

Fingerprinting Leading Indicators of WMD Terrorism: An Integrated Modeling Approach

Lawrence A. Kuznar¹, Victor Asal², Karl Rethemeyer², Krishna Pattipati³, Robert Popp¹, Steven Shellman⁴

¹National Security Innovations, Inc.
8 Faneuil Hall Market Pl 3rd Floor Boston, MA 02109-6114
{lkuznar | rpopp}@natlsec.com

²State University of New York, Albany
Rockefeller College of Public Affairs and Policy, Dept. Political Science
vassal@uamail.albany.edu, kretheme@albany.edu

³University of Connecticut
Department of Electrical and Computer Engineering
Krishna@enr.uconn.edu

⁴Security Analysis Enterprises, Inc.
steve@strategicanalysisenterprises.com

⁴College of William and Mary
Government Department
smshel@wm.edu

Abstract

In this paper, we present a holistic, integrated social science analysis of the leading indicators of WMD terrorism that advances what is known about the social conditions that foster WMD terrorism. This exercise illustrates how a multidisciplinary team of researchers, including political scientists, anthropologists and computer engineers combined state-of-the-art methodologies to fuse disparate data sources to produce research findings not traditionally possible. This report describes the basic theories of WMD terrorism tested by researchers, the sources of their data and special challenges they met, the varied analytical methodologies used, and a brief overview of key findings.

Theory

The literature on WMD terrorism is voluminous, but there is actually little research based on analysis of actual empirical data. In part, this is due to the fact that WMD terrorism is, fortunately, extremely rare. However, the few empirical studies that exist do point toward several key variables that have a causal connection to WMD terrorism. Interestingly, state sponsorship and specific religious motives are not statistically related to WMD terrorism (Asal and Rethemeyer 2008, Ivanova and Sandler 2006, LaFree, Dugan, and Franke 2005, Parachini 2001). The factors that are most related to terrorist WMD activity include the number of connections a group has with other violent groups and a group's previous level of lethality

(Asal and Rethemeyer 2008). Globalization appears to create more opportunities to pursue WMD (Asal and Rethemeyer 2008) and democracy may provide greater freedom for WMD terrorists to operate (LaFree, Dugan, and Franke 2005).

These preliminary findings guided the research team's selection of independent variables, focusing team efforts on measures of governance, freedom, group violence and connectivity, globalization, as well as economic conditions that may foster an environment conducive to illegal and disruptive activity (crime, corruption, civil unrest, ethnic and religious factionalism). The research team's efforts were global in scope, analyzing environmental conditions that influenced WMD terrorism in all countries and all violence non-state actor groups for which data were available. A specialized effort in event extraction focused on analyzing environmental conditions in Russia by province level.

A wide variety of dependent variables were chosen, but for the purposes of illustrating our methodology and for presenting some of our sounder findings, we will focus on studies of nuclear smuggling (smuggling fissile nuclear materials) and possessing and plotting to use biological agents.

Data

A variety of databases were used for this study, including standard political science structured databases and unstructured news reports that required processing to create structured databases amenable to analysis.

Structured databases

Structured databases focused on two units of analysis: regions where WMD terrorism takes place, and terrorist groups themselves. For regions, the characteristics of countries were analyzed in order to identify environmental conditions that were associated with WMD terrorism. General data on countries, were gleaned from publicly available data sets, which included International Atomic Energy Agency data, Heritage US Troops stationed abroad data, 2007 World Bank Development Indicators, DSTO, CIA World Factbook, KOF Globalization Scores data, 1970-2007, and especially the Quality of Government (QoG) dataset.

Sources of data on terrorist groups included the Monterey WMD Terrorism Database, Monterey Institute for International Studies <http://montrep.miis.edu/databases.html>, and the Tactical Terrorism Dataset, compiled by the Institute for the Study of Violent Groups (ISVG) at Sam Houston State University. Dr. Asal compiled these data into the Big, Allied and Deadly (BAAD) terror group data set at SUNY Albany. The researchers compiled data on 395 terrorist organizations, and coded organizational variables for the period 1998-2005 as one time period (no yearly data at present).

Missing Data and the Imputation of Values

Nuclear smuggling and bio-terrorism data sources are characterized by sparsity (estimating low probability events ranging from 0.03 to 0.1 for nuclear smuggling data, 0.0027 to 0.013 for bio-terrorism data) and missing factor data (typically, 3-79% of missing values).

We estimate missing data via support vector machine regression (SVMR) and auto-regression. As a good data pre-processing practice, we also normalize the factor data so that it has zero mean and unit variance for each factor. Scaling of data avoids undue influence of factors having large values on conflict assessment. We also perform data reduction via partial least squares (PLS) prior to classification. Data reduction reduces noise in the data, improves classification/forecasting accuracy and computational efficiency, and reduces memory requirements (Choi et al. 2006).

If a few data points are missing for a factor, a three-step auto-regressive model is used to fill the missing values. This method is also used to forecast each factor. If a country has missing values for all the years, SVMR is used to fill the missing values. In this process, the given data is divided into training, validation and test sets. Then, complete patterns (with no missing values) from the training set are selected. For each factor, we develop a nonlinear regression model using support vector machine regression (SVMR). After training SVM regression on complete training patterns, missing values in the training set are filled in sequentially by setting each factor (with

missing values) as the dependent variable. This process is repeated until all the missing values in the training data set are filled. After the process for the training data is completed, the same procedure is repeated on the validation and testing patterns with missing values.

Event Extraction from Unstructured Data

A specialized examination of nuclear smuggling in Russia was conducted to evaluate the environmental conditions that influenced nuclear smuggling by Russian province. The data were obtained from Project Civil Strife using automated methods to extract events of interest from text reports. The source of the text came from 1 million BBC monitor stories (i.e., 4.7 million lines/sentences of text) from 1995-2005. BBC Monitor includes over 500 English & foreign language sources. Dr. Shellman then added the Moscow Times and the Moscow News as data sources.

To extract the text Shellman created actor and verb dictionaries for the Russian case and regions and used Text Analysis By Augmented Replacement Instructions (TABARI) to extract the events of interest. From the texts the researchers coded over 751,000 political events. Events range from positive and negative statements to political negotiations and armed conflict. Special events of interest included nuclear smuggling activities. The repression, protest, and cooperation variables are coded using the CAMEO scheme (see the following web page for more detail: <http://web.ku.edu/keds/data.dir/cameo.html>). Russian socioeconomic data were obtained from the Russia & CIS Statistics Online Database compiled by Planet Inform (<http://www.planetinform.com/Analytic/default.aspx>).

Methods

Information Gain Algorithm

Which factors have the most relevant information for each NS and bio dependent variable? We address this question via the Information Gain (IG) algorithm, informed by social science experts. Information gain is related to the concept of mutual information: how much information does one random variable (a factor in our case) reveal about another one (output or response variable in our case)? To formalize this concept, we need to understand the information-theoretic concept of entropy.

Entropy is a measure of how “disorganized” a set of data related to random variable is; it is a measure of uncertainty. Let $y \in Y$ represent a dependent (response, output) variable, where Y is a discrete random variable (binary or multi-valued). Formally, entropy $H(Y)$ is given by

$$H(Y) = -\sum_{y \in Y} p(y) \log_2 p(y) \quad (1)$$

Factors provide information on a dependent variable, thereby reducing uncertainty in our knowledge of the

dependent variable. Letting $x \in X$ denote a factor. Then, entropy of Y after observing X is

$$H(Y|X) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(y|x) \quad (2)$$

The mutual information (information gain) is related to $H(Y)$ and $H(Y|X)$ via

$$I(X;Y) = H(Y) - H(Y|X) = H(X) - H(X|Y) \\ = H(X) + H(Y) - H(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 \left[\frac{p(x,y)}{p(x)p(y)} \right] \quad (3)$$

Note that when X and Y are independent, $p(x,y) = p(x)p(y)$ (definition of independence), $I(X;Y) = 0$. This makes sense: if they are independent random variables, then X can tell us nothing about Y . Indeed, mutual information is the relative entropy or Kullback-Leibler distance between the joint distribution $p(x,y)$ and the product distribution $p(x)p(y)$.

Mutual information provides a quantitative measure to rank order factors in a decreasing order of mutual information (information gain). If the costs of acquiring data related to factors vary widely, one can rank order factors in decreasing order of information gain per unit cost of acquiring factor data. Formally, if $c(X)$ denotes the cost of acquiring factor X , information gain per unit cost is $I(X;Y)/c(X)$.

Hidden Markov Modeling

HMMs provide a systematic way to make inferences about the evolution of probabilistic events. The premise behind a HMM is that the true underlying sequence (represented as a series of Markov chain states) is not directly observable (hidden), but it can be probabilistically inferred through another set of stochastic processes (or sequences). In our problem, the “hidden” sequence refers to the true dependent variable sequence that describes the behavior of a particular event over time. HMMs are perhaps a natural choice for this problem, because we can evaluate the probability of a sequence of events given a specific model, determine the most likely state transition path, and estimate parameters which produce the best representation of the most likely path. An excellent tutorial on HMMs can be found in Baum (1970).

Integration of HMM and IG for Forecasting

Figure 1 shows how country-specific HMMs for each dependent variable are fused together with normalized (observed or forecasted) factors to produce dependent variable classifications/forecasts. The predicted probabilities from the dependent variable HMM model are additional input factors to the statistical classifier, SVM.

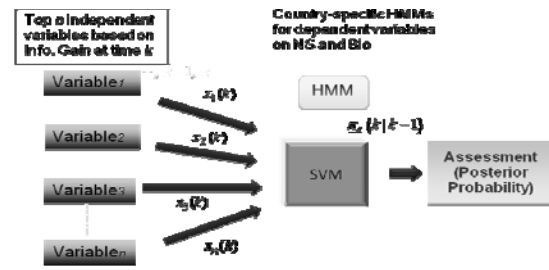


Figure 1: Fusion of HMM probability predictions and additional input indicators using SVM

Support vector machines (SVM) transform the data to a higher dimensional feature space, and find an optimal hyperplane that maximizes the margin between the classes (conflict levels in our case) via quadratic programming (Bishop 2006, Choi et al. 2005). There are two distinct advantages of using the SVM for classification. One is that it is often associated with the physical meaning of data, so that it is easy to interpret. The second advantage is that it requires only a small amount of training data. A kernel function, typically a radial basis function, is used for fitting non-linear models by transforming the data into a higher dimensional space (via “Kernels”) before finding the optimal hyperplane to separate the classes. Data from 1998 to 2001 were used to train the model, and the model was then tested on data from 2002 to 2007.

A nice feature of our modeling process is that it extends naturally to forecasting: The forecasting steps are as follows:

1. Forecast Factors: If the factor is modeled as time series, forecast it using an auto-regressive or nonlinear time series forecasting methods (e.g., SVMR) models.
2. Forecast dependent variables
 - a. Compute predicted dependent variable probabilities from the HMM dynamics.
 - b. Combine factor forecasts with the predicted dependent variable probabilities.
3. Display dependent variable forecast.

Logistic Regression

In addition to SVM, HMM, IG forecasting analysis, more traditional statistical analyses were performed on structured databases. Logistic regression methods were used since the dependent variables were discrete events, and we were interested in predicting the probabilities of these events.

Findings

Space does not permit an exhaustive presentation of the research findings and their specific coefficients, goodness of fit values, etc. However, a brief review of the project’s major findings illustrate that this innovative teaming of researchers, disciplines and methods clearly extended our

understanding of WMD terrorism. Findings were very similar across all studies, regardless of scale (country or Russian province), methodology or dependent variable.

The following factors were statistically related to increases in the probability that nuclear smuggling or bioterrorism will take place in a region.

- A strong black market
- High unemployment
- A globalized economy
- A degree of industrial development
- Sources of Chemical, Biological, Radiological, and Nuclear (CBRN) material
- Ethnic/religious factionalism

Regions with a high likelihood of involvement in nuclear smuggling or bioterrorism tend to be somewhat Westernized and economically developed, but also have high crime rates, worsening economies, and lack security. The countries where these conditions occurred most strongly included, Russia, Caucasus countries, Eastern Europe, Central Asia, Turkey, India. Additionally, the United States, Israel, Japan and the United Kingdom are at risk of bioterrorism; the presence of religious cults in the U.S., U.K. and Japan is a risk factor.

The attributes of terrorist groups that were statistically related to their pursuit of nuclear smuggling or bioterrorism included:

- A history of highly lethal attacks
- A high level of connectedness to other VNSA groups
- Experience with violent terrorist activities

Additionally, religious cults are strongly associated with bioterrorism.

Specific groups predicted on the basis of historical data to be involved with nuclear smuggling and/or bioterrorism include:

- Al-Qaeda
- Abu Sayyaf Group (ASG)
- Jemaah Islamiya (JI)
- Riyad us-Saliheyn Martyrs' Brigade (Chechen Militants)
- Hamas
- Jamatul Mujahedin Bangladesh
- Revolutionary Armed Forces of Colombia (FARC)
- Self-Defense Forces of Colombia (AUC)
- National Liberation Army (Colombia) (ELN)
- Irish Republican Army (IRA)
- PKK
- Armed Islamic Group
- Al Aqsa Martyrs Brigade

Distribution A: Approved for public release; distribution is unlimited (88ABW-2008-0660). The information and opinions stated represent the views of the authors and not the United States Government or Department of Defense.

References

Asal, V., and R. K. Rethemeyer. 2008. "Connections Can Be Toxic: Terrorist Organizational Factors and the Pursuit and Use of CBRN Terrorism." Manuscript, Rockefeller College, University at Albany, State University of New York. .

Baum, L., T. Petrie, G. Soules, and N. Weiss. 1970. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *Ann. of Math. Stat.* 41:164-171.

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. New York: Springer.

Choi, K., M. Azam, M. Namburu, J. Luo, K. R. Pattipati, and A. Patterson-Hine. 2005. Fault Diagnostics in HVAC Chillers using Data-driven Techniques. *IEEE Instrumentation and Measurement Magazine* 8(3):24-32.

Choi, K., J. Luo, K. R. Pattipati, L. Qiao, and S. Chigusa. 2006. "Data reduction techniques for intelligent fault diagnosis." *Proc. Of the IEEE Auotestcon, Anaheim, CA, 2006*.

Ivanova, K., and T. Sandler. 2006. CBRN Incidents: Political Regimes, Perpetrators, and Targets. *Terrorism and Political Violence* 18:423-448.

Lafree, G., L. Dugan, and D. Franke. 2005. The Interplay between Terrorism, Nonstate Actors, and Weapons of Mass Destruction: An Exploration of the Pinkerton Database. *International Studies Review* 7:155-158.

Parachini, J. V. 2001. Comparing Motives and Outcomes of Mass Casualty Terrorism Involving Conventional and Unconventional Weapons. *Studies in Conflict and Terrorism* 24:389-406.