

# Trial Design for Testing and Evaluation of Metal Detectors Used in Humanitarian Landmine Clearance

Mate GAAL, *Bundesanstalt für Materialforschung und –prüfung (BAM), Berlin, Germany*

Christina MUELLER, *BAM, Berlin, Germany*

Uwe EWERT, *BAM, Berlin, Germany*

Peter-T. WILRICH, *Freie Universität, Berlin, Germany*

Wolfgang SPYRA, *Brandenburgische Technische Universität Cottbus, Germany*

**Abstract.** In spite of many efforts directed to new technologies, the metal detector is still the main detection tool used in humanitarian landmine clearance. The most detailed proposal for a standard for testing metal detectors is the CEN Workshop Agreement 14747:2003. Since 2003 a series of tests has been executed with the purpose of optimisation and more detailed specification of the testing procedures. The main purpose of the test procedures is the identification of the most appropriate device for certain conditions of application.

The reliability of a detection system is influenced by the intrinsic physical capability of the method, the application factor and the human factor. The physical capabilities and some influences of application conditions can be evaluated in laboratory measurements, but the influence of the human factor and other environmental influences can be derived correctly from blind tests only, since they are conducted under field conditions.

This paper discusses the design of the experiment for testing and evaluation of metal detectors based on the experiences gained from several field tests performed from 2003 to 2005. These tests evaluated the latest models of four metal detector manufacturers. Both pulse induction and continuous wave detectors were tested. Also the shapes of their coils are different: most of the detectors use a single coil, but some use the “double-D” configuration. Some of the detectors are static mode detectors, some dynamic mode ones. They all have some data processing to compensate the soil background effects.

The influences of soil, target type and its depth, as well as the human operator influence are evaluated with the application of an appropriate design of experiment. A full factorial design was applied to the maximum detection height measurements and a fractional factorial design to the blind tests, also called reliability tests. The results of both tests are discussed in this paper. The performance indicators of the blind tests are the probability of detection (POD) and the false alarm rate (FAR). The methods for the design of experiment and data evaluation for multiple parameter influences are applicable also in other areas of non-destructive testing.

## 1. Test and Evaluation of Metal Detectors in Humanitarian Demining

The metal detector is the main detection tool in humanitarian demining [1]. In recent years many efforts of the demining community were directed to testing and evaluation of demining equipment. A standard for testing and evaluation of metal detectors has been proposed in 2003. The CEN Workshop Agreement CWA 14747:2003 specifies standardised proce-

dures for measuring the influences of many factors to the performance of metal detectors [2]. The purpose of testing is to find the most suitable detector for a given set of conditions.

Many tests described in the CWA 14747:2003 are based on maximum detection height measurements. The maximum detection height is the distance between the search head of the metal detector and the top of the target at which the detector starts to give clear signals in an experiment in which the position of the target is known to the operator. This quantity provides the information about the depths at which the mines in a minefield can still be detected.

The blind in-field tests called detection reliability tests are the tests receiving the highest attention by detector end users, since they are performed in conditions as close as possible to the real minefield conditions. The operators are not familiar with the positions of the targets. They mark the assumed positions of the targets and the positions of their indications are compared with the actual target positions. If an indication falls inside a prescribed circle called halo, the indication is counted as a true positive. If it falls outside, it is counted as a false positive. The halo radius, according to CWA 14747:2003, is “the half of the maximum horizontal extent of the metal components in the target plus 100 mm”.

The two values to be estimated by the results of a detection reliability test are the probability of detection (POD) and the false alarm rate (FAR). The estimated probability of detection for a particular choice of targets, operators or detectors is the number of detected targets (true positives) divided by the total number of targets. The estimated false alarm rate is the number of false alarms (false positives) on a certain area divided by the size of that area. If we assume a binomial distribution for the number of detections, we can find 95% confidence limits for the probability of detection. Similarly, if we assume Poisson distribution for the false alarms, we can construct 95% confidence limits for the false alarm rate [3]. A ROC diagram is a diagram with the POD on the ordinate and the FAR on the abscissa, and it is a modification of a ROC diagram used in non-destructive testing, where the probability of detection is plotted against the probability of false alarms [5][6][7].

The other kind of diagram is the POD curve, presenting the dependence of the POD on a parameter, in our case the depth of the target. The POD curves and the corresponding regression model are described in detail in the final report of the metal detector trial performed in Benkovac, Croatia, 2005 [8].

The detection reliability tests owe their name to the concept of reliability, which is defined (CWA 14747:2003) as “the degree to which the metal detector is capable of achieving its purpose, which is to have maximum capability for giving true alarm indications without producing false alarm indications”. Detection reliability tests are the only tests that include the evaluation of the ability of metal detectors to deal with false alarms. In the context of testing, the detection of metal clutter is not considered as false alarm, since metal detectors are designed to detect metal. The main source of false alarms are the soils with frequency dependent magnetic susceptibility, causing alarms without presence of a metal fragment. Most metal detectors today have some ground compensating abilities, which means that they can decrease their sensitivity to soils with much smaller decrease of sensitivity to metal.

There are many factors influencing the performance of metal detectors. They are all described in a concept called “reliability formula”, first applied in non-destructive testing [5][6]. The total reliability (R) of a detection system is described by three factors: intrinsic capability (IC) describing the physics and the basic technical capabilities of the device, and representing an upper limit of R; factors of application (AP) such as specific environmental conditions in the field generally diminishing R and finally the human factor (HF), which also lowers R. The mutual interaction of these factors is usually very complex.

An important advantage of the detection reliability tests is that they include most of the factors influencing the performance of metal detectors. In the maximum detection

height measurements the influence of the operator (or the “human factor” from the reliability model) is much smaller, but it is not entirely excluded. The operator sets up the metal detector, performs the ground compensation, operates the device and decides whether the audio signal is a detection or noise.

## **2. Design of Experiment, Basic Principles**

The statistical design of experiments [3][4] (sometimes called experimental design) is the process of planning the experiment considering all influencing factors so that appropriate data will be collected, enabling objective conclusions. A scientific approach to planning an experiment results in unambiguous results, which are as little affected by experimental error as possible. The use of experimental design in industry can result in products with better performance, reliability, lower production costs or shorter development time.

In a designed experiment we make a difference between the predictor variables and the response variables describing a process. The predictor variables included in the experiment by controlling their values are called factors and the specific values that these factors can take in an experiment are called factor levels, or simply levels.

The basic principles of experimental design are replication, randomisation and blocking. Replication is the repetition of the experiment. It allows the experimenter to estimate the experimental error. Randomisation is crucial for each design of experiment. Both the allocation of the experimental material and the order of execution of measurements are determined randomly. As a result, errors are usually values of independently distributed random variables. Another important consequence of randomisation can be “averaging out” the effects of extraneous factors that might be present. A block is a portion of the experimental material that is more homogeneous than the entire set of material. Comparisons can be made within each block. This way the variability between blocks does not affect the experimental error.

## **3. Metal Detector Trial, Croatia, May 2005**

### *3.1 Overview*

The metal detector trial performed in Croatia in May 2005 is the last in the series of trials organised by BAM (German Federal Institute for Materials Research) [7][8]. The trial was conducted at the test site of the Croatian Mine Action Centre – Centre for Testing, Development and Training (HCR-CTRO) in Benkovačko Selo near Benkovac in Croatia. It was conducted according to the procedures prescribed in CWA 14747:2003, CEN Workshop Agreement on testing and evaluation of metal detectors for humanitarian demining [2], with the aim to verify the proposed testing procedures. Maximum detection height measurements and detection reliability tests were performed.

The metal detector models tested in the trials were products of the following companies, listed alphabetically: CEIA, Ebinger, Foerster and Vallon. It has been agreed with the manufacturers to keep the detector models anonymous, so that the models are labelled U, X, Y and Z. The detectors operated on different principles: some were time domain, some frequency domain ones; some used a single coil, some a “double-D” coil; some were static mode detectors, and some dynamic mode ones. All detectors had the possibility of ground compensation.

Two soil types were present in the trials in May 2005. There were four lanes with dimensions 1 m × 29 m, lanes 1 and 2 containing a highly magnetic soil from the surround-

ings of the town Obrovac, and lanes 3 and 4 containing a magnetically rather neutral soil from the area around the town Sisak.

Two mine types modified to be safe were used as targets: PMA-2 and PMA-1A. The PMA-2 is a minimum metal mine, the hardest to detect in south-eastern Europe. A surrogate of PMA-2 labelled PMA-S was used too, but only for the maximum detection height measurements. Each lane contained five mines PMA-2 buried just below the surface, five buried to 5 cm depth and five to 10 cm depth, measured to the top of the mine. The PMA-1A mines were buried similarly, but to depths 5, 10 and 15 cm.

The operators were experienced deminers of CROMAC (Croatian Mine Action Centre). They went through a one-day training for each detector model.

### *3.2 Maximum Detection Height Measurements, Design of Experiment*

In earlier investigations it has been conjectured that repeated maximum detection height measurements give very similar results. The experiment described in this article had been designed to check this conjecture of the stability of metal detectors.

The maximum detection height measurements were performed with a full factorial design. The investigated factors were: detector model, operator, target-soil combination. Two series of maximum detection height measurements were performed: one during the detection reliability test and the other afterwards. They both contain the same measurements; they were just performed in a different order.

The goals of this test were:

- To assess the variability of the maximum detection height measurements.
- To compare the detecting capabilities of four metal detectors in two soil types separately.
- To compare the surrogate of the PMA-2 with the real mine, using the in-air measurements.
- If the surrogate faithfully represents the real mine, to use it for comparing the two soils used in the experiment.

The first series of measurements was performed according to the design of the reliability test (see section 3.3) and it is described in the trial final report [8]. The measurements of the second series were performed according to the design presented in Table 1. PMA-S is a surrogate of the PMA-2. This is a full factorial design, meaning that all combinations of factor levels are present. The in-air measurements on the PMA-2 and the PMA-S were performed one after the other, in a random order. The execution order of starts was arranged to avoid bias, i.e. a systematic influence of unknown factors related to time (for example, gradual increase of the deminers' concentration or fatigue). If there was such an influence, it was "distributed" to all detector models equally. The operators set up the detector before each start and performed the ground compensation before the in-soil measurement. The targets were buried to depths up to 15 cm in steps of 1 cm and their positions were visibly marked.

### *3.3 Detection Reliability Tests, Design of Experiment*

The four factors investigated in this test are detector model, operator, lane (related to the soil type), and start. A factorial design including all factor level combinations would be the optimum choice to achieve an unbiased estimate of the detectors in each soil type separately. However, such a test would require a lot of time. This is why a fractional factorial design had been proposed: each detector is tested with each level of each factor, but not with all the possible combinations of factor levels. This design is based on a Graeco-Latin

square and it is shown in Table 2. In the first week two operators (A, B) tested two detectors (alpha, beta) and the other two operators the other two detectors. In the second week they switched (A and B tested gamma and delta). Such a design, compared with an ordinary Graeco-Latin square, allowed the operators to concentrate on two in instead of four detector models at a time.

**Table 1. Design of the maximum detection height measurements. The variable “start” indicates the order of execution. Numbers 1, 2, 3 indicate the random order of execution of the measurements within a start. The in-air measurements were executed before the in-soil measurements, but within the same start.**

start	operator	detector	in-air			in soil of lanes 1, 2	in soil of lanes 3, 4
			PMA-1A	PMA-2	PMA-S	PMA-2	PMA-S
1	C	beta	1	2	3	1	2
2	D	alpha	1	3	2	2	1
3	A	delta	1	2	3	2	1
4	B	gamma	3	2	1	1	2
5	A	alpha	1	3	2	1	2
6	B	beta	1	3	2	1	2
7	C	gamma	1	2	3	1	2
8	D	delta	1	3	2	1	2
9	A	gamma	3	2	1	2	1
10	B	delta	3	1	2	1	2
11	C	alpha	3	2	1	2	1
12	D	beta	1	3	2	2	1
13	C	delta	3	2	1	2	1
14	D	gamma	1	2	3	1	2
15	A	beta	1	2	3	1	2
16	B	alpha	3	2	1	1	2

**Table 2. Design of the detection reliability test. The design is based on a Graeco-Latin square, letters A, B, C, D representing the operators, and  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$  the detectors.**

Week 1  
18-20 May

		Start			
		1	2	3	4
Lane	1	A alpha	C delta	B beta	D gamma
	2	C gamma	A beta	D delta	B alpha
	3	B beta	D gamma	A alpha	C delta
	4	D delta	B alpha	C gamma	A beta

Week 2  
25-27 May

		Start			
		1	2	3	4
Lane	1	C alpha	A delta	D beta	B gamma
	2	A gamma	C beta	B delta	D alpha
	3	D beta	B gamma	C alpha	A delta
	4	B delta	D alpha	A gamma	C beta

## 4. Results of the Metal Detector Trial Performed in Croatia, May 2005

### 4.1 Results of the Maximum Detection Height Measurements

This section discusses some results of both series of maximum detection height measurements.

One of our goals was to compare the metal detectors in a specific soil. Assuming that the maximum detection heights for a specific detector-target-soil combination are independently normally distributed, we can perform some statistical tests to examine the differences between the detectors. Before doing that, it is usually helpful to present the results in a simple diagram. Figure 1 contains the results of both measurement series. The indicated error bars are the sample standard deviations. The results of a t-test with a significance level  $\alpha = 0.05$  for the soil from lanes 1 and 2 show a significant difference between all detectors except between detectors U and X and between Y and Z. In the soil from lanes 3 and 4 the only significant differences are between X and Y and between X and Z.

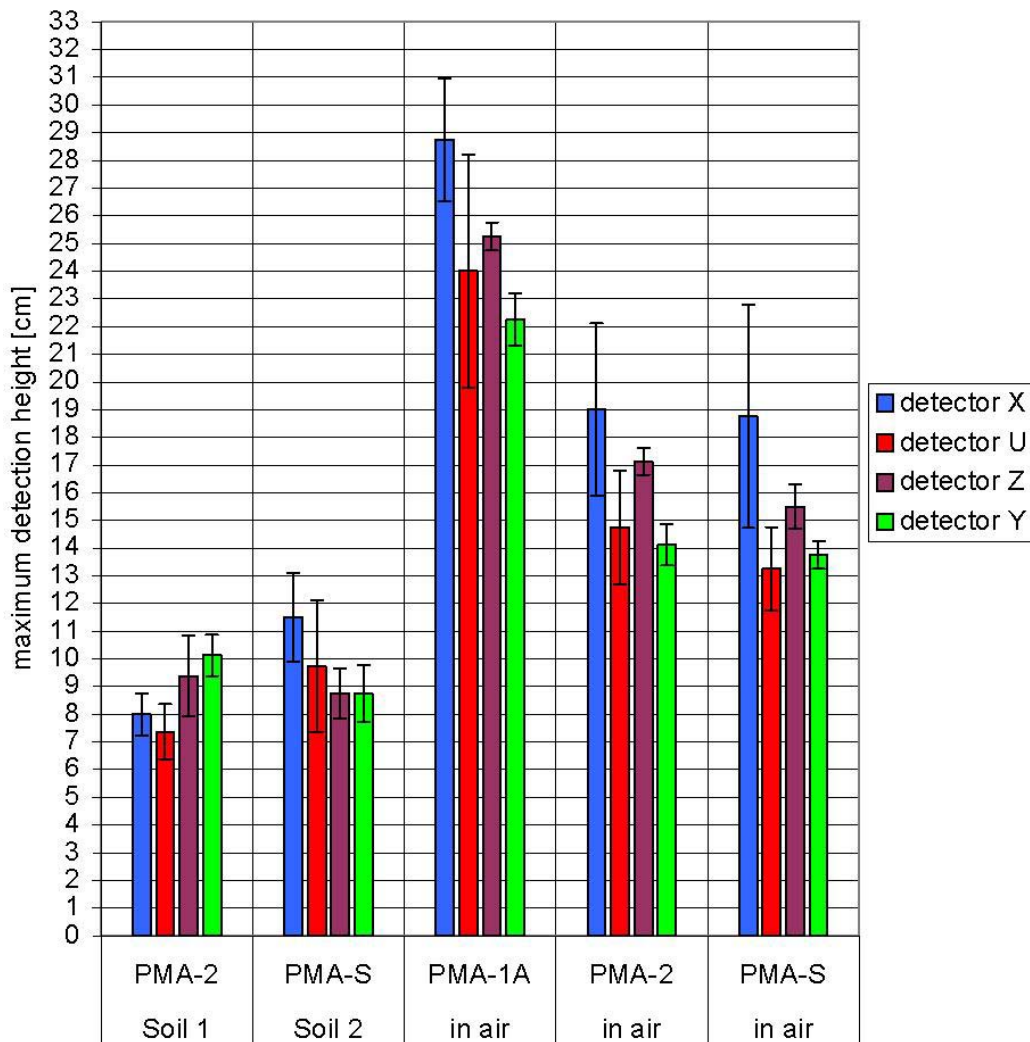


Figure 1. Results of the maximum detection height measurements. The error bars indicate the estimated standard deviations.

Let us compare the PMA-2 with its surrogate, PMA-S. The in-air measurements were analysed as paired measurements. This means that the difference between the measurements on these two targets within the same start (see Table 1) is analysed as a normally distributed variable. The result indicates that the maximum detection height of the real mine is  $(9.4 \pm 6.6)$  mm larger than that of the surrogate, where  $(9.4 \pm 6.6)$  mm marks the 95% confidence limits. This result indicates that in many cases the PMA-S can be used as a surrogate of the PMA-2, since it is just a bit more difficult to detect. However, the difference is not small enough to allow the comparison of the results in two soils as if the two targets were identical.

Many metal detector trials performed before 2005 included some maximum detection height measurements, but almost all of them were performed without repetitions. However, already a quick view on the diagram in Figure 1 reveals that the variance of the maximum detection height measurements cannot be neglected, thus pointing to the conclusion that repetitions are necessary. Further investigations have shown that the variability of the measurements is not caused only by the operators (deminers) nor by their personal differences, but mostly by the instability of the hardware of the devices.

#### *4.2. Results of the Detection Reliability Tests*

This section presents some of the results of the detection reliability tests performed near Benkovac in Croatia, in May 2005.

The following three diagrams are ROC diagrams based on the complete data set, i.e. all levels of all factors: both mine types, both soil types, all four detectors and all four operators. The first diagram, Figure 2, shows the differences between the four detector models tested in the trials. The difference between the soils is the subject of the next diagram, Figure 3. It can be seen that the false alarm rate increases in soils with stronger magnetic properties, that is, in lanes 1 and 2 (soil from Obrovac). It is likely that the probability of detection is reduced as a consequence of ground compensation. The diagram on Figure 4 clearly shows that all operators performed similarly: no significant differences were detected between the deminers' results. This is a consequence of the choice of skilled deminers, an adequate training, and the applied working procedures, which were close to the actual working conditions. The same was observed throughout the whole analysis, for any choice of targets or soils. The importance of operators involved in the trials has proven in all trials of demining equipment, but also in tests of NDT (non-destructive testing) detection systems [5][6]. It has been shown that the human factor (skill of the operators and the testing procedure) can significantly influence the test results. A comparison with earlier tests in Benkovac and Oberjettenberg [7][8] reveals a notable improvement of the test results, mostly due to human factor improvements.

The next diagram (Figure 5) presents the estimated POD curve (POD versus depth) with the corresponding confidence bounds for the selected case of PMA-2 in the soil from Obrovac area. The mine PMA-2 has the smallest metal content in the region of South-Eastern Europe. The soil from Obrovac is very difficult for metal detectors due to its magnetic properties, causing many false alarms, so that ground compensation is necessary. We can see from the diagram that detector Y, which has the highest total POD, can reliably detect the PMA-2 in the soil from Obrovac only if it is buried shallowly. Fortunately, the vast majority of mines is buried very near the surface.

Let us compare this result with the maximum detection height measurements with the same choice of detector, soil and target. The largest value being detected is  $(9.6 \pm 0.6)$  cm. The best estimate of the maximum detection height is that value increased by 0.5 cm, because the targets were buried to depths in steps of 1 cm. Thus we have the estimated

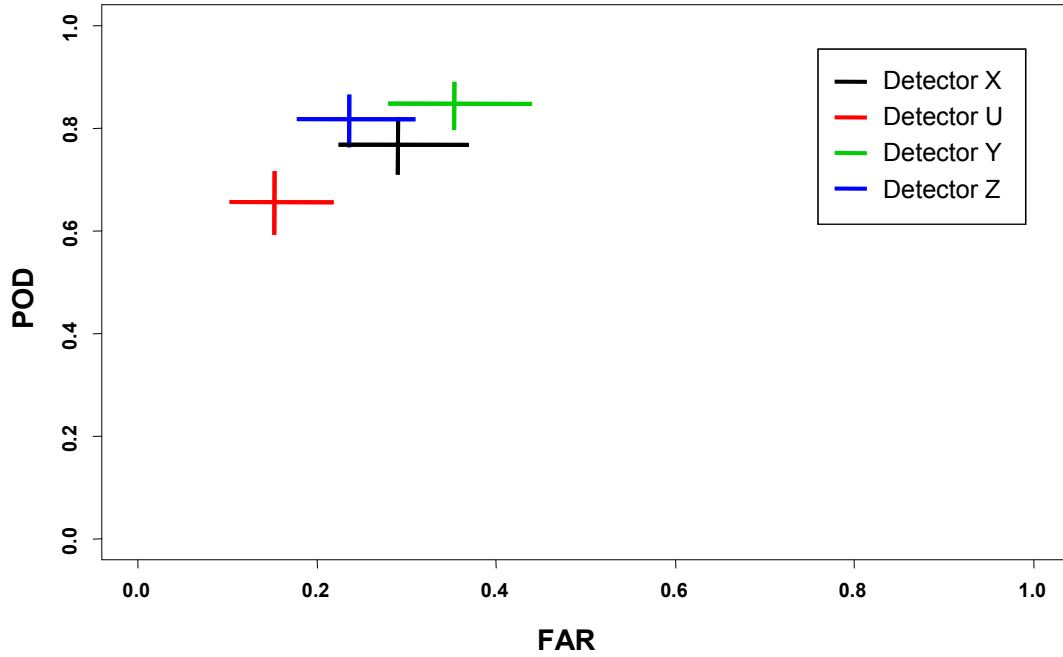


Figure 2. Overall results, comparison of detector models. The size of the crosses indicates the 95% confidence intervals.

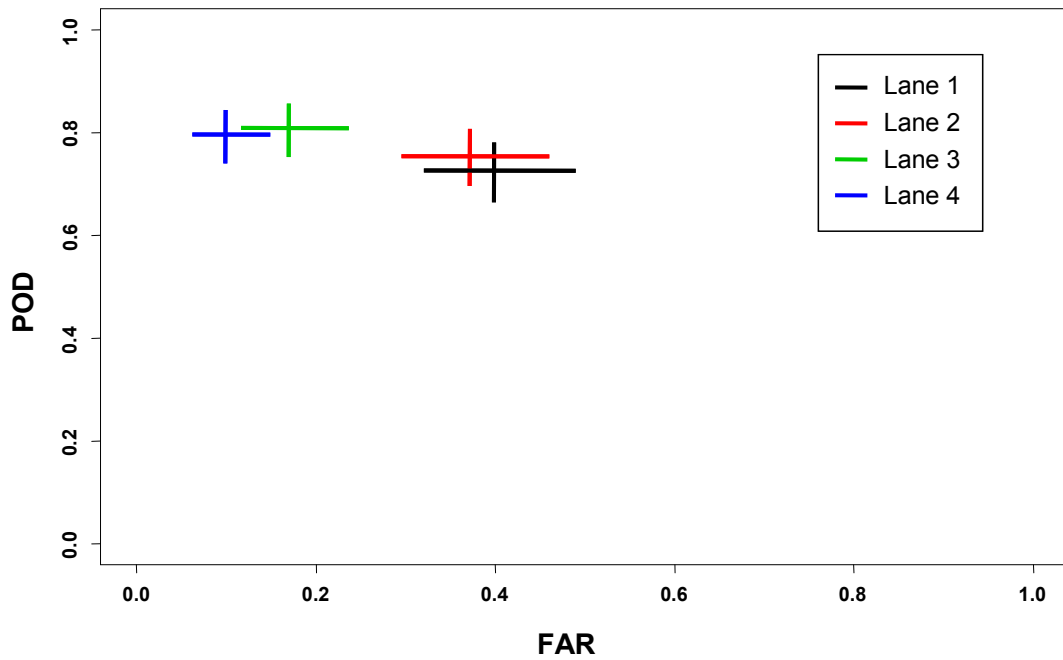


Figure 3. Overall results, comparison of lanes. The size of the crosses indicates the 95% confidence intervals.

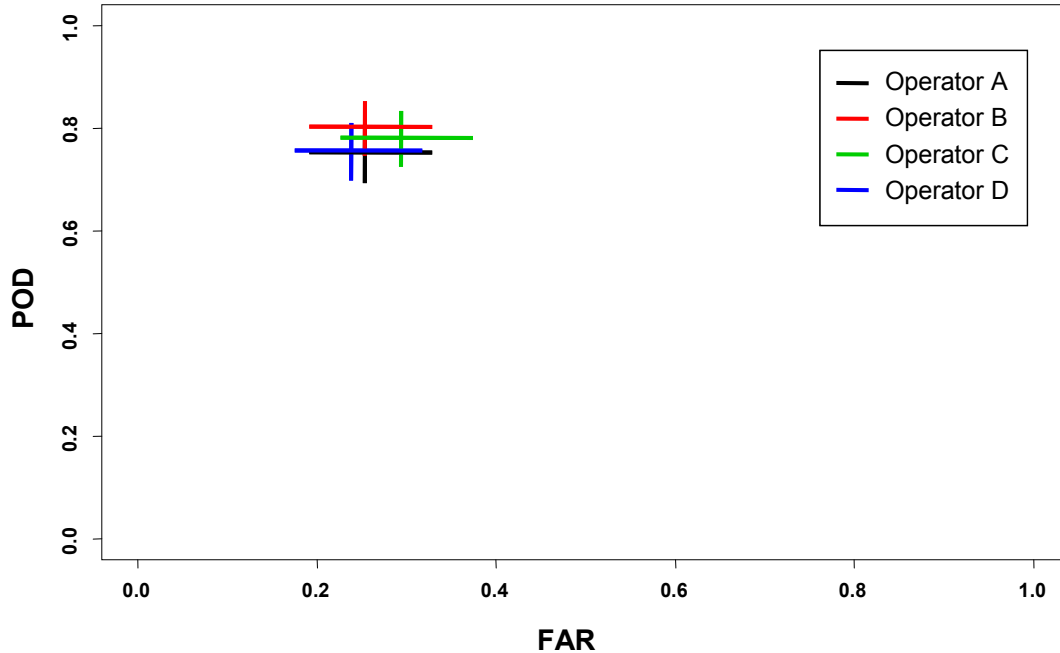


Figure 4. Overall results, comparison of operators. The size of the crosses indicates the 95% confidence intervals.

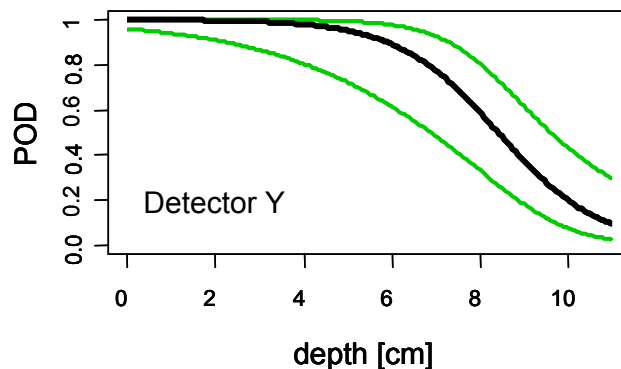


Figure 5. POD curve, the results of detector Y in Lanes 1 and 2 on the target PMA-2. The green curves indicate the 95% confidence bounds.

maximum detection depth:  $(10.1 \pm 0.6)$  cm. We see from Figure 5 that many mines at that depth and even at smaller depths than  $(10.1 - 0.6)$  cm = 9.5 cm were not detected in the reliability test. The main reason is that the operators did not know the positions of the targets during the reliability test. However, the human factor is not the only cause of this effect. There are some other sources of variation involved with blind trials. The depths of the targets can not be controlled so well as during the maximum detection height measurements. Each target is placed on another location, and the local properties of soil can vary. All these influences introduce higher unavoidable errors, both for the POD and for the target depths. This is why the POD curves obtained from a reliability trial will not fall with depth that abruptly as expected from the maximum detection height measurements. The analysis of the other detectors leads to the same conclusions.

## 5. Conclusions

Statistical design of experiment enables unbiased conclusions and minimises the experimental error, thus enabling shorter trials with lower expenses. Conclusions are unbiased when it is possible to separate the influences of different factors, for example, the influence of the detector model from the influence of the operator. Design of experiments is the only way to deal with complex experimental problems like metal detector tests.

The maximum detection height measurements are easier to perform than the more time consuming detection reliability tests. However, detection reliability tests include a large part of the human factor influences. Besides, the false alarm rate can be estimated only in reliability tests. The maximum detection height measurements should be performed with several operators and with repetitions, since they have a variance that cannot be ignored. ROC diagrams, POD curves and the results of maximum detection distance measurements are planned to be included in annual catalogues of demining detection equipment.

## References

- [1] C. Bruschini. A Multidisciplinary Analysis of Frequency Domain Metal Detectors for Humanitarian Demining. PhD thesis, Vrije Universiteit Brussel, September 2002. Available at <http://www.eudem.vub.ac.be/>.
- [2] CWA 14747:2003, CEN Workshop Agreement, Humanitarian Mine Action – Test and Evaluation – Metal Detectors, June 2003. Available at <http://www.itep.ws/>.
- [3] G. E. P. Box, W. G. Hunter, and J. S. Hunter. Statistics for Experimenters. John Wiley & Sons, 1978.
- [4] D. C. Montgomery. Design and Analysis of Experiments. John Wiley & Sons, third edition, 1991.
- [5] C. Mueller, T. Fritz, G. R. Tillack, C. Bellon, and M. Scharmach. Theory and Application of the Modular Approach to NDT Reliability. Materials Evaluation, 59(7):871–874, 2001.
- [6] C. Mueller, M. Scharmach, V. Konchina, D. Markucic, and Z. Piscenec. General Principles of Reliability Assessment of Nondestructive Diagnostic Systems and its Applicability to the Demining Problem. In 8th European Conference on Non-Destructive Testing (ECNDT 2002), Barcelona, Spain, Jun 17-21 2002.
- [7] C. Mueller, M. Gaal, M. Scharmach, U. Ewert, A. Lewis, T. Bloodworth, P.-T. Wilrich, and D. Guelle. Reliability Model for Test and Evaluation of Metal Detectors. Final report of the ITEP Project 2.1.1.2, Federal Institute for Materials Research and Testing (BAM), Berlin, Germany, September 2004. Available at <http://www.itep.ws/>.
- [8] The final report of the ITEP Project 2.1.1.8 (describing the metal detector trial, Croatia, May 2005) will be soon available at <http://www.itep.ws/>.