SANDIA REPORT

SAND98–1769/2
Unlimited Release
Printed August 1998

Statistical Considerations in Designing Tests of Mine Detection Systems: II – Measures Related to the False Alarm Rate

Katherine M. Simonson

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Prices available from (615) 576-8401, FTS 626-8401

Available to the public from
National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Rd
Springfield, VA 22161

NTIS price codes
Printed copy: A03
Microfiche copy: A01

SAND98-1769/2
Unlimited Release
Printed August 1998

Statistical Considerations in Designing Tests of Mine Detection Systems: II - Measures Related to the False Alarm Rate

Katherine M. Simonson
Signal & Image Processing Systems Department
Sandia National Laboratories
Albuquerque, NM 87185-0844

Abstract

The rate at which a mine detection system falsely identifies man-made or natural clutter objects as mines is referred to as the system's false alarm rate (FAR). Generally expressed as a rate per unit area or time, the FAR is one of the primary metrics used to gauge system performance. In this report, an overview is given of statistical methods appropriate for the analysis of data relating to FAR. Techniques are presented for determining a suitable size for the clutter collection area, for summarizing the performance of a single sensor, and for comparing different sensors. For readers requiring more thorough coverage of the topics discussed, references to the statistical literature are provided. A companion report addresses statistical issues related to the estimation of mine detection probabilities.

This work performed under sponsorship of the DoD/DOE
Memorandum of Understanding on Non-Nuclear Munitions Technology.

CONTENTS

LIST OF FIGURES

1 - INTRODUCTION

The primary statistical metrics used to assess performance for mine detection systems are related to the probability of detecting actual targets (PD), and the false alarm rate (FAR). A companion report [Simonson, 1998] outlines some of the standard statistical techniques and concepts used in assessing PD. In this paper, models and calculations used to quantify the false alarm rate are discussed. These methods can be useful both in designing tests of mine detection systems, and an analyzing the data collected during such exercises.

In target recognition applications, every detection is classified as "true" or "false". A detection is said to be true if an actual target (mine) is present at or near the indicated location. Otherwise, it is false. When controlled tests of mine detection systems are conducted, the investigator knows the number, type, and location of actual mines encountered by each system. Thus, it is sensible to characterize performance on target mines in terms of PD, with inferences based on the ratio of the number of detections to the number of targets emplaced. Typically, PD varies with mine type and size.

All true detections are, by definition, caused by target mines. By contrast, false detections may arise from two different sources: known decoy objects (e.g., bolts, cans) intentionally buried at surveyed locations in the range of the sensor, and the background characteristics and natural variability of the test region itself. In order to ensure that experiments designed to test multiple detection systems are unbiased, it is important to distinguish between these two sources.

During detection system tests, decoys (or "confusers") are sometimes emplaced to provide diagnostic information about the types of objects that can cause a system to false alarm. A discrete number of items is emplaced, and the detection probability can be estimated separately for each different type of object. This probability may be referred to as the "probability of a false alarm" (PFA) for the particular decoy type. Statistically, it is equivalent to the "probability of detection" for a particular target type, and may be analyzed in the same manner, using the methods outlined in Simonson [1998]. The distinction between PD and PFA

is that a high-performing system will have a high PD for relevant targets and a low PFA for common non-targets.

The analysis of "clutter" or "background" false alarms (those representing neither an actual mine nor a known decoy) proceeds differently. Because the investigator is unlikely to have a full characterization of the subsurface of the test region, the cause of false alarms not corresponding to decoy objects is generally unknown. Depending on the particular sensor used, false alarms may be due to rocks, pockets of loose sand, small metal scraps, buried organic material, or other subsurface phenomena. Rather than try to assign a cause to each such detection during the test, the investigator simply notes how many of them occur. Coupled with information about the area (or time) covered by the sensor, this data can be used to characterize the *rate* at which false detections occur in clutter. The purpose of this report is to introduce appropriate statistical methods for assessing such rates.

Figure 1 may help to clarify the distinctions between detections due to targets, decoys, and clutter. The schematic plot shows the spatial layout of an experiment designed to estimate PD for three different target types, PFA for two different decoy types, and FAR for one region that is assumed to be fairly homogeneous. Data related to all three measures is collected in a single run of a sensor over the full test region. Each detection is classified as to cause (known target, known decoy, or unknown clutter) and the analysis proceeds separately for each type.

In the next section, the Poisson model for data related to clutter false alarm rates is introduced. Computations used in estimating the false alarm rate of a single detector are described in section 3. Methods for constructing confidence intervals, conducting hypothesis tests, and specifying the size of the study area are all covered. A technique for comparing the false alarm rates of two different systems is given in section 4. Graphical methods for presenting results related to PD and FAR simultaneously are presented in section 5, and section 6 concludes the report with a few additional considerations for data analysis.

2 - THE POISSON MODEL

2.1 - Background and Assumptions

The Poisson distribution [Johnson, Kotz, and Kemp, 1992; Ripley, 1981] and the closely related Poisson process [Çinlar, 1975; Taylor and Karlin, 1994] are often used to model experimental data related to occurrence rates as a function of time or area. While this report focuses on spatial rates, extension to temporal rates is straightforward.

For applications in mine detection, the experimenter must develop a sensible protocol for determining what constitutes a detection. In addition, rules are needed for determining when a detection corresponds to a known object (target mine or decoy) and when it is a clutter false alarm. Typically, if a detection occurs at a point on the surface lying within a circle of fixed radius about the surface point corresponding to the center of a known buried object, that detection is characterized as being due to the buried object. (For analytical purposes, multiple detections occurring within the same circle are generally treated as a single detection.) To compensate for targets of different size, the radius of each detection region is adjusted to the dimensions of the corresponding object. In order to eliminate ambiguity when assigning causes to detections, decoys and mine targets should be emplaced in such a manner that their detection regions do not overlap. Detections occurring outside of all target and decoy circles are classified as clutter false alarms.

In analyzing false alarm rates, two basic measurements are used. The first measurement is the number of clutter false alarms occurring while a detection system is under test. The second measurement is the difference between the total area covered by the sensor, and the combined area of all covered regions corresponding to targets and decoys. This latter figure represents the area of the region covered within which any detection occurring would be characterized as a clutter false alarm. It is referred to as the clutter area. Note that all decoy detections are omitted from the clutter false alarm count, and all regions lying within decoy detection circles are excluded from the clutter area. This allows the experimenter planning a test comparing several

different systems to include decoys that are attractive to one sensor only, without biasing the reported false alarm rates.

The Poisson model assumes that each detection system being tested has a fixed (but unknown) rate of clutter false alarms per unit area. In statistics, this rate is usually represented by the parameter λ, and is referred to as the intensity of the clutter false alarm process. It is assumed that λ is constant across the test region.

Some discussion of this assumption is in order. To ensure that valid estimates of the FAR can be obtained, tests should be conducted in regions that are believed to be fairly homogeneous. When substantial variation is known to exist, due to factors like changing surface vegetation or moisture content, clay versus sandy soil, or shade versus direct sunlight, the full test region should be divided into smaller sub-regions for performance analysis. A different false alarm rate is then computed for each set of conditions, and statistical tests (see Section 4) can be conducted to determine whether the various environmental factors significantly affect performance.

2.2 - Notation

Suppose that a mine detection system has a true intensity of λ false alarms per unit area under certain conditions. In a system test, let R be the total area covered by the sensor, while the combined area of all regions corresponding to mine targets is R_T, and the combined area of all regions corresponding to decoys is R_D. The clutter area, R_C, is equal to $R - R_T - R_D$. Let K be a random variable representing the number of clutter false detections; K is said to have the Poisson distribution with parameter λR_C [Çinlar, 1975; Johnson, Kotz, and Kemp, 1992]. It is of interest to make inferences about the intensity λ.

According to the Poisson model, the probability of observing k clutter false alarms is given by:

$$\text{Prob}(K = k) = \frac{e^{-\lambda R_C} (\lambda R_C)^k}{k!}, \qquad (1)$$

for $k = 0, 1, 2, \cdots$. The quantity K/R_C is an estimator of the unknown λ. The uncertainty in this estimator decreases as the clutter area covered increases: an experiment measuring 50 clutter false alarms in 100 square meters is more informative than an experiment giving a single false alarm in two square meters. The mean and variance of K/R_C are as follows [Johnson, Kotz, and Kemp, 1992]:

$$E\left(\frac{K}{R_C}\right) = \lambda \tag{2}$$

$$var\left(\frac{K}{R_C}\right) = \frac{\lambda}{R_C}. \tag{3}$$

The variance [Larsen and Marx, 1981] is a common measure of the uncertainty present in an estimator. The quantity (3) decreases as R_C increases, demonstrating that experiments covering a large clutter area will provide more certain information than smaller experiments. Intensity also affects variance: for a fixed clutter area, variance increases with λ.

3 – ESTIMATION AND TESTING FOR A SINGLE POISSON PARAMETER

3.1 - Confidence Intervals for λ

As discussed in Simonson [1998], it is common practice among statisticians, scientists, and engineers to report parameter estimates along with uncertainty measures in the form of confidence intervals. Each such interval is associated with a specified degree of confidence, representing the a priori probability that the interval will contain the true parameter value. Informally speaking, 95% confidence intervals are constructed in such a manner that they will have a 95% chance of containing the true value. Formal definitions of confidence intervals can be found in numerous texts [Bickel and Doksum, 1977; Cox and Hinkley, 1974; Silvey, 1975].

The degree of confidence in an interval is frequently represented algebraically in terms of the quantity α, which is equal to one minus the a priori probability that the interval will contain the true parameter value. Thus, for a 95% confidence interval, α is equal to 0.05. The standard

notational convention uses the expression $100(1 - \alpha)\%$ to represent the certainty corresponding to a generic confidence interval.

Two different approaches are employed to construct confidence intervals for a single Poisson parameter, λ. When the number of detections is large, the (continuous) normal distribution is used to approximate the (discrete) Poisson [Larsen and Marx, 1981; Johnson, Kotz, and Kemp, 1992]. This approximation makes the construction of confidence intervals straightforward for large false alarm counts. For small counts, the normal approximation is inappropriate and a different format for confidence intervals is required. The large sample method is acceptable when the number of clutter detections exceeds 15 [Johnson, Kotz, and Kemp, 1992].

3.1.1 - Small Sample Approach

The small sample method for computing confidence intervals for the Poisson parameter is as follows. Denote the lower limit of an interval by Λ_L, and denote the upper limit by Λ_U. If k false alarms occur in a clutter region of area R_C, an approximate $100(1 - \alpha)\%$ confidence interval for λ is defined by [Hald, 1952; Johnson, Kotz, and Kemp, 1992]:

$$\Lambda_L = \frac{1}{2R_c}\chi^2_{2k,\alpha/2} \tag{4}$$

$$\Lambda_U = \frac{1}{2R_c}\chi^2_{2(k+1),1-\alpha/2}. \tag{5}$$

Here the quantity $\chi^2_{2k,\alpha/2}$ is equal to the $\alpha/2$ quantile of the chi-square distribution with $2k$ degrees of freedom, with $\chi^2_{2(k+1),1-\alpha/2}$ defined in a similar manner. Quantiles of the chi-square distribution are tabulated in many statistics textbooks [e.g. Larsen and Marx, 1981], and are readily available from most commercial statistical software packages.

Equation (4) holds only for $k > 0$. When no clutter false alarms are observed ($k = 0$), the lower limit Λ_L is set to 0.0, while the upper limit is still computed from (5).

A simple example illustrates the small sample method. Suppose that $k = 12$ clutter false alarms occur in a clutter region covering $R_C = 40$ square meters. Then the observed FAR is given by 12/40 = 0.30 per m^2. To construct 95% confidence intervals, α is set at 0.05. From a table of the chi-square distribution, $\chi^2_{24,0.025}$=12.401 and $\chi^2_{26,0.975}$ = 41.923. It follows from (4) and (5) that (0.155, 0.524) is a 95% confidence interval for the false alarm rate (per square meter) of the system under test. Values of λ falling within this interval are deemed to be consistent with the observed data.

Confidence interval width ($\Lambda_U - \Lambda_L$) is a natural measure of the uncertainty present in an estimate. Figure 2 shows lower and upper 95% confidence bounds, as well as interval widths, for clutter false alarm counts ranging from zero to 15 in experiments with $R_C = 50$ m^2 and R_C = 250 m^2. All of the values plotted are calculated from equations (4) and (5).

3.1.2 - Large Sample Approach

The large-sample method of computing confidence intervals uses the normal approximation to the Poisson distribution. When k exceeds 15 clutter false alarms, an approximate $100(1 - \alpha)\%$ confidence interval for λ is given by [Johnson, Kotz, and Kemp, 1992]:

$$\Lambda_L = \frac{k}{R_C} + \frac{1}{2R_C} z^2_{1-\alpha/2} - \frac{z_{1-\alpha/2}}{R_C}\sqrt{k + \frac{1}{4} z^2_{1-\alpha/2}} \qquad (6)$$

$$\Lambda_U = \frac{k}{R_C} + \frac{1}{2R_C} z^2_{1-\alpha/2} + \frac{z_{1-\alpha/2}}{R_C}\sqrt{k + \frac{1}{4} z^2_{1-\alpha/2}} \; . \qquad (7)$$

Here, the quantity $z_{1-\alpha/2}$ represents the quantile of the standard normal distribution corresponding to probability $1 - \alpha/2$. For example, to get a 95% confidence interval, choose $\alpha =$

0.05, and use the value $z_{0.975} = 1.960$ in equations (6) and (7). Tables of the standard normal distribution are found in many statistics textbooks [e.g. Larsen and Marx, 1981], and are readily available from most commercial statistical software packages.

Figure 3 shows upper and lower 95% confidence bounds, along with interval widths, for k ranging from 15 to 250 detections in experiments with $R_C = 50$ and 250 m^2. All of the values shown are computed using Equations (6) and (7). As in the small sample case (Figure 2), the confidence interval widths computed here vary with both k and R_C.

3.2 - Hypothesis Tests

Statistical tests provide a mechanism for choosing among two conflicting hypotheses about the model underlying an observed data set [Koopmans, 1987; Silvey, 1975]. The null hypothesis (H_0) is accepted in the absence of strong evidence to the contrary. The alternative hypothesis (H_1) is accepted when experimental data are deemed to be inconsistent with H_0. The probability of rejecting H_0 when H_0 is true is referred to as the level of a test, and is often denoted by α.

In the case of a single Poisson parameter, one may wish to test whether observed data are consistent with the hypothesis that λ is equal to some specified value, λ_0. The appropriate null and alternative hypotheses are given by:

$$H_0 : \lambda = \lambda_0 \tag{8a}$$

$$H_1 : \lambda \neq \lambda_0 \ , \tag{8b}$$

and the following test statistic is used:

$$z = \frac{|\hat{\lambda} - \lambda_0|}{\sqrt{\lambda_0 / R_C}} \ . \tag{9}$$

Here, $\hat{\lambda} = k/R_C$ is the observed false alarm rate. If the null hypothesis is true, the statistic (9) has a distribution that is approximately standard normal for large k (greater than 15) [Johnson,

Kotz, and Kemp, 1992]. If H_1 is true, z will tend to be large. An α-level test rejects H_0 when z exceeds $z_{1-\alpha/2}$, the $1 - \alpha/2$ quantile of the standard normal distribution.

As an example, consider a test of H_0: $\lambda = 0.10$ versus H_1: $\lambda \neq 0.10$, and suppose that the available data show $k = 18$ clutter false alarms in $R_C = 100$ m^2. This gives an observed false alarm rate of $\hat{\lambda} = 0.18$. The test statistic (9) takes on the value $z = 2.530$. To test at level $\alpha = 0.05$, compare z to $z_{0.975} = 1.960$. Because $z > 1.960$, H_0 is rejected at level 0.05: the observed data are not consistent with a false alarm rate of 0.10 per square meter.

The test statistic (9) is based on the normal approximation to the Poisson distribution and is only appropriate when the number of false alarms is large. For small k, the recommended testing procedure would be to reject (8a) for values of λ_0 not lying in the small-sample confidence interval computed from (4) and (5).

3.3 - Clutter Area Calculation

The normal approximation to the Poisson distribution can be used to calculate the approximate clutter area needed to keep the uncertainty in estimates of λ below some specified level. Here, uncertainty is expressed in terms of $100(1 - \alpha)\%$ confidence interval width. The experimenter begins by specifying an observed clutter false alarm rate, $\hat{\lambda}_0$, and a tolerable confidence interval width, W, corresponding to $\hat{\lambda}_0$. From equations (6) and (7), the width of a $100(1 - \alpha)\%$ confidence interval on λ, when the observed proportion is $\hat{\lambda}$, is given by:

$$c.i.\ width = \frac{2z_{1-\alpha/2}}{R_C}\left[\hat{\lambda}R_C + \frac{1}{4}z_{1-\alpha/2}^2\right]^{1/2} \tag{10}$$

Setting the width equal to the desired value W, and solving (10) for R_C gives:

$$R_C = z_{1-\alpha/2}^2\left[\frac{2\hat{\lambda} + (4\hat{\lambda}^2 + W^2)^{1/2}}{W^2}\right]. \tag{11}$$

As an example, to obtain confidence intervals with width no greater than 0.10 for an observed clutter false alarm rate of $\hat{\lambda}_0 = 0.5$ per m^2, it follows from (11) that the clutter area must cover at least 771 square meters.

Of course, during the test planning phase the investigator will not know the observed FAR. Thus, the choice of an appropriate value of $\hat{\lambda}_0$ for use in (11) is not clear. One option is to select a value that is believed to be a reasonable upper bound. Because (11) increases with $\hat{\lambda}_0$, this approach will give a conservative estimate of the required clutter area.

4 - COMPARING TWO POISSON PROPORTIONS

In practical applications of mine detection technology, false alarms can be costly in dollars, time, and operational success. Thus, the development of new sensors and processing methods with reduced false alarm rates is a goal of much ongoing research. Determining when one sensor has significantly out-performed another is an important step in the analysis of data from a multi-system demonstration or test. In this section, a statistical hypothesis test for comparing the performances of two different sensors is discussed. The same technique may also be used to assess the performance of a single system under different experimental conditions (e.g., dry sandy soil versus wet clay), or in different replications of the same system over the same clutter region.

4.1 - An Hypothesis Test for Two Poisson Rates

Suppose that system A covers a clutter area of R_{C_A} during test, with k_A clutter false alarms observed. Let λ_A represent the true FAR underlying system A. The quantities R_{C_B}, k_B, and λ_B are defined similarly for system B.

The FAR performance comparison is based on the question: "Is the FAR demonstrated by system A significantly different from the FAR demonstrated by system B?" In statistical terms, this question is phrased as an hypothesis test, with null and alternative hypotheses given by:

$$H_0 : \lambda_A = \lambda_B \tag{12a}$$

$$H_1 : \lambda_A \neq \lambda_B . \tag{12b}$$

One method for choosing between H_0 and H_1 conditions on the total number of false alarms ($k_A + k_B$) and examines the percentage of this total that is due to each system [Hald, 1952; Lampton, 1994]. Let π_A represent the probability that any one false alarm is due to system A. Conditional on the total number of false alarms, the experimental data may be viewed as a series of $k_A + k_B$ binomial trials, with two possible outcomes (system A or system B) at each trial. The probability of system A is equal to π_A at every trial.

Under the null hypothesis (12a) the clutter false alarm rates underlying the two systems are the same, and π_A should therefore be equal to:

$$\pi_{A_0} = \frac{R_{C_A}}{R_{C_A} + R_{C_B}}, \tag{13}$$

which is simply the percentage of the total clutter area that was covered by system A. (If $R_{C_A} = R_{C_B}$, it follows that $\pi_{A_0} = 1/2$).

Under this framework, the hypotheses given in (12) are equivalent to:

$$H_0 : \pi_A = \pi_{A_0} \tag{14a}$$

$$H_1 : \pi_A \neq \pi_{A_0} . \tag{14b}$$

Methods for testing hypotheses about a single binomial proportion are discussed in Simonson [1998]. If both $\pi_A(k_A + k_B)$ and $(1 - \pi_A)(k_A + k_B)$ exceed five, the test statistic:

$$s = \frac{\left| \frac{k_A}{k_A + k_B} - \pi_{A_0} \right| - \frac{1}{2(k_A + k_B)}}{\sqrt{\frac{\pi_{A_0}(1 - \pi_{A_0})}{k_A + k_B}}} \tag{15}$$

may be used. When the null hypothesis (14a) is true, s has a distribution that is approximately standard normal [Fleiss, 1981]. If H_1 is true, (15) will tend to be large. An α-level test rejects H_0 when s exceeds the quantile of the standard normal distribution corresponding to probability $1 - \alpha/2$.

As an example of the method, suppose that system A covered a clutter area of $R_{C_A} = 250$ m^2 and gave $k_A = 118$ false alarms. Suppose further that system B gave $k_B = 72$ false alarms in a clutter area of $R_{C_B} = 200$ m^2. If the two systems have the same underlying false alarm rate, we would expect that the fraction of false alarms due to system A would be approximately $\pi_{A_0} = 250/450 = 0.556$. The actual fraction due to A was $118/190 = 0.621$. From (15), the test statistic has value $s = 1.744$. To test at level 0.05, s is compared to $z_{0.975} = 1.960$. Because $s < 1.960$, the null hypothesis is not rejected: the difference in rates between the two systems is not significant at the 5% level.

In applying the method outlined in this section to experimental data, it is important to be aware of the assumptions used to derive the statistic (15). Specifying $\pi_{A_0} = R_{C_A}/(R_{C_A} + R_{C_B})$ implies that the two systems were tested under equivalent conditions. If system A were tested in a more challenging environment than system B, (15) would not represent a fair test of the relative performances of the two systems. Ideally, the two systems should cover the same clutter region, so that R_{C_A} equals R_{C_B}. In many experimental situations, this is not possible. However, every effort should be made to ensure that the conditions encountered by the two systems are comparable.

5 - SUMMARY PERFORMANCE MEASURES: PD, PFA, AND FAR

As discussed in this report and the companion document [Simonson, 1998], the common metrics used to characterize the performance of detection systems are PD (the probability of detection), PFA (the probability of a false alarm), and FAR (the false alarm rate). All three are of vital interest, and they must be considered together in evaluating detection systems. A system that achieves a high PD but is subject to frequent false alarms may have little value in time-critical applications. Conversely, a system that rarely makes false detections but misses a substantial proportion of real mines is unlikely to gain acceptance among users. In this section, some methods are presented for graphically summarizing system performance for targets, decoys, and clutter.

Figure 4 is an example of a display style that can be used to show detection rates, along with confidence intervals, for different types of known objects. The viewer can see at a glance that the hypothetical sensor in question had little difficulty in detecting large and small metal mines, but frequently missed plastic mines. While it rarely false alarmed on the wooden decoys, it misclassified bolts as targets about 60% of the time. Due to the relatively small number of decoy objects utilized, the 95% confidence intervals for PFA on the decoys were considerably wider than intervals characterizing PD for mine targets. Variations of this general type of plot can be used to display the performance of multiple sensors on the same test data, or the performance a single sensor under varying experimental conditions. In producing such a plot, it is implicitly assumed that the same detection threshold was used across all target and decoy types.

For many detection systems, the primary output at any given point in space is not a binary (mine/no mine) decision, but rather a continuous one-dimensional variable that is thresholded to determine whether or not a detection has occurred. This variable may simply be the magnitude of a received physical or chemical signal, or it may represent a measure of the similarity between the received signal and a known signal that is characteristic of targets. For such data, it is common to display sensor performance in terms of PD and FAR, as a function of

a varying threshold. Displays of this type are called receiver-operator characteristic (ROC) curves [Andrews, George, and Altshuler, 1997; Poor, 1988].

Assume that the variable in question tends to be large in the presence of the target. When a threshold level is chosen, detection is said to occur at each location in the test region giving a value at or above this level. As the threshold is reduced, the number of detections increases. ROC curves are constructed by varying the threshold, and plotting observed false alarm rate (per unit area) versus observed PD. The points representing PD/FAR pairs are then connected to give a smooth-looking curve. Figure 5 is an example summarizing the performance of a hypothetical sensor for small metallic mines under two different conditions: dry sandy soil, and wet sandy soil. For each condition, the PD corresponding to any given FAR can be read from the appropriate curve. For example, with a FAR of one per 100 m^2, the sensor achieves a PD of 100% for dry soil and about 78% for wet soil. Multiple curves can be used to graphically illustrate performance differences observed across various sensor types, target types, and experimental conditions.

6 - DISCUSSION

This report and its companion [Simonson, 1998] outline a variety of statistical techniques pertinent to mine detection problems. Methods are introduced for the estimation of PD, PFA, and FAR. Hypothesis tests are developed for assessing performance differences between two different sensors, or between two different environments for the same sensor. Basic formulas for calculating sample sizes and clutter areas are included. In selecting the material to be presented, the goal has been to choose a few straightforward and broadly applicable techniques - not to provide an exhaustive catalog and review of statistical methodology. Of necessity, some relevant subjects have been neglected. In this section, two particularly important (and closely related) topics are briefly addressed, with references provided to more complete accounts. These topics are the design of multi-factor experiments and the analysis of variance (ANOVA) for multiple comparisons.

When planning experiments to test the capabilities of different mine detection systems, the designer is frequently interested in determining whether (and how much) a variety of different experimental factors impact sensor performance. Controlled factors may include target type and size, soil type and moisture content, and burial depth. Due to cost considerations, it may not be feasible to conduct a large number of replications of every possible combination of experimental factors. However, by careful selection of the combinations to be tested, the experimenter can ensure good estimates of the effects that are deemed the most important. The general field of statistical experimental design is concerned with the planning of efficient experiments to provide the desired information in a readily extractable form. Two standard references in this area are the texts by Box, Hunter, and Hunter (1978) and Cox (1958).

The first step in analyzing data from multi-factor experiments often involves using the analysis of variance to determine which factors (and combinations of factors) significantly impact performance. While ANOVA is a familiar and widely used technique, it is based on several assumptions that may be questionable for mine detection data. In particular, the response variable is assumed to be normally distributed. When the actual response represents a percentage of mines detected, the exact distribution is binomial and the normal approximation may be inadequate if the observed percentage is close to zero or one, or if the number of target mines emplaced is small. In such cases, it is necessary to transform the observed percentages to a new domain in which the normality assumption is more nearly met. The book by Box, Hunter, and Hunter (1978) provides an excellent introduction to the mathematics of ANOVA. The collection edited by Hoaglin, Mosteller, and Tukey (1991) discusses a number of more advanced topics, including transformation and graphical display.

It is hoped that the material covered in the present report and its companion, along with the referenced statistical literature, will provide some useful guidance to the mine detection community in the areas of experimental planning and statistical data analysis.

7 - REFERENCES

Andrews, A.M., George, V., and Altshuler, T.W. (1997). Quantifying performance of mine detectors with fewer than 10,000 targets. *SPIE Proceedings*, vol. 3079, 273-280.

Bickel, P.J., and Doksum, K.A. (1977). Mathematical Statistics. Oakland, CA: Holden-Day.

Box, G.E.P., Hunter, W.G., and Hunter, J.S. (1978). Statistics for Experimenters. New York: John Wiley & Sons.

Çinlar, E. (1975). Introduction to Stochastic Processes. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Cox, D.R. (1958). Planning of Experiments. New York: John Wiley & Sons.

Cox, D.R., and Hinkley, D.V. (1974). Theoretical Statistics. London: Chapman and Hall.

Fleiss, J.L. (1981). Statistical Methods for Rates and Proportions, Second Edition. New York: John Wiley & Sons.

Hald, A. (1952). Statistical Theory with Engineering Applications. New York: John Wiley & Sons.

Hoaglin, D.C., Mosteller, F., and Tukey, J.W., eds. (1991). Fundamentals of Exploratory Analysis of Variance. New York: John Wiley & Sons.

Johnson, N.L., Kotz, S., and Kemp, A.W. (1992). Univariate Discrete Distributions, Second Edition. New York: John Wiley & Sons.

Koopmans, L.H. (1987). Introduction to Contemporary Statistical Methods. Boston: Duxbury Press.

Lampton, M. (1994). Two sample discrimination of Poisson means. *The Astrophysical Journal*, vol. 436 (#2/pt. 1), 784-786.

Larsen, R.J., and Marx, M.L. (1981). An Introduction to Mathematical Statistics and Its Applications. Englewood Cliffs, NJ: Prentice-Hall.

Poor, H.V. (1988). An Introduction to Signal Detection and Estimation. New York: Springer-Verlag.

Ripley, B.D. (1981). Spatial Statistics. New York: John Wiley & Sons.

Silvey, S.D. (1975). Statistical Inference. London: Chapman and Hall.

Simonson, K.M. (1998). Statistical considerations in designing tests of mine detection systems: I - Measures related to the probability of detection, SAND98-1769/1. Sandia National Laboratories, Albuquerque, NM, August 1998.

Taylor, H.M., and Karlin, S. (1994). An Introduction to Stochastic Modeling. San Diego: Academic Press.

SCHEMATIC REPRESENTATION: LAYOUT FOR ESTIMATING PD, PFA, AND FAR

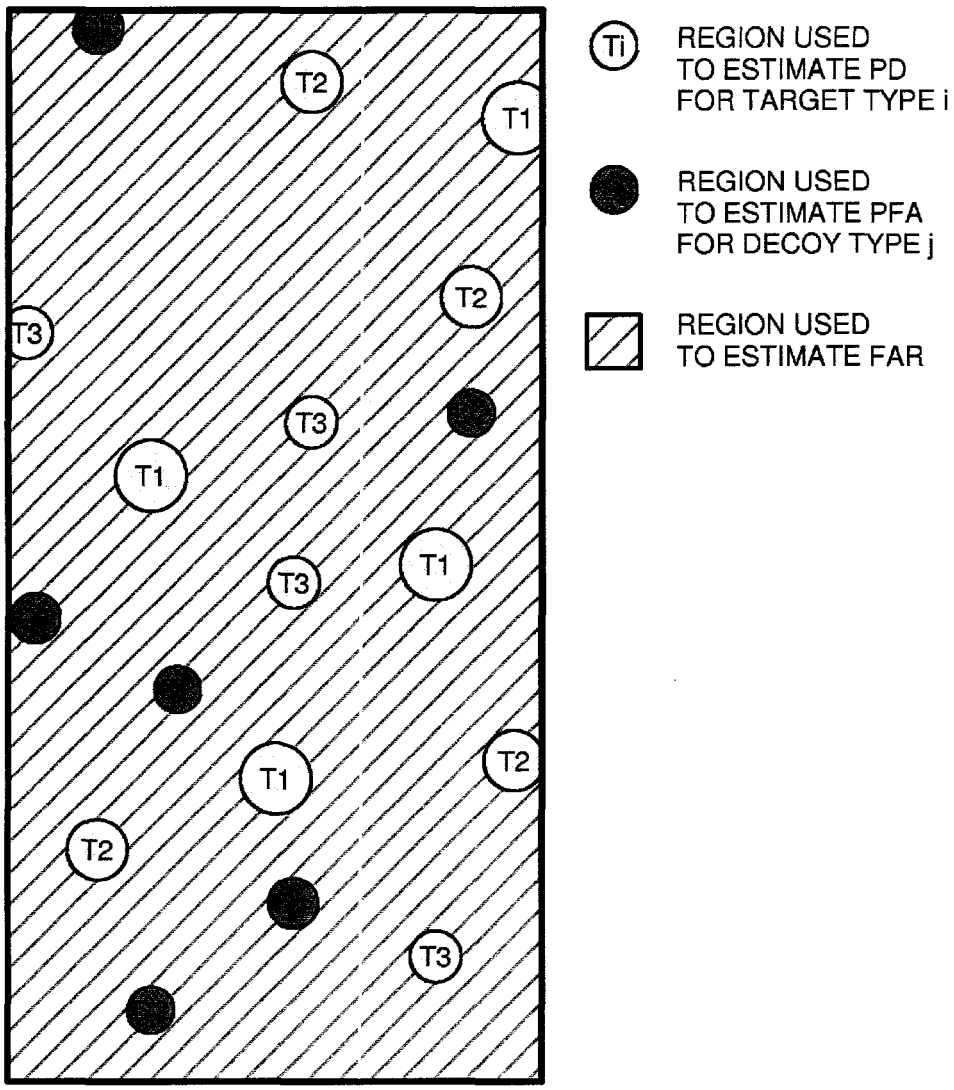

Figure 1 - Schematic diagram of an experimental layout. This experiment is designed to provide estimates of PD for three different mine target types, PFA for two different decoy types, and FAR over a clutter area believed to be homogeneous. Non-overlapping detection regions are used to compute the different estimates. Detections occurring within the dark circles are classified as target hits, while those occurring within the light circles are classified as decoy hits. Detections occurring within the striped region are designated as clutter false alarms.

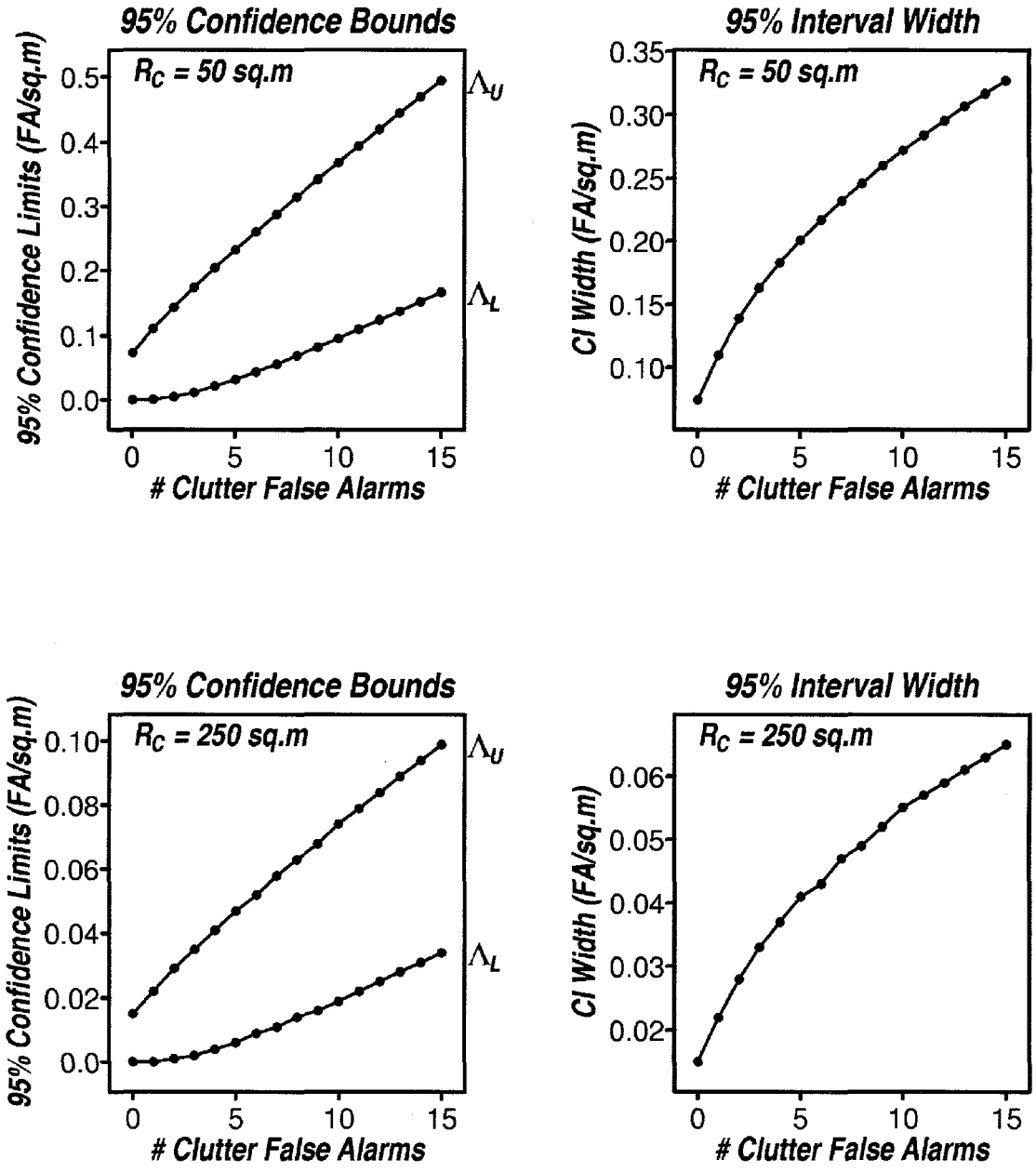

Figure 2 – Small sample 95% confidence intervals for a single Poison parameter. Bounds and widths are shown for clutter areas of $R_C = 50$ m^2 and $R_C = 250$ m^2. All of the values plotted were computed using the small sample method of equations (4) and (5). For a fixed number of false alarms, uncertainty decreases as area increases.

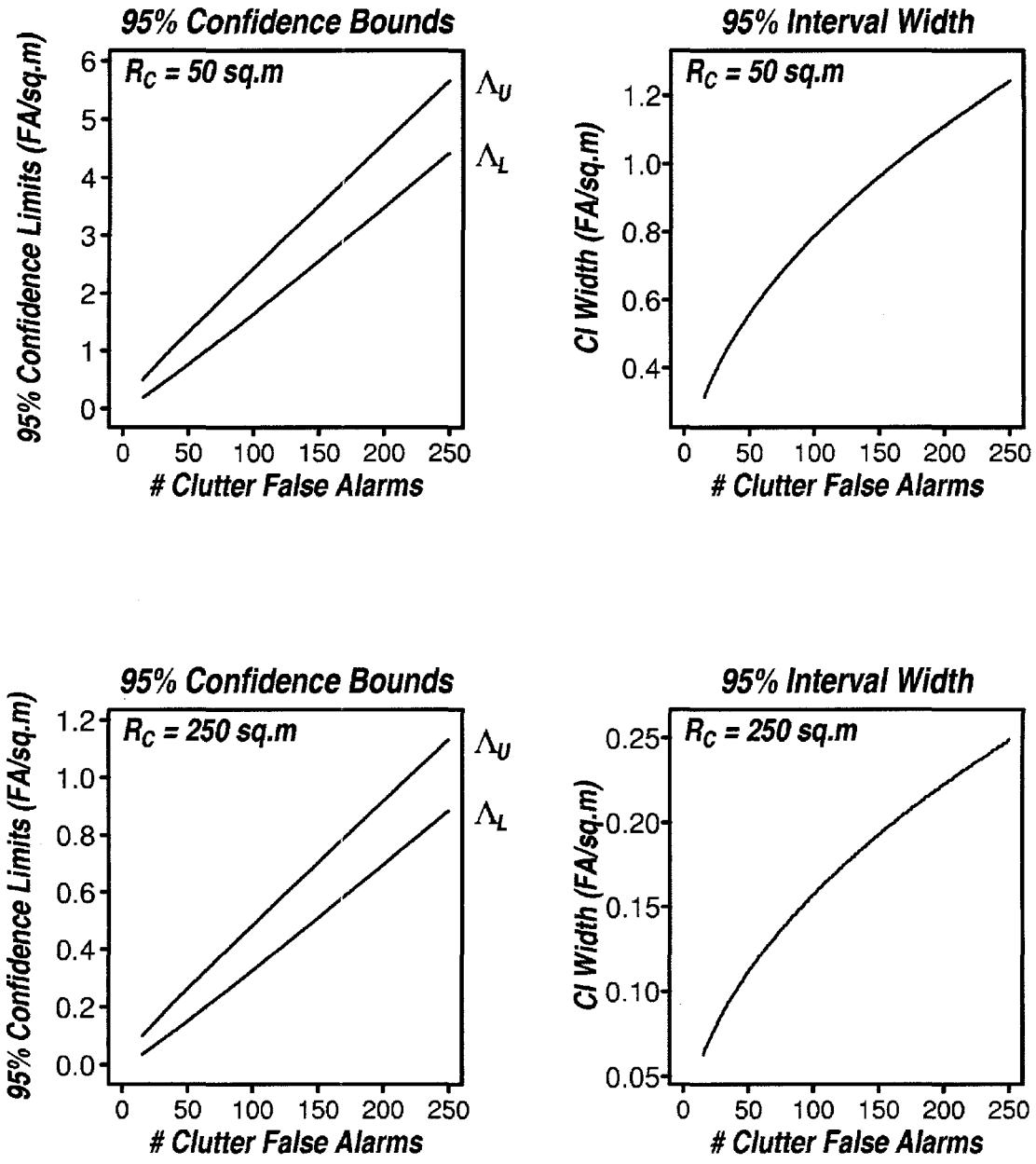

Figure 3 - Large sample 95% confidence intervals for a single Poisson parameter. Bounds and widths are shown for clutter areas of $R_C = 50$ m^2 and $R_C = 250$ m^2. All of the values plotted were computed using the large sample method of equations (6) and (7).

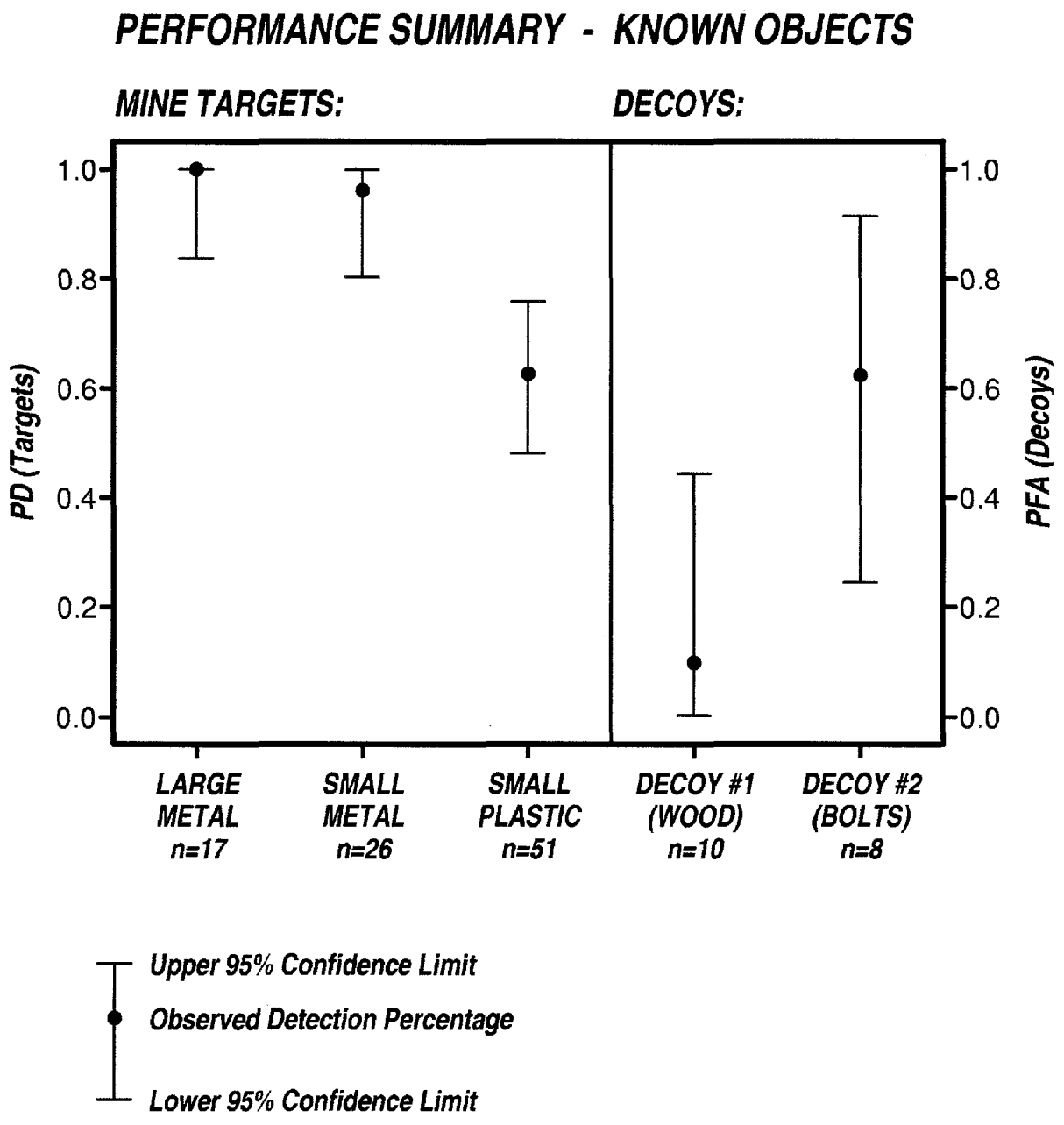

Figure 4 - Sample graphical summary. Results are shown for a hypothetical sensor tested against three different types of mine targets and two types of decoys. For each target or decoy type, the observed detection percentage is shown as a solid dot and the corresponding uncertainty is represented in terms of a 95% confidence interval. To provide the viewer with some information about the experimental design, the number of times each object type was encountered is also listed.

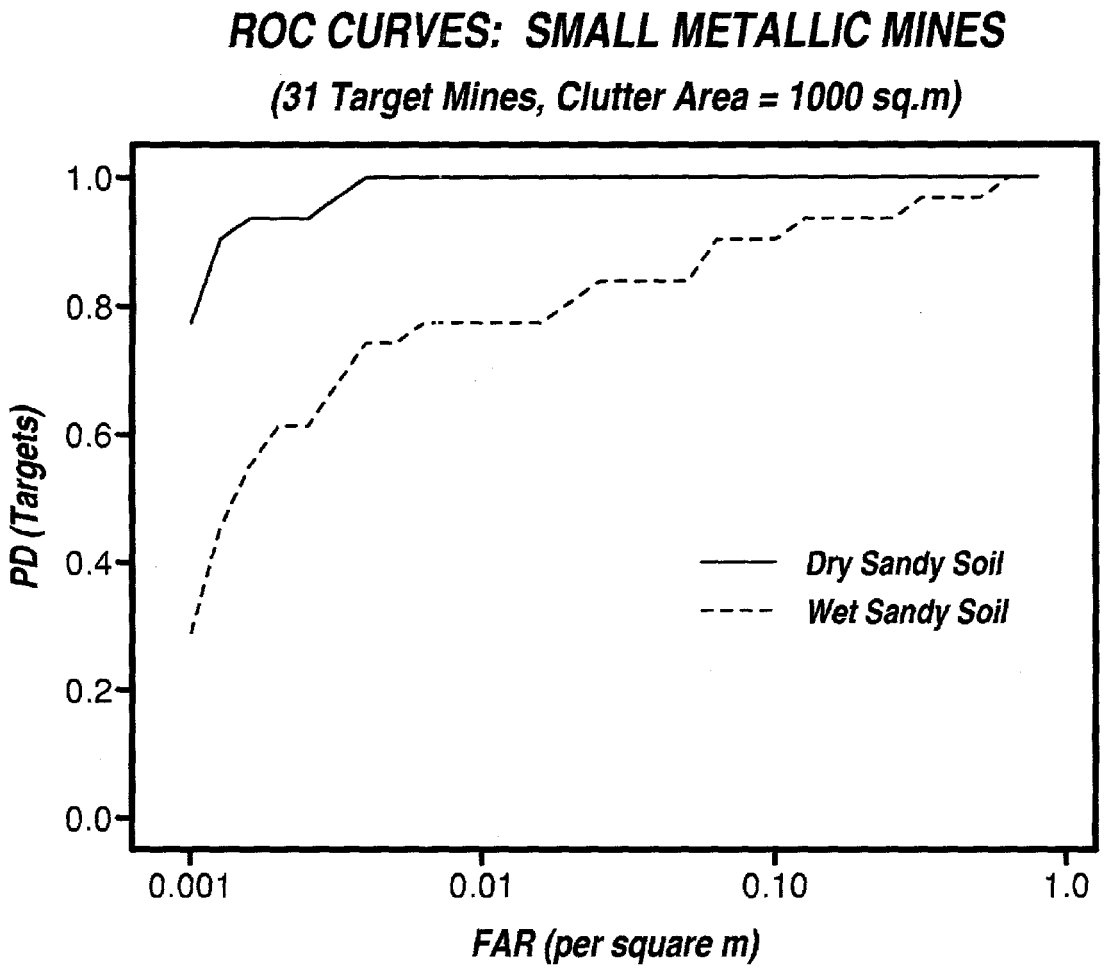

Figure 5 - Sample ROC curve. The curve illustrates the performance of a hypothetical sensor against a specific target type (small metallic mines) under two different environments. For each test condition, 31 target mines were encountered, and a clutter area of 1000 square meters was covered. For a clutter area of this size, false alarm rates below $0.001/m^2$ cannot be estimated, so the curves are left truncated at this point.