Periodic Spatio-Temporal Improvised Explosive Device Attack Pattern Analysis

Matthew Benigni
benigni@gmail.com

Reinhard Furrer
rfurrer@mines.edu

MCS-04-08

July 2008

Department of Mathematical and Computer Sciences
Colorado School of Mines
Golden, CO 80401-1887, USA
Phone: (303) 273-3860
Fax: (303) 273-3875

Periodic Spatio-Temporal Improvised Explosive Device Attack Pattern Analysis

Matthew Benigni
Major, Armor
Dept. of Mathematical Sciences, USMA
West Point, NY

Dr. Reinhard Furrer
Department of Statistics
Colorado School of Mines
Golden, CO

Improvised Explosive Devices (IEDs) are the number one killer of coalition combat forces in the Iraq Theater of Operations (ITO). A unique characteristic of this terrain is that attacks happen almost exclusively on roads. This allows us to reduce location to one dimension and consider historical attacks to quantify periodic, spatio-temporal clusters. The end result for a set of specified routes, is a set of inhomogeneous, bi-variate rate functions that aid the patrol leader in his or her route selection and/or intelligence preparation of the battlefield.

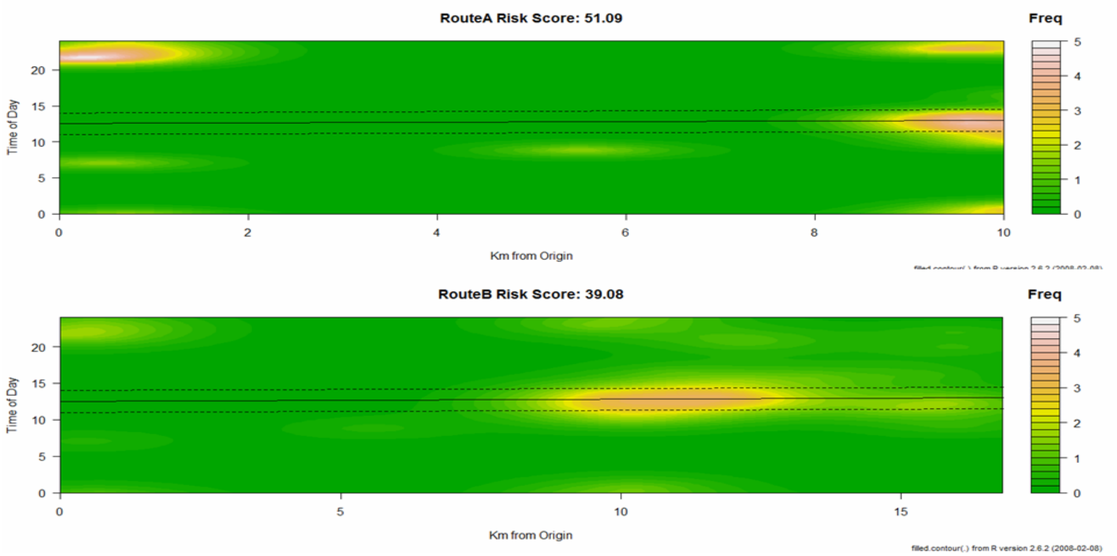

Figure 1: *Shaded contour plots of our estimated attack rate functions for two potential route to a specific objectives. The x-axis depicts road distance from the route origin, and the y-axis depicts time of day in hours. The solid lines depict the predicted routes in time and space and the dashed lines represent the area of interest used to generate risk scores. The risk scores are the integrated attack rates between the two dashed lines.*

1 Introduction

Long-term, counter-insurgent operations in urban terrain force patrol leaders to conduct missions that routinely cross brigade and division boundaries. Intelligence officers at the brigade-level and below commonly assess the likelihood of enemy activity in unfamiliar areas of operation, but objectively assessing risk in unfamiliar areas is time and resource intensive. Although historical data is available, often times we assess tactical risk subjectively.

Furthermore, many of these ‘out of sector’ missions are conducted at the platoon level, and not given significant support from the battalion or brigade staff.

We have developed a methodology that objectively quantifies risk levels associated with specific travel routes at specific times of day. Because attacks occur almost exclusively on roads, we are able to define location in one dimension, similar to modeling defects along a wire. We then replace our second spatial dimension with time of day to characterize events. By isolating a specific stretch of road and analyzing historical enemy activity, we develop a bivariate, inhomogeneous rate function that is proportional to the attack density function along our route in space and time. We can then integrate under this surface to compare risk levels associated with choosing different routes and/or start times. The end product for a patrol leader, depicted in Figure 1, provides a tool that not only allows him or her to quickly and objectively assess risk levels associated with different rout selections, but also indicates areas of higher probability of contact.

This report is structured as follows: Section 2 quantifies data requirements of this new methodology, Section 3 discusses this statistical methodology in detail, Section 4 discusses two different validation techniques and our model’s performance using data from Route Predator and Route Tampa from 2007, and conclusions are given in Section 5.

2 Data Requirements

For this methodology, we define historical attacks for a given route in terms of the following attributes:

- Location: We define location as road distance along our route from the start point or origin of the route.
- Time of Day: time of day in hours
- Date: Julian date

Attack location conversions from MGRS database to an actual road distance from a route origin were completed using SASS, but a computationally inexpensive means to convert and store locations in this format is required to implement this methodology. The date time groups of each attack are obviously readily available in the SIGACTS database.

For this technical report we used all IED events from the SIGACTS database from 01JAN07 to 04FEB08. For future research efforts we could also attach device type and effects, but due to the challenges of working with classified secret information we have not included these attributes in this research (See Section 5).

3 Methodology

Parameter Selection:

This methodology requires the following four parameters:

- *s*: the ratio s defines “close” in terms of space and time. In our model we defined 15 minutes to be analogous to 100 meters in proximity. This ratio allows us to define clusters in the plane in terms of space and time.
- *period*: time period over which we want to develop our pattern. This research focused on *time of day*, but any period could be analyzed using this methodology (weekly, monthly, annual, etc.).
- h_0: the smoothing constant for a standard Gaussian Kernel-Smoother. In this specific case it is the standard deviation for the Gaussian density used to non-parametrically estimate the surface of a the bi-variate rate function generated by our point pattern of attacks.
- *n*: the number of events used to generate modeled risk

In our case we have chosen an h_0 and *n* to maximize the relationship between predicted risk and future attacks. Further analysis of our models sensitivity to these parameters will be offered in Section 5.

Statistical Framework:

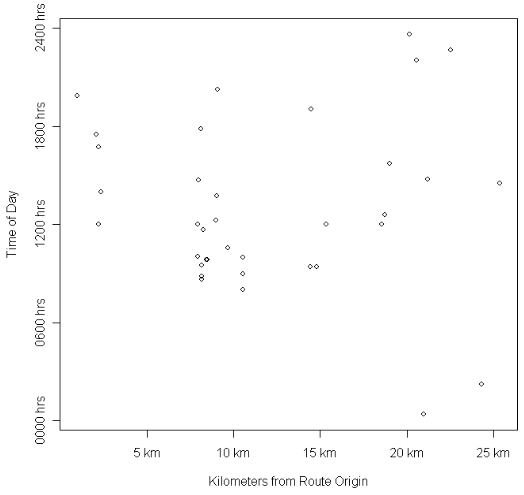

Figure 2*: A sample of 36 attack locations characterized by kilometers from route origin on the x-axis and time of day in hours on the y-axis.*

Because attacks occur almost exclusively on roads, we reduce location to one dimension and reduce our sample space. We then select *n* attacks along a specific stretch of road or route and classify them in terms of road distance from an origin point, time of day, and date of attack. We can select these *n* attacks using either a time threshold as outlined in Keefe and Sullivan (2007), or by fixing the number of attacks and looking as far back in time as needed to find a sample of size *n*. For the purpose of this paper, we will fix our number of points and record how long it took to generate those points. We subsequently develop a "daily rate" for our point process. We can then define a set of *n* points and consider them in the plane as a spatio-temporal point pattern (see Figure 2) and more specifically, as a Cox Process as outlined in Diggle (2007). To do so we make the following assumptions:

Cox Process:

1. $\left\{\Lambda(l,t) : l,t \in \Re^2\right\}$ is a non-negative-valued stochastic process.

2. Conditional on $\left\{\Lambda(l,t) = \lambda(l,t) : l,t \in \Re^2\right\}$, the events form an inhomogeneous Poisson process with intensity function $\lambda(t,l)dtdl$.

Inhomogeneous Poisson Process

1. N(A) has a Poisson distribution with mean $\int_A \lambda(t,l)dtdl$

2. Given N(A)=n, the n events in A form an independent random sample from the distribution on A with pdf proportional to $\lambda(t,l)dtdl$

Analysis of wait times between IED events on Routes Predator and Tampa from the year 2007 validate these assumptions. Figure 3 shows the quantile-quantile plots for IED wait times vs. the exponential distribution for Route Predator in 2007 (left panel) and Route Tampa in 2007 (right panel). Over-all daily IED rates decreased significantly over the course of 2007, and therefore there appear to be outliers toward the larger quantiles of each plot.

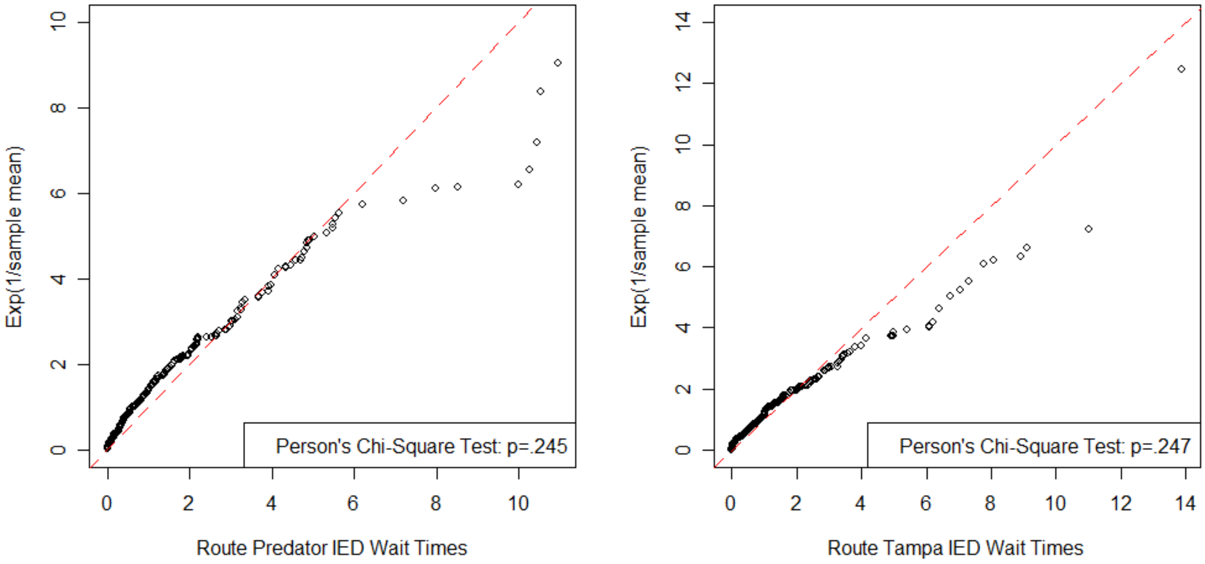

Figure 3. *Quantile-quantile plots for the wait times between IED events on Routes Predator and Tampa respectively for the year 2007. Both compare the wait time to a simulated exponential distribution with the same mean rate.*

Once our set of points is defined in the plane, we scale location by s to define segments in space and time as equally close. This technique is similar to those presented in Knox (1963), where he sets t_o and s_o as thresholds for proximity in time and space respectively. In the case of our models we used $s=.4$ which sets 100 meters and 15 minutes as equally close. We do not define specific thresholds for "closeness", we simply adjust the relationship between space and time to one that is of equal interest to the ground commander. Our analysis technique extends from well established methods by accounting for the periodic nature of our temporal axis. For example an attack that occurs at 2300 hours is not 22 hours away from an attack that occurs at 0100 hrs; it is 2 hours away. To account for the periodic nature of our temporal axis we decompose Euclidian distance to develop the spatio-temporal separation between attacks i and j:

$$attack_i = (\, location_i \, , time_i \,)$$
$$attack_j = (\, location_j \, , time_j \,)$$

$$temporaldist_{i,j} = argmin(\, |time_i - time_j| \, , \, |time_i - time_j + period| \, , \, |time_i - time_j - period|)$$

$$d_{i,j} = \sqrt{(location_i - location_j)^2 + temporaldist_{i,j}^2}$$

This characteristic removes the edges from our plane at both ends of the temporal axis. In effect, we consider these points on a cylinder.

Since our temporal axis is periodic, we no longer need to account for edge effects as we approach 2359 hrs. or 0001 hrs. Additionally, we assume that each route starts and ends at a secure Forward Operating Base or Logistics Support Area and do not need to account for edge effects at the boundaries of our spatial axis. Although this assumption is doubtful for our Route Tampa and Route Predator datasets, it will be valid if this methodology is fielded and routes start and end at secure facilities. On the other hand ordinary corrections for edge effects along the spatial dimension can be easily implemented.

In our problem, we are interested in the rate function λ, a function with respect to location and time. Because of the potential complex structures in the rate function, we use non-parametric methods to fit a rate function $\hat{\lambda}$ to our point pattern as outlined in (Diggle 2003). To do so, we use a Gaussian kernel where the parameter h_0 is defined as the standard deviation of the kernel. We define a fine grid with bins 500 meters by 15 minutes, and predict (or estimate) the rate function $\hat{\lambda}(l,t)$ at the point (l,t). $\hat{\lambda}_j = \hat{\lambda}(l,t)$ in the following manner:

$$\hat{\lambda}(l,t) = \textit{predicted risk at gridpoint}_{(l,t)} = \alpha\sum_{i=1}^{n}\Phi(d_{gridpoint_{(l,t)}},attack_i,h_0)$$

$$\alpha = \frac{n}{\tau\iint_A \hat{\lambda}(l,t)dldt} \Rightarrow \alpha\iint_A \hat{\lambda}(l,t)dldt = DailyAttackRate$$

where:

$\Phi\left(d_{j,i},h_0\right)$ = *cumulative density of a normal distribution with mean=0 and standard deviation=*h_0 *evaluated at* $d_{j,i}$

τ = *the number of days required to generate our set of n attacks.*

A= the area of integration defined by our route velocity and start times (see Figure 5).

Figure 4 illustrates $\hat{\lambda}$ for the set of n=36 points represented in Figure 2. Note the periodicity in Figures 4 and 5 along the temporal edges. Additionally, we scale our smoothed surface such that the area underneath $\hat{\lambda}$ is the daily rate defined by our *n* points.

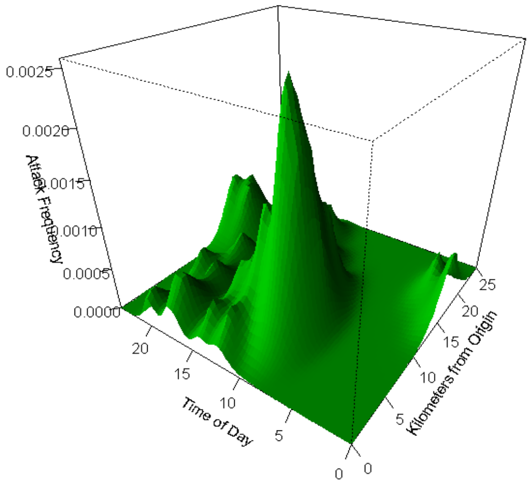

Figure 4*: $\hat{\lambda}$ based on the attack locations characterized in Figure 1, (n=36, h_0=1).*

Scaling $\hat{\lambda}$ allows us to define an area of integration and develop a “risk score”. To do so, we assume constant velocity and establish a bandwidth δ that represents the +/- time under which to integrate. The solid line in Figure 5 depicts the predicted route in time and space; this particular route departs at 1300 hrs at speed 25 kilometers per hour. The dashed lines represent the actual area of integration after we apply our δ.

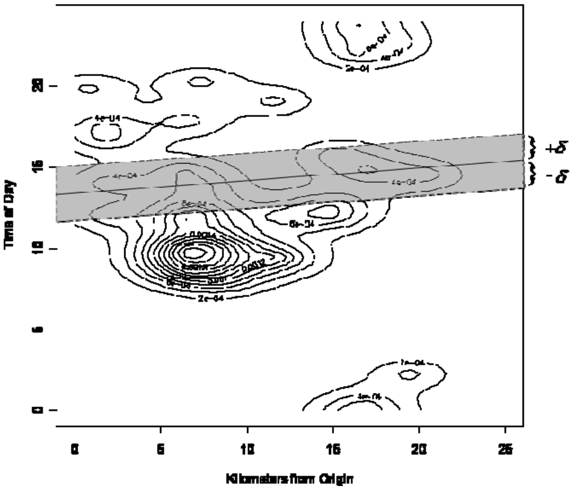

Figure 5*: A contour plot of the same surface pictured in Figure 4,(n=36 and h_0=1). The solid line depicts the predicted route in time and space and the dashed lines represent the actual area to integrate under after we apply our δ. This route departs at 1300 hrs. at a constant velocity of 25 km/h.*

The end result is an objective "risk score" based on travel time of our route and the associated area under the estimated rate function $\hat{\lambda}$. For the specific examples depicted in Figure 1, the risk scores for routes A and B are 51.09 and 31.08 respectively. Notice that the optimal departure times (in the sense of a minimal risk score) are at 1700 hrs and 0300 hrs for Routes A and B yielding scores of 17.08 and 9.48 respectively. Additionally, these risk scores can then easily be used as a tool to assist in route selection. Figure 1 compares two routes that have simultaneous start times (1230 hrs.). It looks as though Route B would pose greater risk, but the density of events and subsequently the elevation of the peak at the end of Route A is significantly higher. Therefore, Route A possesses a higher probability of contact and greater risk. Even with the added ability to look at where clusters occur in time, when we simply use a graphical approach subjective bias can enter the decision. Computing the areas under these surfaces provides the patrol leader with an objective tool to assist in decision making.

4 Model Performance

To measure the predictive power of the estimated rate function $\hat{\lambda}$ and to validate our risk calculation we used two non-parametric tests. Both tests look at the next ten attacks (which are depicted in red in Figure 6) after we have generated a $\hat{\lambda}$ using a specific h_0 and n points. We then generate independent samples of ten completely spatially random (CSR) "non-events" (one such sample is depicted in gray in Figure 6). We record the rate values at these events and non-events, i.e. $\hat{\lambda}(event_i)$ and $\hat{\lambda}(non - event_i)$, i=1,…,10. Subsequently we compare the relationship between them.

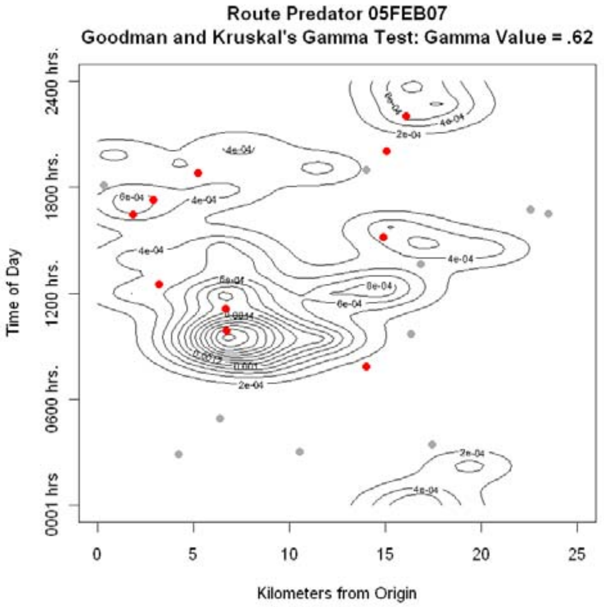

Figure 6*: A contour plot of the surface pictured in Figure 3, (n=36 and h*$_0$*=1). The red points represent the next 10 attacks that occurred after this surface was generated, and the grey points represent a sample of 10 completely spatially random non-attacks.*

Methodology 1 – Goodman and Kruskal's Gamma

For our first test, we define our sample of n attacks as follows:

$$AttackSet_{i,n} = \{attack_{i-n+1},\ldots,attack_i\}$$

n = the number of attacks included in our Attack Set

i = the index number of the last attack included in our Attack set.

For each start point *i*, we estimate $\hat{\lambda}_{i,n}$ based on $AttackSet_{i,n}$ and define

$$a_{i,n,k} = \hat{\lambda}_{i,n}(attack_{i+k}) \text{ for } k=1,\ldots,10$$

We then record the risk at *m=1000* samples of 10 randomly generated non-attacks.

$$\eta_{i,n,k,m} = \hat{\lambda}_{i,n}(non-attack_{k,m}) \text{ for } k=1,\ldots,10 \text{ and } m=1,..,1000 \text{ and define}$$

$$n_{i,n,k} = \frac{1}{1000}\sum_m \eta_{i,n,k,m} \text{ for } k=1,\ldots,10;\ m=1,\ldots,1000$$

Due to the limitations of standard correlation tests when dealing with binary data, we use the ordering based Goodman-Kruskal's Gamma test to measure association (Goodman and Kruskal, 1954) for the following two vectors:

$y = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ where 1 represents an attack and 0 represents a non-attack, and $x = \begin{bmatrix} a_{i,n,1} \\ \vdots \\ a_{i,n,10} \\ n_{i,n,1} \\ \vdots \\ n_{i,n,10} \end{bmatrix}$

The resultant Goodman-Kruskal's Gamma association score is scaled from -1 to 1, and measures the association between "elevation" of future attacks and random events in $\hat{\lambda}_{i,n}$. Although we could use methods based on the second moment to find optimal values for h_0 as outlined in Chetwynd and Diggle (1995), we have chosen to investigate parameter combinations of n and h_o in terms of association with future attacks. To do so we run and test our model on a fine grid of parameter values for $n \in \{25,...,65\}$and $h_0 \in \{.5,.75,...,4\}$, and analyze the characteristics of this surface at each instance. This enables us to see trends in sensitivity and consistency for different parameter combinations in our model.

Results

In Figure 8 we see three surfaces generated from Route Predator data. The x-axis represents the range of n values we observed association scores over and the y-axis represents values of h_0. The model shows sensitivity to sample size n and smoothing constant h_0, but the surfaces are relatively flat. Furthermore, our best choice of h_0 stays relatively fixed at 1, but the best choice of sample size changes event to event. The same observation holds true for Route Tampa at an h_0 of 1.75. The predictive capability of the model depends on how many historical events we use to generate our surface, and the optimal number changes over time.

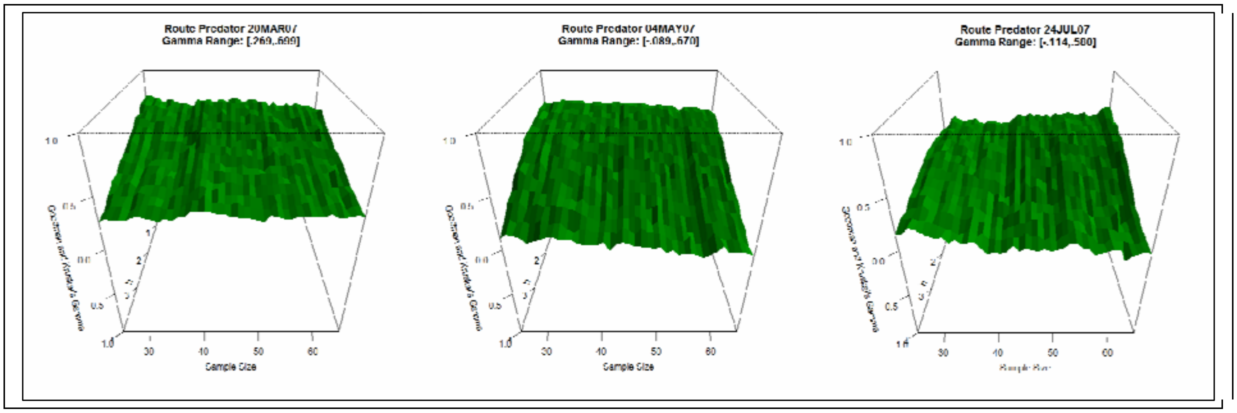

Figure 8*: Goodman-Kruskal's Gamma scores for model performance for Route Predator data on 20MAR07, 04MAY07, and 24JUL07. The plots evaluate association across the following grid of parameter values: n=25,…,65 and* h_0*=[0.5:0.25:4].*

These results are intuitive. Clearly enemy TTPs (tactics, techniques, and procedures) change over time, and that change is likely not continuous or constant. To better understand the sensitivity of model performance with respect to different parameter values we define the following 4 models:

> Maximized Model : chooses the n and h_0 that provide the strongest association with future attacks based on Goodman and Kruskal's Gamma at each:
>
> *Attack Set*$_{i,n}$ *, i= nmax+k,…..,eventmax*
> *nmax= the largest n in our parameter space (65 in our case)*
> *k= the number of test points (10 in our case)*
> *eventmax= the index number of the last event for a particular route in our dataset*
>
> Fixed Model : chooses the n and h_0 that provide the strongest association with future attacks based on Goodman and Kruskal's Gamma across all Attack Sets.
>
> Minimized Model : chooses the n and h_0 that provide the weakest association with future attacks based on Goodman and Kruskal's Gamma at each Attack Set.
>
> Fixed Model : chooses the n and h_0 that provide the strongest association with the $i\text{-}10^{th}$ Attack set. The fixed model is the only model that uses no future knowledge in parameter selection.

Figure 9 depicts the performance of each model throughout 2007. Route Predator is shown in the left panel, and Route Tampa is shown in the right panel. The dates on each communicate the date of the last event included in the Attack Set. These dates indicate the date that this model, generated using $\hat{\lambda}_{i,n}$,would be used in the field. Because of the large range of sample sizes our first Attack Sets start in MAR07 for both the Route Predator and Route Tampa datasets. In both cases the predictive capability of the maximized model appears consistently strong.

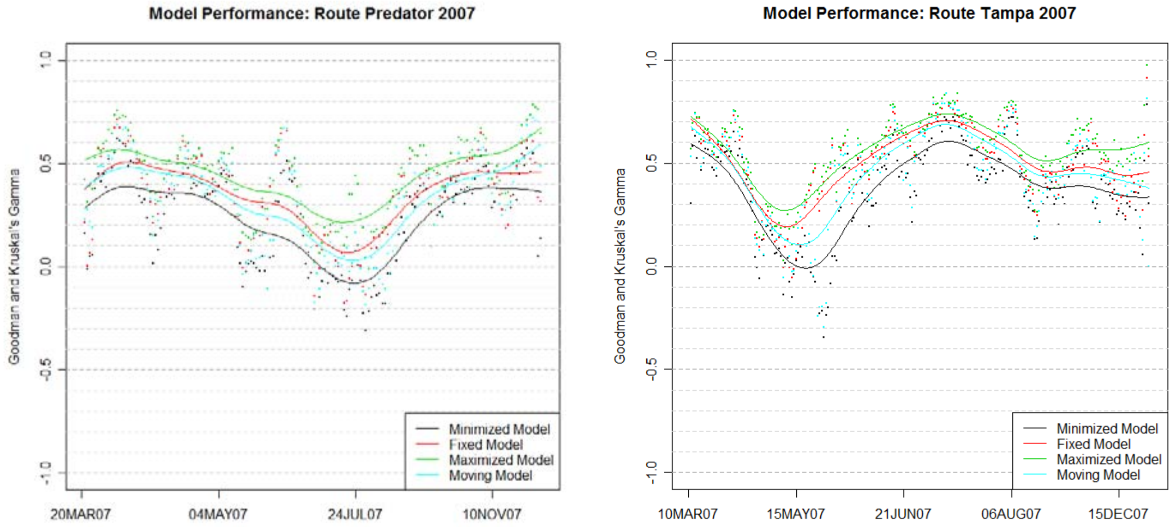

Figure 9*: colored points represent the Goodman and Kruskal's Gamma score for the Minimized, Fixed, Maximized, and Moving Models for Route Predator (left panel) and Route Tampa (right panel). The solid lines represent the smoothed trend of each. The dates on the x-axis indicate the date each model, based on $\hat{\lambda}_{i,n}$,would be used in the field.*

Furthermore, all four models follow the same trend, which indicates that dips in model performance occur when there are changes in the tactical situation as opposed to parametric sensitivity. For example, the dip in performance of the Route Predator model coincide with the end of Operation Arrowhead Ripper in Baquba and the beginning of Muqtada Al Sadr's August cease-fire. These dips in predictive capability found in all models could be due to changes in insurgent patterns do to ongoing blue force actions. Equally as interesting is that the fixed model consistently performs better than the moving model for which a potential explanation is the bias-variance tradeoff. This indicates that our optimal parameter values are fairly stable. The strong associations shown in Figure 9, indicate that the risk score gleaned from integrating under our generated surface $\hat{\lambda}_i(l,t)$ would be a beneficial planning tool for patrol leaders and useful in allocation of ISR assets.

Methodology 2 – Randomization Test

We will refer to our second non-parametric testing methodology as a randomization test. It is based on sums of the estimated risk function evaluated at the 10 next attacks, denoted as $a_{i,n} = \sum_{k=1}^{10} a_{i,n,k} = \sum_{k=1}^{10} \hat{\lambda}_{i,n}(attack_{i+k})$. We compare $a_{i,n}$ to 10,000 independent samples of sums of the estimated risk function $\hat{\lambda}_i$ evaluated at 10 CSR non-events, $n_{i,n,m} = \sum_{k=1}^{10} \hat{\lambda}_{i,n}(non - attack_{k,m})$; $m=1,\dots,10,000$. We then look at the quantile of our attack based sum with respect to the vector $n_{i,n,m}$. Figure 7 is a histogram of 10,000 simulated $n_{i,n,m}$ samples $\hat{\lambda}_i$. The red line depicts $a_{i,n}$, resulting in a quantile of 91**%.**

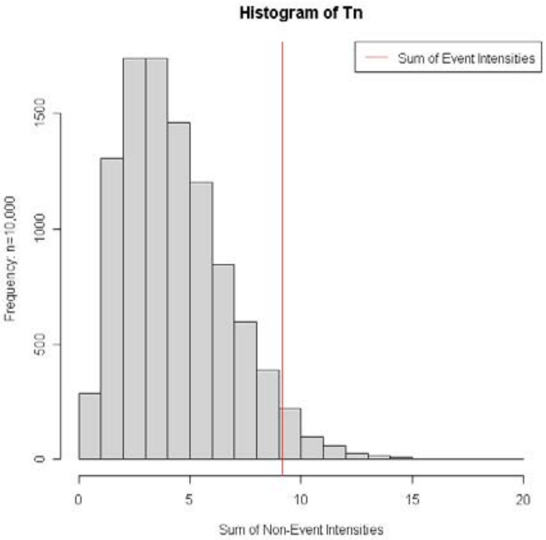

Figure 10*: Histogram of 10,000 simulated $n_{i,n,m}$, based on sums of 10 CSR non-events evaluated at $\hat{\lambda}_{i,n}(l,t)$. The red line depicts $a_{i,n}$, resulting in a quantile of 91***%.**

Results

Our randomization test provides similar results to those found in our Goodman-Kruskal based tests. The predictive power of this test dips during the same time periods. For the Route Predator model we fix $n=26$ and $h_0=1$, and for the Route Tampa model we fix $n=63$ and $h_0=1.75$. The black lines in both panels of Figure 11 represent the quantile values of our $a_{i,n}$ with respect to 10,000 independent samples $n_{i,n,m}$. The date along the x-axis is the date of the i^{th} event. This quantile is above .95 at 69% of our events on Route Predator and 78% of our events on Route Tampa. Both models seem to drop in performance during the build-up of "surge" forces, but appear to regain predictive capability over time. This could indicate that once the "surge" forces were in place and operations were ongoing, enemy TTPs evolved and the past *n* events, once again provided predictive information for future attacks.

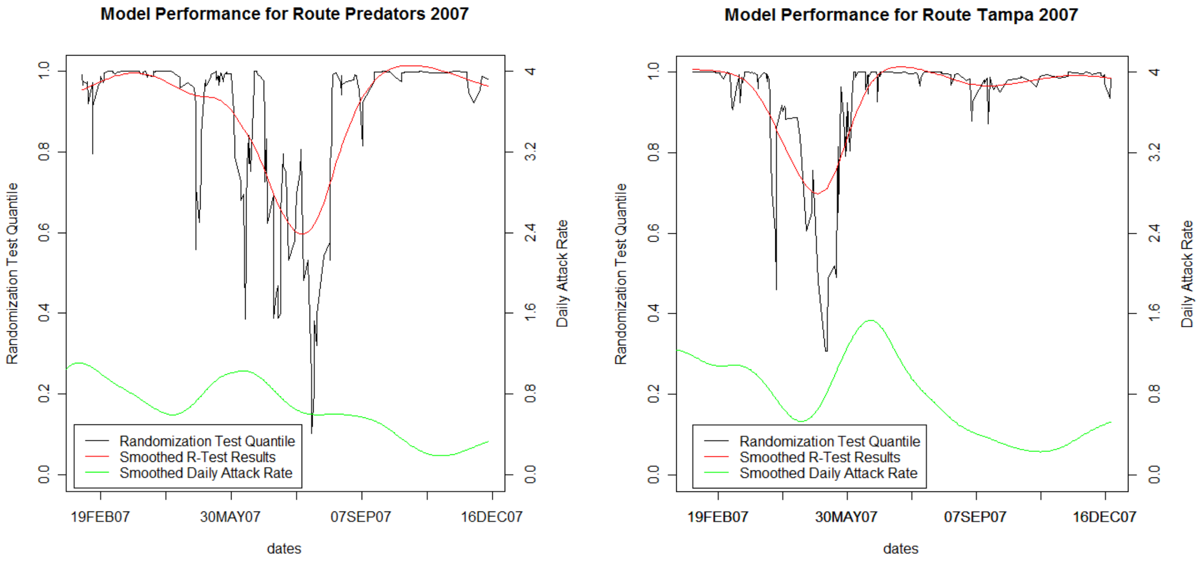

Figure 11*: The left panel illustrates our randomization test quantiles for Route Predator in 2007. The black line is the actual quantile values generated by summing the rate values of the next 10 attack locations and comparing that sum to 10,000 sums of 10 random non-event locations. The red line is a smoothed trend-line, and the green line is the smoothed daily attack rate (with scale on the right). The right panel depicts the exact same plot for Route Tampa 2007.*

6 Conclusion

This research extends standard non-parametric spatio-temporal density fitting techniques, and attempts to assess risk levels for differing routes of travel based on past attacks. As attacks happen almost exclusively on roads we reduce the spatial location to one dimension and embed a cyclical, temporal axis. The attack rate function is estimated with a kernel approach. Additionally, we develop two tests to measure model performance illustrated with current data from the Iraqi Theater of Operations. Both testing procedures provide strong evidence that our statistic based on $\hat{\lambda}_{i,n}$ has predictive power in the short term. In both testing procedures model performance drops during the periods of major changes in battlefield conditions. It is intuitive that the last *n* events would hold less predictive information during more dynamic periods of combat operations, and the model does not account for blue actions. In other words, if blue forces chose to deny a piece of terrain, our model makes no adjustment to the likelihood of attack at that place and/or time. In fact, these drops in correlation/association offer useful information as well. These "dips" in model performance quantify changes in enemy activity. It is also possible that additional noise could be removed from this model by filtering device type. It is likely that sub-surface devices cluster spatio-temporally in different "locations" than surface laid devices. Due to the complications of dealing with secure data, this analysis did not account for device type. For similar reasons we did not model casualties, but this methodology could easily be extended to minimize likelihood of casualties instead of likelihood of attack.

This methodology is a computationally inexpensive tool able to assist in tactical intelligence preparation of the battlefield. Not only does it provide an objective assessment of one route's probability of contact versus another's, it also helps the analyst or patrol leader quickly visualize historical clusters in space and time. In the Figure 1 example, the natural question would be, "Why not change our SP time to avoid the large clusters of events centered at 1200 hrs, 11 kilometers into Route B?" These are vital pieces of information that are particularly elusive when a patrol leader is planning an out of sector operation.   Even the models' dips in performance provide useful information to analysts. In effect, our model identifies when there are significant changes to how IED's are clustered in space and time, and could validate intelligence indicating a change in insurgent cell structure in the area.

Near term research goals should focus on models that filter for specific device types and/or casualty numbers. Due complications associated with classified information at the Colorado School of Mines, we did not consider these models for this paper. There is a strong possibility of removing additional noise from our data once we add this information to our model, and Major Benigni's follow-on assignment to the Department of Mathematical Sciences at the United States Military Academy will remove these information security constraints.

Further analysis also needs to be done on roads that do not have attack frequencies as high as those on Route Predator and Route Tampa. Additionally, because we can define any road segment in this fashion, we could extend this concept to establish arc weights for a network of roads. Ultimately we could find "shortest path" with respect to casualties using conventional network path algorithms.

7 References

Chetwynd, A.G., Diggle, P. J. (1995), “Second Order Analysis of Space-Time Clustering,” Statistical Methods in Medical Research, 4, 124-136.

Diggle, P.J. (2003), Statistical Analysis of Spatial Point Patterns, New York, NY: Oxford University Press.

Diggle, P.J. (2007), “Spatio-Temporal Point Processes: Methods and Applications,” Statistical Methods for Spatio-Temporal Systems, Boca Raton, FL: Chapman & Hall.

Goodman, L., Kruskal, W., (1954), “Measures of Association for Cross Classifications”, Journal of the American Statistical Association, 49, 748-751.

Keefe, R., Sullivan, T. (2007), “Building Time Sensitive Clusters in Time and Space”, Technical Report, JIEDDO.

Knox G., (1963), “Detection of Low Density Epidemicity”, British Journal of Preventative and Social Medicine, 18, 121-127.