachusetts: Pearson Education.

United Nations Statistical Division. 2013. "National Accounts Main Aggregates Database, 2013,"
<http://unstats.un.org/unsd/snaama/dnllist.asp>, 2013.

Ward, Michael D., Brian D. Greenhill, and Kristin M. Bakke. 2010. "The Perils of Policy by P-value: Predicting Civil Conflicts. *Journal of Peace Research*." 47(5): 363-375.

Walsh, James I., and James A. Piazza. 2010. " Why Respecting Physical Integrity Rights Reduces Terrorism." *Comparative Political Studies* 43(5): 551-557.

Watt, Celia. A., Jill W. Lassiter, and Douglas M. Scheidt. (2009). "The Use of Logistic Regression Analyses and Data Classification Mining to Examine Variables Predictive of Long-term Healthcare Staff Giving Cessation Advice." In Proceedings of the 2009 International Conference on Data Mining." Las Vegas, NV, 561-565.

Witten, Ian H. and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edition. California: Morgan Kaufmann.

World Bank. 2013. *World Development Indicators 2013*. Washington, DC: World Bank.

Figure 1: Partial Religious Terrorist Attacks Tree

Table 1: Partial Religious Terrorist Attacks Tree Rules

Rule	IF	AND	AND	AND	AND	THEN
	Pop_BIN	Religion	Area_BIN	Foreign _Occupation	Period	Religious Terrorist Attack
1	<i>Medium</i>	Muslim	<i>Large</i>	<i>No</i>		<i>NO</i>
2	<i>Medium</i>	Muslim	<i>Large</i>	<i>Yes</i>	<i>B-2001</i>	<i>NO</i>
3	<i>Medium</i>	Muslim	<i>Large</i>	<i>Yes</i>	<i>2001-A</i>	<i>YES</i>

**When a value is not indicated for a rule, that variable's value is not relevant to the rule.*

Table 2: Gain Ratio with Respect to *Religious Attacks*

Gain Ratio Value	Variable
0.0471709	<i>Pop_BIN</i>
0.0438213	<i>Foreign_OccupationYN</i>
0.0344472	<i>Area_BIN</i>
0.0305113	<i>Region</i>
0.0237119	<i>Religion</i>
0.0043076	<i>GDP_BIN</i>
0.0017033	<i>BIN-Polity</i>
0.0000773	<i>Period</i>

Table 3: Gain Ratio with Respect to *Secular Attacks*

Gain Ratio Value	Variable
0.041875	<i>Pop_BIN</i>
0.024980	<i>Period</i>
0.019507	<i>Area_BIN</i>
0.010686	<i>Region</i>
0.006843	<i>Foreign_OccupationYN</i>
0.005485	<i>BIN-Polity</i>
0.004267	<i>Religion</i>
0.000620	<i>GDP_BIN</i>