

Classification of Terrorist Group Events in the Philippines:
Location, Location, Location

By

John M. Miller

Joshua Hill

The Institute for the Study of Violent Groups

Sam Houston State University

Huntsville, Texas

March 2010

Abstract

In the ongoing effort to assign culpability for terrorists' violent attacks, the situation in the Philippines is typical. There are large numbers of widely-scattered attacks following varied tactics, techniques and procedures (TTP). Many, but not all, are claimed or definitively attributed publically by authoritative sources to specific groups; many are not. The ISVG database is a huge compendium of public sources on attacks around the globe, and is relatively complete since 2003. Using the data in the ISVG database, we submit the roughly two thousand violent attacks (armed assaults, bombings, hijackings, hostage-takings) reported between 1/1/2003 and 6/3/2009 to a modern update of one of the earliest, but still reputable, classification technique - nearest neighbor. The technique turns out to be remarkably accurate, and serves well to underscore the critical, even definitive, role of location for culpability attribution in the Philippines.

Introduction

There has been significant recent interest in identifying quantitative methods for classifying group culpability. Simply put, researches have been looking for a way to determine the question of "whodoneit" in regards to unclaimed terrorist attacks (Graham et al., 2009; Hill, Miller, & Mabrey, 2009). While several of these efforts have been relatively successful, it is worthwhile to examine the use of different techniques to determine not only their level of accuracy, but also their utility within the context of operations (Mabrey, 2006). The items that seem significant within the context of a given technique may seem counterintuitive to an analyst working in the field. It is within this ability to challenge the preconceptions of experienced personnel that we may find the highest level of utility for these classification techniques.

In an effort to further this area of research, this white paper examines a technique that has not yet been utilized within the context of culpability analysis, nearest neighbor classification. Relying on a dataset that has been proven with other techniques, the analysis demonstrates both the power of the technique for inferring culpability and the importance of identifying powerful predictor variables within a given context to assist an analyst in assigning culpability within a given group.

Literature Review

Culpability

The literature on statistical identification of terrorist culpability is a relatively new addition to the wide literature on terrorism studies in general. However, despite the newness of the topic, great strides have been made in using statistical and machine learning techniques to identify perpetrators of terrorist attacks who may otherwise remain unknown.

Although the application of statistical techniques is relatively new, studies of culpability, and problems in determining it, are not. Traditionally, extensive post hoc investigations are made of an attack, with analytical tradecraft coupled with forensic matches of weapons, explosives, and/or bomb components being the primary mode of culpability assessment (Hill, Miller & Mabrey, 2009). In a secondary method of culpability assignment, counterterrorism analysis make estimates of culpability based on their expertise in terrorist group tactics, operational histories and targeting strategies (Mabrey, 2006).

While anecdotally successful, Heuer (1999) examined intelligence analysts and the various biases they can run against in interpreting events to determine culpability. Further work done in this area has suggested that analysts frequently are unaware of their bias and that it can lead to incorrect

assignment of culpability (Heuer, 1999). As is evidenced by the studies above, these human-centric techniques for the assignment of culpability have significant drawbacks, and have lead to recent developments in the statistical identification of unknown perpetrators of terrorist incidents (Hill, Miller & Mabrey, 2009; Mabrey, 2006).

Statistical Classification for Culpability

While Hale (2005) developed an early empirical framework using Chi-Squared Automated Interaction Detection (CHAID), Mabrey (2006) first presented the classification of unknown terrorist events as an analytical challenge. He examined several techniques for the classification of culpability, with a particular focus on a comparison of classical statistical techniques, to the newer machine learning techniques finding, in general, that the newer techniques performed better. Specifically, he examined several analytical scenarios and classical statistical techniques (logistic regression) were compared to newer machine learning techniques including bagging, boosting, and support vector machines.

In 2009, Graham et al. examined the ability of naïve Bayes classification based on statistically derived "group profiles" to correctly identify culpability on a world-wide dataset. This was significant as it was the first test of a statistical

technique for culpability assessment tested on a world-wide dataset. Additionally, the innovative use of group profiles, driven by the data itself, made the study highly successful correctly identifying roughly 80% of the perpetrators.

Hill, Miller & Mabrey (2009) built on previous work done by both Mabrey (2006) and Graham et al. (2009) by examining a new machine learning technique, Random Forests, in comparison with both traditional statistical techniques (logistic regression) and naive Bayes, as examined by Graham et al. The analysis demonstrated not only the power of Random Forests for classification of culpability, but showed the potential to assist analysts in the field by identifying significant variables.

The current study seeks to expand the knowledge-base on culpability classification through the use of a previously untried technique, Nearest Neighbor, on a large dataset from the ISVG database regarding terrorism in the Philippines.

Terrorism in the Philippines

The Philippines remains a hotbed of terrorist activity, despite years of fighting the problem. These terrorist organizations operating in the Philippines stem from a variety of ideologies, including secular communism, such as the New People's Army (NPA), to the religious separatist Moro Islamic

Liberation Front (MILF). Additionally, the country has seen a significant rise in operations directly related to profit, specifically kidnaps-for-ransom, whether the proceeds are for religious terrorist organizations or simply criminal operations.

Despite the variety of conflicts happening within the Philippines, they are perhaps known best for their religious conflict coupled with "enormous social and political instability" (Church, 2006, 124). Despite these significant social issues, the country remains one of the most important in regards to regional security (Strobel, 2008). Currently, there are three terrorist groups who dominate the scene in terms of violence in the Philippines, the MILF, the NPA, and the Abu Sayyaf Group (ASG).

The MILF, while originally a splinter organization formed out of the Moro National Liberation Front (MNLF) in 1977, has become one of the more dominant terrorist organizations in the Philippines, though they have engaged in sporadic talks with the Philippino government for many years (Vitung & Gloria, 2000). Most recently in 2008, talks with the government begun by the MILF have recently deteriorated into more-or-less outright conflict. Part of the reason for the recent devolvement of peace talks is the MILF's continual provision of sanctuary for terrorist from other regional terrorist groups such as Jemmah Islmiyah (JI), who have previously engaged in operations with

the MILF, such as the Super Ferry Bombing in 2004 that killed 200 people (Mingxin, 2008). Further, the groups inability to control its members has resulted in "rogue attacks," which can be difficult to distinguish from attacks by other groups in the region, particularly ASG.

In contrast to the religious MILF, the NPA's roots as the armed wing of the Communist Party of the Philippines (CPP) remain largely intact (Furtuna, 2007). Also unlike the MILF, there has been virtually no attempt regarding peace with the NPA or the CPP. The government has taken a hard-line against the group, seeking its eradication and frequently engaging in military offensives against the group resulting in a high rate of incidents. The NPA, in turn, has responded by attacking symbols of power within the country, including government buildings to security patrols.

The ASG, rather than having an ideological motivation, straddles the line between terrorist group and criminal gang. The group, while stemming from religious beginnings, has performed extortion and kidnappings for money in recent years, though they have also engaged in more indiscriminate violence, such as bombings (Lingui, 2009). Additionally, and perhaps because of their ideological flexibility, the membership of ASG has frequently overlapped with that of other groups,

particularly MILF, making it difficult to distinguish its members (and attacks) from other organizations'.

In addition to the three main groups mentioned above, there are several smaller groups operating in the Philippines. These smaller groups more frequently engage in criminal activity on a smaller scale, rather than overtly engage in terrorism. Nevertheless, as they tend to engage in some similar activity to the larger groups, particularly kidnappings, they remain an important element of analysis for any culpability assessment of the Philippines. However, despite these groups continuing presence, the low number of attacks that they engage in are low enough so the regional dynamics of conflict remain unaffected by their activities.

Nearest Neighbor

The nearest neighbor classifier is one of the older yet more accurate classifiers utilized within the statistical learning literature (Desarathy, 1991; Fix & Hodges, 1951; Ripley, 1996). It is deceptively simple, makes few assumptions, but frequently performs with accuracy rivaling many far more sophisticated techniques. It is a non-parametric technique with few of the distributional requirements of other classification techniques, thus making it desirable for datasets that depart significantly from normality, which is frequently the case

within the context of terrorism (Desarathy, 1991). Both empirically and theoretically the model has proven that it is an effective discriminator. For instance, Cover and Hart (1967) have shown that, under some conditions, nearest neighbor has a generalization error rate that is bounded above by less than twice the optimal Bayes rule error. In practice, the StatLog project (Michie et al., 1994) found that nearest neighbor actually outperformed all of the other techniques on four of the 23 datasets in the project. It ranked in the top five on seven of the datasets it was tested on. Overall, the technique was above average - ranking below just five of the other techniques.

Nearest neighbor can experience significant problems, particularly in reference to scaling issue. These issues, in turn, are a result of the difficulty in establishing "true" distances between observations in a given parameter space (Gorman & Sejnowski, 1988). Specifically of import within the context of terrorism studies, nearest neighbor struggles when the dataset has either redundant variables (such as multiple measures of a given construct), or irrelevant dimensions (any independent variable that has no statistical relationship to the outcome variable). This is in stark contrast to tree methods such as CART and Random Forest, which may actually benefit from a large number of sometimes very similar variables, drawing from these to construct alternative models (Brieman, 2001). Nearest

neighbor, on the other hand, treats each variable as a viable contributor to the model, thus frequently reducing accuracy. Additionally, unlike naïve Bayes, which while theoretically suffering the same limitation usual performs well with redundant or superfluous variables, nearest neighbor seems to be significantly affected by the introduction of parallel variables (Ripley, 1993). This situation can perhaps be best described as adding a new dimension to the existing space that is perpendicular to all the previous dimensions within the model, when it should be nearly parallel to one or more of the previous dimensions. The placement of the observations in this new, now expanded space, will add distance between the previously extant observations in new and unpredictable ways, thus reducing classification accuracy.

Methods and Data

The Data

ISVG has been collecting data on terrorist attacks since 2004, first under the aegis of the Department of Justice, more recently working with additional partners such as the Department of Defense (ISVG, 2010). The relational database contains the product of numerous trained analysts, proficient in over 30 languages, who scour the world's press, online sites and a number of governmental sources for open-source information

whenever a terrorist attack occurs - wherever it occurs. This unique, human-centric collection process has allowed ISVG to collect detailed information on over 160,000 terrorist incidents, 2,500 hundred groups and over 25,000 individuals (since 2003).

Information collected includes the date of the attack itself, its general type, its location, time, time of day, target characteristics, details such as bomb type, weapon type, groups responsibility, and casualties (deaths and injuries). A total of 1500 variables can be recorded, if present in the public reporting. In addition, if additional information is later exposed on a given incident, the database record is updated. Complete citations are maintained, including the actual text of the source, author and organization responsible for the reporting. The database also includes photos and other media.

This paper explores a specific problem derived from the Philippines as the basis of analysis, in which the number of groups that are assigned culpability, the dependent variable, is four, the New People's Army (NPA), the Moro Islamic Liberation Front (MILF), the Abu Sayaaf Group (ASG) and Other groups. The dataset contains 941 violent events that occurred in the Philippines during the period from 1 January 2004 to 30 June 2008 (the "training" dataset).

Philippines Data

There were a total of 2356 violent attacks in the Philippines between January 2003 and June 2009, with an average of more than one attack a day occurring somewhere in the Philippines. The most frequent attack type was armed assault, but there were significant numbers of bombings and hostage takings as well as is shown in Table 1, below.

Table 1 - Attacks by Type in the Philippines
Incident Type

	Frequency	Percent
Armed Assault	1621	68.80
Bombing	494	20.97
Hijacking	8	0.34
Hostage Taking/Kidnapping	233	9.89
Total	2356	100.00

The attacks in the Philippines were attributed to a total of 17 different groups, however, only three had significant presence over the half-decade examined. One group, in particular, was responsible for nearly half - the NPA. The NPA carried out about a thousand (1012, or 43.0 %) of the violent attacks contained within the dataset. By comparison, the next two groups were responsible for much less activity: the MILF was responsible for 269 (or 11.4%) attacks, and the ASG was responsible for only 177 (or 7.5%) attacks. For the purposes of this analysis the remaining groups were collapsed into an

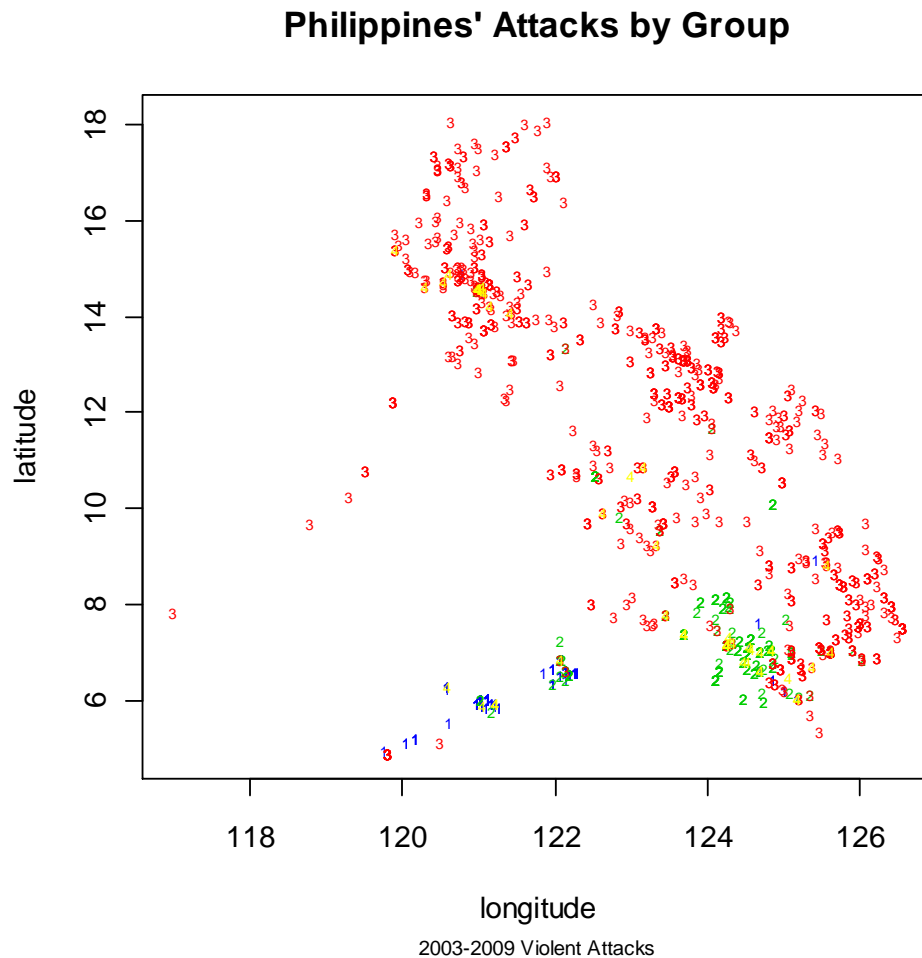
“Other” category. The remaining (n=827) attacks for which no groups were held responsible were omitted from this analysis.

Table 2 - Known Group Attack Frequency

Group Culpability	Frequency	Percent
Valid	827	35.10
Abu Sayyaf Group (ASG) - Philippines	177	7.51
Abu Sofia - Philippines	1	0.04
Al Khobar (Philippines)	9	0.38
Al Qaeda (Afghanistan)	2	0.08
Alex Boncayao Brigade / Breakaway Revolutionary Proletarian Army (RPA-ABB) - Philippines	6	0.25
Bungkatol Liberation Front (BULIF) - Philippines	1	0.04
Communist Party of the Philippines	1	0.04
Farmer's Movement of the Philippines (KMP)	3	0.13
Jemaah Islamiya (JI) - Indonesia	12	0.51
Masses and the Military (Philippines)	6	0.25
Moro Islamic Liberation Front (MILF) - Philippines	269	11.42
Moro National Liberation Front (MNLF) - Philippines	9	0.38
National Democratic Front (NDF) - Philippines	1	0.04
New People's Army (NPA) - Philippines	1012	42.95
Pentagon Kidnap Gang (Philippines)	10	0.42
People's Revolutionary Front (Philippines)	2	0.08
Rajah Solaiman Movement (Philippines)	2	0.08
Revolutionary People's Army (RHB) - Philippines	1	0.04
Waray-Waray Kidnap Gang (Philippines)	4	0.17
Unknown	1	0.04
Total	2356	100.00

A plot of these attacks for which we have longitude and latitude measures (1929 or the 2356) shows that they are spread out over virtually all of the country, Inspection of the groups identified on these attacks (there were 1253 attacks with group culpability known), shows some definite concentrations of group violence, but also that there are overlapping attack zones for some of the groups.

Figure 1 - Philippines' Attacks by Group



Key: 1 - ASG, 2 - MILF, 3 - NPA, 4 - Other. Note that this is the scale used - and does not equate to distance.

Statistical Analysis

A training module (available in the `kkn` package from the R Language (Schliep and Hechenbichler, 2009) was carried out on these data with the objective of optimizing the parameters for the technique. The objective function here is the proportion accurately assigned. There are three critical issues: 1) which dimensions to use; 2) how many neighbors to use; and 3) the kernel to use for the density estimates.

The issue of dimensionality is perhaps the most critical. Many classification techniques (e.g., recursive partitioning, or "trees," especially Random Forests and other recent ensemble techniques) are not seriously impaired by variable specification errors (Breiman, 2001). They are often able to effect high accuracy, even when the predictor variables are highly intercorrelated - even when irrelevant variables are included for prediction. Others like least squares multiple regression, can be seriously degraded by issues of multicollinearity, and misspecified variables. Nearest neighbor is an example of the latter, more sensitive, class.

Nearest neighbor techniques depend on geometry. Each attack is represented as a point in p -dimensional space, where p is the number of predictor variables used. Each additional predictor adds another dimension, usually expressed orthogonal to all the other dimensions in the model. In addition to the

presumption of orthogonality (independence), the researcher is further required to address the scaling problem: the implementations of nearest neighbor usually assumes whatever scale is implied by the actual numerical scores. In our case, longitude and latitude have a natural scaling and an obvious orthogonality nature, which is not the same as implying Euclidean distance; however, over the scale of the one country - the Philippines - can be reasonably approximated as the same as straight-line distance (Desarathy, 1991). When other variables are added in the future though, the size and orientation of the new variables will need to be addressed. For now, we assume that the measure of propinquity is a Minkowski distance $d(x, y)$:

$$d(x, y) = \left(\sum_{i=1}^p (x_i - y_i)^q \right)^{\frac{1}{q}}$$

where x_i and y_i are the i^{th} coordinates of two points (attacks), p is the number of predictor variables, and q is the Minkowski order parameter, equal to 2 for Euclidean distance and one for the Manhattan, or city block, distance.

Repeated runs of the training module with different sets of predictors determined that the optimal set of variables contained only longitude and latitude, and found the standard Euclidean distance to be optimal.

The number of "neighbors" to be used in the analysis was also identified in the context of the training module. It is

possible that most nearest neighbor implementations in the past used only the single closest neighbor - a so-called "1-nn" model. However, this approach is subject to error when a small number of erroneously labeled points are allowed to predict the group of their neighbors. In this case of "outliers" it seems desirable to allow the inclusion of more than one neighbor, then look for a consensus among them. As implemented here, a majority of the "k" nearest neighbors was required for an assignment to be made. The runs of the training module consistently showed the optimal k to be either 10 or 11. We used k = 10 for the results contained in this analysis.

Finally, there was the choice of kernel to use. In fact, the implementation used by this study (Schliep and Hechenbichler, 2009) actually constructed a (nonparametric) density estimate from the k nearest neighbors, and then called the culpability by finding the mode of that density (Scott, 1992).

In kernel density estimation, the density function is estimated as a sum of individual "kernels" or geometric shapes, much like building blocks, piled up wherever there was an observation (wherever an attack took place). The size and shape of those kernels are determined dynamically in order to optimize the overall shape. Kernels under considerations were the rectangular, triangular, Gaussian, and Epanichnikov.

Rectangular and triangular kernels are just that, small boxes and triangles placed on the horizontal axes (latitude and longitude considered as a plane) wherever an attack took place. The Gaussian is a bell-shaped object derived from the normal distribution and the Epanechnikov kernels are small, inverted, parabolic objects. In this case, the kernels were solids, or two-dimensional. The training module found the triangular kernel to be optimal.

Findings

Using the parameters as explicated above resulted in a very successful prediction scheme. The technique was able to accurately predict over 90% of the training data. More impressively, its generalization error (how it would be expected to perform on a new data set) was estimated to be 12.5% with a standard deviation of 1.665%. This generalization error was estimated by a 10-fold cross-validation technique in which the training dataset is divided into tenths and the model is tested by using each of those tenths as a test dataset to gauge the accuracy of the model as generated by the remaining nine-tenths. The reported generalization error is the average of those ten iterations, and its standard deviation is computed the ordinary way from those ten.

The details of how well the nearest neighbor performed are given in the "Hit-or-Miss" matrix below in Table 3. The technique was most accurate in predicting NPA. It correctly predicted 830 (98.8%) of the NPOA attacks. It was nearly as accurate predicting ASG (139 of 154, or 90.3%). It did not do quite as well finding the MILF attacks (165 of 204, or 80.9%) and was very poor predicting the "Other" category (11 of 55, or 20.0%). There appears to have been a slight bias to put ambiguous attacks into the NPA category (882 predictions, compared with only 840 actual attacks).

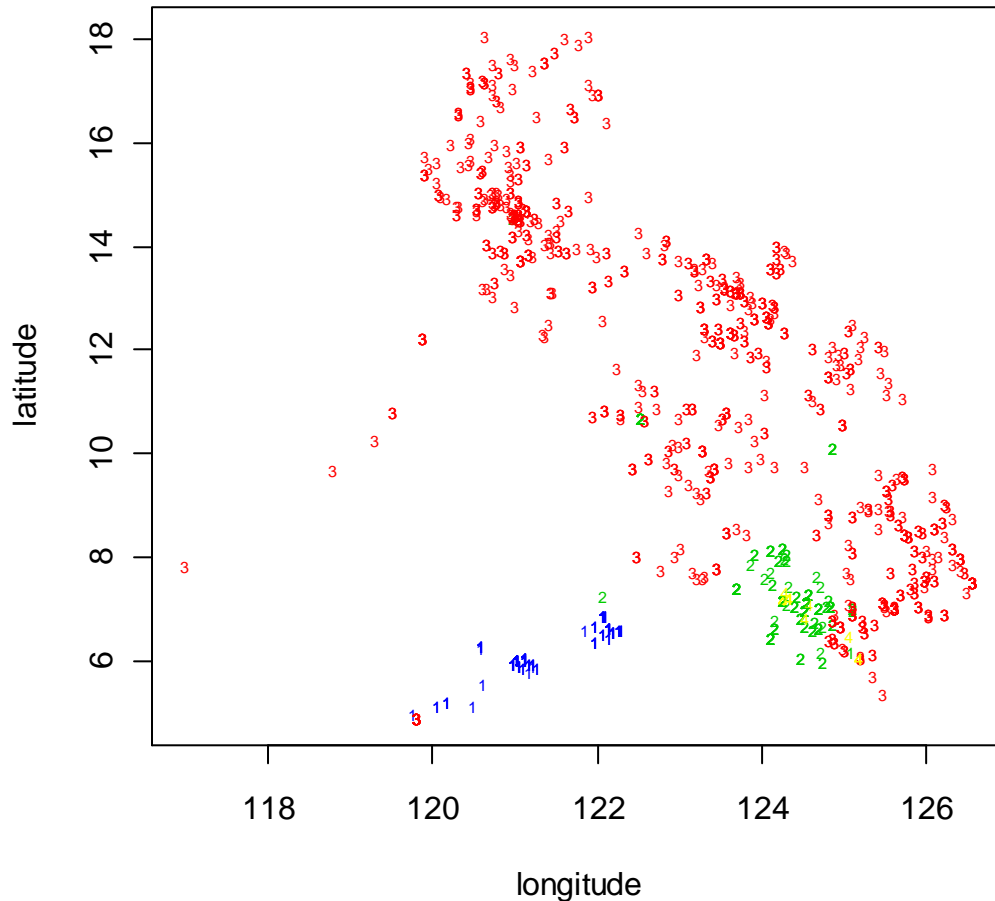
Table 3 - Nearest Neighbor "Hit or Miss" Matrix

Training Data					
Results of 10-Nearest Neighbor Classification					
Predicted					
Actual:	ASG	MILF	NPA	Other	Total
ASG	139	4	10	1	154
MILF	17	166	18	3	204
NPA	3	6	830	1	840
Other	6	14	24	11	55
Total	166	189	882	16	1253

The performance of this classification is visually impressive, as shown in the Map below in Figure 2 (and in comparison with Figure 1, above).

Figure 2 - Nearest Neighbor Classification Results

Nearest Neighbor Classification: Philippines' Attacks



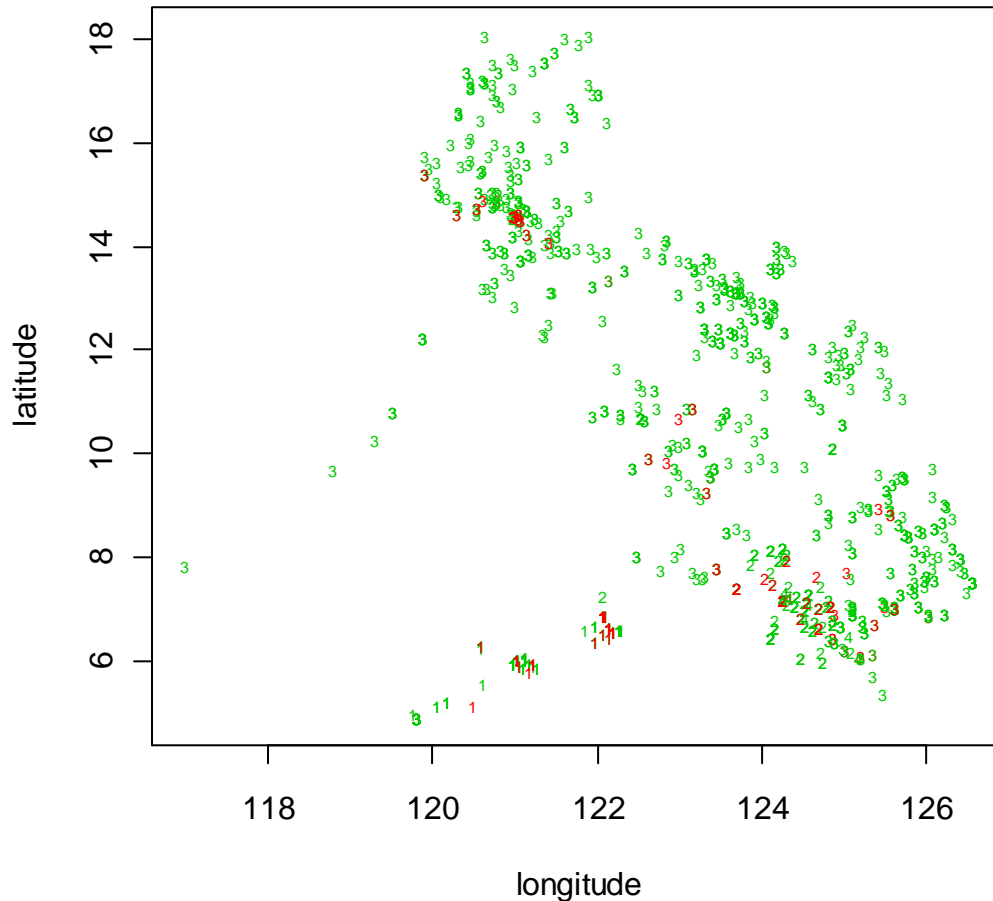
2003-2009 Violent Attacks - 10 Nearest Neighbors

Key: 1 - ASG, 2 - MILF, 3 - NPA, 4 - Other. Note that this is the scale used - and does not equate to distance.

The next map, Figure 3, shows where the erroneous predictions occurred.

Figure 3 - Nearest Neighbor Misclassifications

Nearest Neighbor Classification: Philippines' Attacks



2003-2009 Violent Attacks - 10 Nearest Neighbors - Test Sample

Key: 1 - ASG, 2 - MILF, 3 - NPA, 4 - Other. Note that this is the scale used - and does not equate to distance. The number is the group predicted. Red indicates error.

Discussion and Conclusions

We conclude that the nearest neighbor technique, in its modern guise with kernel-estimated density and cross-validated generalization error estimates, performs extremely well in the case of culpability classification in the Philippines. We have

noted the difficulties that arise when the predictors are not on target, and have indicated that our distance measures are only approximate.

This research is explicitly exploratory in nature. It would be helpful, in the context of future studies, to include theoretically directed variables for confirmation of theory. Additionally, given nearest neighbors' susceptibility to misclassification when misspecified, further integration with models that explicitly identify appropriately strong classification variables, such as Random Forests, would be beneficial to the field.

In addition to combinations with other classification techniques, it would seem appropriate in the next iteration of this research to convert the latitude and longitudes to actual planar coordinates so as to more closely approximate the distances between attacks. However, it should be noted that this raises a much more complicated issue of "effective" distance in which the closest distance between two points may not be a straight line, but would need to take into account infrastructure (roads and waterways) as well as, possibly, political boundaries that might impede traffic. In short, it may not be possible to more accurately measure distance.

References

- Breiman, L. (2001). Random Forests. *Mach. Learn.* 45:5-32.
- Church, Peter. (2006). *A Short History of South-East Asia* 4th ed. John Wiley and Sons, Hoboken, New Jersey.
- Covder, T.M. & Hart, P.E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 13, 21-27.
- Desarathy, B.V. (Ed.). (1991). *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. Los Alamitos, CA: IEEE Computer Society Press.
- Devijver, P.A., & Kittler, J.V. (1982). *Pattern Recognition. A Statistical Approach*. Englewood Cliffs, NJ: Prentice-Hall.
- Fix, E. & Hodges, J. L. (1951). Discriminatory analysis - nonparametric discrimination: consistency properties. Report No. 4, US Air Force School of Aviation Medicine, Random Field, Texas (Reprinted in Desaraty, 1991).
- Gorman, R. P., and Sejnowski, T. J. (1988). "Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets" in *Neural Networks*, Vol. 1, pp. 75-89.
- Graham, S., Ruths, D., Bronk, C., & Subramanian, D. (2009). *The Event-Participant Inference Problem: Using open source information and Bayes' rule to select for the most likely participants in a terrorist incident* [White paper].

- Hale, W.C. (2005). *Twenty-first century terrorism, twenty-first century answers: The why and how of collection, analysis, and dissemination of open source intelligence*. Unpublished doctoral dissertation, Sam Houston State University, Huntsville, Texas. Retrieved July 20, 2007 from ProQuest Dissertations and Theses database.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *Elements of Statistical Learning: Data Mining, inference and Prediction* (2nd ed.). New York: Springer.
- Heuer, R. J. (1999). *Psychology of intelligence analysis*, Center for the Study of Analysis, MacLean, Virginia.
- Lingui, Xu. (2009). Ransom, Politics Embolden Philippine Kidnappers. *Xinhua*. February 8, 2009. Retrieved from http://news.xinhuanet.com/english/2009-02/08/content_10782123.htm on June 12, 2009.
- Mabrey, D.J. (2006). *Tactical terrorism analysis: A comparative study of statistical learning techniques to predict culpability for terrorist bombings in two regional low-intensity conflicts*. Unpublished doctoral dissertation, Sam Houston State University, Huntsville, Texas. Retrieved June 10, 2007 from ProQuest Dissertations and Theses database.
- Michie, D., Spiegelhalter, D.J., & Taylor, C.C. (Eds.) (1994). *Machine Learning, Neural and Statistical Classification*. Chichester, UK: Ellis Horwood.

- Ripley, B.D. (1993). Statistical aspects of neural networks In: *Chaos and Networks - Statistical and Probabilistic Aspects*, (Barndorff-Nielsen, O., Cox, D., Jensen, J. & Kendall, W., eds.) London: Chapman and Hall.
- Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*. Cambridge, UK: Cambridge University Press.
- Schliep, K., & Hechenbichler, K. (2009). KNN: Weighted k-Nearest Neighbors. R package version 1.0-7. <http://CRAN.R-project.org/package=kkn>
- Scott D.W. (1992). *Multivariate Density Estimation: Theory, practice, and visualization*. New York: Wiley.
- Vitug, Marites D. and Glenda M. Gloria. (2000). *Under the Crescent Moon: Rebellion in Mindanao*. Ateneo Center for Social Policy & Public Affairs and Institute for Popular Democracy Quezon City, the Philippines.
- Wenceslao, M. (2008). Commentary: Jemaah Islamiya, an Obstacle to Peace in RP and Beyond. *Philippine Information Agency*. November 17, 2008. Retrieved from <http://www.pia.gov.ph/default.asp?m=12&r=&y=&mo=&fi=p081107.htm&no=13> on June 12, 2009.