FINAL REPORT

Integration of Advanced Statistical Analysis Tools and Geophysical Modeling

SERDP Project MR-1657

AUGUST 2012

Lawrence Carin
Duke University

Douglas Oldenburg
University of British Columbia

Stephen Billings
Leonard Pasion
Laurens Beran
Sky Research

This document has been cleared for public release

SERDP
DOD ▪ EPA ▪ DOE

This report was prepared under contract to the Department of Defense Strategic Environmental Research and Development Program (SERDP). The publication of this report does not indicate endorsement by the Department of Defense, nor should the contents be construed as reflecting the official policy or position of the Department of Defense. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the Department of Defense.

Acronyms

- AUC: Area Under Curve (area under the ROC curve)
- EMI: Electromagnetic Induction
- EM: Expectation Maximization
- FAR: False Alarm Rate
- MAP: Maximum A Posterior
- MRTDB: Munitions Response Target Database
- QC: Quality Control
- ROC : Receiver Operating Characteristic
- SOI: Single Object of Interest
- SVM: Support Vector Machine
- TEM: Time Domain Electromagnetics
- TEMTADS: Time Domain Electromagnetic Towed Array Detection System
- TOI: Target of Interest
- UBC: University of British Columbia
- UXO : Unexploded Ordnance

Abstract

Background. This research program has been focused on advanced technologies for detection and discrimination of military munitions. The underlying premise of the program has been that there is an inherent limitation in the information content associated with magnetometer and EMI sensors deployed for UXO cleanup. To optimize UXO classification one must integrate all available information, both within the measured data itself and within *a priori* knowledge one may possess. An important class of prior knowledge is represented by the sensor physics, and by placing as much physics as possible into the models and classification features, one removes the need to rely on the limited sensor data to *infer* such phenomenology. Statistical classifiers are also required to maximize the information extracted from the measured data to infer the unknown model *parameters*. Further, the statistical classifiers may be used to appropriately exploit other forms of information inherent to the data. For example, while performing classification one may exploit the contextual information provided by all of the unlabeled data at a given site, while also appropriately leveraging related information in data measured at previous sites.

Objective. The overall objective of the research has been to integrate advanced Bayesian statistical models and classifiers with leading geophysical models, to enhance the ability to extract information from limited sensor data, with the goal of markedly improving UXO classification performance on complex cleanup missions. The technology has been directed toward general magnetometer and EMI sensors. A key aspect of the research is to develop sophisticated but practical technology, appropriate for real-world UXO cleanup. The technology is directed toward difficult geology, terrain, and complex ordnance and clutter distributions.

Technical Approach. The research program has exploited the complementary skills of the Duke and UBC/Sky investigators. In the research program a focus has been placed on integrating the statistical inference engines developed at Duke with the sophisticated physics-based models developed at UBC/Sky. The particular statistical techniques into which the advanced geophysical models have been integrated include semi-supervised learning, multi-task and life-long learning, and active learning. We also have developed new techniques that explicitly account for the imbalance in UXO and non-UXO items at a typical site,

with this of significant importance when computing the risk associated with leaving an item unexcavated.

Benefits. By integrating the Duke and UBC/Sky technology, the Bayesian statistical models have been aided by improved geophysical models, and *vice versa*. This new technology has the potential to significantly improve the DoD's ability to do practical UXO cleanup. The experience of the investigators within the ESTCP Demonstration Studies has guided selection of the open research questions to be investigated, advancing the likelihood that the research products will constitute new science while also being of importance to practical UXO cleanup.

Contents

Acronyms i
Abstract i
Background i
Objective i
Technical Approach i
Benefits ii
List of Figures iv
List of Tables vi
1. Objective 1
2. Background 1
2.1. The TEM dipole model 3
2.2. Parameter estimation with the dipole model 4
2.3. Classification 8
3. Methods 10
3.1. The Semi-supervised Learning Algorithm 10
3.2. The Graph Representation of a Partially Labeled Data Manifold 10
3.3. Neighborhood-Based Learning 11
3.4. The Learning Algorithm 13
3.5. Active Learning 14
3.6. Active Learning with Semi-Supervised Classifier 14
4. Results and Discussion 15
4.1. Comparison of Expert QC, Auto QC and No QC using MetalMapper data 15
4.2. Development and testing of active learning algorithms using Sky/UBC features 51
4.3. Development of a munitions response target database 61
References 63
Appendix 64

List of Figures

1 EM sensor geometries and channels 2
2 Flow chart for advanced discrimination of UXO. 3
3 Display for quality control of MetalMapper data fits 7
4 Anomaly 1951 of the Beale C MetalMapper dataset 16
5 Anomaly 2015 of the Beale C MetalMapper dataset 17
6 Decay versus size feature space plot for Beale P data 20
7 Official scoring for Beale P using Expert-QCed data 21
8 ROC curves for Beale P using Expert-QCed data with L1 match 22
9 ROC curves for Beale P using Expert-QCed data with L1,L2, L3 match 23
10 Predicted polarizabilities for the two most difficult TOI of the Beale P dataset 24
11 ROC curves for Beale P using No QC 25
12 ROC curve for Beale P using No QC, threshold on decay parameter 26
13 ROC curves for Beale P using No QC, threshold on L1 misfit and decay parameter 27
14 Automated QC decision (auto QC Test 1) flowchart for passing/failing models based on data and model metrics. 28
15 Decay versus size feature space plots for Beale P data for auto QC test 1 29
16 ROC curves for Beale P using Auto QC 30
17 Automated QC decision flowchart for failing deep 2OI models 31
18 Decay versus size feature space plots for Beale P data for auto QC test 4 32
19 ROC curves for Beale P using Auto QC Test 4 to eliminate unrealistic deep 2OI models 33
20 Decay versus size feature space plot for Beale C data showing all passed and failed models as determined by expert QC 34
21 Official scoring for Beale C using Expert-QCed data 35
22 ROC curve for Beale C using Expert-QCed data, dig order based on L1 match 36
23 Polarizabilities for difficult TOI in Beale C data set 37
24 ROC curves for Beale C using No QC, dig order based on L1 match 37
25 ROC curves for Beale C using No QC, dig order based on L1 match and decay 38
26 ROC curves for Beale C using Auto QC, dig order based on L1 match 38
27 ROC curves for Beale C using Auto QC Test 4 to eliminate unrealistic deep 2OI models 39
28 Decay versus size feature space plot for Butner data 41
29 Official scoring for Butner MM using Expert-QCed data 42
30 ROC curves for Butner MM using No QC 42
31 ROC curves for Butner MM using No QC, dig order based on L1 match and decay 43
32 ROC curves for Butner MM using Auto QC Test 4 to eliminate unrealistic deep 2OI models 43
33 ROC curves for Butner MM using Auto QC Test 4 to eliminate unrealistic deep 2OI models, dig order based on L1 match and decay 44
34 Decay versus size feature space plots for Butner data, with no QC and auto QC 44

35 Decay versus size feature space plots for Pole Mountain data 46
36 ROC curve that would be obtained with expert QC for Pole Mountain 47
37 ROC curve that would be obtained with no QC for Pole Mountain 47
38 ROC curve that would be obtained with auto QC for Pole Mountain 48
39 Pole Mountain diglists using all thre polarizabilities 48
40 Camp Butner MetalMapper size decay features 51
41 Comparison of myopic and submodular learning performance applied to Camp Butner MetalMapper size-decay features 53
42 Boxplots summarizing AUC and FAR performance statistics for myopic and submodular learning algorithms applied to Camp Butner MetalMapper test data 54
43 Comparison of myopic, submodular and SVM performance applied to Camp Butner MetalMapper size-decay features 55
44 Comparison of Duke active learning algorithms and SVM active learning on Butner size decay features 56
45 Comparison of myopic and submodular learning performance applied to Camp Butner MetalMapper size-decay features, with artificial clusters of TOI seeded in the test data. 58
46 Comparison of myopic, submodular and SVM active learning performance applied to Camp Butner MetalMapper total polarizability features 59
47 Comparison of myopic, submodular and SVM active learning performance for Beale MetalMapper data sets 60
48 MRTDB interface and example search results 62

List of Tables

1 MetalMapper datasets used for testing 18
2 Summary of QC test results 50

1. Objective

This research program has been focused on advanced technologies for detection and discrimination of military munitions. The underlying premise of the program has been that there is an inherent limitation in the information content associated with magnetometer and EMI sensors deployed for UXO cleanup. To optimize UXO classification one must integrate all available information, both within the measured data itself and within *a priori* knowledge one may possess. An important class of prior knowledge is represented by the sensor physics, and by placing as much physics as possible into the models and classification features, one removes the need to rely on the limited sensor data to *infer* such phenomenology. While advanced physical models are critical, they are however not enough. Statistical classifiers are required to maximize the information extracted from the measured data, to infer the unknown model *parameters*. Further, the statistical classifiers may be used to appropriately exploit other forms of information inherent to the data. For example, while performing classification one may exploit the contextual information provided by all of the unlabeled data at a given site, while also appropriately leveraging related information in data measured at previous sites. One may also exploit prior knowledge concerning the density of UXOs and non-UXOs at typical cleanup sites.

The overall objective of this research program has been to integrate advanced Bayesian statistical models and classifiers with leading geophysical models, to enhance the ability to extract information from limited sensor data, with the goal of markedly improving UXO classification performance on complex cleanup missions. The technology has been directed toward general magnetometer and EMI sensors, including the new generation of EMI sensors becoming available. A key aspect of the research has been to develop sophisticated but practical technology, appropriate for real-world UXO cleanup. The technology is directed toward difficult geology, terrain, and complex ordnance and clutter distributions.

2. Background

The 2003 Defense Science Board report on unexploded ordnance (UXO) projected that a reduction in false alarm rates from 100:1 to 10:1 would save $36 billion on remediation projects within the United States (Delaney and Etter, 2003). This cost reduction was expected to be achieved by improvements in sensor and data processing technologies. These goals have been met, and sometimes exceeded, in recent demonstration projects conducted by the Environmental Security Technology Certification Program (ESTCP) (e.g. Billings et al. (2010)).

Advances in electromagnetic (EM) sensors have been crucial to these successes: the data provided by multi-static, multi-component EM platforms are much improved inputs into the inversion and discrimination algorithms applied to this problem. Figure 1 compares the

geometry and time channels of the commercial standard Geonics EM-61 with two multi-static EM instruments designed for UXO discrimination. The Time Domain Electromagnetic Towed Array Detection System (TEMTADS) is comprised of an array of 25 horizontal transmitter loops arranged in a 5x5 grid, with horizontal receivers measuring the vertical field arranged concentric to these transmitters. The transmitters are fired sequentially and the secondary field response is recorded in all receivers simultaneously. This configuration provides a diverse data set which is better able to constrain target parameters. The MetalMapper sensor has also greatly improved the reliability of estimated parameters by transmitting orthogonal primary fields and measuring all components of the secondary field in multiple receivers. Both MetalMapper and TEMTADS systems are deployed in a static (or cued) mode: previously-detected targets are interrogated with a stationary sensor. This removes the requirement for accurate geolocation that complicates data acquisition with a moving sensor such as the EM-61.

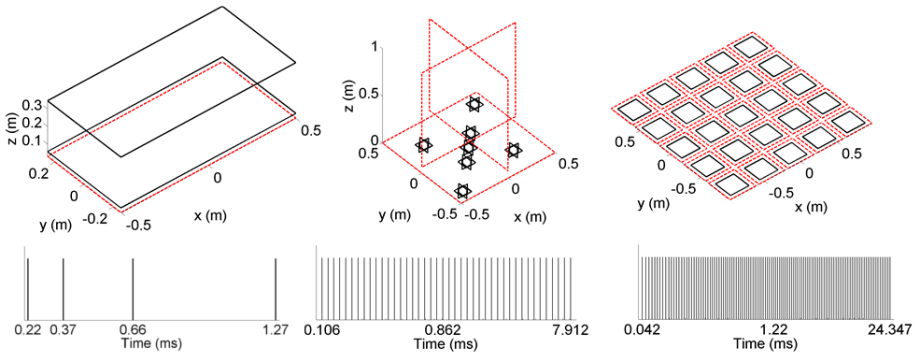


FIGURE 1. Left to right: Mono-static EM-61 and multi-static MetalMapper and TEMTADS sensors for unexploded ordnance detection and discrimination. Top row shows sensor geometry, with solid and dashed lines indicating receiver and transmitter coils, respectively. Bottom row shows time channels.

FIGURE 2. Flow chart for advanced discrimination of UXO.

Given digital geophysical data acquired with a sensor, a number of processing steps are required to produce an ordered list of targets for excavation. Figure 2 shows the typical processing involved in advanced discrimination. In the following sections we provide brief descriptions of the forward modelling, inversion, and discrimination required to generate a dig list.

2.1. **The TEM dipole model.** Essential to most electromagnetic data processing for UXO discrimination is the time (or frequency) dependent dipole model (Bell and Barrow (2001), Pasion and Oldenburg (2001), Zhang et al. (2003)). This model provides a simple parametric representation of the response of a confined conductor. The secondary magnetic field is computed as

$$\mathbf{B_s}(\mathbf{r},t) = \frac{p(t)}{r^3}\left(3(\hat{\mathbf{p}}(t)\cdot\hat{\mathbf{r}})\hat{\mathbf{r}} - \hat{\mathbf{p}}(t)\right) \tag{1}$$

with $\mathbf{r} = r\hat{\mathbf{r}}$ the separation between target and observation location, and $\mathbf{p}(t) = p(t)\hat{\mathbf{p}}(t)$ a time-varying dipole moment

$$\mathbf{p}(t) = \frac{1}{\mu_o}\mathbf{P}(t)\cdot\mathbf{B_o}. \tag{2}$$

The induced dipole is the projection of the primary field $\mathbf{B_o}$ onto the target's polarizability tensor $\mathbf{P}(t)$. The polarizability tensor is assumed to be symmetric and positive definite and so can be decomposed as

$$\mathbf{P}(t) = \mathbf{A}^T\mathbf{L}(t)\mathbf{A} \tag{3}$$

with $\mathbf{A}$ an orthogonal matrix which rotates the coordinate system from geographic coordinates to a local, body centered coordinate system. The diagonal eigenvalue matrix $\mathbf{L}(t)$ contains the principal polarizabilities $L_i(t)$ $(i = 1, 2, 3)$, which are assumed to be independent of target orientation and location.

Features derived from the dipole model have been successfully used to discriminate between targets of interest (TOI) and non-hazardous metallic clutter. In particular the amplitude and decay of the principal polarizabilities provide a simple parameter set for discrimination. For a sensor with N channels, these target features can be computed as

$$\begin{aligned} \text{amplitude} &= \sum_{j=1}^{N} L_{total}(t_j) \\ \text{decay}(t_k, t_j) &= \frac{L_{total}(t_k)}{L_{total}(t_j)} \end{aligned} \tag{4}$$

with the total polarizability $L_{total}(t_j)$ defined as the sum of the polarizabilities at each time channel

$$L_{total}(t_j) = \sum_{i=1}^{3} L_i(t_j). \tag{5}$$

The decay parameter is a ratio of total polarizabilities at selected channels. For $t_k > t_j$ we have decay$(t_k, t_j) < 1$, so that a larger decay parameter is diagnostic of a slow decaying total polarizability.

The amplitude and decay parameters are physically meaningful because, to first order, a confined conductor can be modelled as a simple LR loop which is inductively coupled to transmitters and receivers on the surface. The current response of this loop is a decaying exponential which is fully described by an amplitude and time constant (West and Macnae, 1991). In practice, UXO are characterized as large, thick-walled items and so produce large amplitude, slow decaying polarizabilities relative to metallic debris.

2.2. **Parameter estimation with the dipole model.** The dipole forward model described in the previous section is an example of the forward modelling operation

$$\mathbf{d} = F\{\mathbf{m}\}.$$

The data vector $\mathbf{d}$ is generated by a forward modelling operator F operating on the model vector $\mathbf{m}$. When real data are acquired, the related inverse problem is to estimate model parameters which produced the observed data. In the presence of noise, the inverse problem can be written as

$$\hat{\mathbf{m}} = F^{-1}\{\mathbf{d^{obs}}\}.$$

where the observed data $\mathbf{d^{obs}}$ are the true data plus noise ϵ

$$\mathbf{d^{obs}} = \mathbf{d} + \epsilon.$$

For electromagnetic data the number of observations typically outnumbers the number of model parameters in a parametric forward model. The inverse problem is therefore overdetermined and the solution involves minimizing an objective function which quantifies the misfit between observed and predicted data. A common choice is the least squares (L2) misfit function

$$\phi_d = \|\mathbf{W_d}(\mathbf{d}^{obs} - F\{\mathbf{m}\})\|^2. \tag{6}$$

The diagonal data weighting matrix W_d weights the contribution of a datum based on its estimated standard deviation σ_i

$$\mathbf{W_{d}}_{ii} = \frac{1}{\sigma_i}. \tag{7}$$

Minimization of the L2 norm is equivalent to maximizing the likelihood function of the data given the model (Menke, 1989). This assumes that

$$d_i^{obs} = d_i^{pred} + \epsilon_i, \tag{8}$$

the noise on the data is independent and Gaussian distributed ($\epsilon_i \sim N(0, \sigma_i)$). While the central limit theorem can be employed to justify the assumption of Gaussian noise, it is often difficult in practice to characterize the uncertainties on the data. Data uncertainty is usually estimated as a percentage of each observed datum plus a noise floor. This weighting is particularly important for inversion of time-domain electromagnetic data, which can decay over several orders of magnitude in the range of measured channels. Weighting the data by an estimated standard deviation ensures that early time, large amplitude data do not dominate the misfit. In addition, an appropriate floor value ensures that small amplitude data do not dominate the misfit after scaling by a percentage. The choice of data standard deviations remains something of an educated guess which can be informed by data pre-processing. For example, a noise floor can be estimated for each time channel by windowing regions where no significant signal is observed. In contrast, magnetic data have much less dynamic range and it is often sufficient to specify a noise floor of a few nanotesla when inverting for dipole model parameters.

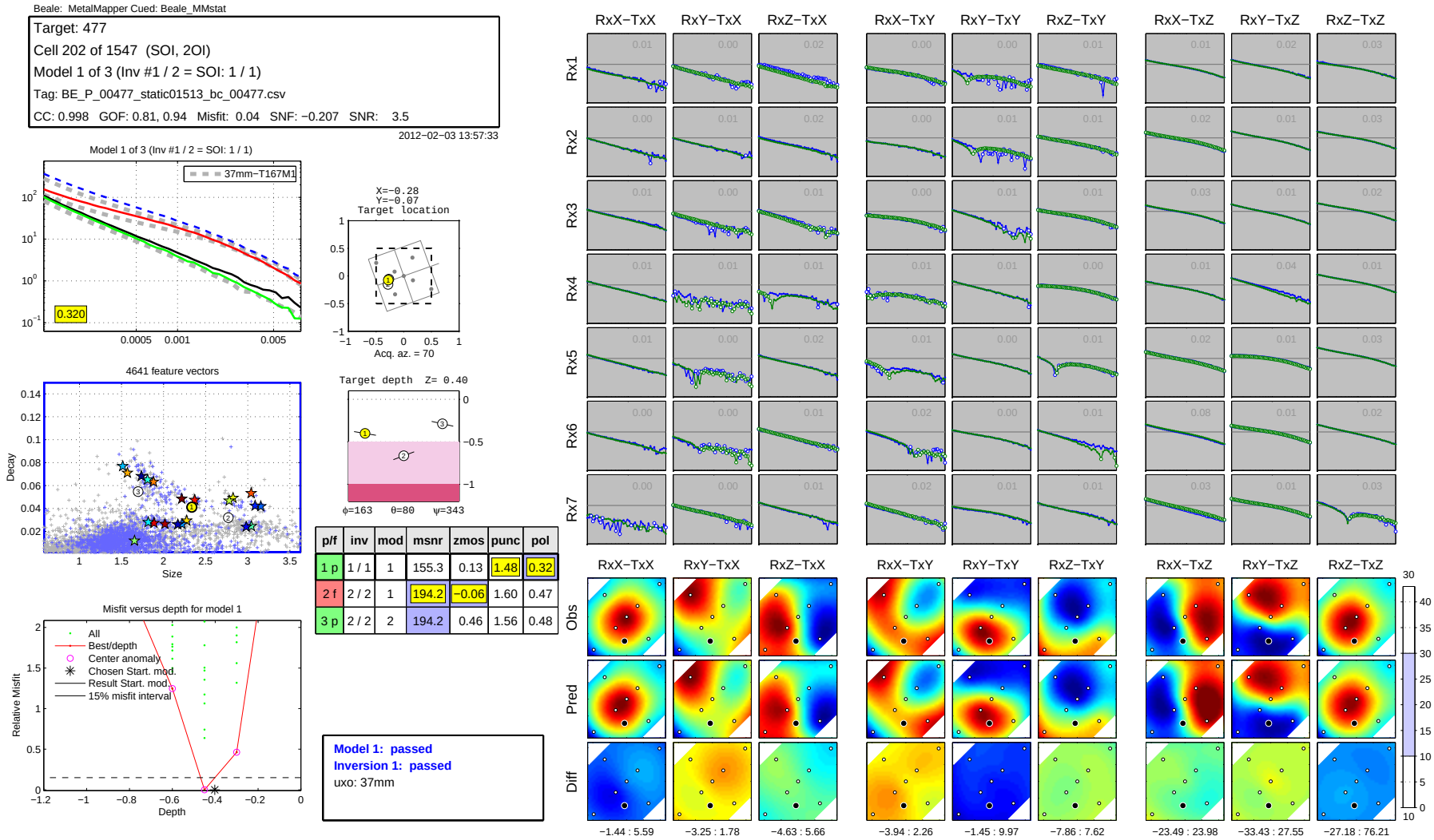
If the forward modelling operator is linear, then there is a single minimum to the misfit function and the best-fitting model can be obtained in one step by solving a linear system of equations. For a nonlinear forward model there may be multiple minima of the misfit function and the solution of the inverse problem cannot be obtained in one step. This is usually the case in UXO applications: all forward models described above are nonlinear functions of the input model parameters. Iterative approaches to the nonlinear inverse problem involve minimizing a quadratic approximation to the objective function with respect to the model perturbation ($\delta\mathbf{m}$) at each iteration. For example, the Gauss-Newton method solves

$$\mathbf{J}^T \mathbf{W_d}^T \mathbf{W_d} \mathbf{J} \delta\mathbf{m} = -\mathbf{J}^T \mathbf{W_d}^T \mathbf{W_d} (\mathbf{d^{obs}} - F\{\mathbf{m}\}) \tag{9}$$

with $\mathbf{J}$ the Jacobian matrix of sensitivities. Given an initial guess for the model parameters, we can solve the above equation for a model perturbation which will reduce the misfit. We then update our model with this perturbation and repeat the procedure until a convergence criterion is achieved (e.g. $\|\delta\mathbf{m}\| < \epsilon$). Iterative methods can converge to local, suboptimal minima and so it is common practice to initialize these algorithms from multiple starting models.

We emphasize that quality control (QC) of fits to observed data is a necessary and important step. Because we often have a poor handle on the noise, metrics such as the final data misfit and correlation coefficient may not always be reliable for deciding whether a fit is successful. QC'ing magnetic data is relatively quick, as there is only one channel of data to consider, but TEM data often requires visual inspection of multiple channels in plan view,

lines, and individual soundings to determine whether a fit is adequate. Figure 3 shows a display used for QC of MetalMapper data fits. Quality control is presently a major bottleneck in UXO data processing, and in section 4.1 we present a detailed analysis of methods for automating the QC process using MetalMapper data sets.



p/f	inv	mod	msnr	zmos	punc	pol
1 p	1 / 1	1	155.3	0.13	1.48	0.32
2 f	2 / 2	1	194.2	−0.06	1.60	0.47
3 p	2 / 2	2	194.2	0.46	1.56	0.48

FIGURE 3. Display for quality control of MetalMapper data fits.

2.3. **Classification.** The end product of geophysical data processing is a diglist which ranks targets from most to least likely to be ordnance, as well as a "stop dig" point (or operating point) on the diglist where digging can be safely terminated. In some cases targets beyond the operating point will be left in the ground. Some sites, however, will require total clearance in order to satisfy environmental regulations. In this case the template for field operations is excavation of high risk targets (as identified on the diglist) by expert disposal teams, with low risk targets excavated by labor under expert supervision. The cost savings of advanced discrimination is realized in the reduction in the number of targets dug by EOD (explosive ordnance disposal) technicians and the choice of operating point is less critical because we are guaranteed to find all detected ordnance.

To rank targets for digging, we use the information in our observed geophysical data. Features of the observed data, estimated without resorting to inversion with a physics-based model, can sometimes suffice as criteria to classify ordnance and non-ordnance targets. For example, in Williams et al. (2007) a bivariate Gaussian distribution is fit to observed EM61 data at each time channel and the average width of the anomaly, as measured by the estimated covariance matrix, is then used as a criterion to rank ordnance (wide anomaly) ahead of clutter (small anomaly). This approach significantly outperformed a statistical classification approach employing features estimated with the dipole model. This can work when ordnance is significantly larger than clutter, but may fail if there are large, deep clutter which can generate broad anomalies. Furthermore, a horizontal target can sometimes produce an anomaly which is better described by a bimodal distribution (i.e. two Gaussians, see Pasion (2007)). Data features are nonetheless useful when data quality is not sufficient to support estimation of useful parameters in an inversion or when time constraints preclude processing with inversion.

Parameters of models estimated from inversion can resolve some of the ambiguities of data features because model parameters can be related to intrinsic target properties. An intuitive template matching approach to classification compares estimated model parameters with those previously derived from a library of known targets. Classification with TEM data is often performed by comparing estimated polarization decays with library responses and then ranking a target based on some measure of closeness between observed and expected responses. Care must be taken here to use parameters which can be reliably estimated: late time polarizations are more susceptible to noise and poor polarization estimates may unduly affect the discrimination decision. Pasion et al. (2007) solve this problem with a fingerprinting algorithm that inverts for target location and orientation while holding polarizations fixed at their library values. Reducing the model's degrees of freedom in this way makes the inversion less susceptible to fitting the noise. Targets are then dug based upon the proposed library item which produces the best fit to the observed data. We can regard this method as incorporating information from our target library directly into the inversion, whereas conventional template matching uses library information in the classification stage.

Library methods assume that there is a true set of model parameters that, under ideal circumstances, can be perfectly reconstructed from an observed data set. Statistical classification algorithms which have been applied to UXO classification can be regarded as Bayesian solutions to the classification problem: we treat the parameters of interest as fundamentally uncertain random variables which are characterized by probability distributions. We then try to learn these probability distributions from a sample of labelled targets for which ground truth is known (the training data), and then formulate a decision rule that tries to minimize the probability of making an incorrect decision for unlabelled targets (the test data). One approach to formulating the decision rule is to fit some assumed parametric distributions to each class of targets in the training data, and then assign a test target to the class distribution which is most likely. The class distributions are defined in a multidimensional feature space spanned by some subset of estimated model parameters, or transformations thereof. The success of a statistical classifier is measured by its ability to generalize to the unseen test data (i.e. correctly classify), and having a training data set which is representative of class variability in the test data set is crucial. In Aliamiri et al. (2007), for example, class distributions are generated by simulating data for each target class in a range of orientations and depths, and then inverting these synthetic data. This assumes that simulations can capture the noise conditions which are encountered in experimental data. Alternatively, training data can be generated by full clearance of selected grids in a geophysical prove-out. Active learning techniques for iteratively selecting targets to build the training data set, based upon reducing uncertainties in the resulting classifier, are developed in Zhang et al. (2004b). In section 4.2 we further investigate active learning for UXO classification using recent methods developed at Duke.

3. Methods

3.1. **The Semi-supervised Learning Algorithm.** We introduce the details of a graph-based semi-supervised algorithm applied to UXO sensing. Semi-supervised learning is applicable to any sensing problem for which all of the unlabeled data are available at the same time, and therefore this approach is applicable to most wide-area sensing problems of interest to the UXO community. In practical applications semi-supervised learning has been found to yield superior performance relative to the widely applied supervised algorithms. However, all of the discussion simplifies to the case for which we consider purely supervised classifiers, and in the experiments with real data we have found that results with supervised classifiers are often adequate (due to sufficient training data). The presentation below for semi-supervised classifiers presents the framework in its most general sense.

3.2. **The Graph Representation of a Partially Labeled Data Manifold.** Let $G = (\mathcal{X}, \mathbf{W})$ be a graph, where $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ is the set of vertices and $\mathbf{W} = [x_{ij}]_{N\times N}$ is the affinity matrix with the (i, j)-th element w_{ij} indicating the strength of immediate connectivity between vertices $\mathbf{x}_i$ and $\mathbf{x}_j$. For the purpose of data classification, the vertex set $\mathcal{X}$ coincides with the set of data points (labeled or unlabeled), and w_{ij} is a quantitative measure of the closeness of data points $\mathbf{x}_i$ and $\mathbf{x}_j$. In the semi-supervised setting, only a subset of $\mathcal{X}$ are provided with class labels, and the remaining data points are unlabeled, and therefore we have a partially labeled graph.

Although there are many alternative ways of defining the connectivity w_{ij}, here we consider a radial basis function

$$w_{ij} = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma_i^2}) \tag{10}$$

where $\|\cdot\|$ represents the Euclidean norm; selection of the parameter σ_i is detailed below. While the affinity matrix may provide a reasonable local similarity among the data points, it is not a good representation of the global similarity measure of the data sets. Following Szummer and Jaakkola (2002), we construct a Markov random walk based on the affinity measure, which is capable of incorporating both the high-density clustering property and the manifold structure of the data set. Specifically, we induce a Markov transition matrix $\mathbf{A} = [a_{ij}]_{N\times N}$, where the (i, j)-th element

$$a_{ij} = \frac{w_{ij}}{\sum_{k=1}^{N} w_{ik}} \tag{11}$$

gives the probability of walking from $\mathbf{x}_i$ to $\mathbf{x}_j$ by taking a single step. In general we are interested in a t-step random walk, the transition matrix of which is given by $\mathbf{A}$ raised to the power of t, *i.e.*, $\mathbf{A}^t = [a_{ij}^{(t)}]_{N\times N}$. The $\mathbf{A}^t$ is row stochastic, where each element $a_{ij}^{(t)}$ represents the probability that the Markov process starts from $\mathbf{x}_i$ and ends at $\mathbf{x}_j$ by taking t-step random walks. As a special case, $\mathbf{A}^t$ degenerates to an identity matrix when $t = 0$, which means one can only stay at a single data point when no walk is performed.

In specifying the Markov transition matrix in (10) we have used a distinct σ for each data point $\mathbf{x}$. In the random walk, σ can be thought of as the step-size. Therefore location-dependent step-sizes allow one to account for possible heterogeneities in the data manifold — at locations where data are densely distributed a small step-size is enough, whereas at locations where data are sparsely distributed a large step-size is necessary to connect a data point to its nearest neighbor. A simple choice of the heterogeneous σ is to let σ_i to be a fraction of the shortest Euclidean distance between $\mathbf{x}_i$ and all other data points in $\mathcal{X}$. This ensures each data point is immediately connected to at least one neighbor.

3.3. **Neighborhood-Based Learning.** Any two data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are said to be t-step neighbors, denoted as $\mathbf{x}_j \overset{t}{\sim} \mathbf{x}_i$, if $a_{ij}^{(t)} > 0$. Then $\mathcal{N}_t(\mathbf{x}_i) = \{\mathbf{x} : \mathbf{x} \overset{t}{\sim} \mathbf{x}_i\} \subseteq \mathcal{X}$, which represents the set of t-step neighbors of $\mathbf{x}_i$, is called the t-step neighborhood of $\mathbf{x}_i$. When $t = 0$, the neighborhood shrinks to a single data point, $\mathcal{N}_0(\mathbf{x}_i) = \{\mathbf{x}_i\}$. We define the probability of label y_i given the t-step neighborhood of $\mathbf{x}_i$ as

$$p(y_i|\mathcal{N}_t(\mathbf{x}_i), \boldsymbol{\theta}) = \sum_{j=1}^{N} a_{ij}^{(t)}\, p(y_i|\mathbf{x}_j, \boldsymbol{\theta}) \tag{12}$$

where the magnitude of $a_{ij}^{(t)}$ automatically determines the contribution of $\mathbf{x}_j$ to the neighborhood, thus we are allowed to run the index j over the entire $\mathcal{X}$. Expression $p(y_i|\mathbf{x}_j, \boldsymbol{\theta})$ is the probability of label y_i given a single data point $\mathbf{x}_j$ (zero-step neighborhood) and it's represented by a standard probabilistic classifier parameterized by $\boldsymbol{\theta}$. We consider binary classification with $y \in \{-1, 1\}$, and choose the form of $p(y_i|\mathbf{x}_i, \boldsymbol{\theta})$ as logistic regression classifier

$$p(y_i|\mathbf{x}_j, \boldsymbol{\theta}) = \frac{1}{1 + \exp(-y_i \boldsymbol{\theta}^T \mathbf{x}_j)} \tag{13}$$

where we assume a constant element 1 is prefixed to each feature vector $\mathbf{x}$ (the prefixed $\mathbf{x}$ is still denoted as $\mathbf{x}$ for notational simplicity), thus the first element in $\boldsymbol{\theta}$ is a bias term. Arbitrarily one may set $y = 1$ as corresponding to a UXO, and $y = -1$ as corresponding to a non-UXO.

The fundamental difference between the classifier in (12) and the typical logistic regression classifier is that the logistic-regression classifier predicts y_i using $\mathbf{x}_i$ alone, while the semi-supervised approach considered here predicts y_i by using $\mathbf{x}_i$ and the feature vectors in the neighborhood of $\mathbf{x}_i$. The neighborhood of $\mathbf{x}_i$ is formed by all $\mathbf{x}_j$'s that can be reached from $\mathbf{x}_i$ by t-step random walks, with each $\mathbf{x}_j$ contributing to the prediction of y_i in proportion to $a_{ij}^{(t)}$, the probability of walking from $\mathbf{x}_i$ to $\mathbf{x}_j$ in t steps. The role of neighborhoods is then conspicuous — in order for $\mathbf{x}_i$ to be labeled y_i, each neighbor $\mathbf{x}_j$ must be labeled consistently with y_i, in the degree proportional to $a_{ij}^{(t)}$; in such a manner, y_i implicitly propagates over the neighborhood. By taking the neighborhoods into account, it is possible to learn a classifier with only a few labels present and yet the classifier learned is much less subject to over-fitting than when ignoring the neighborhoods. This is addressed in greater detail below.

Let $\mathcal{L} \subseteq \{1, 2, \cdots, N\}$ denote the set of indices of labeled data. Assuming the labels are conditionally independent, we obtain the likelihood function

$$\begin{aligned} p(\{y_i, i \in \mathcal{L}\} | \{\mathcal{N}_t(\mathbf{x}_i) : i \in \mathcal{L}\}, \boldsymbol{\theta}) &= \prod_{i \in \mathcal{L}} p(y_i | \mathcal{N}_t(\mathbf{x}_i), \boldsymbol{\theta}) \\ &= \prod_{i \in \mathcal{L}} \sum_{j=1}^{N} a_{ij}^{(t)} \, p(y_i | \mathbf{x}_j, \boldsymbol{\theta}) \end{aligned} \tag{14}$$

which is the joint probability of observed labels given the t-step neighborhood of each corresponding data point. Estimation of θ may be achieved by maximizing the log-likelihood, which however may yield over-fitting, especially when the number of labeled samples is small. To enforce sparseness of $\boldsymbol{\theta}$ (sparseness has been demonstrated as an important property Tipping (2001), discouraging overfitting), we impose a zero-mean Gaussian prior on each dimension of $\boldsymbol{\theta}$,

$$p(\boldsymbol{\theta} | \Lambda) = \frac{|\Lambda|^{1/2}}{(2\pi)^{d/2}} \exp(-\frac{1}{2} \boldsymbol{\theta}^t \Lambda \boldsymbol{\theta}) \tag{15}$$

where $\Lambda = diag\{\lambda_1, \lambda_2, ..., \lambda_d\}$ are hyper-parameters, d is the dimensionality of $\mathbf{x}$. Each hyper-parameter has an independent Gamma distribution, resulting in

$$\begin{aligned} p(\Lambda | \alpha, \beta) &= \prod_{i=1}^{d} \text{Gamma}(\lambda_i | \alpha_i, \beta_i) \\ &= \prod_{i=1}^{d} \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)} \lambda_i^{\alpha_i - 1} \exp(-\lambda_i \beta_i) \end{aligned} \tag{16}$$

Marginalizing Λ, we obtain the prior distribution conditional directly on α and β,

$$p(\boldsymbol{\theta} | \alpha, \beta) = \int p(\boldsymbol{\theta} | \Lambda) p(\Lambda | \alpha, \beta) \, d\Lambda \tag{17}$$

The posterior of $\boldsymbol{\theta}$ follows from (14) and (17),

$$\begin{aligned} & p(\boldsymbol{\theta} | \alpha, \beta, \{y_i, \mathcal{N}_t(\mathbf{x}_i) : i \in \mathcal{L}\}) \\ &= Z^{-1} \prod_{i \in \mathcal{L}} \sum_{j=1}^{N} a_{ij}^{(t)} \, p(y_i | \mathbf{x}_j, \boldsymbol{\theta}) \int p(\boldsymbol{\theta} | \Lambda) p(\Lambda | \alpha, \beta) \, d\Lambda \end{aligned} \tag{18}$$

where Z is a normalization constant. We are interested in the maximum *a posterior* (MAP) estimate of $\boldsymbol{\theta}$, which maximizes (18) or, equivalently,

$$\begin{aligned} \ell(\boldsymbol{\theta}) &\stackrel{def.}{=} \ln p(\boldsymbol{\theta} | \alpha, \beta, \{y_i, \mathcal{N}_t(\mathbf{x}_i) : i \in \mathcal{L}\}) + \ln Z \\ &= \sum_{i \in \mathcal{L}} \ln \sum_{j=1}^{N} a_{ij}^{(t)} \, p(y_i | \mathbf{x}_j, \boldsymbol{\theta}) \\ &\quad + \ln \int p(\boldsymbol{\theta} | \Lambda) p(\Lambda | \alpha, \beta) \, d\Lambda \end{aligned} \tag{19}$$

The $\boldsymbol{\theta}$ obtained by maximization of $\ell(\boldsymbol{\theta})$ generally is not subject to over-fitting due to two reasons — the neighborhoods incorporated into the first term of $\ell(\boldsymbol{\theta})$ encourages smoothness along the manifold, and the second term of $\ell(\boldsymbol{\theta})$ enforces sparseness of $\boldsymbol{\theta}$.

3.4. **The Learning Algorithm.** We maximize (19) by employing an expectation-maximization (EM) algorithm. For any $\{\delta_{ij} : \delta_{ij} \geq 0, \sum_{j=1}^{N} \delta_{ij} = 1\}$ and $\{q(\Lambda) :\in q(\Lambda) d\Lambda = 1\}$, we apply Jensen's inequality to the righthand side of (19) to obtain the lower bound

$$\begin{aligned} \ell(\boldsymbol{\theta}) \geq Q(\boldsymbol{\theta}|\delta, q) \stackrel{def.}{=} & \sum_{i \in \mathcal{L}} \sum_{j=1}^{N} \delta_{ij} \ln \frac{a_{ij}^{(t)} p(y_i|\mathbf{x}_j, \boldsymbol{\theta})}{\delta_{ik}} \\ & + \int q(\Lambda) \ln \frac{p(\boldsymbol{\theta}|\Lambda) p(\Lambda|\alpha, \beta)}{q(\Lambda)} d\Lambda \end{aligned} \tag{20}$$

where the equality holds when

$$\delta_{ij} = \frac{p(y_i|\mathbf{x}_j, \boldsymbol{\theta}) a_{ij}^{(t)}}{\sum_{k=1}^{N} p(y_i|\mathbf{x}_k, \boldsymbol{\theta}) a_{ik}^{(t)}} \tag{21}$$

$$q(\Lambda) = \frac{p(\theta|\Lambda) p(\Lambda|\alpha, \beta)}{\int p(\theta|\Lambda) p(\Lambda|\alpha, \beta) d\Lambda} \tag{22}$$

The EM algorithm consists of iteration of the following two steps.

(1) E-step: computing $\{\delta_{ij}\}$ and $q(\boldsymbol{\Lambda})$ using (21) and (22);
(2) M-step: compute the re-estimate of $\boldsymbol{\theta}$ as

$$\boldsymbol{\theta} = \arg\max_{\widehat{\boldsymbol{\theta}}} Q(\widehat{\boldsymbol{\theta}}|\delta, q) \tag{23}$$

The convergence is monitored by checking $\ell(\boldsymbol{\theta})$, which is guaranteed to monotonically increase over the EM iterations.

There are two noticeable points regarding the technical details. First, since (16) is conjugate to (15), $q(\Lambda)$ is of the same form as (16) with updated hyper-parameters α, β,

$$\begin{aligned} q(\Lambda) &= \prod_{i=1}^{d} \text{Gamma}(\lambda_i|\alpha_i + \frac{1}{2}, \beta_i + \frac{1}{2}\theta_i^2) \\ &= \prod_{i=1}^{d} \frac{(\beta_i + \frac{1}{2}\theta_i^2)^{\alpha_i + \frac{1}{2}}}{\Gamma(\alpha_i + \frac{1}{2})} \lambda_i^{\alpha_i - \frac{1}{2}} e^{-\lambda_i(\beta_i + \frac{1}{2}\theta_i^2)} \end{aligned} \tag{24}$$

and the integral in the dominator of (22) has an analytic form

$$\begin{aligned} & \int p(\boldsymbol{\theta}|\Lambda) p(\Lambda|\alpha, \beta) d\Lambda \\ & = \frac{1}{(2\pi)^{d/2}} \prod_{i=1}^{d} \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)} \frac{\Gamma(\alpha_i + \frac{1}{2})}{\left(\beta_i + \frac{1}{2}\theta_i^2\right)^{\alpha_i + \frac{1}{2}}} \end{aligned} \tag{25}$$

which is useful in checking the convergence of $\ell(\boldsymbol{\theta})$ in (19).

Secondly, in computing $Q(\widehat{\boldsymbol{\theta}}|\delta, q)$ by (20), one needs to compute $\gamma(\widehat{\boldsymbol{\theta}}) \stackrel{def.}{=} \int q(\Lambda) \ln p(\widehat{\boldsymbol{\theta}}|\Lambda) d\Lambda$, and it is found that

$$\begin{aligned} \gamma(\widehat{\boldsymbol{\theta}}) &= -\frac{1}{2}\widehat{\boldsymbol{\theta}}^T \mathbb{E}_q(\Lambda|\theta)\widehat{\boldsymbol{\theta}} \\ &= -\frac{1}{2}\widehat{\boldsymbol{\theta}}^T \mathrm{diag}\left[\mathbb{E}_q(\lambda_1), \mathbb{E}_q(\lambda_2), \cdots, \mathbb{E}_q(\lambda_d)\right]\widehat{\boldsymbol{\theta}} \end{aligned} \tag{26}$$

with

$$\mathbb{E}_q(\lambda_i) = \frac{\alpha_i + \frac{1}{2}}{\beta_i + \frac{1}{2}\theta_i^2}. \tag{27}$$

3.5. **Active Learning.** In the UXO-classification problem, it is a given that excavation will ultimately be performed. The principal objective is to excavate as high a percentage of UXO as possible, while leaving as much of the non-UXO as possible unexcavated. Recall that the primary expense in UXO cleanup is the excavation of non-UXO items, since the density of such is typically much higher than the amount of UXO, and the sensor signatures of UXO are often very similar to those of many types of non-UXO. Given that excavation will be performed in any case, one may ask whether the initial set of excavations may be performed with the purpose of improving the performance of the algorithm. Specifically, one may ask which unlabeled sensor signature would be most informative to improved classifier performance if the associated label could be made available. As discussed below, this question is answered in a quantitative information-theoretic manner. When the expected information content of such an excavation drops below a prescribed threshold, excavation for the purpose of improved learning is terminated, and then the algorithm is used to define the probability that all remaining unlabeled signatures correspond to UXO. Importantly, in active learning the algorithm desires to learn about the properties of the UXO *and* non-UXO at the site, and therefore in this phase an excavated non-UXO should not be termed a "false alarm". Such active learning has been performed previously in a related UXO-cleanup study Zhang et al. (2004a); the distinct character of the algorithm discussed below is that this process is here placed within the context of semi-supervised learning.

3.6. **Active Learning with Semi-Supervised Classifier.** For active label selection, we consider a Gaussian approximation of the posterior of the classifier

$$p(\theta|D) \simeq \mathcal{N}(\theta|\hat{\theta}, \mathsf{H}^{-1}) \tag{28}$$

where $\hat{\theta}$ is the estimate of the classifier learned from the above EM algorithm, and H is the posterior precision matrix $\mathsf{H} = \nabla^2(-\log p(\boldsymbol{\theta}|\{y_i, \mathcal{N}_t(\mathbf{x}_i) : i \in \mathcal{L}\})$. By treating $\gamma(\widehat{\boldsymbol{\theta}})$ in (26) as deterministic, we obtain an evidence-type approximation Tipping (2001):

$$\begin{aligned} \mathbf{H} = &\sum_{i\in\mathcal{L}}\sum_{j=1}^{N}\delta_{ij}\, p(y_i|\mathbf{x}_j, \boldsymbol{\theta})(1 - p(y_i|\mathbf{x}_j, \boldsymbol{\theta}))\mathbf{x}_j\mathbf{x}_j^T \\ &-\nabla^2 \ln \gamma(\widehat{\boldsymbol{\theta}}) \end{aligned} \tag{29}$$

With one more data point x_{i*} with label y_{i*} as the next labeled data, assuming that the MAP estimate of $\widehat{\boldsymbol{\theta}}$ remains the same after including the new data point, then the posterior precision changes to

$$\begin{aligned}\mathbf{H}' = &\sum_{i'\in\mathcal{L}\cup\{i*\}}\sum_{j=1}^{N}\delta_{i'j}\, p(y_{i'}|\mathbf{x}_j,\boldsymbol{\theta})(1-p(y_{i'}|\mathbf{x}_j,\boldsymbol{\theta}))\mathbf{x}_j\mathbf{x}_j^T \\ &-\nabla^2 \ln\gamma(\widehat{\boldsymbol{\theta}})\end{aligned} \tag{30}$$

For active label selection, we could further simplify the equation for the precision matrix by considering the degenerated connectivity matrix $A^{(t=0)}$, which is an identity matrix, such that

$$\delta_{ij} = \{ \begin{array}{l} 1, \text{ for } i=j \\ 0, \text{ for } i\neq j \end{array} \tag{31}$$

Following this, the new precision matrix becomes

$$\mathbf{H}' = \mathbf{H} + p(y_{i*}|\mathbf{x}_{i*},\boldsymbol{\theta})(1-p(y_{i*}|\mathbf{x}_{i*},\boldsymbol{\theta}))\mathbf{x}_{i*}\mathbf{x}_{i*}^T \tag{32}$$

Our criterion for active learning is to choose the feature vector for labeling that maximizes the mutual information between the classifier $\boldsymbol{\theta}$ and the new data point to be labeled, which is the expected decrease of the entropy of $\boldsymbol{\theta}$ after x_{i*} and y_{i*} are observed,

$$\begin{aligned} I &= \frac{1}{2}\log\frac{|\mathbf{H}'|}{|\mathbf{H}|} \\ &= \frac{1}{2}\log\big\{1+p(y_{i*}|\mathbf{x}_{i*},\boldsymbol{\theta})[1-p(y_{i*}|\mathbf{x}_{i*},\boldsymbol{\theta})]\mathbf{x}_{i*}^T H^{-1}\mathbf{x}_{i*}\big\}\end{aligned} \tag{33}$$

The mutual information I is large when $p(y_{i*}|\mathbf{x}_{i*},\boldsymbol{\theta}) \approx 0.5$, therefore, our active learning prefers label acquisition on samples with uncertain classification, based on the current classifier based upon available labeled data. Further, considering the term $\mathbf{x}_{i*}^T H^{-1}\mathbf{x}_{i*}$, the mutual information criterion prefers samples with high variance.

The assumption that the mode of the posterior distribution of the classifier remains unchanged with one more labeled data point is not good at the beginning of the active learning procedure. However, empirically we have found that it is a very good approximation after the active learning procedure has acquired as few as 15 labels, for the examples considered here. In practice the computational cost associated with retraining the classifier with each active-labeled-acquired labeled data is insignificant relative to the time required for excavation, and therefore the classifier weights are updated with each new acquired label.

4. Results and Discussion

4.1. Comparison of Expert QC, Auto QC and No QC using MetalMapper data.

4.1.1. *Introduction.* Prior to construction of a dig list (classification), data and inversion results usually undergo a quality control (QC) check with the primary objective being to fail models (or entire inversion results) deemed unreliable and which may negatively impact

the performance of the classification process. With MetalMapper data we typically run two inversions to solve for model parameters associated with (1) a single object (SOI); and (2) two objects (2OI); these produce three different models of the underlying putative target. The two inversions produce three independent models for each anomaly. A model that is failed during QC is not used during classification. An anomaly for which all models are failed during QC is categorized as "cannot analyze". Anomalies in this category must be dug and accordingly are placed at the top of a dig list. During QC an inversion may, for example, be failed if the fit between the predicted and observed data exceeds some misfit criteria, or visually if the fit is judged to be poor (e.g., Figure 4). A model may be failed if, for example, the predicted location falls on an inversion boundary and/or the predicted polarizabilities are judged to be unrealistic. This commonly occurs in 2OI solutions and is characterized by a model that is very deep, frequently lying on or near a horizontal inversion boundary, with polarizabilities that are relatively large in amplitude (e.g., Figure 5). It is not uncommon that such a model has the minimum polarizability misfit with respect to reference polarizabilities. Because classification is typically based on polarizability matching, these types of models must be omitted (failed).

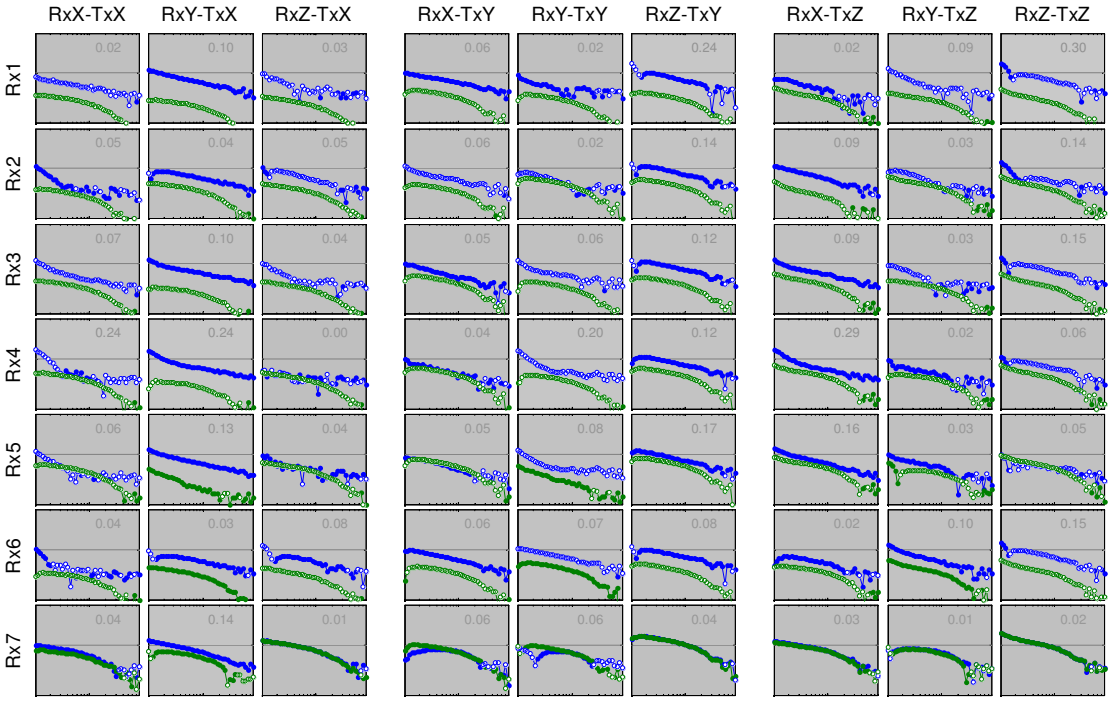


FIGURE 4. Anomaly 1951 of the Beale C MetalMapper dataset (37mm projectile at 11cm depth). The misfit between observed (blue lines and dots) and predicted (green lines and dots) is very large for almost all receiver/transmitter combinations. This inversion result should be classified as "cannot analyze."

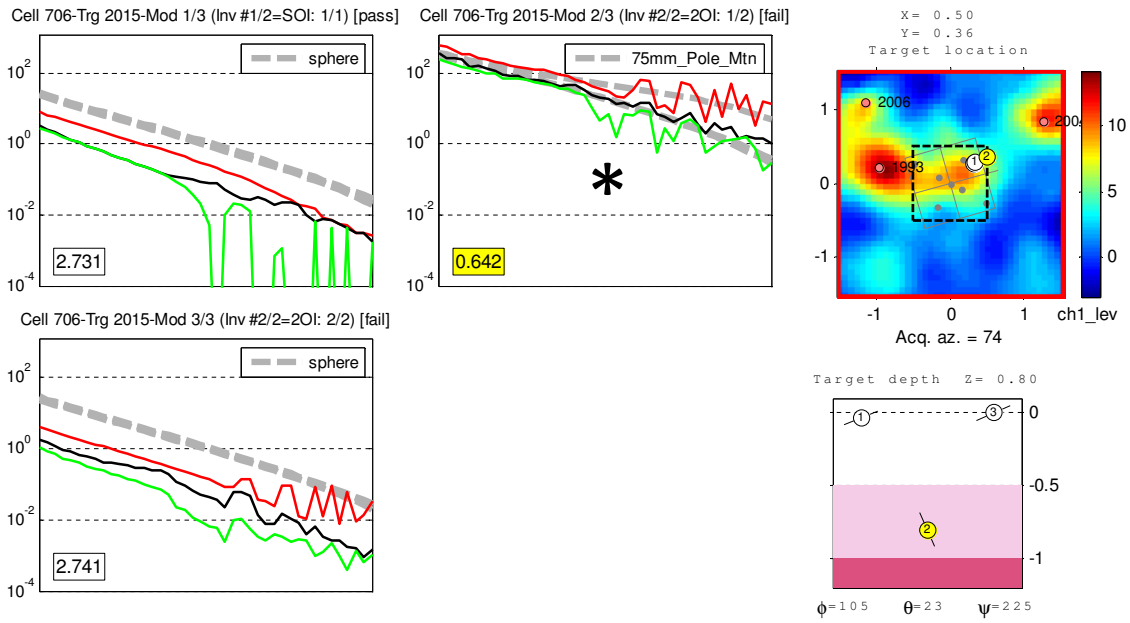


FIGURE 5. Anomaly 2015 of the Beale C MetalMapper dataset (small frag at 4cm depth). In this example one of the models from the 2OI (model 2) is unrealistic. The predicted location (yellow circled numbered "2" in target location map; top right) lies on an inversion boundary (dashed line), just outside the frame of the MetalMapper (grey square). The location map is underlain by the gridded EM61 data, which shows that the anomaly is very weak. The predicted depth for model 2 (lower right) is very deep (80cm). The predicted polarizabilities for model 2 (red, black and green lines in panel with asterisk) are very large in amplitude in relation to the polarizabilities predicted for the other models, and with respect to the weak EM61 anomaly. These are classic symptoms on an unrealistic model which should be failed and not considered in the classification stage. Because model 2 provides the best fit to one of the reference polarizabilities (75mm; broken grey lines in the polarizability plots) a dig list based on polarizability matching would place this anomaly much earlier in the list if model 2 was included in the classification process.

Visual QCing of a dataset can be a tedious and time consuming process, particularly for large datasets. Because of this, even with the best QC tools at hand, QCing is a process that is subject to errors, one of which may prove costly by resulting in a TOI not being dug. Inconsistency is also an issue; due to the somewhat subjective nature of the QC process, a dataset QCed by different analysts will invariably result in different model selections, which may result in dig lists of varying levels of success. At some level visual QC of data may always be desirable due to the ability of the human eye of an experienced analyst to detect issues with the data or model that a specific set of quantitative measures may not pick up.

However, as datasets become larger, or for working with data in the field under tight time constraints, some element of automated QC would be beneficial for decreasing the overall analysis time and providing reliable QC decisions based on a specific set of criteria based, for example, on measures of data and/or model quality. To investigate this, we use cued MetalMapper data from recent live site demonstrations to investigate the performance of dig lists created from datasets that have been QCed using different methods:

(1) Expert QC: visual QC performed by an experienced analyst.
(2) Auto QC: automated QC based on a specific set of rules relating to measures of data or model quality.
(3) No QC: all models are used; no models are failed. The dig lists we generate for our tests are based on simple criteria such as polarizability match with known reference items and/or polarizability decay. The four datasets we use, and measures of dataset quality, are listed in Table 1.

Dataset	**N (All)**	**N (TOI)**	**DS (All)**	**DS (TOI)**	**MSNR (All)**	**MSNR (TOI)**	**Pol. Qual. (TOI)**	**L123 Msft (TOI)**
Beale P	1438	131	1.42	0.30	40.60	157.00	3.48	0.25
Beale C	1438	131	1.68	0.48	17.30	146.00	3.15	0.32
Butner	2304	171	1.20	0.07	60.10	192.00	2.64	0.41
Pole	2370	160	0.66	-0.69	146.00	250.00	6.78	0.12

TABLE 1. MetalMapper datasets used for testing. These are described in more detail in the text. All/TOI refers to all anomalies (from all passed models as determined by expert QC) and TOI anomalies, respectively. N is the number of anomalies. DS is median data shoddiness - an ad hoc measure of data/model inferiority (described in more detail below) - lower values are better. MSNR is median model signal-to-noise ratio calculated using predicted and residual data - higher values are better. Pol. Qual. is median polarizability quality - an ad hoc measure of polarizability smoothness and shape - higher values are better. L123 Msft is the median minimum misfit with all reference items calculated using all three polarizabilities (L1, L2 and L3) - lower values are better. Numbers highlighted in green/red correspond to the best/worst values for each measure. Beale P refers to data collected by Parsons at Camp Beale; Beale C refers to data collected by CH2M Hill at Camp Beale using the same instrument.

4.1.2. *Test Sets 1 and 2: Beale MetalMapper P and C.* MetalMapper (MM) data were collected at the Camp Beale live site demo (July 2011) by two different production groups: (1) Parsons (P); and (2) CH2M Hill (C). The two groups used the same instrument and, as far as is known, acquisition parameters. Differences in the two datasets should be due primarily to field practices which could, for example, affect the accuracy with which the instrument was centered over an anomaly, or processing approach (such as selection of appropriate background files for background noise subtraction). Total number of anomalies in

the Beale dataset is 1438 with 131 of these being TOI. TOI fall into five classes: (105mm, 81mm, 60mm, 37mm and ISO). Smaller items such as fuzes are treated as clutter in these tests. In the first test set we use the Parsons MetalMapper data which, by most measures of data and model quality, is slightly better than the CH2M Hill dataset (Table 1). Even for TOI with the poorest quality data the recovered primary polarizabilities using Parsons data are reasonably accurate with respect to the polarizabilities of the known item based on ground truth. For the CH2M Hill data, there are 2 TOI for which the recovered primary polarizabilities do not closely match any reference polarizability.

Test Set 1: Beale MetalMapper P. A decay versus size feature space plot for the expert-QCed Beale P data, including ground truth information, is shown in Figure 3. Based on the ground truth the generally good separation between TOI and non-TOI suggests that classification should be relatively straightforward. Notice that the expert visual QC resulted in failing a large number (approximately two-thirds) of the models, as defined by a human expert in the viewing of UXO data.

Figure 7 shows ROC curves for two dig lists created independently by different analysts using different approaches. Both dig lists were based on a dataset that had undergone the same visual QC by an expert analyst. One of the dig lists used a simple approach based primarily on a match between all three polarizabilities, as well as polarizability size, decay and quality. This list did not find all TOI before the stop dig point. All TOI were found after 595 non-TOI digs. The second dig list used a Support Vector Machine (SVM) two stage discrimination strategy, with early digs trained on all polarizabilities and later digs trained on total polarizability (L1+L2+L3). This list was more successful, finding all TOI before the stop dig point after 264 non-TOI digs. The latter represents our best result for the Beale P dataset and can be considered as the baseline for comparisons with the tests presented below.

Figure 8 shows ROC curves for dig lists derived from expert-QCed data using matching on the primary polarizability (L1) to determine dig order. Two results are shown using (1) all 42 time channels (maximum t=7.91ms); and (2) the first 30 time channels (maximum t=2.23ms) for computing the polarizability misfit fit. Surprisingly, the performance of this very simple approach to dig list construction are significantly better than the best of the officially submitted dig lists (Figure 7), with all TOI found after 124 and 153 non-TOI digs, respectively.

Figure 9 shows an equivalent set of ROC curves based on matching of all three polarizabilities. Note the performance is much poorer because the data are not capable of constraining the secondary and tertiary polarizabilities for some of the TOI (Figure 10). For the remainder of the results presented for the Beale datasets we will omit ROC curves derived based on a match to all three polarizabilities because these results are all inferior to the results based on L1 matching. We will also omit ROC curves based on matching on the first 30

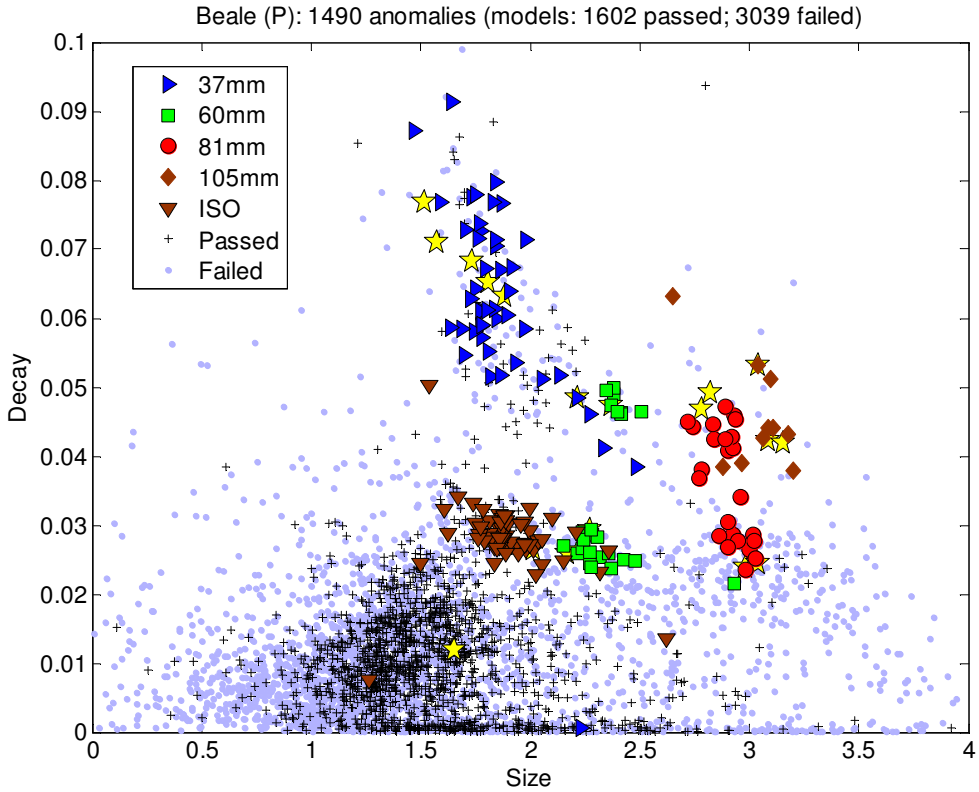


FIGURE 6. Decay versus size feature space plot for Beale P data showing all passed ("+") and failed (blue dot) models as determined by visual QC performed by an expert analyst (expert QC). Yellow stars represent reference items. Other large symbols represent TOI for passed models. Passed models indicated by "+" are non-TOI.

time channels because the results obtained using all time channels are consistently either better or approximately the same.

Figure 11 shows ROC curves based on L1 matching for data with no QC. Using both SOI and 2OI models results in all TOI being found after 268 non-TOI digs. This performance is similar to that of the SVM-based dig list shown in Figure 4. Interestingly, using only the SOI model for each anomaly provides much better performance, with all TOI found after 126 non-TOI digs. Clearly the non-QCed dataset with both SOI and 2OI models contains several non-TOI items with 2OI models that provide a good L1 match to a reference item. The performance of the SOI-only dataset is similar to that obtained with the expert-QCed dataset (Figure 8).

As an alternative to polarizability matching, a more conservative dig list can be created based wholly, or in part, on the decay of the total polarizability (measured between time channels 1 = 0.106ms and 29 = 2.006ms). Figure 12 shows the ROC curve for a dig list

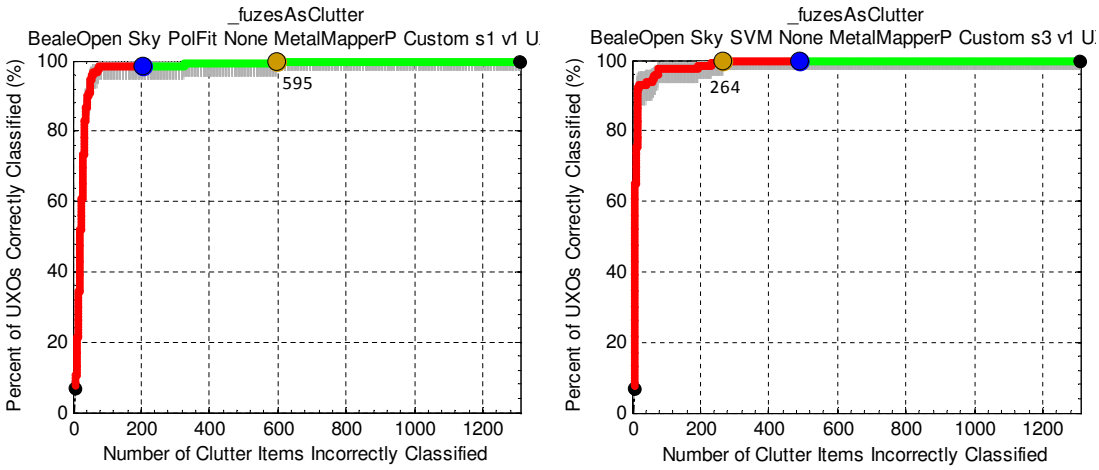


FIGURE 7. Official scoring for Beale P using Expert-QCed data. Dig list order for the ROC curve on the left was based primarily on a simple match to all three polarizabilities, as well as polarizability size, decay and quality. The ROC curve on the right is based on a dig list constructed using a Support Vector Machine (SVM) two stage discrimination strategy with early digs trained on all polarizabilities and later digs trained on total polarizability (L1+L2+L3). Blue dot denotes stop dig point. Yellowish dot denotes point at which all TOI are found. The simple approach missed two TOI; the final TOI was found after 595 non-TOI digs. The SVM approach found all TOI after 264 non-TOI digs.

based only on decay using data with no QC. The performance is significantly worse than the dig lists based on L1 matching which used the expert-QCed dataset (Figure 8) or the SOI-model-only dataset with no QC (Figure 11).

A less conservative approach would be to base the dig list order on polarizability matching for early digs, and decay for later digs. Figure 13 shows ROC curves for two dig lists that employ this approach with data that have undergone no QC. The dig list that transitions to using decay after 200 digs performs well, but still not as good the dig lists based on L1 matching which used the expert-QCed dataset (Figure 8) or the SOI-model-only dataset with no QC (Figure 11). However, we shall see below that with the Beale C dataset, in

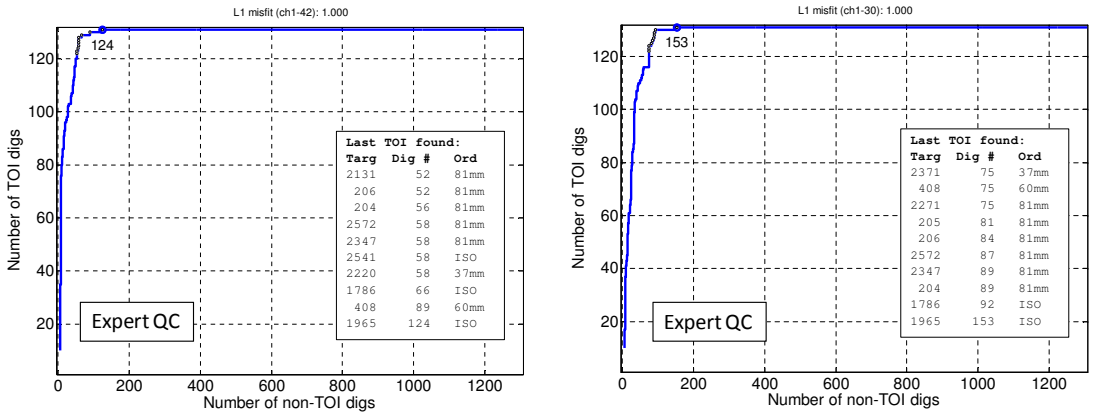


FIGURE 8. ROC curves for Beale P using Expert-QCed data. Dig list order is based on match between primary polarizability (L1) of the predicted and best fitting reference item. The ROC curve on the left used all 42 time channels (0.11-7.91ms) when computing fits; the ROC curve on the right used the first 30 time channels (0.11-2.23ms). Labeled point on the ROC curve denotes the last TOI to be dug. Number refers to the number of non-TOI digs. Inset table lists the anomaly number (Targ), the corresponding non-TOI dig number (Dig #) and the type of ordnance (Ord) for the last ten TOI dug.

which the quality of the recovered polarizabilities is poor for a number of TOI, the approach of transitioning to a list based on decay is more beneficial.

For testing the performance of automated QC we first use a simple decision process for passing or failing a model based on three data and model metrics (Figure 14):

(1) Model SNR (MSNR) is a measure of SNR using the ratio of the size of the predicted data to the (smoothed) data residuals.

(2) Data shoddiness (DS) is an ad hoc measure of data/model inferiority, combining several different measures: (1) data misfit (residual divided by observed); (2) correlation between observed and predicted data; (3) jitter (point-to-point difference) in the observed data; (4) fraction of data above the standard deviation; and (5) size of the difference between L2 and L3.

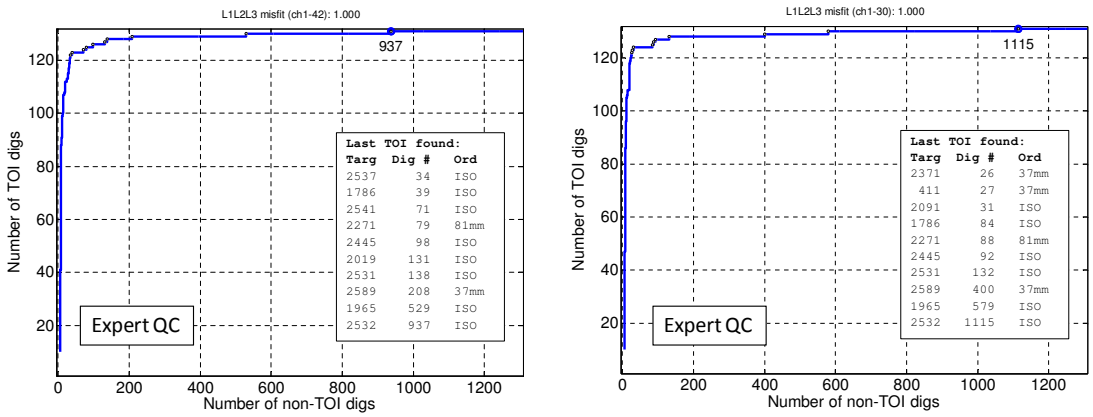


FIGURE 9. ROC curves for Beale P using Expert-QCed data. Dig list order is based on match between all three polarizabilities (L1, L2 and L3) of the predicted and best fitting reference item. The ROC curve on the left used all 42 time channels (0.11-7.91ms) when computing fits; the ROC curve on the right used the first 30 time channels (0.1-2.23ms).

(3) Predicted target depth (Z).

The decision process comprises three criteria (Figure 14). The no contact criterion tries to identify cases where the data are of very poor quality because there is no object within the instrument's field of view. The model-based criterion fails models with unrealistically deep predicted depths. The data-based criterion fails models based on poor quality data and with non-UXO like polarizabilities.

We tried three different variations based on the scheme shown in Figure 14. Test 1 used the criteria shown in Figure 14. The resulting dig list found all TOI after 235 non-TOI digs. Figure 15 shows the result, in feature space, of applying Test 1 versus no QC. The auto QC process resulted in 672 failed models. Many of these are large in size and lie in a position in feature space that is typical of a relatively strong ground response. Note that in comparison to expert QC, the auto QC Test 1 failed far fewer models (672 versus 3039). In Test 2 the model based criteria was changed to $Z > 0.6$m. Models with large predicted depths tend

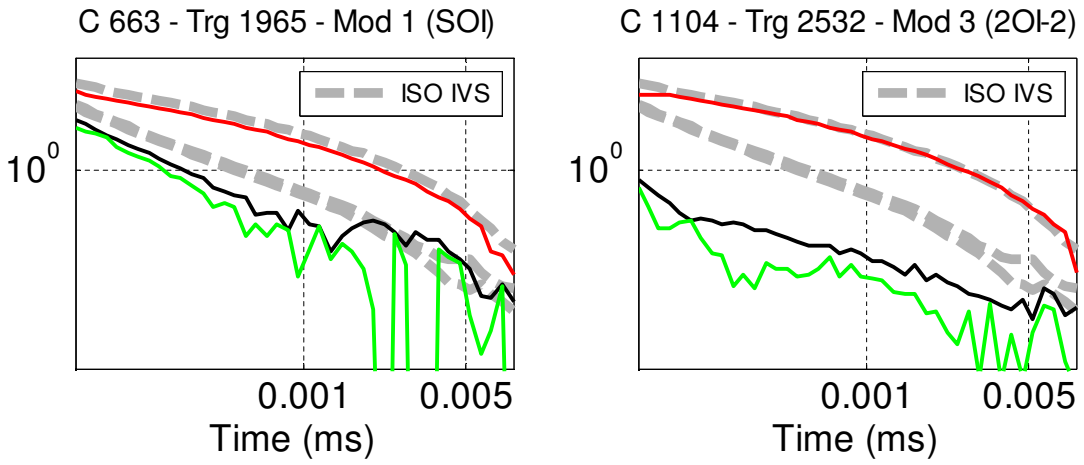


FIGURE 10. Predicted (red, black green lines) and best fitting reference polarizabilities (broken grey lines) for the two most difficult TOI of the Beale P dataset. Anomaly 1965 (left) is an ISO at 20cm depth; anomaly 2532 (right) is an ISO at 19cm depth. Note the poor quality of L2 and L3 (black and green lines, respectively); however, both of these anomalies show a reasonably good L1 (red line) match with the ISO reference polarizabilities.

to be unrealistic. The risk in reducing the depth cutoff is that a valid, and perhaps best, model will be eliminated. The resulting dig list found all TOI after 184 non-TOI digs. Test 3 used the same criteria as Test 2, but the auto QC was applied only to the 2OI models; all SOI models were passed. The resulting dig list found all TOI after 169 non-TOI digs. ROC curves for these tests are shown in Figure 16.

All three auto QC tests produced results which perform better than the SVM-based dig list (Figure 7). However, none of auto QC tests performed as well as the expert-QCed dig list, (Figure 8) or the dig list using only SOI models with no QC (Figure 11).

In Figure 17 we show another simple decision process (auto QC Test 4) designed specifically to eliminate unrealistic deep 2OI models. If model a is a 2OI model, it is failed if it is either (1) absolutely deep; or (2) relatively deep in relation to the other 2OI model (b) and the data/model are of poor quality. Figure 18 shows the result, in feature space, of applying Test 4 versus no QC. The auto QC process resulted in 241 failed models.

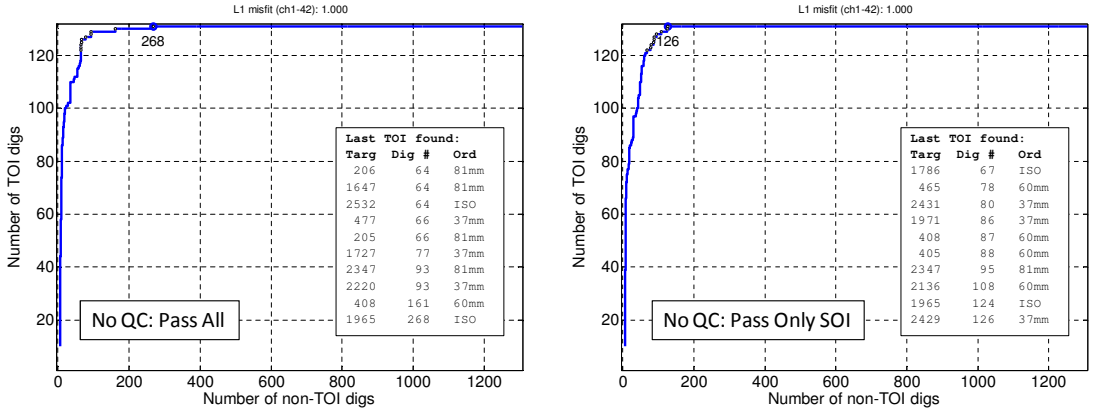


FIGURE 11. ROC curves for Beale P using No QC (no models were failed). Dig list order is based on match between primary polarizability (L1) of the predicted and best fitting reference item. For the ROC curve on the left both SOI and 2OI models were used; the ROC curve on the right used only the SOI model for each anomaly.

Figure 19 shows ROC curves for dig lists based on L1 match, decay, and combinations of L1 match and decay using auto QC Test 4. For all of these dig lists, the performance is marginally better than not applying auto QC. However, none of these lists perform as well as the dig list based on L1 matching which used the expert-QCed dataset or the SOI-model-only dataset with no QC . Note that while auto QC Test 4 failed significantly fewer models than Test 1 (241 versus 672), the resulting dig list based on L1 matching for Test 4 (Figure 19 top left) performs better than Test 1 (Figure 16 top left), with all TOI found after 202 non-TOI digs (compared to 235 non-TOI digs for Test 1).

Using other metrics and/or different parameters for the decision criteria may result in better performance - further research is required. It is also necessary to investigate how the different QC approaches work with different datasets. To address this we now present results using the Beale C dataset.

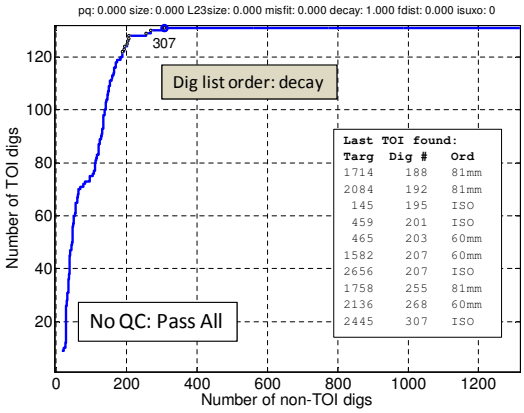


FIGURE 12. ROC curve for Beale P using No QC (no models were failed). Dig list order is based on decay of the total polarizability.

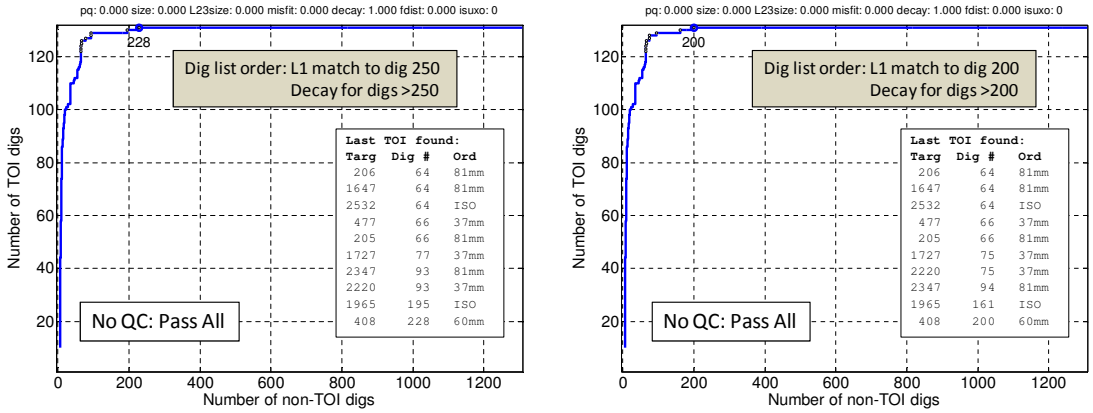


FIGURE 13. ROC curves for Beale P using No QC (no models were failed). Dig list order is based on L1 matching for early digs, then decay of total polarizability for later digs. The transition point occurs after 250 digs for the curve on the left, and after 200 digs for the curve on the right.

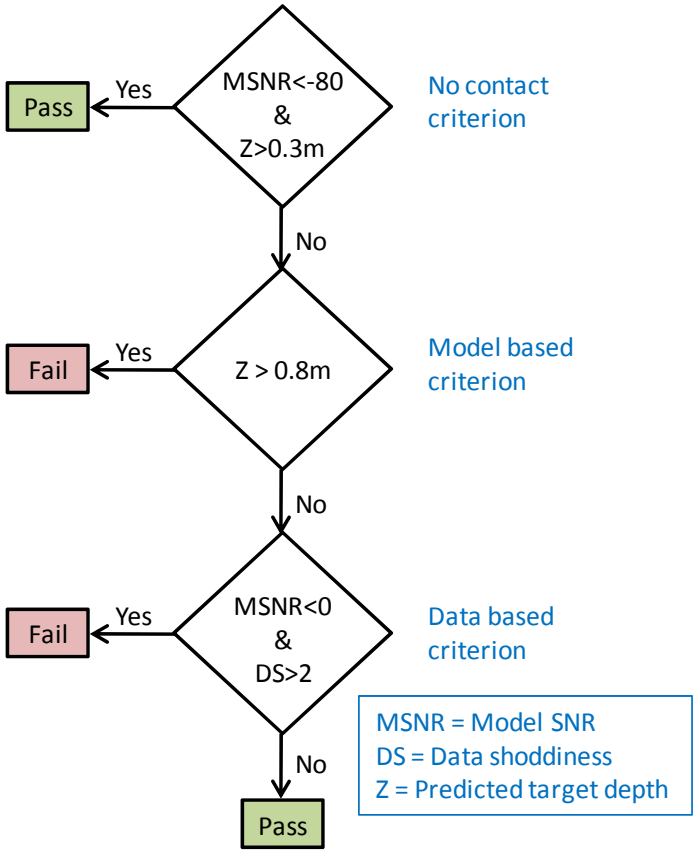


FIGURE 14. Automated QC decision (auto QC Test 1) flowchart for passing/failing models based on data and model metrics.

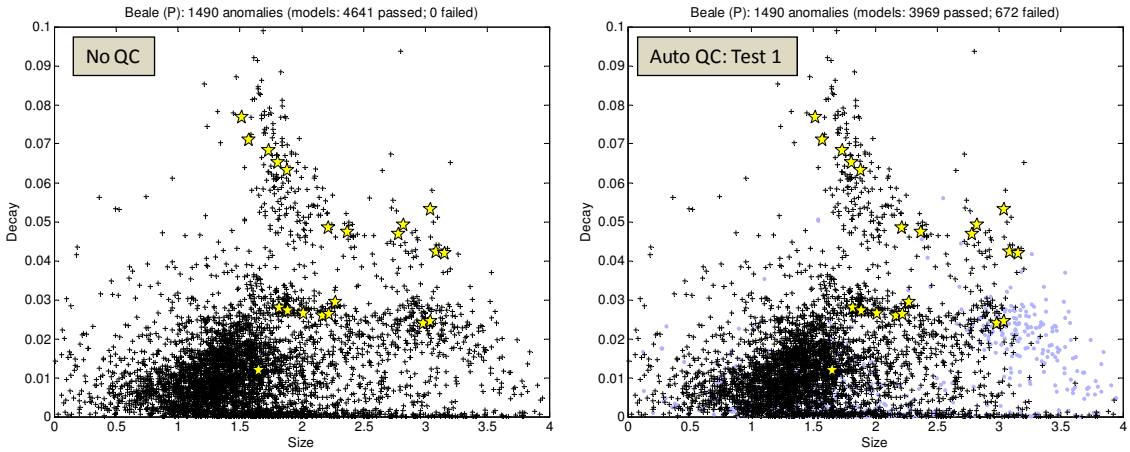


FIGURE 15. Decay versus size feature space plots for Beale P data. Size is the size of the total polarizability at the first time channel. Decay is the ratio of size of the total polarizability at channel 1 (0.106ms) to the size at channel 29 (2.006ms). Left: no QC, i.e., all models are passed. Right: auto QC Test 1. "+" symbols are passed models; blue dots are failed models. Yellow stars represent reference items. Auto QC resulted in 672 models being failed.

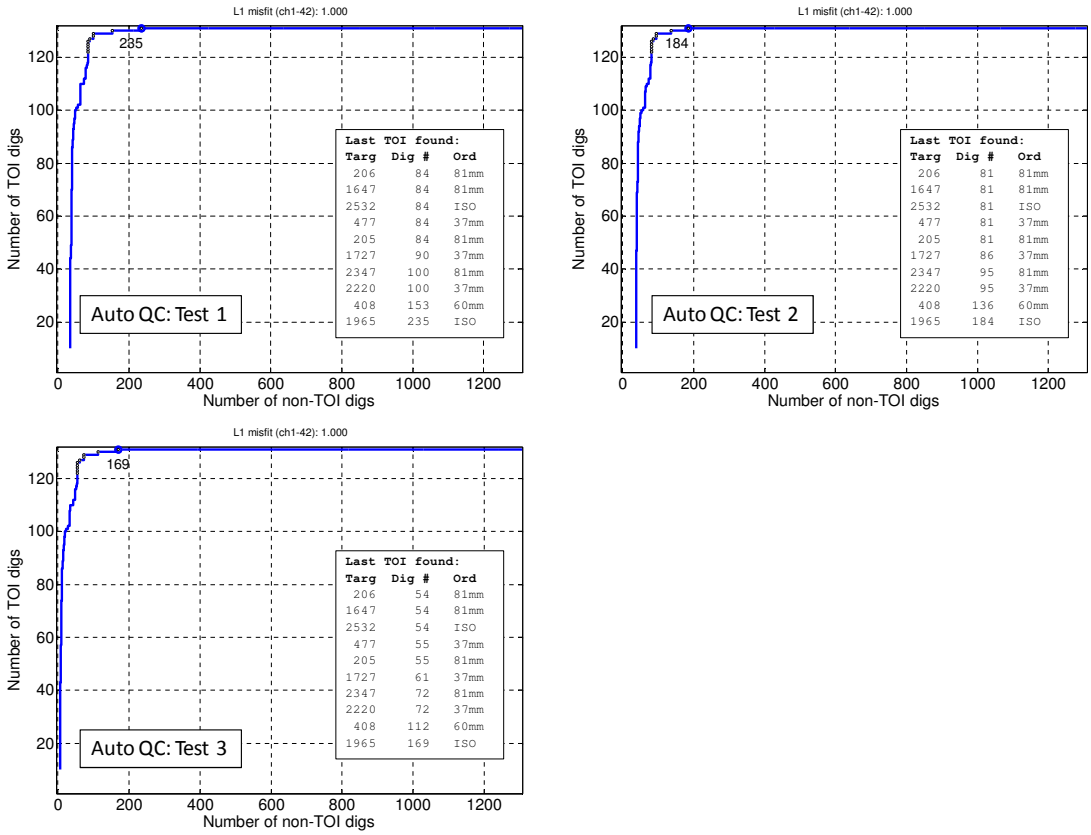


FIGURE 16. ROC curves for Beale P using Auto QC. Dig list order is based on match between primary polarizability (L1) of the predicted and best fitting reference item. Test 1 used the criteria shown in Figure 14. In Test 2 the model based criteria was changed to $Z > 0.6$m. Test 3 used the same criteria as Test 2, but the auto QC was applied only to the 2OI models; all SOI models were passed.

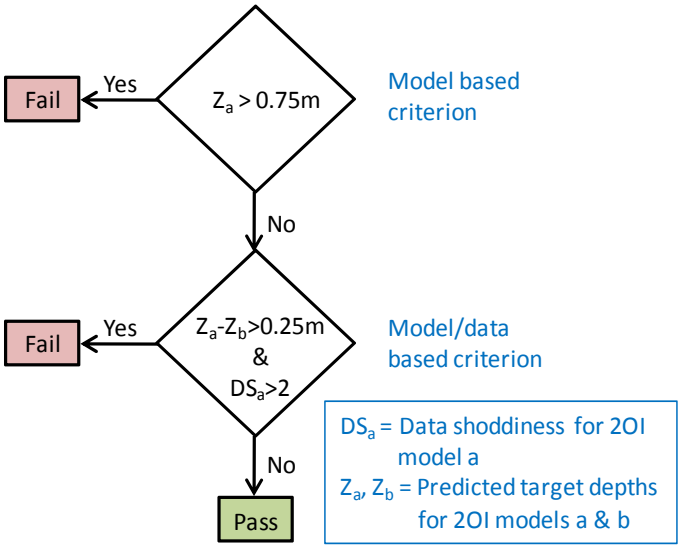


FIGURE 17. Automated QC (auto QC Test 4) decision flowchart for failing deep 2OI models. 2OI model a is failed if it is absolutely deep (model based criterion) or relatively deep in relation to 2OI model b and the data quality is low (model/data-based criterion).

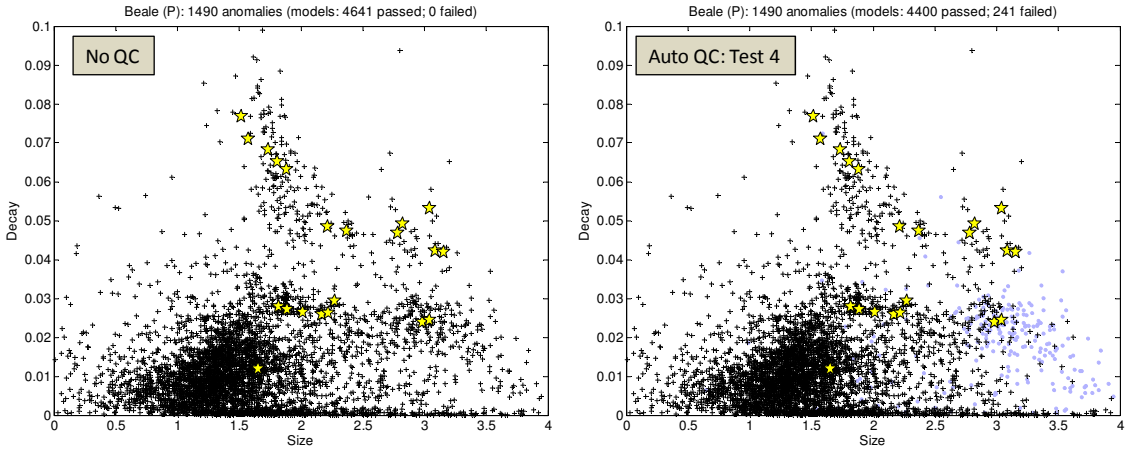


FIGURE 18. Decay versus size feature space plots for Beale P data. Left: no QC, i.e., all models are passed. Right: auto QC Test 4 to eliminate unrealistic deep 2OI models. "+" symbols are passed models; blue dots are failed models. Yellow stars represent reference items. Auto QC resulted in 241 models being failed.

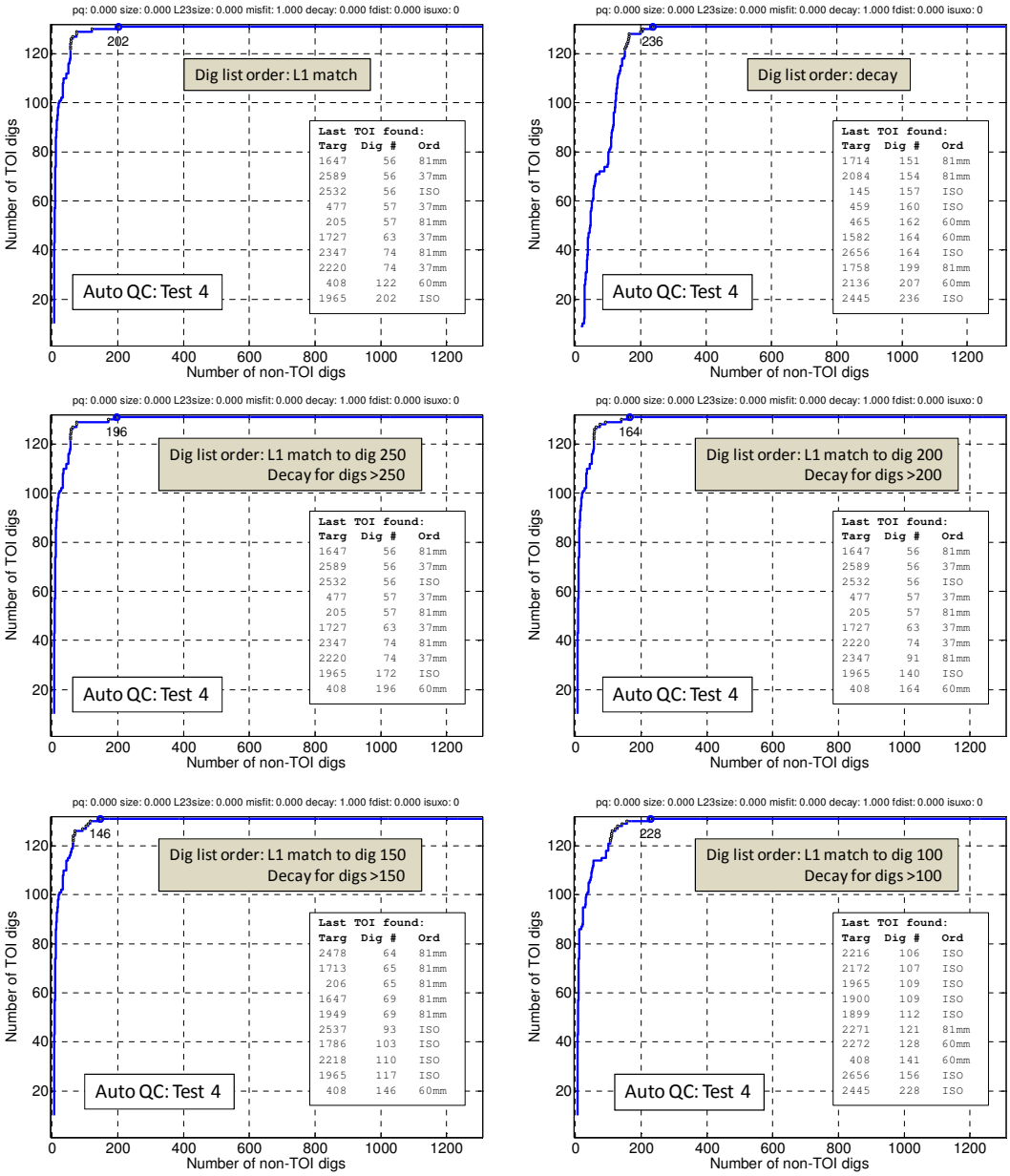


FIGURE 19. ROC curves for Beale P using Auto QC Test 4 to eliminate unrealistic deep 2OI models. The auto QC decision process is shown in Figure 17.

Test Set 2: Beale MetalMapper C. A decay versus size feature space plot for the expert-QCed Beale C data, including ground truth information, is shown in Figure 20. The feature space plot shows the separation of TOI from non-TOI items is similar to the Beale P dataset (Figure 6), but there are a few challenging TOI that are quite distant from their expected location in feature space (e.g., anomalies 1951 and 2091). The expert visual QC resulted in similar number of model failures (approximately two-thirds of the models).

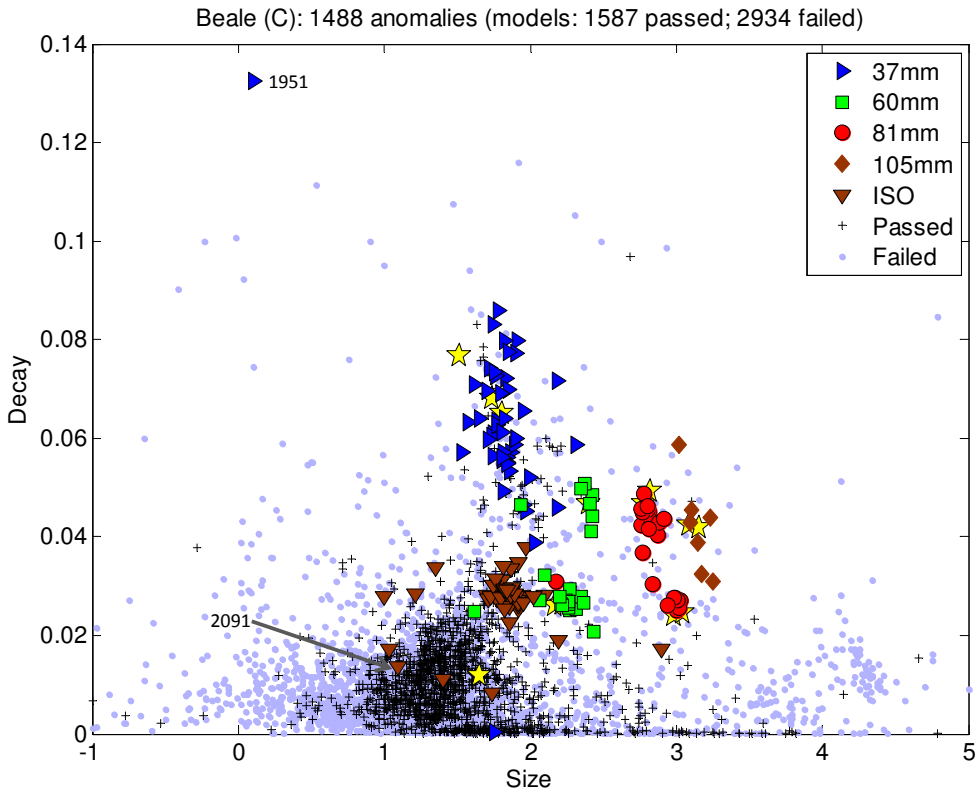


FIGURE 20. Decay versus size feature space plot for Beale C data showing all passed ("+") and failed (blue dot) models as determined by visual QC performed by an expert analyst (expert QC). Yellow stars represent reference items. Other large symbols represent TOI for passed models. Passed models indicated by "+" are non-TOI. The two most difficult items (anomalies 1951 and 2091) are identified.

Figure 21 shows ROC curves for two dig lists created independently by different analysts using different approaches. Both dig lists were based on a dataset that had undergone the same visual QC by an expert analyst. One of the dig lists used a multi-stage approach featuring matching all three polarizabilities for early digs, matching the primary polarizability for later digs, and decay for still later digs. This list missed one TOI; all TOI were found after 513 non-TOI digs. The second dig list used a Support Vector Machine (SVM) two stage discrimination strategy with early digs trained on all polarizabilities and later digs trained

on total polarizability. This list missed two TOI; all TOI were found after 764 non-TOI digs. The former represents our best result for the Beale C dataset and can be considered as the baseline for comparisons with the tests presented below.

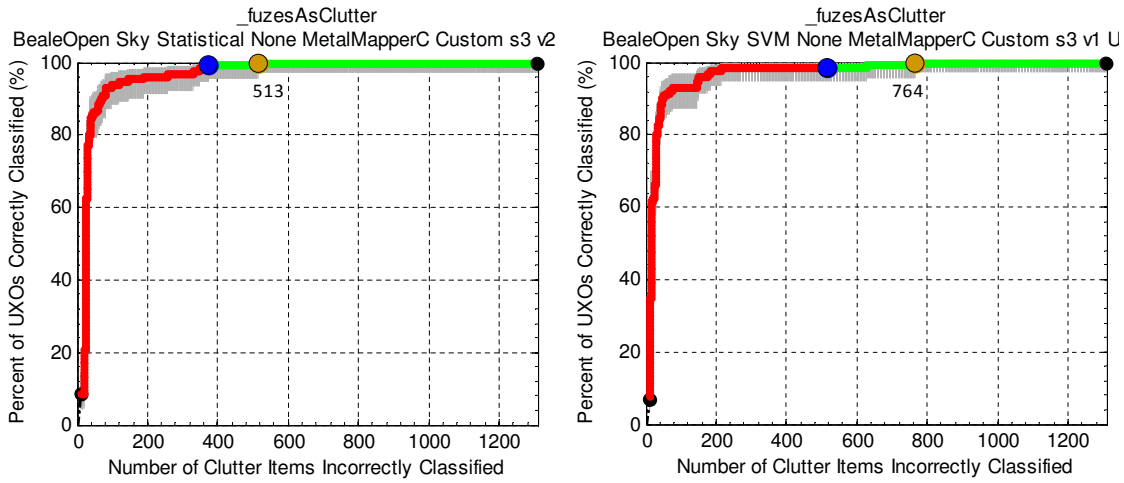


FIGURE 21. Official scoring for Beale C using Expert-QCed data. Dig list order for the ROC curve on the left was based on a multi-stage approach featuring matching all three polarizabilities for early digs, matching the primary polarizability for later digs, and decay for still later digs. The ROC curve on the right was constructed using a Support Vector Machine (SVM) two stage discrimination strategy with early digs trained on all polarizabilities and later digs trained on total polarizability (L1+L2+L3). Blue dot denotes stop dig point. Yellowish dot denotes point at which all TOI are found. The multi-stage approach missed one TOI; the final TOI was found after 513 non-TOI digs. The SVM approach missed two TOI; the final TOI was found after 764 non-TOI digs.

Figure 22 shows the ROC curve for a dig list derived from expert-QCed data using matching on the primary polarizability (L1) to determine dig order. As with the Beale P data we also created dig lists using matching on all polarizabilities and matching using only the first 30 time channels; however, the dig lists based on L1 match using all time channels consistently performed best. In the ROC curves for many of the Beale C dig lists for which we present results, anomalies 2091 and 1951 (Figure 23) occur very late in the list. The

recovered polarizabilities for these anomalies bear no resemblance to the reference polarizabilities, so any dig list based solely on polarizability matching will have these items very late in the list. For judging the performance of the different QC approaches based solely on polarizability matching with these data, it is best to ignore these two anomalies. In so doing, the dig list based on expert-QCed data finds all other TOI after 118 non-TOI digs.

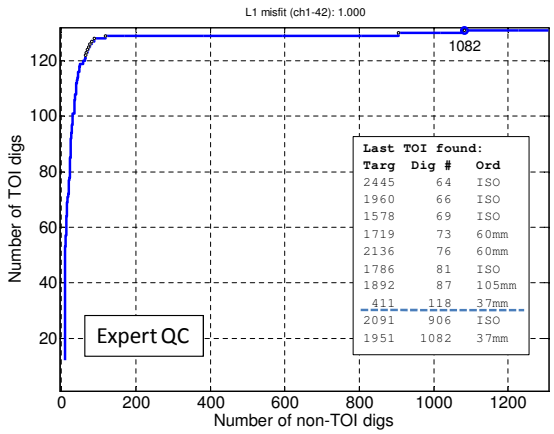


FIGURE 22. ROC curve for Beale C using Expert-QCed data. Dig list order is based on match between primary polarizability (L1) of the predicted and best fitting reference item. Dashed blue line on the inset table marks point at which all TOI except difficult anomalies 2091 and 1951 (Figure 23) are found.

Figure 24 shows the ROC curves based on L1 matching for data with no QC. Using both SOI and 2OI models results in all TOI (except anomalies 2901 and 1951) being found after 102 non-TOI digs. Unlike with the Beale P data, using only the SOI model for each anomaly does not provide better performance: all TOI (except anomalies 2901 and 1951) are found after 129 non-TOI digs. This suggests that with the Beale C dataset there are relatively fewer scrap items with 2OI models that provide a good L1 match to a reference item. In addition, for some of the TOI, one of the 2OI models provides a significantly better polarizability match than the SOI model. The performance of these two dig lists

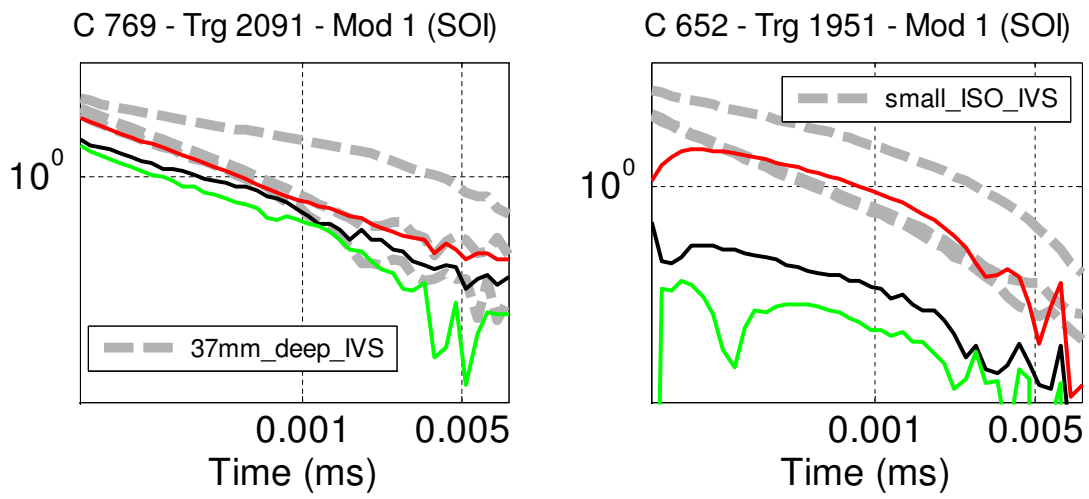


FIGURE 23. Predicted (red, black green lines) and best fitting reference polarizabilities (broken grey lines) for the two most difficult TOI of the Beale C dataset. Anomaly 2091 (left) is an ISO at 10cm depth; anomaly 1951 (right) is a 37mm projectile at 11cm depth. All three polarizabilities for both of these anomalies are so poorly recovered that any dig list based on polarizability matching alone will have these anomalies late in the list.

are only slightly better/worse than the performance obtained with the expert-QCed dig list (Figure 22), respectively.

In Figure 25 we show ROC curves for dig lists based solely and partly on decay with no QC. Note that these lists do significantly better at finding all TOI than the ones based only on polarizability matching. In particularly a strategy of switching from matching L1 polarizability to decay after 250 digs finds all TOI after 216 non-TOI digs. This is significantly better than the submitted dig list which used expert-QCed data and employed a multi-stage classification approach (Figure 21).

For the Beale C data we tried auto QC tests based on the criteria shown in Figure 14, but with the model-based criteria changed to $Z > 0.6m$. In addition to anomalies 2091 and 1951, anomaly 1786 also appears late in the dig list (Figure 26; left). Test 2b (Figure 26; right) used the same criteria as Test 2, but the cutoff for MSNR in the data-based criterion

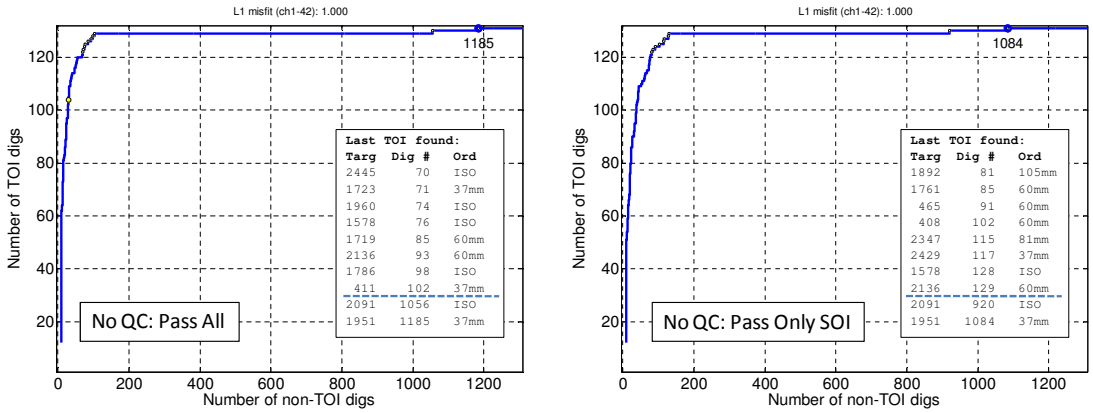


FIGURE 24. ROC curves for Beale C using No QC (no models were failed). Dig list order is based on match between primary polarizability (L1) of the predicted and best fitting reference item. For the ROC curve on the left both SOI and 2OI models were used; the ROC curve on the right used only the SOI model for each anomaly.

was increased to 25. The results of both of these auto QC tests are inferior to the no QC and expert QC results.

Results of using the auto QC process shown in Figure 17 to eliminate unrealistic deep 2OI models are presented in Figure 27. Dig lists based on entirely or partly on decay perform marginally better than the same dig lists using data with no QC (Figure 25). The dig list based on L1 match for early digs and decay for later digs finds all TOI after 202 non-TOI digs. This is the best performance of all of the Beale C dig lists.

The Beale C dataset is slightly more challenging than the Beale P dataset. The recovered polarizabilities for a few of the TOI are not of sufficient quality to support a dig list based only on polarizability matching. However, even with no QC, a simple dig list that is based on L1 polarizability match for early digs and decay for later digs performs very well. Auto QCing to eliminate some of the unrealistic, deep 2OI models gives a marginal increase in performance.

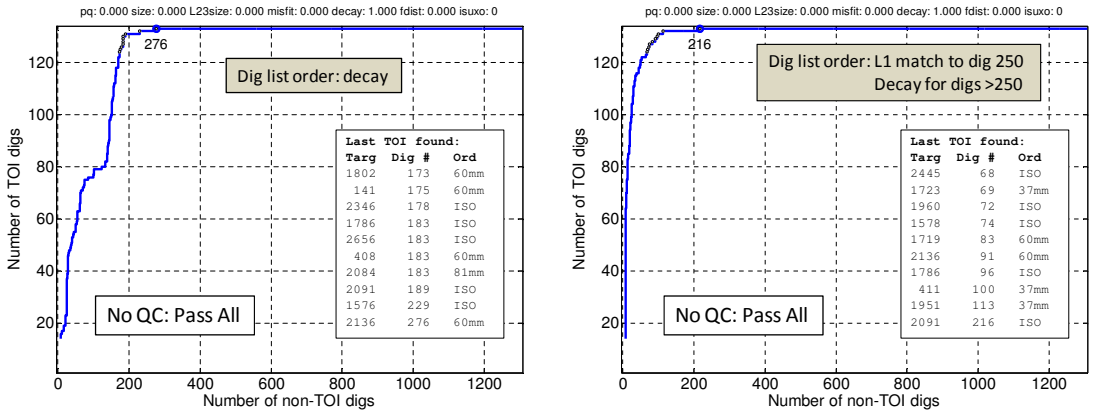


FIGURE 25. ROC curves for Beale C using No QC (no models were failed). For curve on left, dig list order is based on decay of total polarizability. For curve on right, dig list order is based on L1 matching for early digs (1-250), then decay of total polarizability for later digs.

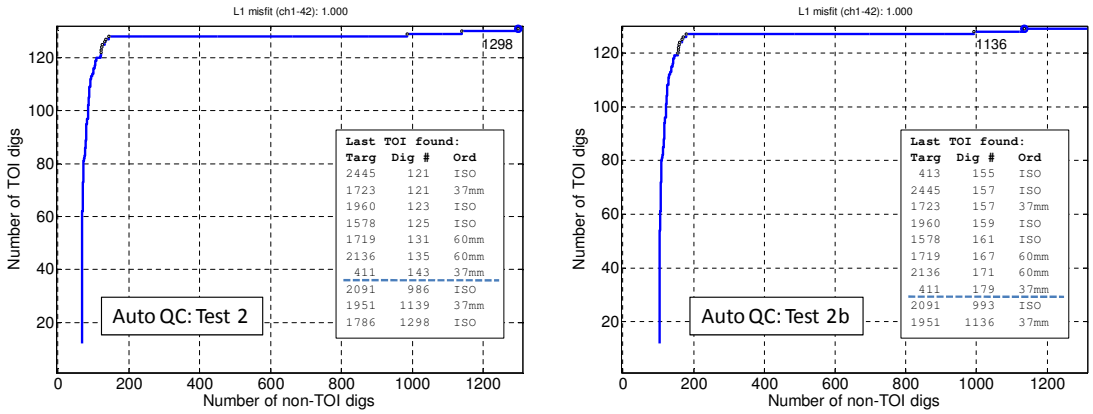


FIGURE 26. ROC curves for Beale C using Auto QC. Dig list order is based on match between primary polarizability (L1) of the predicted and best fitting reference item. Test 2 used the criteria shown in Figure 14, but with the model-based criteria changed to $Z > 0.6$m. Test 2b used the same criteria as Test 2, but cutoff for MSNR in the data-based criterion (Figure 14) was changed to $MSNR < 25$.

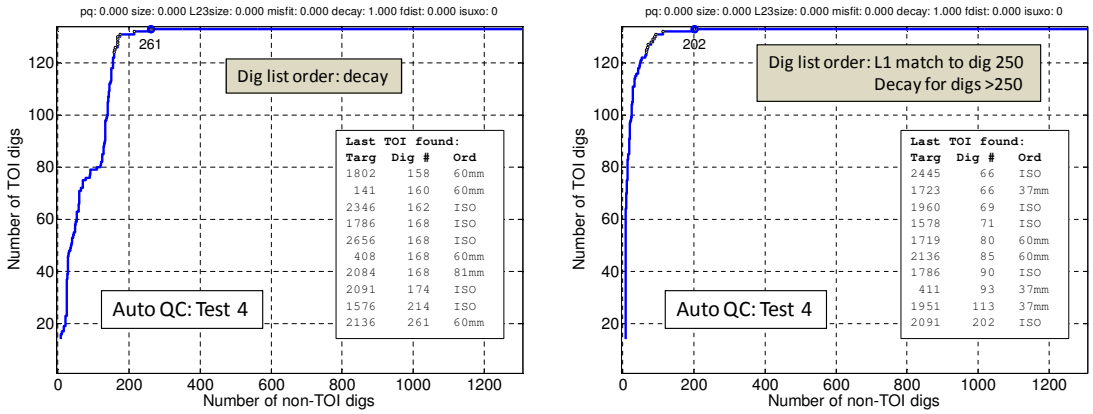


FIGURE 27. ROC curves for Beale C using Auto QC Test 4 to eliminate unrealistic deep 2OI models. The auto QC decision process is shown in Figure 17. These results are slightly better than the equivalent lists that used no QC (Figure 25).

4.1.3. *Test Sets 3 and 4: Camp Butner and Pole Mountain.* In this section we apply the same QC tests to two other MetalMapper datasets. Relative to the Beale datasets, the Butner data present a more challenging discrimination problem. By some objective measures the Butner data are better than the Beale data, but the recovered models for TOI tend to be poorer in quality resulting in larger misfits with respect to reference items (Table 1). In contrast, the Pole Mountain dataset is of excellent quality and did not present a challenge from a discrimination point of view due.

Test Set 3: Camp Butner. The Former Camp Butner (North Carolina) cued MetalMapper dataset was collected in September 2010. Two different instruments were used to collect data. About 60 percent of the anomalies were collected with an instrument that performed noticeably worse than the other instrument, in part because some of the receiver/transmitter components tended to malfunction. A fairly large number of the anomalies (15 percent) were recollected. Total number of anomalies in the Butner dataset is 2304 with 171 of these being TOI. TOI fall into three classes: (105mm, 37mm and large M48 fuzes).

A decay versus size feature space plot for the expert-QCed Butner data, including ground truth information, is shown in Figure 25. Relative to the Beale datasets, some of the TOI (fuzes and faster decaying 37mm projectiles) overlap the main cluster of non-TOI, suggesting that classification will be more challenging. The expert visual QC resulted in a similar number of model failures (approximately two-thirds of the models).

Figure 29 shows the IDA-scored ROC curve for a dig list based on a dataset that had been visually QCed by expert analysts. The dig list was created using a Support Vector Machine (SVM) two stage discrimination strategy with early digs trained on all polarizabilities and later digs trained on total polarizability. This list performed very well but missed two TOI, with one (anomaly 1346) occurring very late in the list (after 1669 non-TOI digs).

Figure 30 shows ROC curves based on dig lists using L1 matching (left) and total decay (right) for data with no QC. Although neither of these perform well, the last TOI to be dug in both lists is found earlier than with the SVM list (Figure 26). A simple two stage dig list using L1 matching for early digs (1-500) and decay for later digs performs significantly better, with all TOI dug after 658 non-TOI digs (Figure 31).

Figures 32- 33 show ROC curves for the same three dig lists with the auto-QC Test 4 process applied to eliminate deep, unrealistic 2OI models. In all cases the performance is better. In particular the two-stage (L1 match/decay) dig list (Figure 30) performs very well with all TOI found after 500 non-TOI digs. Results obtained using the auto-QC process described in Figure 14 (not shown) were worse, for example the two-stage dig list found all TOI after 605 non-TOI digs. Figure 34 shows the decay versus size feature space plots for no QC versus auto QC Test 4. The latter has removed many (734) of the dubious 2OI models. The improvement in performance gained by using auto QC Test 4 with the Butner data is more significant compared to the Beale datasets.

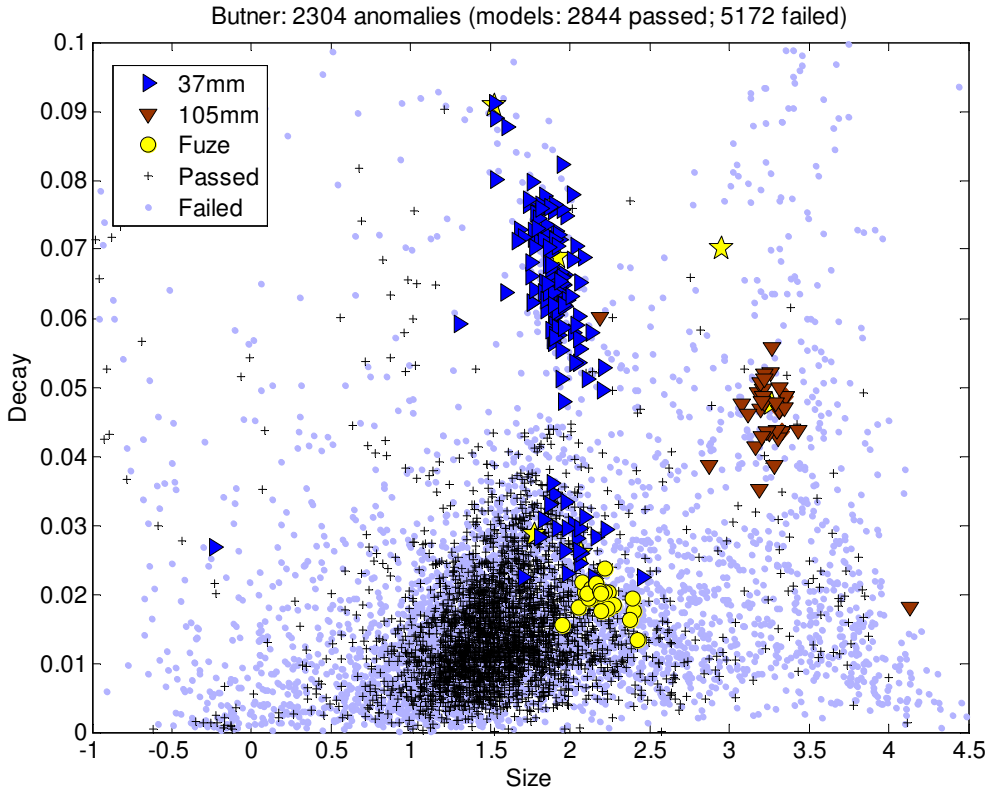


FIGURE 28. Decay versus size feature space plot for Butner data showing all passed ("+") and failed (blue dot) models as determined by visual QC performed by an expert analyst (expert QC). Yellow stars represent reference items. Other large symbols represent TOI for passed models. Passed models indicated by "+" are non-TOI.

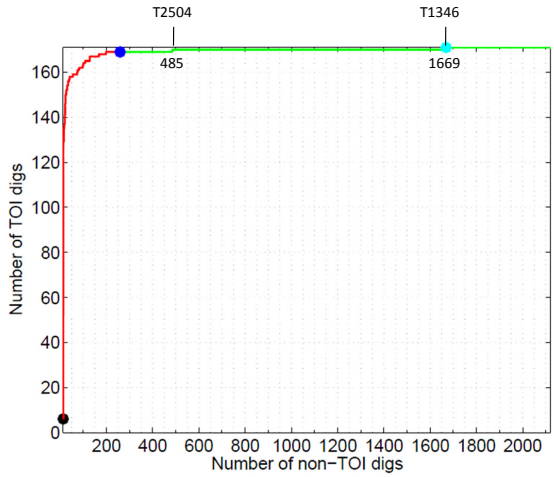


FIGURE 29. Official scoring for Butner MM using Expert-QCed data. The ROC curve is based on a dig list constructed using a Support Vector Machine (SVM) two stage discrimination strategy with early digs trained on all polarizabilities and later digs trained on total polarizability (L1+L2+L3). Blue dot denotes stop dig point. Light blue dot denotes point at which all TOI are found. The SVM approach missed two TOI (labeled at top of plot). The final TOI was found after 1669 non-TOI digs.

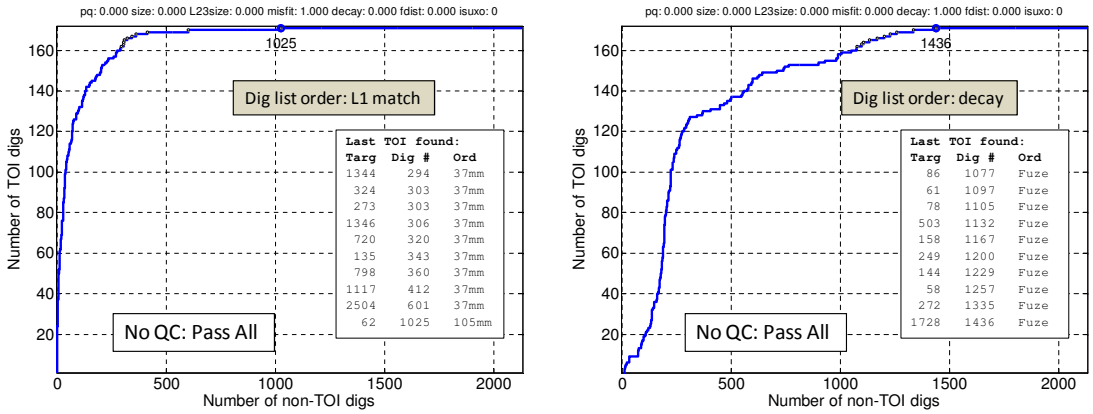


FIGURE 30. ROC curves for Butner MM using No QC (no models were failed). Dig list order for the ROC curve on the left is based on match between primary polarizability (L1) of the predicted and best fitting reference item. Dig list order for the curve on the right is based on decay of the total polarizability.

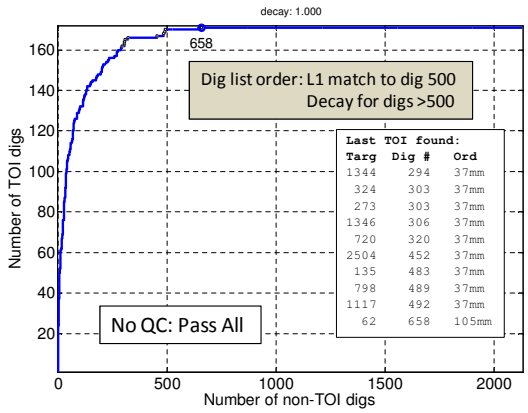


FIGURE 31. ROC curve for Butner MM using No QC (no models were failed). Dig list order is based on L1 matching for early digs (1-500), then decay of total polarizability for later digs.

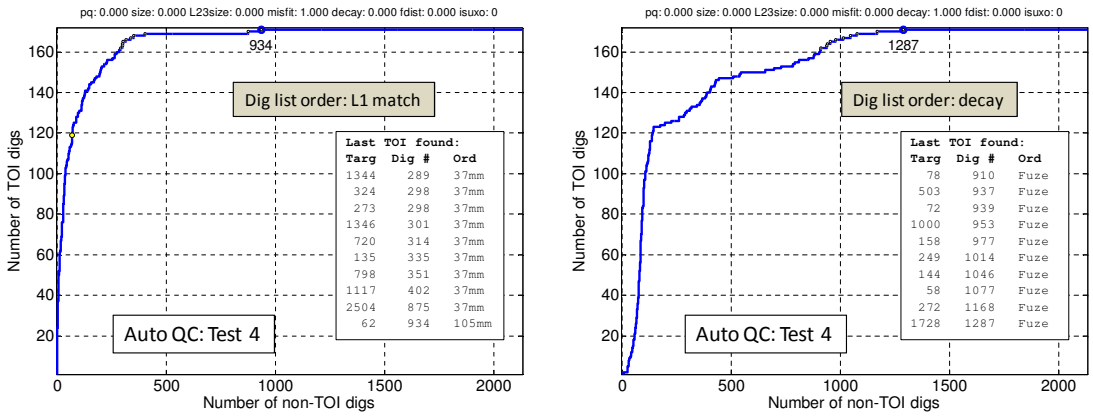


FIGURE 32. ROC curves for Butner MM using Auto QC Test 4 to eliminate unrealistic deep 2OI models. Dig list order for the ROC curve on the left is based on match between primary polarizability (L1) of the predicted and best fitting reference item. Dig list order for the curve on the right is based on decay of the total polarizability.

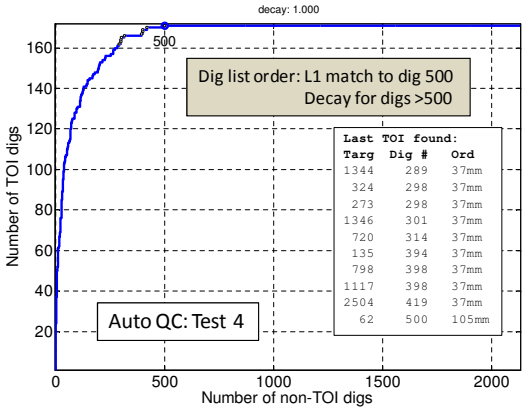


FIGURE 33. ROC curve for Butner MM using Auto QC Test 4 (Figure 17) to eliminate unrealistic, deep 2OI models. Dig list order is based on L1 matching for early digs (1-500), then decay of total polarizability for later digs.

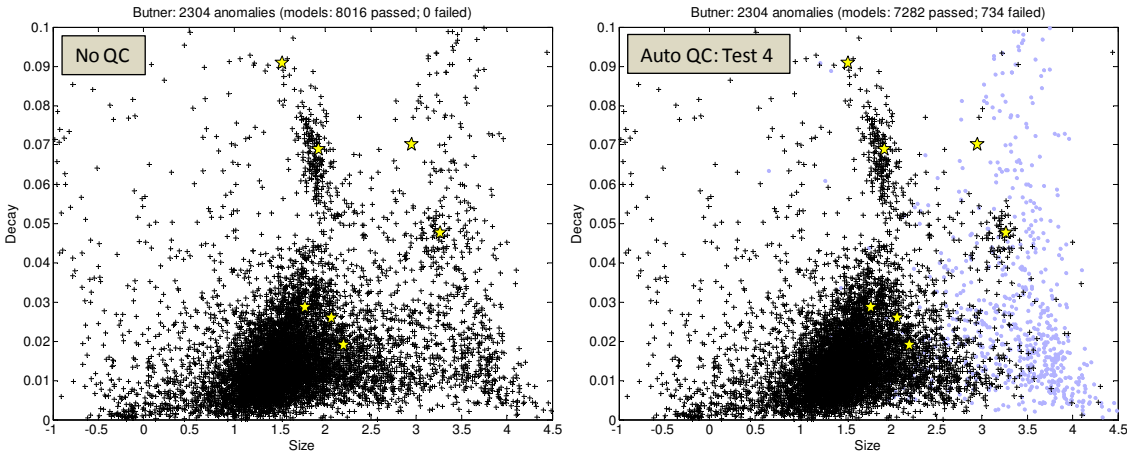


FIGURE 34. Decay versus size feature space plots for Butner data. Left: no QC, i.e., all models are passed. Right: auto QC Test 4 to eliminate unrealistic deep 2OI models. "+" symbols are passed models; blue dots are failed models. Yellow stars represent reference items. Auto QC resulted in 734 models being failed.

Test Set 4: Pole Mountain. The Pole Mountain (Wyoming) cued MetalMapper dataset was collected in July-August 2011. By any objective measure, the quality of this dataset is excellent, being superior to both the Beale and Butner datasets (Table 1). Total number of anomalies in the Pole Mountain dataset is 2370 with 160 of these being TOI. TOI fall into six classes: (Stokes mortar, 75mm, 60mm mortar, 57mm, 37mm and small ISO). In the actual analysis performed by Sky Research, this dataset was divided into two parts representing a 2 year study (thus we do not have official scoring for the combined dataset). For the purposes of this study, we use the combined dataset.

A decay versus size feature space plot for the expert-QCed Pole Mountain data, including ground truth information, is shown in Figure 35. The good separation between TOI and non-TOI and tight clustering of the TOI attest to the high quality of the dataset, and suggest that classification should not be too difficult, especially compared to the more difficult Butner dataset. The expert visual QC resulted in similar number of model failures (approximately two-thirds of the models).

Figure 36 shows the ROC curve that would be obtained with expert QCed data using the same approach taken to analyze the separate Pole Mountain years 1 and 2 datasets. The dig list order is based on a combination of polarizability matching (using all three polarizabilities), decay, size, and polarizability quality. All TOI are found after 80 non-TOI digs. The excellent performance of this list using a simple discrimination approach reflects the high quality of the dataset. The results of using the same procedure to develop the dig list, but using data that have not been QCed, is shown in Figure 37. This list performs slightly better, with all TOI found after 67 non-TOI digs, suggesting that for this very high quality dataset, QC is not necessary. Again using the same procedure to develop the dig list, but using auto-QC Test 4 (Figure 17) to remove deep, unrealistic 2OI models results in no improvement (Figure 38).

Dig lists based on matching only the primary polarizability (L1), or on decay, using either no QC or auto-QC all perform significantly worse than the results presented above, with the last TOI being found after 300 non-TOI digs for lists based on L1 matching, and after 650 digs for lists based on decay (not shown). Similarly, dig lists based on a two stage approach, with matching on polarizabilities for early digs and decay for later digs, do not perform as well as the results presented in Figures 36- 38.

Dig lists based on matching all three polarizabilities with no QC, or with auto QC to remove deep, unrealistic 2OI models perform reasonably well (Figure 39), but not, however, as well as the dig lists based on a slightly more sophisticated and aggressive approach to anomaly ranking (Figures 36- 38). This shows that for high quality data, dig lists based on matching all three polarizabilities out-performs matching on L1 only, and an approach which uses more features of the data can out-perform matching on all three polarizabilities.

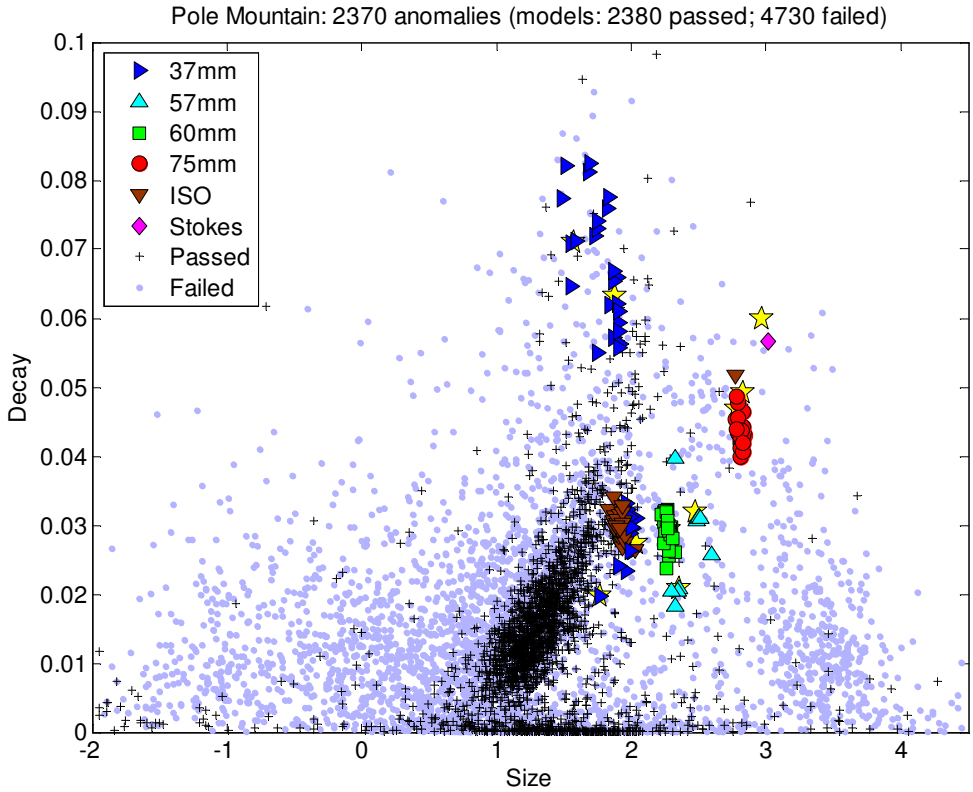


FIGURE 35. Decay versus size feature space plot for Pole Mountain data showing all passed ("+") and failed (blue dot) models as determined by visual QC performed by an expert analyst (expert QC). Yellow stars represent reference items. Other large symbols represent TOI for passed models. Passed models indicated by "+" are non-TOI.

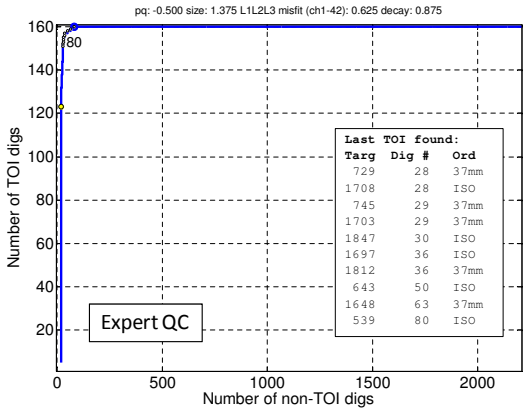


FIGURE 36. ROC curve that would be obtained with expert QCed data using the same approach taken to analyze the separate Pole Mountain years 1 and 2 datasets. Dig list order is based on a combination of polarizability matching (using all three polarizabilities), decay, size, and polarizability quality. All TOI are found after 80 non-TOI digs.

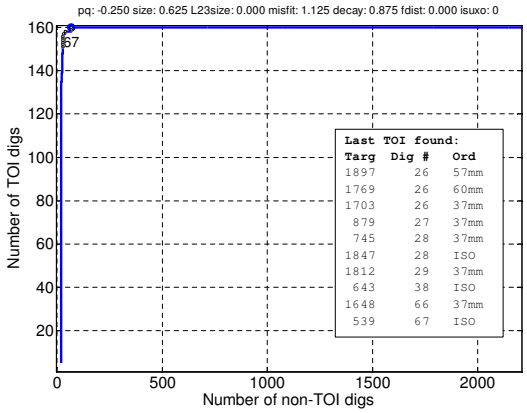


FIGURE 37. ROC curve for Pole Mountain MM using no QC. Dig list order is based on a combination of polarizability matching (using all three polarizabilities), decay, size, and polarizability quality. All TOI are found after 67 non-TOI digs. Performance is slightly better than using expert-QCed data (Figure 36).

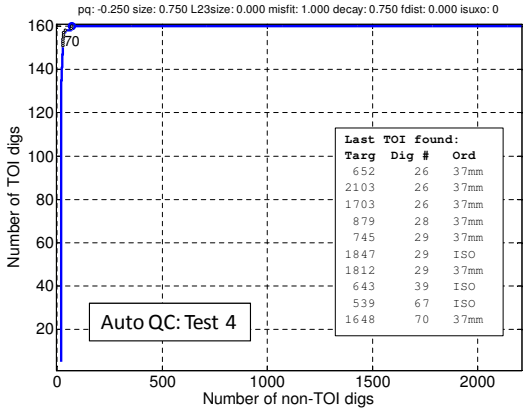


FIGURE 38. ROC curve for Pole Mountain MM using auto QC Test 4 to eliminate unrealistic, deep 2OI models. Dig list order is based on a combination of polarizability matching (using all three polarizabilities), decay, size, and polarizability quality. Performance is no better than using no QC.

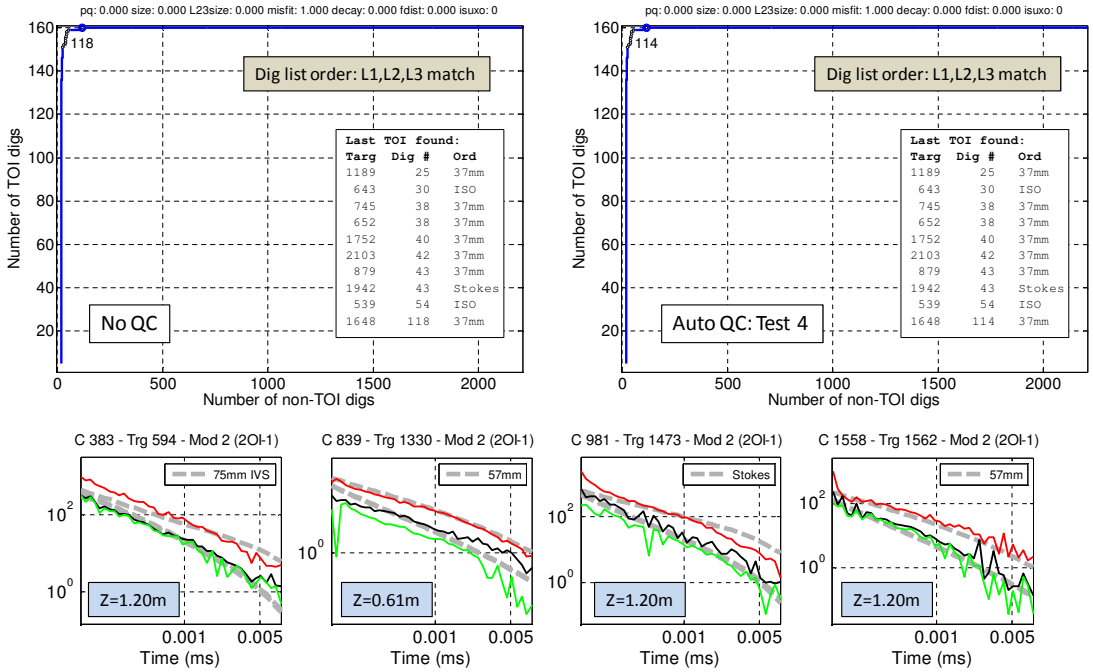


FIGURE 39. Top: ROC curves Pole Mountain MM using no QC (left) and auto QC to eliminate unrealistic, deep 2OI models. Dig list order is based on match to all three polarizabilities. Note that auto-QC for the Pole Mountain data provides minimal improvement. In this case four deep anomalies (plots at bottom), all corresponding to scrap and all fitting reference items reasonably well, have been removed from the early part of the dig list.

4.1.4. *Summary.* A summary of test results is presented in Table 4.1.4. For all four datasets, good results can be obtained with either no QCing, or using auto QC Test 4 to eliminate some of the unrealistic, deep 2OI models. For the two datasets with the lowest quality data (Butner and Beale C), using auto QC and a two stage approach for dig list order (with L1 matching for early digs, and decay for later digs), provides the best results. This approach also performed well with the Beale P dataset, although the best results were obtained using either expert-QCed data or no QC and a dig list based on L1 matching. For the high quality Pole Mountain dataset, excellent results were obtained using either the expert QCed data, no QC or auto QC, and a dig list based on more features of the data. The absolute best result, however, was obtained using no QC. This dataset is so excellent that a variety of approaches for dig list construction would likely work very well.

Dataset	QC method	Dig list order	Non-TOI digs at last TOI	Note
Beale P	Expert	L1,L2,L3 match decay, size, quality	595	Offically scored result
Beale P	Expert	SVM: (1) L1,L2,L3 match; (2) L1 match	264	Offically scored result
Beale P	**Expert**	**L1 match**	**124**	*** Best for Beale P**
Beale P	No QC	L1 match	268	
Beale P	**No QC (SOI only)**	**L1 match**	**126**	*** Best for Beale P**
Beale P	No QC	Decay	307	
Beale P	No QC	(1) L1 match digs 1-200; (2) Decay digs >200	200	
Beale P	Auto QC Test 1	L1 match	235	
Beale P	Auto QC Test 2	L1 match	184	
Beale P	Auto QC Test 3	L1 match	169	
Beale P	Auto QC Test 4	(1) L1 match digs 1-150; (2) Decay digs >150	146	
Beale C	Expert	(1) L1,L2,L3 match: (2) L1 match; (3) Decay	513	Offically scored result
Beale C	Expert	SVM: (1) L1,L2,L3 match; (2) L1 match	764	Offically scored result
Beale C	Expert	L1 match	1082	
Beale C	No QC	L1 match	1185	
Beale C	No QC (SOI only)	L1 match	1084	
Beale C	No QC	Decay	276	
Beale C	No QC	(1) L1 match digs 1-250; (2) Decay digs >250	216	
Beale C	Auto QC Test 2	L1 match	1298	
Beale C	Auto QC Test 2b	L1 match	1136	
Beale C	Auto QC Test 4	Decay	261	
Beale C	**Auto QC Test 4**	**(1) L1 match digs 1-250; (2) Decay digs >250**	**202**	*** Best for Beale C**
Butner	Expert	SVM: (1) L1,L2,L3 match; (2) L1 match	1669	Offically scored result
Butner	No QC	L1 match	1025	
Butner	No QC	Decay	1436	
Butner	No QC	(1) L1 match digs 1-500; (2) Decay digs >500	658	
Butner	Auto QC Test 4	L1 match	934	
Butner	Auto QC Test 4	L1 match	1287	
Butner	**Auto QC Test 4**	**(1) L1 match digs 1-500; (2) Decay digs >500**	**500**	*** Best for Butner**
Pole Mtn	Expert	L1,L2,L3 match decay, size, quality	80	Equiv. to officially scored result
Pole Mtn	**No QC**	**L1,L2,L3 match decay, size, quality**	**67**	*** Best for Pole Mtn**
Pole Mtn	Auto QC Test 4	L1,L2,L3 match decay, size, quality	70	
Pole Mtn	Expert	L1,L2,L3 match	103	
Pole Mtn	No QC	L1,L2,L3 match	118	
Pole Mtn	Auto QC Test 4	L1,L2,L3 match	114	

TABLE 2. Summary of test results for all datasets using different methods for QCing and dig list ranking. Highlighted lines correspond to the best result for each dataset.

4.2. **Development and testing of active learning algorithms using Sky/UBC features.** In this section we show applications of active learning algorithms to MetalMapper data sets from ESTCP demonstrations conducted at Camp Butner and Camp Beale.

4.2.1. *Application to Camp Butner MetalMapper data.* As a first test of Duke active learning algorithms using Sky/UBC features, we consider a two-dimensional decay versus size feature space extracted from ESTCP Camp Butner MetalMapper data. Figure 40 shows the distributions of TOI and non-TOI in this feature space. In figure 40 the decay parameter is

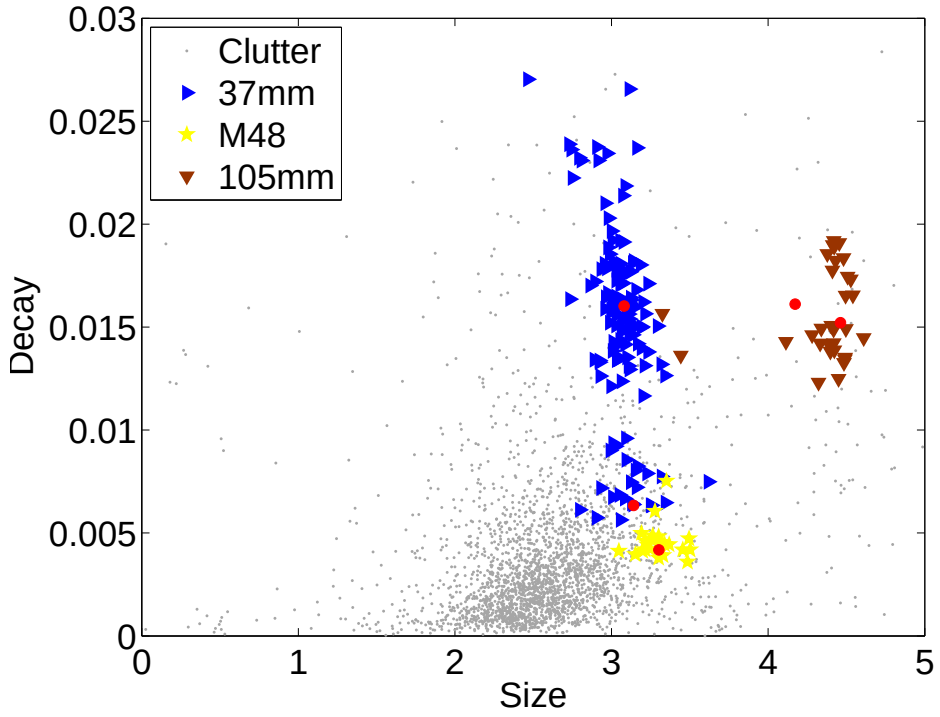


FIGURE 40. Camp Butner MetalMapper size decay features. Red circles are test pit TOI.

computed at MetalMapper channels 1 (0.1 ms) and 36 (4.2 ms).

In this experiment we randomly seed the myopic and submodular learning algorithms with two labelled items, then dig in batches of ten targets until 50 items are labelled. We then train a semi-supervised classifier using the labelled training data and remaining unlabelled test data. Figure 41 shows the performance of myopic and submodular learning approaches for a single trial (i.e. for an initial realization of two randomly-selected training items). We find that both learning algorithms are quite slow when applied to the full feature data set (approximately 3000 feature vectors) and so for each realization we downsample the test data by randomly selecting a subset of 800 clutter items. We retain all 171 TOI for every realization.

The myopic algorithm tends to select redundant items for labelling, resulting in clusters of labelled feature vectors and limited information from the region of overlap between TOI and non-TOI classes. However, even in the worst case realization (top row of figure 41),

the myopic algorithm ROC is not dramatically worse than the submodular result. This is likely because the semi-supervised classifier exploits the unlabelled test data and so is less sensitive to the particular realization of training data.

The submodular algorithm produces a much more sensible distribution of labelled feature vectors, and in the best case example shown in the top row of figure 41 the algorithm does produce an improvement in both false alarm rate (FAR) and area under the curve (AUC). Conversely, the worst case ROC for the submodular algorithm (bottom row of figure 41) is not significantly different from the corresponding myopic ROC, suggesting that the former is robust to an unfavorable initial seeding of training data.

For the submodular algorithm there is a reasonable exploration of the region bordered by the smallest TOI (37 mm). It perhaps focuses too much effort on large, slow decaying targets that are obviously TOI (105 mm) and on small, fast decaying items. In the case of Camp Butner this latter category of targets can safely be assumed to be non-TOI and so we might not need to dig these items. Instead, we can provide a subset of fast-decaying test targets as *assumed* non-TOI. While this is a viable approach for Camp Butner, it does risk mislabelling smaller TOI that might be hidden in the "cloud" of clutter. For example, at the recent Camp Beale demonstration fuzes and fuze parts similar in size to small clutter were encountered. Querying small, fast decaying targets, as in figure 41 is therefore a prudent practice, provided the labelling algorithm has the ability to find concealed clusters of small TOI.

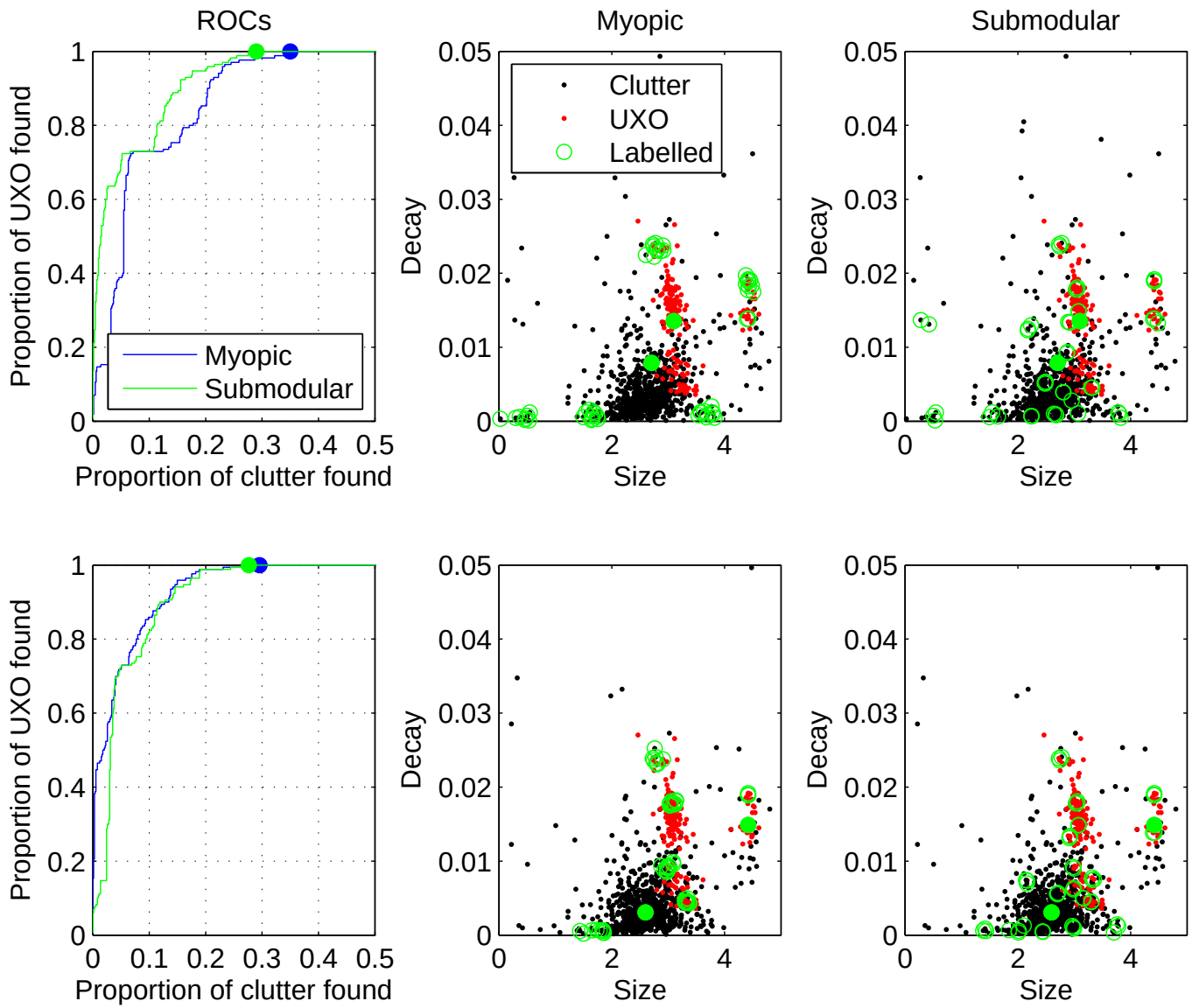


FIGURE 41. Comparison of myopic and submodular learning performance applied to Camp Butner MetalMapper size-decay features. Top row (l to r): ROCs for realization with maximum improvement in AUC for submodular algorithm relative to myopic algorithm, selected training data for myopic algorithm, selected training data for submodular algorithm. Bottom row: as above, but showing the realization with the maximum improvement in AUC for myopic algorithm relative to submodular algorithm. In feature plots solid green circles indicate the initial labelled training data.

The statistics for 50 realizations of Camp Butner test and training data are summarized in figure 42. We see that the myopic algorithm is more susceptible to producing large outlying false alarm rates (or small AUC), while the submodular algorithm has only one realization that produces an outlying AUC. Preventing outlying TOI is crucial to successful UXO discrimination, and from this experiment we can conclude that the submodular active learning algorithm will be less susceptible to false negatives than the myopic approach.

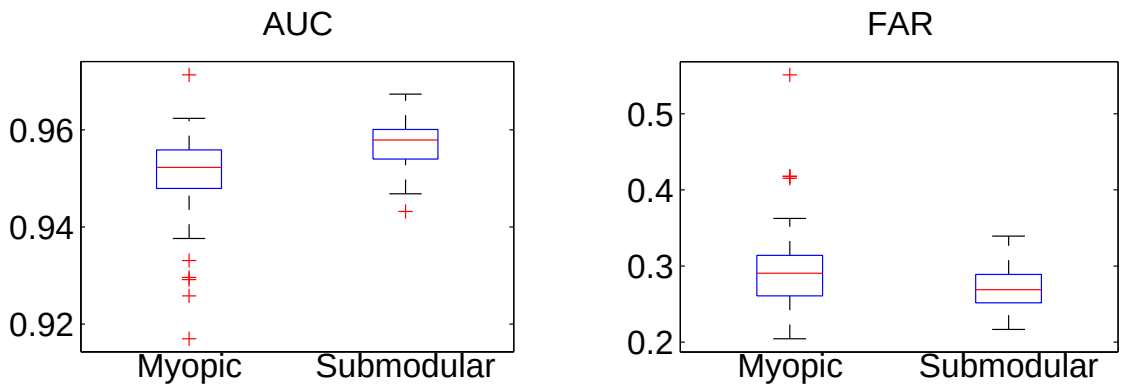


FIGURE 42. Boxplots summarizing AUC and FAR performance statistics for myopic and submodular learning algorithms applied to Camp Butner MetalMapper test data. Central mark indicates the median, the edges of each box are the 25th and 75th percentiles and whiskers extend to the most extreme data points not considered outliers. Outliers are shown as red crosses.

How well do active learning algorithms perform relative to conventional classification with limited training data? Performance comparisons with randomly selected training data sets seem somewhat biased in favor of active learning: the rarity of UXO at most sites means that a random sample is unlikely to produce an adequate sample of TOI features for training. Even in the absence of any initial groundtruth, obvious clusters of target features are often evident in the test data and this clustering can be used to guide target sampling when building the training data set. Furthermore, at most sites testpit measurements of known munitions classes provide useful information about the distributions of TOI features. In figure 43 we compare the performance of active learning algorithms with a support vector machine (SVM) classifier trained only using 5 feature vectors estimated from TOI testpit measurements. Binary decision rules always require features from both classes (TOI and non-TOI). However, rather than directly sampling from the non-TOI class, we assume that small, fast-decaying targets are clutter, without actually digging those targets during the training stage. To identify these items, we form a matrix with element M_{jk} the misfit between the jth training and k^{th} test vectors

$$M_{jk} = \sum_{i=1}^{N} \left(x^j(i) - x^k(i)\right)^2 . \tag{34}$$

In this context the feature vectors $\mathbf{x}$ are size-decay parameters (equation 4) normalized by standard deviations estimated from the test data (without this normalization the size

parameter will dominate the misfit). We then identify test feature vectors with the largest misfit relative to training vectors and use these as assumed non-TOI when training the SVM classifier.

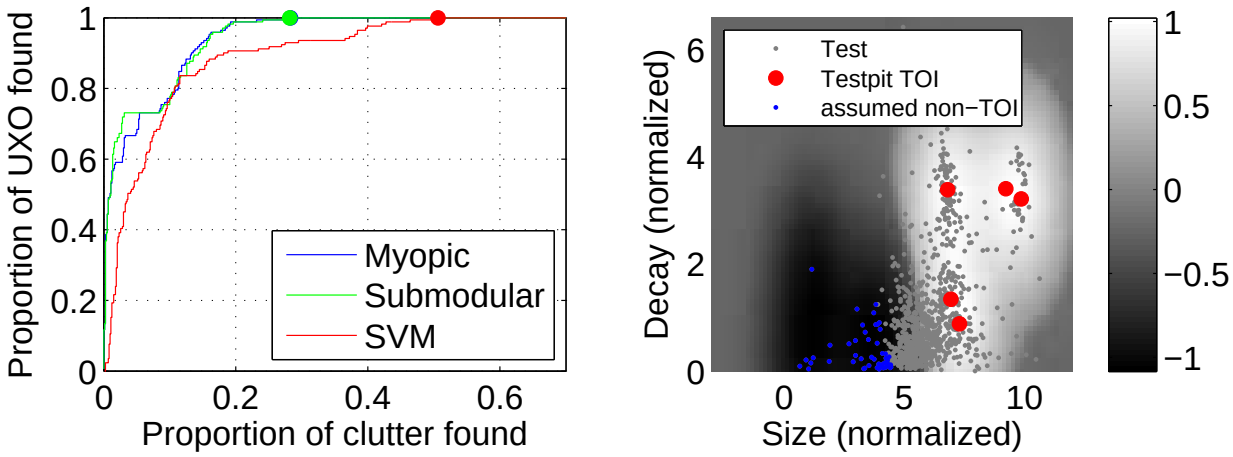


FIGURE 43. Left: comparison of active algorithm performance when seeded with testpit training items, plus an additional 50 labelled targets identified by each algorithm. Support vector machine (SVM) performance is shown for comparison. Right: support vector machine decision surface (grayscale image), with testpit items (red circles) and test data. The SVM is trained using only the 5 testpit TOI feature vectors and uses unlabelled fast-decaying test vectors as non-TOI (blue markers)

The ROC of the SVM classifier in figure 43 is a baseline for measuring the classification performance of active learning algorithms on the Camp Butner MetalMapper data. It represents the performance that is obtained using size-decay parameters and without any additional labelling of the test data. Relative to this classifier, the active learning algorithms trained with 50 additional digs significantly reduce the false alarm rate and increase the AUC. In 43, both active learning algorithms are initially seeded with the 5 TOI testpit feature vectors. This has no significant effect on the resulting ROCs relative to the simulations in shown in figure 41: both algorithms can identify TOI clusters automatically and do not require additional testpit information to succeed on these data.

4.2.2. *Active learning with the SVM.* To further validate the performance of Duke active learning algorithms on ESTCP demonstration data, we develop and apply an intuitive approach to active learning using the support vector machine. The SVM formulation assumes that the optimal decision function f_{SVM} is a weighted sum of support vectors defining the maximum extents of TOI and non-TOI classes. To achieve good discrimination performance with this algorithm we must therefore query test feature vectors in the region of overlap between the two classes, i.e. close to the decision boundary $f_{SVM} = 0$. Active learning with the SVM can then proceed as follows:

(1) Train SVM algorithm with labelled training data.
(2) Label n_{dig} test feature vectors closest to the SVM decision boundary.

(3) Append newly labelled items to training set and return to (1) until n_{batch} batches of training requests have been labelled.

In the active learning stage we wish to initially query feature vectors close to known testpit TOI, and so we use a small kernel width σ when training the SVM (here $\sigma_{active} = 0.1$). Once labelling is finished, we train our final SVM classifier using a much larger kernel width ($\sigma_{final} = 1$) to avoid overfitting the training data. Figure 44 compares the resulting ROC with the previous Duke active learning results. As in figure 43, all algorithms have only testpit TOI as initial training data, and an additional 50 test items are labelled separately by each algorithm. The SVM achieves comparable performance to the Duke algorithms, with a slight reduction in false alarm rate. As expected, many of the queried feature vectors for the SVM occupy the overlapping region between TOI and non-TOI classes. While the SVM result in this example is quite promising, we emphasize that this algorithm relies upon good initial knowledge of the TOI classes from test pit data. If there are unknown TOI clusters far from known TOI in the feature space, then the SVM active approach will likely overtrain on the known TOI, producing a large false alarm rate. In contrast, the Duke active learning algorithms are relatively insensitive to the initial training data. In the next section, we investigate the ability of these algorithms to find novel clusters of TOI within the test data.

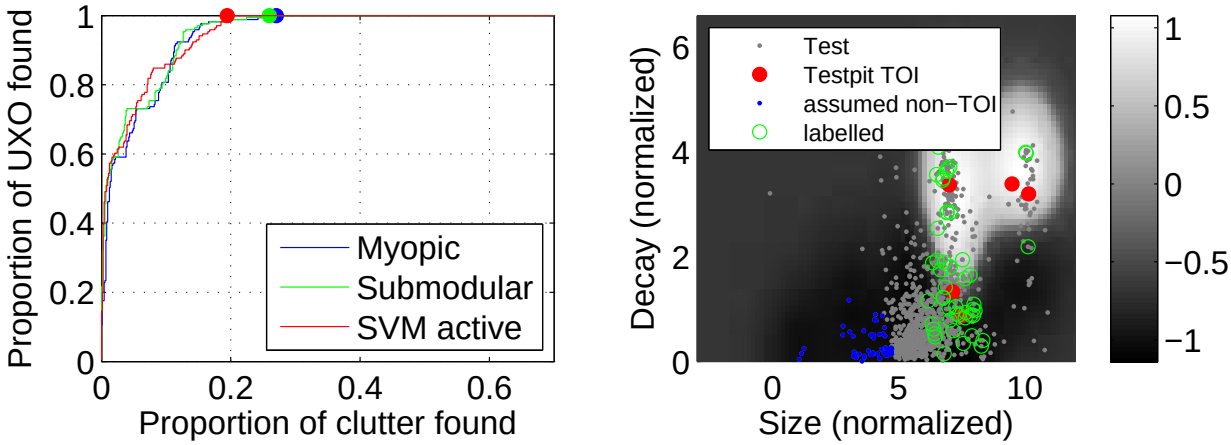


FIGURE 44. Left: Comparison of Duke active learning algorithm performance with support vector machine (SVM) active learning. Right: final support vector machine decision surface (grayscale image), as in figure 43, with additional labelled features identified by SVM active learning shown.

4.2.3. *Identification of "hidden" TOI clusters.* An important characteristic of a successful active learning algorithm is the ability to identify novel TOI classes in the test data. In the majority of ESTCP demonstrations conducted to date the TOI encountered in the field were known a priori. However, at the San Luis Obispo (SLO) and Camp Beale demonstrations unexpected TOI were encountered. If the new TOI are large (i.e > 81 mm), then they can be readily identified by their size-decay parameters as well as by features diagnostic of target

shape. Novel TOI of this type were present at SLO and were easily found by classifiers working with size-decay parameters. Camp Beale was a much more challenging scenario: fuzes and fuze parts similar in size to clutter were found.

In this section we focus on this problem in detail and test the ability of active learning algorithms to find targets of interest that are not apparent in the test data as distinct clusters.

Figure 45 shows two scenarios where a novel TOI cluster is seeded in the test data. In both cases the cluster is comprised of twenty items with nearly identical size-decay features. In the first scenario (top row of 45)) the TOI cluster is squarely within the cloud of clutter items. The submodular algorithm succeeds in finding this cluster in the training stage, while the myopic algorithm fails. Interestingly, the resulting ROC curves are not significantly different for the two algorithms. This is likely because the final semi-supervised classifier uses the test data to generate the decision function, so that a labelled UXO embedded within clutter will be bumped down the dig list. When discrete clusters of small TOI occur, it may therefore be appropriate to initially train a classifier that overfits the training data (i.e. with very small kernel widths). This will ensure that initial digging efforts focus on known, training UXO. We can then revert to a classifier with larger kernels to achieve good generalization to the test data.

In the second scenario in figure 45 (bottom row), we introduce an even smaller, faster decaying TOI cluster that lies near the edge of the clutter distribution. Both active learning algorithms find this cluster, but only the submodular algorithm results in an acceptable ROC curve. The training data generated from the myopic learning seemingly leads the semi-supervised classifier to ignore the seeded cluster.

From this experiment we conclude that the submodular active learning algorithm developed at Duke is capable of finding novel clusters of TOI. However, when TOI clusters are embedded within clutter, it may be necessary to adapt the final classifier to aggressively overfit the training data in the early stages of digging.

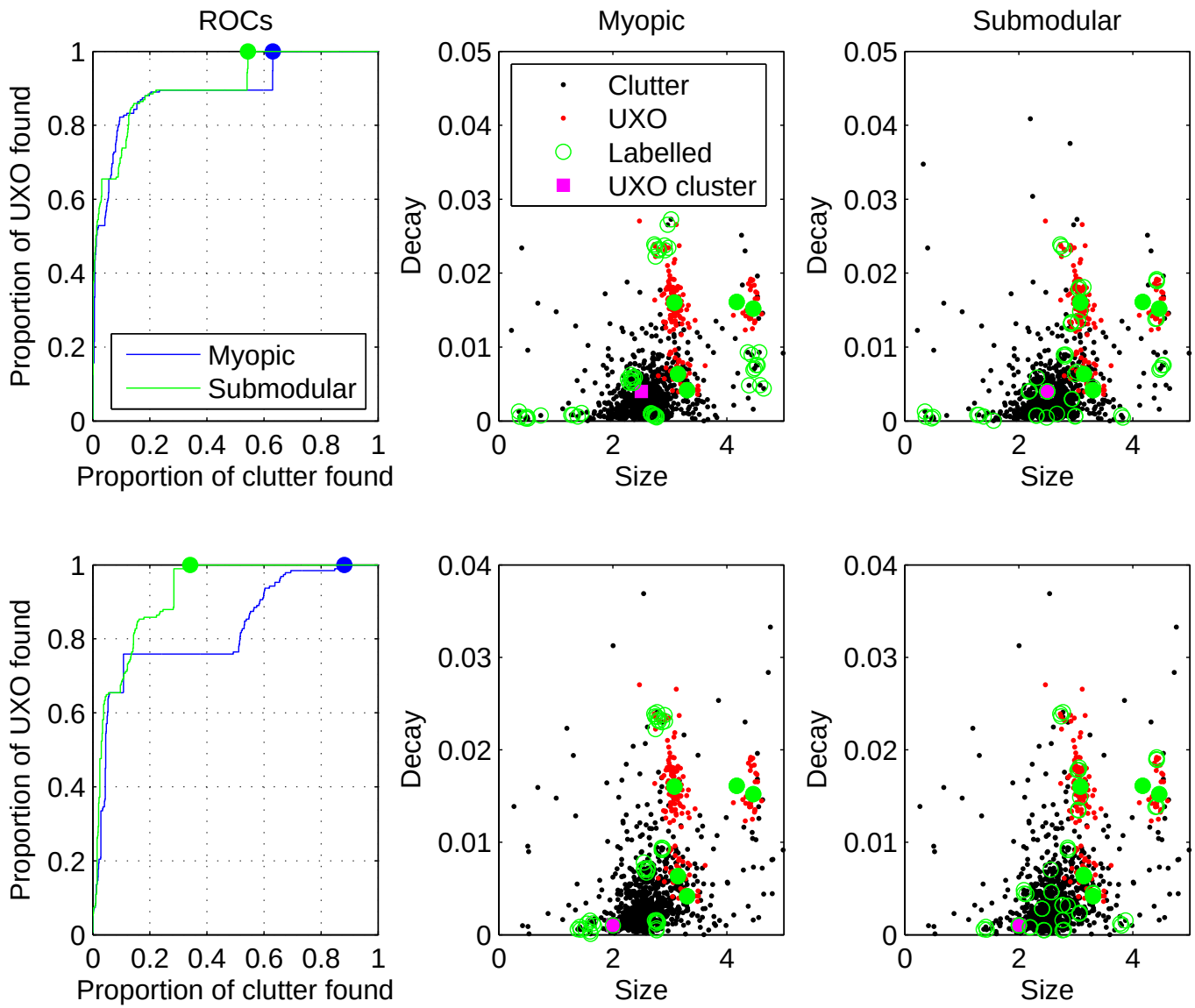


FIGURE 45. Comparison of myopic and submodular learning performance applied to Camp Butner MetalMapper size-decay features, with artificial clusters of TOI seeded in the test data.

4.2.4. *Active learning with polarizabilities.* Thus far we have focused on active learning in a simple two-dimensional feature space. However, polarizabilities estimated with next generation sensor data (e.g. MetalMapper) are sufficiently well constrained that excellent discrimination performance can be achieved by training classifiers directly on these parameters. For example, Shubitidze (2010) achieved near perfect discrimination performance on the Camp Butner MetalMapper data. This is in contrast to monostatic sensors (e.g. the Geonics EM-61): these instruments produce poorly constrained model estimates over a limited time range and so it is advisable to work with size-decay parameters (or even just the decay parameter) when ranking targets.

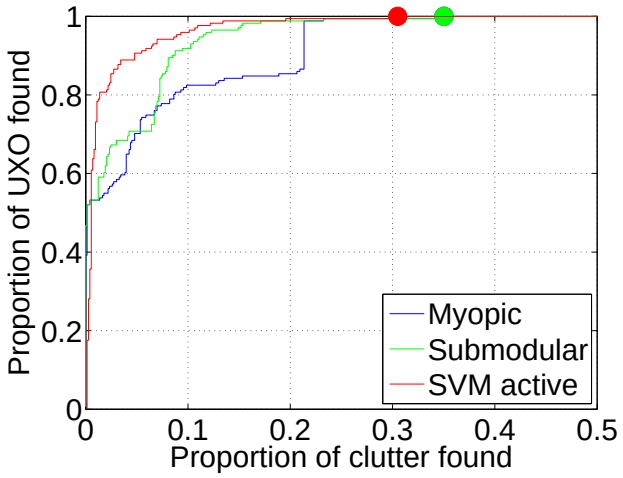


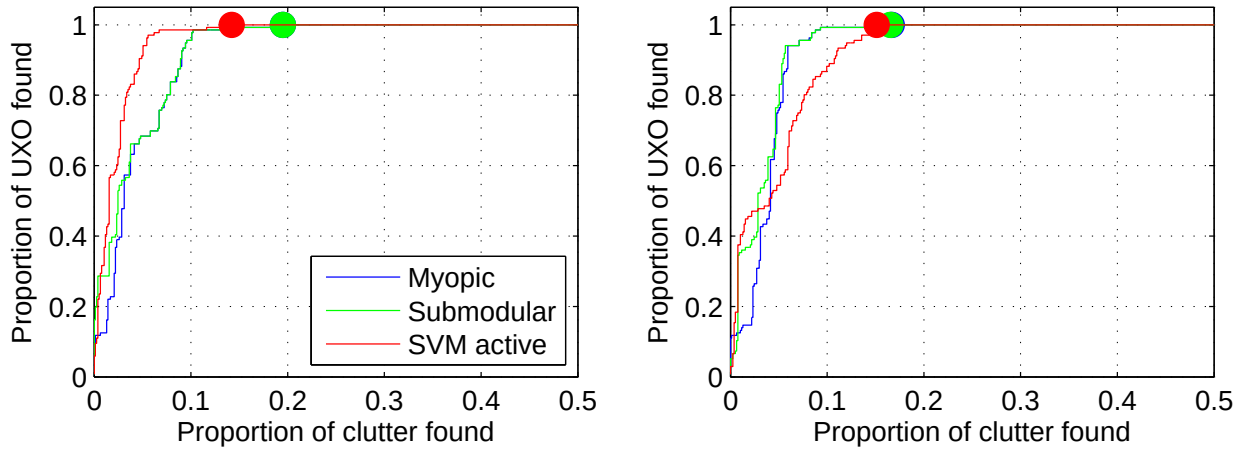
FIGURE 46. Comparison of myopic, submodular and SVM active learning performance applied to Camp Butner MetalMapper total polarizability features. Myopic and submodular algorithms have the same false alarm rate.

Figure 46 shows discrimination performance for active learning algorithms applied to log-transformed total polarizabilities (equation 5) from Camp Butner. In this example we use estimated polarizabilities at all 42 MetalMapper channels, and we initialize all algorithms with features from test pit TOI. Training submodular and myopic algorithms on these features does not produce any performance improvement relative to using size-decay features (shown in figure 43). In contrast, the SVM active algorithm has an increased AUC when trained on total polarizabilities. However, there is some difficulty finding the final TOI and consequently the FAR for the SVM active is increased relative to the analogous size-decay result.

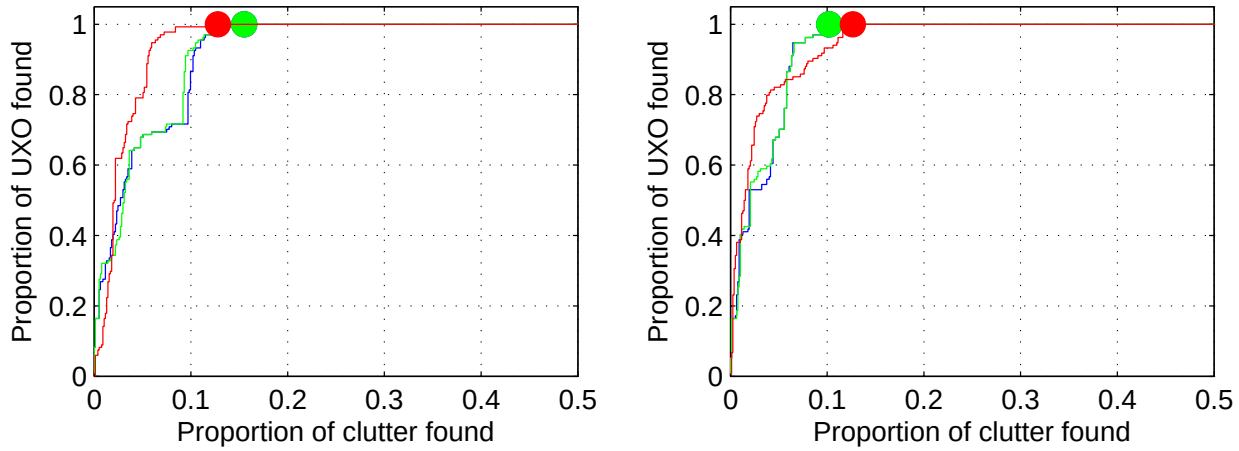
Figure 47 shows a second comparison of active learning algorithms, here applied to MetalMapper data from the 2011 Camp Beale demonstration. In this example we consider classifiers trained on either total or primary polarizabilities (log-transformed in both cases). The summation of polarizabilities will necessarily be dominated by the primary polarizability, so that there is a strong correlation between total and primary polarizability

feature sets. However, the total polarizability may be affected by poorly constrained tranverse (secondary and tertiary) polarizabilities. In these cases we may obtain a good library match with the primary, but not with the total.

In figure 47 all active learning algorithms achieve comaparable false alarm rates, with myopic and submodular algorithms attaining identical performance in all cases. These algorithms do slightly better when trained on primary polarizabilities. Active learning with the SVM has good initial performance (high AUC) when trained on total polarizabilities, with a marginal reduction in false alarm rate for primary polarizabilities.



(a) Beale P



(b) Beale C

FIGURE 47. Comparison of myopic, submodular and SVM active learning performance for Beale MetalMapper data sets. Left: classification on total polarizabilities, right: classification on primary polarizabilities. Myopic and submodular algorithms have the same false alarm rate in these examples.

4.3. **Development of a munitions response target database.** A primary objective of this project has been the integration of feature estimation capabilities developed by the SKY/UBC group with advanced classification algorithms from Duke University. To this end, we have implemented a web-accessible munitions response target database (MRTDB) comprised of sensor data, ground truth and estimated features from all ESTCP demonstrations dating back to San Luis Obispo in 2009. This is intended as a platform for researchers at Duke, UBC, and the broader UXO community to further test algorithms for feature extraction and classification. Similar standardized data sets within the machine learning community serve as testbeds for algorithm development and also promote reproducible research.

While this effort to some extent parallels development of a library of TOI polarizabilities within UX-Analyze (Keiswetter, 2009), our database includes responses for *all* demonstration targets, including both TOI and non-TOI. This provides researchers with the tools to, for example, characterize the variability of TOI polarizabilities across sites, or test the ability of a discrimination algorithm to distinguish between TOI and clutter of similar size.

The MRTDB is hosted at `www.skyresearch.com/mrtdb`, and users can log in with the username `estcpuser` and password `estcpuser!`. Arbitrary queries of the database can then be constructed by specifying desired fields in the browser interface, as illustrated in figure 48. The following fields can be specified:

- Site
- DigType (TOI, Cultural Debris, Munitions Debris, No Contact)
- Target number
- Class (e.g. 37 mm, 105 mm, etc.)
- Length
- Depth
- Dip

The database then returns a table of targets meeting the query criteria. Sensor data, images, or features for selected targets can be downloaded in a zip file. In addition, users can view individual inversion results in a PDF file. The PDF contains images of all inversion results, including both passed and failed models. PDFs for each data type are accessed via links in the `Data` column, as shown in figure 48.

Importing new data sets and ground truth into the database is straightforward, and we will maintain this resource as the ESTCP demonstrations continue beyond 2012. We will also leverage this work in the new start SERDP project MR-2226 (Decision support tools for munitions response performance prediction and risk assessment). Successful classification performance prediction given arbitrary site conditions will exploit data sets and features from previous demonstrations.

SERDP ESTCP Munitions Response Target Database

Home | Query Functions | Tools

Site: Beale
Class: *
Identification: *
Target: *
DigType: *
Depth(cm): >
Dip: >
Azimuth: >
Length(cm): >

Enable Sorting Enable Paging *Rows per Page:* 10 Apply Filter and Create Table

Nr	Photo	Site	Target	DepthCm	Dip	Azimuth	Identification	Class	DigType	LengthCm	EastingCm	NorthingCm	Data
1	View	Beale	1	24.000	-21.000	149.000	105mm	105mm	TOI	65.000	647324.839	4331165.313	TEMTADS2x2 MPV2 BUDhh
2	View	Beale	2	24.000	-49.000	52.000	81mm	81mm	TOI	49.000	647329.651	4331144.507	TEMTADS2x2 MPV2 BUDhh
3	View	Beale	3	0.000			Survey Nail	Cultural Debris	CD	21.000	647330.102	4331167.248	TEMTADS2x2 MPV2 BUDhh
4	View	Beale	4	4.000			Frag	Munitions Debris	MD	50.000	647331.828	4331175.988	TEMTADS2x2 MPV2 BUDhh
5		Beale	4	46.000	-42.000	250.000	81mm	81mm	TOI	6.000			
6	View	Beale	5	19.000	-25.000	310.000	60mm	60mm	TOI	13.000	647331.978	4331166.296	TEMTADS2x2 MPV2 BUDhh
7	View	Beale	6	12.000	-5.000	13.000	Frag	Munitions Debris	MD	13.000	647332.582	4331163.112	TEMTADS2x2 MPV2 BUDhh
8	View	Beale	7	9.000	90.000	0.000	Frag	Munitions Debris	MD	14.000	647347.015	4331159.335	TEMTADS2x2 MPV2 BUDhh
9	View	Beale	8	6.000	-45.000	205.000	Frag	Munitions Debris	MD	12.000	647320.760	4331167.344	TEMTADS2x2 MPV2 BUDhh
10	View	Beale	9	10.000	-40.000	160.000	Frag	Munitions Debris	MD	13.000	647315.962	4331163.083	TEMTADS2x2 MPV2 BUDhh
1 2 3 4 5 6 7 8 9 10 ...													Page 1/328

Download IMAGES for selected Items Download MAT Data for selected Items * Export Selected Items to CSV File

Logout

SERDP DOD • EPA • DOE ESTCP

This website was developed by Sky Research Inc. and the University of British Columbia under project MR-1657

FIGURE 48. MRTDB interface and example search results

References

A. Aliamiri, J. Stalnaker, and E. L. Miller. Statistical classification of buried unexploded ordnance using nonparametric prior models. *IEEE Trans. Geosci. Remote Sensing*, 45: 2794–2806, 2007.

T. Bell and B. Barrow. Subsurface discrimination using electromagnetic induction sensors. *IEEE Trans. Geosci. Remote Sensing*, 39:1286–1293, 2001.

S. D. Billings, L. R. Pasion, L. Beran, N. Lhomme, L. Song, D. W. Oldenburg, K. Kingdon, D. Sinex, and J. Jacobson. Unexploded ordnance discrimination using magnetic and electromagnetic sensors: Case study from a former military site. *Geophysics*, 75:B103–B114, 2010.

W. P. Delaney and D. Etter. Report of the Defense Science Board on Unexploded Ordnance. Technical report, Office of the Undersecretary of Defense for Acquisition, Technology and Logistics, 2003.

D. Keiswetter. Description and features of UX-Analyze. Technical report, ESTCP, 2009.

W. Menke. *Geophysical Data Analysis: Discrete Inverse Theory*. Academic Press, 1989.

L. R. Pasion. *Inversion of time-domain electromagnetic data for the detection of unexploded ordnance*. PhD thesis, University of British Columbia, 2007.

L. R. Pasion and D. W. Oldenburg. A discrimination algorithm for UXO using time domain electromagnetic induction. *Journal of Environmental and Engineering Geophysics*, 6:91–102, 2001.

L. R. Pasion, S. D. Billings, D. W. Oldenburg, and S. Walker. Application of a library-based method to time domain electromagnetic data for the identification of unexploded ordnance. *Journal of Applied Geophysics*, 61:279–291, 2007.

F. Shubitidze. Camp Butner UXO data inversion and classification using advanced EMI models. SERDP-ESTCP Symposium, 2010.

M. Szummer and T. Jaakkola. Partially labeled classification with markov random walks. In *Advances in Neural Information Processing Systems (NIPS)*, 2002.

M. Tipping. Sparse bayesian learning and the relevance vector machines. *Journal of Machine Learning Research*, 1:211–244, 2001.

G. F. West and J. C. Macnae. *Electromagnetic methods in applied geophysics*, chapter Physics of the electromagnetic exploration method, pages 5–45. SEG, 1991.

D. Williams, Y. Yu, L. Kennedy, X. Zhu, and L. Carin. A bivariate gaussian model for unexploded ordnance classification with EMI data. *IEEE Geosci. Remote Sensing Letters*, 4:629–633, 2007.

Y. Zhang, L. M. Collins, H. Yu, C. E. Baum, and L. Carin. Sensing of unexploded ordnance with magnetometer and induction data: theory and signal processing. *IEEE Trans. Geosci. Remote Sensing*, 41:1005–1015, 2003.

Y. Zhang, X. Liao, and L. Carin. Detection of buried targets via active selection of labeled data: Applications to sensing subsurface uxo. *IEEE Trans. Geosience and Remote Sensing*, 42:2535–2543, 2004a.

Y. Zhang, X. Liao, and L. Carin. Detection of buried targets via active selection of labeled data: Application to sensing subsurface UXO. *IEEE Trans. Geosci. Remote Sensing*, 42: 2535–2543, 2004b.

Appendix

List of Scientific/Technical Publications

Q. Liu, X. Liao, H. Li, J. Stack and L. Carin, "Semi-supervised multitask learning," IEEE Trans. Pattern Analysis Machine Intelligence, vol. 31, pp. 1074-1086, June 2009

J. Paisley, X. Liao and L. Carin, "Active learning and basis selection for kernel-based linear models: A Bayesian perspective," IEEE Trans. Signal Processing, vol. 58, pp. 2686-2700, 2010

C. Wang, X. Liao, D. Dunson and L. Carin, "Multi-task learning for incomplete data," J. Machine Learning Research, vol. 11, pp. 3269-3311, 2010

S. Han, X. Liao and L. Carin, "Cross-Domain Multitask Learning with Latent Probit Models," Proc. Int. Conf. Machine Learning (ICML), 2012

L. Beran, S.D. Billings and D. Oldenburg, "Incorporating Uncertainty in Unexploded Ordnance Discrimination," IEEE T. Geoscience and Remote Sensing 49(8): 3071-3080 (2011)

Patents

None