INTERIM REPORT

Integration of Advanced Statistical Analysis Tools and Geophysical Modeling

SERDP Project MR-1657

DECEMBER 2010

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.

Contents

I Abstract 2

II Feature extraction and comparison of classification performance 2
- II-A Feature extraction: data features . . . 2
  - II-A.1 Data features . . . 3
  - II-A.2 Feature extraction and discrimination with a simplified fingerprinting method . . . 5
- II-B Comparison of classification performance . . . 12

III Workflow development 20

IV Numerical simulation of sensor data for input into feature extraction and classification 21

V Spatially-Constrained Classifiers for UXO Detection 24

VI Motivation of Spatially-Constrained Classification 25

VII Spatially-Constrained Classification Models 29
- VII-A Spatially-Constrained Gaussian Mixture Models . . . 30
- VII-B Spatially-Constrained Mixture of Experts . . . 36

VIII Offline-learning and Online Learning 38

IX Experimental Results 39
- IX-A Experimental setup . . . 39
- IX-B Detection Results . . . 40

X Discussion and future work 48

References 51

I. Abstract

This report details progress on a joint project between Duke University, the University of British Columbia and Sky Research, on the development of advanced statistical signal processing and geophysical-modeling tools for sensing buried unexploded ordnance (UXO). The research reported here discusses feature extraction, workflow development, and exploitation of additional knowledge about UXO (that they are rare). Further, we have performed a detailed comparison of different classification strategies, across the different components of the research team. The methods are tested throughout using measured data from actual sites, often utilizing data collected under related ESTCP projects.

II. Feature extraction and comparison of classification performance

UBC and Sky have processed a number of data sets (either retrospectively or as part of ongoing demonstration projects) and provided these to the Duke University group for input into their classification algorithms. Data sets and features provided in Year 1 of the project include:

1) Camp Lejeune. EM63 data collected at this site were re-inverted using our latest inversion codes and data and features were provided to Duke. These data were used by Duke for a synthetic demonstration of a classifier which incorporates the spatial distribution of targets of interest into discrimination decisions. UBC/Sky have subsequently used these data to develop and test a simplified fingerprinting method, as described later in this section.
2) San Luis Obispo. Extracted features were also provided for MTADS EM61, MTADS magnetics, EM61 cart, and TEMTADS data sets from SLO. For all TEM data sets, an instantaneous amplitude three dipole model was used to fit the data. In this formulation, the three principal dipole polarizabilities are estimated at each time channel and can be subsequently input into a discrimination algorithm. We compare discrimination performance at SLO in Section II-B.

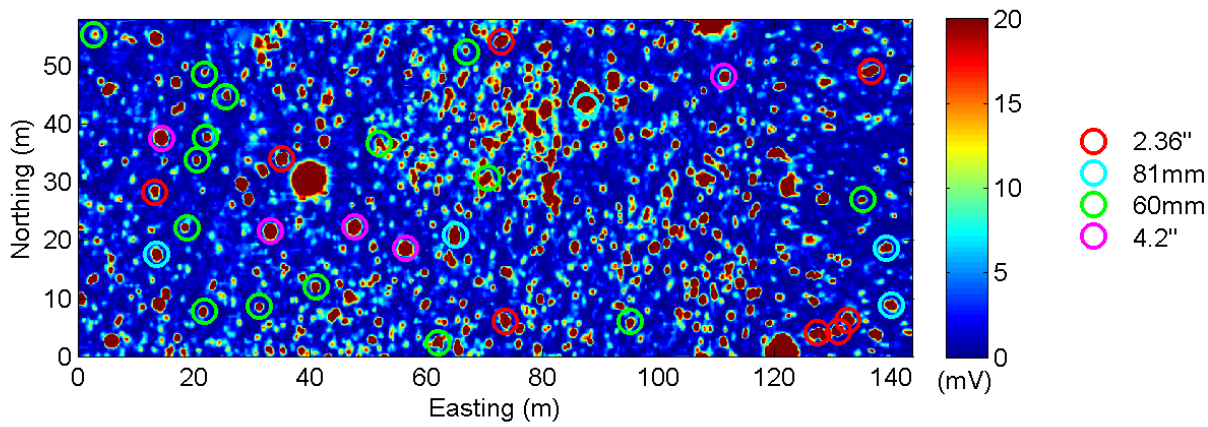
A. *Feature extraction: data features*

Most current UXO research is focused upon the estimation of model-based features and subsequent training of statistical classifiers using these features. Results of discrimination studies at Camp Sibert and San Luis Obispo have shown that this data-processing effort is worthwhile: discrimination performance with next-generation sensors consistently outperforms the 10:1 false alarm rate envisioned in the Defence Science Board report on UXO [1]. There are scenarios, however, where model-based discrimination is not feasible or necessary. While industry is increasingly adopting advanced discrimination techniques, surveys with production sensors (e.g. the Geonics EM61) will likely still predominate when the discrimination task is relatively simple or when more advanced technology and expertise are unavailable.

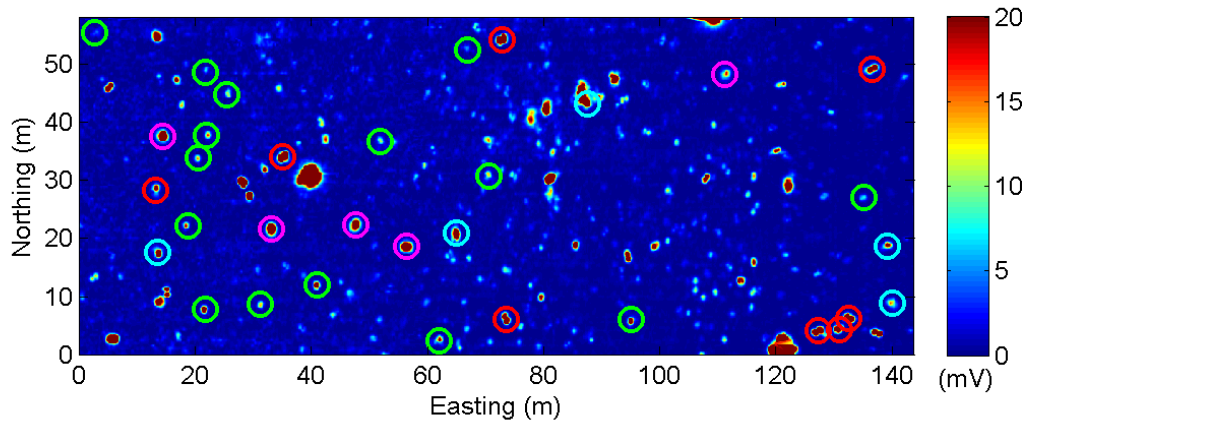
Considerable time is required for target masking and inversion QC when processing data from production sensors, and this may preclude model-based discrimination when a short turn-around is required between the geophysical survey and digging. Motivated by these thoughts, we investigate the performance of data-based discrimination strategies. Our aim is to develop and test fast and robust approaches to discrimination using data features (e.g. anomaly width) or a simplified (i.e. linear) form of the TEM dipole model. Here a data feature is any parameter which is estimated from observed data but which does not involve inversion with the TEM dipole model. This UBC/Sky-driven work is intended to complement research at Duke on

dictionary methods (ongoing in parallel under this project), which similarly avoid the use of the dipole model.

1) Data features: Figures 1 and 2 show data and data features computed for a subsection of EM61 cart data acquired at SLO. In 2(a) we compute a decay constant β for each sounding d by fitting a power



(a) Channel 1



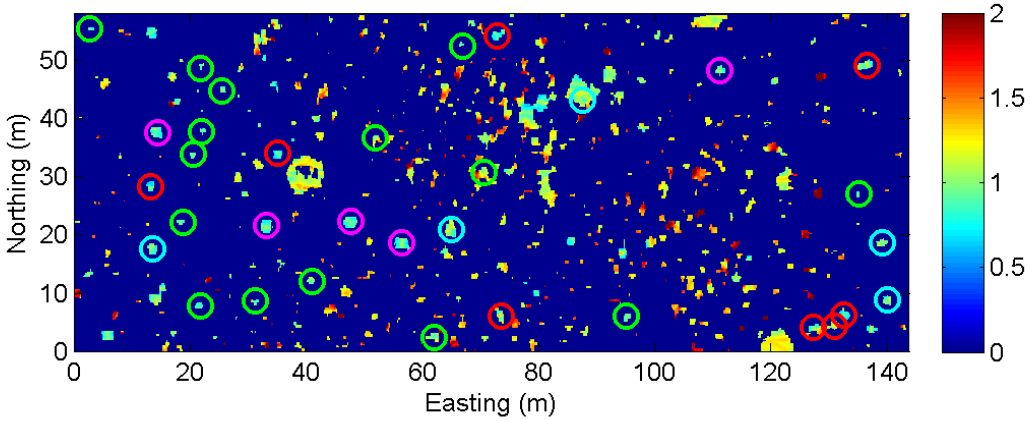
(b) Channel 4

Fig. 1. Data, SLO EM61.

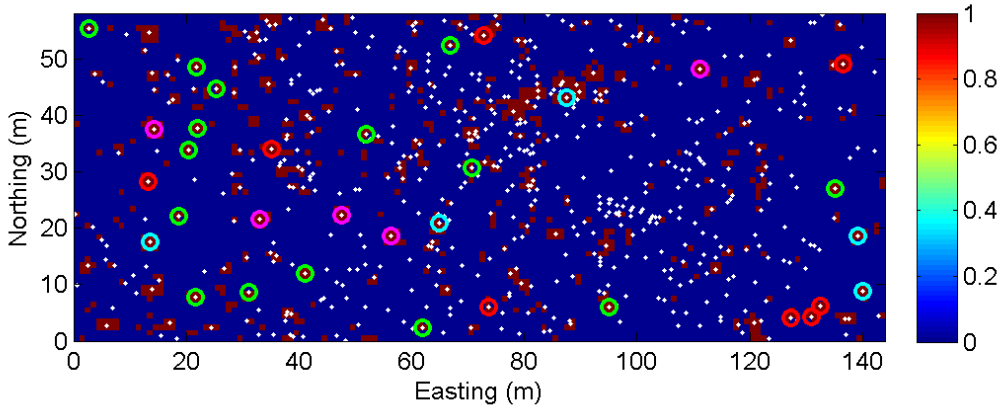
law of the form

$$d(t) = kt^{-\beta} \tag{1}$$

to soundings exceeding a predefined minimum threshold (here selected to be 10 mV at the first time channel. This cutoff corresponds to the detection threshold for targets of interest at this site). In addition, if the correlation coefficient between the observed and predicted data for a given sounding does not exceed 0.95 then the decay constant for that sounding is not included when generating the gridded image in 2(a). This type of decay constant analysis is commonly used in exploration geophysics to identify geologic targets of interest. In Figure 2(a) larger targets (4.2" mortars) are evident as regions of slow decay with a somewhat broader spatial extent, but small targets such as 60mm may appear as only a few slow decaying pixels and are not clearly detectable in the image. In Figure 2(b) we translate the image of the decay



(a) Decay Constant



(b) Dig map. White points are original targets picked from EM61 data for SLO demonstration

Fig. 2. Data features, SLO EM61.

constant into a binary "dig map". Within each 1 x 1 m pixel in this image, we compute the median decay constant. If the decay constant is less than a predefined threshold (here 1.1), then that pixel has a value of 1 in the dig map (colored red in 2(b)). [2] shows that for conductive targets, the time constant has a lower bound of 1/2. An effective upper bound for the decay constant is 3/2, representing the "late early time" decay for ferrous targets derived by Weichman. Here we have selected a decay constant threshold that is (approximately) intermediate to these bounds. Forward modelling of UXO responses could also be used to objectively determine a threshold.

We see that all TOI in this image coincide with red (dig) pixels, while 399 of 540 non-TOI targets picked for the demonstration within this grid coincide with blue (no-dig) pixels. The dig map is an alternative to the usual target-based diglist: rather than providing dig teams with an ordered list of targets we generate a set of discrete regions which must be cleared. This requires that field technicians have a means to accurately delineate a region, perhaps using a template which can be positioned using GPS. In Figure 2(b) we have not prioritized the regions, though of course a ranking based on median decay rate could be generated if necessary.

The dig map avoids the need for target picking and inversion, which can become problematic in highly cluttered regions. However, the result is quite sensitive to the threshold on median decay rate, and this parameter must be carefully inferred from testpit measurements and simulations. While this preliminary result seems promising, the method is likely to work best with sensor data extending later in time (e.g. TEMTADS data) where slow-decaying TOI are easily discriminated from fast-decaying clutter. The efficacy of this approach on next generation sensor data will be investigated in further work.

A second data feature which has been proposed for UXO discrimination is the width of the target anomaly. [3] showed that 4.2" mortars at Camp Sibert can be distinguished from clutter by fitting a bivariate Gaussian to the data at a single time channel. Here the data at location $r = (x, y)$ are modelled as

$$d(r) = A \exp(-(r-\mu)^T S^{-1} (r-\mu)) + b \tag{2}$$

with A an amplitude, μ and S the mean and covariance, and b a bias term. While this simple model can generally reproduce observed data acquired with a monostatic sensor (Figure 3), it is easy to envision scenarios where it might produce a poor fit. For example, a horizontal target may sometimes have an

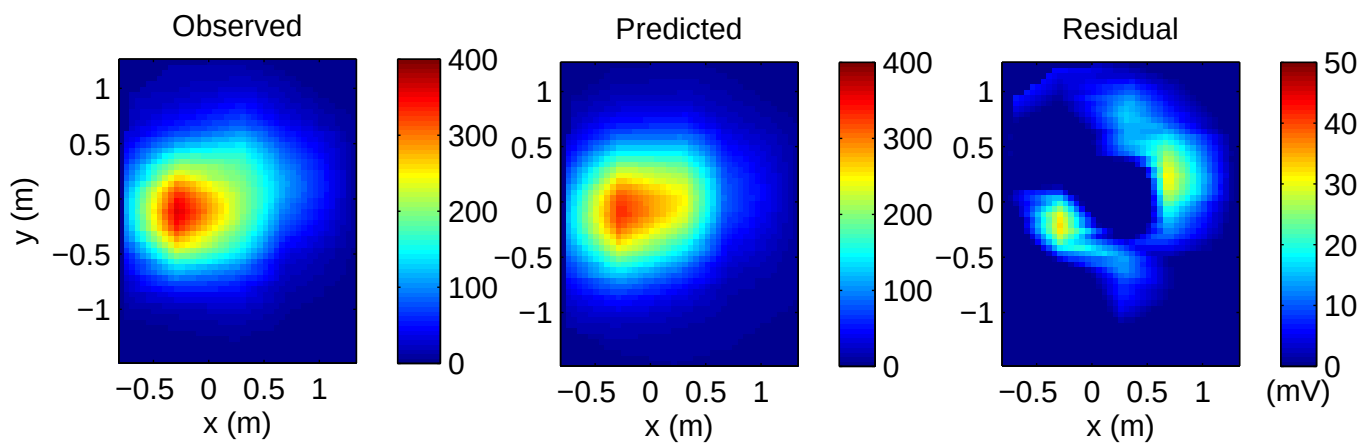


Fig. 3. Example fit of a bivariate Gaussian to channel 1 of EM61 anomaly at SLO.

anomaly which is better described as a bimodal mixture of Gaussians. However, the single Gaussian expressed above can still provide a first order indication of anomaly width, as quantified by the trace of the estimated covariance. At Camp Sibert this data feature outperformed discrimination with a statistical classifier trained on dipole model parameters. However, this positive result may be specific to the Camp Sibert data - excellent discrimination of 4.2" mortars can be achieved with most approaches. SLO presents a more challenging problem, and in Figure 4 we see that anomaly width estimated via equation 2 provides no significant separation between TOI and clutter. However, anomaly width may be used to pre-screen very small anomalies prior to inversion and to thereby reduce time required for QC.

2) Feature extraction and discrimination with a simplified fingerprinting method: Testpit and instrument verification strip (IVS) measurements can provide valuable information which should be exploited when processing field data. Here we demonstrate how these measurements can be used to generate a library (or, synonymously, a dictionary) of expected responses which can then be used to fit observed field data. In this approach we fit each sounding independently, resulting in a linear inversion problem which

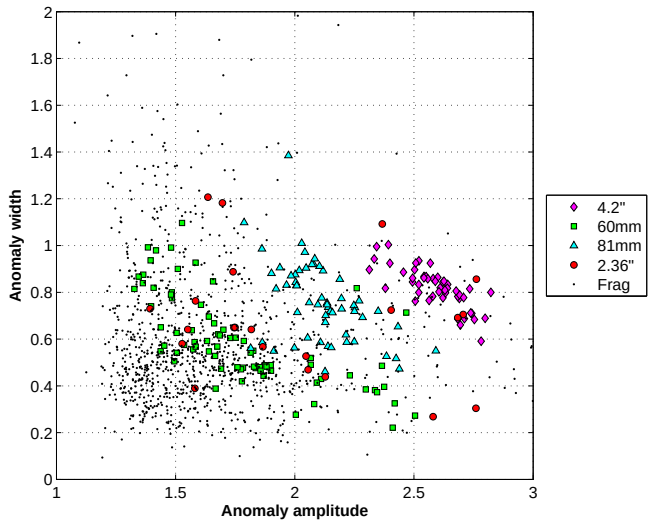


Fig. 4. Spatial data features: a bivariate Gaussian is fit to each anomaly, providing target width (trace of the covariance), and amplitude (coefficient A in equation 2).

eliminates extrinsic parameters (i.e. target location and orientation) from the model. Strictly speaking this is a model-based method, but it is very fast and can potentially eliminate the requirement for target picking and masking. We regard it as a compromise between discrimination with data features and discrimination with features derived from the full TEM dipole model. We first demonstrate the viability of the method on a simple discrimination problem from an older data set (Camp Lejeune, North Carolina).

The goal of the Camp Lejeune study was to identify pre-existing and emplaced ordnance items from ubiquitous clutter on an active range. Clutter items were predominantly non-ferrous (aluminum) "adapters" used as casings for armor-piercing sabot rounds (figure 5). Here we focus on analysis of an EM63 data

Fig. 5. Adapter excavated at Camp Lejeune

set acquired at Camp Lejeune.

The TEM dipole model assumes that the observed data are a linear superposition of a target's polarizations $L_i(t)$

$$d^{pred}(t_i) = \sum_{j=1}^{N} w_j L_j(t_i). \tag{3}$$

$N = 2$ or 3 for an axisymmetric or non axisymmetric target, respectively. The coefficients w_j vary nonlinearly from sounding to sounding as a function of target and sensor location and orientation. However, if the target polarizations are known (i.e. via inversion of high quality testpit data) then for a single sounding the coefficients can be estimated by solving a linear inverse problem. We use these polarizations to discriminate between adapters and ordnance as follows:

1) Identify all soundings above a predefined threshold. At Camp Lejeune we used 15 mV at the first time channel.
2) For each sounding identified in (1), minimize

$$\phi = \|\mathbf{W}_d(\mathbf{d}^{obs} - \mathbf{d}^{pred})\|^2 \text{ subject to } w_j \geq 0,\ j = 1, 2. \tag{4}$$

with $\mathbf{d}^{pred}$ given by equation 3. The data weightings $\mathbf{W}_d$ are assigned as the inverse data standard deviations, here estimated as 10 percent of each observed datum plus a 1 mV floor. We impose a positivity constraint on the coefficients w_j based upon the expression for the measured voltage over an axisymmetric target at depth z illuminated by a vertical primary field with magnitude B^p

$$V(t) = 2\kappa \frac{B^p(z)}{z^3} \left[L_1 \cos^2(\theta) + L_2 \sin^2(\theta)\right] \tag{5}$$

with κ a non-negative constant depending upon sensor geometry, and θ the dip angle of the target [4]. The coefficients of the polarizations L_1 and L_2 are therefore non-negative for a monostatic sensor. High SNR soundings identified in step (1) correspond to locations where the horizontal displacement of the sensor from the target is small, so that the primary field is approximately vertical at the target and equation 5 is a good approximation to the measured voltage. Figure 6 indicates that the coefficients w_j for the EM63 are non-negative regardless of target-sensor separation. The optimization problem (equation 4) is solved with the `lsqlin` routine in Matlab.
3) Compute the correlation coefficient (CC) between the (weighted) observed and predicted data, and for each target retain the median correlation coefficient over all soundings. Other CC statistics such as the mean or maximum were also investigated, we choose the median because it is a robust statistic and so is less influenced by noisy soundings with low CC.
4) Dig targets based upon a ranking of median CC.

At Camp Lejeune high quality test pit measurements were made to recover estimates of the polarization decays for adapters (Figure 7(a)). Figure 7(b) shows observed and predicted data for soundings corresponding to the median CC for representative adapter (clutter) and ordnance targets at Camp Lejeune. A much better fit is obtained for the adapter than for the ordnance target, and so the former will occur much later in the diglist (ordered from likely TOI to non-TOI) generated by the simplified fingerprinting technique.

Figure 8 compares the expected performance of quadratic discriminant analysis (QDA) classifiers trained on features derived from the TEM dipole model and the fingerprinting technique. QDA assumes that the class distributions in the feature space are Gaussian, with individual means and covariances estimated for each class [5]. For this study, dipole polarizabilities were fit with a Pasion-Oldenburg parameterization of

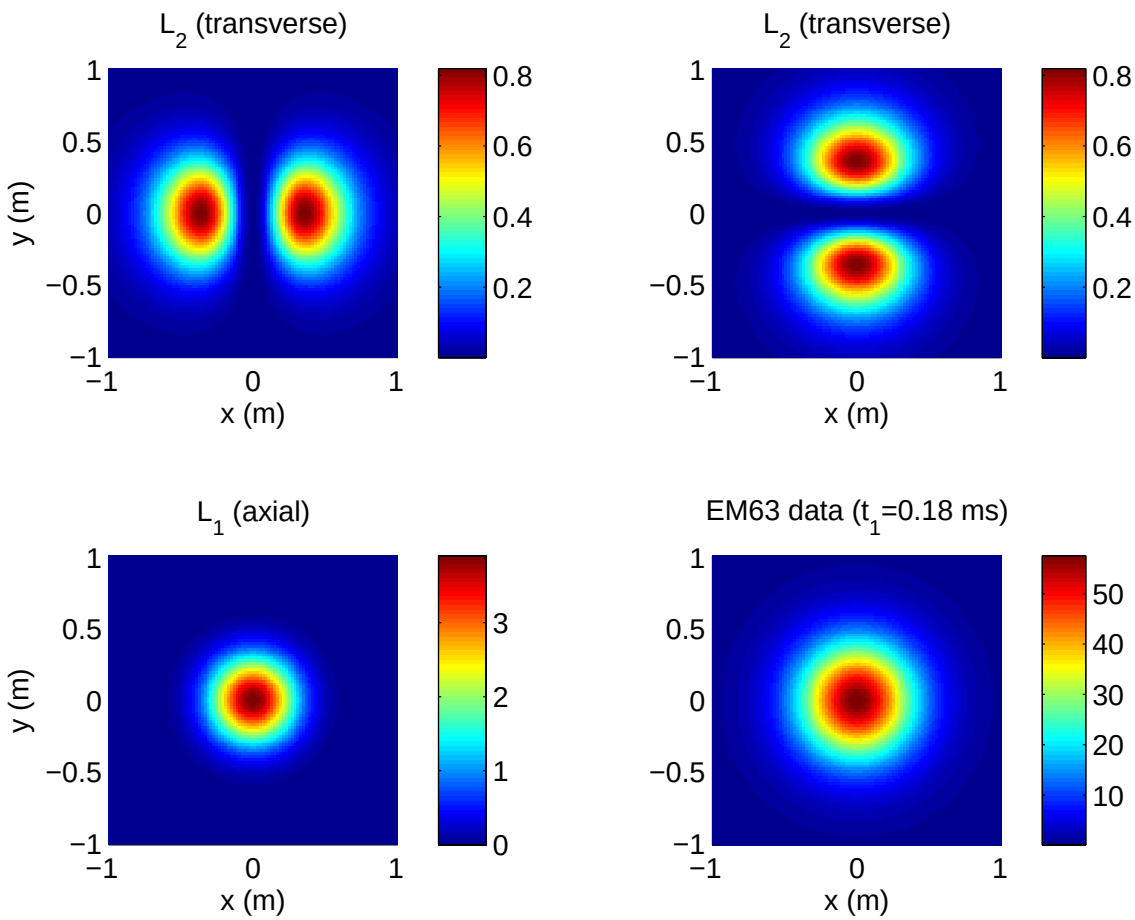


Fig. 6. EM63 polarization coefficients for a vertical target at 30 cm depth. The coefficients are non-negative regardless of sensor-target separation. Predicted EM63 data are for a vertical adapter. The predicted data are a weighted sum of these coefficients, with the weightings given by the polarizations $L_i(t_j)$.

the form

$$L_i(t) = k_i t^{-\beta_i} exp(-t/\gamma_i) \tag{6}$$

and the estimated parameters (k, β, γ) were input into QDA.

On the basis of the average performance shown in Figure 8, we conclude that for this particular data set fingerprinting is the best choice of discrimination algorithm, followed closely by statistical classification with the β parameters. This does not imply, however, that fingerprinting will always be the best technique for discrimination with electromagnetic data. Indeed, [6] found that statistical classifiers outperformed "data-space" discrimination algorithms (analogous to fingerprinting) for discrimination of EM63 data.

The simplified fingerprinting technique can also be applied to next generation sensor data to generate a dig map similar to that shown in Figure 2(b). In Figure 9 we show a dig map generated by fitting 37 mm library polarizations to dynamic MetalMapper data acquired at Camp Butner. In this case a pixel is colored red (dig) if the median correlation coefficient between the observed and predicted data exceeds 0.99 for soundings within that pixel. This processing risks missing lower SNR 37mm which may have a

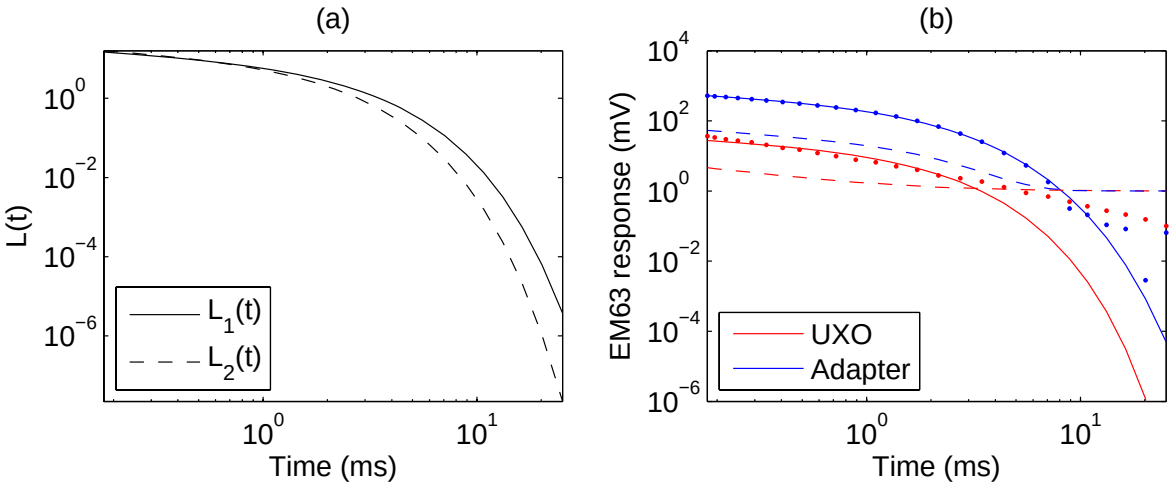


Fig. 7. (a) Adapter library polarizations. (b) Fingerprinting fits to soundings for ordnance and adapter targets. Solid line is predicted data, dots are observed data, and dashed line indicates the estimated standard deviations.

reduced correlation coefficient, but should, as a first pass, identify regions where isolated 37mm targets are most likely. It also circumvents the rather tricky process of target picking with MetalMapper data and is very fast (approximately 30 minutes processing time for these data on a dual quadcore desktop).

The simplified fingerprinting approach presented here is a simplification of the method developed by [7] and is similar to methods considered in [8] and [9]. By eliminating extrinsic target properties (location and orientation) from the forward model, we obtain a linear inverse problem which can be solved very quickly. The cost of this simplification is that the coefficients are no longer constrained by the physics of the dipole model, although the imposed positivity constraint does at least ensure that the predicted decay is physical. However, it is possible that the coefficients estimated for two different soundings cannot be generated by a single target. This ambiguity can only be resolved by considering all soundings simultaneously in a nonlinear inversion. Despite these limitations, the simplified fingerprinting approach does perform quite well on these data and its speed and simplicity make it a candidate for real-time discrimination.

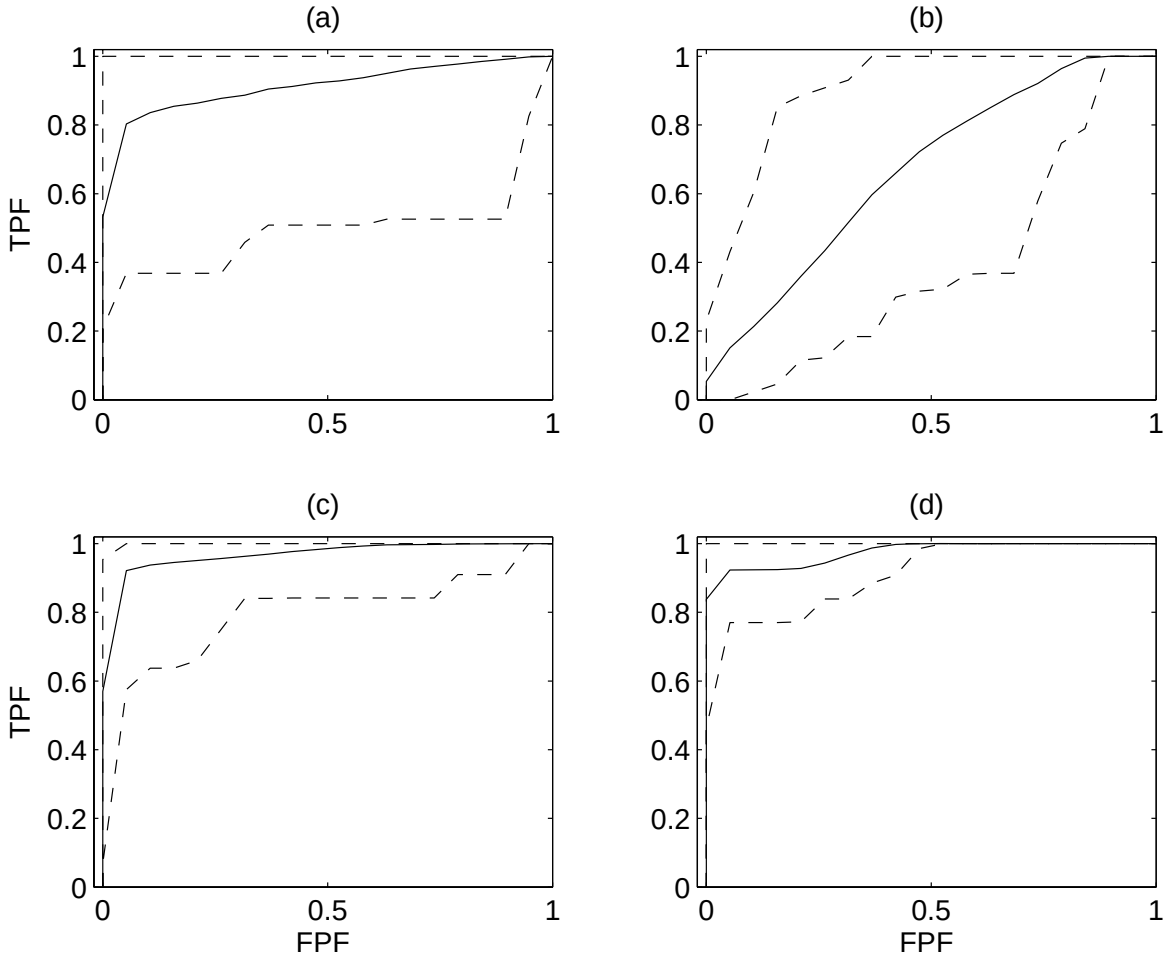


Fig. 8. Receiver operating characteristics (ROC) showing true positive fraction (TPF) versus false positive fraction (FPF) for quadratic discriminant analysis (QDA) classifiers (a-c) and library method (d). Here QDA is applied to discrimination between ordnance and adapters using features estimated from EM63 data. Solid line is the mean bootstrapped ROC, dashed lines are best and worst case ROCs from bootstrapping. Feature spaces are: (a) all polarization parameters k_i, β_i, γ_i, $i = 1, 2, 3$, (b) polarization amplitude parameters k_1, k_2, k_3 , (c) decay parameters β_i, $i = 1, 2, 3$, (d) fingerprinting.

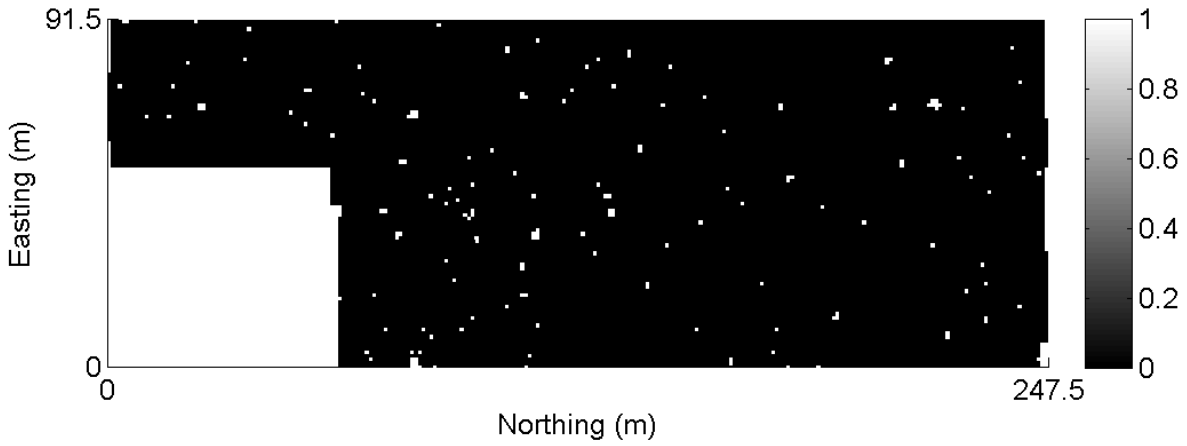


Fig. 9. Dig map for 37 mm targets computed from Dynamic MetalMapper data acquired at Camp Butner. White pixels correspond to regions where a high SNR 37mm target is likely.

B. Comparison of classification performance

Figures 10 through 13 show receiver operating characteristics for data sets acquired at San Luis Obispo. Subplot titles indicate which group produced the result:

- SIG: Duke/SIG group
- SIGUBC: Feature extraction by Sky/UBC, classification by Duke/SIG.
- Sky: Sky/UBC group.

The classification algorithms denoted by subplot titles are:

- PNBC: Semi-supervised classifier
- RVM: Relevance vector machine
- RVMKW: Relevance vector machine with active learning
- TimeDecay: Threshold on polarization time decay.
- Library: Fingerprinting method of [7].
- StatisticalClassifier: nonparametric classifier trained on polarizability amplitude and decay.

The most salient metric for evaluating discrimination performance in the context of UXO discrimination is the false alarm rate (FAR) at which all TOI have been dug. Discrimination algorithms can often achieve good initial detection of UXO, but may have difficulty identifying a few outlying ordnance in the feature space. This drives up the FAR and so does not achieve the primary goal (for a regulator) of reducing the number of non-TOI digs. A secondary metric for evaluating discrimination is the area under the ROC curve (AUC), which is representative of average discrimination performance throughout digging. Given two discrimination algorithms with similar FARs, the algorithm with larger AUC will likely be preferred by a regulator. This is because a larger AUC corresponds to an ROC that tends to identify the majority of TOI in the first stages of digging. This would allow EOD teams to focus their initial efforts almost exclusively on disposal of UXO and hence make field operations more efficient.

The best performance - in terms of area under the ROC and false alarm rate (FAR) - for the MTADS array (Figure 10) was achieved by a threshold on polarizability decay rate. The semi-supervised classifier trained on features extracted by the Duke group had the best performance of statistical classifiers applied to these data. Collaborative SIGUBC results are comparable to SIG classifiers in terms of FAR, but have slightly worse AUC. For the MSEMS EM61 analysis, the semi-supervised PNBC classifier trained on SIG features had the best performance, followed by a threshold on decay rate. Here the SIGUBC results were significantly worse than the independent classifications from either group. We conclude that the most robust classification approaches for production sensor data at SLO are thresholding on polarizability decay and semi-supervised classification. Based on these results we have adopted a hybrid classification approach for the Camp Butner demonstration: for all targets with polarizability decay above a threshold, we apply a statistical classifier to rank targets (Figure 20). Below the threshold, we revert to a simple threshold on decay rate.

¿From the comparisons of TEMTADS and MetalMapper performance at SLO (Figures 12 and 13), we conclude that while most classifiers yield excellent results, no single classifier performs best on all data sets. For the TEMTADS analysis it does appear that features estimated by Sky/UBC do provide a slight advantage in the performance of Sky and SIGUBC classifiers relative to independent SIG results.

An excellent result is obtained with the semi-supervised classifier applied to SIG MetalMapper features (Figure 13).

The mediocre performance of the collaborative SIGUBC efforts (perhaps with the exception of SIGUBC TEMTADS results) at SLO underline the necessity for good communication and feedback between analysts working on feature extraction and discrimination. Ideally, all workers should be involved to some extent in feature extraction so that they understand how parameters are estimated and the complications encountered in inversion QC. All analysts should similarly have some exposure to discrimination procedures. A simple step we have taken in this direction is to display inversion results together with all estimated test data features in a single QC window (Figure 14). Showing the analyst where the current inversion lies in feature space can aid in deciding whether further processing is required. For example, if the fit is mediocre but the polarizability decay is quite fast, then we may safely conclude that the target is likely not a TOI and so avoid further masking and inversion.

In Figure 15 we show a retrospective performance comparison for classifiers trained on UBC size/decay features extracted from SLO MetalMapper data. Note that the number of non-TOI in this Figure differs from that in Figure 13 because IDA used only a subset of test targets to generate their ROCs. For this comparison the PNBC and RVM classifiers were provided to us by Duke and all classifier training was carried out independently at UBC. We also trained a PNN (probabilistic neural network) classifier on the same data. In this case, classification performance is good for all algorithms, with the PNN achieving the best FAR and AUC. Consistent with the original SIG MetalMapper result, the PNBC semi-supervised algorithm has a lower false alarm rate than the RVM. Retrospective application of Duke classifiers to TEMTADS size/decay features yielded similar results (Figure 16). The RVM outperforms the PNN and is is able to detect an outlying TOI in this case, but the retrospective classifiers all have a worse AUC than the original submitted library and statistical methods.

We adjusted default classifier parameters (e.g., kernel widths) very little to obtain these results and it is possible that improved performance can be achieved with the Duke classifiers if parameters are carefully optimized. An important task for the upcoming year will be to gain further experience with these and other Duke/SIG algorithms so that they can be confidently included in workflow processing.

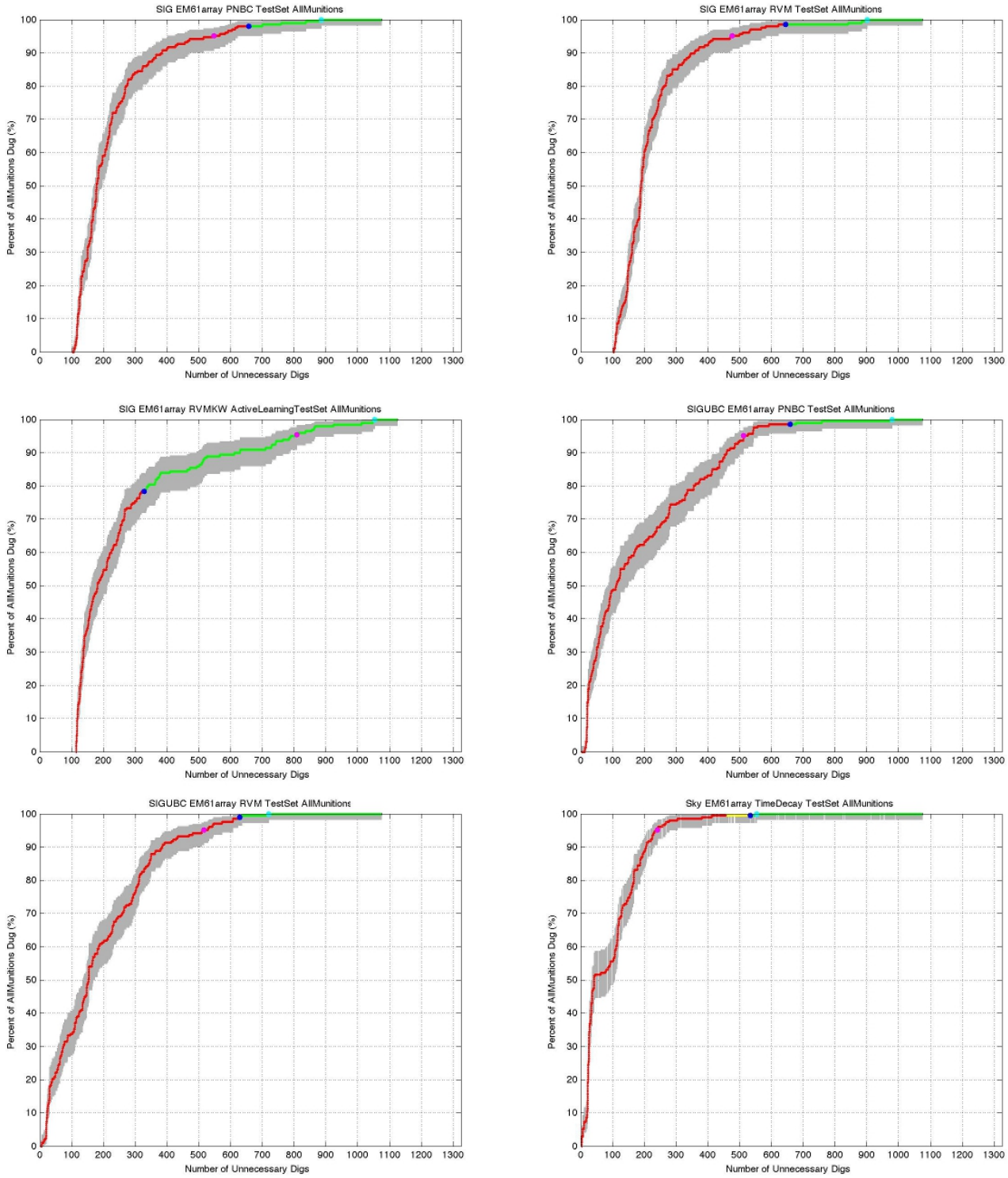


Fig. 10. Performance comparison, MTADS EM61.

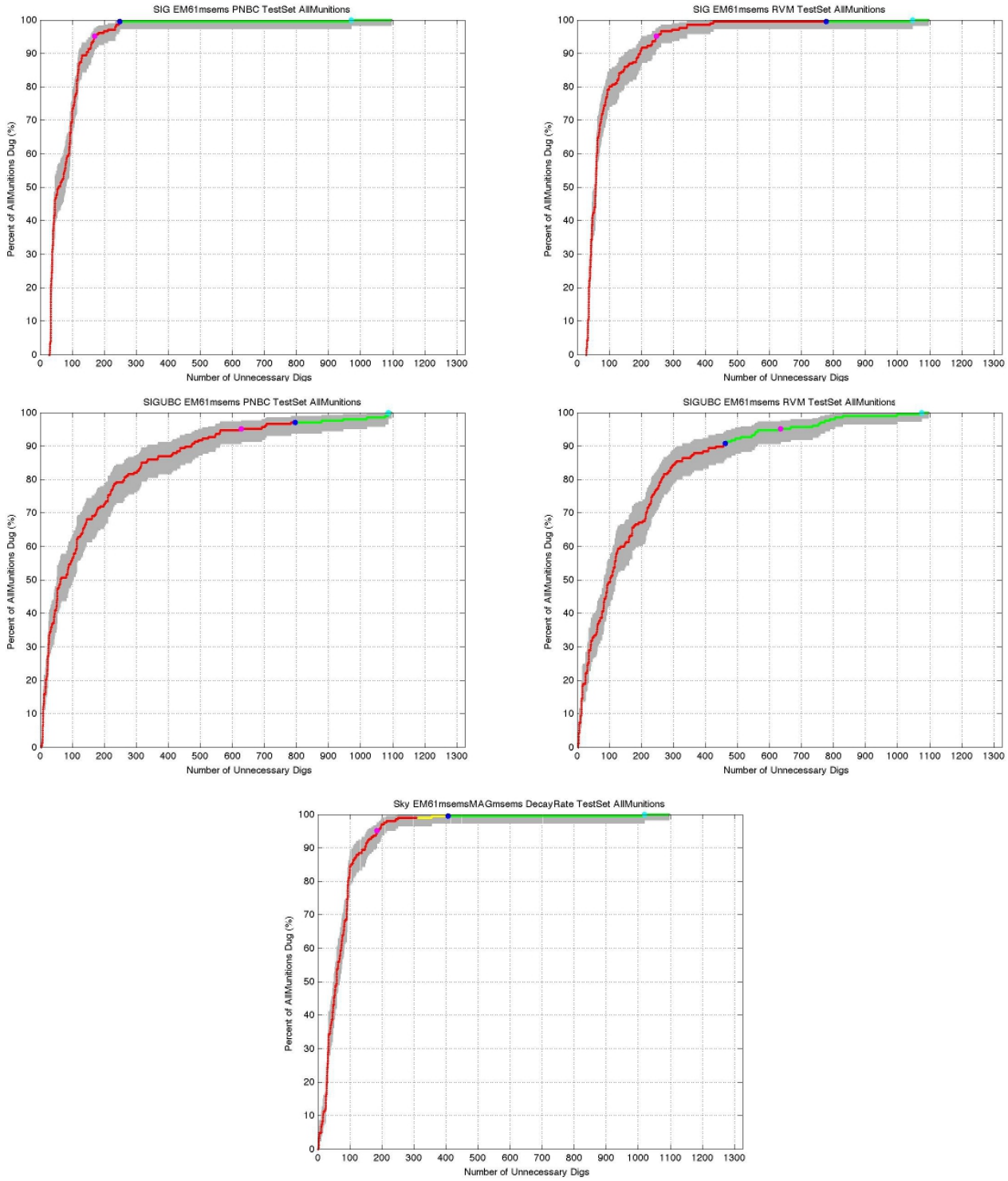


Fig. 11. Performance comparison, MSEMS.

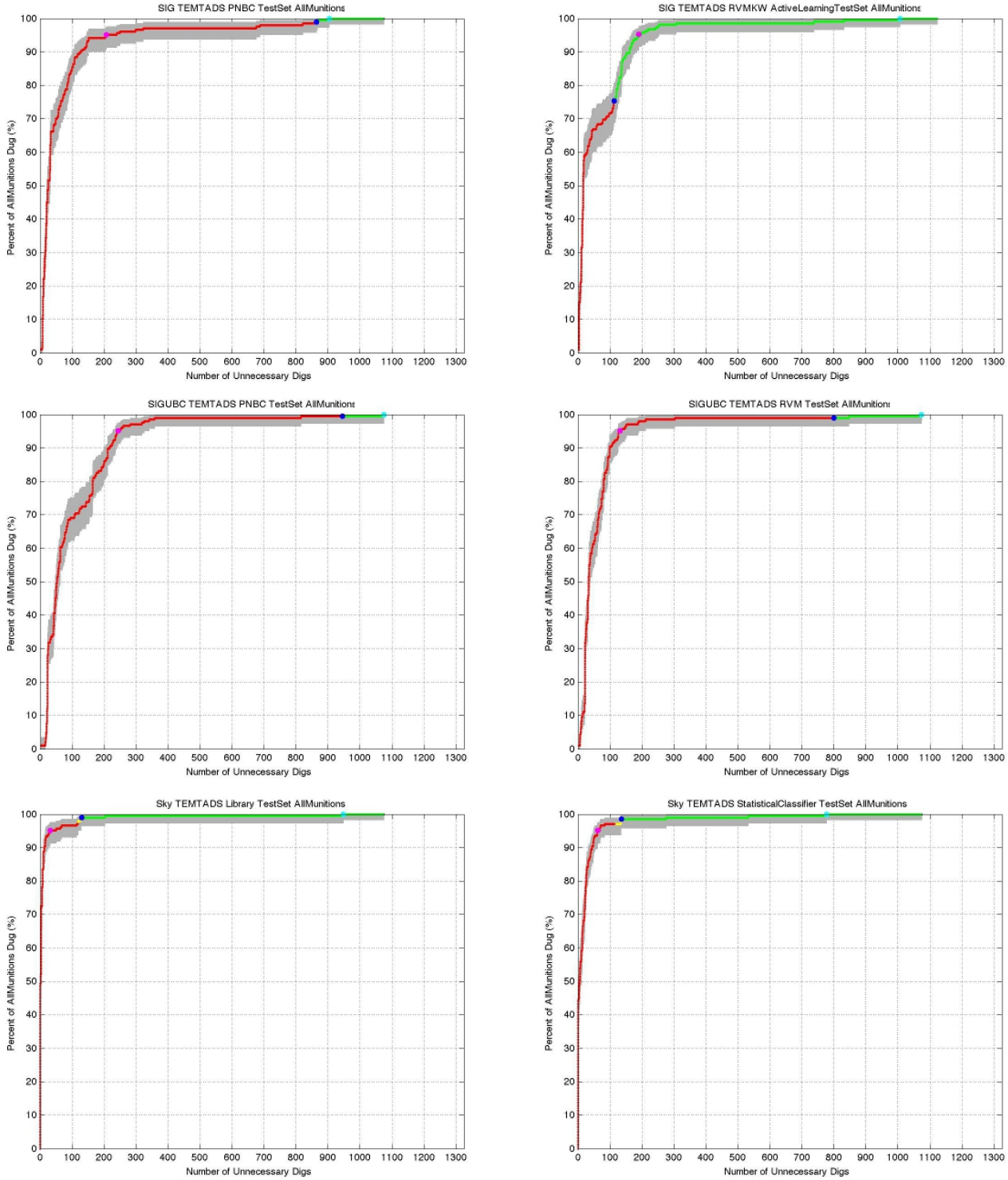


Fig. 12. Performance comparison, TEMTADS.

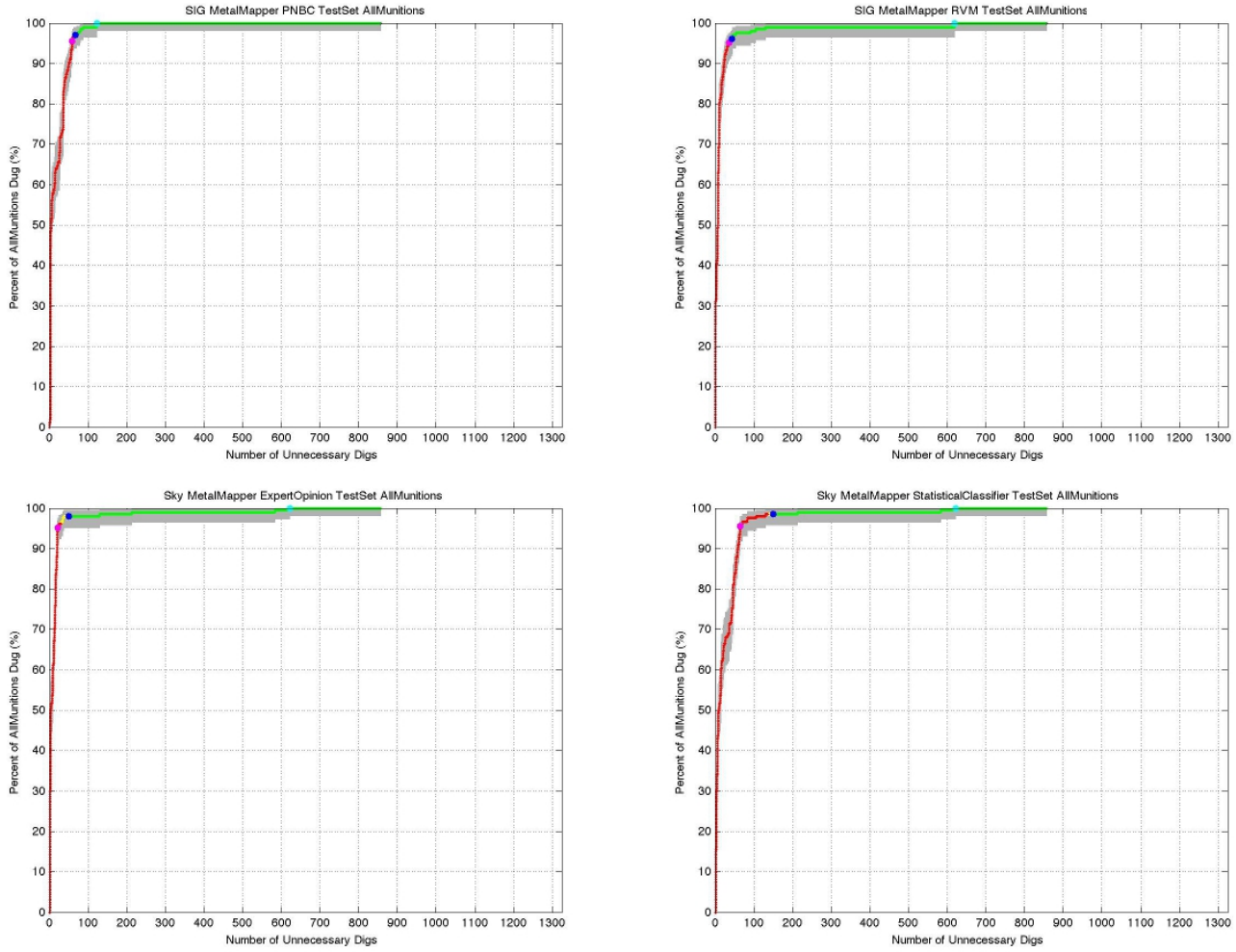


Fig. 13. Performance comparison, MetalMapper.

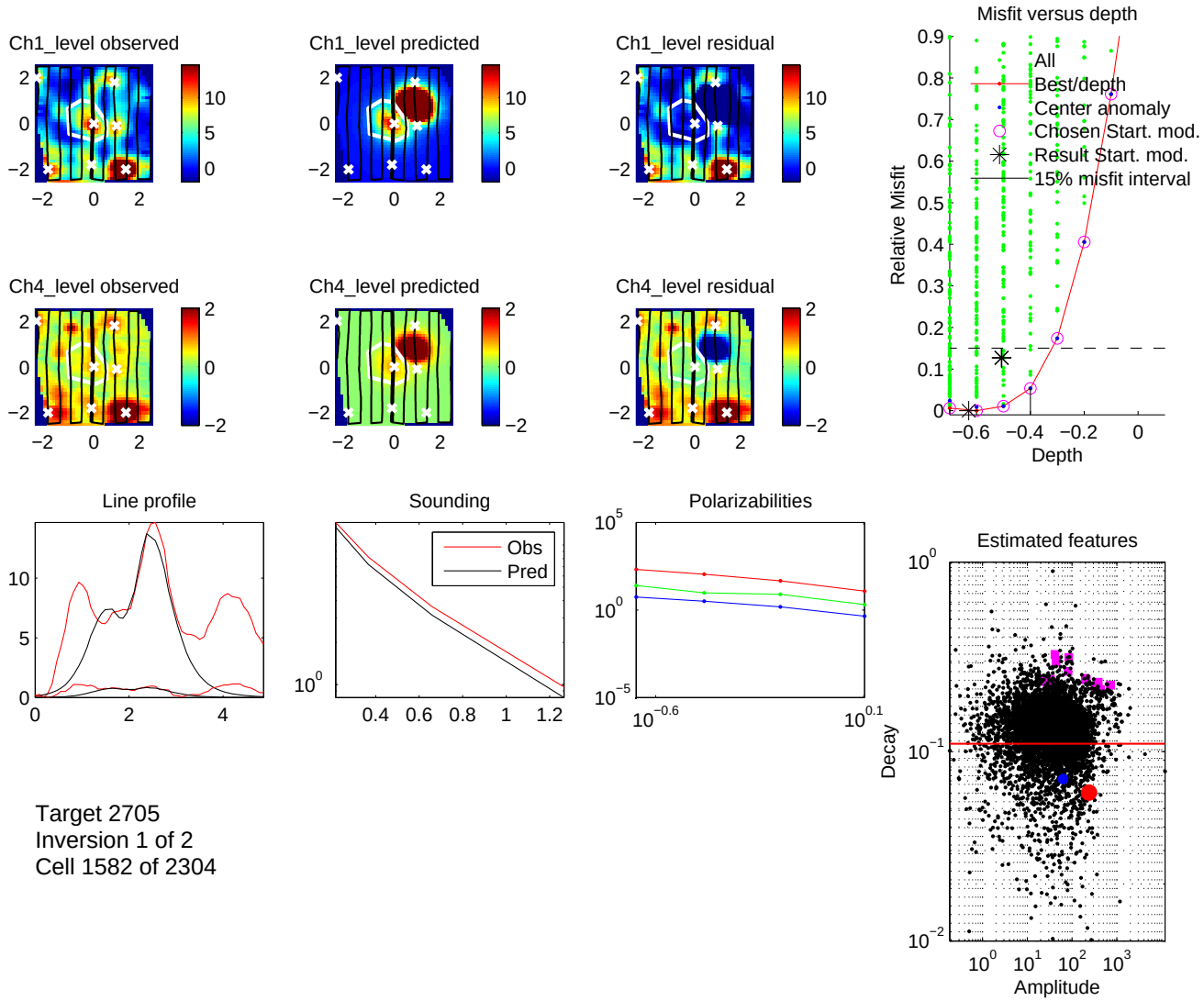


Fig. 14. EM61 QC display. In feature plot (bottom right) feature vector for current inversion is shown as a red circle, decay rate cutoff of 0.11 as a horizontal red line, and testpit features for TOI (37 mm, 105 mm, etc.) as magneta squares. Test data features are plotted as black dots.

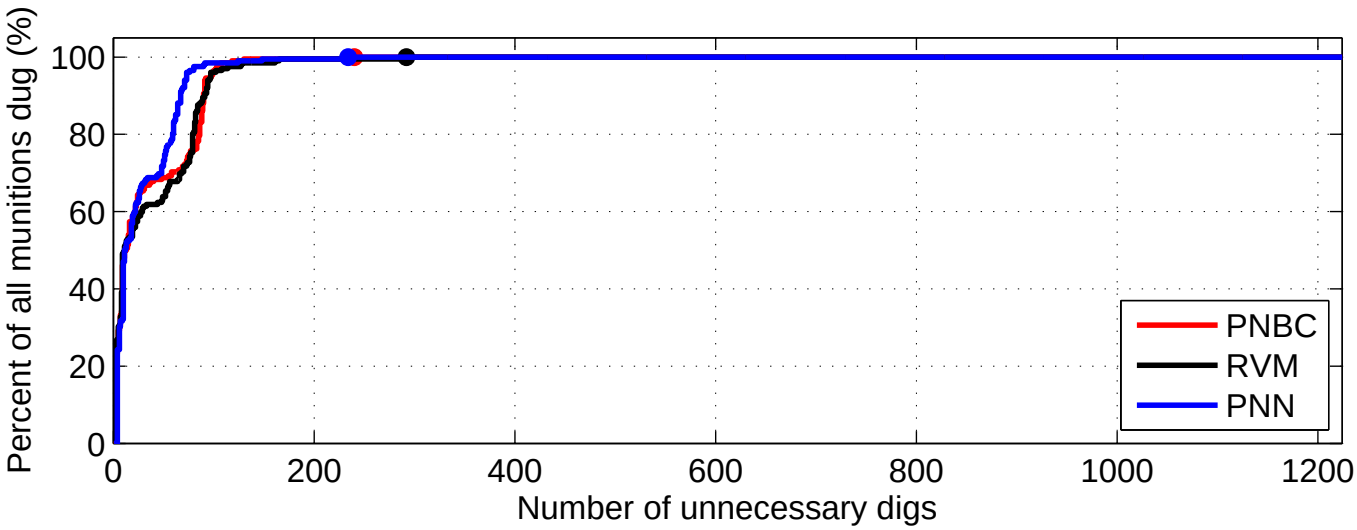


Fig. 15. Retrospective performance comparison for Duke (PNBC and RVM) and UBC (PNN) classifiers applied to Sky MetalMapper features (polarizability amplitude and decay). Circles indicate $Pd = 1$ point on each ROC.

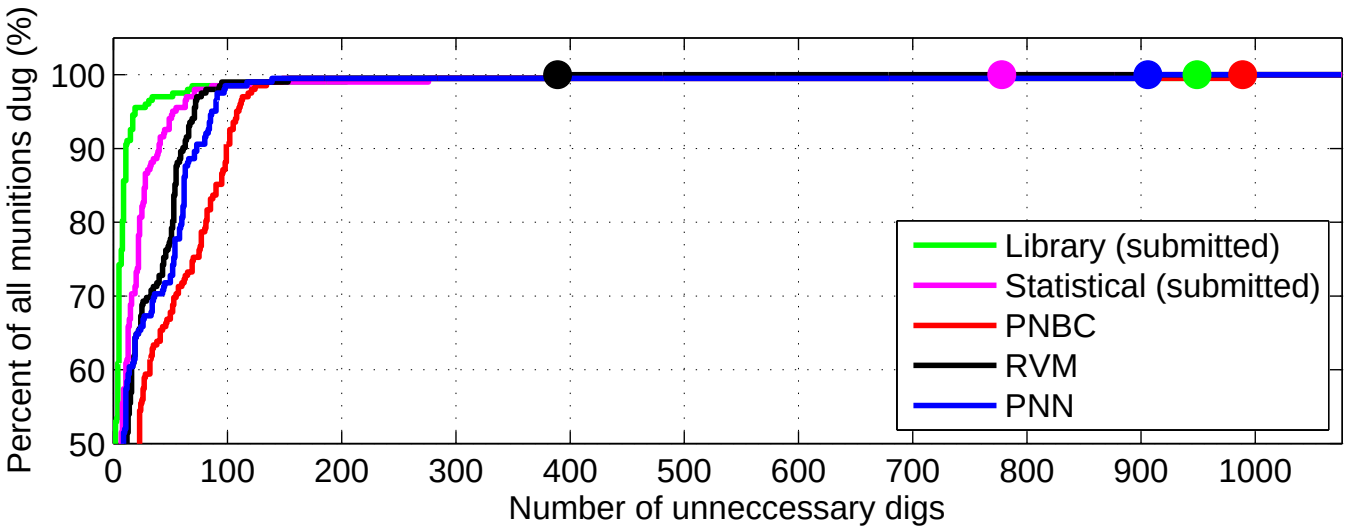


Fig. 16. Retrospective performance comparison for Duke (PNBC and RVM) and UBC (PNN) classifiers applied to Sky TEMTADS features (polarizability amplitude and decay). Submitted diglists are original diglists submitted to IDA. Circles indicate $Pd = 1$ point on each ROC.

III. Workflow development

Workflows are intended to guide a data analyst through the process of inversion and classification. While inversion procedures are well established for most data types, discrimination still depends upon the expertise of the analyst, with considerable variety in the features and algorithms employed by different research groups. For the SLO demonstration most groups achieved comparable performance when processing next generation sensor data. At Camp Sibert the relatively straightforward discrimination problem similarly produced excellent performance across most data types and discrimination approaches. These results suggest that well-constrained model parameters input into any sensible discrimination algorithm will yield good performance. Discrimination still requires care and a measure of expertise, but reliable features are the key to success.

The aim of discrimination workflow development in MR-1657 is then to transition classification expertise so that the "less than expert" data analyst can independently use a variety of advanced discrimination algorithms (statistical, rule-based, etc.) and achieve good results. There are two possible routes to implementation of practical and useful workflows. In a "top-down" approach we might develop functionality and GUI interfaces which can be used to chain together a series of operations (e.g. feature computation from model parameters, classifier training, etc.). Early iterations of our classification codes in UXOLab emphasized GUI functionality in an attempt to make discrimination accessible to all users in our group. However, this approach involved considerable coding effort and produced a package which, while functional, was not flexible enough to anticipate the processing steps necessary for different data sets. For the SLO demonstration we used a "bottom up" approach: Matlab scripts external to UXOlab were written to compute features, train classifiers and generate diglists. This was more efficient and flexible than working in a GUI, but the resulting scripts were poorly documented. A recurring problem in retrospective analysis was reproducing our original, submitted diglists. Data files and scripts are continually modified during processing and it is easy for the two to become incompatible. From these experiences we conclude that workflows should be

- Reproducible: a novice should be able to run a workflow on a data file and achieve the same result as an expert.
- Flexible: the user should be able to easily modify a template workflow to carry out any necessary operation.

To realize these properties for our MR-1657 work, we have started to implement a set of workflow scripts which automate the processing steps from feature computation to diglist generation. As for the SLO work, these scripts are external to the UXOLab GUI. To ensure reproducibility of original results during retrospective analyses, we have implemented functionality that stores all relevant workflow scripts within the associated data file, so that all processing steps applied to a data set are always stored within that data set. We have also started to rewrite our discrimination codes so that they are "decoupled" from the UXOLab GUI. Existing functions will still be accessible through the GUI interface, or they can be called directly in a script. The core discrimination codes have the following functionality:

- Feature computation: for each inverted target, a set of features is extracted from the observed data (for data-based features) and estimated model. These features are stored in a "Classification definition" structure which also keeps track of whether a given feature vector is associated with a passed inversion.

Multiple feature vectors may be present for any target to account for multiple passed inversion results and multi-object inversions.

- Class assignment: feature vectors associated with a labelled target are assigned a class based upon a standardized and intuitive numeric ID code (e.g. 60 mm mortars have $ID = 60$). A necessary step prior to classifier training is the merging of target classes to form new UXO and clutter classes.
- Feature plotting: any subset of features can be plotted, and features from pre-existing classification definitions (e.g. TOI features from past demonstrations) can be overlain. Plotting defaults for TOI classes are hard coded to ensure consistency when displaying results (see plotting conventions for TOI used throughout this report).
- Classifier training: classifier parameters are learned from the training data and stored in a "Classifier". This classifier can then be used to generate predictions for any test classification definition with the required features.
- Diglist generation: an ordered list of targets can be output using the predictions of any statistical or rule-based classifier. The diglist is divided into high confidence non-TOI, can't decide, and high confidence TOI categories based upon a user-specified threshold. Functionality to select this stop-digging threshold will also be implemented.

An important component of workflow implementation for MR-1657 is ensuring that features from older data sets remain up to date and accessible for classifier training. As we proceed to more realistic scenarios where training data become increasingly scarce, it is important to exploit all prior knowledge so that a previously encountered TOI can be readily detected. We have generated retrospective classification definitions from SLO data sets, and this has already proved valuable for the ongoing Camp Butner demonstration. In Figure 20 we show a classifier trained on synthetic TOI features (as described in the next section of this report) as well as using TOI features from SLO. For each TOI class we estimate a bivariate multimodal distribution, and we then weight the relative importance distributions based upon our prior expectation of finding targets in that class at Camp Butner (i.e. we weight the classifier to look for 37mm and 105mm before 60mm and 2.36" rockets).

Although our development of discrimination workflows is specific to UXOLab, this work will provide a template for discrimination functionality which will be implemented in the inversion and discrimination API being developed as part of ESTCP 1004.

IV. Numerical simulation of sensor data for input into feature extraction and classification

Obtaining a reliable and representative training data set is essential to the success of any discrimination algorithm. Statistical classifiers in particular require a training set from which a decision function can be learned. The behavior of this decision function in the tails (low probability regions) of the distribution of TOI determines the false alarm rate of the classifier at the $Pd = 1$ point on the ROC (i.e. where all TOI have been detected by the classifier). However, it is a difficult task to characterize the tails of any distribution from a finite sample, and this is especially true for UXO applications where training data sets are typically small. Active learning techniques have been developed by the Duke group (and others) to address this problem. These methods iteratively query the test data set by identifying feature vectors

which are most informative to the classifier. Here information is quantified by the Fisher information matrix, which is inversely proportional to the uncertainty in the classifier parameters. Semi-supervised discrimination algorithms also address the paucity of labelled data by using both test and training data to learn a decision function.

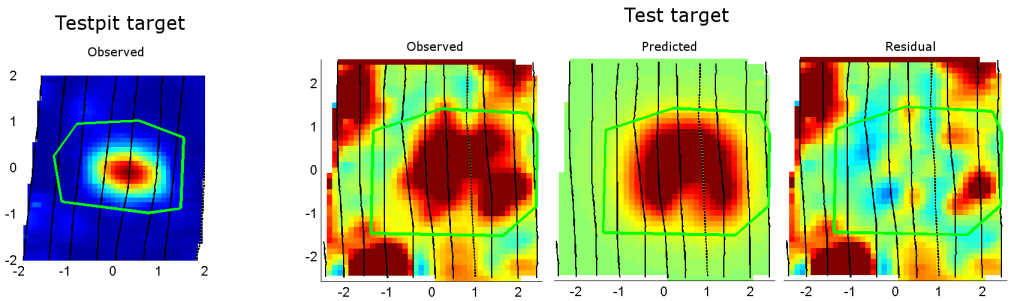
As part of MR-1657, we are tasked with developing numerical capabilities in UXOLab that allow the user to simulate data, from different surveys using different instruments, and to carry through with feature extraction and classification. As a first step in this work we have developed a technique which we call "simulated seeding" which attempts to simulate the variability of TOI features at a given site and to thereby supplement limited training data from testpit and IVS measurements.

Simulated seeding requires a set of testpit measurements over TOI items expected at a given site, as well as field measurements with targets picked and inverted. Ideally the testpit measurements will include each TOI type at the maximum detection depth required by the project and at the least favorable orientation (i.e. horizontal cross-track). Simulated seeding then involves the following steps:

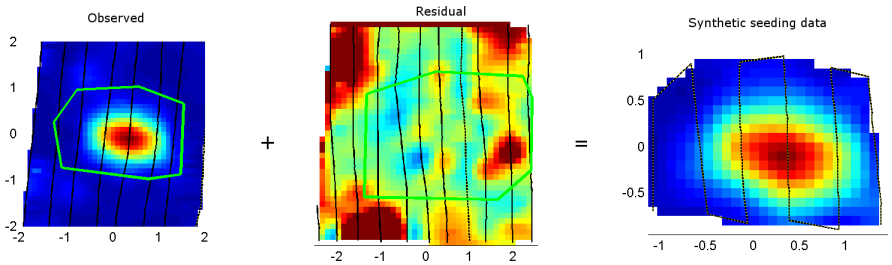
1) For a given testpit measurement, identify a field target (i.e. from the test data) with a mask completely enclosing the testpit mask.
2) Within the testpit mask, interpolate the testpit data at all channels to measurement locations over the field target.
3) Add the residuals from inversion of the field target data to the interpolated testpit data. At this step the residuals from the field data should not exceed a predefined percentage (e.g. 50 percent) of the maximum signal amplitude in the testpit data. This constraint is necessary to ensure that residuals from large amplitude field targets are not added to the data for small testpit items. If this constraint is not imposed then the residuals may completely swamp any signal, and the resulting target features from an inversion will be unrealistic.

These steps are illustrated in Figure 17.

This processing generates a number of synthetic data sets from each testpit measurement which can then be inverted to generate an ensemble of synthetic features for TOI. In step 2, we interpolate the testpit data to the observation locations in the field data. This reproduces the actual data coverage encountered in the field data and so may help to identify areas where data coverage does not support feature estimation via inversion. By adding the residuals from inversion of field data to the interpolated testpit data (step 3), we are in effect adding the actual noise (signal which cannot be explained by the dipole model) encountered in the field. The features generated by synthetic seeding are therefore likely to be more representative of the test data TOI features than simulations which rely on some assumed noise model and nominal data coverage. Figure 18(a) shows features estimated from 25 testpit datasets acquired with the EM61 cart at SLO. The feature space is spanned by our usual parameters representative of target size (abscissa) and decay rate (ordinate). While these features provide good initial information with which an active learning classifier might query the test data, they are certainly not sufficient for learning a decision function with a conventional supervised algorithm. Features estimated by synthetic seeding (Figure 18(b)) give a much better indication of the potential variability within each TOI class. Comparison of synthetic seeding features with the EM61 test data from SLO (Figure 18(c)) shows a good correspondence between the synthetic and actual distributions of TOI. We remark that the synthetic distribution of 60mm features



(a) We identify an inverted target in the field data with a mask larger in area than that of a testpit TOI, and with residuals not exceeding 50 percent of the maximum signal amplitude in the testpit data.

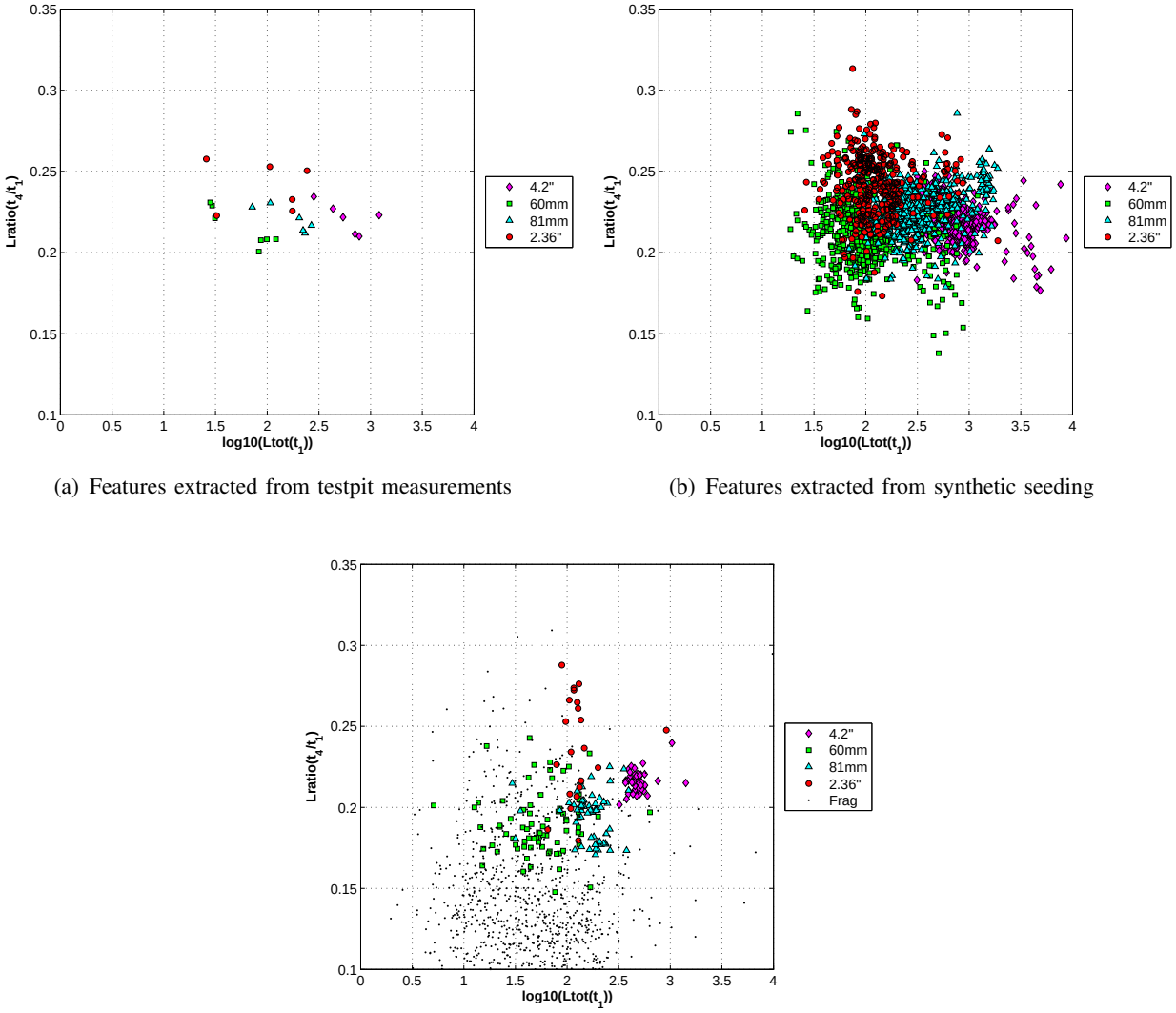


(b) The testpit observations are interpolated to measurement locations for the selected field target. The field target residuals are then added to the interpolated data to produce a synthetically seeded target. The mask for this synthetic data is the mask used for the original testpit data.

Fig. 17. Schematic of simulated seeding processing

anticipates some of the faster-decaying (smaller Lratio) outliers to the actual (test) distribution. Some care should be taken in synthetic seeding to ensure that a given testpit measurement is not overemphasized in the simulations, thereby skewing the resulting mode of the synthetic distribution. This can be done by verifying that no single testpit item dominates the population of synthetic seedings for its target class.

Figure 19 shows the results of the same procedure applied to the Camp Butner EM61 cart data. Here synthetic seeding allowed us to fully exploit the information in the testpit data and to avoid requesting training data altogether. Of course, this can be risky if there are ordnance types in the test data which are not in the set of testpit items. For this reason at Camp Butner we dig all slow-decaying targets above a threshold on the decay parameter ($Lratio > 0.11$). In this case we selected a very cautious threshold which is likely to identify the majority of TOI, even if they do not occur in the testpit set. Targets are prioritized above the threshold on the basis of a classifier trained on synthetic seeding distributions (Figure 20). This will (hopefully) have the effect of improving the area under the curve of the resulting ROC. Interestingly, there are no obvious clusters corresponding to TOI in the Butner EM61 test data (Figure 20(b)). This suggests that TOI comprise a much smaller proportion of targets at Camp Butner than at SLO. Another possibility is that the data are insufficient to constrain the model parameters, so that no clusters occur. This could actually be the case for Camp Butner where the TOI are much smaller than at SLO.



(a) Features extracted from testpit measurements

(b) Features extracted from synthetic seeding

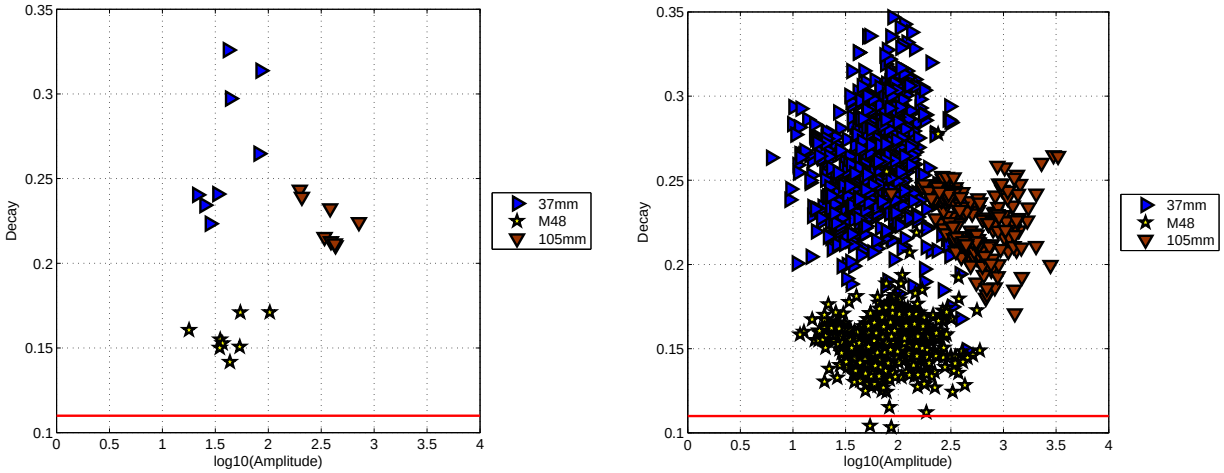
(c) Test data features

Fig. 18. Synthetic seeding results, SLO EM61 cart data.

V. Spatially-Constrained Classifiers for UXO Detection

An important part of MR-1657 is to investigate the utility of spatial information in UXO detection. The team at Duke University takes a leading role in this task. Our major findings are that the use of spatial information can significantly reduce the number of false alarms. The subsequent sections describe the work we have accomplished on this task.

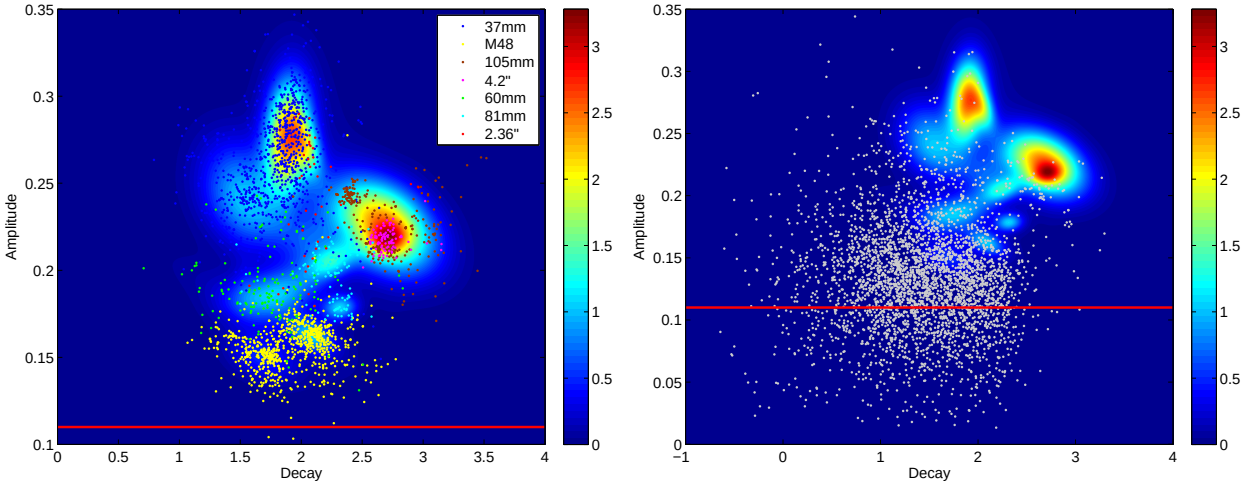
We present an approach to exploiting spatial information to improve detection of unexploded ordnance (UXO). The approach is motivated by the observations that UXO targets tend to be proximate to each other in the spatial domain and that they are significantly outnumbered by the clutter items. Our approach aims to exploit the correlation between the spatial location and the class label and yet emphasize that spatial locations are not characteristic features and cannot independently realize a correct classification. Such a goal leads to classification models in which the decision boundary is primarily realized in the feature space, with the spatial locations used as auxiliary information to constrain the parameter space of the feature-domain model, and help to find the optimal parameters. We give two formulations of such



(a) Testpit data.

(b) Simulated seeding

Fig. 19. Testpit and simulated seeding training features for Camp Butner EM61 data. Horizontal red line shows decay rate cutoff applied to test data.



(a) Training data, including features from synthetic seeding and test data TOI from SLO.

(b) Test data

Fig. 20. Non-parametric classifier surface for discrimination of Camp Butner EM61 data. Horizontal red line indicates decay rate cutoff applied in generation of diglist - targets above the cutoff are labelled high confidence TOI.

spatially-constrained classifiers, and develop algorithms to learn their parameters in a semi-supervised fashion, by using labeled data from previous UXO sites as well as unlabeled data from the test site. We consider both offline learning wherein the labels of test items remain unknown during learning, and online learning wherein an item declared as UXO is excavated to reveal its true label, which is used to refine the classifier before the next declaration is made. The results on real UXO data demonstrate that, the use of spatial information leads to significant reduction of false alarms, and at the same time achieving a complete detection of all UXO targets at the test site.

VI. Motivation of Spatially-Constrained Classification

Unexploded ordnance (UXO) are explosive munitions that failed to detonate at the time of being employed [10]. They are left intact or partially intact, posing a risk of explosion even long time after

they were fired. Unexploded ordnance are typically found on or below the surface of the earth [11], and they constitute a serious environmental contamination around the world. Because of the hazardous nature of UXOs, the contaminated lands demand thorough cleanup before they can be returned to public use or private ownership [12], [13].

Disposing of surface UXOs (i.e., those existing on the surface) is relatively easy, usually involving a manual inspection to detect the targets, followed by a rigorous procedure of removing them [14]. Subsurface UXOs, however, pose a much more challenging task for both detection and removal. Because these UXOs are buried, as deep as a meter below the surface, they require specialized sensors to detect and locate them. Once a UXO target has been found and the location determined, the target must be excavated before it can be disarmed and removed. Excavation of a UXO target is a costly procedure because of the potential risk of explosion; it is also a time-consuming process, because one must proceed slowly and carefully to avoid disturbing the target. Therefore it is important to ensure that most excavations turn out to be UXO targets such that the cost associated with excavating false targets is minimized. While the cost of excavating a false target is high, that of leaving a true UXO undetected is even higher given the hazard it can bring [12]. Thus it is imperative to seek detection technologies that can detect all UXO targets at the minimum false alarm rate.

A key component dictating the UXO detection performance is the sensor used to collect the data about the potential targets. Over the past decades, a number of sensing techniques have been developed for detecting subsurface UXOs, among which magnetometers [15] and electromagnetic induction (EMI) sensors [16], [17] have been most widely used. Magnetometers can effectively detect ferrous targets, by measuring the perturbations that these targets introduce to the static magnetic field of the Earth. Typical UXOs are rich in ferrous contents and can thus be detected by magnetometers. However, many clutter items (fragments left from munition operations or other metallic debris) also have rich ferrous contents, and these clutter are easily confused with UXO targets by magnetometers, contributing a large set of false alarms. EMI sensors provide more detailed information about general conducting targets, by sending wideband signals to the targets and measuring their responses. The wideband EMI responses prove more competent for discriminating between UXO targets and non-UXO conducting targets [18], than the static magnetic measurements provided by magnetometers.

Typical UXO detectors do no operate directly on raw measured data but on the parameters of a physical model that best explain the raw data. A number of models have been developed for the data measured by magnetometers [19] or EMI sensors [16], [17], [18], [20], based primarily on a dipole approximation, although higher order approximations have also been considered [21]. The parameters of these models have been shown to carry information that are relevant to distinguishing UXO targets from clutter [14], [18]. The EMI data have a subset of parameters in common with the magnetometer data, and the common parameters include the target spatial location, the utility of which will be investigated in this study. The commonality can be exploited to improve the accuracy of the parameters, by using the parameters inverted from one data type to constrain the inversion of another [18], [22], or by inverting both data types jointly to obtain agreements on the common parameters [22].

The parameters inverted from the magnetometer and/or EMI data yield a compact representation of the targets, and various approaches have been proposed to address the problem of detecting UXO based on

using these parameters [23], [24], [25], [26], [14], [27], [28], [29]. These approaches typically select a subset of parameters to constitute a feature vector, which represents each interrogated target as a point in a Euclidean space (i.e., the feature space) [30]. Some of the parameters included in the feature vector may eventually be found irrelevant to the detection task by the algorithms, but which parameters are irrelevant are not known *a priori* and therefore all potentially relevant ones must be included to allow their relevances to be examined. Some other parameters, however, are *knowingly* irrelevant, and these are treated as nuisance parameters and are excluded from the feature vector from the beginning. The spatial coordinates of a target are usually treated as nuisance parameters, because they provide no essential information about whether the target is UXO or clutter, and because their values depend on the coordinate system employed at a specific site, thus do not have consistent meanings across different UXO sites (because the coordinate system changes from site to site).

In this study we demonstrate that, although spatial locations are not appropriate to be used as features, they provide important information that can be exploited to enhance the performance of a UXO detector. The work here is motivated by two observations: (*i*) the UXO targets at a given site are typically significantly outnumbered by the clutter items, and (*ii*) the UXO targets tend to form *noisy clusters* in the spatial space, where a noisy spatial cluster is defined as a spatial region containing UXO targets and possibly also clutter items. In intuitive words, we may equivalently say that the UXO targets are rare and spatially proximate. The rarity arises from the fact that most munitions targeted at a given site successfully detonated at the time of being employed, leaving only a small fraction of them unexploded, and the fact that the fragments resulting from the exploded munitions and other natural metallic debris contribute a large number of clutter items. The spatial proximity may be explained by that a UXO target found at a location indicates an attack to the neighborhood, leading to an increased chance of seeing more unexploded munitions at nearby locations. Motivated by the above observations, we propose to formulate UXO detection as a problem of binary classification based on using both feature vectors and spatial locations. For the reasons stated below, we employ a customized structure to combine the feature vector and the corresponding location, instead of concatenating the two and building a model for the concatenation.

The spatial locations of targets are site specific parameters, therefore one must model the target locations for each individual site separately, and the model built at one site cannot be used at another site. Feature vectors, by contrast, represent the targets' fundamental properties, therefore they are relatively site-insensitive, although geological change may introduce small variations in the features. As a result, one can build a model for the feature vectors collected across different sites. Moreover, a feature vector may not necessarily be associated with a spatial location – this happens, for example, when the UXO targets are measured (and features are extracted) in a laboratory. Because of these reasons, we propose models which assume that the feature vector of a target and the associated spatial location are respectively governed by two sub-models, with the sub-models linked together through the class label, a binary indicator variable indicating whether the target is UXO or clutter. Under such an assumption, the feature vector and the location is conditionally independent given the class label, which expresses the intuition that the features, of either UXO or clutter, are insensitive to spatial changes. Thus, any dependence between the feature vector and the location must be explained by the class label, a property consistent with the spatial

proximity of UXO targets.

The conditional independence makes our models akin to the co-training models [31], [32], [33] proposed for semi-supervised learning. In terms of co-training, the feature vector and the spatial location correspond to two different views of a target. The two views in co-training are assumed to be each adequate for realizing a correct classification of the targets, therefore they play symmetric roles, with the model of one view furnishing noisy labels to improve the model of another. In the UXO detection, however, the spatial locations cannot independently realize a correct classification, because clutter are ubiquitous and thus the spatial boundary between the UXO and the clutter is difficult to realize or even to define. On the contrary, as discussed earlier, the feature vectors provided by EMI sensors and magnetometers, possess adequate discriminating power to realize a correct classification and thus they play the primary role in defining the decision boundary, with the spatial locations providing relevant information to help to find the decision boundary. For these reasons, the models we propose here are better described as spatially-constrained classifiers in the feature space, where the spatial information is used to constrain the parameter space of the feature-domain model and facilitate search for the optimal parameters.

We give two formulations of such spatially constrained classifiers, in both of which the feature vectors and the spatial locations are governed, respectively, by a mixture of Gaussian distributions, with the class labels serving as links between the feature-domain mixture and the spatial-domain mixture. Both of the two models yield a conditional distribution of the label given the feature vector and/or the spatial location, which classifies the targets into two classes: UXO or clutter. However, the second model also employs a discriminative classifier within each component in the feature-domain mixture, to constitute a more accurate UXO-vs-clutter boundary in the feature space. We consider semi-supervised learning, seeking maximum likelihood (ML) estimates of the model parameters using both training data and test data, where the training data consist of labeled feature vectors alone and the test data consist of unlabeled feature vectors along with the associated spatial locations. The training feature vectors either have no associated locations (when they are collected from laboratories), or their locations are based on coordinate systems inconsistent with that of the test site (when they are from previous UXO sites). We show that likelihood maximization on the test data effectively enforces the spatial constraints by maximizing the label-consistency between the feature-domain mixture model and the spatial-domain mixture model. We consider two learning scenarios: offline learning and online learning. In offline learning, the test items declared as UXO are not excavated until all items at the test site have been declared, thus the labels of test items remain unknown during learning. In online learning, a test item declared as UXO is excavated right away to reveal its true label, and the true label, along with the feature vector and the location of the item, is added to the training set to retrain the model. The retrained model is then used to make the next declaration.

The subsequent sections are organized as follows. In Section VII-A we present the model in which the feature-domain classifier is implemented by a Gaussian mixture model (GMM), with this model appropriate for use when initially no training examples are available for the clutter. In Section VII-B, we replace the feature-domain GMM with a discriminative classifier to allow a more accurate representation of the decision boundaries. The experimental results are reported in Section IX based on Camp Lejeune data and Sibert data. Discussions and conclusions are provided in Section X.

VII. Spatially-Constrained Classification Models

We construct classification models based on the feature vectors and as well as the spatial locations, with the feature vectors playing the primary role in defining the decision boundary between the UXO class and the clutter class, and the associated spatial locations playing the role of constraining the feature-domain decision boundary to yield improved classification.

We assume access to a set of labeled feature vectors before we embark on the task of detecting the UXO at the test site. The labeled feature vectors may be collected from previous UXO sites or from laboratories. The UXO class is relatively well defined, with the constituent targets assumed known and their feature vectors robust to the change of geological conditions. The clutter class, in contrast, exhibits greater variability, consisting of the fragments left from exploded munitions, natural metal debris, and virtually all metal items that are not UXO. Thus it is more difficult to constitute a set of training feature vectors for the clutter than for the UXO.

In addition to the labeled feature vectors, we assume to have a set of unlabeled feature vectors and the associated spatial locations, collected from the site currently under test. These unlabeled examples represent all targets of potential interest and the prediction of their labels constitutes the detection task. That all the data at the test site can be collected at once before the excavation begins is based on the assumption that all surface UXO targets have been removed and therefore, the remaining targets are buried and do not pose danger to a cart-based system used to collect the data [15].

For ease of exposition, we denote by $\mathcal{D}^r$ the set of labeled data from previous sites and denote by $\mathcal{D}^e$ the unlabeled data from current test site. We can further write $\mathcal{D}^r = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N^r}$, $\mathcal{D}^e = \{(\mathbf{x}_i, \mathbf{s}_i)\}_{i=N^r+1}^{N^r+N^e}$, where $\mathbf{x}$ is a feature vector with associated spatial location $\mathbf{s}$, and $y \in \{0, 1\}$ is the corresponding class label with $y = 1$ indicating the UXO class and $y = 0$ indicating the clutter class. We have numbered the data examples in $\mathcal{D}^r$ and $\mathcal{D}^e$ consecutively, assuming that the examples in $\mathcal{D}^r$ precede those in $\mathcal{D}^e$. The examples in $\mathcal{D}^r$ may also have associated spatial locations, but which have been purposely removed because they are based on coordinate systems different from that used at the current site. The labels in $\mathcal{D}^e$ are the missing variables that one aims to predict.

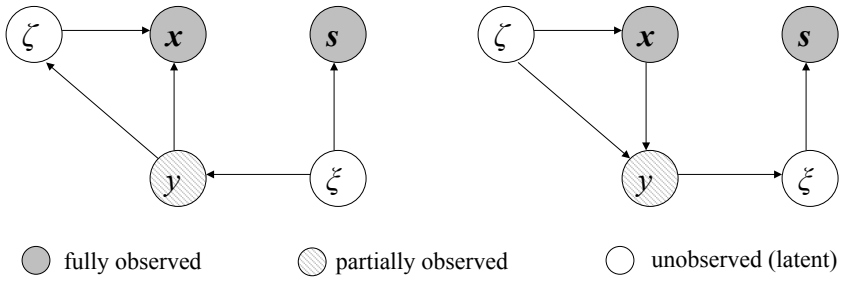


Fig. 21. Graphical Models of, (left) the spatially-constrained Gaussian mixture models (SC-GMM), and (right) the spatially-constrained and quadratically gated mixture of experts (SC-QGME). A shaded circle denotes a variable that is observed on all examples (fully observed), a partly shaded circle denotes a variable that is observed on some examples but unobserved on others (partially observed), a hollow circle denotes a variable that is unobserved on all examples (latent).

We propose two models to represent the joint probability of $\mathcal{D}^r$ and $\mathcal{D}^e$. In the first one, the feature vectors of each class are governed by an independent Gaussian mixture model (GMM) whose parameters are specific to the class label, while in the second, the feature vectors and the associated labels are jointly

governed by a quadratically gated mixture of experts (QGME) [34]. In both models, the spatial locations of the UXO and those of the clutter are respectively governed by two GMMs with shared mixture components but label-specific proportions. The two models are represented graphically in Figure 21, where each circle denotes a variable with the strength of shading indicating the degree of observability of the variable, an arrow denotes direct dependence of the head variable on the tail variable, and ξ and ζ are discrete latent variables with ξ indicating the spatial-domain mixture components and ζ indicating the feature-domain mixture components.

It is seen that the two models have the same set of variables, but the dependencies between the variables are different. In the first model, the feature-domain GMM for UXO is independent of that for clutter, thus the two GMMs can be learnt independently, using the respective training (labeled) feature vectors in $\mathcal{D}^r$. This is appealing when it is difficult to constitute a set of feature vectors for clutter (as mentioned earlier), because the absence of clutter feature vectors does not affect updating the GMM for UXO (which requires only the UXO feature vectors). The second model learns a mixture of experts in the feature-domain, each expert implementing a local discriminative classifier between UXO and clutter in a particularl region of the feature space. Training of the classifiers is based on maximizing the discrepancy between the two classes, which requires the training set $\mathcal{D}^r$ to contain feature vectors from both the UXO class and the clutter class. For both models, the unobserved labels of the test data are treated as latent variables which, along with other latent variables, are handled by expectation maximization (EM) [35].

A. Spatially-Constrained Gaussian Mixture Models

In the first model, the labeled data (feature vectors) are assumed to be governed by two independent GMMs, respectively associated with the class of UXO and the class of clutter,

$$p(\mathbf{x}_i|y_i) = \sum_{\zeta_i=1}^{K_{y_i}} p(\mathbf{x}_i|\zeta_i, y_i)p(\zeta_i|y_i), \tag{7}$$

for $i = 1, 2, \cdots, N^r$, with

$$\begin{aligned} p(\zeta = k|y = t) &= \pi_{kt}, \\ p(\mathbf{x}_i|\zeta_i = k, y_i = t) &= \frac{1}{\sqrt{2\pi|\Sigma_{kt}|}} e^{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_{kt})^T \boldsymbol{\Sigma}_{kt}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_{kt})}, \end{aligned}$$

where $|\boldsymbol{\Sigma}_{kt}|$ denotes the determinant of the matrix $\boldsymbol{\Sigma}_{kt}$ and the superscript T denotes matrix transpose. Given the label, the feature vectors are distributed according to a Gaussian mixture model, with its parameters specific to the label.

The unlabeled data (feature vectors along with the associated spatial locations) are assumed to be governed by

$$p(\mathbf{x}_i, \mathbf{s}_i) = \sum_{y_i=0}^{1} p(\mathbf{x}_i|y_i) \sum_{\xi_i=1}^{J} p(y_i|\xi_i)p(\mathbf{s}_i|\xi_i)p(\xi_i), \tag{8}$$

for $i = N^r + 1, N^r + 2, \cdots, N^r + N^e$, where $p(\mathbf{x}_i|y_i)$ is the same as defined in (7), and

$$p(y_i = t|\xi_i = j) = \theta_{tj},$$

$$\begin{aligned} p(\xi_i = j) &= \eta_j, \\ p(\mathbf{s}_i|\xi_i = j) &= \frac{1}{\sqrt{2\pi|\Omega_j|}} e^{-\frac{1}{2}(\mathbf{s}_i - \boldsymbol{\beta}_j)^T \boldsymbol{\Omega}_j^{-1}(\mathbf{s}_i - \boldsymbol{\beta}_j)}, \end{aligned}$$

The model in (8) is seen to be composed of one spatial-domain GMM and two label-dependent feature-domain GMMs. The spatial-domain GMM is obtained by integrating out $\mathbf{x}_i$ in (8),

$$p(\mathbf{s}_i) = \int p(\mathbf{x}_i, \mathbf{s}_i) d\mathbf{x}_i = \sum_{\xi_i=1}^{J} p(\mathbf{s}_i|\xi_i) p(\xi_i), \tag{9}$$

which probabilistically partitions the spatial space into J regions, with the probability of any spatial location $\mathbf{s}_i$ belonging to region j given by

$$p(\xi_i = j|\mathbf{s}_i) = \frac{p(\mathbf{s}_i|\xi_i = j) p(\xi_i = j)}{p(\mathbf{s}_i)}. \tag{10}$$

The two label-dependent feature-domain GMMs, each conditional on a specific class label, are given in (7). The unconditional feature-domain GMM can be obtained by integrating out $\mathbf{s}_i$ in (8),

$$p(\mathbf{x}_i) = \sum_{y_i=0}^{1} \sum_{\zeta_i=1}^{K_{y_i}} p(\mathbf{x}_i|\zeta_i, y_i) p(\zeta_i, y_i), \tag{11}$$

which consists of $K_0 + K_1$ components, with the mixing proportions given by

$$\begin{aligned} p(\zeta_i, y_i) &= p(\zeta_i|y_i) p(y_i) \\ &= p(\zeta_i|y_i) \sum_{\xi_i=1}^{J} p(y_i|\xi_i) p(\xi_i). \end{aligned}$$

Since the marginal distribution of $\mathbf{x}_i$ and $\mathbf{s}_i$ are both GMMs, one may conjecture that their joint distribution is also a GMM. This is indeed true, as can be verified by substituting (7) into (8) and rearranging the terms under the summations, which yields a GMM with $2J(K_0+K_1)$ components, where each component Gaussian distribution has a block-diagonal covariance matrix reflecting that $\mathbf{x}_i$ and $\mathbf{s}_i$ are conditional independent given (y_i, ζ_i, ξ_i). As a matter of fact, one does not have to condition on ζ_i or ξ_i to obtain the independence between $\mathbf{x}_i$ and $\mathbf{s}_i$, since they become independent of each other as long as y_i is observed, as discussed further below.

The spatial-domain GMM and the feature-domain GMMs are connected through a set of Bernoulli distributions $p(y_i|\xi_i = j)$, $j = 1, 2, \cdots, J$, with $p(y_i = 1|\xi_i = j)$ defining the probability that UXO occurs in spatial region j. This connection induces the following dependence between the feature vector and the associated spatial location,

$$p(\mathbf{x}_i|\mathbf{s}_i) = \sum_{y_i=0}^{1} \sum_{\zeta_i=1}^{K_{y_i}} p(\mathbf{x}_i|\zeta_i, y_i) p(\zeta_i, y_i|\mathbf{s}_i), \tag{12}$$

where

$$p(\zeta_i, y_i|\mathbf{s}_i) = p(\zeta_i|y_i) p(y_i|\mathbf{s}_i)$$

$$= \quad p(\zeta_i|y_i)\sum_{\xi_i=1}^{J} p(y_i|\xi_i)p(\xi_i|\mathbf{s}_i), \tag{13}$$

with $p(\xi_i|\mathbf{s}_i)$ given in (10). It is clear from (12) that, given the spatial location $\mathbf{s}_i$ of target i, its feature vector $\mathbf{x}_i$ is governed by a mixture of K_0+K_1 Gaussian distributions. The dependence of the feature vector on its spatial location is solely manifested in the mixing proportions which has a location-dependency; each of the $K_0 + K_1$ component distributions is location-independent.

The location-dependent proportions in (12) play a role similar to that of the gating network in a mixture of experts [36]. However, each expert here is a Gaussian distribution representing the density of feature vectors in a local spatial region, and thus, the experts operate in the feature domain but the gating network operates in the spatial domain. This is different from a standard mixture of experts, in which the experts and the gating network operate in the same space. The spatial gating network, represented in (13), allocates target i to a feature-domain Gaussian expert by first assigning it to the GMM for UXO or the GMM for clutter, i.e., $p(\mathbf{x}_i|y_i = t)$ with $t \in \{0, 1\}$, and then assigning it to a particular component of the assigned GMM. The first assignment, represented by $p(y_i|\mathbf{s}_i)$, is obtained by summing over all probable spatial regions that the target might reside in, taking into account the proportions of UXO and clutter in each probable spatial region. The second assignment is directly made by using $p(\zeta_i = k|y_i = t) = \pi_{kt}$. Clearly, the spatial location $\mathbf{s}_i$ only provide information for the first assignment, but it does not provide the detailed information as to which particular Gaussian expert to use, the latter assignment determined solely by π_{kt}. This is consistent with the fact that the spatial variation of target feature vectors is mainly contributed by the spatial variation of the associated class labels. For a given class label (UXO or clutter), the variation of feature vectors is spatially independent. It is here assumed that the target features are relatively insensitive to the changes in soil conditions and other geological factors, and any variation arising from such changes can be ignored.

Focusing on the first assignment, which involves using the spatial information, one actually can gain further insight into the model. Integrating out ζ_i in (12), one obtains

$$p(\mathbf{x}_i|\mathbf{s}_i) = \sum_{y_i=0}^{1} p(\mathbf{x}_i|y_i)p(y_i|\mathbf{s}_i), \tag{14}$$

from which it is clear that the dependence of a feature vector on the associated spatial location is solely contributed by the non-observability of the class label. If the label is observed, the two becomes independent of each other. The dependence of $\mathbf{x}_i$ on $\mathbf{s}_i$ reflects the degree to which $p(\mathbf{x}_i|y_i)$ and $p(y_i|\mathbf{s}_i)$ agrees with each other on the unobserved label y_i. The dependence is strong when $p(\mathbf{x}_i|y_i)$ and $p(y_i|\mathbf{s}_i)$ match each other, i.e., they have similar shapes when each of them is viewed as a function in y_i. This requires that small $p(\mathbf{x}_i|y_i)$ implies small $p(y_i|\mathbf{s}_i)$ and large $p(\mathbf{x}_i|y_i)$ implies large $p(y_i|\mathbf{s}_i)$, and vice versa, for any given $y_i \in \{0, 1\}$. Recall that $p(\mathbf{x}_i|y_i = t)$, $t \in \{0, 1\}$, are the feature-domain GMMs, and $p(y_i|\mathbf{s}_i)$ arises from the spatial-domain GMM. Thus, the dependence between $\mathbf{x}_i$ and $\mathbf{s}_i$ is in essence an indicator of the label-consistency between the feature-domain sub-model and the feature-domain sub-model, and maximization of the dependence effectively enforces the label-consistency between the two sub-models.

The idea of enforcing consistency on common latent variables between multiple sub-models has been used in various applications. For example, [37] considered approximating an intractable model as a product

of multiple tractable sub-models, and train the sub-models jointly such that they agree on the a set of common latent hidden variables and thus the integrity of the overall model is preserved. The sub-model in [37] all live in the same data space, each capturing a particular aspect of the overall model. The co-training models (and more generally, multi-view models) [31], [32], [33] consider two or more sub-models, each living in a distinct data space, corresponding to a particular view of the complete data space. In co-training, all sub-models are equally important, each assumed to be adequate for achieving the learning objective (for example, realizing a correct classification of the targets). Our model described above is different from the model in [37] in that the sub-models in our case live in different spaces, one in the feature space and the other in the spatial space. Thus our model is more related to co-training models. However, the two sub-models in our case do not play symmetric roles: the feature-domain sub-model plays the primary role in realizing the decision boundary between UXO and clutter, with the spatial-domain sub-model playing the role of constraining the parameter space of the feature-domain sub-model and facilitating the search for the optimal parameters. To emphasize this fact, the model specified in (8) is referred to as a *spatially-constrained GMM* (SC-GMM).

As discussed above, SC-GMM enforces the consistency on unobserved labels by maximizing the dependence between $\mathbf{x}_i$ and $\mathbf{s}_i$. Instead of directly maximizing $p(\mathbf{x}_i|\mathbf{s}_i)$, we maximize $p(\mathbf{x}_i, \mathbf{s}_i)$. This gives us the convenience of obtaining a simple and computationally efficient algorithm. In addition, this is also necessary given that the parameters of $p(\mathbf{s}_i)$ is a subset of the parameters of $p(\mathbf{x}_i|\mathbf{s}_i)$. To make it precise, we introduce the notations: $\Theta_{f0} = \{\pi_{kt}, \boldsymbol{\mu}_{kt}, \boldsymbol{\Sigma}_{kt} : k = 1, 2, \cdots, K_t,\ t = 0\}$, $\Theta_{f1} = \{\pi_{kt}, \boldsymbol{\mu}_{kt}, \boldsymbol{\Sigma}_{kt} : k = 1, 2, \cdots, K_t,\ t = 1\}$, $\Theta_s = \{\eta_j, \boldsymbol{\beta}_j, \boldsymbol{\Omega}_j : j = 1, 2, \cdots, J\}$, and $\Theta_{sf} = \{\theta_{tj} : j = 1, 2, \cdots, J,\ t = 0, 1\}$, where Θ_{f0}, Θ_{f1}, Θ_s, and Θ_{sf} denote the parameters for $p(\mathbf{x}_i|y_i = 0)$ (the feature-domain GMM for clutter), the parameters for $p(\mathbf{x}_i|y_i = 1)$ (the feature-domain GMM for UXO), the parameters for $p(\mathbf{s}_i)$ (the spatial-domain GMM), and the parameters establishing the connection between the feature-domain GMMs and spatial-domain GMM, respectively. Clearly, the parameters of $p(\mathbf{x}_i, \mathbf{s}_i)$ are $\Theta = \Theta_{f0} \cup \Theta_{f1} \cup \Theta_s \cup \Theta_{sf}$. Since $p(\mathbf{x}_i|\mathbf{s}_i) = p(\mathbf{x}_i, \mathbf{s}_i)/p(\mathbf{s}_i)$, the parameters of $p(\mathbf{x}_i|\mathbf{s}_i)$ are Θ, which contain Θ_s, the parameters of $p(\mathbf{s}_i)$. When considering maximum likelihood (ML) estimation of $(\Theta_f, \Theta_s, \Theta_{sf})$, the likelihood function used in the estimation must take into account all the information that observed data carry about these parameters. Suppose we have data $\{\mathbf{x}_i, \mathbf{s}_i\}$ and desire an ML estimate of Θ. Clearly, using $\prod_i p(\mathbf{x}_i|\mathbf{s}_i)$ as the likelihood function ignores the information provided by $\prod_i p(\mathbf{s}_i)$, which has a dependence on Θ_s. The likelihood function satisfying the complete-information condition is therefore $\prod_i p(\mathbf{x}_i, \mathbf{s}_i) = \prod_i p(\mathbf{x}_i|\mathbf{s}_i)p(\mathbf{s}_i)$.

For UXO detection, we have both labeled data $\mathcal{D}^r$ from previous sites and unlabeled data $\mathcal{D}^e$ from the test site. Following the notational conventions introduced earlier, we can write $\mathcal{D}^r = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N^r}$, and $\mathcal{D}^e = \{(\mathbf{x}_i, \mathbf{s}_i)\}_{i=N^r+1}^{N^r+N^e}$. Note that the labeled data have only feature vectors – the associated spatial locations are purposely removed because they are based on coordinate systems different from that used at the current site. Given $\mathcal{D}^r$ and $\mathcal{D}^e$, the likelihood function satisfying the complete-information condition is

$$p(\mathcal{D}^r, \mathcal{D}^e|\Theta) = \prod_{i=1}^{i=N^r} p(\mathbf{x}_i|y_i) \prod_{i=N^r+1}^{i=N^r+N^t} \sum_{y_i=0}^{1} p(\mathbf{x}_i|y_i) \sum_{\xi_i=1}^{J} p(y_i|\xi_i)p(\mathbf{s}_i|\xi_i)p(\xi_i). \tag{15}$$

Substituting (7) into (15), the likelihood function can be expanded as

$$p(\mathcal{D}^r, \mathcal{D}^e|\Theta) = \prod_{i=1}^{i=N^r} \sum_{\zeta_i=1}^{K_{y_i}} p(\mathbf{x}_i|\zeta_i, y_i)p(\zeta_i|y_i) \times \prod_{i=N^r+1}^{i=N^r+N^t} \sum_{y_i=0}^{1} \sum_{\zeta_i=1}^{K_{y_i}} \sum_{\xi_i=1}^{J} p(\mathbf{x}_i|\zeta_i, y_i)p(\zeta_i|y_i)p(y_i|\xi_i)p(\mathbf{s}_i|\xi_i)p(\xi_i). \tag{16}$$

The objective is to predict $\{y_i\}_{i=N^r+1}^{N^r+N^e}$, the unobserved class labels for the targets at the test site, which is accomplished by first maximizing the likelihood function in (16) to obtain an ML estimate of Θ and then predicting the labels based on the estimated Θ. The likelihood function has a special structure, because it involves at least one summation for each data point, where each summation arises from the marginalization with respect to a discrete latent variable. As a matter of fact, the likelihood function in (16) can be seen as resulting from the complete-data likelihood function,

$$\begin{aligned} &p(\mathcal{D}^r, \mathcal{D}^e, \{y_i, \zeta_i, \xi_i\}|\Theta) \\ &= \prod_{i=1}^{i=N^r} p(\mathbf{x}_i|\zeta_i, y_i)p(\zeta_i|y_i) \prod_{i=N^r+1}^{i=N^r+N^t} p(\mathbf{x}_i|\zeta_i, y_i)p(\zeta_i|y_i)p(y_i|\xi_i)p(\mathbf{s}_i|\xi_i)p(\xi_i), \end{aligned}$$

by marginalizing out all latent variables $\{y_i, \zeta_i, \xi_i\}$, where the indices are taken over a subset of $\{1, 2, \cdots, N^r + N^e\}$ as appropriate for each type of variables. The complete likelihood function, which expresses the joint probability distribution of observed data and all latent variables for any given parameters Θ, has a nice mathematic from, and maximization of this function (equivalently its logarithm) with respect to Θ admits a unique and analytic solution. To exploit this convenience to the benefit of ML estimation, one constructs

$$Q(\widehat{\Theta}|\Theta, \mathcal{D}^r, \mathcal{D}^e) = \mathbb{E}_{\{y_i,\zeta_i,\xi_i\}|\mathcal{D}^r,\mathcal{D}^e,\Theta} \left[\ln p(\mathcal{D}^r, \mathcal{D}^e, \{y_i, \zeta_i, \xi_i\}|\widehat{\Theta}) - \ln p(\{y_i, \zeta_i, \xi_i\}|\mathcal{D}^r, \mathcal{D}^e, \Theta) \right]. \tag{17}$$

which, for SC-GMM, inherits the nice property of admitting a unique and analytic solution when it is maximized with respect to $\widehat{\Theta}$. It is straightforward to verify that

$$\begin{aligned} p(\mathcal{D}^r, \mathcal{D}^e|\Theta) &= Q(\Theta|\Theta, \mathcal{D}^r, \mathcal{D}^e), \\ p(\mathcal{D}^r, \mathcal{D}^e|\widehat{\Theta}) &\geq Q(\widehat{\Theta}|\Theta, \mathcal{D}^r, \mathcal{D}^e). \end{aligned} \tag{18}$$

Let

$$\widehat{\Theta} = \arg\max_{\widehat{\Theta}} Q(\widehat{\Theta}|\Theta, \mathcal{D}^r, \mathcal{D}^e) \tag{19}$$

Thus, $Q(\widehat{\Theta}|\Theta, \mathcal{D}^r, \mathcal{D}^e) \geq Q(\Theta|\Theta, \mathcal{D}^r, \mathcal{D}^e)$, which, along with (18), gives

$$p(\mathcal{D}^r, \mathcal{D}^e|\widehat{\Theta}) \geq p(\mathcal{D}^r, \mathcal{D}^e|\Theta).$$

Therefore, maximization of (17) achieves a one-step improvement (from Θ to $\widehat{\Theta}$) of the parameters. By iteraively performing this maximization, $\widehat{\Theta}^{(n)} = \arg\max_{\widehat{\Theta}} Q(\widehat{\Theta}|\Theta^{(n-1)}, \mathcal{D}^r, \mathcal{D}^e)$, for $n = 1, 2, \cdots$, starting from some initialization $\Theta^{(0)}$, one obtains a sequence of successively improved estimates of Θ, which lead to monotonic increases of the likelihood function, until convergence is attained. Such iterative procedures

constitutes an expectation maximization (EM) [35] algorithm.

It is convenient to describe each iteration of EM as consisting of two step: the E-step and the M-step. In the E-step, one constructs the local objective function in (17), by taking expectation of the logarithmic complete-data likelihood function, with respect to $p(\{y_i, \zeta_i, \xi_i\}|\mathcal{D}^r, \mathcal{D}^e, \Theta)$, the joint posterior of all latent variables conditional on the observed data and the estimate of Θ in the most recent iteration. In the M-step, one maximizes the local objective function to obtain an improved estimate of the parameters Θ.

Because the data points in SC-GMM are assumed independent of each other, the latent variables associated with one data point are *a posteriori* independent of those associated with other data points. Thus, the E-step consists of calculating the posterior of latent variables for each data point separately. The posterior of latent variables associated with the labeled data points are given by

$$p(\zeta_i|\mathbf{x}_i, y_i, \Theta) = \frac{p(\mathbf{x}_i|\zeta_i, y_i)p(\zeta_i|y_i)}{\sum_{\zeta_i=1}^{K_{y_i}} p(\mathbf{x}_i|\zeta_i, y_i)p(\zeta_i|y_i)},$$

for $i = 1, 2, \cdots, i = N^r$. The posterior of the latent variables associated with the unlabeled data points are given by

$$p(y_i, \zeta_i, \xi_i|\mathbf{x}_i, \mathbf{s}_i, \Theta) = \frac{p(\mathbf{x}_i|\zeta_i, y_i)p(\zeta_i|y_i)p(y_i|\xi_i)p(\mathbf{s}_i|\xi_i)p(\xi_i)}{\sum_{y_i=0}^{1}\sum_{\zeta_i=1}^{K_{y_i}}\sum_{\xi_i=1}^{J} p(\mathbf{x}_i|\zeta_i, y_i)p(\zeta_i|y_i)p(y_i|\xi_i)p(\mathbf{s}_i|\xi_i)p(\xi_i)}, \tag{20}$$

for $i = N^r + 1, N^r + 2, \cdots, i = N^r + N^e$.

To state the parameter updating formulae in the M-step, we first introduce the following abbreviations and intermediate notations:

$$\begin{aligned}
\omega_k^i &= p(\zeta_i = k|\mathbf{x}_i, y_i, \Theta),\ i = 1, 2, \cdots, i = N^r, \\
\omega_{tkj}^i &= p(y_i = t, \zeta_i = k, \xi_i = j|\mathbf{x}_i, \mathbf{s}_i, \Theta),\ i = N^r + 1, N^r + 2, \cdots, N^r + N^e, \\
c_{kt}^i &= \begin{cases} \omega_{kt}^i \delta_{y_i,t}, & i = 1, 2, \cdots, N^r \\ \frac{\sum_{j=1}^{J}\omega_{tkj}^i}{\sum_{t'=0}^{1}\sum_{j=1}^{J}\omega_{t'kj}^i}, & i = N^r + 1, N^r + 2, \cdots, N^r + N^e \end{cases}, \\
d_j^i &= \sum_{t=0}^{1}\sum_{k=1}^{K_t}\omega_{tkj}^i,\ i = N^r + 1, N^r + 2, \cdots, N^r + N^e.
\end{aligned}$$

Solving the maximization in (19) gives the updated parameters as

$$\begin{aligned}
\widehat{\pi}_{kt} &= \frac{\sum_{i=1}^{N^r+N^e} c_{kt}^i}{\sum_{k=1}^{K_t}\sum_{i=1}^{N^r+N^e} c_{kt}^i}, \\
\widehat{\boldsymbol{\mu}}_{kt} &= \frac{\sum_{i=1}^{N^r+N^e} c_{kt}^i \mathbf{x}_i}{\sum_{i=1}^{N^r+N^e} c_{kt}^i} \\
\widehat{\boldsymbol{\Sigma}}_{kt} &= \frac{\sum_{i=1}^{N^r+N^e} c_{kt}^i (\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_{kt})(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_{kt})^T}{\sum_{i=1}^{N^r+N^e} c_{kt}^i} \\
\widehat{\eta}_j &= \frac{\sum_{i=N^r+1}^{N^r+N^e} d_j^i}{\sum_{j=1}^{J}\sum_{i=N^r+1}^{N^r+N^e} d_j^i}, \\
\widehat{\boldsymbol{\beta}}_j &= \frac{\sum_{i=N^r+1}^{N^r+N^e} d_j^i \mathbf{s}_i}{\sum_{i=N^r+1}^{N^r+N^e} d_j^i},
\end{aligned}$$

$$\widehat{\boldsymbol{\Omega}}_j = \frac{\sum_{i=N^r+1}^{N^r+N^e} d_j^i (\mathbf{s}_i - \widehat{\boldsymbol{\beta}}_j)(\mathbf{s}_i - \widehat{\boldsymbol{\beta}}_j)^T}{\sum_{i=N^r+1}^{N^r+N^e} d_j^i},$$
$$\widehat{\theta}_{tj} = \frac{\sum_{i=N^r+1}^{N^r+N^e} \sum_{k=1}^{K_t} \omega_{tkj}^i}{\sum_{t=0}^{1} \sum_{i=N^r+1}^{N^r+N^e} \sum_{k=1}^{K_t} \omega_{tkj}^i}.$$

Upon convergence of the EM iterations, the estimate of Θ is (locally) optimal and one can predict the class labels for the targets at the test site, based on using

$$p(y_i|\mathbf{x}_i, \mathbf{s}_i, \Theta) = \frac{\sum_{\zeta_i=1}^{K_{y_i}} \sum_{\xi_i=1}^{J} p(\mathbf{x}_i|\zeta_i, y_i) p(\zeta_i|y_i) p(y_i|\xi_i) p(\mathbf{s}_i|\xi_i) p(\xi_i)}{\sum_{y_i=0}^{1} \sum_{\zeta_i=1}^{K_{y_i}} \sum_{\xi_i=1}^{J} p(\mathbf{x}_i|\zeta_i, y_i) p(\zeta_i|y_i) p(y_i|\xi_i) p(\mathbf{s}_i|\xi_i) p(\xi_i)},$$

for $i = N^r + 1, N^r + 2, \cdots, N^r + N^e$, which follows directly from (20). The predictions above are based on using both the feature vectors and the associated spatial locations, which is possible only for the unlabeled data from the current test site (these data participated in the ML estimation). For unlabeled data from a future test site, the spatial locations cannot be used, because the spatial coordinate system there may not be same as the one used at the current site. In this case, the predictions are based on using only the feature vectors,

$$p(y_i|\mathbf{x}_i, \Theta) = \frac{\sum_{\zeta_i=1}^{K_{y_i}} p(\mathbf{x}_i|\zeta_i, y_i) p(\zeta_i|y_i) p(y_i)}{\sum_{y_i=0}^{1} \sum_{\zeta_i=1}^{K_{y_i}} p(\mathbf{x}_i|\zeta_i, y_i) p(\zeta_i|y_i) p(y_i)}, \tag{21}$$

where

$$p(y_i) = \sum_{\xi_i=1}^{J} p(y_i|\xi_i) p(\xi_i).$$

B. Spatially-Constrained Mixture of Experts

In the model of SC-GMM, the feature-domain sub-model consists of two density functions, each represented by a GMM, one for the feature vectors of UXO and the other for the feature vectors of clutter. To make the prediction for a label, one applies Bayes rule as in (21). A drawback of making prediction in such a way is that the accuracy of such predictions is dependent on the accuracy of the estimated densities, here the two feature-domain GMMs. To obtain accurate densities, one needs a large amount of sample feature vectors from both UXO and clutter, to continually cover all probable areas. Many of these samples, however, would be irrelevant to the purpose of learning the decision boundary between UXO and clutter. This is particularly true for the samples far away from the boundary, because only the samples near the decision boundary are relevant to defining an accurate classifier [38]. Therefore, SC-GMM is not efficient in the usage of feature-domain samples, and does not perform satisfactorily in UXO detection when the training samples in the feature domain are inadequate.

To ameliorate this situation, we give an alternative formulation for the spatially-constrained classifier. A particular attraction of the new formulation is that it has a local probit model associated with each Gaussian component in the feature-domain. The conditional distribution of class label given a feature vector is directly represented as a mixture of local probit models, without having to resort to Bayes rule as in (21). Thus the new model can yield a more accurate decision boundary between UXO and clutter than SC-GMM when trained on a given number of labeled feature vectors. The discriminative nature of

the new model makes it possible to further reduce the false alarm rate in UXO detection.

In the new formulation, the labeled data are assumed to be governed by a quadratically gated mixture of experts (QGME) [34],

$$p(\mathbf{x}_i, y_i) = \sum_{\zeta_i=1}^{K} p(y_i|\mathbf{x}_i, \zeta_i) p(\mathbf{x}_i|\zeta_i) p(\zeta_i), \tag{22}$$

with

$$\begin{aligned} p(\zeta = k) &= \pi_k, \\ p(\mathbf{x}_i|\zeta_i = k) &= \frac{1}{\sqrt{2\pi|\Sigma_k|}} e^{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)}, \\ p(y_i|\mathbf{x}_i, \zeta_i = k) &= \int_{z_i : z_i(2y_i - 1) \geq 0} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z_i - \mathbf{w}_k^T \mathbf{x}_i)^2} dz_i, \end{aligned}$$

for $i = 1, 2, \cdots, N^r$, where K is the number of experts and the k-th expert, represented by $p(y_i|\mathbf{x}_i, \zeta_i = k)$, is a probit regression model [39]. The unlabeled data are governed by

$$p(\mathbf{x}_i, \mathbf{s}_i) = \sum_{y_i=0}^{1} p(\mathbf{x}_i, y_i) p(\mathbf{s}_i|y_i), \tag{23}$$

with

$$p(\mathbf{s}_i|y_i) = \sum_{\xi_i=1}^{J} p(\mathbf{s}_i|\xi_i) p(\xi_i|y_i), \tag{24}$$

for $i = N^r + 1, N^r + 2, \cdots, N^r + N^e$, where $p(\mathbf{x}_i, y_i)$ is the same as in (22), and

$$\begin{aligned} p(\xi_i = j|y_i = t) &= \eta_{jt}, \\ p(\mathbf{s}_i|\xi_i = j) &= \frac{1}{\sqrt{2\pi|\Omega_j|}} e^{-\frac{1}{2}(\mathbf{s}_i - \boldsymbol{\beta}_j)^T \boldsymbol{\Omega}_j^{-1} (\mathbf{s}_i - \boldsymbol{\beta}_j)}. \end{aligned}$$

The new model formulated above is referred to spatially-constrained QGME (SC-QGME). In reference to SC-GMM, the SC-QGME model defined above has several major changes: (*i*) the density of feature vectors is represented by a single GMM, instead of two label-dependent GMMs; (*ii*) the density of spatial locations is represented by two label-dependent GMMs, instead of a single GMMs, and the two spatial-domain GMMs have shared mixture components but label-dependent mixing proportions; (*iii*) class labels are generated in the feature-domain, instead of in the spatial-domain.

The dependence between the spatial location and the feature vector can be represented by

$$p(\mathbf{s}_i|\mathbf{x}_i) = \sum_{y_i=0}^{1} p(\mathbf{s}_i|y_i) p(y_i|\mathbf{x}_i)$$

where

$$p(y_i|\mathbf{x}_i) = \sum_{k=1}^{K} p(\zeta_i = k|\mathbf{x}_i) p(y_i|\mathbf{x}_i, \zeta_i = k)$$

with

$$p(\zeta_i = k|\mathbf{x}_i) = \frac{p(\mathbf{x}_i|\zeta_i = k) p(\zeta_i = k)}{\sum_{k=1}^{K} p(\mathbf{x}_i|\zeta_i = k) p(\zeta_i = k)}$$

Similar to the case with SC-GMM, the dependence reflects the degree to which the feature-domain sub-model and the spatial-domain sub-model agrees on the class label. Based on the same arguments as made

in Section VII-A about the complete-information condition, we seek the parameters Θ that maximize the following likelihood function,

$$p(\mathcal{D}^r, \mathcal{D}^e|\Theta) = \prod_{i=1}^{i=N^r} p(\mathbf{x}_i, y_i) \prod_{i=N^r+1}^{i=N^r+N^t} \sum_{y_i=0}^{1} p(\mathbf{s}_i|y_i)p(\mathbf{x}_i, y_i).$$

Substitution of (22) and (23) expands the likelihood function as

$$\begin{aligned} p(\mathcal{D}^r, \mathcal{D}^e|\Theta) = \quad & \prod_{i=1}^{i=N^r} \sum_{\zeta_i=1}^{K} p(y_i|\mathbf{x}_i, \zeta_i)p(\mathbf{x}_i|\zeta_i)p(\zeta_i) \\ & \times \prod_{i=N^r+1}^{i=N^r+N^t} \sum_{y_i=0}^{1} \sum_{\zeta_i=1}^{K} \sum_{\xi_i=1}^{J} p(y_i|\mathbf{x}_i, \zeta_i)p(\mathbf{x}_i|\zeta_i)p(\zeta_i)p(\mathbf{s}_i|\xi_i)p(\xi_i|y_i). \end{aligned}$$

The associated complete-data likelihood function is given by

$$p(\mathcal{D}^r, \mathcal{D}^e, \{y_i, \zeta_i, \xi_i\}|\Theta) = \prod_{i=1}^{i=N^r} p(y_i|\mathbf{x}_i, \zeta_i)p(\mathbf{x}_i|\zeta_i)p(\zeta_i) \prod_{i=N^r+1}^{i=N^r+N^t} p(y_i|\mathbf{x}_i, \zeta_i)p(\mathbf{x}_i|\zeta_i)p(\zeta_i)p(\mathbf{s}_i|\xi_i)p(\xi_i|y_i).$$

Maximization of the likelihood function, with respect to Θ, can be accomplished by expectation maximization, similar to the case with CS-GMM. The technical details are omitted here.

VIII. Offline-learning and Online Learning

Given the model of SC-GMM or SC-QGME, one proceeds to learn the model parameters based on observed data, with the goal of predicting unobserved class labels. In UXO detection, the observed data include those from previous sites (labeled feature vectors) as well as those from the site currently under test (unlabeled feature vectors and the associated spatial locations). There are two approaches that one can take to learn the model parameters.

The first is off-line learning, which distinguishes two different phases: the learning phase and the testing phase. In the learning phase, one learns the model parameters based on a given set of observed data. Upon convergence of the parameters, one then switches to the testing phase and employs the learned model to predict the labels for the unlabeled targets at the test site.

The second is online learning, which interweaves learning and testing into one single process, without treating them as two separate phases. Specifically, online learning works with a incremental set of labeled data. Initially the labeled data are $\mathcal{D}^r$, the labeled feature vectors from previous sites. Right after each interrogated target is declared, the label is submitted for verification to a UXO expert who compares it to the ground truth uncovered by performing experiments. Each new ground truth label produces a new labeled feature vector along with the associated spatial location. The new labeled data are added into $\mathcal{D}^r$ (note that the new ones have spatial locations and the initial ones do not have). Clearly, each example moved into $\mathcal{D}^r$ must come from $\mathcal{D}^e$, so eventually $\mathcal{D}^e$ will become empty, at which point all targets at the test site have been declared. In off-line learning, where one does not have access to the true label after each declaration, $\mathcal{D}^e$ remains in its initial state until the end of the learning phase.

Online learning is appropriate for any application in which observed data are sequentially augmented, typically each time when a prediction is made. In UXO detection, the additional data observed (which is

the label of an unknown target) could be obtained by performing excavation around the target to reveal its class identity. In general, it is desirable to uncover the class label regardless of what the prediction is. However this is not true for UXO detection, because excavation of a clutter item is a waste of time and resources. Therefore one proceeds with excavation only when the unknown target is predicted as potentially to be a UXO target. The validity of this "dig-UXO-only" approach is supported by the fact that the UXO is a rare class. The ground truth of UXO targets being significantly outnumbered by clutter items dictates that, the occurrences of targets predicted as UXO may also be small; this is particularly true after a number of excavations have been made, in which time the model is better-trained and the predictions become more accurate.

IX. Experimental Results

A. Experimental setup

The data we use in the experiments are the Sibert data set, which consists of a training set and testing set. The training set has 166 labeled feature vectors, of which 58 are associated with UXO, and the feature dimensionality is 5. The test set contains 714 unlabeled feature vectors along with their associated spatial locations, of which 595 correspond to clutter items and 119 correspond to UXO targets. The spatial distribution of the 714 targets under test are displayed in the left panel of Figure 22, with red stars indicating UXO targets and blue circles indicating clutter items. The UXO targets in this test set are artificially planted for testing purposes, therefore their spatial distribution deviate from those that would be encountered in practical scenarios. To make it more realistic, we modify the spatial distribution by perturbing the UXO locations around their original values. The modified spatial locations are shown in the right panel of Figure 22. The modification is guided by the spatial distribution of UXO targets detected at Yekau Lake, Alberta, shown in Figure 23. This is a real UXO site, where the ordnance density tails away from the epicenter of bombing. Other than the spatial locations, the data sets remain intact, which yield a training set D^r constituting 166 labeled feature vectors, and a test data D^e consisting of 714 unlabeled feature vectors along with the associated spatial locations modified as described above. The goal is to predict the labels of examples in D^e.

We compare the following six models in the experiments:

1) SC-GMM, the model defined in Section VII-A, trained on $\mathcal{D}^r \cup D^e$;
2) GMM, the two label-dependent GMMs as specified in (7), trained on $\mathcal{D}^r$. In online learning, the two GMMs are additionally trained on labeled feature vectors acquired from $\mathcal{D}^e$.
3) SS-GMM, the semi-supervised version of the two label-dependent GMMs as specified in (7), trained on all feature vectors in $\mathcal{D}^r$ and $\mathcal{D}^e$, regardless of whether they are labeled.
4) SC-QGME, the model defined in Section VII-B, trained on $\mathcal{D}^r \cup \mathcal{D}^e$;
5) QGME, the quadratically gated mixture of experts [34], trained on $\mathcal{D}^r$, as well as labeled feature vectors available from $\mathcal{D}^e$ when in online learning.
6) SS-QGME, semi-supervised QGME, trained on all labeled, and unlabeled feature vectors in $\mathcal{D}^r$ and $\mathcal{D}^e$.

In plain words, GMM, SS-GMM, QGME, and SS-QGME are trained on (labeled and unlabeled) feature vectors alone, without using the spatial locations. SC-GMM and SC-QGME are trained on feature vectors

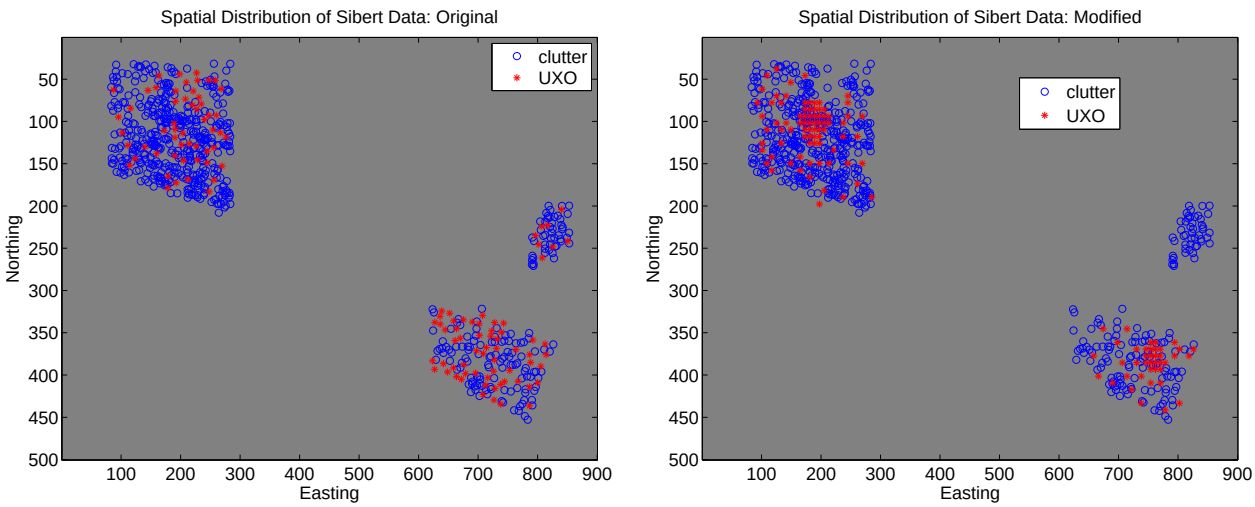


Fig. 22. The spatial distribution of targets in the Sibert test set: (left) the original set, (right) the modified test set, obtained by perturbing the original spatial locations of the UXO targets.

along with the associated spatial locations, labeled and unlabeled.

We also compare two learning approaches:

1) Offline learning, in which each model is trained as described above, and is then employed to rank the test targets in the descending order of

$$p(y_i = 1|\mathbf{x}_i, \mathbf{s}_i) \quad \text{or} \quad p(y_i = 1|\mathbf{x}_i), \quad i = N^r + 1, N^r + 2, \cdots, N^r + N^e.$$

The list of ranked test targets is used to produce a receiver operating characteristic (ROC) curve.

2) Online learning, in which an initial model is learned as in offline learning, and the initial model is successively improved using an augmented version of $\mathcal{D}^r$. The $\mathcal{D}^r$ is augmented by adding in a data point from $\mathcal{D}^e$, which is the most probable one for being UXO,

$$(\mathbf{x}, \mathbf{s}) = \arg \max_{(\mathbf{x},\mathbf{s}) \in \mathcal{D}^e} p(y = 1|\mathbf{x}, \mathbf{s}) \quad \text{or} \quad (\mathbf{x}, \mathbf{s}) = \arg \max_{(\mathbf{x},\mathbf{s}) \in \mathcal{D}^e} p(y = 1|\mathbf{x})$$

and the true label y of the selected target is acquired and added to $\mathcal{D}^r$ along with $(\mathbf{x}, \mathbf{s})$. Online learning stops when $\mathcal{D}^e$ becomes empty. During online learning, the targets selected to be added to $\mathcal{D}^e$, in the order of being added, form a ranked list to produce the ROC curves.

B. Detection Results

A total of twelve ROC curves are generated, corresponding to the twelve cases resulting from the Cartesian product of the set of six models with {offline-learning, online-learning}. The orders of the models are set as: $K = 2$ for QGME, SS-QGME, and SC-QGME; $K_0 = K_1 = 2$ for GMM, and SS-GMM, SC-GMM; $J = 20$ for SC-GMM and $J = 50$ for SC-GQME. The setting is based on intuition, without careful tuning. Te ROC curves are presented in Figure 24 and Figure 25.

The best performance is achieved by online learning of SC-QGME, followed by offline learning of SC-QGME. These two both employ spatial information and take advantage of local probit models in the feature domain. The only difference lie in that the first has additional training from the augmentation of $\mathcal{D}^r$, and

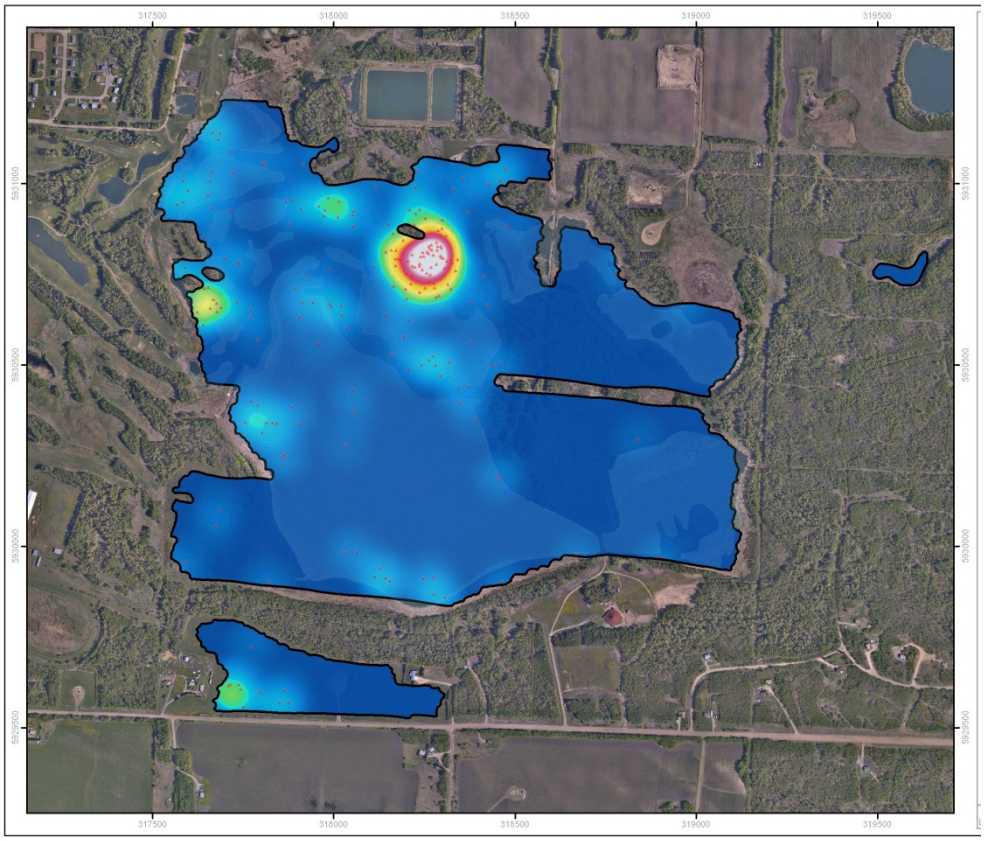


Fig. 23. Distribution of UXO detected at Yekau Lake, Alberta. Ordnance cluster around bombing targets.

the second does not. Thus this comparison highlights the benefits gained by additional amount of training examples. Note that the online learning approach here acquires the labels only if the target is predicted as being UXO; in other words, it actively acquires the labels of the most probable UXO candidates. This special-form online learning is based on the fact that UXO is a rare class.The improvement from online learning therefore also implicitly shows the benefits of using the rarity property of UXO.

The comparison of SC-QGME against QGME, in both offline learning or online learning, demonstrate the improvements brought about by using the spatial information. The degraded performances of SS-QGME and SS-GMM indicate that unlabeled examples may not always be helpful. For SS-QGME, the unlabeled examples, which outnumber the labeled ones, can exert a dominant effect on the gating network in QGME, and thus make the model suffer from over-fitting. For SS-GMM, the latent labels of test examples provide extra freedom of over-fitting the model to the density of feature vectors, which is known to lead to the drawback of self-training in supervised learning [32]. It is interesting to note that SC-QGME is also semi-supervised in the feature domain, but it does not suffer from self-training because semi-supervision is simultaneously performed in the spatial domain, leading to “co-training” between two (relatively independent) domains, which is known to be beneficial to supervised learning [31].

The fact that SC-GMM is outperformed by GMM indicates that utility of spatial information heavily depend on the feature-domain model. The GMM, which is a purely generative model, may require more stringent conditions on the UXO spatial distribution in order to become an appropriate feature-domain

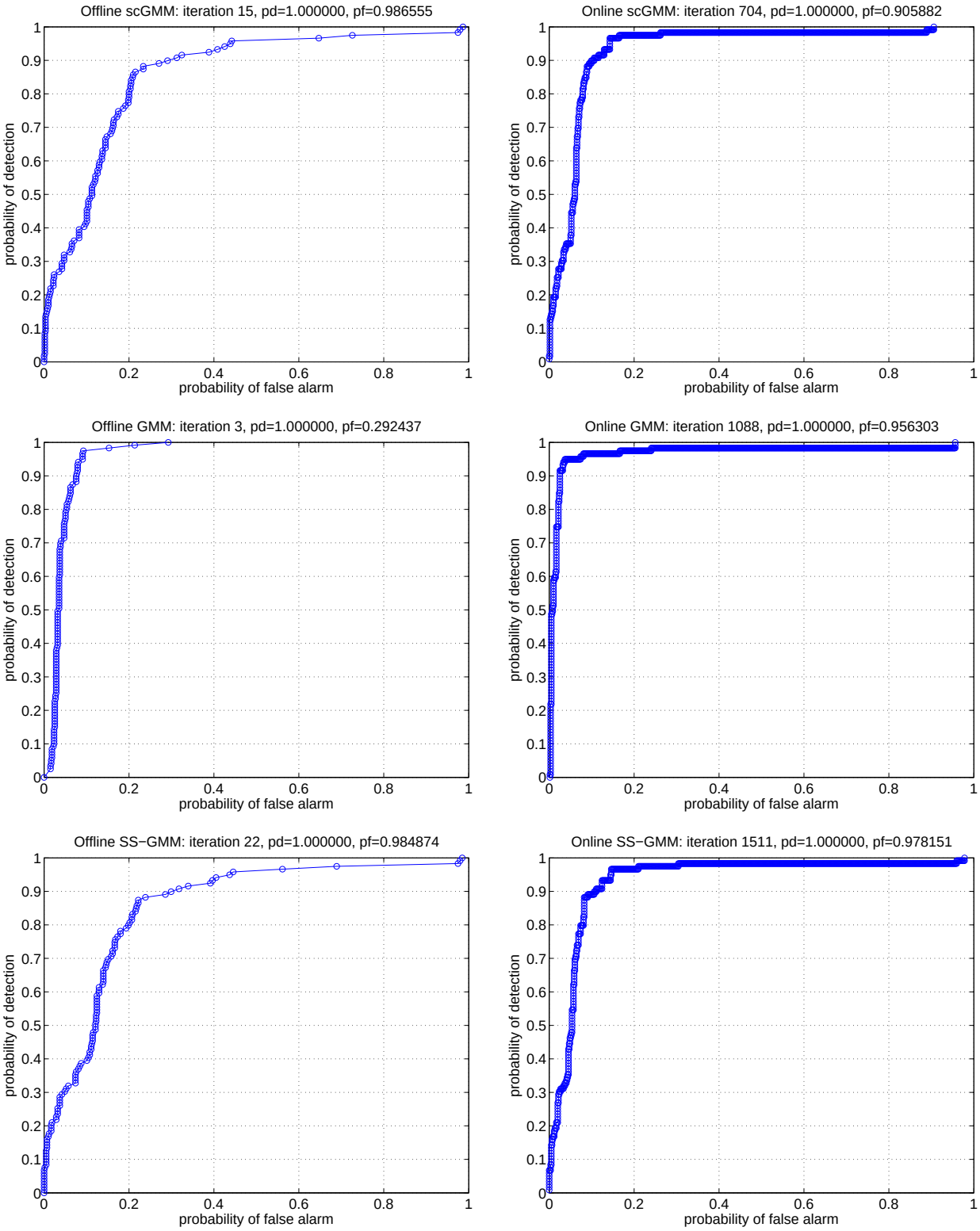


Fig. 24. A comparison of six models in terms of ROC curves, where the results in the top row are based on the perturbed UXO spatial locations shown in the right panel of Figure 22. Each model is indicated in the title of the associated subfigure. To be continued in Figure 25.

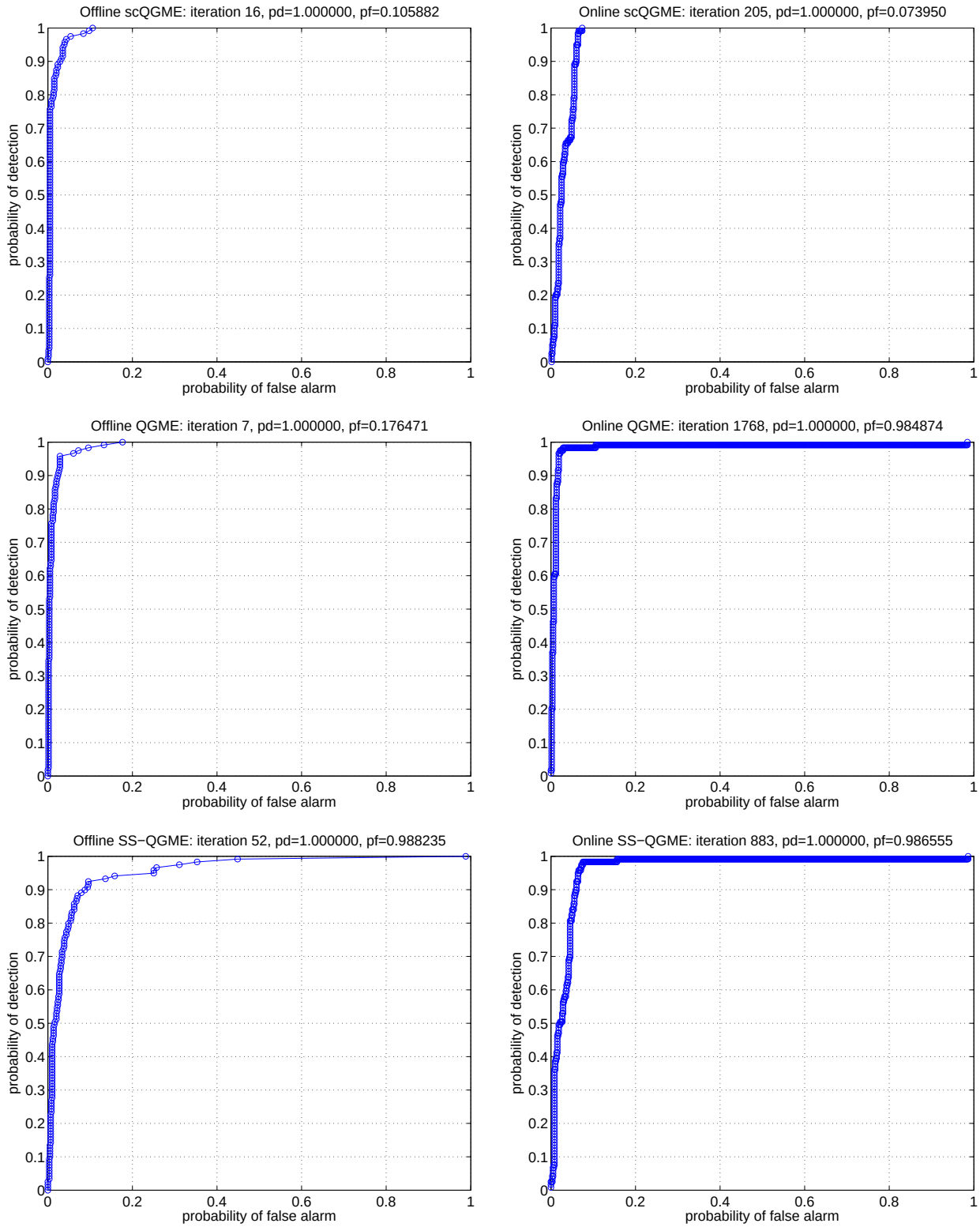


Fig. 25. Continued from Figure 24. A comparison of six models in terms of ROC curves, where the results in the top row are based on the perturbed UXO spatial locations shown in the right panel of Figure 22. Each model is indicated in the title of the associated subfigure.

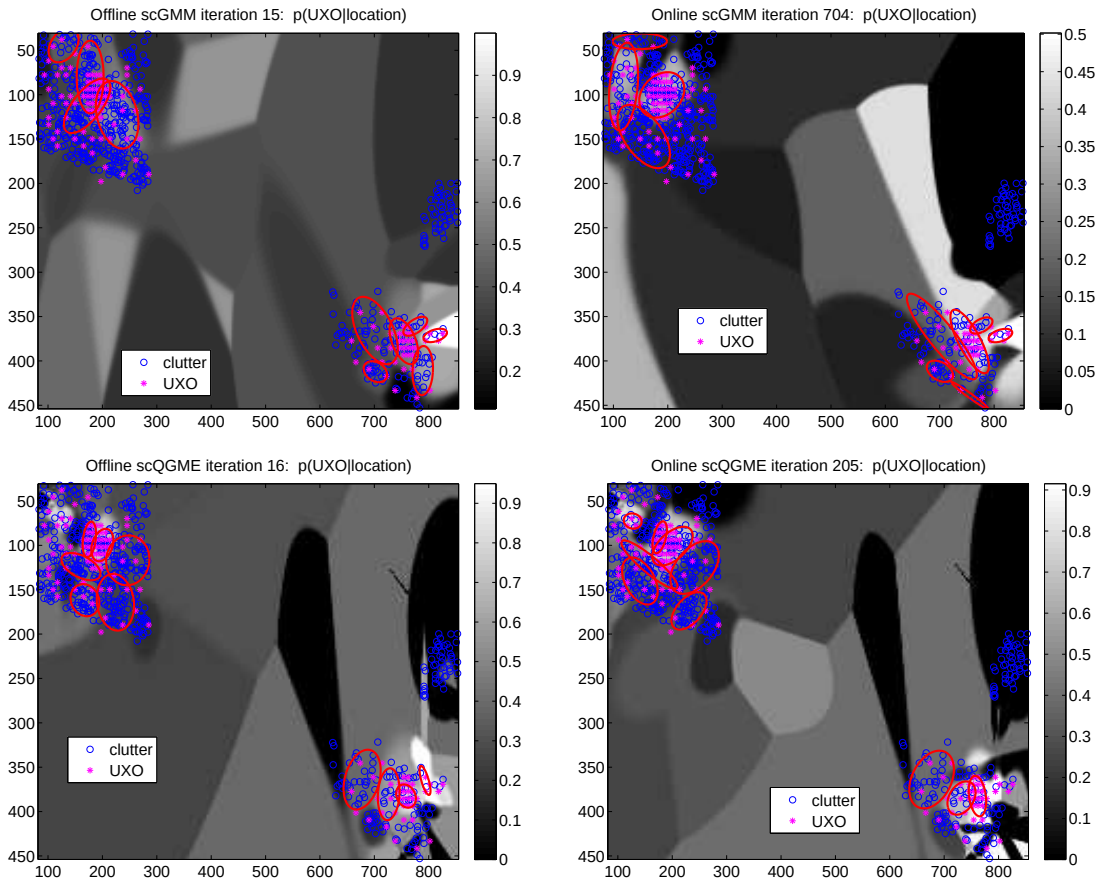


Fig. 26. The probability of being UXO given the spatial location, i.e., $p(y = 1|\mathbf{s})$, learned by: (top left) offline CS-GMM, (top right) online CS-GMM, (bottom left) offline CS-QGME, (bottom right) online CS-QGME. The results are based on the perturbed UXO spatial locations shown in the right panel of Figure 22.

model to incorporate the spatial information in a useful manner. Our previous studies showed that, when the UXO targets fall in tight clusters, SC-GMM could be a useful model. In the present data sets, the UXO targets do not form tight clusters, rendering the spatial information unhelpful. This, along with the fact that the model is semi-supervised in the feature-domain, makes SC-GMM degenerate to a self-training SS-GMM and yield similarly unsatisfactory performances.

The results consistently show that QGME is a better feature-domain model to take advantage of the spatial information, which is attributable to the local discriminative classifiers (probit regression models) contained in the model.

The performance comparison among those models employing spatial information can also be intuitively understood by visualizing $p(y = 1|s)$ as a spatial UXO-map in s, where $p(y = 1|s)$ is obtained when all UXOs have been found. These maps are shown in Figure 26, where the ellipses represent the top ten spatial local regions dominated by UXO. From these maps one can draw similar comparisons as from the ROC curves.

Finally, Figure 27 shows the mixing proportions of the two label-dependent spatial-domain GMMs. The bottom plot, which corresponds to the GMM for UXO, has few significant nonzero values, showing the GMM for UXO has sparse mixing over the spatial regions (each represented by a Gaussian component). This again demonstrates that UXO is a rare class, sparsely distributed in the spatial space. As a comparison,

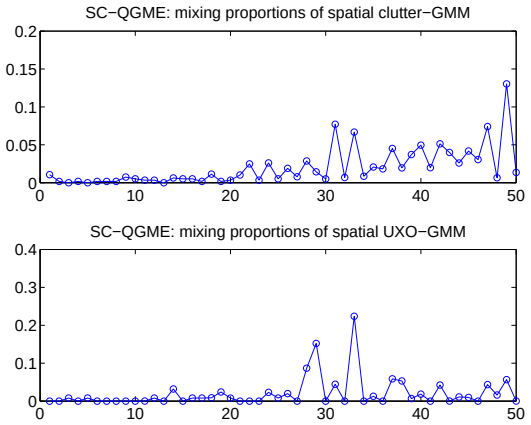


Fig. 27. The mixing proportions of the spatial-domain GMMs in SC-QGME, for: (top) clutter, (bottom) UXO. The results are based on the perturbed UXO spatial locations shown in the right panel of Figure 22.

the GMM for clutter has dense mixing proportions.

The above discussions show that the use of spatial information can help to reduce the number of false alarms, particularly at the point when all UXO targets are identified. The utility of spatial information depends on how much prior knowledge we have about the spatial distribution of UXO targets. Clearly, the more we know *a priori*, the greater the utility will be. The spatially-constrained classifiers we have designed assume that the UXO targets are proximate to each other. Here the proximity means that, the occurrence of UXO at a location implies high probability of observing more UXO's in the neighborhood. This assumption is different from that the UXO's cluster together, because we assume that clutter may reside between UXO targets.

The UXO targets in the right penal of Figure 22 are not spatially grouped – instead they have a density that tails away from the epicenter of UXO occurrences; such densities have been observed at real UXO sites such as that shown in Figure 23 and are generally true for sites with a broad area (tens of acres, for example). The left penal of Figure 22 shows the original UXO locations of Sibert data, which are seen almost uniformly distributed. This uniform distribution violates the above assumption and we expect that the spatial information will not provide too much help. To verify this, we run the comparison of the six models on the original Sibert test data, with UXO locations shown in the left penal of Figure 22. The results are reported in Figures 28-31, which parallel the results in Figures 24-27, with the only difference in UXO locations of the test data. The results not using spatial information are exactly the same as those in previous figures – they are duplicated in the new figures to allow a shoulder-by-shoulder comparison. The results show that spatial information does help when the UXO targets are distributed in a non-informative format (here nearly uniformly distributed). Therefore, the best performance here is given by QGME based on using only feature vectors. Nevertheless we notice that, for QGME operating in the offline mode, the use of spatial information does not bring detrimental effect on the performance. Thus, even though one may not know *a priori* if the UXO locations are distributed to satisfy the assumption, one can still safely run scQGME and get no clearly degraded performance in any case. From Figure 31, we also observe that the UXO targets are not sparse spatially – the uniform occurrences of UXO provides no information to further improving the detection.

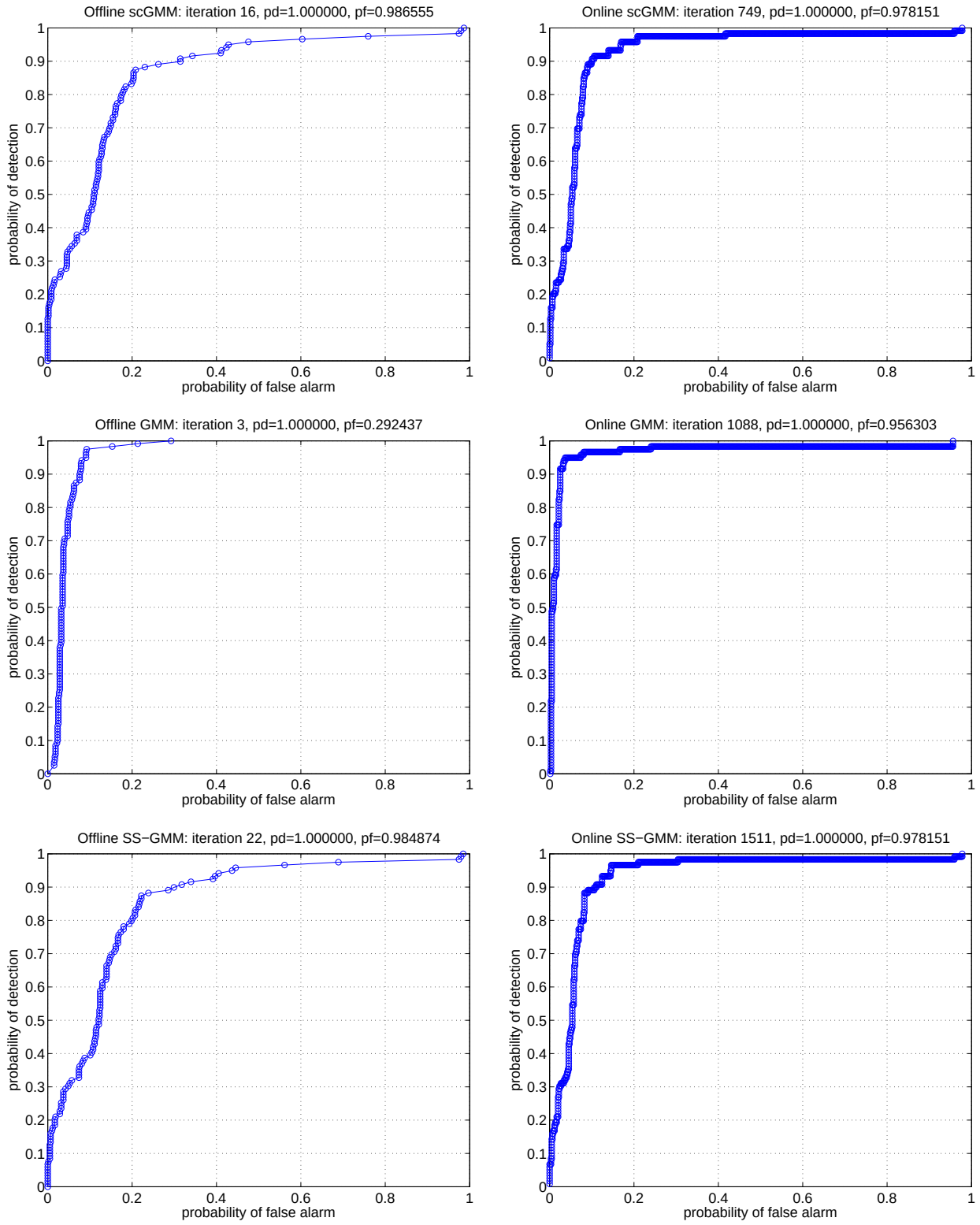


Fig. 28. A comparison of six models in terms of ROC curves, where the results in the top row are based on the original UXO spatial locations shown in the left panel of Figure 22. Each model is indicated in the title of the associated subfigure. To be continued in Figure 25.

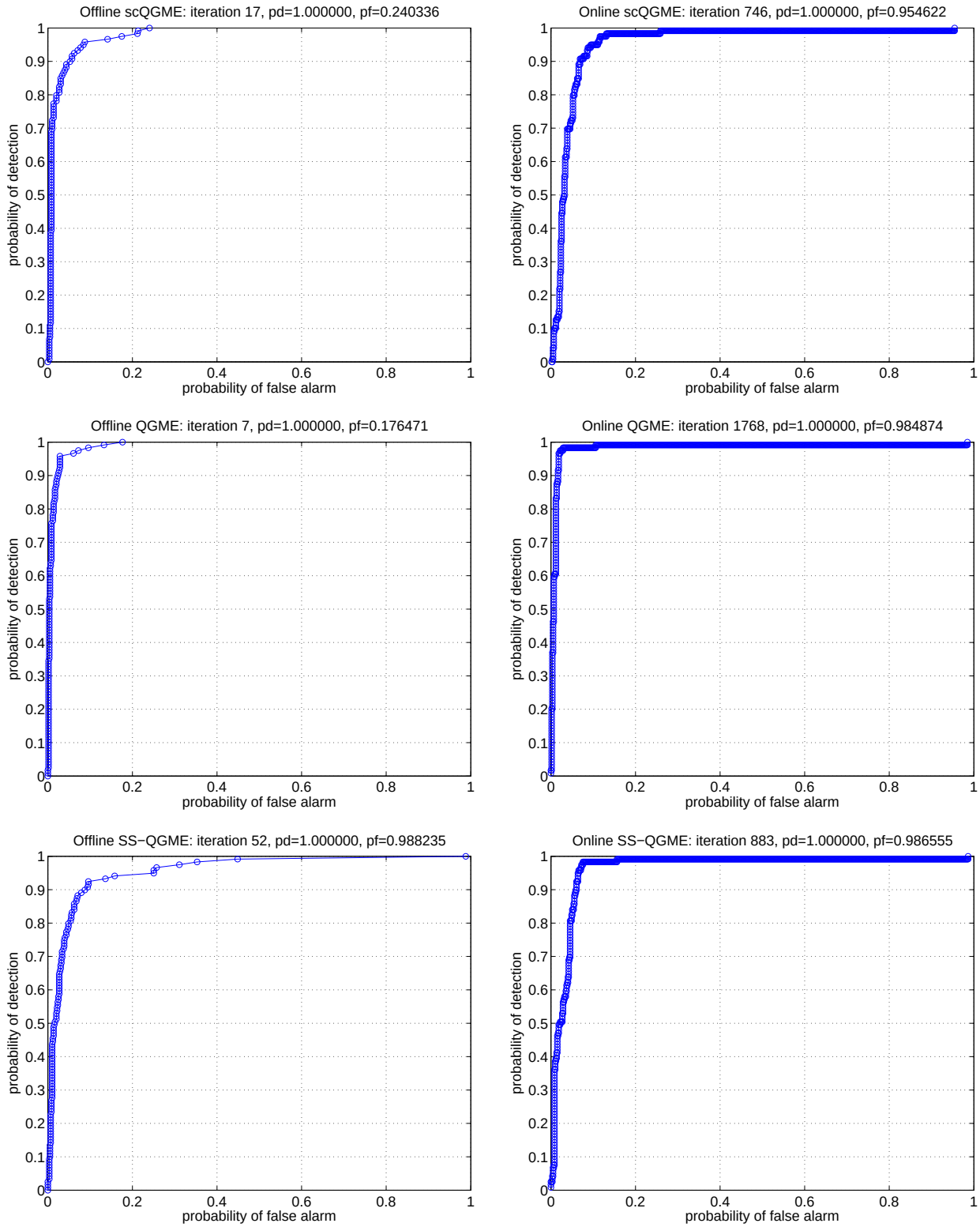


Fig. 29. Continued from Figure 24. A comparison of six models in terms of ROC curves, where the results in the top row are based on the original UXO spatial locations shown in the left panel of Figure 22. Each model is indicated in the title of the associated subfigure.

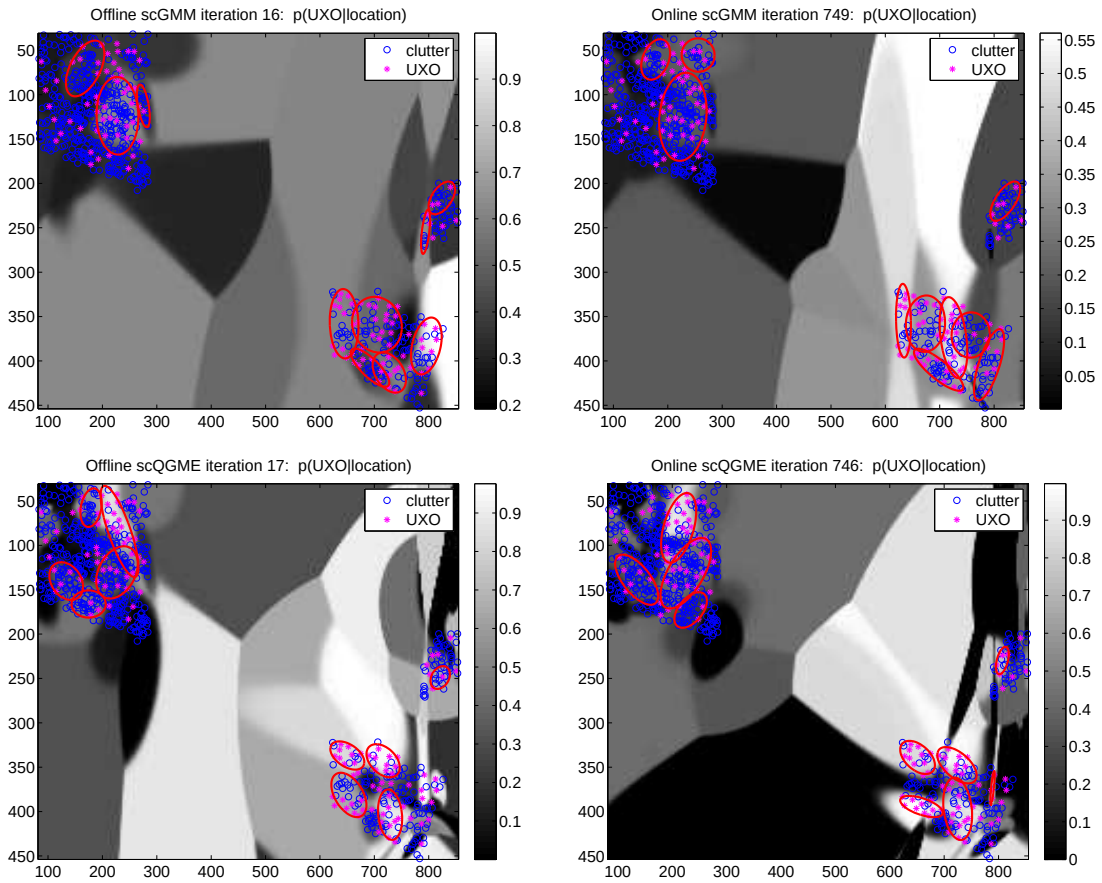


Fig. 30. The probability of being UXO given the spatial location, i.e., $p(y = 1|\mathbf{s})$, learned by: (top left) offline CS-GMM, (top right) online CS-GMM, (bottom left) offline CS-QGME, (bottom right) online CS-QGME. The results are based on the original UXO spatial locations shown in the left panel of Figure 22.

X. Discussion and Future Work

Our progress this year has focused on the following areas:

- Feature extraction. We have investigated a number of data features which can be used to pre-screen targets or to generate a dig map of regions in a survey area where ordnance are likely to occur. We have developed a simplified fingerprinting technique which fits library polarizations to individual soundings and applied the technique to EM63 and MetalMapper data. An important task for the upcoming year will be direct comparison of this method with our standard inversion and discrimination workflow, and with methods developed at Duke.
- Performance comparison. We have compared performance of Duke/SIG and UBC/Sky discrimination strategies at SLO. We have also independently applied Duke PNBC and RVM classifiers in a retrospective analysis of MetalMapper and TEMTADS data. A priority for next year will be close communication between the two groups when working to generate collaborative results.
- Workflow development. We have reorganized our discrimination codes to make them accessible from workflow scripts. Upcoming collaboration with CH2M Hill on the Camp Butner demonstration will allow us to develop and test specific workflows for processing of EM61 and MetalMapper data.
- Simulated seeding. We have simulated the variability of TOI features by adding residuals from field

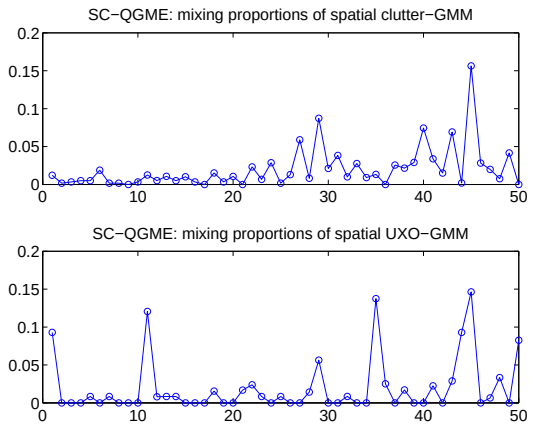


Fig. 31. The mixing proportions of the spatial-domain GMMs in SC-QGME, for: (top) clutter, (bottom) UXO. The results in the top row are based on the original UXO spatial locations shown in the left panel of Figure 22.

data to testpit measurements. This procedure uses interpolation to generate data at field locations. Further investigation is needed to determine how this interpolation affects the simulated features. Another option which will be considered is to simply forward model the synthetic data using polarizations extracted from testpit data. This avoids potentially non-physical interpolation artifacts, but risks the "inverse crime" (wherein the same model is used to forward model and invert synthetic data).

- Utilization of additional UXO information. Most previous research on UXO detection has focused on modeling the data, feature extraction and classifier design. An important piece of information is widely known in practice but is rarely accounted for in classifier design. Specifically, actual UXO are rare, and this knowledge should be accounted for when making classification decisions and when designing the algorithm. In this project we have developed a new class of algorithms that achieve this goal. As this SERDP project proceeds, we will further test and refine this approach based upon additional data sources. We are also placing the algorithm under a broader framework, via use of spatially-dependent Poisson processes.
- Multi-scale, data-driven physical features. As discussed in this report, an important direction for new UXO-detection research involves development of techniques that move beyond the simple dipole-based feature extraction framework. This has at least two motivations: (i) in many realistic scenarios the data of interest simply do not fit a dipole model well, and in this case the inferred dipole-model parameters are of limited value, and may be misleading; (ii) as SERDP/ESTCP develop new more-sophisticated sensors, the dipole-source assumption, and the ability to fully/accurately model the sensor may become a challenge. We have reported here on new techniques for classification, that do not rely on a dipole model. There is ongoing research in this project, which will be detailed in the next report, which is based on dictionary learning for feature extraction. Such a technique is data driven, and has recently proven to yield state-of-the-art results in image classification tasks. Additionally, the spatial character of the data signatures (magnetometer/EMI) may have multiple scales/resolutions. For this problem we are developing multi-scale statistical models, which yield a physics-based wavelet-like decomposition. These so-called "deep" or multi-scale constructions are

also yielding state-of-the-art results in many inference challenges.

REFERENCES

[1] W. P. Delaney and D. Etter, "Report of the Defense Science Board on Unexploded Ordnance," Office of the Undersecretary of Defense for Acquisition, Technology and Logistics, Tech. Rep., 2003.

[2] P. B. Weichman, "Surface modes and multi-power law structure in the early-time electromagnetic response of magnetic targets," *Phys. Rev. Lett.*, vol. 93, p. 023902, 2004.

[3] 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*, vol. 4, pp. 629–633, 2007.

[4] L. R. Pasion, "Inversion of time-domain electromagnetic data for the detection of unexploded ordnance," Ph.D. dissertation, University of British Columbia, 2007.

[5] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference and prediction*. Spring-Verlag, 2001.

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

[7] 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*, vol. 61, pp. 279–291, 2007.

[8] S. L. Tantum and L. M. Collins, "A comparison of algorithms for subsurface target detection and identification using time-domain electromagnetic induction data." *IEEE Trans. Geosci. Remote Sensing*, vol. 39, pp. 1299–1306, 2001.

[9] S. Norton, I. J. Won, and E. Cespedes, "Spectral identification of buried unexploded ordnance from low-frequency electromagnetic data." *Subsurface Sensing Technologies and Applications*, vol. 2, pp. 177–189., 2001.

[10] L. Carin and et al., "Wide-area detection of land mines and unexploded ordnance," *Inverse Problems*, vol. 18, pp. 575–609, 2002.

[11] S. Billings, "Discrimination and classification of buried unexploded ordnance using magnetometry," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 6, 2004.

[12] K. O'Neill, "Discrimination of uxo in soil using broadband polarimetric gpr backscatter," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 2, 2001.

[13] S. Billings, "Field measurements of induced and remanent moments of unexploded ordnance and shrapnel," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 3, 2009.

[14] Q. Liu, X. Liao, and L. Carin, "Detection of unexploded ordnance via efficient semisupervised and active learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 9, 2008.

[15] H. H. Nelson and J. R. McDonald, "Multisensor towed array detection system for uxo detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 6, 2001.

[16] N. Geng, C. E. Baum, and L. Carin, "On the low-frequency natural response of conducting and permeable targets," *IEEE Trans. Geosci. Remote Sensing*, vol. 37, p. 347359, Jan 1999.

[17] L. Carin, H. T. Yu, and C. E. Baum, "On the wideband electromagnetic inductance signature of conducting and permeable targets," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, pp. 1206–1213, 2001.

[18] 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*, vol. 41, pp. 1005–1015, May 2003.

[19] S. Billings, C. Pasion, S. Walker, and L. Beran, "Magnetic models of unexploded ordnance," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 8, 2006.

[20] K. O'Neill, K. Sun, F. Shubitidze, I. Shamatava, and K. Paulsen, "Accounting for the effects of widespread discrete clutter in subsurface emi remote sensing of metallic objects," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 1, 2006.

[21] V. Sanchez, Y. Li, M. Nabighian, and D. Wright, "Numerical modeling of higher order magnetic moments in uxo discrimination," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 9, 2008.

[22] L. R. Pasion, S. D. Billings, and D. W. Oldenburg, "Joint and cooperative inversion of magnetics and electromagnetic data for the characterization of uxo," in *Proc. Symp. Appl. Geophys. Eng. Environ. Problems*. San Antonio, TX, Sep. 2003.

[23] A. Aliamiri, J. Stalnaker, and E. Miller, "Statistical classification of buried unexploded ordnance using nonparametric prior models," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 9, 2007.

[24] D. Williams, C. Wang, X. Liao, and L. Carin, "Classification of unexploded ordnance using incomplete multisensor multiresolution data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 7, 2007.

[25] W.-H. Lee, P. D. Gader, and J. N. Wilson, "Optimizing the area under a receiver operating characteristic curve with application to landmine detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 2, 2007.

[26] L. Beran and D. Oldenburg, "Selecting a discrimination algorithm for unexploded ordnance remediation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 9, 2008.

[27] B. Zhang, K. O'Neill, J. A. Kong, and T. Grzegorczyk, "Support vector machine and neural network classification of metallic objects using coefficients of the spheroidal mqs response modes," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 1, 2008.

[28] X. Liao and L. Carin, "Migratory logistic regression for learning concept drift between two data sets with application to uxo sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 5, 2009.

[29] Y. Zhang, X. Liao, and L. Carin, "Detection of buried targets via active selection of labeled data: application to sensing subsurface uxo," *IEEE Trans. Geosc. Remote Sensing*, 2004.

[30] R. O. Duda and P. E. Hart, *Pattern classification and scene analysis*. New York: John Wiley, 1973.

[31] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *The Annual Conference on Learning Theory (COLT)*, 1998, pp. 92–100.

[32] U. Brefeld and T. Scheffer, "Co-EM support vector learning," in *Proceedings of the twenty-first international conference on Machine learning*. ACM Press, 2004.

[33] B. Krishnapuram, D. Williams, Y. Xue, A. Hartemink, L. Carin, and M. Figueiredo, "On semi-supervised classification," in *Advances in Neural Information Processing Systems (NIPS)*, 2005.

[34] X. Liao, H. Li, and L. Carin, "Quadratically gated mixture of experts for incomplete data classification," in *Proceedings of the 24th International Conference on Machine learning (ICML'07)*, 2007, pp. 553–560.

[35] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistical Society B*, vol. 39, pp. 1–38, 1977.

[36] M. Jordan and R. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," *Neural Computation*, vol. 6, pp. 181–214, 1994.

[37] P. Liang, D. Klein, and M. I. Jordan, "Agreement-based learning," in *Advances in Neural Information Processing Systems 20*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. Cambridge, MA: MIT Press, 2008, pp. 913–920.

[38] B. Scholkopf and A. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, 2002.

[39] J. H. Albert and S. Chib, "Bayesian analysis of binary and polychotomous response data," *Journal of the American Statistical Association*, vol. 88, no. 422, pp. 669–679, June 1993.