INSTITUTE FOR DEFENSE ANALYSES

Scoring Underwater Demonstrations for Detection and Classification of Unexploded Ordnance (UXO)

Shelley M. Cazares

November 2020

IDA Document D-19436

Log: H 20-000452

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

The Institute for Defense Analyses is a nonprofit corporation that operates three Federally Funded Research and Development Centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical, and analytic expertise.

About This Publication

This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-14-D-0001, Project AM-2-1528, "Assessment of Traditional and Emerging Approaches to the Detection and Classification of Surface and Buried Unexploded Ordnance (UXO)," for the Director, Environmental Security Technology Certification Program (ESTCP) and Strategic Environmental Research and Development Program (SERDP), under the Deputy Assistant Secretary of Defense (Environment). The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

For More Information

Shelley M. Cazares, Project Leader
scazares@ida.org, 703-845-6792

Leonard J. Buckley, Director, Science and Technology Division
lbuckley@ida.org, 703-578-2800

Copyright Notice

INSTITUTE FOR DEFENSE ANALYSES

IDA Document D-19436

Scoring Underwater Demonstrations for Detection and Classification of Unexploded Ordnance (UXO)

Shelley M. Cazares

Executive Summary

The Strategic Environmental Research and Development Program (SERDP) and Environmental Security Technology Certification Program (ESTCP) are sponsoring the development of novel systems and processes for the detection and classification of unexploded ordnance (UXO) in underwater environments. SERDP is also sponsoring underwater testbeds to demonstrate the performance of these novel systems and processes. Scoring these demonstrations is a complicated process. The Institute for Defense Analyses designed and implemented the scoring process for SERDP's and ESTCP's previous terrestrial demonstrations in the 2000s and 2010s. In some cases, the lessons learned from the terrestrial demonstrations can be leveraged in the underwater demonstrations. In other cases, new solutions must be found, due to the added logistical, engineering, and safety challenges of the underwater environment. This document was written for underwater testbed managers to document frequently asked questions regarding the scoring of underwater demonstrations for UXO detection and classification:

A. Why do we need to score underwater demonstrations?

B. What kind of scores do we need to calculate?

C. What is the difference between detection and classification?

D. How can we tell if the system "missed" a target of interest (TOI)?

E. How can we tell if the system has a false alarm?

F. What is the detection halo and how do we set its radius?

G. What is ground truth, where do we get it, and what should we do with it?

H. Can any ground-truth information be released to the demonstrators before data collection?

I. Why do we need to emplace true TOI objects?

J. Do we also need to emplace true non-TOI objects?

K. In what pattern should we emplace the objects?

L. If demonstrators do not get a perfect score, should we allow them to explain why?

M. If a system performs well at the test site, should we assume it will also do well at other sites?

N. What if environmental conditions change throughout the test area or during the course of a demonstration?

O. Can we combine the scores from different systems to predict how an ideal system may perform?

P. In summary, what does SERDP or its representatives need to provide to the scoring team?

Q. In summary, what does the scoring team need to provide back to SERDP?

Contents

1. Introduction

This document is about novel systems and processes for detecting and classifying unexploded ordnance (UXO) in underwater environments. The Strategic Environmental Research and Development Program (SERDP) and the Environmental Security Technology Certification Program (ESTCP) are sponsoring the development of these novel systems and processes. SERDP is also funding the design and management of underwater testbeds to demonstrate the performance of these novel systems and processes.

Scoring these underwater demonstrations is a complicated process. The Institute for Defense Analyses designed and implemented the scoring process for SERDP's and ESTCP's previous terrestrial demonstrations during the 2000s and 2010s (Cazares, Ayers, and Tuley 2018). In many cases, the lessons learned from the terrestrial demonstrations can now be leveraged in the underwater demonstrations. In other cases, new solutions must be found, due to the added logistical, engineering, and safety challenges of the underwater environment.

This document was written for underwater testbed managers to document frequently asked questions (FAQs) regarding the scoring of underwater demonstrations. These FAQs summarize the main considerations at a high level—they are a mile wide and an inch deep. Subsequent documents may discuss specific issues at a more detailed level, in particular those issues introduced in the footnotes.

2. Frequently Asked Questions

A. Why do we need to score underwater demonstrations?

First, it's good research practice.

Second, it helps secure funding for the program as a whole. U.S. taxpayers and lawmakers would like to understand the return on SERDP's and ESTCP's investments in novel underwater UXO remediation systems. One way to easily communicate the return on investment is to document the systems' performance in relevant underwater environments. Quantitative performance metrics (scores) can quickly summarize how far the novel systems have come—that is, their return on investment. These scores can also clarify which research areas require *additional* investment so that the systems can reach technical maturity. Scoring is needed to calculate these metrics.

B. What kind of scores do we need to calculate?

Scores must quantify how well a system can detect and classify targets of interest (TOIs) from non-TOIs:

- A TOI is a catchall term that includes UXO that already contaminates the test site, as well as inert or surrogate munitions that are purposely emplaced at the test site for the demonstration.
- A non-TOI is any other object, such as rocks, crab pots, fragments of previously exploded munitions, and so forth.

Two types of scores are needed to describe the two types of errors in Figure 1:

1. A metric related to false positives (FPs). An FP is a false alarm. Throughout this document, the terms *false positive* (FP) and *false alarm* are used interchangeably. Many different metrics can be used to summarize FPs, including the false-alarm rate (*FAR*). These metrics describe how often the system creates false alarms.
2. A metric related to false negatives (FNs). An FN is a "missed" TOI—a TOI that the system did not find. Throughout this document, the terms *false negative* (FN) and *missed TOI* are used interchangeably. FNs are often summarized by the metric probability of detection (P_d). This metric describes how well the system finds TOIs (i.e., avoids missing TOIs) (Cazares, Ayers, and Tuley 2018).

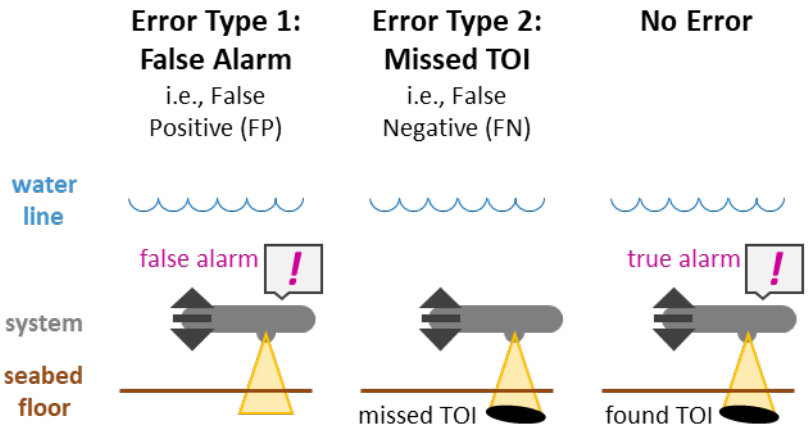

Figure 1. Two Types of Errors in UXO Remediation

Generally, FNs and FPs trade off of each other—as one count goes down, the other can go up. As do P_d and *FAR*—as one metric gets better, the other can get worse. That is why both types of scores must be calculated and reported *together*. Often, the two scores are plotted against each other, as the classification threshold is varied. For example, P_d and *FAR* are often plotted against each other to form a free-response receiver-operating characteristic (ROC) curve. Figure 2 shows an example ROC curve from a previous terrestrial demonstration. ROC curves can quickly convey a large amount of information about a system's detection and classification performance. The vertical axis (P_d) represents how well the system can find TOIs (i.e., how well it can avoid missing TOIs), an indication of how *safe* the system is. In contrast, the horizontal axis (*FAR*) represents how many false alarms are caused by the system, an indication of how many unnecessary costs are created by the system. The large blue dot represents the final TOI-versus-non-TOI classification threshold selected by the demonstrator. The shape of the ROC curve illustrates how safety and cost would trade off of each other if other classification thresholds were selected instead. Cazares, Ayers, and Tuley (2018) describe the ROC curves generated for the terrestrial demonstrations in detail.

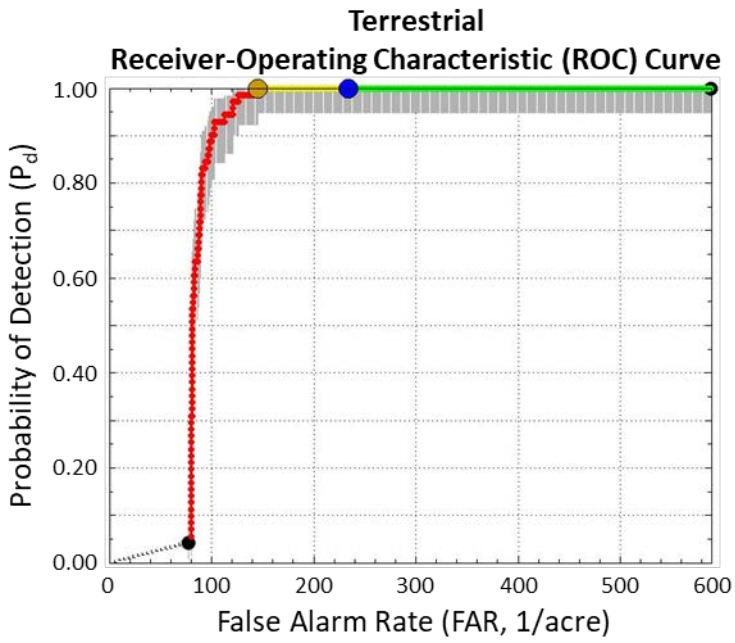

Figure 2. Free-Response Receiver-Operating Characteristic Curve from a Previous Terrestrial Demonstration

C. What is the difference between detection and classification?

UXO remediation involves two main data processing steps:

1. *Detection:* The system analyzes the data collected across the entire test site and determines that one or more objects are likely to be present. The system outputs an estimate of the likely location of each detected object.

   The demonstrator submits a detection list for scoring, consisting of the position coordinates of each detected object. To ease scoring, the demonstrator should provide position coordinates in flat-plane easting and northing coordinates in Universal Transverse Mercator (UTM) units.[1]

2. *Classification:* The system reanalyzes the data collected around each detected object and determines if that object is likely to be a TOI or non-TOI. In a real remediation project:

   a. TOIs must be excavated or closely monitored.

   b. Non-TOIs do not need to be excavated or closely monitored.

   The demonstrator submits a ranked detection list for scoring, consisting of the same detected objects as on the original detection list. Now, however, detected

[1] Fiducial monuments emplaced on the seabed floor before beginning the demonstration may be helpful in estimating the detected objects' locations as accurately and precisely as possible.

objects on the ranked detection list must be ordered according to their likelihood of being a TOI. The first detected object is the one most likely to be a TOI; the last object is the one most likely to be a non-TOI.

For most systems, the detection step must always come before the classification step—a system cannot classify an object until it has detected it.[2]

D. How can we tell if the system missed a TOI?

There are two different ways in which a system can miss a TOI:

1. *TOI-Miss Detection Error*: The system fails to detect that an object is present even when a true TOI is actually there. To determine if a TOI-miss detection error has occurred, the scoring team must know the true locations of all TOI objects present at the site (● in the notional bird's-eye view map of Figure 3), as well as the estimated locations of all objects detected by the system (× in Figure 3). The scoring team must also decide in advance how close an estimated location must be to a true location to declare that the system "found" the TOI. This is done by drawing a circle or *detection halo* around each true location (--- in Figure 3). Then:

   a. If no estimated locations are in or on the detection halo (such as the left-most TOI in Figure 3), then the scoring team should conclude that the system "missed" the TOI. All TOI-miss detection errors must be counted as FNs and included in the P_d metric.

   b. On the other hand, if at least one estimated location is in or on the detection halo (such as the other three TOIs in Figure 3), then the scoring team must conclude that the system found the TOI. Note, however, that at this point in the process, the system does not yet know if the detected object is a TOI or non-TOI. Therefore, the system must then pass the detected object onto the next step of the process (classification).

[2] Some systems may perform detection and classification simultaneously. To date, such systems are not common.

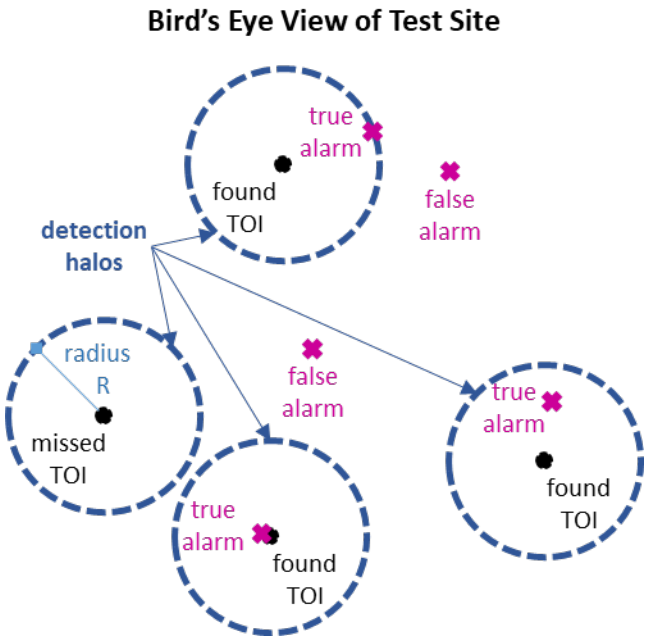

Figure 3. Bird's-Eye View of a Test Site (notional), Showing Two False Alarms, One Missed TOI (left), and Three Found TOIs, Based on a Detection Halo with Radius *R*

2. *TOI-Miss Classification Error:* The system classifies a detected object as non-TOI even though it is actually a TOI. To determine if this is the case, the scoring team must know the true and estimated types of the detected object, that is, whether the objects are truly TOI or non-TOI and whether the detected objects were classified as TOI or non-TOI. All TOI miss classification errors must also be counted as FNs and included in the P_d metric.

In summary, both types of TOI misses—TOI-miss detection errors and TOI-miss classification errors—must be included in the FN count and the P_d metric. At the same time, however, each type of FN must also be tallied separately to explain to stakeholders why the system missed a TOI—either it didn't detect the TOI at all or it detected the TOI but misclassified it.[3]

E. How can we tell if the system has a false alarm?

A system can have a false alarm in two different ways:

[3] For some demonstrations, it may be useful to calculate two metrics relating to missed TOIs: (1) the probability of detection (P_d) to summarize the TOI-miss detection errors and, separately, (2) the probability of correct classification (P_c) to summarize the TOI-miss classification errors. We could then quantify the overall performance of the system as $P_d \times P_c$. There are pros and cons to this scoring approach.

1. *False-Alarm Detection Error:* The system detects that an object is likely to be present even though no object is actually there. To determine if a false-alarm detection error has occurred, the scoring team must:

    a. First, compare the true locations of all TOI objects to the estimated locations of all detected objects (i.e., the system alarms), as described in the previous question. That is, the scoring team must determine if each true TOI object was "found" or "missed" based on whether or not at least one system alarm was in or on its detection halo, as illustrated in the bird's-eye view plot of Figure 3.

    b. Then, all remaining alarms that are not in or on any detection halo must be considered as potential false alarms, such as the two false alarms in Figure 3.

    Note, though, that all detected objects, even the incorrectly detected ones, are then passed on to the next step of the process (classification), where the system has a second chance to correctly classify them as likely non-TOI. Therefore, false alarm detection errors should not be counted as FPs or included in the *FAR* summary metric.

2. *False-Alarm Classification Error:* The system classifies a detected object as TOI even though it is actually non-TOI (or not present at all). To determine if this is the case, the scoring team must know the detected object's true type (i.e., true TOI vs. true non-TOI/not present) and the system's estimate of the detected object's type (i.e., TOI vs. non-TOI). All false-alarm classification errors should be counted as FPs and included in the *FAR* summary metric.[4]

In summary, only one type of false alarm must be included in the FP count and the *FAR* metric—the false-alarm classification errors. This is due to the fact that, after the detection step, *all* detected objects, including the incorrectly detected ones, are passed on to the classification step, where the system gets a second chance to correctly classify them as non-TOI.[5] However, it can still be useful to informally tally both types of FPs, to indicate

---

[4] This sentence is still subject to debate in the underwater UXO remediation community. In terrestrial UXO remediation projects, geolocation errors are small (on the order of centimeters), and so the remediation team can, and therefore must, reacquire and investigate each and every alarm individually. Multiple false alarms in a small area are therefore nuisances that eat up time and resources—and these individual nuisances should be counted and added up. In underwater projects, however, geolocation errors are much larger (on the order of meters), so multiple false alarms in a small area may be able to be condensed into one, in the minds of the remediation team, while the team investigates the area as a whole. Therefore, one could argue that the condensed false alarms should only be counted as one. There are pros and cons to both scoring approaches.

[5] That is, during the classification step, the system gets a second chance to correct the mistakes made in the detection step. In fact, from a philosophical point of view, there were not "mistakes" at all. They were not "false" positives, per se, because the system never claimed they were likely TOI, only that they were worthy of further analysis—which is the entire point of the classification step.

how well the classification step can reduce the potential false alarms initially introduced by the detection step.[6]

F. What is the detection halo and how do we set its radius?

The *detection halo* is a concept used to identify missed TOIs, found TOIs, and false alarms, as discussed in the previous two questions. The scoring team must define the radius of the detection halo, labeled *R* in Figure 3. This radius *R* sets the maximum acceptable distance between an object's true (●) and estimated (×) location during detection scoring.[7]

In an ideal world, the detection halo radius *R* would be based solely upon the requirements of the subsequent steps of the UXO remediation process (e.g., reacquisition for retrieval/neutralization). For example, consider a case in which the remediation team has instructed its divers to swim to a particular coordinate location and then search for a buried UXO within a certain clearance radius (e.g., 1 m). Then the detection halo radius *R* should be based on this clearance radius (e.g., 1 m).

In the real world, though, the detection halo radius must also be based on two other concepts: the geolocation error and the sensor resolution:

- *Geolocation Error* refers to the uncertainty in the survey instrumentation used to measure location—both for (1) the ground truth, when the testbed crew's survey instrumentation measures the true locations of the objects, and also for (2) the detection step, when the system's survey instrumentation measures the estimated locations of the detected objects. It is impossible for the scoring team to determine if the system can detect an object any more accurately than the overall geolocation error. The scoring team must consider both sources of geolocation

---

6 For some demonstrations, it may be useful to calculate the ratio between the number of false-alarm classification errors and the number of false-alarm detection errors. In informal scoring exercises, this ratio is often colloquially called the probability of false alarms (P_{fa}). While this ratio is different in nature to the traditional P_{fa} metric defined in the statistics literature, it can still be an informative metric. There are pros and cons to this P_{fa}-like metric.

7 Early terrestrial demonstrations in the 2000s at standardized sites like Aberdeen Proving Ground used oblong, rather than circular, detection halos to tightly circumscribe the shape of the cylindrical munitions. This was due to a concern that the geolocation errors on land (on the order of centimeters) were small with respect to the largest dimension of the objects (on the order of tens of centimeters) and therefore real-world remediation teams would tire of systems that produce many false alarms within a few centimeters of each other. As it turned out, oblong detection halos were difficult to implement and did not provide much benefit to scoring, given the ability of the systems to avoid false alarms in the first place. Therefore, oblong detection halos were later abandoned and replaced with circular detection halos for the terrestrial live-site demonstrations in the 2010s. For underwater demonstrations, oblong detection halos are not expected to provide much benefit, since the geolocation errors underwater (on the order of meters) are large with respect to the largest dimension of the objects (on the order of tens of centimeters). Therefore, circular detection halos will be used to score underwater demonstrations for the foreseeable future.

error (ground truth and detections) when setting the appropriate value for the detection halo radius *R*.[8]

- *Sensor Resolution* refers to the ability of the sensor to tell multiple closely spaced objects apart. The sensor can only tell two objects apart if they are spaced farther than the sensor resolution. Therefore, the detection halo radius *R* should also be based on the sensor resolution.[9]

Note that each subsequent step of the UXO remediation process may have a different requirement (e.g., clearance radius). Also, each system may have a different type of survey instrumentation with a different geolocation error and a different type of sensor with a different sensor resolution. Therefore, the scoring team must consider all reasonable values of clearance radius, geolocation error, and sensor resolution for all systems, when determining the appropriate value for the detection halo radius *R*. This is likely to be the trickiest part of scoring underwater UXO demonstrations.[10] *System developers should begin communicating their expected geolocation error and sensor resolution to SERDP as early as possible, so that appropriate testbeds and scoring protocols for their systems can begin to be developed.*

G. What is ground truth, where do we get it, and what should we do with it?

Ground truth is true information about the TOI and non-TOI objects that are actually present in the demonstration area of the testbed. SERDP or its representatives must compile the ground truth and provide it to the scoring team. To preserve the scientific integrity of

---

8 Geolocation error is of more concern to underwater than terrestrial demonstrations. *Terrestrial* demonstrations can use real time kinematic (RTK) Global Positioning System (GPS) technology to provide *centimeter*-level accuracy in location estimates. Therefore, geolocation error is not a large consideration in terrestrial demonstrations. In contrast, underwater demonstrations must use alternative surveying technology, for which geolocation error is much larger, likely on the order of meters. Thus, geolocation error becomes a much larger consideration for underwater demonstrations.

9 Sensor resolution is of more concern to underwater than terrestrial UXO remediation projects. Terrestrial projects make use of electromagnetic induction sensors for which a data-processing technique called "multi-source dipole inversions" can resolve multiple, closely spaced objects—therefore, sensor resolution is no longer a large consideration for terrestrial demonstrations. In contrast, some underwater remediation projects are likely to use acoustic sensors, for which sensor resolution may be a larger issue.

10 Several methods can potentially be used to help set the detection halo radius *R*. Some methods can be employed during the data collection part of the demonstration, such as having the demonstrators collect data over known objects several times throughout the day, detect those objects, estimate their locations, and compare the estimated locations to ground truth. Other methods can be employed during the scoring part of the demonstration, such as defining a standardized set of detection halo radii, ranging from the tightest stakeholder requirement (e.g., 1 m for diver reacquisition) to the loosest consideration (e.g., 3 m for geolocation error) and then repeating the scoring several times, once with each radii in the standardized set, to explore how sensitive the performance metrics P_d and *FAR* are to detection halo radius *R*.

the demonstration, the ground truth must be closely held and hidden from the demonstrators throughout the demonstration.

There are two main pieces of ground truth information, used to generate the scores:

- *The true location of each object* (i.e., true Easting and Northing coordinates), used to score the detection step of the demonstration. *To ease scoring, the testbed manager should provide position coordinates in flat-plane Easting and Northing coordinates in UTM units.*[11]
- *The true type of each object* (i.e., true TOI vs. true non-TOI), used to score the classification step of the demonstration.

In addition, there are at least three other pieces of ground truth information that can help stakeholders interpret the scores (i.e., understand why some objects are more or less difficult to detect and classify):

- The true burial depth of each object.
- The true orientation of each object (i.e., azimuth and inclination; roll is irrelevant).
- Specific characteristics of each object, including:
  - For TOI objects, the caliber and type of the munition[12] (e.g., 37 mm projectile with rotating band, 4.2″ mortar, M48 fuze, 2.36″ rocket motor, small industry standard object, and so forth), as well as the overall condition of the object (i.e., degree of corrosion and/or biofouling).
  - For non-TOI objects, a quantitative description of the object (e.g., 3″ × 8″ flat piece of munitions frag, 12″ diameter steel crab pot, and so forth), as well as the overall condition of the object (i.e., degree of corrosion and/or biofouling).

SERDP or its representatives should compile this ground-truth information into a single spreadsheet and deliver it to the scoring team. Compiling this information can be a challenge, especially in dangerous and logistically constrained environments such as underwater sites. Therefore, not all the information listed above may be available. Furthermore, not all this information will remain static over the course of the demonstration (e.g., nearby boats may drag anchors that disturb the positions, depths, and orientations of the objects, and natural processes may lead to increased biofouling or corrosion of the objects over time). In such cases, the scoring team must make assumptions about what the missing or changing ground truth may be (e.g., "All objects that were not deliberately

[11] Fiducial monuments emplaced on the seabed floor before beginning the demonstration can be helpful in estimating the objects' locations as accurately and precisely as possible.

[12] If possible, the specific make and model of the munition also should be provided.

emplaced in the testbed are assumed to be true non-TOI" and "all locations and orientations are assumed to be static over time"). *The scoring team should clearly document all assumptions about missing ground truth and clearly communicate them to all stakeholders as part of the scoring report.*

H. Can any ground-truth information be released to the demonstrators before data collection?

The answer depends on whether the test makes use of an instrument verification strip (IVS) or a blind site:

- An IVS is located to the side of the test area. All ground-truth information about the objects in the IVS should be freely shared with the demonstrators, including the objects' true locations, types, burial depths, 3D orientations, and other specific characteristics discussed above in the previous question (e.g., munition caliber and type, condition, etc.). SERDP or its representatives must determine how to set up the IVS such that these objects are representative of those emplaced in the demonstration area. The demonstrators may periodically collect data over the IVS throughout the demonstration to optimize or calibrate their systems or check that their systems are operating correctly. IVSs are often called calibration strips or calibration lanes. Demonstrators can detect and classify the objects in the IVS to produce detection lists and ranked detection lists. These lists can be scored, either formally or informally (self-scored). Tests like these are often called engineering tests.[13]

- In contrast, a blind site constitutes the main test area. Most ground-truth information about the objects in the blind site, such as the number of objects and their true locations, types, burial depths, 3D orientations, and specific characteristics discussed in the previous question, should be withheld from the demonstrators. Demonstrators collect data over the blind site and then process these data without any knowledge of ground truth to form their official detection list and ranked detection list, which are then formally scored. Tests like these are called blind tests.

  However, despite the "blind" nature of a blind test, limited ground-truth information can and should be released to the demonstrators, including:

[13] Some testbeds may choose to make use of a geophysical prove-out (GPO) instead of, or in addition to, the IVS. While an IVS usually consists of a small number of objects (e.g., around 10) emplaced at regular intervals along a straight line that has been previously cleared of objects, a GPO usually consists of a larger number of objects (e.g., around 30) emplaced in a more realistic distribution (e.g., random) without first clearing the area of any objects. That is, a GPO is often a smaller replica of the test area itself. IVSs and GPOs each have pros and cons.

- Information that would be obtainable through a remedial investigation/feasibility study in a real remediation project, such as the expected types of TOIs or ranges of their expected burial depths.
- Environmental conditions throughout the test area, since such information would be available to remediation team during a real remediation project. This information includes bathymetry; current velocity; wave height; water salinity/conductivity, temperature, and turbidity; sediment types and layering; vegetation and inhabiting biota; seabed floor and surrounding topography; local ships' schedules; and rates of biofouling/corrosion on known object types. The demonstrators may use this information to optimize or calibrate their systems for particular environmental conditions. Since environmental conditions may change over time, this information must be time- and date-stamped.

The scoring team should document which ground truth has been released to the demonstrators, and why, and communicate this to all stakeholders as part of the scoring report.

After all scoring is completed, *all* blind site ground truth should be provided to the demonstrators, so that they can use it to perform failure analyses on their missed TOIs and recommend corrective actions.

I. Why do we need to emplace true TOI objects?

UXO is rare.[14] Therefore, any given site is likely to have little to no UXO. Additional TOI objects must be emplaced at the testbed site to provide the systems with enough opportunity to detect and classify them as part of the demonstration.

For example, if the scoring report states "P_d = 0.80," how confident can the scoring team be that the system can detect and correctly classify TOIs 80% of the time? The confidence depends on the total number of TOIs:

- If the testbed contains only five TOIs, and the system correctly detected and classified four of them, then the scoring team can estimate that the system's P_d = 4/5 = 0.80. However, this is only an estimate of P_d. Theoretically, if the demonstration were repeated several times, the system may find a different subset of TOIs each time, potentially resulting in a different value for P_d each

[14] Exceptions to this rule involve facilities adjacent to ammunition plants, at which munitions were once routinely fired for test purposes. However, locations like that are unlikely to be chosen as SERDP or demonstration sites in the next few years, due to the high density of clutter that would make detection and classification extremely difficult, if not impossible. Even state-of-the-art technologies and processes for UXO detection and classification are not likely to work well on locations like these, nor are they designed to.

time. Based on the results of a single demonstration, however, the scoring team can use statistical methods (i.e., the binomial distribution) to state that there is a 95% probability that the true P_d value lies anywhere between 0.28 and 0.99. This range is called the *95% confidence interval* around P_d. The small number of TOIs at the site (five) leads to a very wide confidence interval (0.28 – 0.99), indicating little confidence in the P_d estimate of 0.80.

- In contrast, if the testbed contains 500 TOIs, and the system correctly detected and classified 400 of them, then the scoring team can estimate that P_d = 400/500 = 0.80, the same value as before. However, this is a better estimate of P_d, since it is based on a larger number of TOIs. The scoring team can state that there is a 95% probability that the true P_d lies within a much tighter range, between 0.76 and 0.83. The larger number of TOIs (500) leads to a tighter 95% confidence interval (0.76 – 0.83), indicating more confidence in the P_d estimate of 0.80.

As a rule, the more TOIs that are emplaced at the site, the more confident the scoring team can be when estimating the P_d metric. To estimate and interpret the P_d metric, several pieces of ground-truth information must be recorded for each emplaced TOI, as discussed in question G. SERDP must tell the testbed managers how many TOIs, and what types of TOIs, should be emplaced at the site. But to be a truly blind test, the demonstrators themselves should not be told how many TOIs have been emplaced at the site.

J. Do we also need to emplace true non-TOI objects?

The answer depends on the testbed itself:

- Often, there are already a very large number of non-TOI objects present in the demonstration area of a testbed site that are similar to the types of non-TOIs that would likely be present in a real remediation project—rocks, crab pots, fragments of previously exploded munitions, and so forth. In that case, no additional non-TOI objects need to be emplaced.
- However, it may be possible that the demonstration area of a testbed site was cleared of all debris before the demonstration setup (like the IVS area) or is otherwise lacking the types of non-TOI objects that would likely be present in a real remediation project. In that case, representative non-TOI objects need to be emplaced to provide the system with enough opportunity to detect and classify them as non-TOI during the demonstration.

Therefore, for each demonstration, SERDP or its representatives must decide if some non-TOI must be emplaced. Such decisions must consider the type of system participating in the demonstration. If the system is based on magnetometry, then some truly magnetic non-TOI objects should be emplaced to intentionally stress the system. Similarly, if the system is based on electromagnetic induction, then some truly metallic (but not magnetic)

non-TOI objects should be emplaced, for the same reason—to stress the system. Finally, if the system is based on acoustic sensors, then some truly non-magnetic, non-TOI objects should be emplaced to stress the acoustic systems. Unfortunately, if the demonstration budget is limited, the more resources devoted to emplacing non-TOIs, the fewer resources available for emplacing TOIs. Therefore, trade-offs will have to be made in the testbed design.

K. In what pattern should we emplace the objects?

It depends:

- For the IVS, demonstrators may choose a pattern to ease their calibration methods (i.e., a straight line).
- For the blind site, SERDP or its representatives must choose the pattern, and this pattern should be withheld from the demonstrator. If the objects are emplaced in too regular a pattern, then the demonstrators may (consciously or unconsciously) learn this pattern over the course of the demonstration, thus biasing their detection performance. Therefore, objects should be emplaced in as realistic a pattern as possible in the blind site. For many types of munitions, a random distribution will simulate a realistic pattern. For other munition types, the demonstration may want to use a different pattern, such as a Gaussian distribution around a central target point.

For both the IVS and the blind site, the objects should *not* be emplaced too close together. If objects are emplaced too close together, then the sensor may not be able to tell them apart, due to the limitations associated with geolocation error and sensor resolution (both of which are discussed in question F).[15] Each system participating in the demonstration may have a different type of survey instrumentation with a different geolocation error and a different type of sensor with a different sensor resolution, as discussed in question F. Therefore, when determining how far apart objects should be emplaced from each other, SERDP and its representatives must consider all reasonable values for geolocation error and sensor resolution for all systems participating in the demonstration.

L. If demonstrators do not get a perfect score, should we allow them to explain why?

Yes. In fact, the demonstrators must be *required* to do so. After all the scoring is complete, all ground truth should be provided to the demonstrators for this purpose. Each demonstrator must perform a failure analysis for each missed TOI (i.e., each FN) and

[15] Some demonstrations may choose to emplace some non-TOI objects very close to a TOI object (even closer than the sensor resolution may be capable of resolving) to intentionally stress the system.

suggest a corrective action to avoid such an error in the future. This failure analysis/corrective action should be presented at the next in-progress review and discussed in a collegial group setting with all other demonstrators. These types of discussions were helpful in pushing forward the state of the art during ESTCP's previous terrestrial demonstrations.

Furthermore, if the corrective action concerns changes to the data processing, then the demonstrators should be encouraged (and possibly even required) to reprocess their data and try again. The scoring team must then be willing and properly resourced to re-score the demonstrators' second try to determine whether the corrective action was truly corrective (i.e., it eliminated the missed TOIs without adding too many more false alarms).

Finally, the demonstrators may be required to document good performance on an IVS (or another area with known ground truth, such as a GPO) before being allowed to participate in a blind test. This rule would avoid the unnecessarily expending of testbed resources on a demonstrator who is not yet ready for a blind test.

M. If a system performs well at the test site, should we assume it will also do well at other sites?

Not necessarily. We cannot assume that the system will perform similarly at other sites with different TOIs, different non-TOIs, or different environmental conditions. To help stakeholders understand when they can and cannot extrapolate demonstration results to other sites, SERDP or its representatives must quantitatively characterize each test site at the time of the demonstration—the size of the demonstration area; the number, types, and characteristics of TOIs and non-TOIs in the demonstration area; the environmental conditions in the demonstration area; and so forth. SERDP must communicate this information to the scoring team, and the scoring team must include it in its scoring report for context.

In the future, stakeholders in real UXO remediation sites can compare the demonstration area to their site—the more similar their site is to the demonstration area, the more confidence they can have in extrapolating the system's demonstration performance to their site. This process could be helpful in obtaining the buy-in of environmental regulators in real UXO remediation projects, since the regulators will have a quantitative way of gauging how suited the system is for a particular project.

N. What if environmental conditions change throughout the test area or during the course of a demonstration?

When scoring the blind test, the FNs (missed TOIs) and FPs (false alarms) must be counted over stable conditions. Similarly, the P_d and *FAR* metrics must be calculated over stable conditions as well. Therefore, if the environmental conditions change over space and

time, then the demonstration should be segmented into different sub-demonstrations, each of which has stable conditions:

- For example, if the sediment type changes rapidly from one part of the blind site to another (i.e., sandy vs. silty), then one set of FN and FP counts (and, similarly, one set of P_d and *FAR* metrics) should be calculated over the sandy part of the blind site, and another set of FN and FP counts (and P_d and *FAR* metrics) should be calculated over the silty part. This type of segmentation was often done in ESTCP's previous terrestrial demonstrations, such as when the former Camp Beale and Camp Spencer were segmented into "open" and "tree" areas.
- In another example, if the wave and current conditions change over time (e.g., from calm to stormy weather), then one set of counts and metrics should be calculated over the calm days of the blind test and another set of counts and metrics should be calculated over the stormy days.

Before data processing begins, SERDP should inform the demonstrators if the blind site is to be segmented in space or time and, if so, how. The demonstrators should then process the data from each sub-demonstration separately, creating a separate detection list and ranked detection list for each sub-demonstration.[16]

O. Can we combine scores from different systems to predict how an ideal system may perform?

Maybe, but with caution. The ideal system will likely be based on multiple sensor types; current systems may use only one sensor type. The scoring team should consult with SERDP and its representatives to consider what statistical methods would be best suited for estimating what the performance of an ideal system *could* be, based on the scores of multiple current systems. Such an estimate would allow SERDP to better project the value (or lack thereof) in funding the development of the ideal system. Once built, though, the ideal system itself must then be tested in a demonstration.

P. In summary, what does SERDP or its representatives need to provide to the scoring team?

SERDP or its representatives must provide the following:

[16] Some of the ground-truth information may change, as well (e.g., position or orientation may be changed by storms, sweeping boat anchors, etc.). When available, this information should be clearly documented for the scoring team, such that it can aid the partitioning of the demonstration by space and time.

- Quantitative characteristics of the blind site, including environmental conditions over space and time, and how the blind test may (or may not) be segmented by space or time into multiple sub-demonstrations.
- Definition of scoring metrics (e.g., FN, FP, P_d, *FAR*).
- Detection halo radius, (e.g., *R*).
- Ground truth (with appropriate time and date stamps).
- Position coordinates outlining the blind site demonstration area and each sub-demonstration area (if any).
- Ranked detection lists submitted by demonstrators (with appropriate time and date stamps to aid with sub-demonstration segmentation). Each ranked detection list must:
  - Contain the estimated locations of *all* objects detected by the system, including those that the system classified as TOI *and* non-TOI.
  - Be ordered by increasing likelihood of being a TOI. That is, the first detected object on the list should be the object classified as most likely to be a TOI, and the last detected object should be the one most likely to be a non-TOI.

Q. In summary, what does the scoring team need to provide back to SERDP?

The scoring team must return the following:

- A reference to the quantitative characteristics of the blind site that were originally provided by SERDP.
- A reference to the definition of scoring metrics (e.g., FN, FP, P_d, *FAR*) that were originally provided by SERDP.
- A reference to the detection halo radius (e.g., *R*) that was originally provided by SERDP.
- A reference to the position coordinates outlining the blind site demonstration area and each sub-demonstration area (if any) that were original provided by SERDP.
- Assumptions about missing ground-truth information (if any).
- For each ranked detection list:

- List of all FPs (false alarms) and associated information (e.g., target ID number, estimated location and type, seed ID number of closest true TOI object, and so forth) for each possible classification threshold.
- List of all FNs (missed TOIs) and associated information (e.g., seed ID number, true location and type, target ID number of closest detected TOI object, and so forth) for each possible classification threshold.
- Estimated P_d metric and 95% confidence interval for each possible classification threshold.
- Estimated *FAR* metric for each possible classification threshold.
- Free-response ROC curve plotting P_d vs. *FAR* as the classification threshold is retrospectively varied over all possible values.

3. Conclusions

Scoring underwater demonstrations is a complicated process. Several considerations should be kept in mind:

- To ease scoring, position coordinates should be provided in flat-plane Easting and Northing coordinates in UTM units on both the ground-truth list (provided by the testbed manager) and the ranked detection list (provided by the demonstrator).
- System developers should begin communicating their expected geolocation error and sensor resolution to SERDP as early as possible, so that appropriate testbeds and scoring protocols for their systems can begin to be developed.
- The scoring team should clearly document all assumptions about missing ground truth and clearly communicate them to all stakeholders as part of the scoring report.
- The scoring team should document which ground truth has been released to the demonstrators, and why, and communicate this to all stakeholders as part of the scoring report.

This document was intended to be a high-level description of the main considerations of underwater scoring. Many details are beyond the scope of this document and may be explored in subsequent related documents, including:

- The proper selection, emplacement, and surveying of fiducial monuments to ease the surveying of position coordinates.
- The proper methods to measure ground truth (position, depth, azimuth, and inclination of each object).
- The definition of the metrics (P_d, P_c, FAR, P_{fa}, etc.) and ROC curves used to quantify the performance of a system for detection and/or classification.
- The proper selection of the detection halo radius R.
- Whether or not multiple alarms in a detection halo should be counted as multiple false alarms.
- The pros and cons of IVSs versus GPOs.

Reference

Cazares, Shelley M., Elizabeth L. Ayers, and Michael T. Tuley. 2018. "ESTCP UXO Live Site Demonstrations 2007 to 2017." IDA Document, D-9193. Alexandria, VA: Institute for Defense Analyses.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED *(From–To)*
November 2020	Final	

4. TITLE AND SUBTITLE

Scoring Underwater Demonstrations for Detection and Classification of Unexploded Ordnance (UXO)

5a. CONTRACT NUMBER
HQ0034-14-D-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Cazares, Shelley M.

5d. PROJECT NUMBER
AM-2-1528

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311-1882

8. PERFORMING ORGANIZATION REPORT NUMBER

IDA Document D-19436

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

SERDP/ESTCP
4800 Mark Center Drive, Suite 16F16
Alexandria, VA 22350-3605

10. SPONSOR/MONITOR'S ACRONYM(S)

SERDP/ESTCP

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited (24 November 2020).

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The Strategic Environmental Research and Development Program (SERDP) and the Environmental Security Technology Certification Program (ESTCP) are sponsoring the development of novel systems and processes for the detection and classification of unexploded ordnance (UXO) in underwater environments. SERDP is also sponsoring underwater testbeds to demonstrate the performance of these novel systems and processes. Scoring these demonstrations is a complicated process. The Institute for Defense Analyses designed and implemented the scoring process for previous terrestrial demonstrations in the 2000s and 2010s for SERDP and ESTCP. In some cases, the lessons learned from the terrestrial demonstrations can be leveraged in the underwater demonstrations. In other cases, new solutions must be found, due to the added logistical, engineering, and safety challenges of the underwater environment. This document, which was written for underwater testbed managers, documents frequently asked questions regarding the scoring of underwater demonstrations for UXO detection and classification.

15. SUBJECT TERMS

acoustic color; demonstration test; electromagnetic induction (EMI); False Alarm Rate (FAR); Probability of Detection (Pd); Probability of False Alarm (Pfa); Receiver Operating Characteristic (ROC) curve; scoring; testbed; underwater; unexploded ordnance (UXO)

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
					Bradley, David
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
Uncl.	Uncl.	Uncl.	SAR	27	571-372-6388

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18