FINAL REPORT

Robust Statistics and Regularization for feature Extraction and UXO Discrimination

SERDP Project MR-1629

JULY 2011

Stephen Billings
Laurens Beran
Doug Oldenburg
Sky Research, Inc. and
University of British Columbia

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.

Executive Summary

Current methods for Unexploded Ordnance (UXO) discrimination using magnetic and electromagnetic induction (EMI) data generally rely on feature vectors extracted from physics based dipole models. These feature vectors are obtained by solving an inverse problem that provides a "best-fit" to the observed data. Typically, this best-fit is defined as the model that minimizes the sum-of-squares of the residuals between observed and predicted data, with each residual weighted by an estimated standard deviation (the-so-called L2 norm). Thus, there is an implicit assumption that the residuals are normally distributed (Gaussian) and that the maximum likelihood solution is the most appropriate model to extract from the data. This assumption of Gaussian statistics may not be appropriate if the residuals have outliers (due to sensor or positional glitches) or if the residuals contain significant structure (model not adequate to represent the data). In those cases, the predicted feature vectors may be significantly in error and should not be relied upon for discrimination. In addition, the maximum likelihood solution does not account for any uncertainty in the recovered feature vectors and may not be the most appropriate criterion to use to assess UXO likelihood.

In this project we researched the statistical structure of the underlying inversion process and developed methods for more accurate extraction of feature vectors from multi-time, multi-frequency and multi-component EMI data.There were four main areas explored with the first three involving different treatments of Bayes equation for combining a-priori knowledge with the constraints imposed by the observed data.

Topic 1: Robust-statistical methods. The L2 norm involves minimizing the sum of squares of the residuals between observed and predicted data. Any outliers in the data (due to glitches, positional error, model mismatch) then excerpt undue influence on the fitted model parameters: the model parameters can change significantly just to accommodate one outlier. The first approach explored in this project was to use robust-statistical norms that down-weight the influence of outliers and result in recovered model parameters that are less sensitive to a few abnormal data-points. Robust-statistical methods effectively use a likelihood function that has fatter tails than the Gaussian distribution corresponding to the L2 norm. Robust statistical methods were able to improve the false-alarm rates encountered at Camp Butner when using both the EM61 production mode data and the MetalMapper cued-interrogation data. In both cases, the primary contribution of the robust-statistical method was the prevention of outliers in the TOI class. For the EM61 we found that the robust statistical method did not result in a significant improvement in the accuracy of the depth (and hence size) estimates. Monte-Carlo simulations revealed that correlated position errors appear to be the dominant cause of depth uncertainty. To militate against this effect, methods that explicitly account for positional uncertainty would need to be used. For the MetalMapper, improved recovery of secondary and tertiary polarizabilities using the robust-statistics algorithm resulted in more TOI identified early using the classifier built on all polarizabilities, with less dependence on the more robust, but less efficient, classifier built on total-polarizability. On a subset of the MetalMapper data collected with a newer instrument and with better field procedures, there was no significant performance improvement when using robust-statistics. As intuitively expected, the robust-methods are most effective when applied to problem datasets. But they also have the desirable attribute of not degrading performance when applied to data free of outliers.

Topic 2: Regularization methods. The next approach explored was to incorporate prior information into the model parameter estimation problem. UXO are typically ferrous and axially symmetric which results in one large and two smaller and equal-polarizabilities. A parameter extraction routine that is biased towards recovering models with minimal difference between secondary polarizabilities will minimize the chance that a target of interest (TOI) is mistaken as harmless scrap. Incorporation of a-priori information results in a regularization problem that involves a trade-off between fitting the data and satisfying the model parameter penalty term. We developed a regularized inversion algorithm that penalizes the deviation

between secondary polarizabilities. Rather than selecting a single model from this inversion process, we input all models into a support vector machine classifier. This corresponds to test features vectors with multiple values for each element. We compare the elements of each test vector with the training data and retain the model value which best corresponds to a given training vector. We also penalize the match of test and training vectors by the likelihood that the model fits the observed data. We find that the regularized method can improve initial performance on high SNR targets with well-constrained secondary polarizabilities while preventing the occurrence of outlying TOI that arise when we rely on unregularized parameters throughout the diglist. The greatest benefit in discrimination performance is obtained with sensor data which interrogates all polarizabilities with orthogonal (horizontal and vertical) primary fields (i.e. MetalMapper).

Topic 3: Incorporating uncertainty into the classification problem. Using the single model that maximizes the a-posteriori probability does not account for any uncertainty in the recovered model parameters when developing a UXO classification strategy. We explored methods for explicitly incorporating model parameter uncertainty in the classification process. Effectively, this involved using the a-posteriori probability to appraise the ensemble of potential models that could have generated the observed data (to within the limits imposed by noise). We first compare a local uncertainty estimate derived from the curvature of the misfit function with global estimates of model posterior probability density (PPD) obtained with Markov chain sampling. For well-posed experiments (i.e. with high SNR and adequate spatial coverage), the two methods of uncertainty appraisal agree. However, when the inverse problem is ill-posed we find that the PPD can be multimodal. To incorporate these uncertainties in discrimination, we first develop an extension of discriminant analysis which integrates over the posterior distribution of the model. When dealing with multimodal PPDs, we show that an effective solution is to input all modes of the PPD-corresponding to all models at local minima of the misfit - into discrimination, and to then classify on the basis of the model which is most likely a UXO.

The method was applied to EM61 and EM63 data sets acquired at Camp Sibert. For both EM61 and EM63 data sets the area under the ROC and the false alarm rate at $P_d = 1$ (i.e. the proportion of false positives required to identify all true positives) are improved. While the improvement for the EM63 data appears negligible (P_{fp} reduced from 0.03 to 0), the identification of one outlying UXO (4.2" mortar) is a significant result from the perspective of a regulator charged with site remediation. Similarly, the significant reduction in false alarm rate at $P_d = 1$ for the EM61 data (P_{fp} reduced from 0.35 to 0.08) improves the likelihood that all ordnance will be identified with this sensor.

Topic 4: Determining when to stop digging. The final issue addressed in this report was the determination of a stop-digging point. It is clear that regulators do not want to leave hazardous items in the ground, so that any strategy for determining an optimal operating point must attempt to recover all TOI. Ideally the stop-dig point lies just after the last TOI has been excavated to prevent excessive numbers of clutter items from being removed. Here we have developed heuristics and statistical criteria for the operating point problem when the total number of instances which must be labelled is known. We derived an approximate probability distribution for the discrete random variable λ, which we defined as the order statistic at which all T instances are labelled. This probability distribution depends upon the generating distributions, prior probabilities, and sample size. Given this probability distribution, we can select an operating point which corresponds to the most likely value of λ. In addition, we have derived a lower bound to the expected value of λ which is a useful approximation to the most likely value. However, this approach has limited practical applicability because it depends upon accurate estimation of the generating distributions and extrapolation into the extreme tails of these distributions. To address this shortcoming, we have proposed a heuristic for selecting N_L, the number of F instances which must occur sequentially before digging is terminated. This is equivalent to cost minimization, but the proposed heuristic provides an objective means of choosing relative costs based upon sample size. In simulations and applications to

real data we find that this technique has an improved probability of finding all ordnance in a test data set, relative to previously published methods. We have limited our investigations to samples on the order of $N = 10^3$, which is representative of the number of detected targets at many sites. Tests on larger data sets should still be carried out.

In previous work we considered a bootstrapping approach to selecting the operating point. While this method may still be viable we find that it is highly dependent upon the training data realization and does not exploit information in the test data as digging proceeds. In contrast, specifying the parameter N_L does not depend on training data and termination of digging depends upon the test data (rather than some pre-specified point derived from limited training data). If successful in finding all TOI, the proposed approach will always overshoot the last target of interest by N_L items, but this is a necessary expense if we are to have confidence that all TOI have been identified. Furthermore, when digging is terminated with this method, we can accurately estimate the distributions of T and F items and use this to compute a confidence that no more targets of interest remain in the ground. A program of verification digging (e.g. using Visual Sample Plan) should also be employed to provide independent confirmation of the stated confidence level.

At the same time that new methods were being developed under this project, comprehensive tests of discrimination performance were conducted at three sites as part of the ESTCP Discrimination Pilot Study. These included the Former Camp Sibert in Alabama, San Luis Obispo in California and Camp Butner in North Carolina. At the demonstration sites, the highest quality data and best discrimination results were achieved in cued-interrogation mode by instruments that dwell at a fixed location while changing the transmitter excitation pattern (e.g. MetalMapper, TEMTADS). ROC curves from the next generation sensor data at both SLO and Camp Butner were near vertical initially (many TOI recovered with low numbers of false-alarms) but often tended to flatten out, with many false alarms excavated before all TOI were recovered. In effect, one part of the discrimination problem (with next-generation sensor platforms deployed in cued-interrogation mode) is relatively easy, with the second part more challenging.

The principal contribution of the work reported here was in developing algorithms and strategies that minimize or eliminate the discrimination outliers encountered during the live-site tests. That is, the methods were particularly efficacious when applied to the "hard" anomalies encountered at a site. By minimizing or eliminating outliers in a UXO discrimination strategy we can alleviate the greatest concerns of the regulatory community: that hazardous UXO are left in the ground at the end of the remediation process.

Contents

Executive Summary i
List of Acronyms v
1. Introduction 1
1.1. Scope of the work conducted 3
2. Robust statistics applied to UXO discrimination 5
2.1. Abstract from paper on robust inversion 6
2.2. Analysis of EM61 data from Camp Butner 7
2.3. Analysis of MetalMapper data from Camp Butner 9
2.4. Discussion 15
3. Regularizing polarizabilities in time-domain electromagnetic inversion 22
3.1. Introduction 22
3.2. Regularized inversion 26
3.3. Multi-object regularized inversion 29
3.4. Discrimination using transverse polarizabilities 30
3.5. Conclusions 33
4. Including uncertainty in parameter estimates during UXO classification 36
5. Selecting a receiver operating point 39
5.1. Optimal operating points 40
5.2. Optimal operating points for samples 42
5.3. Simulations 49
5.4. Application to real data 49
5.5. Discussion 51
6. Discussion 54
6.1. Principal contribution of the work performed under this project 55
6.2. Potential additional research under the robust statistics and regularization theme 55
References 59

List of Acronyms

AUC	Area Under the Curve
cm	centimeter
dB	Decibel
EM	Electromagnetic
EMI	Electromagnetic induction
ESTCP	Environmental Security Technology Certification Program
FAR	False Alarm Rate
FN	False Negative
FY	Fiscal Year
GPS	Global Positioning System
IDA	Institute for Defense Analyses
IMU	Inertial Measurement Unit
in.	Inches
ITC	Information Theoretic Criteria
m	Meter
mm	Millimeter
MDL	Minimum Description Length
MI	Multi-object inversion
MRS	Munitions Response Site
MM	MetalMapper
ms	Millisececond
MTADS	Multi-Sensor Towed Array Detection System
mV	MilliVolt
NFA	Number of False Alarms
OP	Operating Point
PI	Principal Investigator
POC	Point of Contact
QC	Quality Control
RF	Range Fan
ROC	Receiver Operating Characteristic
RTS	Robotic Total Station
SKY	Sky Research, Inc.
SI	Single object inversion
SLO	San Luis Obispo
SVM	Support Vector Machine
TEM	Time-domain electromagnetic
TEMTADS	Time Domain Electromagnetic Towed Array Detection System
TOI	Target of Interest
UBC-GIF	University of British Columbia - Geophysical Inversion Facility
UXO	Unexploded Ordnance

1. Introduction

The 2003 Defense Science Board report on unexploded ordnance (UXO) discrimination projected that a reduction in false alarm rates from 100:1 to 10:1 would save $36 billion on remediation projects within the United States (Delaney and Etter, 2003). This cost reduction was expected to be achieved by improvements in sensor and data processing technologies. These goals have been met, and sometimes exceeded, in recent demonstration projects conducted by the Environmental Security Technology Certification Program (ESTCP), see e.g. Billings et al. (2010). Advances in electromagnetic (EM) sensors have been crucial to these successes: the data provided by multi-static, multicomponent EM platforms are much improved inputs into the inversion and discrimination algorithms applied to this problem. For example, the Time Domain Electromagnetic Towed Array Detection System (TEMTADS) is comprised of an array of 25 horizontal transmitter loops arranged in a 5x5 grid, with horizontal receivers measuring the vertical field arranged concentric to these transmitters (see figure 1). The transmitters are fired sequentially and the secondary field response is recorded in all receivers simultaneously. This multi-static configuration provides a diverse data set which is better able to constrain target depth and transverse polarizabilities than a mono-static sensor. The MetalMapper sensor has also greatly improved the reliability of estimated parameters by transmitting orthogonal primary fields and measuring all components of the secondary field in multiple receivers. Both MetalMapper and TEMTADS systems are deployed in a static mode: previously-detected targets are interrogated with a stationary sensor. This removes the requirement for accurate geolocation that complicates data acquisition with a moving sensor.

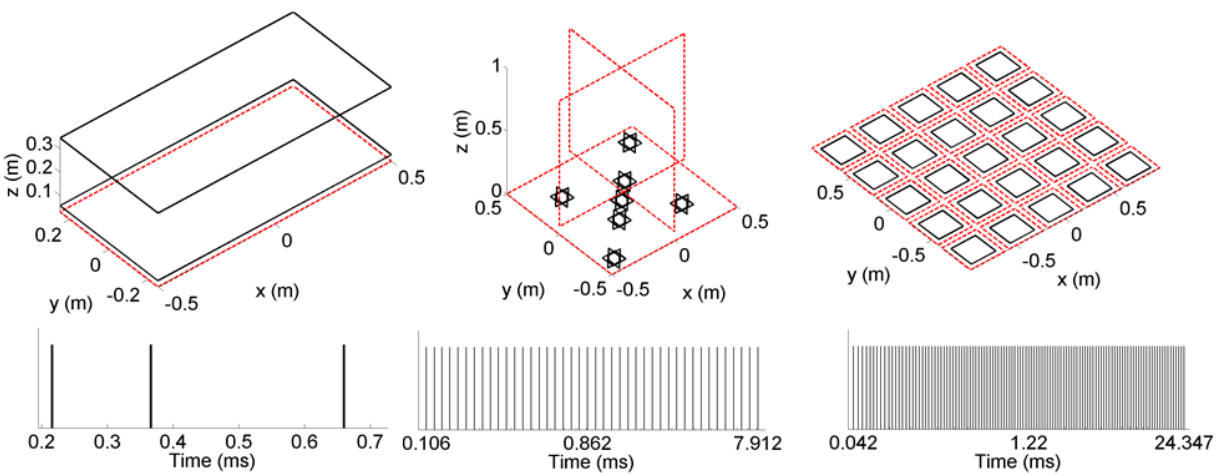


Figure 1. Left to right: Mono-static EM-61 and multi-static MetalMapper and TEMTADS sensors for unexploded ordnance detection and discrimination. Top row shows sensor geometry and bottom row shows time channels

Current methods for UXO discrimination using magnetic and electromagnetic induction data generally rely on feature vectors extracted from physics based dipole models. These feature vectors are obtained by solving an inverse problem that provides a "best-fit" to the observed data. Typically, this best-fit is defined as the model that minimizes the sum-of-squares of the residuals between observed and predicted data, with each residual weighted by an estimated standard deviation. Thus, there is an implicit assumption that the residuals are normally distributed (Gaussian) and that the maximum likelihood solution is the most appropriate model to extract from the data. This assumption of Gaussian statistics may not be appropriate if the residuals have outliers (due to sensor or positional glitches) or if the residuals contain

significant structure (model not adequate to represent the data). In those cases, the predicted feature vectors may be significantly in error and should not be relied upon for discrimination. In addition, the maximum likelihood solution may not be the most appropriate one to recover from the available data. In this project we researched the statistical structure of the underlying inversion process and developed methods for more accurate extraction of feature vectors from multi-time, multi-frequency and multi-component EMI data. Before describing the work conducted we first provide some background information on the forward models used and the basic underlying principles behind our inversion routines.

Throughout this report we use the dipole model (Bell (2005), Pasion (2007)) to predict observed TEM data. The secondary magnetic field is computed as

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

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

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

The induced dipole is the projection of the primary field $\mathbf{B_o}$ onto the target's polarization tensor $\mathbf{M}(t)$. If the primary field is itself modelled as a dipole, then the above expressions indicate that the observed data will have a $1/r^6$ dependence. The polarization tensor is assumed to be symmetric and positive definite and so can be decomposed as

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

with $\mathbf{A}$ an orthogonal matrix which rotates the coordinate system from geographic coordinates to a local, body centered coordinate system.

$$\mathbf{L}(t) = \begin{bmatrix} L_1(t) & 0 & 0 \\ 0 & L_2(t) & 0 \\ 0 & 0 & L_3(t) \end{bmatrix} \tag{4}$$

Typically, the polarization tensor is estimated at a number of discrete time-channels as dictated by the measurement characteristics of the EMI sensor.

Decomposing the polarization tensor with equation 3 parameterizes the model in an orthogonal coordinate system which is assumed to correspond to axial and transverse coordinates of a target. However, this parameterization introduces additional nonlinearity into the forward model: the rotation matrix $\mathbf{A}$ is a nonlinear function of target orientation. An additional source of nonlinearity in the forward model is the $1/r^6$ dependence arising from the dipolar primary and secondary fields. All of these nonlinearities complicate the corresponding inverse problem. Iterative algorithms may converge to local minima of the chosen objective function and thereby produce model estimates which are far from the true parameters. One way to address these complications is to repeatedly solve a related linear problem. If the target location is known, then the forward model for the data at a single time channel is linear in terms of the elements of $\mathbf{M}(t)$. Solving this linear problem over a range of proposed target locations provides a preliminary search of model space and can help identify local minima of the misfit and starting models for subsequent nonlinear inversion.

From a set of N observations, $\mathbf{d}$, of the magnetic or electromagnetic field, the inverse problem is to find the set of model parameters $\mathbf{m} = (x, y, z, \mathbf{M}(\mathbf{t_1}), ..., \mathbf{M}(\mathbf{t_n}))$, that best-fits the data, where (x, y, z) is the estimated dipole position and we assume that there are n time-channels. Predicted data corresponding to the model $\mathbf{m}$ is obtained through the forward modelling operating $\mathbf{d}^{pred} = \mathbf{g}(\mathbf{m})$. This optimization problem is usually solved through maximization of the posterior probability density function (PDF),

$$P(\mathbf{m}|\mathbf{d}) = kp(\mathbf{m})L(\mathbf{d}|\mathbf{m}) \tag{5}$$

where $p(\mathbf{m})$ is the prior-PDF of the model parameters, $L(\mathbf{d}|\mathbf{m})$ is the likelihood-function that represents the fit to the data and k is a normalizing constant. This is the well known Bayes formula. By neglecting the prior (for the present moment) the posterior PDF is determined entirely by the likelihood function which can be written in the form

$$L(\mathbf{d}|\mathbf{m}) = \exp[-\rho(\mathbf{x})] \tag{6}$$

Maximization of the likelihood corresponding to minimization of the norm $\rho(\mathbf{x})$ with

$$\mathbf{x} = \mathbf{W}_d(\mathbf{d} - \mathbf{g}(\mathbf{m})) \tag{7}$$

expressing the (weighted) discrepancy between observed and predicted data. The approach usually taken is to assume that the residuals are normally distributed which corresponds to using an L2, or least-squares norm,

$$\rho(\mathbf{x}) = \mathbf{x}^T\mathbf{x} \tag{8}$$

1.1. **Scope of the work conducted.** There were four main areas explored in the work conducted under this project, with the first three involving different treatments of Equation 5:

(1) *Robust-statistical methods*: The L2 norm involves minimizing the sum of squares of the residuals between observed and predicted data. Any outliers in the data (due to glitches, positional error, model mismatch) then excerpt undue influence on the fitted model parameters: the model parameters can change significantly just to accommodate one outlier. The first approach explored in this project is to use robust-statistical norms that down-weight the influence of outliers and result in recovered model parameters that are less sensitive to a few abnormal data-points. Robust-statistical methods effectively use a likelihood function that has fatter tails than the Gaussian distribution corresponding to the L2 norm. Most of the theoretical material and initial results with robust-statistics are provided in a paper published by Beran et al. (2011a) and are not reproduced here. In section 2, we do present some more recent work undertaken at the Camp Butner discrimination site.

(2) *Regularization methods*: The next approach explored was to incorporate prior information into the model parameter estimation problem of Equation 5. UXO are typically ferrous and axially symmetric which results in one large and two smaller and equal-polarizabilities. A parameter extraction routines that is biased towards recovering models with minimal difference between secondary polarizabilities will minimize the chance that a target of interest (TOI) is mistaken as harmless scrap. Incorporation of a-priori information results in a regularization problem with methods for solution discussed in section 3.

(3) *Incorporating uncertainty into the classification problem*: Using the single model that maximizes the a-posteriori probability does not account for any uncertainty in the recovered model parameters when developing a UXO classification strategy. We explored methods for explicitly incorporating model parameter uncertainty in the classification process. Effectively, this involved using the a-posteriori probability of Equation 5 to appraise the ensemble of potential models that could have generated the observed data (to within the limits imposed by noise). The theoretical basis and results of the methods that incorporate uncertainty in the classification process are described in Beran et al. (2011b) and are not reproduced here, with just a brief summary presented in section 4.

(4) *Determining when to stop digging*: The final issue addressed in this report is the determination of a stop-digging point. It is clear that regulators do not want to leave hazardous items in

the ground, so that any strategy for determining an optimal operating point must attempt to recover all TOI. Ideally the stop-dig point lies just after the last TOI has been excavated to prevent excessive numbers of clutter items from being removed. The analysis we present in section 5 demonstrates that determination of an optimal operating point involves understanding the tails of the distributions of TOI and non-TOI.

2. Robust statistics applied to UXO discrimination

Robust inversion algorithms in the geophysical and statistical literature modify the misfit function so that outlying data have downweighted contributions to the total misfit. In the context of geophysical inversion, Farquharson and Oldenburg have demonstrated the use of various robust norms for measuring both data misfit and model norm (Farquharson and Oldenburg, 1998). Here we follow their development of an iterative approach to minimizing these norms for linear and nonlinear inverse problems. The proceeding section developed some of the preliminary material and introduced the concept of the norm and forward model $\mathbf{g}(\mathbf{m})$. For a linear problem, we can write $\mathbf{g}(\mathbf{m}) = \mathbf{G}\mathbf{m}$. Now to minimize ϕ with respect to the model vector $\mathbf{m}$ we have

$$\begin{aligned} \frac{\partial \rho}{\partial \mathbf{m}} &= \frac{\partial \rho}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial \mathbf{m}} \\ &= \mathbf{B}^T \mathbf{R} \mathbf{x} \end{aligned} \tag{9}$$

with

$$\mathbf{R} = diag\left(\frac{\partial \rho / \partial \mathbf{x}}{\mathbf{x}}\right) \tag{10}$$

and

$$B_{ij} = \frac{\partial x_i}{\partial m_j}. \tag{11}$$

For a linear forward problem then

$$\begin{aligned} \frac{\partial \phi}{\partial \mathbf{m}} &= \mathbf{B}^T \mathbf{R} \mathbf{x} \\ &= \mathbf{G}^T \mathbf{W}_d^T \mathbf{R} \mathbf{W}_d (\mathbf{d} - \mathbf{G}\mathbf{m}). \end{aligned} \tag{12}$$

Setting the above expression equal to zero and solving for $\mathbf{m}$ yields an expression identical to that in a standard least squares problem, except for the presence of $\mathbf{R}$. This matrix depends upon $\mathbf{x}$ the weighted residual, and so is a function of the model $\mathbf{m}$. Hence minimization of an arbitrary norm becomes a nonlinear problem even when the forward problem is linear. This nonlinearity can be circumvented with the "iteratively reweighted least squares" (IRLS) algorithm, which iteratively updates the model according to the following procedure

(1) Set $R_{ii} = 1$ and solve for $\mathbf{m}$ with equation 12.
(2) Update the elements of $\mathbf{R}$ (equation 10) using the current estimate of $\mathbf{m}$
(3) Recompute $\mathbf{m}$ and iterate to 2 until convergence.

For a nonlinear forward problem, we can use the linearization

$$F(\mathbf{m} + \delta\mathbf{m}) \approx F(\mathbf{m}) + \mathbf{J}\delta\mathbf{m} \tag{13}$$

to derive an expression for the model perturbation $\delta\mathbf{m}$ at each iteration, with the sensitivity matrix $\mathbf{J}$ taking the place of the forward modelling operator $\mathbf{G}$ in equation 12. IRLS provides a general procedure for minimizing a norm, table 2 summarizes a number of norms which appear in the literature. Figure 2 compares these norms as a function of $\mathbf{x}$ and also shows the resulting weightings R_{ii}. The bisquare norm is unique amongst the norms considered here in that it has the capability to completely disregard outlying data ($R_{ii} = 0$ for $|x_i| > k$). For this reason it is recommended by statistical practitioners for linear regression applications. However, some care must be taken in initializing IRLS for the bisquare norm, since the algorithm is not guaranteed to converge for this norm (Marrona et al., 2006). We find that initializing IRLS with the least squares solution works well in this application.

Norm	$\rho(x)$
Huber	$\begin{cases} 1/2x^2, & \lvert x\rvert \le k \\ k\lvert x\rvert - 1/2k^2, & \lvert x\rvert > k \end{cases}$
Eckblom	$(x^2+k^2)^{p/2}$
Bisquare (Tukey)	$\begin{cases} \frac{k^2}{6}\left(1-\left[1-(x/k)^2\right]^3\right), & \lvert x\rvert \le k \\ k^2/6, & \lvert x\rvert > k \end{cases}$

TABLE 1. Example norms used for robust inversion

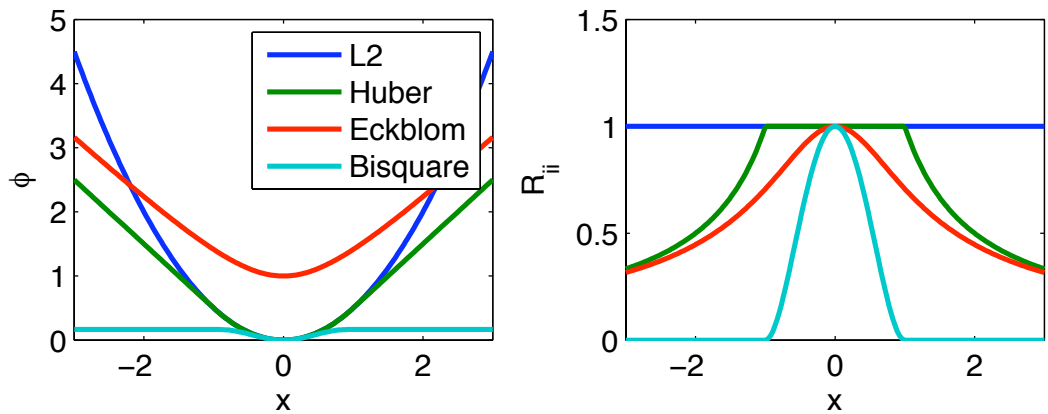


FIGURE 2. Left: comparison of norms as a function of x, the weighted discrepancy between observed and predicted data. Right: weightings (R_{ii}) applied to data for IRLS minimization of robust norms. For all norms $k = 1$, and $p = 1$ for the Eckblom p norm.

The reasoning and algorithms behind our robust statistical methodology and initial applications to field data are described in Beran et al. (2011a). In this section of the final report, we describe new work that we have not published previously on the Camp Butner ESTCP discrimination study data. The production mode EM-61 data and the MetalMapper data collected in cued-mode represent the two most interesting datasets, from the perspective of robust statistics, within that discrimination study. The reader is referred to the relevant ESTCP discrimination reports for additional details on these two datasets.

2.1. **Abstract from paper on robust inversion.** We refer the interested reader to the paper Beran et al. (2011a) for more details of the robust-inversion methods and initial applications to Geonics EM-61 and EM-63 data collected at Camp Sibert, Alabama. The abstract of the paper follows: *We invert time-domain electromagnetic data for the purpose of discriminating between buried unexploded ordnance (UXO) and non-hazardous metallic clutter. The observed secondary magnetic field radiated by a conductor is forward modelled as a linear combination of decaying, orthogonal dipoles. We show via a perturbation analysis that errors in the measurement of sensor position propagate to non-normal errors on the observed data. A least squares (L2) inversion assumes normal errors on the data, and so non-normal errors have the potential to bias dipole parameter estimates. In contrast, robust norms are*

designed to downweight the effect of outlying (noisy) data and so can provide useful parameter estimates when there is a non-normal component to the noise.

When positional errors are modelled as independent Gaussian perturbations, we find that weighted least squares and robust inversions have comparable performance. Both inversion techniques estimate data uncertainties from observed data, and this has the effect of "robustifying" the least squares inversion. However, when simulated errors are correlated, robust inversion with a bisquare norm provides a marked improvement over L2 inversion. Application of robust inversion to real data sets from Camp Sibert, Alabama produced an incremental improvement to the initial L2 inversion, identifying outlying ordnance items and improving discrimination performance.

2.2. **Analysis of EM61 data from Camp Butner.** We reanalyzed the EM-61 data from Camp Butner using a robust-statistical inversion algorithm. The algorithm uses the bisquare norm in place of the weighted least-squares algorithm we used for the discrimination study. The inversion algorithm was improved over our previous incarnations in Beran et al. (2011a) by

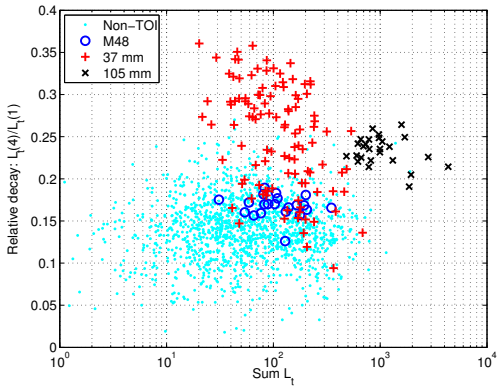
(1) Using the bisquare norm to select an optimal position and depth of the dipole model (previously we had used the least-squares position and depth);
(2) Incorporation of constraints when solving for the non-decomposed polarizability matrix to ensure that M is positive definite.

Figure 3 shows a feature space comprised of a time-decay feature versus a size based feature. In our ESTCP submissions we had used the sum of the polarizabilities to calculate both the size (sum of all polarizabilities at all channels) and time-decay features (ratio of 4th to 1st channels). One improvement we made with this plot was to use the maximum time-decay calculated from either the total-polarizability and the principal polarizabilty. This had the effect of suppressing some potential outliers when the total-polarizability was not well constrained. The results for the least-squares and robust algorithms look reasonably similar. The main difference between the two is that the robust norm has eliminated the worst time-decay outliers: the minimum time-decay of any TOI is 0.13 compared to 0.095 for the least-squares algorithm. For our scoring submissions for the ESTCP study we used the time-decay to rank the digging order: items with slower decay were dug earlier. With our modified time-decay feature and the least-squares algorithm we would need to dig 88% of the non-TOI to recover all TOI (Figure 3c) compared to 58% of the non-TOI for the robust statistics (Figure 3d). This involves a reduction from 1640 to 1080 unnecessary digs.

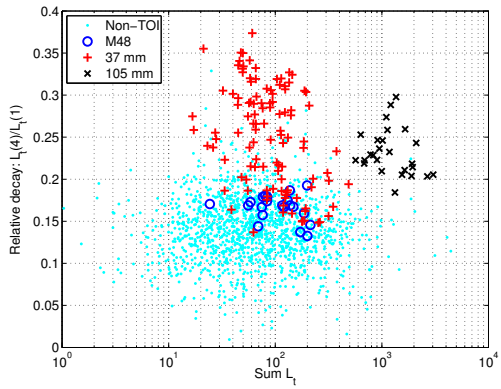
Figure 4a compares the ROC curves for the least-squares and robust norm classification methods that threshold on the time-decay. The performance is virtually identical up to the point where the false-alarm rate is 40%. The superior ability of the robust statistical algorithm to prevent outliers results in it reaching $P_{TP} = 1$ much sooner than the least-squares algorithm. Figure 4b shows how the performance varies if the size parameter is also included in the classification ranking. For both least-squares and robust-statistics the ranking strategy is modified by taking a weighted sum of the size and time-decay parameters. Again the robust statistical method produces lower false-alarm rates than least-squares. Performance gains against the method that used time-decay only for ranking are marginal and the increased complexity and risk associated with including the size parameter are not justified.

One of the reasons the EM61 discrimination ability is so limited is that the object depth, and hence size, is poorly resolved (Figure 5a-b). There is a strong tendency to over-predict the item depth, particularly for the small, shallow non-TOI (Figure 5d). The robust-statistical algorithm displays a slight improvement in the accuracy of the recovered depths for the TOI items but does not improve the depth accuracy for the non-TOI (in fact these appear to be slightly worse).

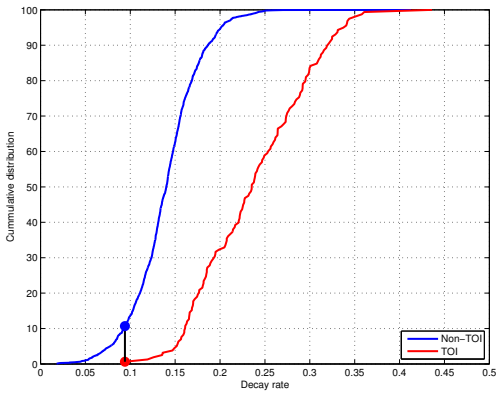
The depth recovery problem is well illustrated by the object that caused anomaly 165. The ground-truth indicates that the item was a fuze at 1 cm depth, but both the robust (56 cm) and least-squares (70 cm) algorithms placed the item at significantly greater depths (Figure 6a). Comparison of observed



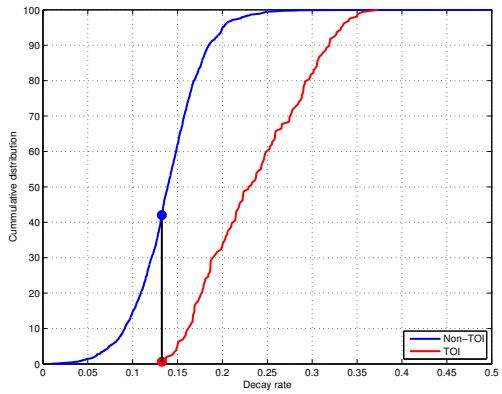
(a) Least Squares feature space



(b) Robust feature space

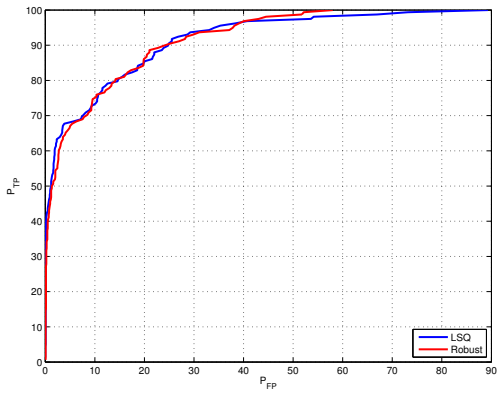


(c) Least Squares time-decay

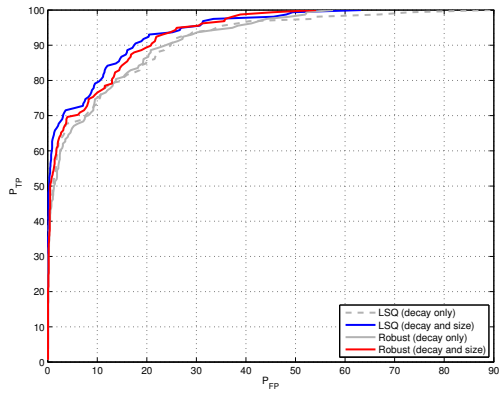


(d) Robust time-decay

FIGURE 3. Comparison of EM61 derived features using least-squares and robust statistical algorithms.



(a) ROC using time-decay



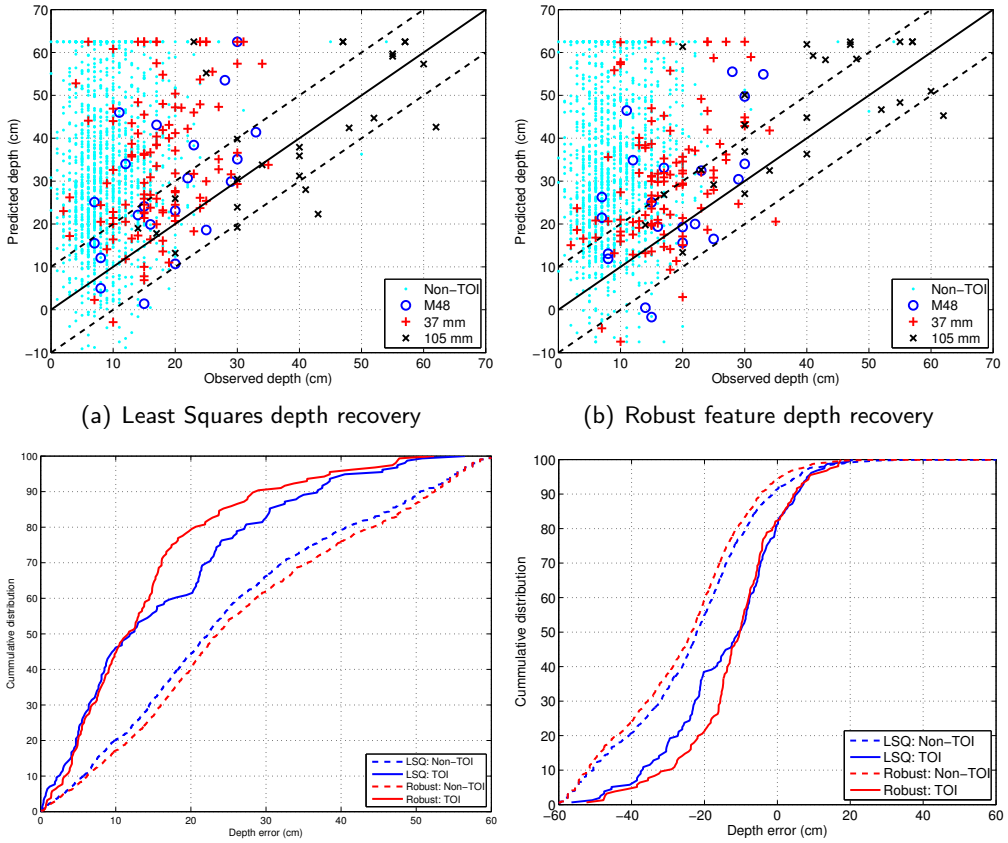
(b) ROC using time and size

FIGURE 4. Comparison of discrimination performance of least-squares and robust statistical algorithms. A slight improvement results if both size and time-decay are utilized.

data with best-fitting predicted data at difference depths (Figure 6b) shows that the deeper models provide a better fit to the observed data. At the shallower depths, the model is unable to match all of the observed peaks in the profile data. We believe that this is caused by correlated position errors of the sensor positions from different transects across the area. To test this assertion, we used the same positions as anomaly 165 and modelled a synthetic 37 mm projectile that was horizontal at a depth of 0 and oriented 45° from North. We then corrupted the positions by both random and correlated position errors of different sizes before calculating a set of synthetic data (which was also corrupted by additive noise of 5 mV standard deviation). For random errors we added normally distributed pertubations of a specified standard deviation to each of the three position coordinates. For the correlated errors we added a randomly generated shift to all the Northing values in each transect, with the sign of the error alternating between each line (that is all Northward collected lines were moved North and all Southward collected lines were moved South). Figure 7 shows the misfit-versus depth curves for two realisations for each of the different error scenarios. An inversion algorithm will seek the model with minimum misfit. With no positional error, there is a local minima at a depth of approximately 50 cm, but the fit is significantly worse than when the model is placed at the surface. With correlated positional errors the deeper model often becomes preferred relative to the shallower model. We repeated the calculation of the misfit versus depth curves for 100 different realisations and kept track of the best-fitting depth in each case (Figure 8). The results indicate that, for this case, sensor noise alone never causes the deeper model to be preferred. The deeper model becomes much more common with increasing levels of correlated positional perturbations.

2.3. **Analysis of MetalMapper data from Camp Butner.** We also applied the robust statistical algorithm to the MetalMapper data collected at Camp Butner. Positions and depths recovered by the least-squares method were generally very accurate and no significant differences between the depths recovered by robust and least-squares methods were found. Figure 9 compares the polarizabilities obtained by the different methods on data collected at a test-pit at Camp Butner. It was discovered after the survey that the Y-component of the 3rd receiver was faulty. The robust inversion method is, as expected, much more tolerant of the outlying data compared to the least-squares inversion. This is particularly evident in the feature space plots of Figure 10, which compare time-decay, size and asymmetry feature vectors obtained by the least-squares and robust methods. The time-decay and size feature vectors were calculated using the total-polarizability: the time-decay feature vector as the ratio of time-channels at 8.1 ms and 160μs, and the size size feature vector as the sum of all 42 time-channels between 160μs and 8.1 ms. The asymmetry parameter was calculated as the sum of the difference between the secondary and tertiary polarizabilities normalized by the secondary polarizability. All asymmetry calculations were made using all time-channels between 160μs and 1 ms. The asymmetry parameter will be zero for radially symmetric steel objects. There is considerably less scatter in the feature-vector plots obtained through robust-statistics and the asymmetry values are generally smaller. If the faulty receiver is excluded from the inversion process, then the polarizabilites recovered by the least-squares and robust methods are almost identical (Figures 9c-d).

Figure 11 compares time-decay and size based feature vectors obtained by the least-squares and robust methods on all of the live-site data anomalies. At the Camp Butner site, two vendors, Geometrics and SKY, were involved in the data collection. In was previously pointed out by us (in our ESTCP report) and others that the SKY data were of a higher quality than the Geometrics data (the Geometrics system was used on the test-plot). This is evident in the plots showing the feature space broken out across the two methods (Figures 11c-d) where there is considerably greater variability shown within the feature vectors derived from the Geometrics data. These differences are even more evident when we plot a feature space comprising the object size and asymmetry (Figure 12). There is considerably less variability of



(a) Least Squares depth recovery

(b) Robust feature depth recovery

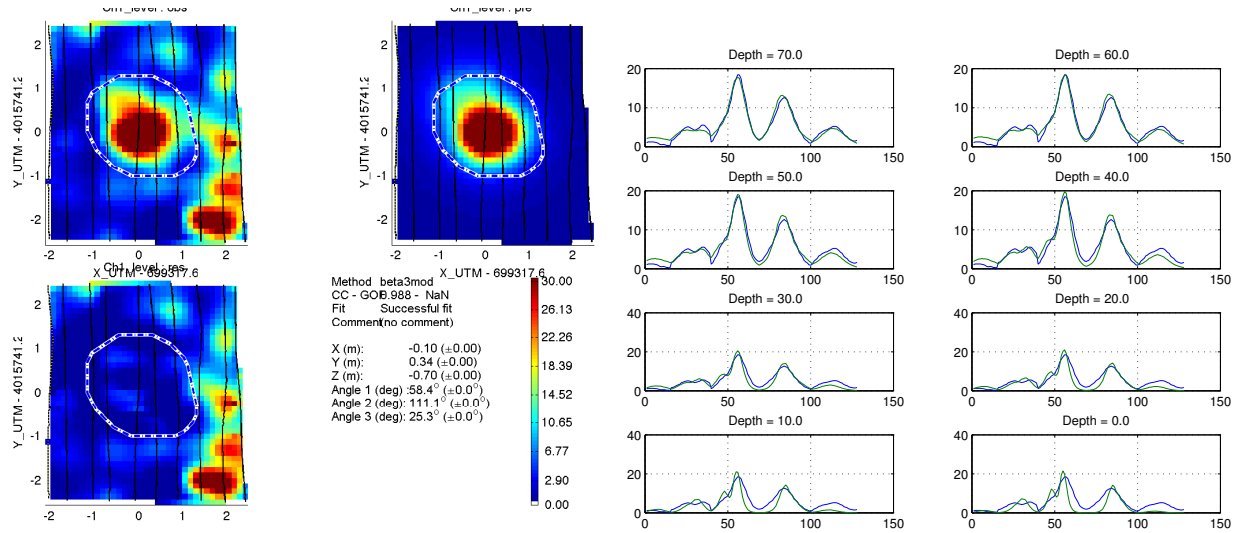
(c) Cumulative depth error (absolute)

(d) Cumulative depth error (signed)

FIGURE 5. Comparison of EM61 derived depths using least-squares and robust statistical algorithms. The signed cumulative depth plot is defined as observed minus predicted depth.

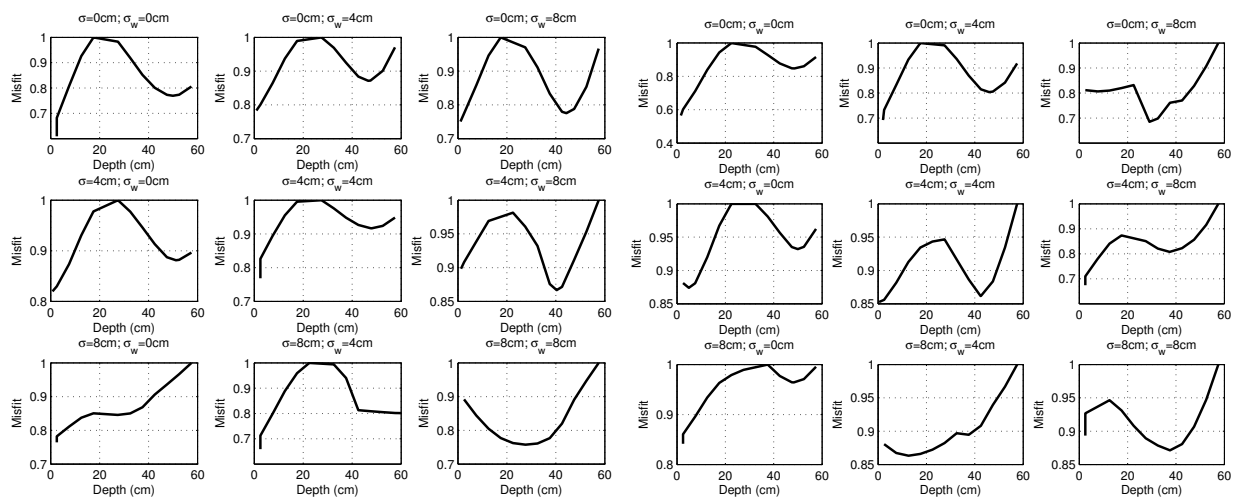
the asymmetry parameter when calculated from the polarizabilities recovered by the robust statistical algorithm.

Even with the robust statistics algorithm, there are still some outliers in the size versus asymmetry space (Figure 12). Figure 13a plots the asymmetry parameter versus the offset from the center of the MetalMapper array. The worst outliers are offset by more than 30 cm from the array center and arise because the MetalMapper cannot constrain all the polarizabilities when the object is too far from the array center. The positional offset does not explain all of the outliers. In an effort to better understand the underlying cause of the outliers, we calculated linearized error estimates of the polarizabilities. These are straightforward to obtain for a fixed position as the problem is linear (Smith and Morrison, 2005) . Figure 13b plots the error in the total-polarizability (square root of the trace of the model covariance matrix) against the asymmetry parameter. The total uncertainty tends to be large for the 105 mm projectiles as these are deeply buried and reflect the increasing uncertainty with depth. This is evident in Figure 13c which plots total polarizability error versus item depth. The percentage error in polarizability Figure 13d is a more useful metric. It identifies all of the asymmetry outliers (7 of them with percentage error $> 10\%$) and anomaly 2504 which was an item that many venders had trouble identifying.



(a) Inversion result (b) Data fits at different depths

FIGURE 6. (a) Least-squares inversion of anomaly 165. (b) Comparison of EM-61 observed and predicted data at time-channel 3 for anomaly 165.



(a) Realization 1 (b) Realization 2

FIGURE 7. Impact of positional errors on misfit versus depth curves. Rows contain increasing amounts of random positional error (σ), while columns have increasing amounts of correlated positional errors (σ_w).

The "model error" metrics computed in Figure 13 do not directly include any information on the misfit between observed and predicted data. We therefore computed a "data error" metric by taking the ratio of the sum of squares of the residuals divided by the sum of squares of the observed data. This metric was calculated using the time-channels between 160 and 460 μs after pulse turn-off. Figure 14 compares the model and data errors against asymmetry for the anomalies collected by Geometrics

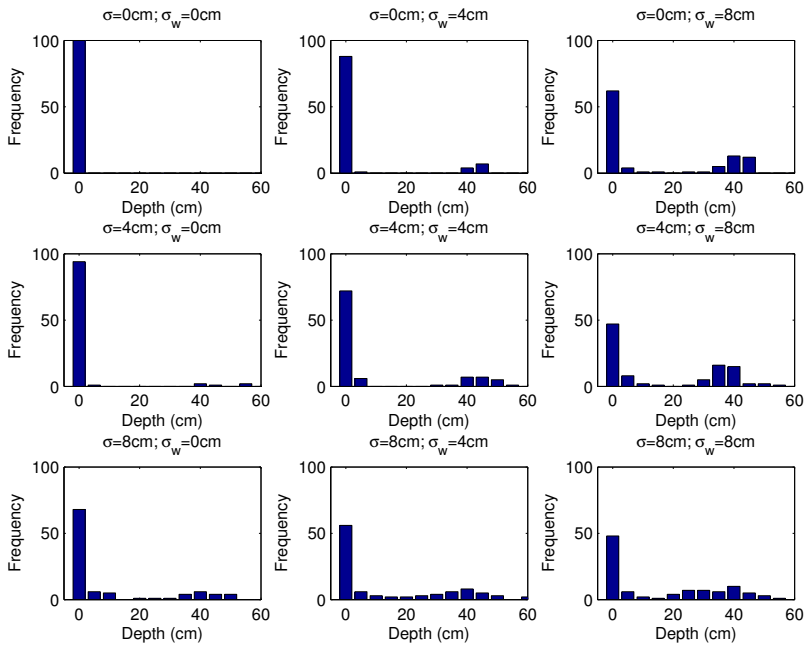
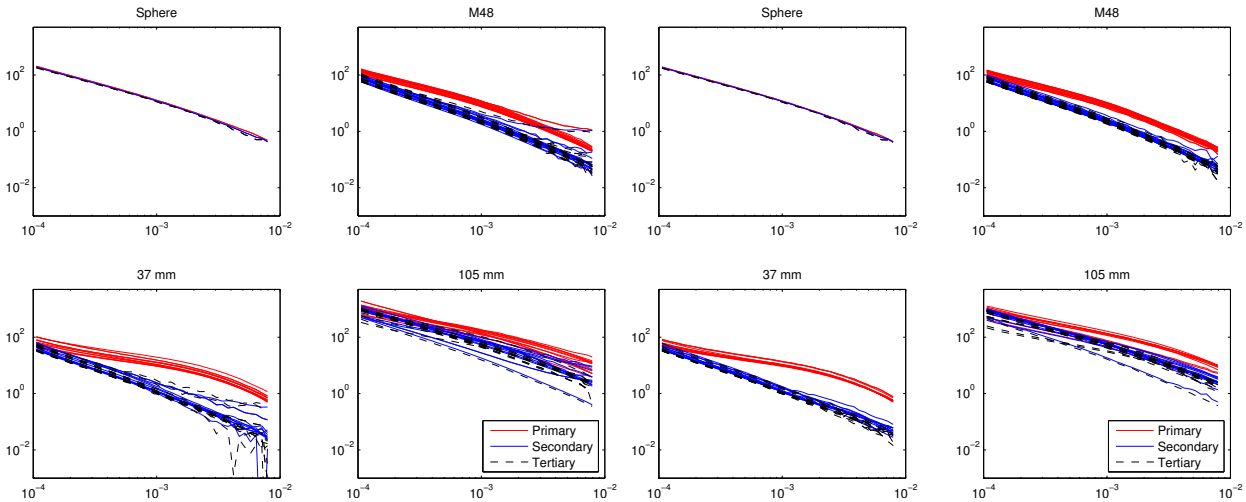


FIGURE 8. Histogram of best fitting depths for 100 realizations of different positional pertubations. Rows contain increasing amounts of random positional error (σ), while columns have increasing amounts of correlated positional errors (σ_w).

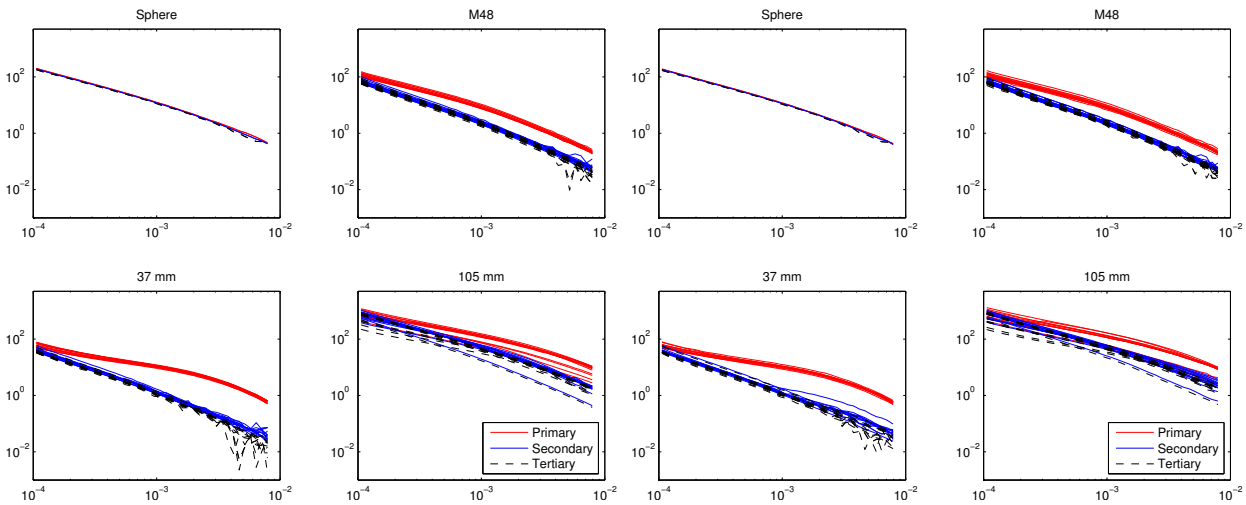
and those collected by SKY. Both error metrics are generally larger for the Geometrics derived datasets, reflecting our earlier comments on the higher quality of the SKY derived data. Percentage data error is larger for the non-TOI items for both datasets: this occurs because most of the clutter is small and data amplitudes are generally low. The cumulative distribution plots in Figures 14e-f show that 85% of TOI and 30% of non-TOI of the Geometrics anomalies have model errors of less than 10% compared to 98% of TOI and 90% of non-TOI for the Sky anomalies. For data errors the comparison is 65% of TOI and 15% of non-TOI for Geometrics anomalies against 78% of TOI and 30% of non-TOI.

Turning now to classification performance, we employ a similar method to that used for our MetalMapper submissions for Camp Butner that were generated under the MM-1004 project. The recovered polarizabilities across all time-channels were used to train two support-vector machine classifiers. The first used all polarizabilities (all channels of the primary polarization and the first 15 channels of the secondary polarizations) while the second used just the total-polarizability (all time-channels). The classifier that used all polarizabilities was trained using 35 items measured in a test-pit (all UXO) and 18 items from the live-site. The 18 items were chosen on the boundaries between the obvious clusters of TOI items and comprised 5 TOI items. The classifier that used total-polarizabilities was trained using the same 40 TOI items, but for non-TOI we did not use the training data and instead used the 500 items that were furthest from the TOI items. These were typically the small, fast decaying items that are obviously non-TOI and produce a more effective classifier than when using the 13 non-TOI in our training set. In our original submission, we combined the two classifiers by selecting the highest ranked classifier output: this procedure was intended to minimize the chance of false-negatives occurring for TOI with poorly constrained secondary polarizabilities. We utilized two variations of this procedure to generate the ROC curves shown in Figures 15:



(a) Least Squares (All receivers)

(b) Robust (All receivers)

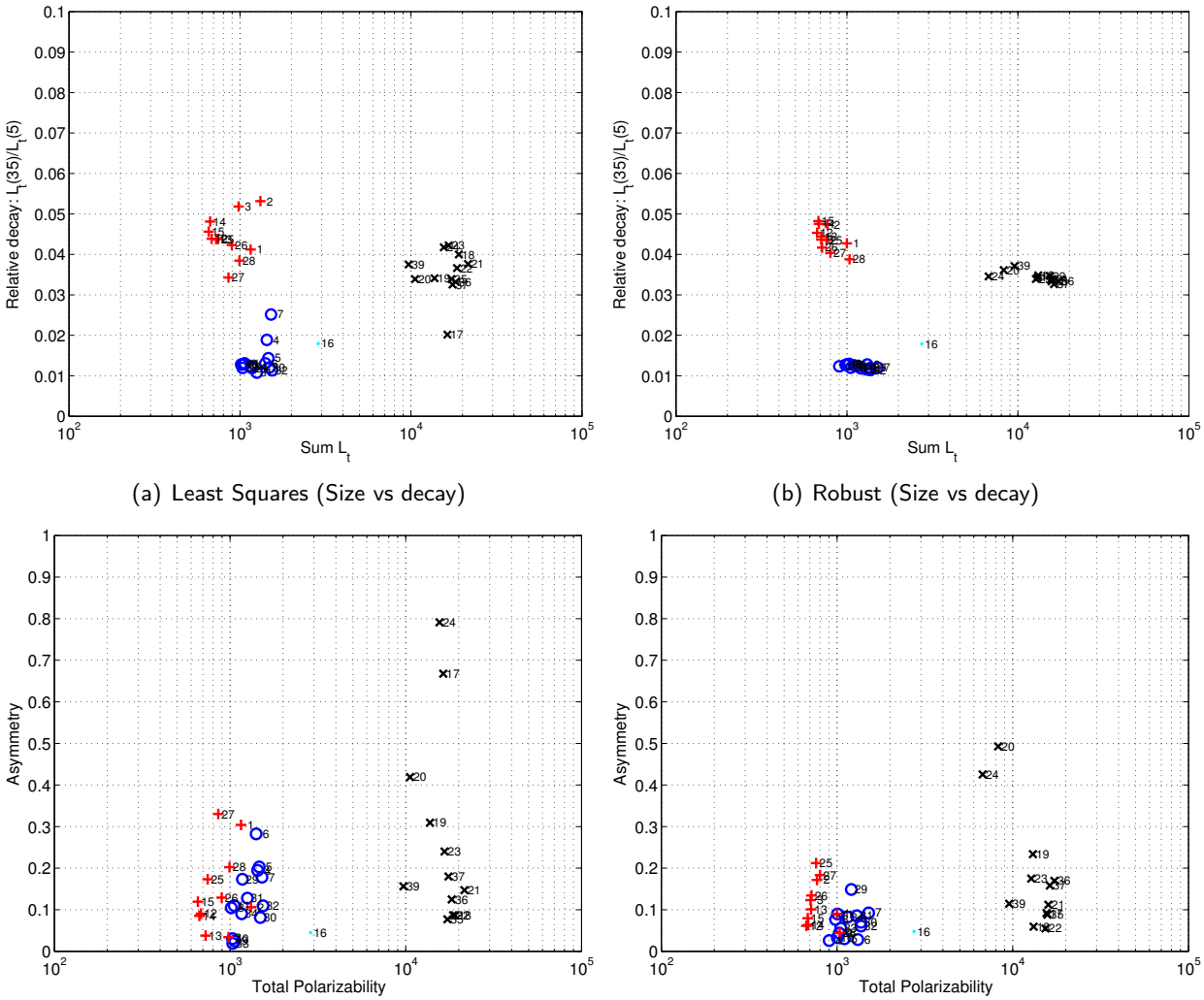


(c) Least Squares (Excluding Rx3Y)

(d) Robust (Excluding Rx3Y)

FIGURE 9. Comparison of polarizabilities extracted over the test-pit using least-squares and robust statistical algorithms. In the bottom row the faulty Y-component of sensor 3 was excluded.

(1) Combined using a cutoff on the SVM applied to all polarizabilities. In this method the highest ranked items were all those anomalies with SVM(all polarizabilities)>0.5, with the SVM(total polarizability) controlling the ranking for the remainder of the items. The 0.5 cutoff was chosen by inspection of the SVM values in the dataset. It appeared to lie just above a large cluster of non-TOI items. The model and data errors did not explicitly influence the ranking scheme.
(2) Combined using percentage data and model errors. In this method the dig-order from the all-polarizability ranking was used except for those anomalies where either data or model error was above the pre-defined thresholds of 15% and 10% respectively and the total-polarizability ranking placed the item higher in the digging order.



(a) Least Squares (Size vs decay)

(b) Robust (Size vs decay)

(c) Least Squares (Size vs asymmetry)

(d) Robust (Size vs asymmetry)

FIGURE 10. Comparison of polarizabilities extracted over the test-pit using least-squares and robust statistical algorithms. In the bottom row the faulty Y-component of sensor 3 was excluded.

Figures 15a-b show the ROC curves for the combined methods when using feature vectors recovered by least-squares and robust-statistics along with the ROC curve for the MetalMapper submission made under project MR-1004[1]. Also shown are the ROC curves that would be obtained by using the SVM trained on all polarizabilities and on the total-polarizability. The first combined method has an identical ROC curve to the all-polarizabilities method until the point where SVM=0.5 at which point it switches to the total-polarizability method. For both the least-squares and robust feature vectors switching over to the total-polarizability causes a significant reduction in the number of false-alarms required to dig all TOI: from 724 down to 633 for least-squares and from 814 down to 434 for robust statistics (Table 2). The final false-alarm rates of the second combined method are similar to those from the first method

[1]Note that we use the revised submission made under MR-1004 that corrected a QC error in the original submission

but the ROC curves do not rise quite as steeply as the first method in the early stages. This occurs because the ranking based on all-polarizabilities is very effective at distinguishing TOI with high-quality data from non-TOI: it just becomes a much less effective method when the data-quality degrades.

The last five TOI recovered by the first method and robust feature vectors are anomalies 1346, 1154, 2405, 1298 and 2504 (very last one). Four of these items are obvious outliers in the model and data error plots provided in figures 13 and 14. Anomaly 2504 is the worst outlier with a model error percentage of almost 60%. Clearly, we should not even rely on the total-polarizability for anomalies with such large model errors. If we assign all anomalies with a model error > 50% to a "can't decide" category (there are 53 of them with just one TOI, anomaly 2504) and insert them at the point where SVM(total polarizabilies)=0, then we get the ROC curves shown in Figures 15c-e. The ROC curve flattens out at the point where the "can't decide" anomalies are dug, but rises to 1 much faster than a ranking scheme that does not use a "can't decide" category. For instance, with robust statistics and the first combined method the false-alarm rate is reduced from 434 to 225 (Table 2).

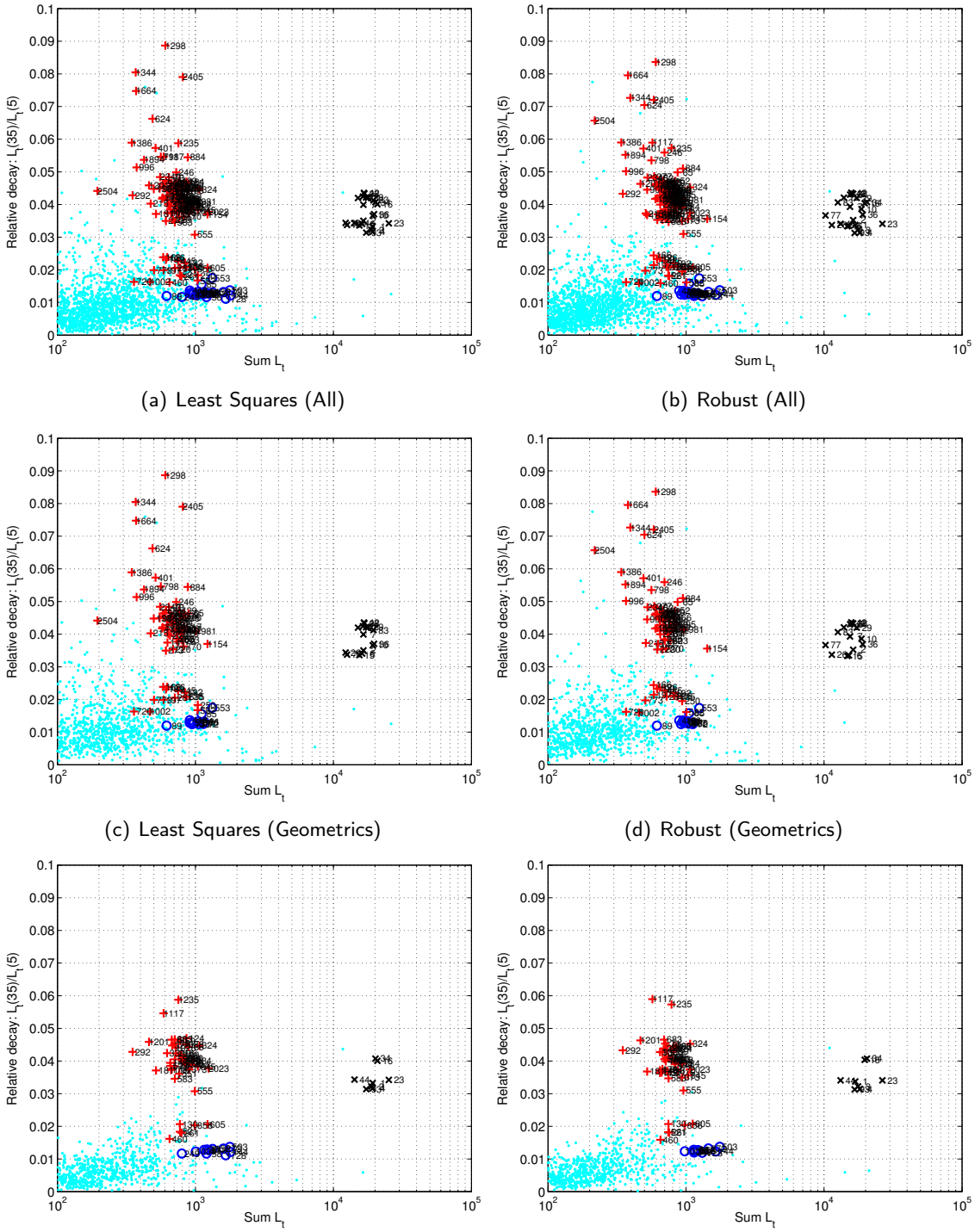
On all combinations of classifier the robust statistics derived feature vectors were better than the least-squares feature vectors (Table 2). The best-performing least-squares method (first method with can't decide) required 319 false-alarms, which was 94 items more than the equivalent robust-statistics derived ranking scheme. As was evident in the feature plots, there is a significant difference in the discrimination performance when the Geometrics and SKY collected data are treated separately (Figure 16 and Table 3). For the Geometrics collected data the false-alarm rate of the robust-statistics derived feature vectors using method 1 (164 items) is 99 items less than the best performing classifier built using least-squares derived feature vectors (263 items). For the SKY collected data, the false-alarm rates of the classifiers built using the different feature vectors are very similar. The good performance of least-squares derived feature vectors implies that the higher-quality data from the SKY system do not contain significant outliers that would benefit from the application of a robust inversion routine (this was evident in the feature vector plots in Figures 11 and 12).

Note that adding the false-alarm rates for the SKY and Geometrics systems in Table 3 does not produce the same false-alarm rate for the combined systems in Table 2. This is because there is a subset of SKY collected false-alarms that occur after the last SKY collected true-positive and before the last Geometrics collected true-positives (and vice-versa, depending on which system has the last ranked TOI). These items are included in the false-alarm rate for the combined system but not when false-alarm rates for each are tallied separately.

	Least Squares		**Robust statistics**	
Method	**No can't decide**	**With can't decide**	**No can't decide**	**With can't decide**
Total polarizabilities	633	427	434	313
All Polarizabilities	724	400	814	818
Combined (SVM < -0.5)	633	368	434	225
Combined (%error)	909	319	555	262
As submitted	736	403		

TABLE 2. False alarm rates for the different ranking methods applied to MetalMapper data at Camp Butner.

2.4. **Discussion.** Robust statistical methods were able to improve the false-alarm rates encountered at Camp Butner when using both the EM61 production mode data and the MetalMapper cued-interrogation data. In both cases, the primary contribution of the robust-statistical method was the prevention of outliers in the TOI class. For the EM61 we found that the robust statistical method did not result in a



(a) Least Squares (All)

(b) Robust (All)

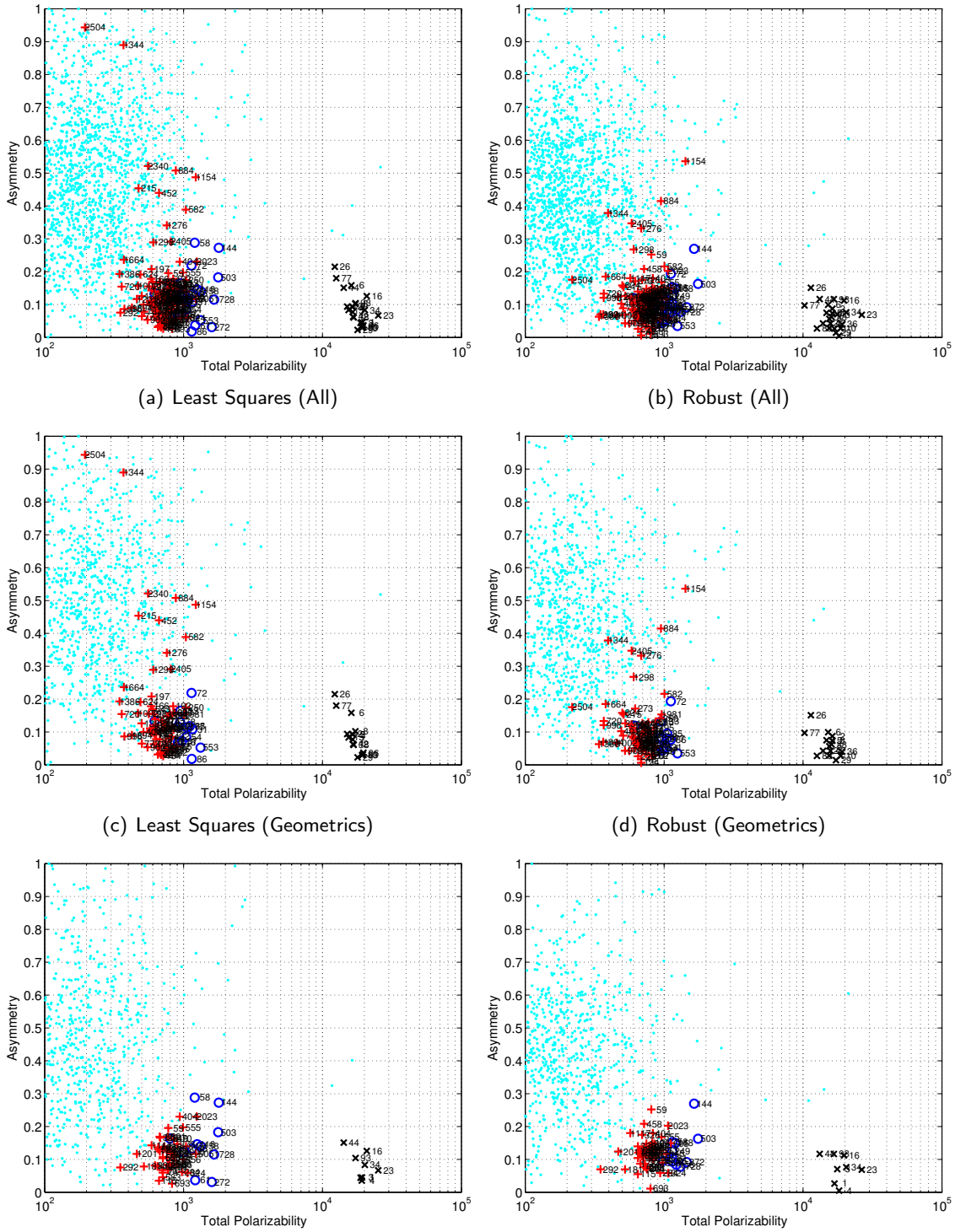
(c) Least Squares (Geometrics)

(d) Robust (Geometrics)

(e) Least Squares (SKY)

(f) Robust (SKY)

FIGURE 11. Comparison of MetalMapper time-decay and size related features using least-squares and robust statistical algorithms.



(a) Least Squares (All)

(b) Robust (All)

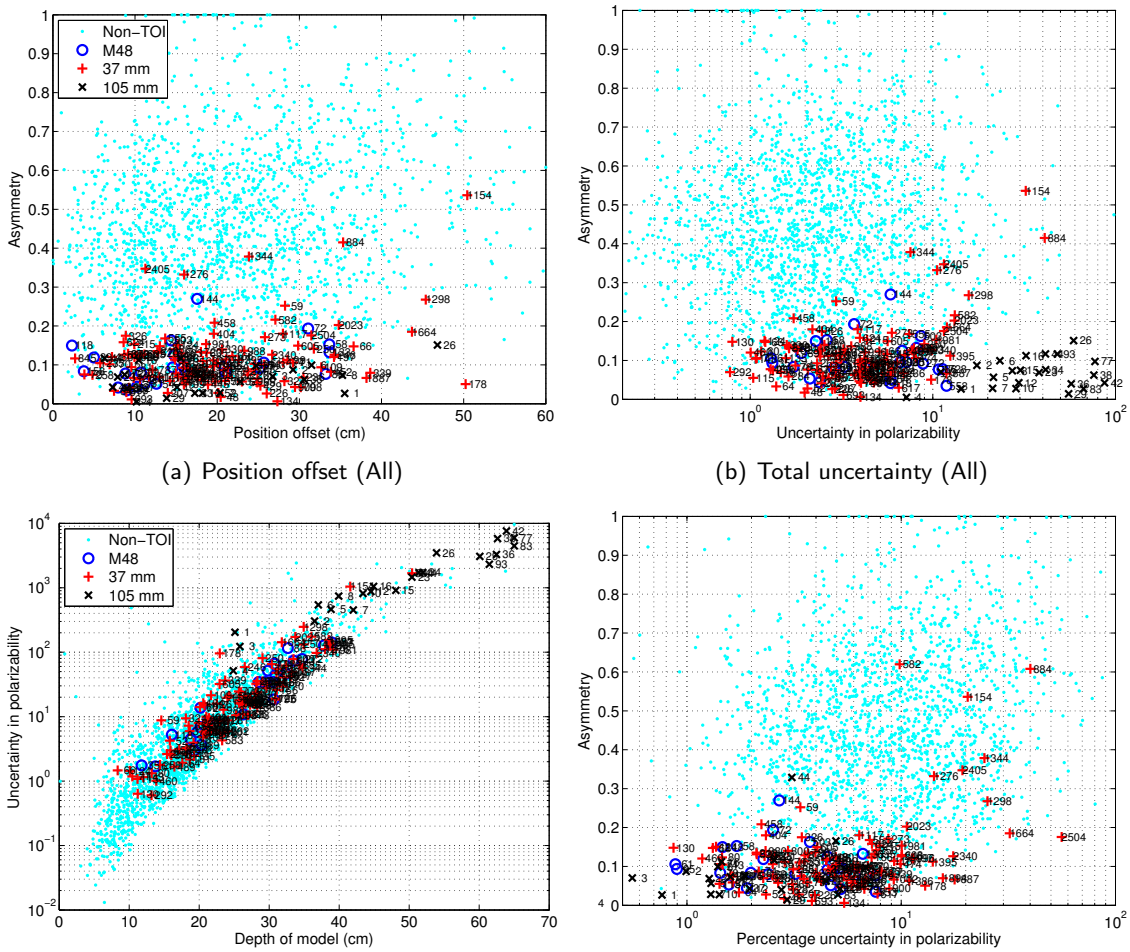
(c) Least Squares (Geometrics)

(d) Robust (Geometrics)

(e) Least Squares (SKY)

(f) Robust (SKY)

FIGURE 12. Comparison of MetalMapper asymmetry and size related features using least-squares and robust statistical algorithms.



(a) Position offset (All)

(b) Total uncertainty (All)

(c) Total uncertainty versus depth (All)

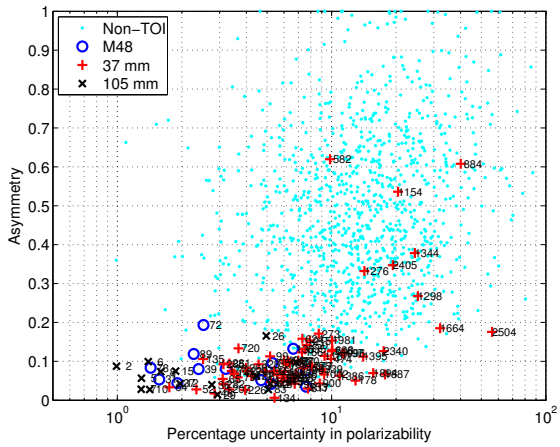
(d) Percentage uncertainty(All)

FIGURE 13. Error analysis of MetalMapper data

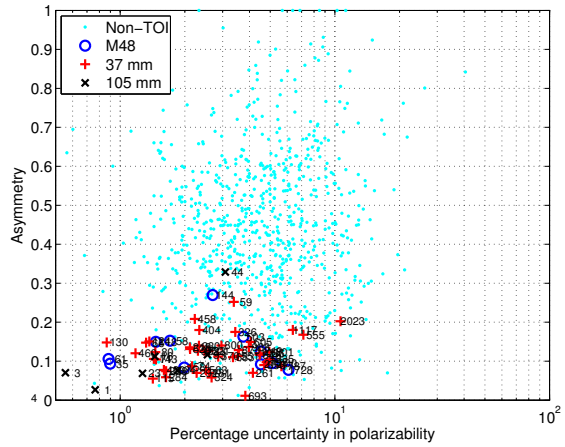
	Geometrics		Sky	
Method	**Least-squares**	**Robust**	**Least-squares**	**Robust**
Total polarizabilities	348	235	18	16
All Polarizabilities	324	551	17	17
Combined (SVM < -0.5)	302	164	23	10
Combined (%error)	263	182	23	22
As submitted	250		21	

TABLE 3. False alarm rates for the different ranking methods applied to MetalMapper data at Camp Butner, split by data collection group.

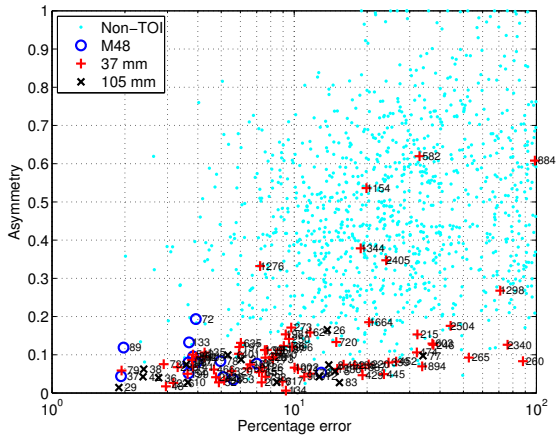
significant improvement in the accuracy of the depth (and hence size) estimates. Monte-Carlo simulations revealed that correlated position errors appear to be the dominant cause of depth uncertainty. To militate against this effect, methods that explicitly account for positional uncertainty would need to be used.



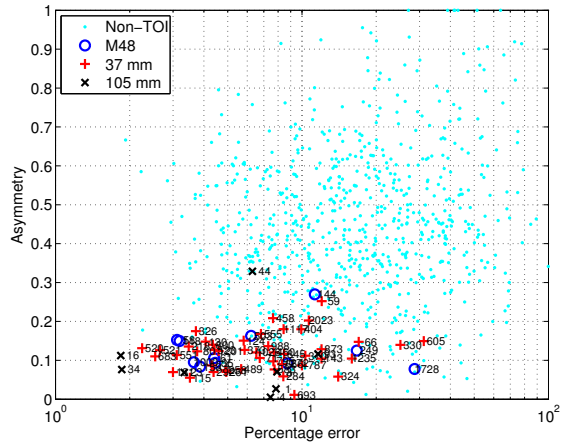
(a) Percentage model uncertainty (Geometrics)



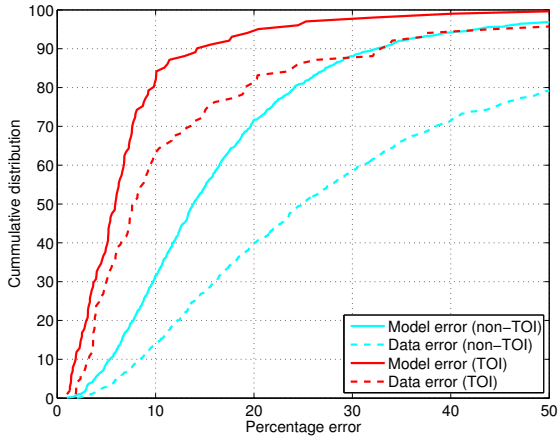
(b) Percentage model uncertainty (SKY)



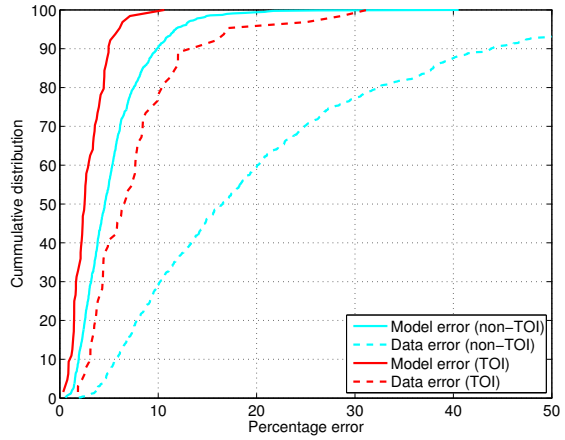
(c) Percentage data uncertainty (Geometrics)



(d) Percentage data uncertainty (SKY)

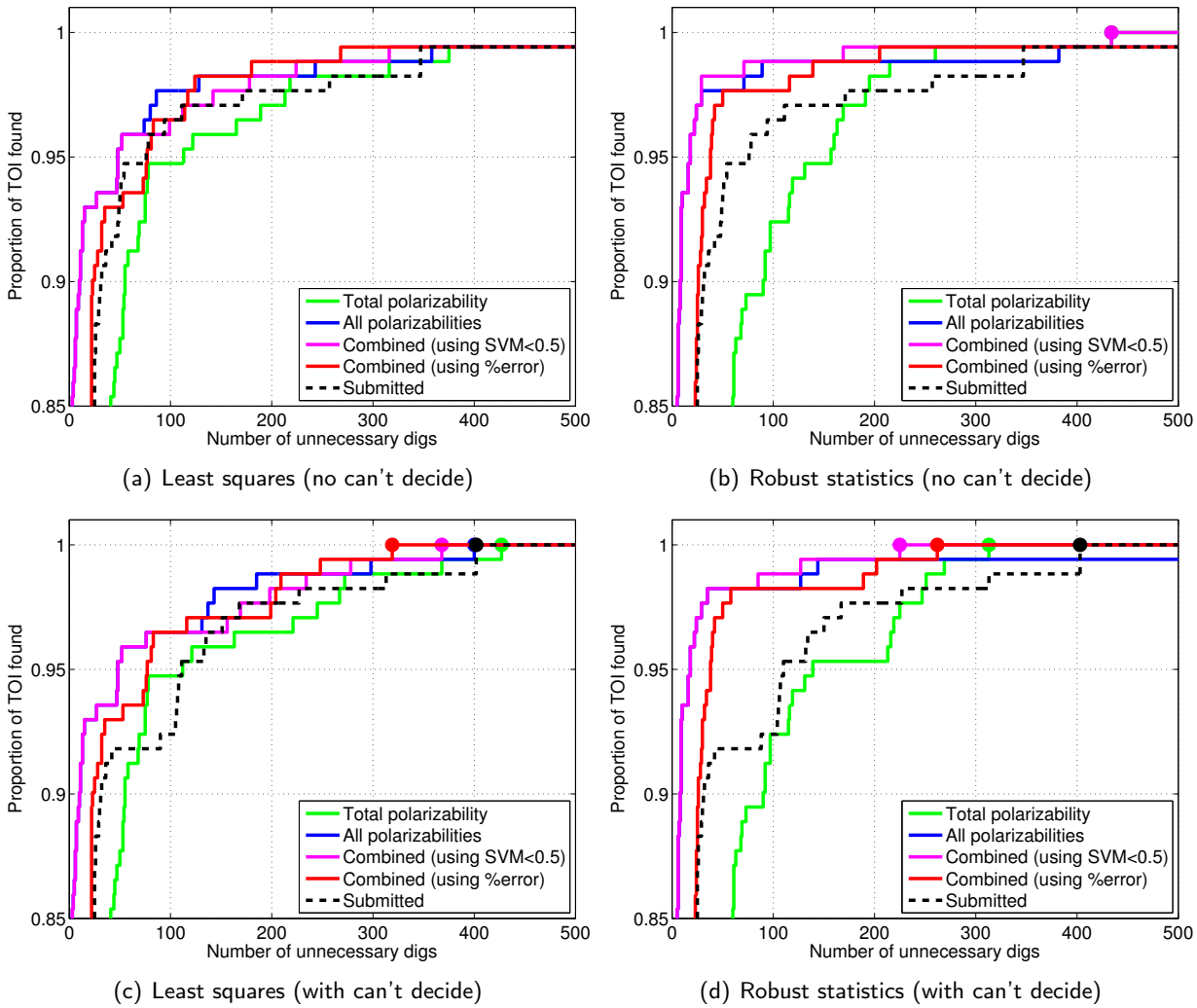


(e) Cummulative error (Geometrics)



(f) Cummulative error uncertainty (SKY)

FIGURE 14. Error analysis of MetalMapper data split between Geometrics and SKY collected datasets



(a) Least squares (no can't decide)

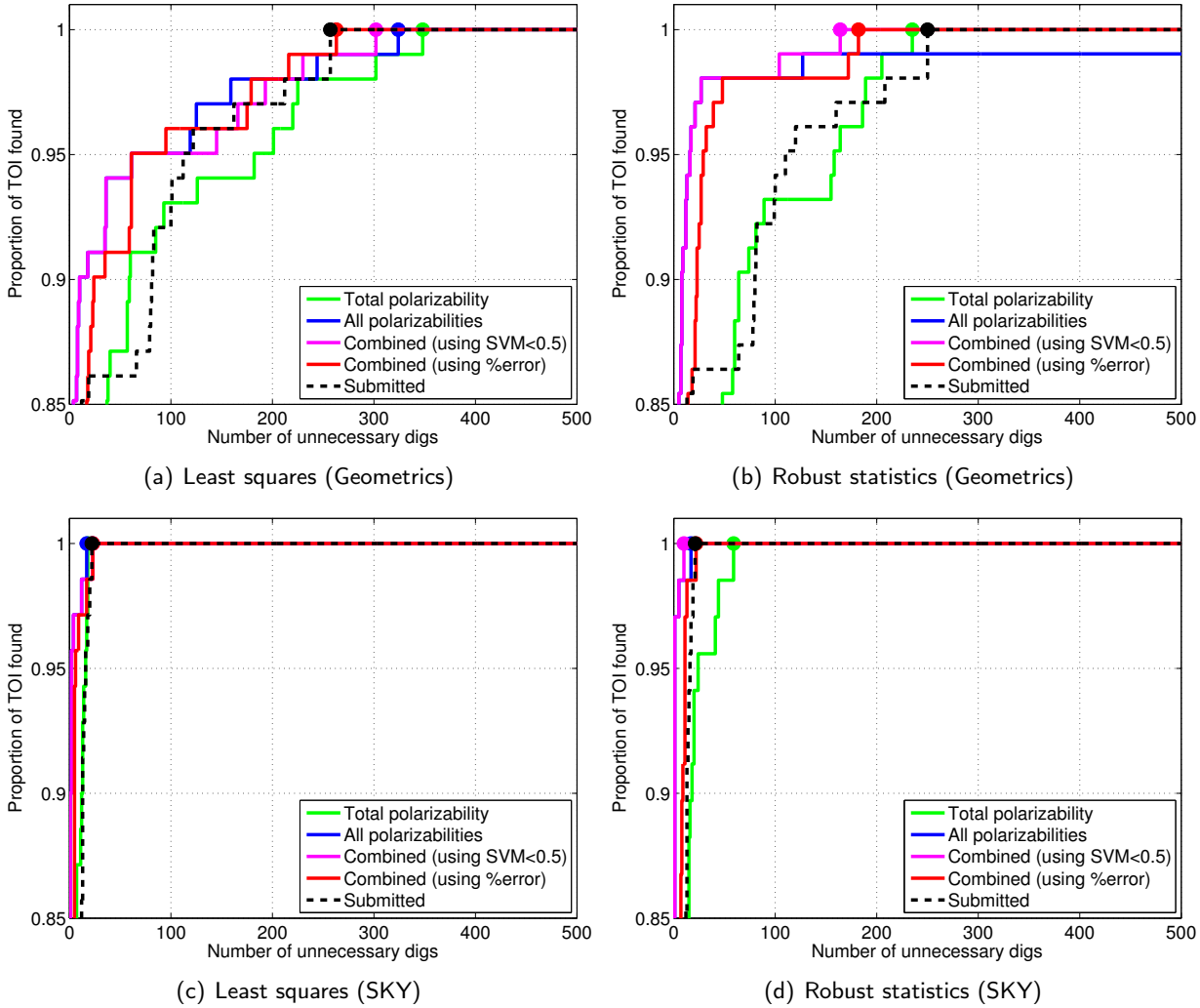
(b) Robust statistics (no can't decide)

(c) Least squares (with can't decide)

(d) Robust statistics (with can't decide)

FIGURE 15. Top left-hand corner of ROC curves for least-squares (left column) and robust statistics (right column) feature vectors either without (top row) or with (bottom row) a "can't decide" category based on model error of $> 50\%$.

For the MetalMapper, improved recovery of secondary and tertiary polarizabilities using the robust-statistics algorithm resulted in more TOI identified early using the classifier built on all polarizabilities, with less dependence on the more robust, but less efficient, classifier built on total-polarizability. On a subset of the MetalMapper data collected with a newer instrument and with better field procedures, there was no significant performance improvement when using robust-statistics. As intuitively expected, the robust-methods are most effective when applied to problem datasets. But they also have the desirable attribute of not degrading performance when applied to data free of outliers.



(a) Least squares (Geometrics)

(b) Robust statistics (Geometrics)

(c) Least squares (SKY)

(d) Robust statistics (SKY)

FIGURE 16. Top left hand corner of ROC curves for feature-vectors derived from least-squares and robust statistics algorithms and incorporating a can't decide category, split by data collection group.

3. Regularizing polarizabilities in time-domain electromagnetic inversion

3.1. **Introduction.** In the introduction to this report we discussed the dipole forward model and the concept of the polarizability tensor $\mathbf{M}(t)$. The polarizability tensor is assumed to be symmetric and positive definite and so can be decomposed as

$$\mathbf{M}(t) = \mathbf{A}^T \mathbf{L}(t) \mathbf{A} \tag{14}$$

with $\mathbf{A}$ an orthogonal matrix which rotates the coordinate system from geographic coordinates to a local, body centered coordinate system. The diagonal eigenvalue matrix $\mathbf{L}(t)$ contains the principal polarizabilities $L_i(t)$ $(i = 1, 2, 3)$, which are assumed to be independent of target orientation and location. Features derived from the dipole model, in particular amplitude and decay of the principal polarizibilities, have been successfully used to discriminate between targets of interest (TOI) and non-hazardous metallic clutter. These parameters are useful because, to first order, a conductor can be modelled as a simple LR loop which is inductively coupled to transmitters and receivers on the surface. The current response of this loop is a decaying exponential which is fully described by an amplitude and time constant (West and Macnae, 1991). The TEM dipole model generalizes this simple circuit model to account for target size and shape. This latter property is represented by the principal polarizabilities, which decay independently in time and are approximately aligned with the semi-major and minor axes of the target.

Equal transverse (secondary and tertiary) polarizabilities indicate an axisymmetric target (Bell and Barrow, 2001). Most ordnance can be treated as bodies of revolution (Shubitidze et al., 2002), and so equality of transverse polarizabilities has been proposed as a useful feature for discriminating between TOI and irregularly-shaped clutter. However, in practice it has been difficult to reliably estimate target shape from observed TEM data. This is because mono-static, vertical-component sensors conventionally deployed for unexploded ordnance detection often cannot adequately interrogate the transverse response of buried targets. Figure 17 illustrates this effect for a spherical target illuminated by a mono-static Geonics EM-61 sensor (geometry and time channels are shown in figure 1). In order to excite the

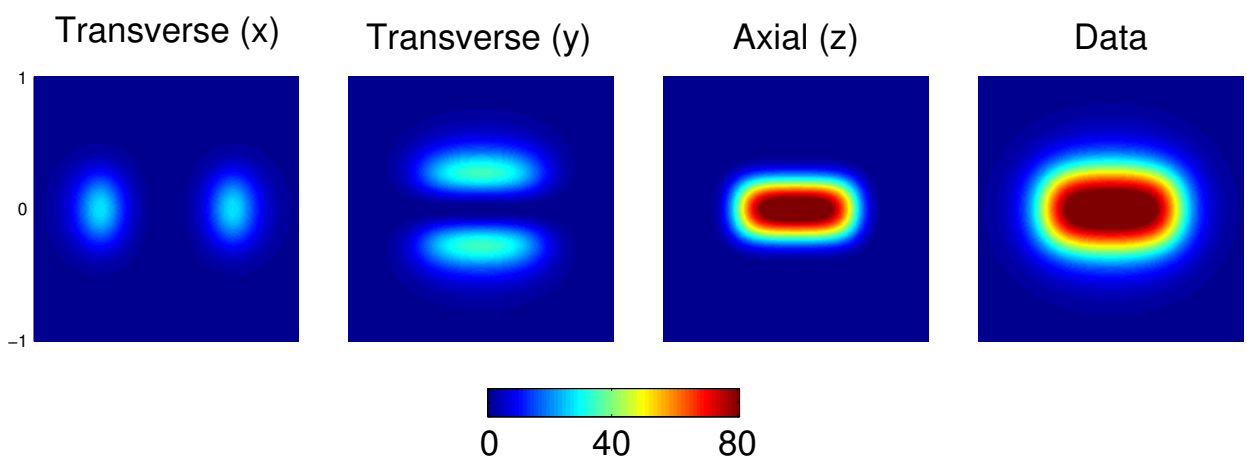


Figure 17. Components of the dipole response over a target positioned at $\mathbf{r} = [0,\ 0,\ -0.3]$ m for the EM-61. Predicted data are linear combination of axial and transverse responses, here for a spherical target with polarizabilities $L_i = 1$, $i = 1, 2, 3$. Excitation of transverse responses requires a horizontal standoff, resulting in a lower SNR than for axial excitation.

transverse response of the target the sensor must be positioned with a horizontal stand-off from the

target. Assuming an approximately dipolar field from transmitter and target, the secondary field decays approximately as $1/r^6$ with increasing sensor-target separation r. For this reason, the axial polarizability response dominates the measured data. Data which are sensitive to transverse polarizations tend to have low signal to noise, so that inverted parameters for axisymmetric targets will not necessarily have approximately equal transverse polarizabilities. Early attempts to estimate target shape from mono-static sensors therefore proved unsuccessful (e.g. Bell and Barrow (2001)).

Multi-static TEM sensors designed for UXO detection have helped address these limitations. Figure 18 shows the components of the dipole response for the TEMTADS sensor over the same target as in figure 17. The data received for transmitters immediately adjacent to the center transmitter are primarily sensitive to a combination of the transverse (x and y) excitations. Inversion of these data therefore produces less uncertain estimates of transverse polarizabilities. We illustrate this in figure 19

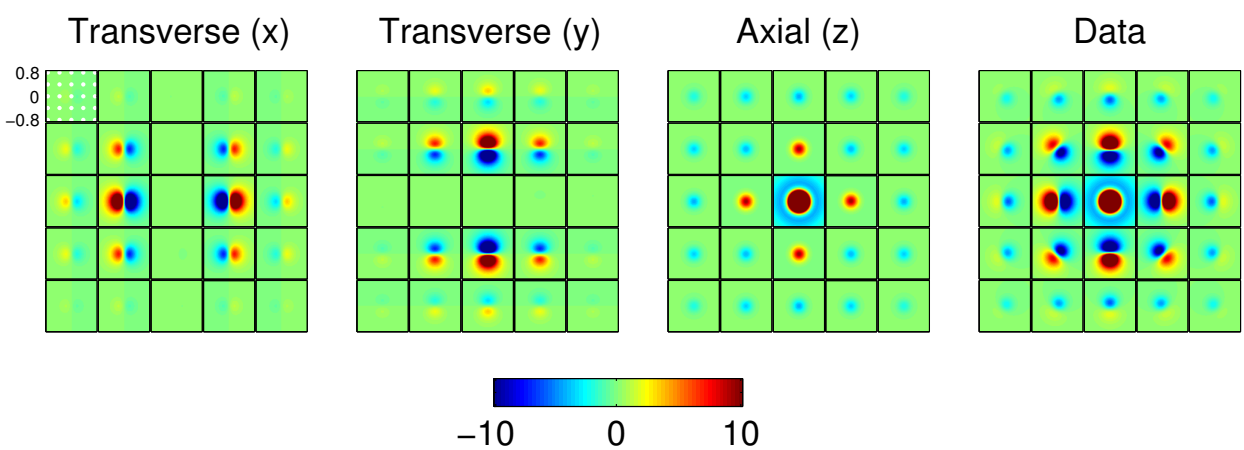


FIGURE 18. Components of the dipole response over a spherical target for the 5 x 5 TEMTADS array. Each subplot shows the received field in all receivers excited by the corresponding transmitter in the array. The locations of receivers are indicated by the white markers in the top left subplot.

by considering the condition number of the linear forward operator $\mathbf{G}$ mapping from model $\mathbf{m_M}$ to predicted data $\mathbf{d}^{pred}$

$$\mathbf{d}^{pred} = \mathbf{G}\mathbf{m_M}. \tag{15}$$

Here the model $\mathbf{m_M}$ is comprised of the six unique elements of the polarizability tensor $\mathbf{M}$ at a single time channel

$$\mathbf{m_M} = [M(1,1),\ M(1,2),\ M(1,3),\ M(2,2),\ M(2,3),\ M(3,3)]^T. \tag{16}$$

For a datum at location $\mathbf{r}_i$ (relative to the target), the corresponding row in $\mathbf{G}$ can be expressed as (Song et al., 2011)

$$\mathbf{G}_i = \begin{bmatrix} B_s^x B_p^x \\ B_s^x B_p^y + B_s^y B_p^x \\ B_s^x B_p^z + B_s^z B_p^x \\ B_s^y B_p^y \\ B_s^y B_p^z + B_s^z B_p^y \\ B_s^z B_p^z \end{bmatrix}^T \tag{17}$$

with B_p the primary field at the target and B_s the secondary field at the receiver, with all fields implicitly dependent upon $\mathbf{r}_i$. Superscripts denote the x, y, z components of the respective fields.

The predicted data in equation 15 are a linear function of the polarizability tensor, so that if target location $\mathbf{r}$ is known, least squares parameter estimates can be directly obtained via the pseudoinverse $\mathbf{G}^\dagger$

$$\hat{\mathbf{m}}_{\mathbf{M}} = \mathbf{G}^\dagger \mathbf{d}^{obs} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{d}^{obs} \tag{18}$$

The least squares solution is a linear combination of the observed data weighted by the (inverse) singular values of $\mathbf{G}$ (see Golub and van Loan (1989)). Small singular values s_i of $\mathbf{G}$ amplify errors on the data and so the condition number

$$\mathrm{cond}(\mathbf{G}) = \max_i s_i / \min_i s_i \tag{19}$$

is a diagnostic of ill-posedness. In figure 19 we see that the condition number for the EM-61 can increase unexpectedly as a function of target depth for data acquired with a 50 cm line spacing (and 10 cm along line spacing). This indicates that the inversion can be ill-posed if the survey is in an unlucky configuration relative to a given target, so that $\mathbf{G}$ has small singular values. This can be alleviated by decreasing line spacing to 25 cm: here data coverage is sufficient that the inversion is well-posed over the entire range of target depths. However, in practice this tight line spacing is prohibitively slow to acquire relative to multi-static sensors that can fully interrogate a target with a single sounding.

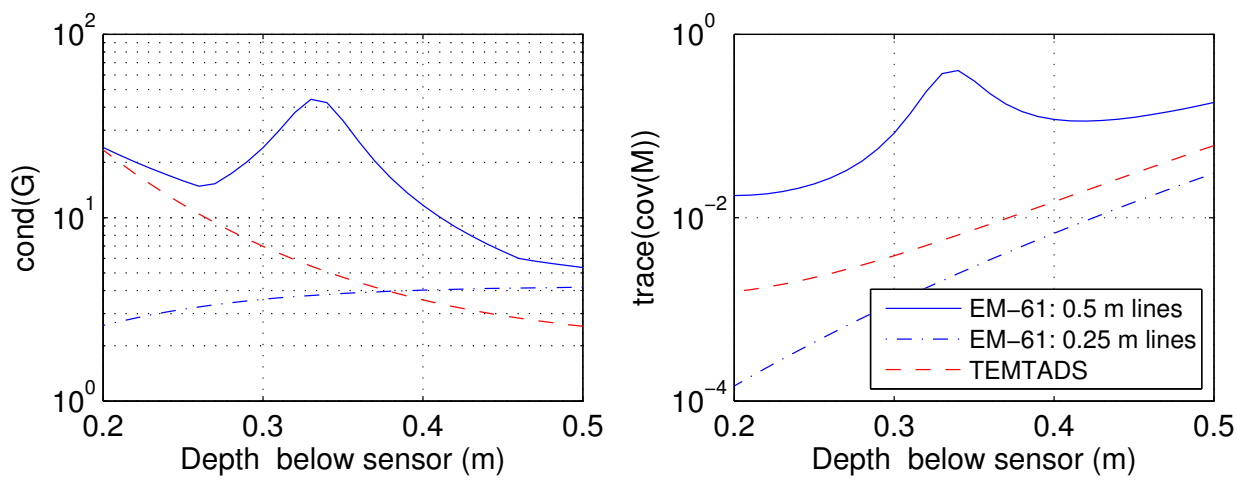


FIGURE 19. Left: condition number of the forward modelling matrix $\mathbf{G}$ as a function of target depth for EM-61 and TEMTADS sensors. Right: trace of the covariance of the polarizability tensor elements as a function of target depth.

For the TEMTADS sensor positioned directly over the target, we see in figure 19 that the condition number decreases monotonically with target depth. This is because the effective number of data is decreased as a shallow point dipole response is restricted to central transmitters and receivers. This can be addressed by eliminating low amplitude data from outer coils, by truncating small singular values in the singular value decomposition of $\mathbf{G}$, or, as will be explored in this paper, by explicitly regularizing the inverse problem. Also shown in figure 19 is the trace of the estimated model covariance for the same scenarios. The model covariance is (Menke, 1989)

$$\mathrm{cov}(\hat{\mathbf{m}}) = (\mathbf{G}^T\mathbf{G})^{-1}. \tag{20}$$

The TEMTADS and cued (0.25 m) EM-61 surveys decrease the total uncertainty relative to the detection mode (0.5 m) EM-61 survey. This covariance, however, does not account for errors on the data and assumes that target depth is perfectly recovered. Data acquired with a moving sensor are susceptible to correlated errors in position, and this can translate to models at local minima corresponding to erroneous depths. In practice, cued data acquired with a stationary multi-static sensor provide better estimates of target depth and so greatly reduce uncertainty in polarizabilities (Billings et al., 2010).

However, small or deep TOI can still be problematic in multi-static data processing. The decreased signal from deep targets counteracts improved conditioning of the inverse operator with depth and can make polarizability estimates ill-posed. This is exacerbated for low SNR channels at late times, as illustrated in figure 20 for inversion of TEMTADS data acquired over an (axisymmetric) 4.2" mortar. The unregularized inversion produces transverse polarizabilities which are approximately equal over most of the channels, but at late times it is evident that low SNR precludes accurate estimation of these parameters, even for the large TOI considered here.

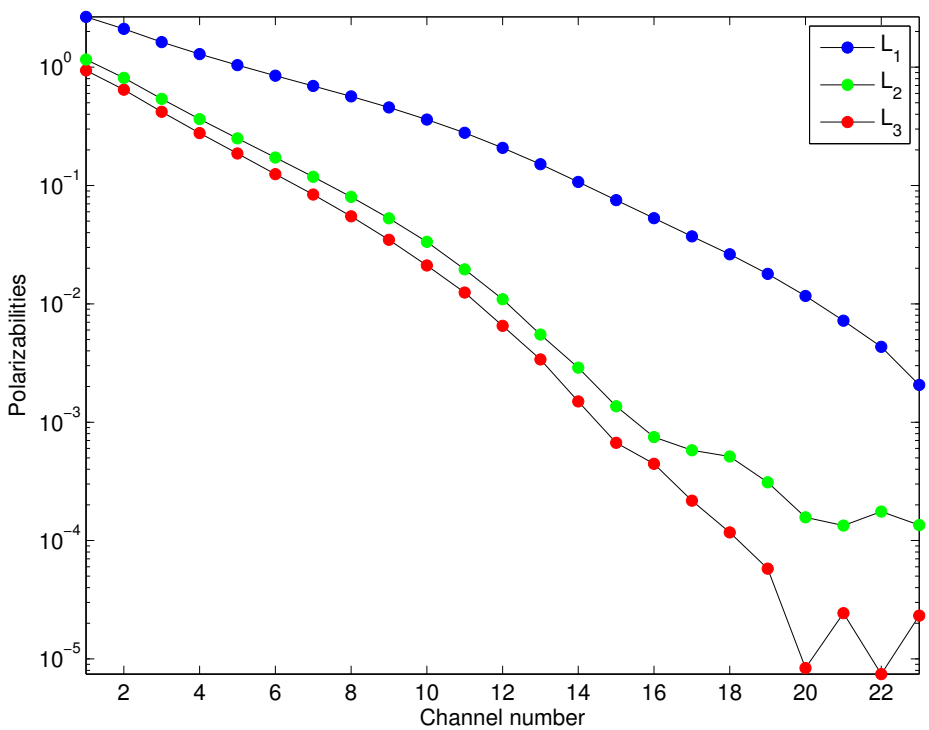


FIGURE 20. Unregularized TEMTADS inversion result: 4.2" mortar. We average TEMTADS channels (see figure 1) in windows of length 5, producing 23 channels spanning the range 0.05-22 ms. This speeds processing and improves SNR at later channels.

Here we investigate techniques for explicitly constraining transverse polarizability estimates in an inversion. An obvious and viable approach to this problem is to simply reparameterize the dipole model so that secondary and tertiary polarizabilities are equal. This is termed a two-dipole model because the secondary response is a superposition of axial and (equal) transverse polarizations. Practical use of the two-dipole model is motivated by the analytic response of a spheroid and by successful fits to high-fidelity test stand data acquired over axisymmetric targets (Pasion, 2007). Of course, the two-dipole model may not provide a good fit to data acquired over an non-axisymmetric target.

A data processing approach which has been proposed to handle this ambiguity is to fit each target using both two or three (i.e. unequal transverse) dipole parameterizations and then to compare the fits to the observed data. The three-dipole parameterization has more degrees of freedom with which to fit observed data and so generally provides a lower misfit than the two-dipole parameterization. The problem is then to determine what constitutes a significant difference in data misfit for the two competing parameterizations. Model selection criteria can be used to select the most parsimonious model parameterization which can explain the data (Hastie et al., 2001).

In this work we instead develop regularization techniques to constrain the recovered polarizabilities. Constraints on model parameters are typically applied in the form of parameter bounds, here we will impose an additional constraint in the form of a penalty on unequal transverse polarizabilities. This approach provides us with a continuum of possible models between constrained and unconstrained models (or, equivalently, between two and three dipole parameterizations). We do not select a single model but rather input all models into a discrimination algorithm and then classify a target based on the model which most probably corresponds to a target of interest. The method is demonstrated on TEMTADS and MetalMapper data sets acquired at San Luis Obispo (SLO), California, and Camp Butner, North Carolina.

3.2. **Regularized inversion.** When solving parametric inverse problems, it is usually sufficient to minimize a data norm quantifying the misfit between observed and predicted data

$$\phi_d = \|d^{obs} - d^{pred}\|^2 \tag{21}$$

with $d^{pred} = F(m)$ generally a nonlinear functional of the model m. Additional prior information can be incorporated in the inversion via parameter bounds (e.g. positivity) or by constructing a model which has specified properties. In the latter case, the optimization problem becomes

$$\begin{aligned} &\min_m \phi_m \\ &\text{s.t. } \phi_d \leq \tau. \end{aligned} \tag{22}$$

The model norm ϕ_m is a regularizer which ensures that the recovered model has, for example, a minimum deviation from some prior reference model. The parameter τ is a target value of the data misfit which is typically specified as the expected value of ϕ_d under the assumption of Gaussian noise (Oldenburg and Li, 2005).

In the case of the dipole model, a simple regularizer which penalizes differences in secondary and tertiary polarizabilities is

$$\phi_m = (L_2 - L_3)^2. \tag{23}$$

To solve the optimization problem in equation 22, we form the Lagrangian

$$\mathcal{L} = \phi_m + \gamma(\phi_d - \tau) \tag{24}$$

subject to the Lagrange multiplier satisfying $\gamma \geq 0$. Minimizing the Lagrangian yields the unconstrained minimization problem (Tikhonov regularization)

$$\min \phi = \phi_d + \beta\phi_m. \tag{25}$$

For a nonlinear forward modelling with fixed β, the solution is obtained by linearizing ϕ and iteratively computing the least squares solution to an overdetermined system of equations for the model perturbation δm

$$\begin{pmatrix} J \\ \sqrt{\beta\phi''_m} \end{pmatrix} \delta m \approx \begin{pmatrix} \delta d \\ 0 \end{pmatrix} \tag{26}$$

We solve this system using the `lsqnonlin` routine in Matlab with sensitivities J computed numerically. The regularization parameter β is inversely proportional to the Lagrange multiplier γ. In practice, the

inverse problem is solved by minimizing equation 25 over a range of βs values until a model satisfying $\phi_d \approx \tau$ is identified. Criteria such as generalized cross-validation or L-curve heuristics can also be used to select a model which balances the trade-off between model and data norms (Oldenburg and Li, 2005). In this work we will employ a different approach to model selection: we allow a discrimination algorithm to select a model which best corresponds to a target of interest. Before turning to the discrimination stage, however, we describe in more detail our inversion algorithm.

Following on work in Song et al. (2011), we use a sequential inversion approach as follows:

1: Solve an inverse problem for target location $\mathbf{r}$. From equation 18, the model $\mathbf{m_r} = \mathbf{r}$ is related to the data via the nonlinear functional

$$\mathbf{d}^{pred} = F[\mathbf{r}] = \mathbf{G}(\mathbf{r})\mathbf{G}^{\dagger}(\mathbf{r})\mathbf{d}^{obs}. \tag{27}$$

We estimate $\mathbf{r}$ by minimization of equation 21 using an iterative Gauss-Newton algorithm in Matlab.

2: Solve the linear inverse problem for the unique elements of the polarizability tensor $\mathbf{M}(t)$ (equation 18) at all time channels using the location obtained at the previous step.

3: Compute the eigenvalue decomposition of the polarizability tensor at each time channel. The eigenvalues are an initial estimate of principal polarizabilities, and eigenvectors correspond to the columns of the Euler rotation matrix $\mathbf{A}$ in equation 14. To estimate the orientation angles we then minimize the least squares difference (Frobenius norm) between the eigenvector matrix and Euler rotation matrix parameterized by orientation angles (ϕ, θ, ψ).

4: Finally, at each time channel solve a nonlinear regularized inverse problem for orientation angles and principal polarizabilities. The model at each time channel is

$$\mathbf{m}_L(t_j) = [\phi,\ \theta,\ \psi,\ L_1(t_j),\ L_2(t_j),\ L_3(t_j)]^T \tag{28}$$

At each time channel we obtain a set of models corresponding to the solution of equation 25 over a range of βs.

We use a "β cooling" technique to generate the models in the final step. In this approach we iteratively minimize 25 starting from an initial, large value of β. The regularization parameter is initialized so that the term $\beta\phi_m \gg \phi_d$ at the initial model obtained in step 3. On convergence at a given β, we lower the regularization parameter by a factor κ (e.g. $\kappa = 0.5$) and initialize the next inversion in step 4 with the final model from the previous step. This procedure is repeated until the relative change in the model parameters achieves some tolerance ϵ. Note that for this problem we do not terminate the inversion on the basis of achieving some target data misfit. This is because in a parametric inversion the model has limited degrees of freedom with which to fit the data and so the β cooling procedure will stall at a model corresponding to the unconstrained ($\beta = 0$) solution. This is in contrast to the overdetermined inverse problem, where decreasing the regularization parameter below the target misfit introduces spurious model structure (Oldenburg and Li, 2005).

We remark that the sequential inversion approach developed by Song is very fast relative to an "all-at once" algorithm which tries to estimate polarizability parameters at all time channels simultaneously. Because the time channels are inverted separately in the sequential method, re-weighting of time channels via estimated errors is unnecessary and parallel processing can greatly reduce computation time.

Figure 21 shows the regularized inversion result for channel 1 for the same TEMTADS data acquired over a 4.2" mortar as in figure 20. As the regularization parameter is decreased with β the objective function ϕ in equation 25 is dominated by the data misfit term, so that ϕ_d decreases monotonically. Conversely, the model norm increases monotonically with decreasing β. The Tikhonov (or Pareto) curve shows ϕ_d versus ϕ_m; typically this curve is used to select a model which balances the trade-off between norms. Finally we see in figure 21 that the initial, regularized model estimate has equal transverse polarizabilities. As the regularization parameter is decreased the secondary and tertiary polarizabilities diverge to their unregularized values, with the primary polarizability unaffected by the regularization.

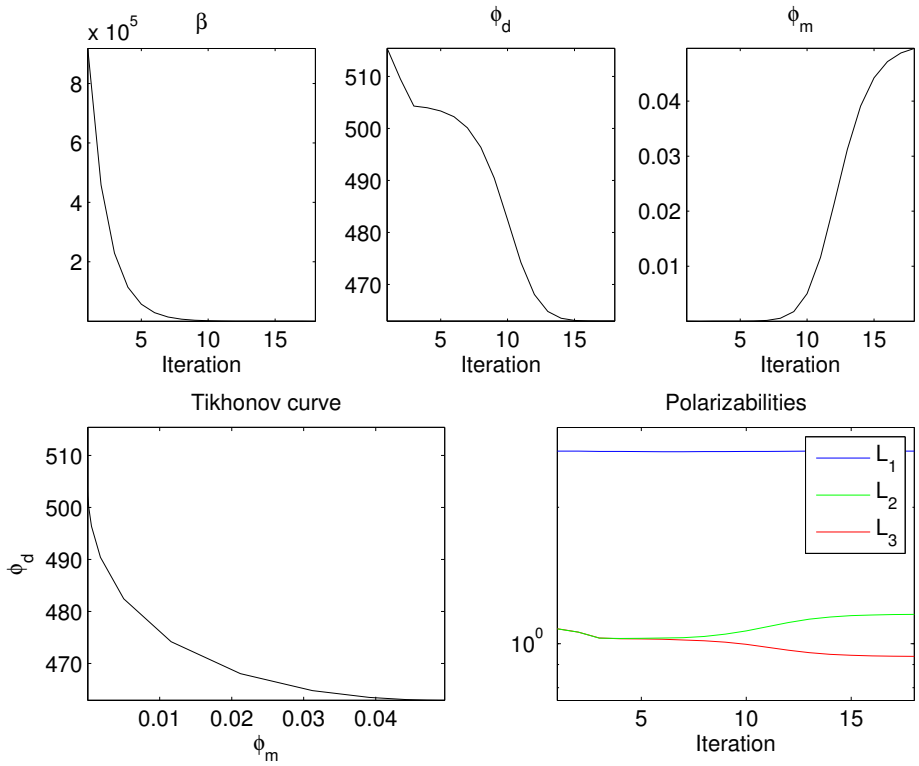


FIGURE 21. Regularized inversion result for channel one of target in figure 20

Regularized inversion results at all channels are shown in figure 22. We show the relative quality of the fit to the observed data by computing a likelihood

$$p(\mathbf{m}_i^j) = \exp\left(-\frac{\phi_d(\mathbf{m}_i^j) - \min(\phi_d^j)}{\min(\phi_d^j)}\right) \tag{29}$$

for the ith model at the jth channel ($\mathbf{m}_i^j$), with $\min(\phi_d^j)$ the minimum misfit obtained over all models at channel j. This form of model likelihood is somewhat arbitrary; rigorous calculation of model likelihoods (or posterior probabilities) requires a full uncertainty appraisal. However, we will show in the following examples that equation 29 is a reasonable metric for weighting competing models. We also remark that the model likelihood is approximately

$$\begin{aligned} p(\mathbf{m}_i^j) &\approx 1 - \frac{\phi_d(\mathbf{m}_i^j) - \min(\phi_d^j)}{\min(\phi_d^j)} \\ &= 1 - \text{relative misfit} \end{aligned} \tag{30}$$

with relative misfit used to compare models in a misfit vs. depth curve (Lhomme et al., 2008). In figure 22 we show the minimum misfit (unregularized) model (with $p(\mathbf{m}) = 1$) and the model from regularized inversion with minimal model norm. At early, high SNR channels, the unregularized model is a slightly better fit to the data than the regularized model, and so the latter has a somewhat reduced likelihood. However, at late time channels (i.e channels 15 and beyond in figure 22) both models have approximately equal likelihoods, indicating that the regularized model with $L_2 \approx L_3$ is as good a fit to the data as the unregularized model.

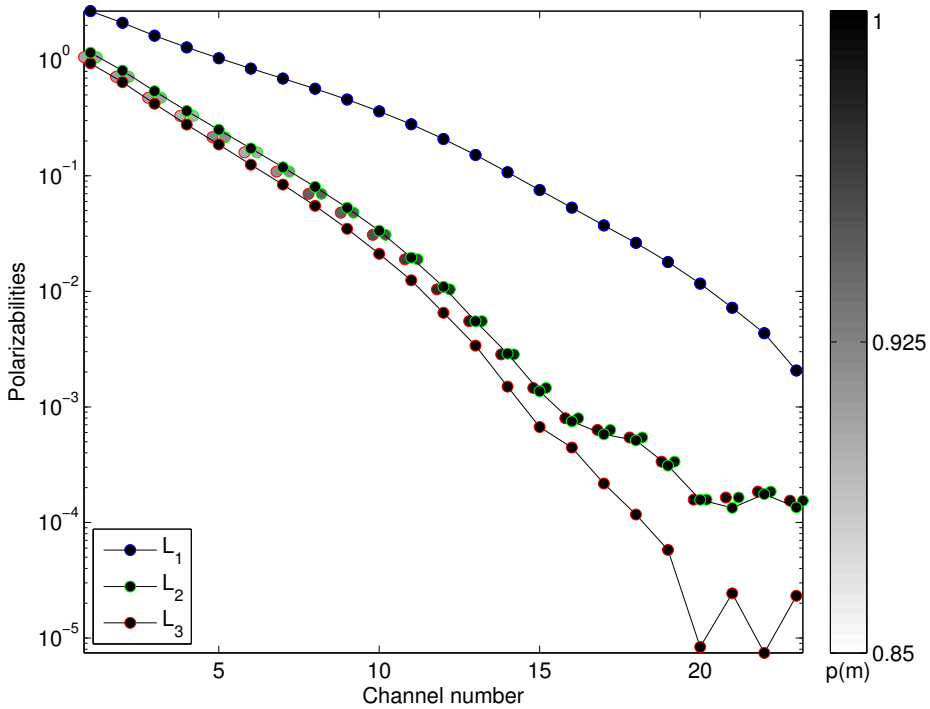


FIGURE 22. Models from regularized TEMTADS inversion: 4.2" mortar. Solid lines indicate unregularized result, jittered L_2 and L_3 correspond to models with the minimum difference in transverse polarizabilities obtained with regularized inversions.

Figure 23 shows a regularized inversion result for a 60 mm mortar at 15 cm depth in the TEMTADS training data at San Luis Obispo. There is a much greater deviation between transverse polarizabilities for the unregularized inversions, particularly at the first time channel. This suggests that discrimination using target shape can still be problematic for small targets at even modest depths. Again, axisymmetric models from the regularized inversions are nearly as likely as the best-fitting (non-axisymmetric) model, as might be expected for this TOI.

If the target is in fact non-axisymmetric, then can we definitively rule it out as a target of interest by comparing the relative likelihoods of models from regularized inversions? It depends. Figure 24 shows a regularized inversion result for a clutter item at the surface. While there is not a marked difference between unregularized transverse polarizabilities in figure 24, at early times a model with equal (or nearly equal) transverse polarizabilities is less likely (i.e. is a significantly poorer fit to the data). However, at later times with reduced SNR, axisymmetric models become almost as probable as non-axisymmetric models. Hence if the SNR is sufficiently high, then we may be able to eliminate a target as a TOI on the basis of shape information. If the target is small or deep (as in figure 23), there may not be enough information in the data to confidently rule out a TOI using target shape alone.

3.3. **Multi-object regularized inversion.** Any practical inversion algorithm for UXO discrimination must consider overlapping responses from multiple targets. The approach in Song et al. (2011) augments the model vector with multiple dipole sources and simultaneously estimates locations for all sources, followed by estimation of all polarizabilities. Applying regularized inversion at this second step would require separate model norm terms for each object, and careful balancing of these terms via separate regularization parameters. A more straightforward route is to decouple the regularized inversions into a

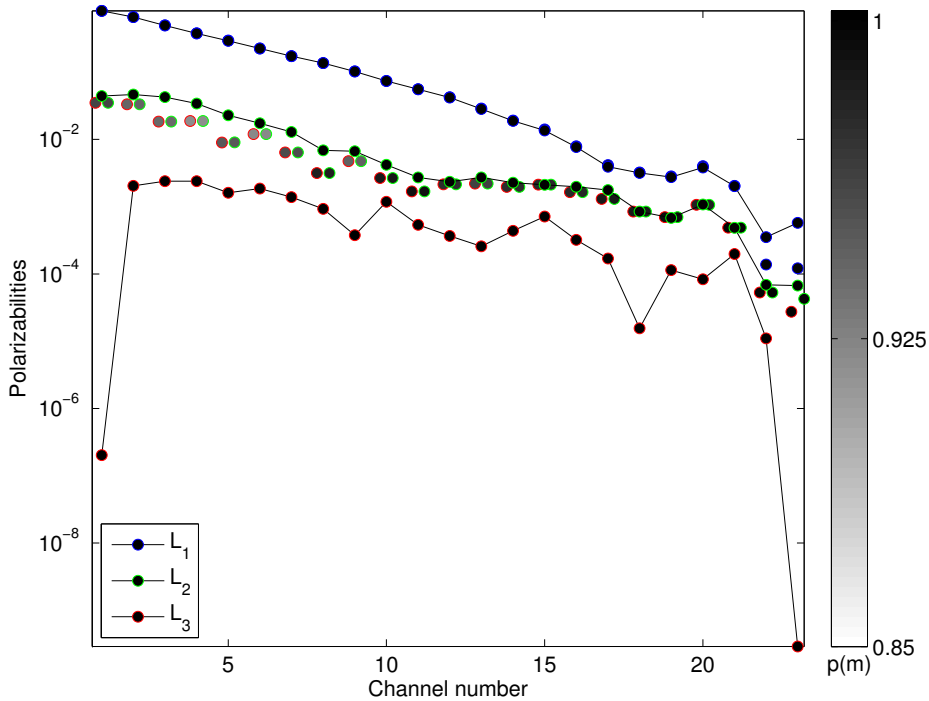


FIGURE 23. Models from regularized TEMTADS inversion: 60 mm mortar. Solid lines indicate unregularized result, jittered L_2 and L_3 correspond to models with the minimum difference in transverse polarizabilities obtained with regularized inversions.

series of single object inversions. In this procedure, termed an "iterative residual fit" (IRF) by Bell (2006), each single object inversion fits the residual data that cannot be predicted by other objects (figure 25). We apply this approach to our initial location estimation problem, iterating between single object inversions until each converges to a final location. We then carry out separate regularized inversions at these locations, with each regularized inversion again fitting the residual data that cannot be predicted by all other dipole sources. Initial testing of the iterative residual fit procedure by Bell (2006) on magnetic and mono-static electromagnetic data proved unsuccessful: the method was consistently outperformed by simultaneous estimation of all sources (the "double happiness" approach). However, the diversity of transmitter excitations and receiver components in multi-static data may better support the iterative approach.

Figure 26 compares single and two-object inversion results for TEMTADS data acquired over a 37mm target at Camp Butner. In this example the two-object (IRF) inversion recovers a set of polarizabilities that are in much better agreement with the training data than the single object result. Despite this improvement, we note that the estimated secondary polarizabilities for this target are poorly constrained at late times; subsequent regularized inversion at the corresponding source location addresses this ill-conditioning.

3.4. **Discrimination using transverse polarizabilities.** To demonstrate the utility of including regularized transverse polarizabilities in discrimination, we consider discrimination with a nonlinear support vector machine (SVM) (see Hastie et al. (2001) or Burges (1998) for a full description) The decision function for the SVM is

$$\mathrm{y}^{test} = \mathrm{w}^T \mathrm{K}(\mathrm{x}^{train}, \mathrm{x}^{test}) + b_0 \tag{31}$$

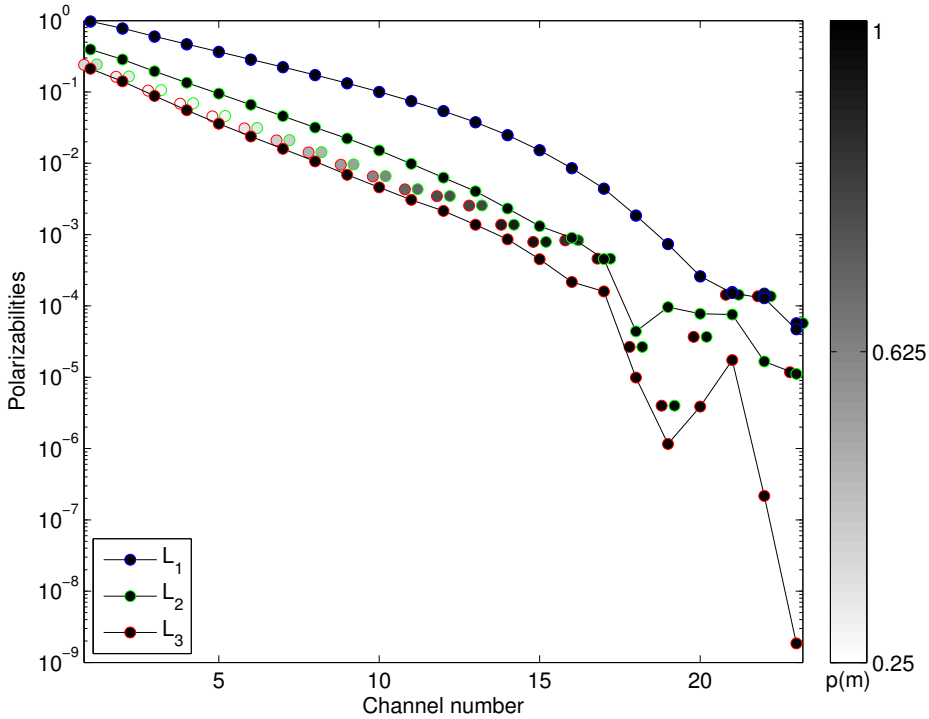


FIGURE 24. Models from regularized inversion: clutter. Solid lines indicate unregularized result, jittered L_2 and L_3 correspond to models with the minimum difference in tranverse polarizabilities obtained with regularized inversions.

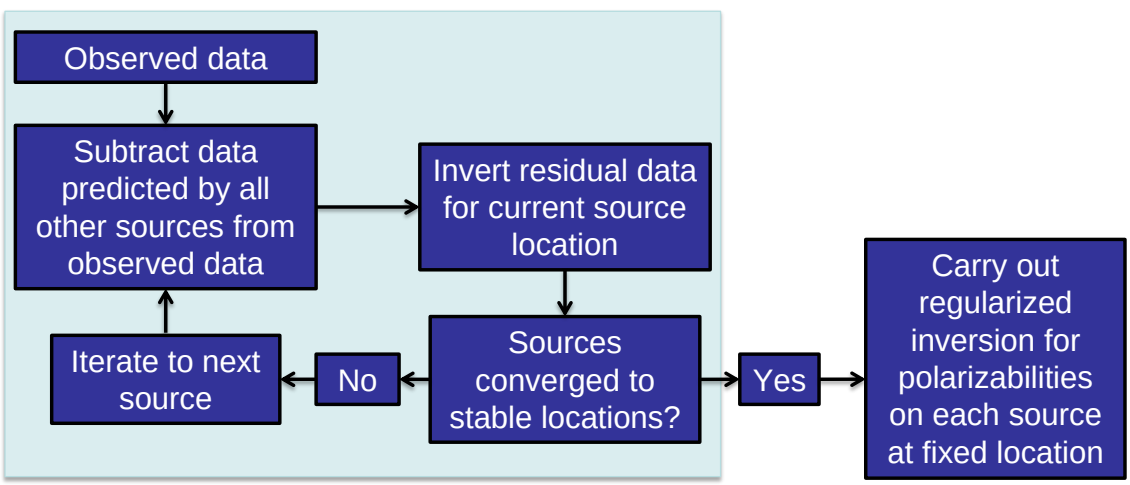


FIGURE 25. Processing flow for iterative residual fit procedure.

with K a kernel matrix, w a (sparse) weight vector, and b_0 a constant bias term. For a radial basis function

$$K_{ij} = \exp(-\nu \|\mathbf{x}_i^{train} - \mathbf{x}_j^{test}\|^2) \tag{32}$$

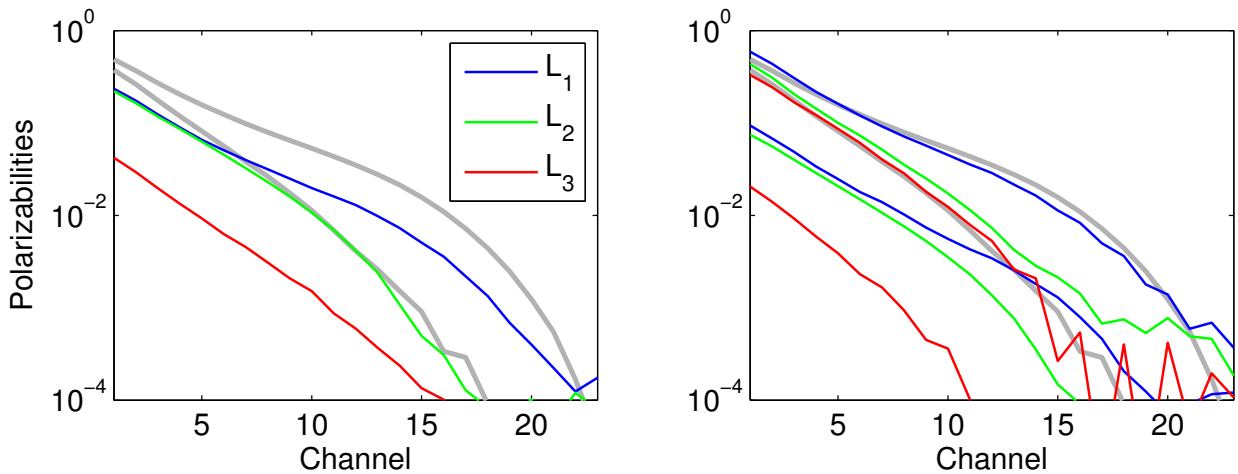


FIGURE 26. Estimated polarizabilities extracted from TEMTADS data for a 37 mm target at Camp Butner. Grey lines are training data polarizabilities for this target class. Left: single object inversion, right: two object inversion with iterative residual fit method

with $\mathbf{x}_i^{train}$ and $\mathbf{x}_j^{test}$ the ith training and j^{th} test vectors, respectively. For discrimination with polarizabilities, the elements of these feature vectors are log-transformed primary, secondary, and tertiary polarizabilities at a subset of the available time channels. Expanding the norm in terms of the elements of training and test vectors, the kernel matrix can be expressed as

$$K_{ij} = \prod_{k=1}^{N} \exp(-\nu(x_{ik}^{train} - x_{jk}^{test})^2). \tag{33}$$

When considering multiple models from a regularized inversion, the kth element of the jth test vector has multiple values, denoted x_{jkl}^{test}, with corresponding likelihoods $p(x_{jkl}^{test})$ from equation 29. For discrimination, we select the element for which the term

$$\kappa_{ijkl} = \left[\exp(-\nu(x_{ik}^{train} - x_{jkl}^{test})^2)p(x_{jkl}^{test})\right] \tag{34}$$

is maximized. The elements of the kernel matrix are then

$$K_{ij} = \prod_{k=1}^{N} \max_l \kappa_{ijkl} = \prod_{k=1}^{N} \max_l \left[\exp(-\nu(x_{ik}^{train} - x_{jkl}^{test})^2)p(x_{jkl}^{test})\right]. \tag{35}$$

The preceding computations compare all possible values of a test vector with a given training vector and retain the test vector elements which are "closest" (in the sense of the radial basis function) to that training vector. The weighting by model likelihoods acts to penalize models which may agree with a training vector but which do not fit the data.

We have applied the above processing to cued MetalMapper and TEMTADS data sets acquired for ESTCP demonstrations at San Luis Obispo (SLO) and Camp Butner. At SLO, the discrimination task was to identify seeded targets ranging in size from 4.2" mortars down to 60 mm mortars (figure 27a). For this site training data were provided by ESTCP and comprised a random sample of 174 detected targets. Targets of interest at Camp Butner ranged from large 105 mm projectiles down to 37 mm projectiles (figure 27b). Beyond test pit measurements of individual items from each class of TOI, no training data were initially available for classifier training. For each data set at Camp Butner, we requested groundtruth

for a small number (< 50) of targets in order to characterize the distributions of TOI polarizabilities. More description of this procedure is given in Pasion et al. (2011).

Figure 28 compares the performance of nonlinear support vector machines using regularized and unregularized test polarizabilities. In all cases, regularization decreases the false alarm rate relative to using all (primary, secondary and tertiary) unregularized polarizabilities. For SLO and Camp Butner MetalMapper data sets the regularized method achieves the maximal area under the ROC (AUC) and a false alarm rate (FAR) comparable to discrimination using only primary polarizabilities. For the TEMTADS data sets, discrimination with regularized inversion has similar AUC and FAR to discrimination with primary polarizabilities.

3.5. **Conclusions.** We have developed a regularized inversion algorithm that penalizes the deviation between transverse polarizabilities. Rather than selecting a single model from this inversion process, we input all models into a support vector machine classifier. This corresponds to test features vectors with multiple values for each element. We compare the elements of each test vector with the training data and retain the model value which best corresponds to a given training vector. We also penalize the match of test and training vectors by the likelihood that the model fits the observed data.

We find that the regularized method can improve initial performance on high SNR targets with well-constrained tranverse polarizabilities while preventing the occurrence of outlying TOI that arise when we rely on unregularized parameters throughout the diglist. The greatest benefit in discrimination performance is obtained with sensor data which interrogates all polarizabilities with orthogonal (horizontal and vertical) primary fields (i.e. MetalMapper).

(a) San Luis Obispo. Clockwise from top left: 60 mm mortar, 81 mm mortar, 4.2" mortar, 2.36" rocket.

(b) Camp Butner. Clockwise from top left: 37 mm projectile, M48 fuze, 105 mm high explosive (HE) projectile, 105 mm high explosive anti-tank (HEAT) projectile.

FIGURE 27. Targets of interest for ESTCP demonstrations at SLO and Camp Butner.

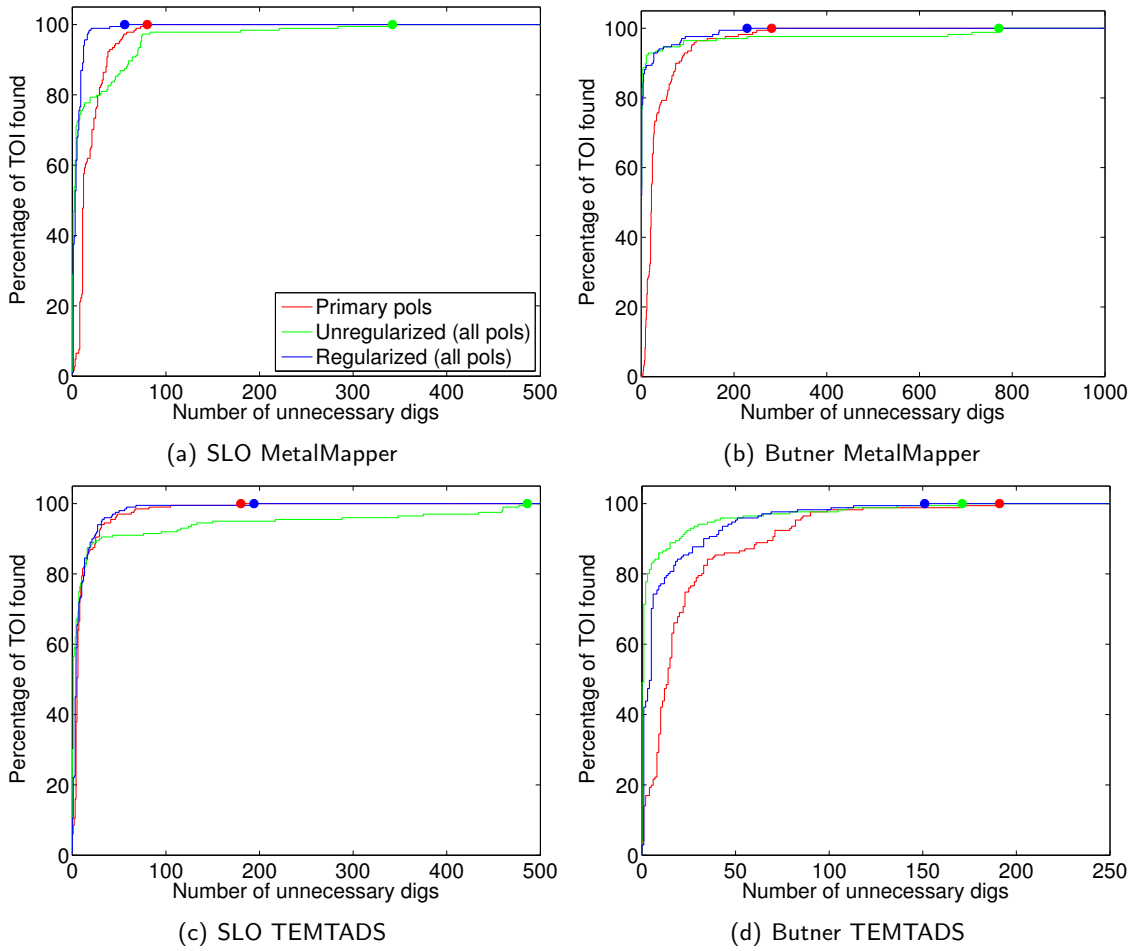


FIGURE 28. Receiver operating characteristics for support vector machines applied to cued data sets at San Luis Obispo (SLO) and Camp Butner.

4. Including uncertainty in parameter estimates during UXO classification

We have investigated how model parameter uncertainty can be incorporated in the discrimination process. A full description of methods and results can be found in Beran et al. (2011b), here we provide a brief summary of the work.

Figure 29 shows a motivating example for the problem of discrimination with features extracted from time-domain electromagnetic (TEM) data. The best-fitting model for one target (indicated by an arrow in figure 29) is far removed from the typical feature vectors we obtain for this type of ordnance. This outlier highlights that the TEM parameter estimation problem is fundamentally ill-posed: a small perturbation in the data caused by noise can produce a large change in the estimated model (ill-conditioning), and multiple models can fit the data equally well (non-uniqueness).

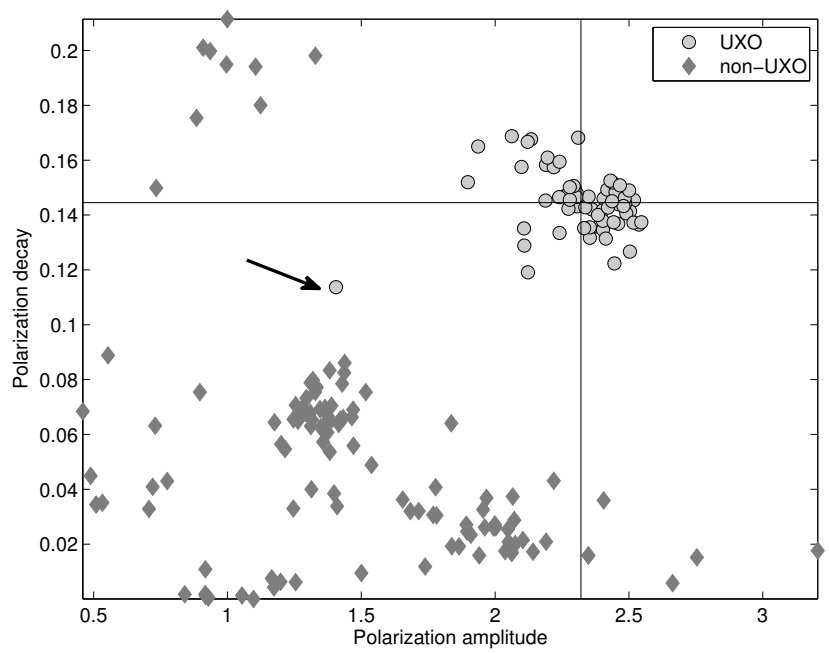


Figure 29. Estimated model parameters for ordnance and non-ordnance targets for Camp Sibert EM63 data. Crosshairs show the location of the mean of the UXO class.

Model uncertainty characterizes whether a given inverse problem is ill-posed. For a nonlinear forward modelling $\mathbf{d}^{pred} = \mathrm{F}(\mathbf{m})$ the misfit can be minimized iteratively by solving for the model perturbation $\delta\mathbf{m}$

$$\mathbf{H}\delta\mathbf{m} = \mathbf{J}^T\mathbf{W}_d^T\mathbf{W}_d\delta\mathbf{d} \tag{36}$$

with $\mathbf{H} = \nabla^2\phi$ the Hessian of the misfit, $\mathbf{J}$ the Jacobian matrix of sensitivities, and $\delta\mathbf{d} = (\mathbf{d}^{obs} - \mathrm{F}(\mathbf{m}))$. At the minimizer $\hat{\mathbf{m}}$ we can approximate the model covariance as (Menke, 1989)

$$\mathrm{cov}(\hat{\mathbf{m}}) \approx \mathbf{H}^{-1}. \tag{37}$$

From this result we see that if there is a large curvature to the misfit function at the model estimate $\hat{\mathbf{m}}$, then the model is well constrained and the variance of the model parameters is small. The probability distribution of the model parameters may then be approximated as a normal PDF with mean $\hat{\mathbf{m}}$ and covariance computed with equation 37 (Menke, 1989).

This linearized uncertainty analysis may not be valid if the objective function is nonconvex. In this case the local quadratic approximation provides a poor approximation to the actual objective function and uncertainties estimated with equation 37 may not be reflective of the actual uncertainties in the model. An alternative approach to estimating uncertainties is to use a Bayesian framework (Tarantola (2005), Sen and Stoffa (1995)) to estimate the model posterior probability distribution (PPD)

$$p(\mathbf{m}|\mathbf{d}^{obs}) \propto p(\mathbf{d}^{obs}|\mathbf{m})p(\mathbf{m}). \tag{38}$$

The PPD for a nonlinear problem can be estimated numerically using the Metropolis-Hastings algorithm. This algorithm works by randomly perturbing the current model $\mathbf{m}^{current}$ to a proposed model $\mathbf{m}^{proposed}$ and accepting the proposed model according to the Metropolis criterion (Metropolis et al., 1953). This scheme is a Markov chain; acceptance of the proposed model depends only on the current model. After a sufficient number of samples the chain of accepted models will converge to a stationary distribution which is the posterior distribution.

From linearized and nonlinear uncertainty appraisals of TEM dipole parameters we concluded that:

(1) the minimum misfit model is not always the best model to use for discrimination,
(2) linearized uncertainty appraisal cannot always account for the deviation of an estimated feature vector from the expected value of that target class.

We applied Bayes rule to propagate model uncertainty through to the discrimination stage. This yielded the posterior probabilities of class membership (in classes T (UXO) and F (clutter)) given the observed data $\mathbf{d}^{obs}$

$$\begin{aligned} p(T|\mathbf{d}^{obs}) &= \frac{\int p(\mathbf{x}|\mathbf{d}^{obs})p(\mathbf{x}|T)p(T)d\mathbf{x}}{\int p(\mathbf{x}|\mathbf{d}^{obs})p(\mathbf{x}|T)p(T)d\mathbf{x} + \int p(\mathbf{x}|\mathbf{d}^{obs})p(\mathbf{x}|F)p(F)d\mathbf{x}} \\ p(F|\mathbf{d}^{obs}) &= 1 - p(T|\mathbf{d}^{obs}). \end{aligned} \tag{39}$$

Intuitively, the above expression evaluates Bayes rule over all possible values of the estimated feature vector $\mathbf{x}$, weighted by the probability of each respective value. For multivariate normal distributions the integral can be solved analytically, with the result that the integral of two Gaussians is itself a Gaussian distribution (Brookes, 2005). We therefore term a classifier employing equation 39 with Gaussian distributions for both classes and feature vectors a "Gaussian product" (GP) classifier.

Application of the GP classifier requires estimation of model uncertainty for all inverted targets. While nonlinear appraisal can be carried out for all targets, here we find that an efficient solution is to approximate the multi-modal parameter distributions with an ensemble of models obtained by repeatedly minimizing the misfit with an iterative algorithm. Each model corresponds to a mode of the posterior probability distribution. Figure 30 shows EM63 test data from Camp Sibert, AL generated with this approach. We then evaluate the Gaussian product (equation 39) using the linearized uncertainty about each model in our ensemble (i.e. each kernel in the mixture model). The kernel with maximal probability of membership in the UXO class is then used to classify the respective target.

Figure 31 compares receiver operating characteristics (ROCs) obtained with this approach with conventional quadratic discriminant analysis (QDA) using only the minimum misfit feature vector to classify each target. The method is applied to EM61 and EM63 data sets acquired at Camp Sibert. For both EM61 and EM63 data sets the area under the ROC and the false alarm rate at Pd=1 (i.e. the proportion of false positives required to identify all true positives) are improved by the GP classifier. While the improvement for the EM63 data appears negligible, the identification of one outlying UXO (4.2" mortar) is a significant result from the perspective of a regulator charged with site remediation. Similarly, the significant reduction in false alarm rate at Pd=1 for the EM61 data improves the likelihood that all ordnance will be identified with this sensor.

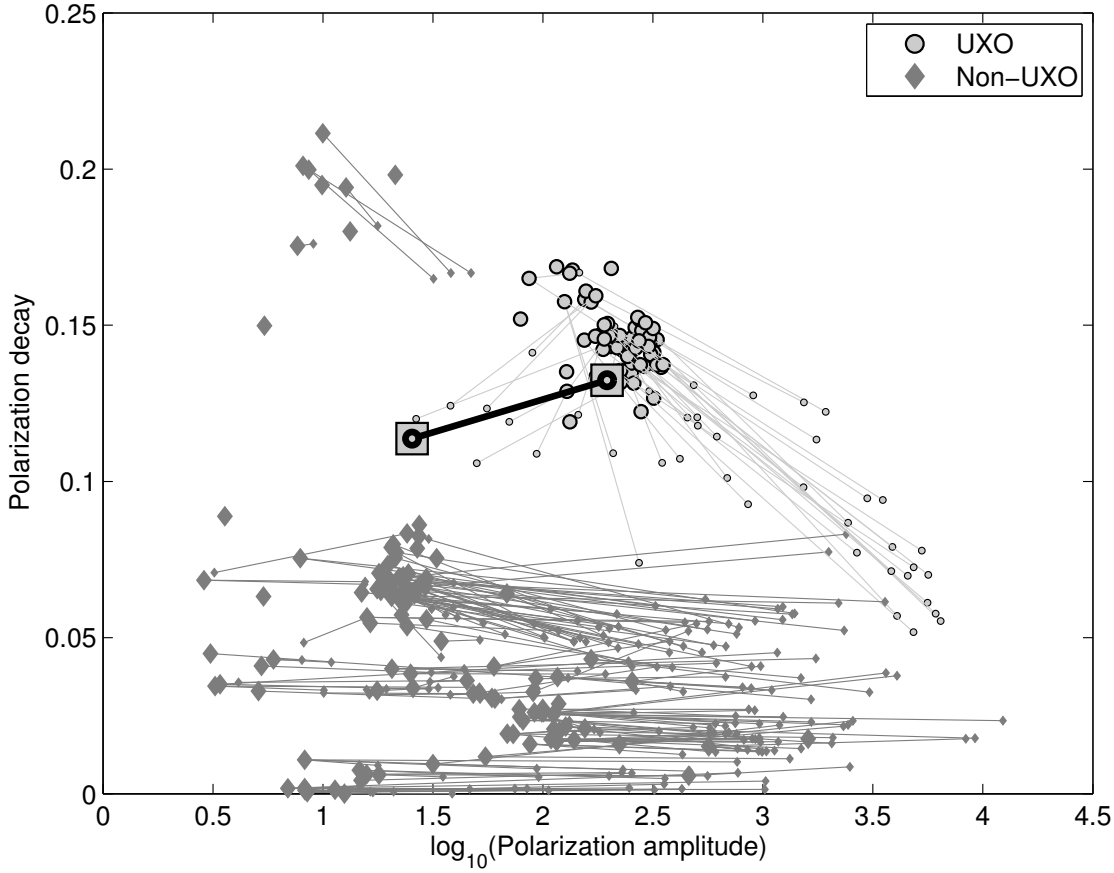


FIGURE 30. Feature vectors estimated from Camp Sibert EM63 data. Feature vectors for a given target are connected by a line, with the largest marker indicating the minimum misfit model. Highlighted feature vector is an outlier to the ordnance class.

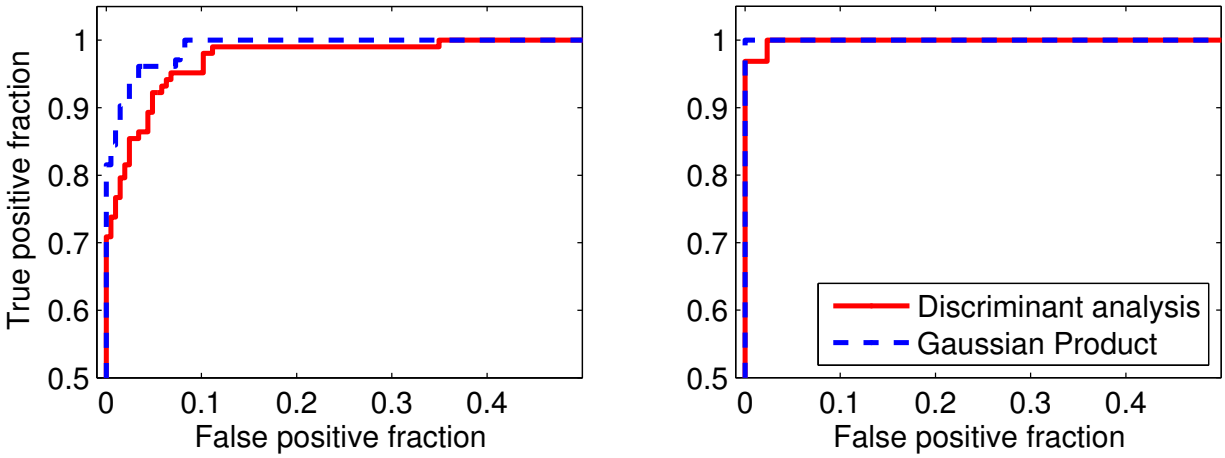


FIGURE 31. Receiver operating characteristics for classifiers applied to EM61 (left) and EM63 (right) test data.

5. Selecting a receiver operating point

The receiver operating characteristic (ROC) curve is a tool for displaying the performance of a binary decision process and is used in applications such as medical diagnosis, economics, and machine learning. The curve is generated by varying the threshold of the decision algorithm and plotting the resulting proportion of true positives as a function of the proportion of false positives. For medical studies a true positive often denotes correctly predicting that a condition is present, while a false positive denotes incorrectly predicting the presence of the condition. The decision statistic, or score, is a single continuous or discrete value which corresponds to the output of a diagnostic test. A second application, which is our focus in this article, is unexploded ordnance discrimination. Here a true positive indicates that we have successfully predicted that an ordnance belongs to the class of UXO targets (T) which must be dug, while a false positive is a non-ordnance item belonging to the class F which we have predicted belongs to T. The decision statistic is the output of any algorithm which can discriminate between T and F classes. For example, the decision statistic might be the probability of membership in the UXO class predicted with a discriminant analysis classifier.

As the decision statistic is varied monotonically for a test data set, we generate a sequence of true and false positives which cumulatively generate points on the ROC. This empirical curve is necessarily an uncertain estimate of the expected performance of a discrimination algorithm, since it is derived from a limited set of observations which is assumed to be representative of the underlying populations of T and F classes. Accordingly, methods for computing ROC confidence intervals are increasingly applied in machine learning contexts (Macskassy et al., 2005).

In this work we address selection of the operating point on the ROC curve. The operating point is a cut-off value of the decision statistic used to generate the ROC. For example, any diagnostic test with a value less than a selected operating point will be classified as "disease present." In the UXO problem the operating point corresponds to the "stop-dig" point, any targets with a decision statistic less than this threshold will be left in the ground. When full clearance of all detected targets is required by a regulator, the operating point might instead represent the division between digging targets with explosive ordnance disposal (EOD) technicians and digging with slightly fewer precautions (e.g. using a backhoe to dig targets which are likely clutter).

In section 5.1 we set up the problem and discuss existing methods for selecting the operating point. These methods use slightly different notions of optimality to balance the trade-off between true and false positive fractions. In unexploded ordnance discrimination we often know a priori how many unlabelled targets must be classified. We assume that this set of unlabelled targets is a sample from hypothetical generating distributions of T and F classes. In section 5.2, we show that the empirical ROC for a sample depends upon the generating and prior distributions of T and F classes, as well as the size of the sample. In particular, as the sample size grows the expected proportion of targets which must be labelled in order to find all T instances in the sample also grows. We then derive the approximate sampling distribution of the point on the ROC at which all T instances are found. In theory, this distribution can be used to select an operating point which attains a specified probability of identifying all T instances in a sample. In practice, however, this depends upon accurate estimation and extrapolation of the tails of the generating distributions from a limited training data set. We find that this approach has limited success when applied to real data sets.

We therefore propose a simple method that does not depend upon estimating distributions, but rather continues labelling instances until a predefined number N_L of false instances occur in sequence. In applications to synthetic and real data this technique has a higher probability of identifying all T instances than other methods, though with a commensurate increase in the proportion of labelled F instances.

5.1. **Optimal operating points.** In this section we review existing theory for choosing the operating point. Following work in Kanungo and Haralack (1995) and Fawcett (2004), denote a positive (e.g. target is ordnance) by T and a negative (e.g. target is not ordnance) as F. At a specified threshold Λ of the decision statistic x, we define the following probabilities

$$P(T|T,\Lambda) = \int_{-\infty}^{\Lambda} p(x|T)dx \qquad \text{(true positive)} \tag{40}$$

$$P(F|T,\Lambda) = \int_{\Lambda}^{\infty} p(x|T)dx = 1 - P(T|T,\Lambda) \qquad \text{(false negative)} \tag{41}$$

$$P(T|F,\Lambda) = \int_{-\infty}^{\Lambda} p(x|F)dx \qquad \text{(false positive)} \tag{42}$$

$$P(F|F,\Lambda) = \int_{\Lambda}^{\infty} p(x|F)dx = 1 - P(T|F,\Lambda) \qquad \text{(true negative)} \tag{43}$$

where $p(x|T)$ and $p(x|F)$ are the distributions of T and F, as shown in figure 32(a). The decision statistic x is a number representing the output of a discrimination algorithm, and $p(x|T)$ and $p(x|F)$ are univariate distributions, sometimes referred to as score distributions, which are specific to that discrimination algorithm. Statistical classification algorithms (e.g. discriminant analysis, support vector machines) may operate in a high-dimensional feature space to generate the decision statistic; we emphasize that the score distributions are a univariate projection of the distributions of T and F in the feature space. In figure 32(a) we have shown the score distributions as normal probability distributions. However, in practice score distributions are rarely normal (Macskassy et al., 2005) and the development which follows is therefore general to any form of $p(x|T)$ and $p(x|F)$. The receiver operating curve shows the dependence of $P(T|T,\Lambda)$ (true positives) upon $P(T|F,\Lambda)$ (false positives) as Λ is varied. The probabilities at a given threshold Λ can be displayed as a "confusion matrix" (figure 32(b)). Hereafter we retain the convention that if the decision statistic x for a test item is less than the threshold Λ then that item classified as T by our discrimination algorithm, as shown in figure 32(a)[2].

In this framework, confusion matrix probabilities alone are not sufficient to select an optimal threshold, we must assign some cost C to each of the decisions we make. We denote the cost of a true positive as $C(T|T)$, and a false positive as $C(T|F)$. The costs of the other elements of the confusion matrix are similarly defined. A notable feature of the ROC is that it is independent of the relative frequencies of the two classes (the prior probabilities $P(T)$ and $P(F)$). However, once we assign a cost to each of the decisions in the confusion matrix, the prior probabilities become crucial because they determine how often we expect to incur the specified costs. Now given our confusion matrix probabilities, specified costs and prior probabilities, the expected cost (also called risk) at Λ will be

$$\begin{aligned} E(C(\Lambda)) &= P(T|T,\Lambda)C(T|T)P(T) + P(F|T,\Lambda)C(F|T)P(T) \\ &+ P(F|F,\Lambda)C(F|F)P(F) + P(T|F,\Lambda)C(T|F)P(F) \\ &= P(T|T,\Lambda)C(T|T)P(T) + (1 - P(T|T,\Lambda))C(F|T)P(T) \\ &+ (1 - P(F|T,\Lambda))C(F|F)P(F) + P(T|F,\Lambda)C(T|F)P(F) \end{aligned} \tag{44}$$

[2]Many statistical classifiers output a probability $P(T|x)$, so that a large decision statistic indicates that an instance is likely to be T. In this case we simply threshold on $-P(T|x)$.

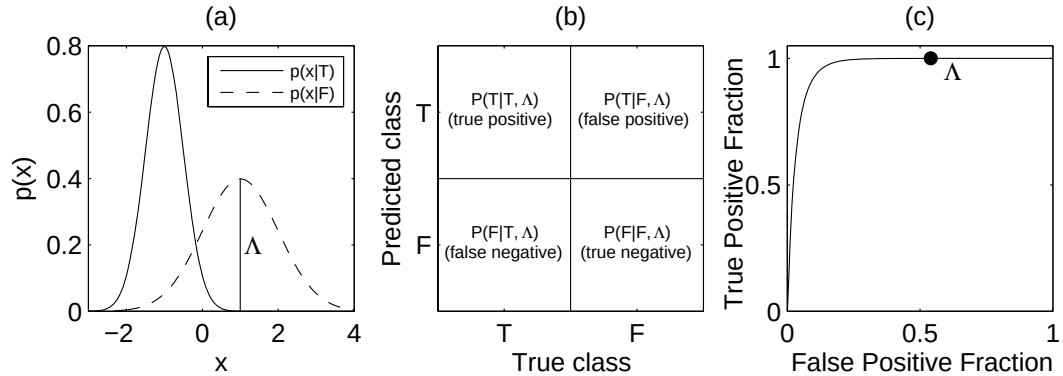


FIGURE 32. (a) The ROC is generated by integrating the distributions of T and F classes. (b) At a given operating point Λ,we have a confusion matrix of probabilities, and the ROC is a plot of the first row of this matrix as a function of Λ. (c) The resulting ROC with operating point Λ.

Minimizing the expected cost with respect to Λ gives

$$\frac{\partial P(T|T,\Lambda)}{\partial P(T|F,\Lambda)} = \frac{(C(F|F) - C(T|F))P(F)}{(C(T|T) - C(F|T))P(T)}. \tag{45}$$

Recalling that the ROC is a plot of $P(T|T,\Lambda)$ versus $P(T|F,\Lambda)$, the optimal operating point is defined by the point where the slope of the curve equals the right hand side of the above expression (Kanungo and Haralack, 1995). The ROC is a non-decreasing function, and a positive slope can be ensured by requiring

$$\begin{aligned} C(T|F) &\geq C(F|F) \\ C(F|T) &\geq C(T|T) \end{aligned} \tag{46}$$

so that misclassification costs more than correct classification. In machine learning it is often assumed that correct classifications do not incur any cost and we must only specify the relative costs of false negatives and false positives. While in some applications (e.g. economics) costs are available, in many cases it is not obvious how to objectively choose these parameters. Of particular concern is the cost of the false negative. What does it cost to leave an ordnance item in the ground? In the short term, not taking action when it is required may save us expenditures on remediation, but in the long term we must account for potential environmental damage, liability, etc. We might encode expert, but ultimately subjective, judgements into costs. Variable costs of remediating targets might be considered: for example deep targets are more time-consuming (and hence more expensive) to excavate than shallow targets. However, deep targets are also likely to pose less risk for accidental detonation. We conclude that there is considerable ambiguity to cost specification and employing expert judgement when specifying costs may not minimize actual risk.

Even if costs can be chosen in a manner that will satisfy regulators, we also require the prior probability for each class to apply equation 45. We are thus faced with the problem of specifying two functions (costs and priors) from limited training data. A further problem is that the empirical ROC generated from observed data is a piecewise constant curve with infinite or zero slope, and so finding an operating

point with slope satisfying equation 45 requires that we fit a smooth function to the empirical ROC. This interpolated function may also achieve the desired slope in more than one location, so that a given specification of costs and priors does not unambiguously define an operating point.

Billings (2004) suggested that an operating point for UXO discrimination be chosen by terminating digging once a predefined number N_L of successive F instances (clutter) are encountered in the ordered list of targets. In this case the sequence of labelled targets will end with one T instance followed by N_L F instances. The slope of the ROC at this operating point is approximately

$$\frac{\partial P(T|T,\Lambda)}{\partial P(T|F,\Lambda)} \approx \frac{1/N_T}{N_L/N_F} = \frac{1}{N_L}\frac{P(F)}{P(T)} \tag{47}$$

Choosing N_L is equivalent to specifying the right hand side of equation 45. Approximating the slope of the ROC in this way circumvents the need to estimate priors and to interpolate the empirical ROC, but still leaves us with the problem of choosing costs via the parameter N_L. In section 5.2 we explore how N_L is influenced by sample size (the total number of instances) and propose objective criteria for specifying this parameter.

Simpler options for selecting an operating point which do not require costs or priors have also been proposed. Two criteria which have been considered are the point on the ROC closest to the point (0,1), and the point on the ROC which is the largest vertical distance from the 45 degree line passing through (0,0) and (1,1) (the Youden index, Perkins and Shisterman (2006)). Figure 33 illustrates these methods. The point closest to (0,1) is also equivalent to choosing the point with the largest perpendicular distance from the 45 degree line. The 45 degree line corresponds to random chance: a discrimination algorithm with this ROC will correctly rank randomly selected items from each class only half of the time (Fawcett, 2004).

Figure 33 compares the location of operating points on a theoretical ROC curve with the three methods discussed thus far. This curve is generated for two univariate Gaussian distributions with standard deviations $\sigma_T = \sigma$ and $\sigma_F = 2\sigma$, and means separated by 4σ, as shown in figure 32(a). To illustrate the effect of unequal costs on the operating point, we also show operating points with hypothetical costs displayed in table 4. The Youden index and cost minimization with equal priors and equal costs to incorrect classification give the same operating point. We can see this by writing the vertical distance d between the ROC and the 45 degree line as

$$d = P(T|T,\Lambda) - P(T|F,\Lambda). \tag{48}$$

Maximizing d with respect to Λ yields equation 45 for equal priors and costs. The operating point for the closest (0,1) method stops slightly before the Youden index, while cost minimization with lower costs for digging clutter has the intuitive result that decreasing the relative costs of digging clutter causes us to dig more clutter.

Method	C(T|T)	C(T|F)	C(F|T)	C(F|F)
C1	0	1	1	0
C2	0	0.1	1	0

TABLE 4. Example costs for risk minimization used in figure 33.

5.2. **Optimal operating points for samples.** The operating points discussed in the previous section are expected to be optimal (according to slightly different criteria) for the underlying population from which our observed data are assumed to be a random sample. However, in practice we are not applying our discrimination algorithm to the inferred population, but to a sample from that population. How do the expected ROC curves for a population and a sample from that population differ? The empirical

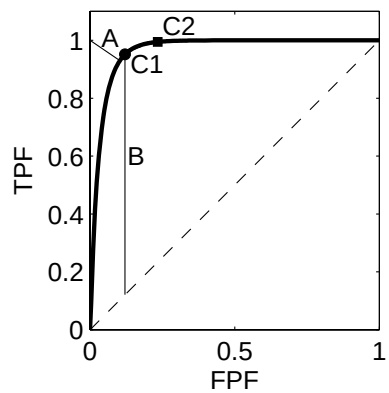


FIGURE 33. Methods for selecting the operating point on an ROC curve. A: closest to point (0,1), B: maximum vertical distance from 45 degree line (dashed). C1: risk minimization with equal priors and equal misclassification costs. C2: risk minimization with equal priors and unequal misclassification costs ($C(F|T) > C(T|F)$) specified in table 4. ROC (heavy line) is generated from two univariate normal distributions with $\mu_T = 2\sigma$, $\mu_F = -2\sigma$, $\sigma_T = \sigma$, $\sigma_F = 2\sigma$, as shown in figure 32(a).

ROC for a sample is generated by ordering the sample from smallest to largest decision statistic, and then incrementing the cumulative counts of T and F instances as we move up the ordered list. The true and false positive fractions are therefore discrete quantities which asymptote to the continuous form (equation 43) as $N \to \infty$.

Consider the example of two univariate Gaussian distributions from the previous section. Figure 34 shows the theoretical ROC curve for the population, as well as ROC curves for sample sizes ranging from $N = 200$ to $N = 800$, with $N/2$ the number of samples from each distribution (i.e. equal priors). As sample size increases, the empirical ROCs are increasingly better approximations to the true ROC. Also shown in figure 34 are the false alarm rates (FAR) for the random samples. Here we define the FAR as the proportion of F items which must be dug in order to find all T items in the sample. The generating distributions require that we dig every F item to guarantee that all T items are found, so that $FAR = 1$ for the population, However, the expected value of the false alarm rate for a given sample size is considerably less than one for finite samples and slowly asymptotes to one as sample size is increased.

Based upon these observations, it seems reasonable to define an optimal operating point λ for a sample of size N as one which finds all T items in the sample. Processing of unexploded ordnance data typically occurs in batches, so that the sample size is known prior to selecting the operating point. If the prior probability of T is known (or is estimated from a training data set) and the total number of instances in the unlabelled sample (test data) is known, then the expected number of T instances in the test data can be computed. This suggests a straightforward criterion for selecting an operating point: continue labelling until the expected number of T instances has been found. However, if the number of T instances is overestimated by just one, then this approach will lead us to label all instances in a fruitless search for a non-existent occurrence of T. This defeats the purpose of discrimination, and so instead we seek to develop methods which have an improved chance of identifying all T instances than the methods previously discussed.

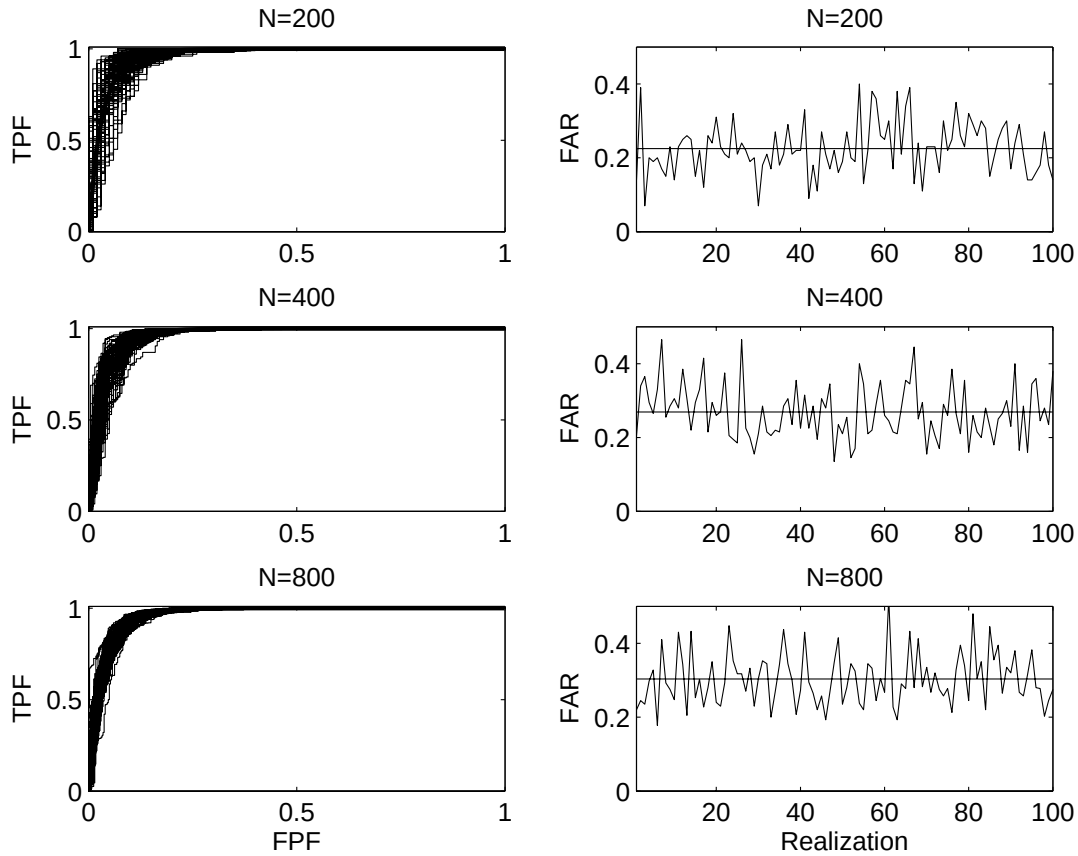


FIGURE 34. Effect of sample size on ROC and false alarm rate. Left column: population ROC (heavy line) and realizations of empirical ROCs for samples of size $N/2$ from each generating distribution. Right column: false alarm rates (FAR) of empirical ROCs in left column. Solid horizontal line is the mean FAR over all realizations.

Here λ is a discrete random variable which corresponds to the index of the last T instance encountered in the ordered sample. A lower bound on the expected value of λ can be derived by finding the critical value x^λ such that

$$1 - P(x^\lambda|T) = \int_{x^\lambda}^{\infty} p(x|T)dx = 1/N_T, \tag{49}$$

with $N_T = NP(T)$ the expected number of T instances in the sample (and similarly, $N_F = NP(F) = N(1 - P(T))$. At the point x^λ, we expect that one T instance will occur somewhere in the interval $[x^\lambda\ \infty]$. This extreme value is the last T instance we will encounter in the sample and therefore defines the desired operating point. At x^λ we expect the false alarm rate to be

$$P(x^\lambda|F) = \int_{-\infty}^{x^\lambda} p(x|F)dx \tag{50}$$

so that the expected number of F instances occurring before the point x^λ is $N_F P(x^\lambda|F)$. Then a lower bound on the expected number of instances which must be labelled in order to find all T occurrences in a sample of size N is

$$E(\lambda) \geq N_T + N_F P(x^\lambda|F). \tag{51}$$

This expression is only a lower bound on the expected value of λ because the extreme T instance can occur anywhere in the interval $[x^\lambda\ \infty]$. We note that computing this lower bound requires that we integrate the distribution $p(x|T)$ up to a small probability $1/N_T$ (equation 49) so that the corresponding value x^λ will be in the tail of $p(x|T)$. In practice the generating distributions must usually be estimated from a small set of labelled training data, and extrapolation beyond the range of observations into the extreme tails of an estimated distribution is prone to large errors (Pandey, 2001). We will demonstrate this problem in section 5.4 with applications to real data sets.

Now to derive an approximate probability distribution for the discrete random variable λ, we consider a test sample comprised of the union of independent samples from $p(x|T)$ and $p(x|F)$. The test data have the score distribution

$$p(x) = p(x|T)P(T) + p(x|F)P(F). \tag{52}$$

The empirical ROC is generated by ordering the test data from smallest to largest score (decision statistic). The ith item in this ordered list of N samples (the ith order statistic $x(i)$) has the probability distribution (Balakrishnan and Cohen, 1956)

$$p(x|x(i)) = \frac{N!}{(i-1)!(N-i)!} P(x)^{(i-1)}(1-P(x))^{(N-i)} p(x). \tag{53}$$

The probability of observing a sample from the T class at x is given by Bayes rule

$$P(T|x) = \frac{p(x|T)P(T)}{p(x|T)p(T) + p(x|F)p(F)}. \tag{54}$$

Figure 35 illustrates the computation of equations 52 through 54. Marginalizing over x we obtain the probability that the ith order statistic in the test data belongs to the T class

$$P(T|x(i)) = \int_{-\infty}^{\infty} P(T|x)p(x|x(i))dx \tag{55}$$

with $P(F|x(i)) = 1 - P(T|x(i))$. Figure 36 verifies computation of equation 55.

To compute the probability that the ith order statistic is the last true item in the sample ($P(i = \lambda)$), we must enumerate all permutations for which this outcome can occur and compute the probability of each permutation. This quickly becomes prohibitively expensive for even modest sample sizes, and so here we approximate the distribution of $P(i = \lambda)$ as follows. If the ith order statistic is the last T item in the sample, then the remaining $i+1$ through N order statistics must all be F. Therefore the probability that the ith order statistic is the last T is approximately equal to

$$P(i = \lambda) \approx P(T|x(i))P(F|x(i+1))P(F|x(i+2))...P(F|x(N)). \tag{56}$$

Figure 37 shows computation of equation 56 for the generating distributions shown in figure 32(a), with equal priors and for sample sizes ranging from $N = 200$ to 800.

The probability mass function $P(i = \lambda)$ can in principle be used to select an operating point λ. For example, we might adopt the order statistic which is most probably the maximum T instance in a sample of size N as an optimal operating point (i.e. the mode of the distribution of $P(i = \lambda)$). Figure 38 shows the dependence of the mode of λ upon sample size and prior probability of the T class. The operating point increases logarithmically with sample size, with the rate of increase weakly dependent upon the

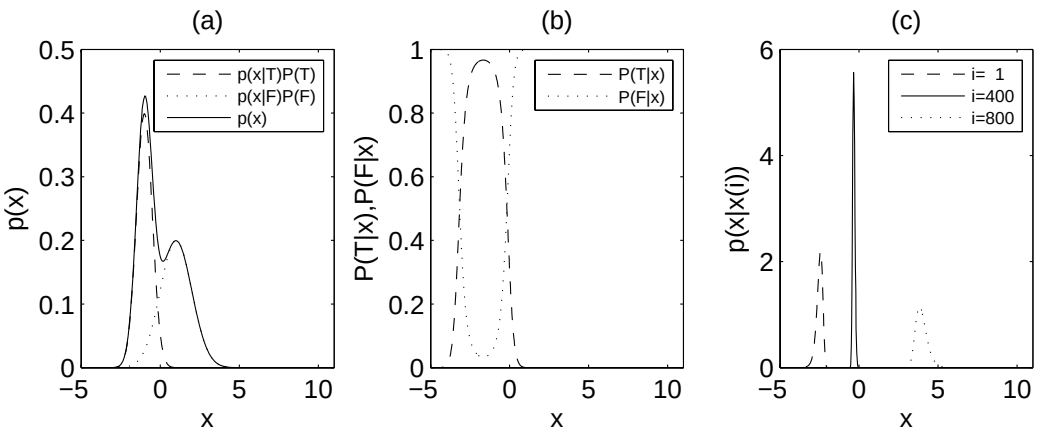


FIGURE 35. Steps in computing the probability $p(T|x(i))$ of a T instance given the ith order statistic from a sample. (a) The probability distribution of the test data $p(x)$ is comprised of the distributions $p(x|T)p(T)$ and $p(x|F)p(F)$ (equation 52). (b) The probability $P(T|x)$ of a T instance given score x (equation 54). (c) The probability distribution $p(x|x(i))$ of obtaining score x given that x is the ith item in an ordered list of N scores sampled from $p(x)$ (here $N = 800$). The probability $P(T|x(i))$ is obtained by multiplying each $p(x|x(i))$ by $p(T|x)$ and integrating over all x.

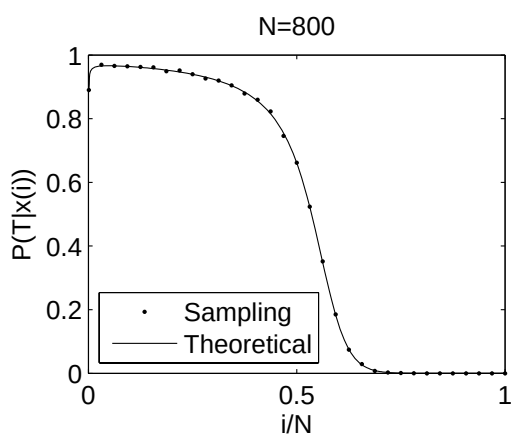


FIGURE 36. Computing the probability $P(T|x(i))$ of a T instance given the ith order statistic. Points are generated by repeated sampling from the distribution $p(x)$, ordering the sample, and then determining the proportion of realizations for which the i/N^{th} item is T. Solid line is the prediction from equation 55, with the required integrals evaluated numerically.

prior probabilities. Also shown in figure 38 is the lower bound on the expected value of λ computed with equation 51. We see that the lower bound asymptotes to the most probable value as sample

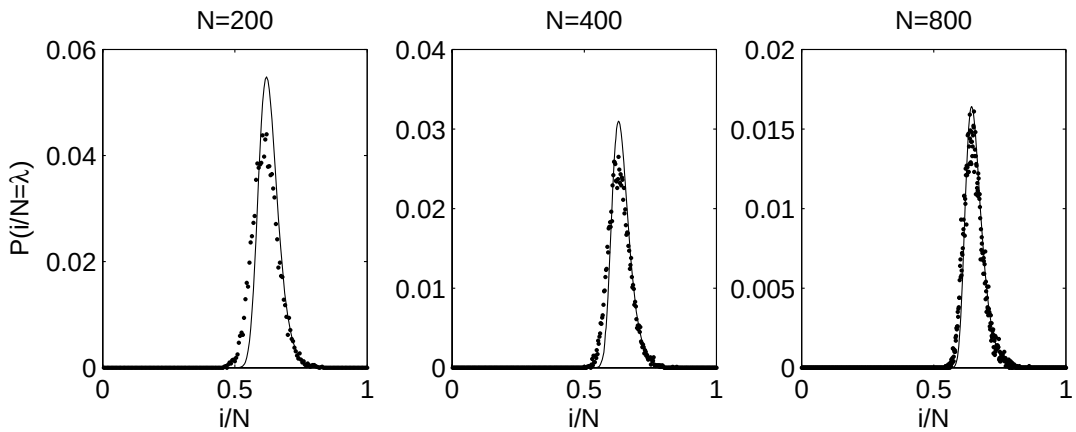


FIGURE 37. Probability that the i/N^{th} order statistic is the maximum T instance (λ) in samples of size N. In all plots dots are the result of simulations, solid lines are theoretical calculations.

size increases. This makes equation 51 a useful approximation for large sample sizes where computing equation 55 over all order statistics can become quite slow.

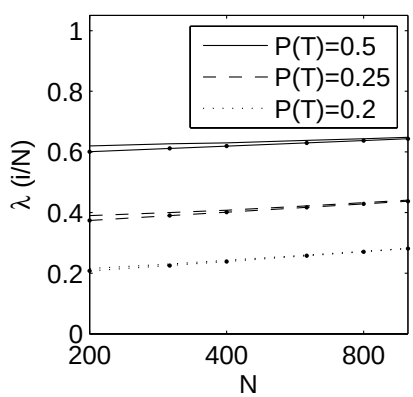


FIGURE 38. Dependence of the optimal operating point λ on sample size N and prior probability T for the generating distributions of figure 32(a). Lines with markers are a lower bound on the expected value of λ (equation 51).

We now return to the idea of specifying a number (N_L) of F instances occurring sequentially in the ordered diglist which will trigger us to stop digging. The approximate distribution of λ can help to understand how a particular choice of N_L affects the probability of finding all T instances. Using the probabilities $P(F|x(i))$ (equation 55), figure 39 shows the probability $P(N_L)$ of observing N_L F instances in sequence immediately prior to the most likely value of λ in an ordered list of length N. If $P(N_L)$ is small then there is a small probability that we will terminate digging prior to the final T

instance. As N_L is increased, $P(N_L)$ will necessarily decrease monotonically since $P(F|x(i)) < 1$. The curves in figure 39 have the character of a Tikhonov curve encountered when solving a regularized inverse problem. This suggests using a heuristic from Tikhonov regularization (such as the point of maximum curvature) to select N_L. However, in practice it will be difficult to construct these curves from limited training data and so we seek heuristics which depend only on the size of the sample N. In figure 39 we find that P(N_L) is bounded by the function

$$P(N_L) \leq \exp\left(-\frac{N_L^2}{N}\right). \tag{57}$$

Setting $P(N_L) = 1/N$ in the above expression and solving for N_L yields

$$N_L = \lceil \sqrt{N \log N} \rceil \tag{58}$$

with the inequality converted to an equality and the ceiling function ($\lceil \cdot \rceil$) mapping to the next largest integer. As shown in figure 39, this approach yields a consistently small value of $P(N_L)$ for the sample sizes considered. Beyond figure 39, we have no theoretical basis for justifying the form of equation 58, and the result in figure 39 is of course particular to the distributions, prior probabilities, and sample sizes in this example. Other choices for the dependence of N_L on sample size can be considered, in the remainder of this paper we demonstrate the efficacy of this particular form with simulations and applications to real data.

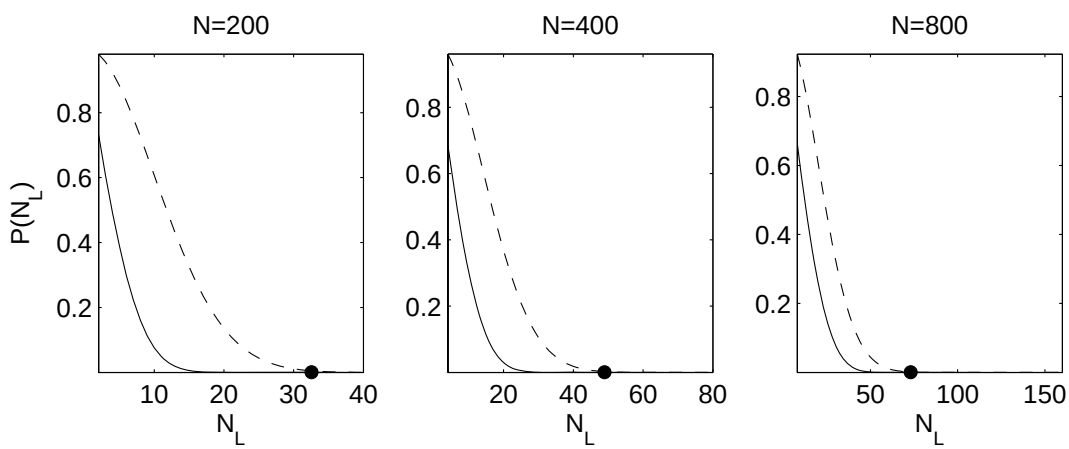


FIGURE 39. Solid lines: probability $P(N_L)$ of observing N_L F instances in sequence, immediately prior to the most likely value of λ (the last T instance), for varying sample sizes N. Dashed lines: an upper bound on $P(N_L)$ (equation 57). Markers indicate N_L computed with equation 58.

5.3. **Simulations.** To illustrate the relative merits of criteria for selecting an operating point, we simulate the performance of the following operating points:

(a) the point on the training data ROC closest to (0,1).
(b) the lower bound on the expected value of λ (equation 51), assuming normal probability distributions and with required probabilities and priors estimated from the training data. This is an approximation to the most probable value of λ, as shown in figure 38.
(c) the number N_L of F instances which must occur sequentially, computed with equation 58.

We estimate all operating points for (a) and (b) in the above list from the same realization of training data, and then apply these operating points to an independent test data set to compute true and false positive fractions (TPF, FPF) at each operating point. For each realization of training data a single realization of test data is used to evaluate all operating points. This is repeated for 100 independent realizations of training and test data, allowing us to estimate means and standard deviations of TPF and FPF at all operating points. Generating distributions for T and F classes are as depicted in figure 32, and we used equal priors when generating the test data. These simulations are repeated for training data sets ranging from $N = 200$ to $N = 800$, as in figure 34, with test sample size fixed at $N^{test} = 1000$.

In table 5.3 we display the proportion of test realizations for which a given operating point is able to identify all T instances, $P(TPF = 1)$, as well as the mean and standard deviations of the false positive fraction at the operating points. Not surprisingly, the closest (0,1) operating point has zero chance of finding all T instances. The λ operating point asymptotes to finding all T instances approximately half of the time. This is consistent with the distribution of λ in figure 39: at the most probable value of λ there is still approximately half of the area under the distribution greater than this value. To achieve a higher $P(TPF = 1)$ we can integrate the distribution of λ up to some desired confidence, though of course this will increase the expected number of F instances which must be labelled. While this approach seems to represent a reasonable compromise between finding all true positives while limiting the number of false positives, it is contingent upon knowing the correct form of the probability distributions. For these simulations normal probability distributions are assumed and so the performance of the λ operating point is quite good. This assumption will not hold in general and so actual performance of this operating point may not be as suggested by these simulations. This is illustrated on a real data example in the next section. Finally, we see that the N_L operating point has a high probability of finding all T instances, though with a corresponding increase in false positive fraction.

N^{train}	(a) Closest (0,1)		(b) λ (Normal)		(c) N_L	
	P(TPF=1)	*FPF*	*P(TPF=1)*	*FPF*	*P(TPF=1)*	*FPF*
200	0.00	0.10 ± 0.02	0.43	0.30 ± 0.04	0.93	0.46 ± 0.06
400	0.00	0.10 ± 0.02	0.45	0.31 ± 0.04	0.93	0.46 ± 0.06
800	0.00	0.10 ± 0.01	0.54	0.32 ± 0.03	0.96	0.47 ± 0.07

TABLE 5. Performance of operating point criteria for simulations with varying training data size N^{train}. $P(TPF = 1)$ is the proportion of test data realizations for which the operating point found all T instances, FPF is the mean and standard deviation of the false positive fraction at the operating point.

5.4. **Application to real data.** We conclude with applications of operating point criteria to unexploded ordnance discrimination. The problem of selecting an operating point has received limited attention in this context: Zhang et al. (2004) derive an expression for a probability threshold which depends only

upon the numbers of true and false instances in the training data

$$P^{\lambda} = \frac{\sqrt{N_T}}{\sqrt{N_T} + \sqrt{N_F}}. \tag{59}$$

This assumes that the score distributions are normally distributed with equal variance. Operating points selected by this method typically identify 90% of the test ordnance items in the published results, and we found that this technique had comparable performance to the closest (0,1) method for the simulations reported in the previous section. We retain the latter method for application to real data sets. For discrimination of UXO using frequency domain electromagnetic data, Huang et al. (2006) maximize a "degree of discrimination" parameter identical to the maximum vertical distance from the 45 degree line described in section 5.1. This method also found approximately ninety percent of ordnance in the test data.

In figure 40 we compare the performance of operating point criteria applied to a data set from Guthrie road, Montana. The discrimination problem at this site was to discriminate 81 mm mortars and 76 mm projectiles from clutter using magnetics data. As described in Billings (2004), for each observed magnetic anomaly we estimate an (apparent) remanence quantifying the deviation of the observed dipole moment from a library of expected moments for ordnance targets. This provides a decision statistic which can be used to discriminate ordnance (small remanence) from clutter (large remanence). We see in figure 40(a) that the empirical distributions of the log transformed remanence are approximately normal. However, the empirical distribution of ordnance is not symmetrical about its mean, such that the upper tail of the normal distribution is a poor characterization of the actual distribution of ordnance. Extrapolation of a fitted normal distribution to determine the operating point λ therefore overshoots the last T (ordnance) instance and results in a large number of unneccesary digs on the ROC (40(b)). In this plot we have not normalized the ROC axes by the numbers of true and false positives, so that the actual numbers of ordnance (80) and clutter (644) are shown. The operating point selected with $N_L = 70$ finds all ordnance in this case, with an acceptable overshoot past the last ordnance item. In comparison, the expected closest (0,1) operating point achieves a $TPF = 0.925$ (6 ordnance are left in the ground). This is consistent with previously reported results for this method. In practice, the λ or closest (0,1) operating points are estimated from a small training set and subsequently applied to the remaining test data. Here we have instead used the full data set to choose these operating points and so the reported performance does not capture the variability of these operating points introduced by the training procedure. A bootstrap analysis can be used to quantify this effect, but the simulations in the previous section are likely indicative of the relative variability of each technique. An obvious advantage of the N_L operating point is that it depends only upon the size N of the data set and so does not require a training data set in order to determine an operating point.

Figure 41 shows example ROCs generated in a demonstration study at San Luis Obispo. The discrimination task at this site was to find a variety of ordnance targets ranging in size from 4.2" mortars down to 60 mm mortars. A number of electromagnetic sensors were deployed at the site, here we show results for

(a) EM61 cart operating in a detection mode survey. The decision statistic is the estimated rate of decay of the induced dipole moment.
(b) EM61 with a magnetometer (MSEMS sensor) operating in a detection mode survey. The decision statistic is as in (a).
(c) TEMTADS array operating in a cued interrogation survey. The decision statistic is the maximum correlation coefficient between the observed data and data predicted by a best fitting item in a library of pre-defined ordnance polarizations (fingerprinting method described in Pasion et al. (2007)).

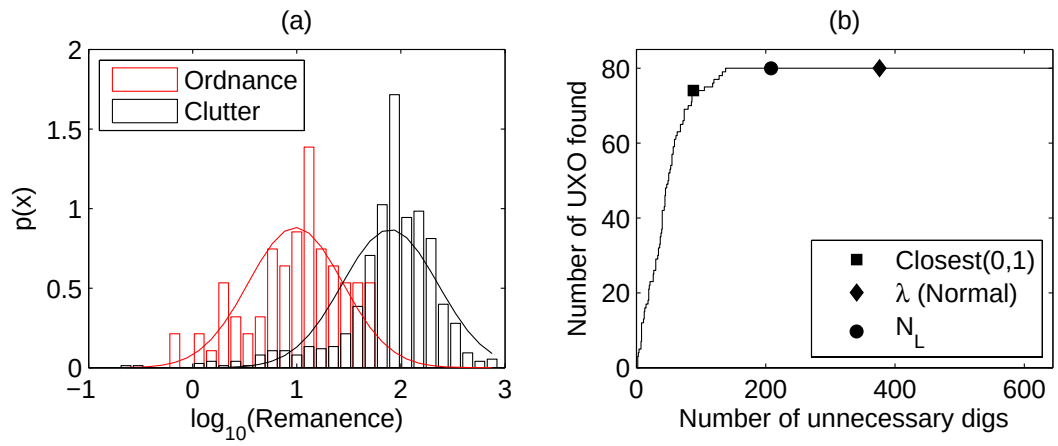


FIGURE 40. Operating points applied to Guthrie road data. Left plot shows the empirical distributions of the decision statistic, apparent remanence, for ordnance and clutter classes. Solid lines are normal distributions fit to these data. Right plot shows the ROC generated by thresholding from smallest to largest decision statistic, with operating points selected by closest (0,1), λ (Normal) and N_L methods.

(d) Same TEMTADS data set as in (c), but with the decision statistic the probability of membership in the ordnance class as predicted by a statistical classifier trained on size and decay rate of the primary polarization (similar to the approach employed in chapter 2 of this thesis on Camp Sibert data).

Given the limitations of the λ operating point demonstrated above, here we show only the closest (0,1) and N_L operating points. In figure 41 (a), (b) and (c) the N_L operating point achieves $TPF = 1$, while in (d) we obtain $TPF = 0.990$ (2 ordnance left in the ground). Again, the closest (0,1) point finds approximately ninety percent of the ordnance in all cases. The slightly different results for (c) and (d), both derived from the same TEMTADS data set, emphasize that the ROC curve depends not only on the geophysical data, but on the discrimination strategy used to process these data. While careful selection of an operating point can improve our chances of finding all ordnance, failures in inversion and quality control can introduce outliers to the ordnance class. For the TEMTADS statistical classification (figure 41(d)), the two outliers to the ordnance class are lower SNR 60 mm mortars for which the decay parameter is difficult to estimate. More advanced discrimination techniques (e.g. robust inversion, or incorporating parameter uncertainty in inversion) may reduce the occurrence of outliers and improve the probability that $TPF = 1$.

5.5. **Discussion.** We have investigated selecting the operating point on a receiver operating curve. Previously published criteria are generally designed to balance the trade-off between true and false positive fractions on the ROC. This type of trade-off occurs in many analogous problems. For example, in Tikhonov regularization of underdetermined inverse problems we seek a model which balances the misfit to the observed data and some measure of model complexity. For the regularization problem a model can be selected with heuristics (e.g. selecting a point of maximum curvature on the Tikhonov curve), cross-validation techniques, or statistical criteria (e.g. achieving a target data misfit).

Here we have developed heuristics and statistical criteria for the operating point problem when the total number of instances which must be labelled is known. We derived an approximate probability

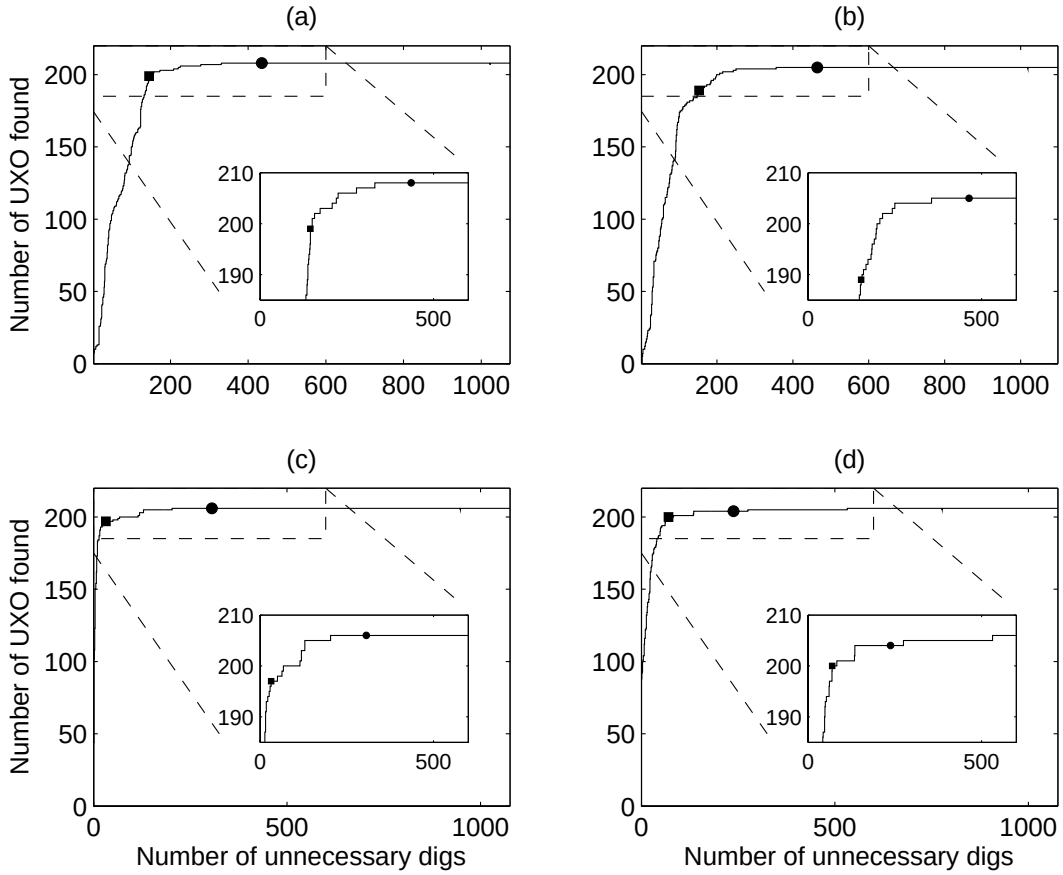


FIGURE 41. Operating points applied to San Luis Obispo data. Squares are closest (0,1) operating points, circles are N_L operating points. (a) EM61 cart (b) MSEMS EM61 (c) TEMTADS fingerprinting (d) TEMTADS statistical classification.

distribution for the discrete random variable λ, which we defined as the order statistic at which all T instances are labelled. This probability distribution depends upon the generating distributions, prior probabilities, and sample size. Given this probability distribution, we can select an operating point which corresponds to the most likely value of λ. In addition, we have derived a lower bound to the expected value of λ which is a useful approximation to the most likely value. However, this approach has limited practical applicability because it depends upon accurate estimation of the generating distributions and extrapolation into the extreme tails of these distributions. To address this shortcoming, we have proposed a heuristic for selecting N_L, the number of F instances which must occur sequentially before digging is terminated. This is equivalent to cost minimization, but the proposed heuristic provides an objective means of choosing relative costs based upon sample size. In simulations and applications to real data we find that this technique has an improved probability of finding all ordnance in a test data set, relative to previously published methods. We have limited our investigations to samples on the order of $N = 10^3$, which is representative of the number of detected targets at many sites. Tests on larger data sets should still be carried out.

In previous work we considered a bootstrapping approach to selecting the operating point. While this method may still be viable we find that it is highly dependent upon the training data realization and does not exploit information in the test data as digging proceeds. In contrast, specifying the parameter N_L does not depend on training data and termination of digging depends upon the test data (rather than some pre-specified point derived from limited training data). If successful in finding all TOI, the proposed approach will always overshoot the last target of interest by N_L items, but this is a necessary expense if we are to have confidence that all TOI have been identified. Furthermore, when digging is terminated with this method, we can accurately estimate the distributions of T and F items and use this to compute a confidence that no more targets of interest remain in the ground. A program of verification digging (e.g. using Visual Sample Plan) should also be employed to provide independent confirmation of the stated confidence level.

6. Discussion

At the same time that new methods were being developed under this project, comprehensive tests of discrimination performance were conducted at three sites as part of the ESTCP Discrimination Pilot Study. These included the Former Camp Sibert in Alabama (mostly production EMI sensors with 4.2" mortars the primary TOI), San Luis Obispo in California (production and next generation sensors with 60 mm, 81 mm and 4.2" mortars, and 2.36" rockets as the primary TOIs) and Camp Butner in North Carolina (production and next generation sensors with 37 and 105 mm projectiles, and m48 fuzes as TOI).

At the demonstration sites, the highest quality data and best discrimination results were achieved in cued-interrogation mode by instruments that dwell at a fixed location while changing the transmitter excitation pattern (e.g. MetalMapper, TEMTADS). ROC curves from the next generation sensor data at both SLO and Camp Butner were near vertical initially (many TOI recovered with low numbers of false-alarms) but often tended to flatten out, with many false alarms excavated before all TOI were recovered. In effect, one part of the discrimination problem (with next-generation sensor platforms deployed in cued-interrogation mode) is relatively easy, with the second part more challenging. The discrimination problem was "easy" when the following conditions were met:

(1) single object in the field of view,
(2) cued-survey centred approximately over item location, and
(3) high SNR.

The discrimination problem was "hard" when the above conditions were not met, typically due to one or more of the following issues:

- Challenge 1: Multiple objects in the field of view;
- Challenge 2: Anomaly response obscured by larger adjacent target or targets (similar to 1)
- Challenge 3: Data insufficient to constrain all components of the polarization tensor due to one or more of the following:
  - Low SNR
  - Deep burial
  - Object not centered under array

While not experienced at SLO or Camp Butner, viscous remanent magnetization of soil can also present a significant challenge to effective discrimination.

The best results at SLO and Camp Butner were obtained with next generation sensors deployed in a fixed location in a cued-interrogation mode. Additional challenges arise when dealing with other survey modes as follows:

- Cued-interrogation by moving sensor: Including instruments such as the MPV where cued-data are collected by moving the sensor and tracking its position (either using a template or a beacon). The challenges are the same as for the static cued systems, but with the additional complication that (1) we need to ensure the anomaly is adequately sampled; and (2) positional error could come into play.
- One-Pass detection and discrimination: Includes instruments such as the One Pass TEM Array (OPTEMA) where measurements are collected while continuously moving and varying the transmitter excitation strategy. Challenges are similar to the MPV type situation, except with the one-pass we can't control the position of the object relative to the survey instrument.
- Detection mode: Includes instruments such as the MetalMapper Dynamic, where the measurements are collected while continuously moving, but the transmitter excitation strategy doesn't change. Challenges are similar to the one-pass system with the extra consideration that these systems were not designed to fully excite and measure the response of the 3 principle axes of an

object. Thus even under ideal conditions (adequate SNR, shallow depth, adequate coverage), the data won't be able to constrain all components of the polarization tensor.

In addition, while discrimination using production quality EM61 data was not very effective at Camp Butner, good results were achieved at SLO and Camp Sibert. This was mostly due to a better separation between the decay characteristics of the TOI compared to the non-TOI items. At suitable sites, we expect that some level of discrimination will be possible with an EM61 cart.

6.1. **Principal contribution of the work performed under this project.** While the discrimination results at SLO and Camp Butner showed the great promise of the combination of next-generation sensor data with high quality parameter extraction and classification algorithms, they also demonstrated that we can't always efficiently recover 100% of all TOI without also digging up a significant number of non-TOI. The principal contribution of the work reported here was in developing algorithms and strategies that minimize or eliminate the discrimination outliers that were often encountered at Camp Sibert, SLO and Camp Butner. That is, the methods were particularly efficacious when applied to the "hard" anomalies encountered at a site. This is because the methods exhibit the following two desirable attributes:

- Parameter extraction: The methods are robust and tolerant of data quality issues; and
- Discrimination: The methods are adaptable and sensitive to the uncertainty in the recovery of extracted parameters.

6.2. **Potential additional research under the robust statistics and regularization theme.** The following are directions of research within the area of parameter extraction that we believe deserve additional attention:

(1) *Robust statistical norms and multiple objects in the field of view*: The algorithms developed here only work when there is a single object in the field of view and need to be adapted for multiple objects. Good and viable multi-object code using the least-squares approach exist (Song et al., 2011) and these would need to be adapted to robust statistical norms.

(2) *Regularization using reference models*: An alternative form of regularization still to be investigated is a penalty on the deviation of the recovered polarizabilities from some reference model. Here the reference model is a library entry for a given ordnance type. This approach would regularize all polarizabilities and represents a compromise between fixing polarizabilities at their library values (as in the fingerprinting method developed under MR-1637) and an unconstrained inversion. This may help to smooth noisy polarizability estimates for low SNR targets. When discriminating multiple ordnance types, a separate regularized inversion (i.e with different reference polarizabilities) will be required for each ordnance type. The proposed approach is similar to that employed in Aliamiri et al. (2007) (where the inverse problem is formulated in the context of Bayesian maximum a priori (MAP) estimation) but has not been tested with field data.

(3) *Include positional uncertainty in the inversion algorithm*: We believe there is merit in specifically targeting the issue of sensor positioning and orientation errors that arise from platform or cart motion when data are collected dynamically.It is known that the measured EMI response is a function of the location and dipole moment of an object as well as sensor positions. A precise knowledge of sensor positions is generally required for accurate estimation of equivalent dipole polarizabilities. For dynamically collected data, this stringent requirement is often hard to meet, with accuracies typically on the order of several centimeters. Perturbations in sensor positions, if not accounted for properly, can significantly degrade inverse results. Thus, it is highly desirable to develop an EMI inversion scheme that can explicitly account for errors in the position and orientation of the sensor.

Several signal processing approaches have been proposed to overcome sensor position uncertainties including Tarokh and Miller (2007) and Tantum et al. (2008). The min-max approach

of Tarokh and Miller (2007) was designed to look for the parameters of interest that minimize the maximum data residual, where the maximum error is computed over ellipsoids or polyhedra of possible sensor locations defined by the bound information. This approach is computationally prohibitive if all possible scenarios need to be trialed in the search for the worst case scenario. To avoid intensive computation, Tarokh and Miller (2007) simplified their min-max approach by using a using a box-shaped uncertainty region, and performed the min-max inversion by testing over the eight corner points of the box for each sensor location. The Bayesian approach of Tantum et al. (2008) accounted for positional error by using a prior probability density function on positions. A high-dimensional integration over all unknown parameters was performed to compute the maximum likelihood estimate of the underlying dipole parameters. This approach is sensitive to the assumed prior density function. Although both these studies were numerical and experimental in nature, they did show the feasibility of improving EMI inversion by incorporating prior knowledge of sensor positioning errors into parameter extraction algorithms.

A potential research topic is the development of a technique which explicitly treats sensor positions/orientations as additional model parameters in the inversion and aims to replace the original sensor positions/orientations with nominally more accurate estimates. This is an extended inverse problem where we will determine not only target parameters but also relocate sensor positions and orientations from observed data.

At first glance it appears that inverting for sensor positions would be computationally prohibitive as it increases the number of unknowns by three times the number of positions (six times if we also invert for sensor attitude). We believe that there are a number of factors that work in our favour. **First**, a UXO survey is a controlled experiment in which nominal sensor positions are close to true sensor locations and can be good starting points for solving the associated nonlinear inverse problem. This is in contrast to the case of determining the location of the obscured object where we have to find a good initial model. **Second**, prior estimates of positioning accuracy can be made and used as useful constraints in the inversion process. **Third**, if we assume that the perturbations in each sensor can be treated as independent of others (Tarokh and Miller, 2007) then the relocation of each sensor position is decoupled from the other positions. This simplifies the problem and makes it tractable. **Fourth**, and more importantly, next-generation sensors typically take multi-time/multi-static/multi-component measurements which provide the additional independent information needed to accurately relocate sensor positions.

In this extended problem the non-linear coupling between sensor positions and target parameters is mathematically challenging and introduces additional degrees of nonlinearity and non-uniqueness into the solution process. We will initially utilize a solution strategy that involves decomposing the extended inverse problem into two major steps in each of which one category of model parameters is estimated. That is, the procedure is first to find locations and dipolar polarizations given nominal sensor positions and then to optimize sensor locations given the current estimate of target parameters. The process can be implemented in a sequential, iterative manner until a pre-defined misfit level is reached (or the model reaches some convergence tolerance).

A field data example of the extended inversion technique using the MetalMapper dynamic data collected at SLO is shown in Figure 42. These are just preliminary tests but show that the new method for accounting for sensor position errors, has promise.

In summary, to develop this extended inversion technique, we would need to: 1) Study the sensitivities of transient electromagnetic signals to perturbations in sensor positions and orientations in order to assess the relative impact of the mis-located or mis-recorded spatial coordinates on the measurements; 2) Investigate and quantify the trade-off and resolvability between two

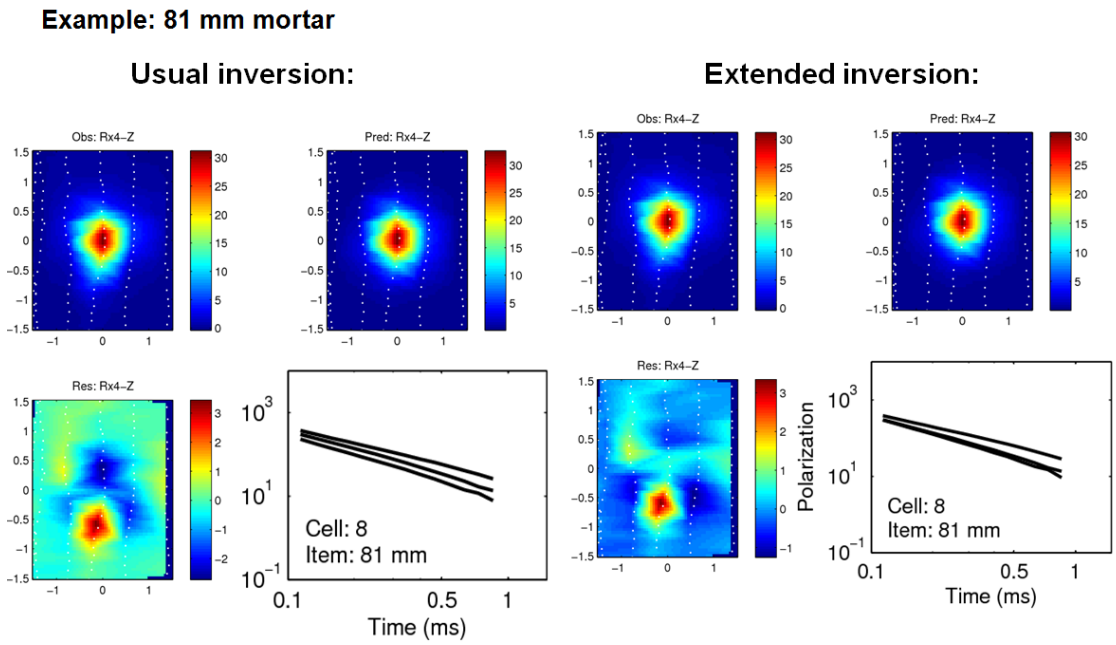


FIGURE 42. Polarizabilities extracted from MetalMapper data collected in dynamic mode at SLO when using the an algorithm that accounts for errors in sensor position. The secondary and tertiary polarizabilities extracted from the extended inversion are approximately equal providing evidence that the obscured object could be an axially symettric ordnance item.

set of model parameters, (i.e, one set that is related to the object, and the other with sensor positions) using a number of sensor systems (e.g., OPTEMA, MPV and dynamic MetalMapper); and 3) Formulate and develop an optimal, robust inverse approach that is able to effectively account for the presence of sensor position uncertainties via a relocation process .

On the classification front we believe the following areas deserve additional attention:

(1) *Methods to determine when to stop digging*: Rather than guessing the stop-dig point prior to digging based upon limited training data, we should make this decision based upon information that becomes available as digging proceeds. When no TOI are encountered in a specified number of digs N then we can stop remediation. The question is then how to determine an appropriate N such that we reach a certain probability that no TOI remain in the ground. In the work performed here we showed that estimation of these quantities requires accurate characterization of TOI and non-TOI distributions, in particular the tails of the TOI distribution. This is impossible with limited training data, and we believe there is merit in pursuing methods to learn these distributions from the test data ground truth. These new methods should be compared with the power analysis approach of Hathaway et al. (2009) for assessing the probability that TOI remain in the ground.

(2) *Methods to choose and/or assess the adequacy of training data*: The Camp Butner demonstration provided a test ground for development of techniques to build the training data set by iteratively digging test items. Our approach was to query test vectors in regions of overlap between TOI and non-TOI in order to determine the extent of the TOI distributions in the feature space. More formally, this corresponds to regions of the feature space where the predicted class

memberships are most uncertain (i.e. have maximum entropy). In addition, we tried to identify clusters of self-similar test targets by computing a symmetric matrix of the misfits between all test vector polarizabilities. A number of feature vectors with small mutual misfits are then indicative of a repeated polarizability fingerprint in the test data. We remark that this unsupervised analysis has connections to the semi-supervised algorithm presented in Liu et al. (2008). While our unsupervised analysis did help to identify 37mm TOI with slightly different polarizability responses from calibration items, it failed to pick up some consistent non-TOI classes in the test data (e.g. spent fuzes). Further investigation of these techniques is therefore required, as well as development of methods for re-training classifiers as additional ground truth becomes available during digging.

(3) *Combining different classification strategies*: Another topic to explore is the update of the discrimination strategy as digging proceeds to account for varying difficulty of the problem. TOI are typically easy to identify initially, and so an aggressive strategy which uses the full polarizability response will produce good performance. However, as we encounter smaller, lower SNR targets later in the diglist, we risk overfitting the training data and must therefore switch to a less aggressive classifier (even when using robust statistical methods). Simpler classifiers trained on total polarizabilities or size and decay parameters can help prevent outlying TOI. The results reported here demonstrated the promise of this technique which needs to be further explored at additional sites and with additional types of sensor data.

References

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

N. Balakrishnan and A. C. Cohen. *Order statistics and inference: estimation methods*. Academic Press, 1956.

T. Bell. Geo-location requirements for UXO discrimination. Technical report, Science Applications International Corporation, 2005.

T. Bell. Adaptive and iterative processing techniques for overlapping signatures: Technical summary report. Technical report, ESTCP, 2006.

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

L. S. Beran, S. D. Billings, and D. Oldenburg. Robust inversion of time-domain electromagnetic data: application to unexploded ordnance discrimination. *Journal of Engineering and Environmental Geophysics*, in press, 2011a.

L. S. Beran, S. D. Billings, and D. Oldenburg. Incorporating uncertainty in unexploded ordnance discrimination. *IEEE Trans. Geosci. Remote Sensing*, in press, 2011b.

S. D. Billings. Discrimination and classification of buried unexploded ordnance using magnetometry. *IEEE Trans. Geosci. Remote Sensing*, 42:1241–1251, 2004.

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

M Brookes. The Matrix Reference Manual, 2005. URL `http://www.ee.imperial.ac.uk/hp/staff/dmb/matrix/`.

C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:955–974, 1998.

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

C. Farquharson and D. W. Oldenburg. Nonlinear inversion using general measures of data misfit and model structure. *Geophysical Journal International*, 134:213–227, 1998.

T. Fawcett. ROC graphs: Notes and practical considerations for researchers. Technical report, HP Labs, 2004.

G. H. Golub and C. F. van Loan. *Matrix Computations*. Baltimore, MD, USA, second edition, 1989.

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

J. Hathaway, R. Gilbert, J. Wilson, and B. Pulsipher. Evaluation of spatially clustered ordnance when using compliance sampling surveys after clean-up at military training sites. *Stoch. Environ. Res. Risk. Assess.*, 23:253–261, 2009.

H. Huang, B. SanFilipo, and I.J. Won. Optimizing decision threshold and weighting parameter for UXO discrimination. *Geophysics*, 71:313–320, 2006.

T. Kanungo and R. M. Haralack. Receiver operating characteristic curves and optimal Bayesian operating points. In *1995 International Conference on Image Processing (ICIP'95) - Volume 3*, 1995.

N. Lhomme, D. W. Oldenburg, L. R. Pasion, D. Sinex, and S. D. Billings. Assessing the quality of electromagnetic data for the discrimination of UXO using figures of merit. *Journal of Engineering and Environmental Geophysics*, 13:165–176, 2008.

Q. Liu, X. Liao, and L. Carin. Detection of unexploded ordnance via efficient semi-supervised and active learning. *IEEE Trans. Geosci. Remote Sensing*, 46:2558–2567, 2008.

S. Macskassy, F. Provost, and S. Rosset. ROC confidence bands: An empirical evaluation. In *Proceedings of the 22nd International Conference on Machine Learning*, 2005.

R. A. Marrona, R. D. Martin, and V. J. Yohai. *Robust Statistics: Theory and Methods*. J. Wiley, 2006.

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

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

D.W. Oldenburg and Y. Li. *Investigations in Geophysics No.13: Near-surface*, chapter Inversion for applied geophysics: A tutorial, pages 89–150. Society of Exploration Geophysics, 2005.

M. D. Pandey. Extreme quantile estimation using order statistics with minimum cross-entropy principle. *Probabilistic Engineering Mechanics*, 16:31–42, 2001.

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

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

L. R. Pasion, L. S. Beran, B. C. Zelt, and K. A. Kingdon. Feature Extraction and Classification of EMI Data, Camp Butner, NC. Technical report, ESTCP, 2011.

N. J. Perkins and E. F. Shisterman. The inconsistency of "optimal" cut-points using two ROC based criteria. *American Journal of Epidemiology*, 163:670–675, 2006.

M. Sen and P.L. Stoffa. *Global Optimization Methods in Geophysical Inversion*. Elsevier, 1995.

F. Shubitidze, K. O"Neill, S. A. Haider, K. Sun, and K. D. Paulsen. Application of the Method of Auxiliary Sources to the Wide-Band Electromagnetic Induction Problem. *IEEE Trans. Geosci. Remote Sensing*, 40:928–942, 2002.

J. T. Smith and H. F. Morrison. Optimizing receiver configurations for resolution of equivalent dipole polarizabilities in situ. *IEEE Trans. Geosci. Remote Sensing*, 43:1590–1498, 2005.

L. P. Song, L. R. Pasion, S. D. Billings, and D. W. Oldenburg. Non-linear inversion for multiple objects in transient electromagnetic induction sensing of unexploded ordnance: Technique and applications. *IEEE Transactions on Geoscience and Remote Sensing*, 99:1–14, 2011.

S. L. Tantum, Y. Li, and L. M. Collins. Bayesian mitigation of sensor position errors to improve unexploded ordnance detection. *IEEE Geosci. Remote Sensing Letters*, 5:103–107, 2008.

A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM, 2005.

A. B. Tarokh and E. L. Miller. Subsurface sensing under sensor positional uncertainty. *IEEE Trans. Geosci. Remote Sensing*, 45:675–688, 2007.

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

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