Discrimination of Munitions and Explosives of Concern at F.E.Warren AFB using Linear Genetic Programming

Frank D. Francone
RML Technologies Inc. and Chalmers Univ. of Technology
7606 S. Newland St.
Littleton CO 80128
+1.720.981.8710
fdf@francone.com

Larry M. Deschaine
Science Applications Int. Corp and Chalmers Univ. of Technology
272 Parker Rd.
Edgefield SC 29824
+1.706.951.2750
Larry.M.Deschaine@SAIC.com

Jeffrey J. Warren
Science Applications International Corp.
6310 Allentown Boulevard
Harrisburg, PA 17112
+1.717.901.8828
Jeffrey.J.Warren@SAIC.com

ABSTRACT

Removing underground, unexploded bombs, mortars, cannon-shells and other ordnance ("MEC" or "UXO") from former military ranges is difficult and expensive. The principal difficulty is discriminating intact, underground ordnance from other metallic items such as fragments of exploded ordnance ("Clutter"), magnetic rocks, and "historic" items such as horseshoes, barbed-wire, and refrigerators.

This study represents the first, large-scale, blind-test of MEC discrimination technology on production-grade, survey-mode data from the cleanup of a real impact site. The results reported here significantly advance the state-of-the-art in MEC discrimination over alternative, reported forward-modeling/inversion approaches to performing MEC discrimination.

We combined Linear Genetic Programming (LGP) and statistical analysis to process data from the cleanup of 600 acres of the F.E.Warren Air Force Base. These data contained almost 30,000 targets of interest identified by geophysicists, including three-hundred thirty-two 75mm projectiles (75mm) and 37mm projectiles (37mm). A little under one-third of the groundtruth was held back by the customer for blind-testing. Our task was to discriminate intact 37mm's and 75mm's from the clutter by ordering the targets from most-likely to be MEC to least-likely to be MEC in what is referred to as a "prioritized dig-list".

We identified all 75mm's by 27% of the way through our prioritized dig-list. We identified all 37mm's by 64% of the way through the prioritized dig list. Thus, depending on ordnance type, we reduced the number of targets that had to be excavated (false alarms) to clear the entire site by between 36% and 73%.

1. INTRODUCTION

Buried unexploded ordnance ("UXO") (or Munitions and Explosives of Concern ("MEC")) on closed military bases poses a hazard to life-and-limb and further prevents huge tracts of land—frequently urban—from being returned to civilian use. The main barrier to cleaning-up closed military bases is cost. The Department of Defense ("DoD") stated: "The UXO cleanup problem is a very large-scale undertaking involving 10 million acres of land at some 1400 sites."[1] One of the key problems is, according to DoD, ". . . instruments that can detect the buried UXO's also detect numerous scrap metal objects and other artifacts, which leads to an enormous amount of expensive digging. Typically 100 holes may be dug before a real UXO is unearthed!"[6]. It is costly to excavate each target location that *might* contain a MEC item. So most of the cost of base-cleanup is spent digging empty holes, horseshoes, magnetic rocks and "MEC Clutter" or "Clutter"—small pieces of ordnance that flew apart on detonation or impact and pose no explosive hazard.

This study represents the first, large-scale, blind test of MEC discrimination technology on production-grade, survey-mode data from the cleanup of a real impact site. The results reported here significantly advance the state-of-the-art in MEC discrimination over alternative forward-modeling/inversion approaches to the subject.

1.1 Digital Geophysical Sensors

Geophysicists locate possible MEC items on a site by pulling passive magnetic or active electromagnetic sensors across the site in parallel lines (also called "transects") spaced 0.5 to 1 meter apart. These data are plotted spatially and potential MEC items show up as anomalies in an otherwise reasonably flat signal field. This process is referred

to as Digital Geophysical Mapping ("DGM"). The anomalous regions in the DGM are referred to as targets. This paper describes a process for determining which Targets are MEC and which are harmless metallic objects.

In this project we used the signal from a four-channel, active electromagnetic sensor, the Geonics EM61 MK2.[1] It records the magnetic signal induced in an underground object by an electromagnetic pulse at three time-decay intervals at a lower coil. The fourth channel is measured at a coil placed above the lower coil—the "upper coil."

MEC items range in mass from a few ounces to 2000 pounds. Almost all buried metallic clutter items on a site fall in the same weight-range (the vast bulk in the lower end of that range). So the task of discriminating between smaller MEC and Clutter cannot rely on mass—it has to identify objects with typical MEC characteristics. For example, most MEC is roughly cylindrical and has a length to diameter ratio of four or five to one. By way of contrast, magnetic rocks and metallic clutter are normally irregular and less elongated.

1.2 Previous Work

Two approaches to MEC discrimination have shown promise. In one approach, a parameterized forward-model derived from the physics of magnetometers and electromagnetic induction is derived. This approach uses theory to reduce all information about a given Target to the set of parameters (between 2 and 5 parameters) that drive the forward model. To discriminate a Target, the forward-model is inverted and the parameters that best fit the actual measured DGM of a Target are extracted. The parameters are then used in a variety of ways to classify the Target as MEC or not-MEC [3][4][9][15].

The advantage of the forward-modeling/inversion approach is that it works from known physics approximations of actively-induced magnetism and passive magnetometers. The principal shortcomings of this approach are: (1) It does not produce a unique solution for the parameters for a large number of targets; (2) For effective discrimination, it requires very accurate positional information (<= 1 cm error), which is not attainable with current, economic positioning technology for sensors moving across the field ("survey" mode) [18]; and (3) It requires higher signal-to-noise ratios than are typically encountered in production data for smaller targets like 37mm MEC.

Thus, successful forward modeling/inversion technology is mostly confined to "cued" data—that is, data collected at very high resolution and positional accuracy with static measurements over a known Target location or research grade survey data. Further, studies using this approach have not reported much success in discriminating small ordnance (20mm and 37mm projectiles)—most of the successes have been with larger ordnance.

The other approach was first reported by Deschaine in 2002. He applied Linear Genetic Programming ("LGP") to research-grade data from a Protem-47 sensor and compared his results to those of the ten other contractors who had attempted discrimination on these same data. Deschaine, using LGP, did a dramatically better job discriminating MEC from not-MEC than had been reported by the other ten contractors [7]. Later, Banks applied a Genetic Programming approach to these same data [1]. His preprocessing and Genetic Programming modality were different than Deschaine's. Nevertheless, he produced results significantly better than the other ten contractors; but not as good as Deschaine's.

Francone and Deschaine then applied LGP to high-quality, survey-mode, production-grade EM61 MK1 data from a DoD test bed. That test-bed attempted to simulate an actual MEC impact-site with known buried items. Francone [12][11] reported in that study that LGP reduced the number of false-alarms (the number of non-MEC items that had to be dug to completely clear the site of MEC items) by about 40%. This was particularly significant as the smallest items on that site were 20mm projectiles—traditionally regarded as very difficult to discriminate from clutter. On the other hand, the sample size in that study was very small (a total of only 17 MEC items and 300 non-MEC targets). So some question remained whether the results would replicable on larger data sets.

The principal advantages of Francone and Deschaine's LGP approach to-date are: (1) It has demonstrated success on field-grade, simulated, production data gathered in survey-mode; and (2) It has demonstrated success in discriminating small ordnance from surrounding clutter [12]. This paper reports the results of applying the methodologies reported by Francone and Deschaine in [12] to a large, MEC cleanup project using only production-grade, survey-mode data.

[1] For more information about the EM61 MK2, please see www.geonics.com.

2. THE WARREN AIR FORCE BASE DATA

F.E.Warren Air Force Base ("Warren") is located near Cheyenne, Wyoming. In the past, a portion of that base served as a practice range, primarily for 75mm and 37mm projectiles.

We analyzed DGM data comprising over 60 million data points in four channels of data from 600 acres of Warren (the "Site"). Each data point consisted of four channels of information gathered from a Geonics EM61 MK2 configured with three time-decay channels on the lower-coil and one upper-coil channel. These data were integrated with a differential global positioning system (GPS). They were collected with one-meter between transects. [2]

We were also provided a table containing geophysicist-designated target-locations on the Site (the "Targets") and ground-truth for some of those Targets (groundtruth is what the dig-teams actually found when they dug that Target up). Altogether, there were 29,130 Targets designated. All Targets were investigated by EOD teams. As a result, the customer had groundtruth for all Targets. The vast bulk of the Targets contained one or more metallic items on investigation. Of these 29,130 Targets, only 332 contained intact 37mm or 75mm MEC items. The remainder contained mostly MEC Clutter.

The intact 75mm's recovered ranged in depth from 0-40 inches while the intact 37mm's ranged from 0-20 inches. The deeper items were close to the instrument's detection range.

3. TARGET DENSITY

The Target density across the 600 acre Site varied widely. Some areas were densely populated with buried metal. Figure 1 shows the Targets-per-acre on the Site sampled in 20x20 meter squares. The density ranges from low to over 400 Targets-per-acre.

Figure 1. Target Density (Targets per Acre)

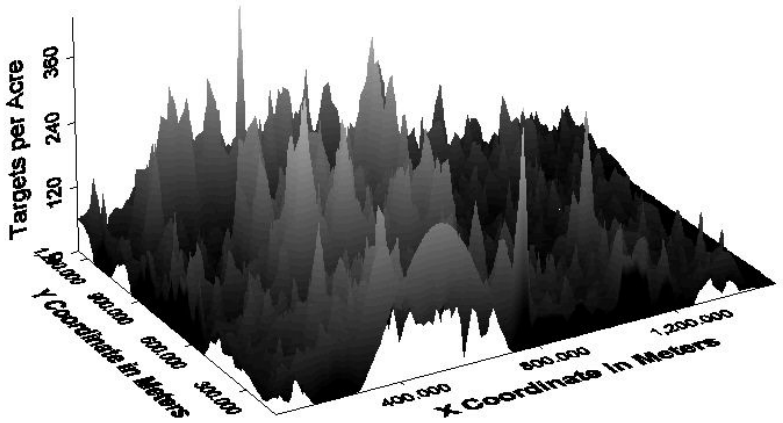
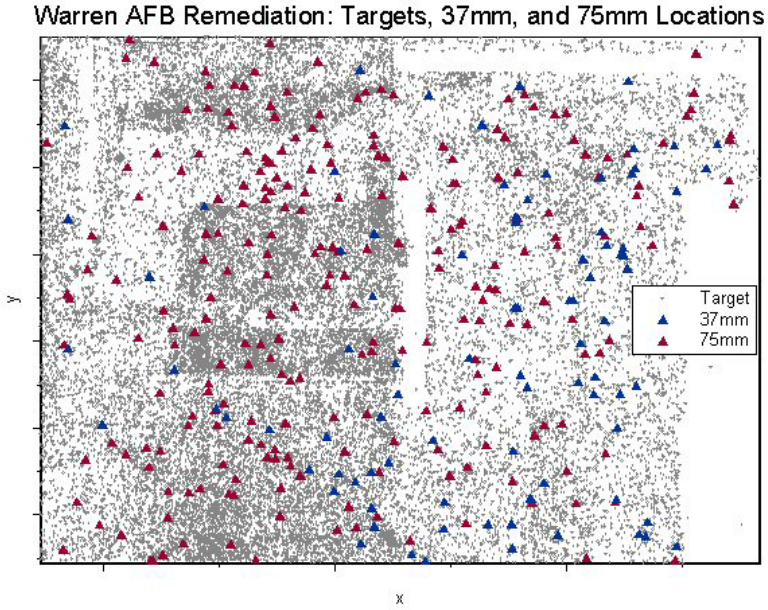


Figure 2 shows a top-down picture of Target and MEC density on the Site. The very small gray dots are the geophysicist-selected Targets. The colored triangles show the 37 and 75mm's. Most of the small, gray circles contained Clutter. In the denser, darker regions of Figure 2, many Targets contained multiple metallic objects in the same hole. In fact, across the entire 600 acres, about 20% of all Target locations contained multiple metallic objects.

[2] The DGM of the Site was delivered to us already lag-corrected by the customer in the Geosoft Oasis-Montaj UXO processing module. Lag correction is the process of adjusting the x,y coordinates of each data point to reflect the fact that alternate lines of data are collected when the sensors are traveling different directions. We assumed, for the purpose of this study that no further lag correction was necessary or desirable.

Figure 2. MEC and Target Density Map—600 Acres
Approximately. X and Y Axes denominated in meters.



Areas containing a large number of Targets pose a particular problem for MEC discrimination because they frequently contain overlapping Target DGM signatures. It is necessary to untangle such overlapping signals before they may be characterized individually. On the Site, there were 7,382 Target pairs (including 36 MEC) that were less than two-meters apart and 1,466 Target pairs (including 20 MEC) that were less than one-meter apart. Depending on the size, depth, and inclination, the DGM signatures of the Targets ranged from 0.5 to more than 3 meters in radius. Thus, overlapping Targets were a significant problem in this project.

4. DISCRIMINATION CHALLENGES

In addition to the overlapping Targets problem, these data posed several significant challenges in producing good discrimination results.

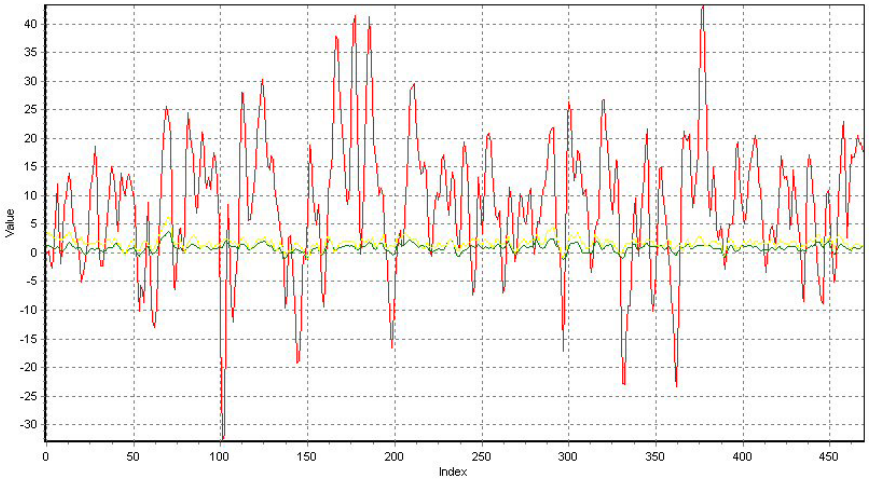
4.1 Data Quality Issues

These data were collected for comprehensive removal of all buried metallic objects on the Site. The Targets for excavation were picked by geophysicists, who examined these data using Geosoft Oasis-Montaj. The intent in picking Targets was to dig every Target. This was a good-quality, commercial data set for that purpose.

These data were not, however, collected for the purpose of machine-based discrimination and therefore posed significant data-quality issues for the present study. Geophysicist Target selection normally needs only one good channel of data over any given Target to make a determination that there is possibly a metallic object there. In that context, it makes no sense to incur extra costs to reacquire data when, for example, one channel has too much noise, so long as there is at least one good channel of data.

By way of contrast, our discrimination approach uses all four channels. A significant portion of the geophysicist-selected Targets had noise or calibration issues with one or two channels of data—usually channels one and/or four. We elected to retain data as long as there were at least two good channels of data and to allow the LGP algorithm to adjust for varying noise contexts. This decision permitted us to rank all 29,130 Targets for the likelihood they were MEC. Figure 3 is an example of one discrimination data-quality issue we encountered that affected 28 of the intact 37mm's that were located on the site.

Figure 3. Single Transect Showing Three Upper Channels from EM61. Two Channels Contain Apparently Good Data. But Channel 1 Exhibits Less-Than-Optimal Data-Quality. Horizontal Axis in Meters. Vertical Axis in Millivolts.



Signal-to-Noise Ratio. Traditional forward-modeling/inversion approaches to discrimination degrade quickly when the signal-to-noise ratio for a Target falls below 30-1. 37mm Targets in the Warren data routinely presented with a signal-to-noise ratio lower than that. Thirty-four percent of the 37mm's had a s/n ratio of less than 5. Eighty-three percent had a s/n of less than 10.

4.2 An Ocean of Clutter

The most significant challenge we faced was discriminating the 37mm MEC items from MEC Clutter. As noted above, 37mm's are rather small ordnance and are much more difficult to discriminate than, say, 500 lb bombs or 155mm artillery shells.

The data at Warren illustrate why this is so. There, the largest ordnance type was the 75mm projectile. Investigators recovered 26,996 pieces of MEC Clutter out of the 29,130 Targets excavated. By far, MEC Clutter from detonated 75mm's was most common buried metallic item found on the site. On the other hand, the dig-teams recovered only 87 intact 37mm's. They reported the intact 37mm's weighed between 2 and 20 ounces.[3] Of the MEC Clutter items recovered, 19,351 weighed between 2 and 20 ounces. Thus, most of the MEC Clutter was about the same size as the 37mm's recovered. Mass was, therefore, not an appropriate or useful discriminator for the 37mm's.

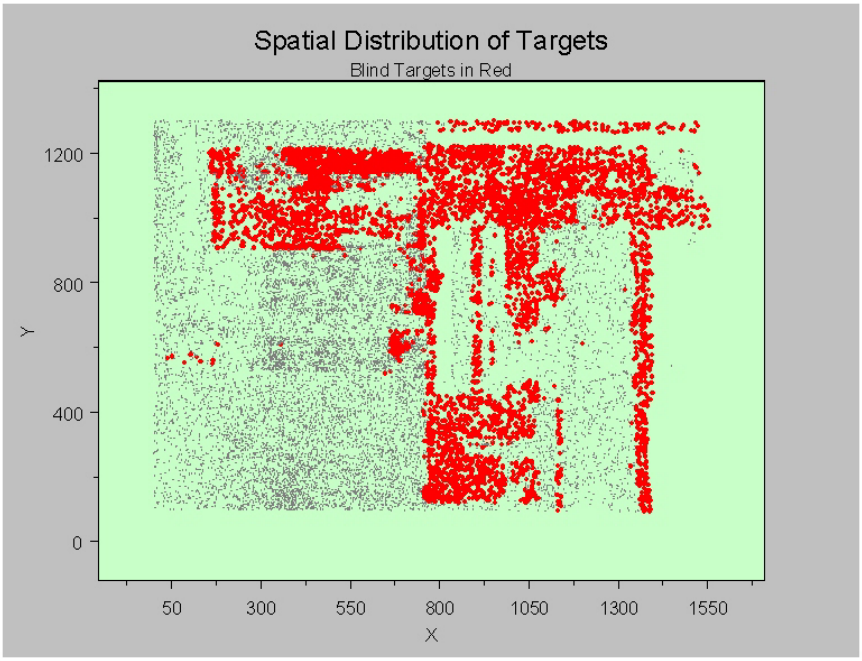
The discrimination results reported below reflect the difference between finding 75mm's and 37mm's at Warren. As the largest metallic item on the Site, intact 75mm's were relatively easy to discriminate. 37mm's, on the other hand, were more difficult to discriminate from similarly sized clutter.

5. BLIND-TESTING PROCEDURES

We were provided with DGM for the entire 600 acre site. In addition, we were provided with groundtruth (that is, what the dig-teams found when they excavated a Target) for 23,085 of the 29,130 Targets, which we used to develop our models. The Targets for which we were provided groundtruth contained fifty-nine, 37mm's and one-hundred eighty-six 75mm's. We were *not* provided groundtruth for the remaining 6,045 Targets, which contained twenty-eight 37mm's and fifty-nine 75mm's. Nor were we involved in selecting the blind Targets. Figure 4 shows the spatial distribution of the training data and the actual blind-data that were withheld by the customer.

[3] Weights were collected in the field by the dig-teams and were estimated manually, not weighed.

Figure 4. Map of F.E.Warren Training vs. Blind-Data. Actual Blind-Data that Was Withheld by the Customer Are Shown in Red. The Training Data for which We Were Provided Groundtruth Are Shown in Gray



We performed two kinds of blind testing, which we distinguish here. First: In training our models to discriminate between MEC and not-MEC, we used four-fold cross-validation. All results on the 23,085 Targets on which we had groundtruth are reported on the held-out, testing data, the "virtual" blind-predictions if you will. Second: We delivered predictions to our customer for all 29,130 Targets—the Targets for which we had groundtruth and the Targets for which we did not. The predictions took the form of rankings of the Targets by the likelihood the Target was MEC. So we ranked the 6,045 Targets for which we had no groundtruth blind. After we delivered the rankings to the customer, the customer then delivered groundtruth for these 6,045 Targets to us. So, as to these 6,045 Targets, the predictions were actual blind predictions and we refer to them as such.

When we compared the Blind-Data and the data for which we had groundtruth, it was clear that the customer had not selected the 6,045 Blind Targets entirely at random—see Figure 4.[4] Accordingly, the blind Targets were not representative of the entire site. Our goal in this project was to determine what portion of Targets had to be excavated in order to clear the whole site of MEC. In order to make that assessment, it was necessary to combine predictions on the training, virtual blind-Targets with the predictions on the "actual" blind Targets. In that way, it was possible to fit the blind-Targets into the context of the entire site. Our reports below were assembled in this manner—all results reported are on either the training, virtual blind-data and/or the actual blind-data where groundtruth was withheld by the customer.

6. EXPERIMENTAL PROCEDURE

Our process involves several steps, described below. Before LGP was applied, considerable preprocessing to remove noise and standardize the Targets was necessary. In addition, for each Target we had to define what data points were in the Target and which were not—that is, we had to define the size, shape, and orientation of each Target. Finally, we extracted the features used for LGP discrimination, analyzed them and prepared preliminary statistical models from them. The process is described in more detail below.

[4] For example, the proportion of 37mm's in the blinded data was 0.0046. The proportion in the training data was .0025. The probability of this happening by random chance is less than p=0.007. In addition, the blinded data had a disproportionately high percentage of 37mm targets where Channel 1 was excessively noisy over the target.

6.1 Preprocessing

EM61 data contains low frequency variation caused by factors like instrument drift or geomagnetic variation. Targets are typically between 0.5 and 3 meters in radius. We filtered the low frequency variation using a robust, high-pass filter, allowing features less than 10 meters in length to pass. This filtering process was entirely automated--the volume of data prevented individual examination of Targets. Then we standardized the background noise regions to a mean level of about zero.

6.2 Defining the Targets

The customer provided us x, y coordinates of the geophysicist-selected Targets. We grouped the EM61 signals read in the vicinity of each Target as being: (1) in the Target; and/or (2) in another (adjacent) Target; or (3) in the background noise. Our process for doing so involved fitting an ellipse to each Target using the preprocessed data converted to localized z-scores. The localized z-scores were calculated for each Target empirically, relative to the background noise level in the ten meter circle surrounding the Target.

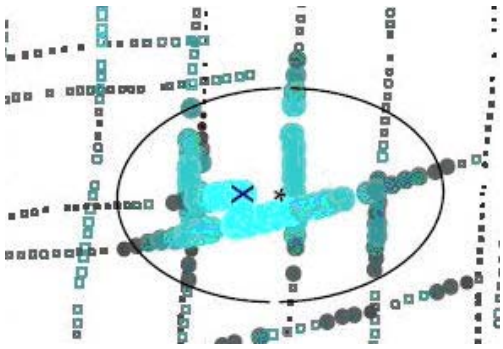
We used a deterministic optimizer using the Lipchitz global optimization algorithm [17] to define the Targets from the localized z-scores. The problem posed to the optimizer was to find the ellipse that best separated the above-background-noise values (z-score > 2) in the vicinity of each Target from the below-background-noise values (z score <= 2) in the vicinity of that Target. For overlapping Targets, points in the area of overlap were excluded from the optimization. Thereafter, the ellipse derived was used to define the size and shape of the Target. We defined such an ellipse for each channel of data for each of the 29,130 Targets. This process was also entirely automated.

Figure 5 shows the result of this process for a single EM61 channel around one Target. Each datum from the EM61 near the geophysicist-picked Target location (marked with an "X") is shown as a point. The size of the points in Figure 5 represents a point's z-score—larger means a higher z-score relative to background noise. The ellipse defines which points we treated as being "in" the Target.

6.3 Feature Extraction

Once the ellipse was defined, we extracted a set of features for each Target. We determined what features to extract using two criteria. First: approximate physics-based models of induced magnetic field provides some guidance as to what ought to be important. See e.g. [4][3]. For example: (1) The shape of the time-decay of the signal from Channel 1 to Channel 3 shows different characteristics for differently shaped objects; (2) The ratio of the upper and lower coil signals ought to provide information about the depth of the object because of the exponential decay of signal strength with distance; and (3) The symmetry of the signal about the major and minor axes is affected by the shape and inclination of the buried object. So the features we extracted provided detailed information about decay, upper/lower coil ratios, and signal symmetry.

Figure 5. Ellipse defining a Target for a single channel of data. The larger, bluer data-points represent higher millivolt values. The ellipse optimially separates above-background-noise values from in-background-noise values.



But existing physics-based models are approximations and do not completely define the problem posed by production-grade survey data. Accordingly, we also extracted features that provide a detailed statistical topology of each Target. For example, we extracted the median signal value for the innermost part of each ellipse and for concentric ellipsoidal donuts outside that innermost part.

6.4 Feature Reduction

We took the extracted feature set and subjected them to statistical analysis to determine which were most predictive of "MEC-ness" and, of those, which features were the least correlated with each other. In addition, we performed a series of preliminary modeling runs to determine which features contributed most to good-quality models [10]. We modeled using the reduced data set.

6.5 Modeling

We then built our models using four-fold cross-validation—that is, the data set was split four different ways, with each data point serving as training data three times and as blind, testing data once. Each split of the data set was generated randomly. The examples of MEC were duplicated until there was an approximate balance between the number of MEC items and the number of non-MEC items in each cross-validation training set. No such duplication was performed for the data used to test the models.

Our principal tool here was LGP [2][10][16]. That software performs multiple runs automatically and selects the best evolved programs out of multiple runs. We used the default parameters of the LGP software for in non-stepping mode [10] and terminated each run when no improvement had occurred for at least 50 runs.

The Target output we wanted to predict was Boolean—a value of 1 represented MEC and a value of 0 represented non-MEC. The output from our models was a probability that an item was MEC.

For each of the cross-validation data sets, we chose the model that our software selected as best on the training data. The numbers reported below represent the performance of the four chosen models on the held-out, testing data—the "virtual" blind-data.

For our predictions on the "actual" blind Targets, we took the four models trained on the four cross-validation data sets, applied them to the features extracted for the "actual" blind Targets, and averaged the predictions from the four models. That average was the value used to order the blind Targets.

We performed this process separately for 37mm and 75mm. Using the outputs of the four models described above on both the "virtual" blind Targets and the "actual" blind Targets, we prepared a list prioritizing the Targets starting with the most likely to be MEC to the least likely to be MEC. We did that on a "whole site" basis to show where the "actual" blind Targets would have been excavated as part of clearing the entire site.

7. RESULTS

7.1 75mm MEC Results

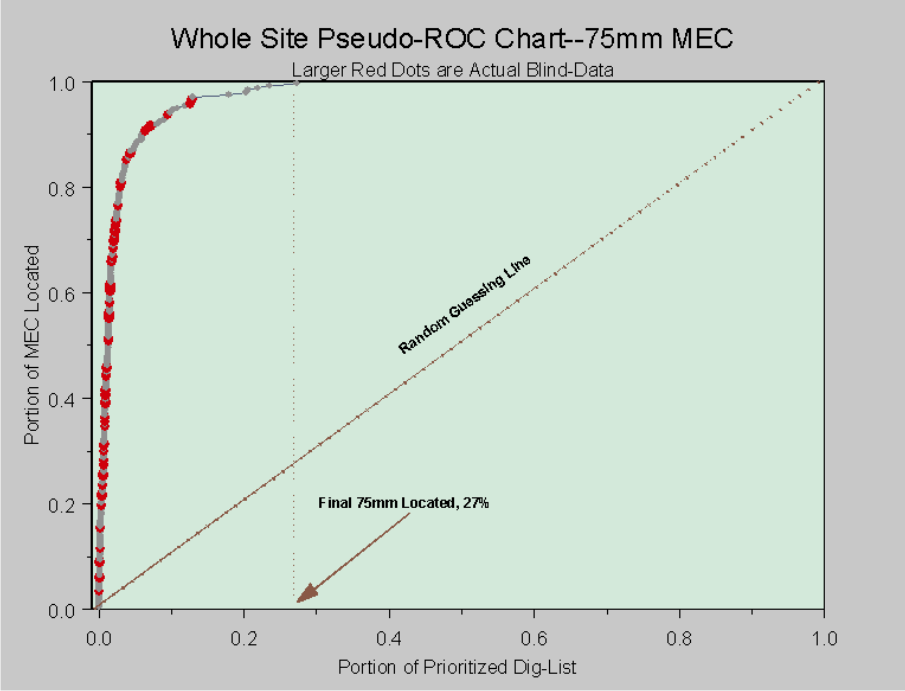
Figure 6 is a pseudo-Receiver Operating Characteristics ("ROC") chart [13] showing the performance of our prioritized dig-list on 75mm MEC. The horizontal axis represents the prioritized dig-list—that is, what portion of all Targets would have been excavated had the site been cleared in the order specified by our prioritized dig-list. The vertical axis shows what portion of the 75mm's would have been cleared had the site been cleared in the order suggested by our prioritized dig-list. The curved line is the ROC curve. The Targets in the training "virtual" blinded-data are shown with small circles. The Targets in the actual blind-data are shown with larger, orange circles.

The diagonal line represents what random guessing would generate. It shows the tradeoff between finding all 75mm's, on the one hand, and the number of Targets that must be dug to do so, on the other hand, given our prioritized dig list. (The horizontal axis in Figure 6 is somewhat different than what would be shown in a true ROC curve. In a true ROC curve, the horizontal axis would show the portion of true-negatives excavated in the order of the prioritized dig-list [13]. We use the pseudo-ROC chart for display as that is the recommended DoD format for assessing rankings to MEC vs. not-MEC in remediation discrimination [5].)

In brief summary, Figure 6 shows that:

1. All 75mm MEC would have been cleared from the site by excavating only 27.3% of the total Targets, had our dig-list been used to order the excavation. Put another way, our approach would have cleared all 75mm's and reduced the number of false-alarms to do so by 72.7%; and
2. The final 75mm in the actual blind-data would have been excavated about 14% of the way through our prioritized dig-list. Thus, our approach actually performed better on the withheld, actual blind-data than it did on the training blinded data.

Figure 6. Whole-Site, Pseudo-ROC Chart for 75mm MEC. Actual Blinded Data Shown in Larger Orange Points. Training (Virtual) Blinded-Data Shown in Small Gray Points



Our test of the statistical significance of these ROC curves begins by converting the pseudo-ROC curve in Figure 6 into a true ROC curve, as described above. Of course, the true ROC curve for these rankings also has a 45 degree random-guessing line that goes thru origin and the point 1,1. The area under that random guessing line is 0.5. So our null hypothesis is that the area under the true ROC curve for our 75mm rankings is not greater than 0.5. The alternative hypothesis is that the area under the ROC curve is greater than 0.5.

The computation of the area under a ROC curve and the standard error of the area are described by Hanley in [13]. For our 75mm rankings, the area under the ROC curve is 0.9568 and the standard error is 0.00918. Thus, 0.5 is more than 49 standard deviations away from the computed area under the ROC curve. The p-value that the area under this ROC curve is not greater than 0.5 is zero to the limits of computational precision. Our 75mm rankings are, therefore, statistically significant at a very high-level for both the "virtual" blind-data and the actual blind-data.

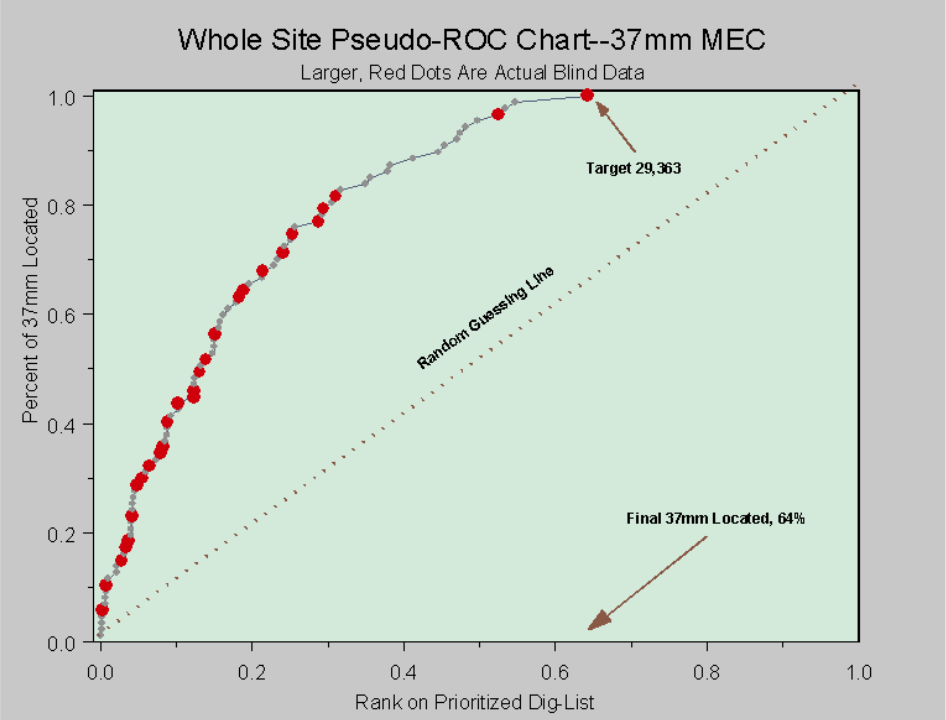
7.2 37mm MEC Results

Figure 7 is a pseudo-ROC chart showing the performance of our prioritized dig-list on 37mm MEC. The horizontal axis represents the prioritized dig-list. The vertical axis shows what portion of the 37mm's would have been cleared had the site been excavated in the order suggested by the prioritized dig-list. The diagonal line again represents what random guessing would generate. We again show the Targets from the actual blind-data as larger orange circles and the Targets from the training "virtual" blinded-data as smaller, gray circles.

In summary, Figure 7 shows that:

1. All 37mm MEC would have been cleared from the site by excavating only 64.2% of the total Targets. Put another way, our approach would clear all 37mm's and reduce the number of false alarms by 35.8%; and
2. The final 37mm amongst the actual blind targets, Target 29,363, was located about 10% later than the next-previously located target. Thus, the performance of the dig-list degraded somewhat as between the training, "virtual" blind-data and the actual blind-data

Figure 7. Whole-Site, Pseudo-ROC Chart for 37mm MEC. Actual Blinded Data Shown in Larger, Orange Points. Training (Virtual) Blinded-Data Shown in Small Gray Points



We applied the same Hanley procedure [13] used for the 75mm ROC curve to test the statistical significance of our 37mm rankings. The results were: (1) The area under the true-ROC curve for the 37mm rankings is 0.8233; (2) the standard error of the area is 0.0275; and (3) The probability that the area under the 37mm true-ROC curve is not greater than 0.5 (random guessing) is 0.0 to the limits of machine precision. Accordingly, our prioritized dig-list produced a highly statistically significant ranking of both the training, 'virtual" blind-data and the actual blind-data.

7.3 Breakdown of Results for Training, Virtual Blind-Data and Actual Blind-Data

Second, this study replicates Francone's findings in [12] on a much larger data set in blind, third-party-conducted testing. Both [12] and the present study reduced false alarms by between 35% and 40% over geophysicist Target-selection in clearing a site of the smallest ordnance. (In [12], the smallest ordnance was 20mm MEC. In the present study, the smallest was 37mm.) Both this study and Francone's previous work used similar methodologies. Thus, the replication of these earlier results on blind-data provides substantial evidence for the efficacy of the LGP-oriented methodology.

Table 1 shows the breakdown of the Areas Under the Curve and the statistical significance of the Areas Under the Curve for all divisions of data. We note that because the training (virtual) blinded data and the actual blind-data were not randomly sampled from the overall site, that these numbers are not valid statistics for the overall site. For numbers that are meaningful with respect to the overall site, please see Figure 6 and Figure 7 and the accompanying discussion. Second, this study replicates Francone's findings in [12] on a much larger data set in blind, third-party-conducted testing. Both [12] and the present study reduced false alarms by between 35% and 40% over geophysicist Target-selection in clearing a site of the smallest ordnance. (In [12], the smallest ordnance was 20mm MEC. In the present study, the smallest was 37mm.) Both this study and Francone's previous work used similar methodologies. Thus, the replication of these earlier results on blind-data provides substantial evidence for the efficacy of the LGP-oriented methodology.

Table 1 demonstrates that, no matter how we slice our results, our methodology produced highly statistically significant rankings of both 37mm and 75mm's on all data sets.

8. DISCUSSION

These results represent a significant step forward in the state-of-the-art for MEC discrimination in several regards.

First, this study extends the approach described by Francone and Deschaine in [12] from small, high-quality data samples derived from a simulated impact site to noisy, production-grade, survey data. To our knowledge, this is by far the largest MEC discrimination test reported and the only one demonstrating substantial success on production-grade, survey data.

Second, this study replicates Francone's findings in [12] on a much larger data set in blind, third-party-conducted testing. Both [12] and the present study reduced false alarms by between 35% and 40% over geophysicist Target-selection in clearing a site of the smallest ordnance. (In [12], the smallest ordnance was 20mm MEC. In the present study, the smallest was 37mm.) Both this study and Francone's previous work used similar methodologies. Thus, the replication of these earlier results on blind-data provides substantial evidence for the efficacy of the LGP-oriented methodology.

Table 1. Area Under the ROC Curve and P-Value for All Data Sets

	Area Under the ROC Curve	Probability
75mm MEC		
Training virtual blinded data	0.97	0.0[5]
Actual blind-data	0.97	0.0
37mm MEC		
Training virtual blinded data	0.84	0.0
Actual blind-data	0.76	0.00000054

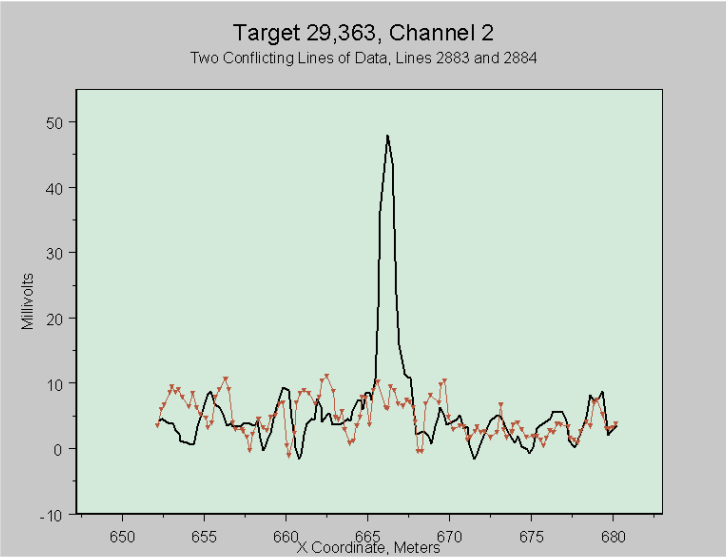
Third, these results improve on results for larger ordnance reported by well-conducted studies using the forward-modeling/inversion approach we described earlier. For example, [9] is a relatively large, well-conducted study of an actual impact site using magnetic and EM61 survey data collected by the MTADS system. MTADS produces much higher quality survey data than we had access to in this project.[6] The discrimination approach tested there produced a 58% reduction in false alarms for 81mm mortars, when the 81mm's were considered alone. (In general, the larger the ordnance, the easier it is to discriminate. So our discrimination of 75mm projectiles in the present study was a roughly comparable task to the discrimination of 81mm's in [9].) Our approach produced a 72.2% reduction in the false alarms for 75mm's on data very much inferior to that group's MTAD's data.

Fourth, the discrimination of small ordnance (such as 37mm's) from Clutter is usually the most difficult discrimination task on a site. We are not aware of any studies that have produced results on small ordnance using survey-based, production-grade data that show results comparable to ours, other than Francone's previous work, which also used LGP and preprocessing methodologies similar to the present study [12].

[5] Values of zero for probability on this chart mean that the probability was zero to the limits of machine precision.

[6] This difference in data-quality is not a result of different data-gathering technique. Rather, it is the result of different hardware platforms. MTAD's is state-of-the-art for survey-based data collection. [8]

Figure 8. Two Conflicting Lines of Data under Target 29,363



Fifth, we examined the data for Target 29,363, the final 37mm on our prioritized dig-list. It was prioritized considerably later than the remainder of the 37mm's (a full 10%, see Figure 7). That appears to have been caused by a QAQC problem we did not detect until our post-project analysis. Although the typical line spacing on this site was one-meter, two transects passed almost directly over Target 29,363 with almost no line separation. Ordinarily, this would not be a problem. However, for Target 29,363, the two lines of data contained conflicting information.

Figure 8 shows those two lines of data under Target 29,363. One is shown in bold, black and the other is shown in red. The bold line shows a clear anomaly. The other shows no anomaly at all. Obviously, our future work will contain procedures to locate situations similar to this. We suspect that the prioritization of Target 29,363 would have been considerably better had we caught this QAQC problem before generating our prioritized dig-lists.

9. ACKNOWLEDGMENTS

Our thanks to John Wright, Chief of Restoration at Warren Air Force Base and to Brian Powers and Andy Gascho of URS Corp. for making this project possible. We also express our gratitude to C. Edward Dilkes and Robert "Tom" Weatherly, without whose direction and vision, MECFinder would not have been developed.

10. REFERENCES

[1] Banks, R. E., Núñez, E., Agarwal, P., Owens, C., McBride, M., and Liedel, R., 2005. *Genetic Programming for Discrimination of Buried Unexploded Ordnance (UXO)*). Late-breaking paper at The Genetic and Evolutionary Computation Conference (GECCO-2005)

[2] Banzhaf, W., Nordin, P. Keller, R. Francone, F. *Genetic Programming, an Introduction*, Morgan Kaufman Publishers, Inc., San Francisco, CA (1998).

[3] Billings, L.R., and Oldenburg, D. W. (2003). Joint and Cooperative Inversion of Magnetic and Time Domain Electromagnetic Data for the Characterization of UXO, *Proceedings of the Symposium on Application of Geophysics to Environmental and Engineering Problems 2003 (CD)*, Environmental and Engineering Geophysical Society, San Antonio, TX, 2003.

[4] Billings, S. D., Pasion, L. R., and Oldenburg, D. W. (2003). Discrimination and Classification of UXO Using Magnetometry: Inversion and Error Analysis Using Robust Statistics, *Proceedings of the Symposium on Application of Geophysics to Environmental and Engineering Problems 2003 (CD)*, E.E.G.S., San Antonio, TX, 2003.

[5] Cespedes, E.: *Advanced UXO Detection/Discrimination Technology Demonstration—U.S. Army Jefferson Proving Ground, Madison, Indiana.* US ACOE, Engineer Research and Development Center, ERDC/EL TR-01-20 (2001).

[6] Defense Science Board Task Force. *Report of the Defense Science Board Task Force on Unexploded Ordnance.* Department of Defense. December (2003).

[7] Deschaine, L. M., Hoover, R. A., Skibinski, J. N., Patel, J. J., Francone, F. D., Nordin, P. and Ades, M. J.: Using Machine Learning to Compliment and Extend the Accuracy of UXO Discrimination Beyond the Best Reported Results of the

Jefferson Proving Ground Technology Demonstration. In: *Proceedings of the Society for Modeling and Simulation International's Advanced Technology Simulation Conference*, April 2002. San Diego, CA, USA (2002).

[8] Environmental Security Technology Certification Program. *Multi-Sensor Towed Array Detection System, Cost and Performance Report (UX-9526)*, September 1999.

[9] Environmental Security Technology Certification Program. *Environmental Induction and Magnetic Sensor Fusion for Enhanced UXO Target Classification, Cost and Performance Report (UX-9812)*, February 2004.

[10] Francone, F. D., *Discipulus Owner's Manual*. RML Technologies, Inc. (2002). Available at www.aimlearning.com.

[11] Francone, F. D., and Deschaine, L.M., Extending the Boundaries of Design Optimization by Integrating Fast Optimization Techniques with Machine-Code-Based Linear Genetic Programming, *Information Sciences Journal—Informatics and Computer Science*, Elsevier Press, Vol. 161/3-4 pp 99-120: 2004. Amsterdam, the Netherlands.

[12] Francone, F. D., Deschaine, L. M., Battenhouse, T., and Warren, J. J., 2004. Discrimination of Unexploded Ordnance from Clutter Using Linear Genetic Programming, *Proceedings of the Genetic and Evolutionary Computation Conference, Late Breaking Papers*, 2004, Seattle, WA, USA.

[13] Hanley, J. and McNeil, B., *The Meaning and Use of the Area under a Receiver Operator Characteristic (ROC) Curve*," Radiology, Vol. 143, pp. 29-36, 1982.

[14] Jolliffe, I.T. *Principal Components Analysis,* Second Edition. (Springer Series in Statistics, NY, NY 2002), p. 1.

[15] Nelson, H. H., Altshuler, T., Rosen, E., McDonald, J.R., Barrow, B., and Khadr, N. *Magnetic Modeling of UXO and UXO-Like Targets and Comparison with Signatures Measured by MTADS*, www.citeseer.ist.psu.edu/273329.html

[16] Nordin, P., Francone, F. Banzhaf, W. *Efficient Evolution of Machine Code for CISC Architectures Using Blocks And Homologous Crossover*. In: Advances in Genetic Programming 3, MIT Press, Cambridge, MA (1998).

[17] Pinter, J., *Global Optimization in Action. Continuous and Lipschitz Optimization: Algorithms, Implementations and Applications*. Kluwer Academic Publishers, Dordrecht / Boston / London, 1996.

[18] Tantum, S., Yu, Y., Zhu, Q., Wang, Y., and Collins, L. http://www.ee.duke.edu/research/collins/uxodocs.html, March 28, 2006