Trial Design for Testing and Evaluation in Humanitarian Mine Clearance

Dissertation approved by the Faculty of Environmental Sciences and Process Engineering, Brandenburg Technical University, Cottbus, for a doctorate degree in engineering sciences

submitted by
graduate engineer
Mate Gaal
from Zagreb, Croatia

Examiner: Prof. Dr.-Ing. Wolfgang Spyra
Examiner: Prof. Dr.-Ing. Heinrich Heidt
Examiner: Prof. Dr. Damir Markučič
Examiner: Prof. Dr.-Ing. Peter-Th. Wilrich

Date of the oral examination: 6 July 2007

Versuchsplanung für Prüfung und Bewertung in der humanitären Minenräumung

Von der Fakultät für Umweltwissenschaften und Verfahrenstechnik der Brandenburgischen Technischen Universität Cottbus zur Erlangung des akademischen Grades eines Doktor-Ingenieurs genehmigte Dissertation

vorgelegt von
Diplom-Ingenieur
Mate Gaal
aus Zagreb, Kroatien

Gutachter: Prof. Dr.-Ing. Wolfgang Spyra
Gutachter: Prof. Dr.-Ing. Heinrich Heidt
Gutachter: Prof. Dr. Damir Markučič
Gutachter: Prof. Dr.-Ing. Peter-Th. Wilrich

Tag der mündlichen Prüfung: 6. Juli 2007

Summary

This thesis proposes a design of experiment for testing and evaluation of the equipment and the methods used in manual mine clearance. The thesis is based on several metal detector trials and a trial of manual demining methods.

The core of this dissertation comprises four metal detector trials performed in Germany and Croatia in 2003 and 2005. The purpose of these trials was to investigate the feasibility of the tests described in the CWA (Comité Européen de Normalisation /CEN/ Workshop Agreement) 14747:-2003, the standard for testing metal detectors for humanitarian demining. The goals of the trials were: to find an appropriate design of experiment for testing metal detectors; to establish the use of ROC diagrams (Receiver Operating Characteristics) and POD curves (Probability of Detection) in the analysis of the experimental results; and to gain practical experience in organising and conducting metal detector trials. A part of this thesis is devoted to a trial of manual demining methods performed in Mozambique in 2004. The main goal of that trial was to compare the speed of various manual demining methods, including the most common excavation methods. The outcome of this work are the proposals and recommendations for an update of the standard for testing metal detectors CWA 14747:2003.

Maximum detection height measurements were performed as a part of the metal detector trial carried out in Croatia in 2005. The results reveal a high variability of the maximum detection height. This high variability needs to be taken into account in all experiments. A part of the variability is caused by the differences between the operators and by the setup of the metal detector. It is therefore recommended that two kinds of experiments with the maximum detection height as a response variable are defined in the next update of CWA 14747:2003. The first kind should include the setup, the soil and the operator as factors in the design of experiment. The in-soil measurements with the same detector should be performed with repeated setups and with several operators. The second kind of experiments should be experiments evaluating the influence of other predictor variables. In those experiments, it is recommended to perform one-factor or multiple-factor in-air measurements with the operators and the setup as a block.

The main part of the metal detector trials described in this thesis were

the detection reliability tests. Detection reliability tests as described in CWA 14747:2003 come closest to representing the real field conditions in demining. They include many environmental influences and, most importantly, many of the human factor influences. However, each test design is a compromise between fully representative conditions and cost effectiveness. In this thesis, a fractional factorial design based on the Graeco-Latin square is proposed as a solution to the experimental problem. The results are reported in the form of ROC diagrams and POD curves. The crossover design enables each operator to work with fewer detector models within a certain time. The variations of the design enabled an unbiased comparison of detectors in each soil and with each target model separately. It is recommended that the solutions proposed in this thesis be incorporated in the standard CWA 14747:2003.

It has been shown that maximum detection height measurements provide the information about the best possible performance of a metal detector in a reliability test.

Zusammenfassung

Diese Doktorarbeit stellt eine Versuchsplanung für Prüfung und Bewertung von Geräten und Methoden vor, die in der manuellen Minenräumung eingesetzt werden. Die Grundlage hierfür wurde mit einer Reihe von verschiedenen Versuchsreihen zum Test von Metalldetektoren und einer Versuchsreihe zur Untersuchung von ausgewählten manuellen Minenräumtechniken erarbeitet.

Im Mittelpunkt dieser Dissertation stehen vier Versuchsreihen zum Test von Metalldetektoren, die in Deutschland und Kroatien in den Jahren 2003 und 2005 durchgeführt wurden. Der Anlass dieser Versuchsreihen war, die Durchführbarkeit der Tests, die im CWA (Comité Européen de Normalisation /CEN/ Workshop Agreement) 14747:2003, dem europäischen Standard zum Test von Metalldetektoren in der humanitären Minenräumung, beschrieben sind, zu untersuchen. Die Ziele waren, eine geeignete statistische Versuchsplanung zum Test von Metalldetektoren aufzustellen, ROC-Diagramme (Receiver Operating Characteristics) und POD-Kurven (Probability of Detection) für die Analyse der experimentellen Ergebnisse einzuführen und praktische Erfahrungen bei der Organisation und Durchführung von Metalldetektortests zu sammeln. Ein weiterer Teil der Arbeit wurde einer Versuchsreihe auf dem Gebiet der manuellen Mineräumung gewidmet, die im Jahr 2004 in Mosambik durchgeführt wurde. Das Hauptziel dieser Versuchsreihe war, die Geschwindigkeit verschiedener manueller Entminungsmethoden zu vergleichen. Eingeschlossen waren die am häufigsten verwendeten manuellen Ausgrabungsmethoden, die keinen Metalldetektor verwenden. Die Ergebnisse dieser Dissertation sind Vorschläge und Empfehlungen zur Aktualisierung des europäischen Standards zum Test von Metalldetektoren CWA 14747:2003.

Als Teil der Versuchsreihe in Kroatien in 2005 wurden Messungen des maximalen Detektionsabstandes durchgeführt. Die Ergebnisse ließen eine hohe Variabilität des maximalen Detektionsabstandes erkennen. Diese hohe Variabilität muss bei allen Experimenten in Betracht gezogen werden. Ein Teil dieser Variabilität wird von den Unterschieden zwischen den bedienenden Personen und der Geräteeinstellung hervorgerufen. Deshalb werden für die nächste Aktualisierung des CWA 14747:2003 zwei verschiedene Arten von Experimenten mit dem maximalen Detektionsabstand als Zielvariable

empfohlen. Die erste sollte die Geräteeinstellung, den Bodentyp und das bedienende Personal als Faktoren in der Versuchsplanung enthalten. Die Messungen im Boden sollten mit wiederholten Geräteeinstellungen und verschiedenen Personen durchgeführt werden. Die zweite Art von Experimenten sollte die Bewertung des Einflusses von anderen Wirkungsvariablen beinhalten. Bei diesen Versuchen wird empfohlen die Experimente mit einem oder mehreren Faktoren in Luft durchzuführen, wobei die bedienenden Personen und die Geräteeinstellung jeweils für sich einen Block bilden.

Die Zuverlässigkeitstests zur Minendetektion, beschrieben im CWA 14747:2003, kommen realen Bedingungen bei der Entminung am nächsten. Darin enthalten sind sowohl viele der Umweltbedingungen als auch viele der überaus wichtigen Einflüsse des Faktors Mensch. Jede Versuchsplanung stellt jedoch einen Kompromiss zwischen vollständig repräsentativen Bedingungen und der Kosteneffektivität dar. Zur Lösung dieses experimentellen Problems wird in dieser Doktorarbeit die fraktionell faktorielle Versuchsplanung basierend auf dem griechisch-lateinischen Quadrat vorgestellt. Die Versuchsergebnisse werden in Form von ROC-Diagrammen und POD-Kurven dargestellt. Die Überkreuz-Planung ("crossover design") ermöglicht, dass jede Person nur wenige Geräte in einem bestimmten Zeitabschnitt bedient. Die Variationen in der Versuchsplanung erlauben weiterhin einen erwartungstreuen Vergleich der Leistungen der Detektoren in jedem Boden und bei jedem Minentyp separat. Es wird empfohlen, die in dieser Dissertation vorgeschlagenen Lösungen in die nächste Version des Standards CWA 14747:2003 einzuarbeiten.

Es wurde nachgewiesen, dass die systematischen Messungen des maximalen Detektionsabstandes die bestmögliche Leistung eines Detektors in einem Zuverlässigkeitstest wiedergeben.

Sažetak

U ovom se doktorskom radu iznosi prijedlog plana pokusa za testiranje i evaluaciju opreme i tehnika koje se koriste u ručnom razminiranju. Temelji se na nekoliko testova detektora metala i na jednom testu tehnika ručnog razminiranja.

Jezgru ovoga rada čine četiri testa detektora metala provedena 2003. i 2005. godine u Njemačkoj i Hrvatskoj. Svrha tih testova bila je ispitati provedivost testova opisanih u CWA (Comité Européen de Normalisation /CEN/ Workshop Agreement) 14747:2003, standardu za testiranje detektora metala za humanitarno razminiranje. Ciljevi testova bili su: odrediti odgovarajući plan pokusa za testiranje detektora metala, uvesti upotrebu ROC-dijagrama i POD-krivulja u obradu rezultata testiranja i steći praktično iskustvo u organiziranju i provođenju testova detektora metala. Dio ove disertacije posvećen je testovima tehnika ručnog razminiranja provedenim 2004. godine u Mozambiku. Glavni cilj tih testova bila je usporedba brzine raznih tehnika ručnog razminiranja, od kojih neke uključuju potpuno otkopavanje tla bez upotrebe detektora metala. Ishod ovoga rada su prijedlozi i preporuke za promjene standarda za testiranje detektora metala CWA 14747:2003.

Kao dio testova provedenih u Hrvatskoj 2005. godine, izvršena su mjerenja najveće visine detekcije ("maximum detection height"). Rezultati ukazuju na visoku varijabilnost najveće visine detekcije. Stoga tu varijabilnost treba uzeti u obzir u svim pokusima. Dio varijabilnosti uzrokovan je razlikama među operaterima i postupkom kalibriranja detektora. Stoga se u ovom radu preporučuje da se pri sljedećoj promjeni CWA 14747:2003 definiraju dvije vrste pokusa s najvećom visinom detekcije kao izlaznom varijablom. Prva bi vrsta trebala uključiti tlo, operatera i kalibriranje detektora kao faktore u plan pokusa. Mjerenja u tlu s jednim detektorom trebalo bi provesti uz ponavljanje kalibracije i s nekoliko operatera. Druga vrsta pokusa trebala bi obuhvatiti testove utjecaja drugih varijabli. Za te se pokuse preporučuje provesti mjerenja u zraku s jednim ili s više faktora, pri čemu bi operateri i kalibracija činili blok.

Većinu testova opisanih u ovom radu čine testovi pouzdanosti detekcije. Testovi pouzdanosti detekcije kakvi su opisani u CWA 14747:2003 su najbliže stvarnim uvjetima u razminiranju. Oni uključuju mnoge utjecaje

okoline i, što je najvažnije, mnoge utjecaje čovjeka. Međutim, svaki je plan pokusa kompromis između potpuno reprezentativnih uvjeta i efikasnog upravljanja troškovima. Kao rješenje eksperimentalnog problema, u ovom se radu predlaže frakcijski faktorijalni plan pokusa temeljen na grčko-latinskom kvadratu. Rezultati testova prikazani su u obliku ROC-dijagrama i POD-krivulja. "Crossover" plan pokusa omogućuje da operateri rade s manje modela detektora u istom vremenskom periodu. Varijacije plana pokusa omogućile su nepristranu usporedbu detektora u svakom tlu i sa svakim tipom mete. Preporučuje se da rješenja predložena u ovome radu budu uključena u standard CWA 14747:2003.

Pokazalo se da mjeranja najveće visine detekcije pružaju informaciju o najvišoj mogućoj performansi detektora metala u testovima pouzdanosti detekcije.

Acknowledgments

I would like to thank all who helped me during my work on this thesis. Most of all, I would like to thank Christina Müller, my first supervisor, for encouraging me, for trusting me and for pushing me always to the fore. I am most grateful to Prof. Peter-Th. Wilrich for his invaluable help in statistics. I would also like to thank my other supervisors, Prof. Wolfgang Spyra, Prof. Damir Markučič and Prof. Heinrich Heidt, for their help with the preparation of the manuscript.

My deep gratitude goes to my colleagues, from whom I have learned a lot about demining and metal detectors, and who supported me generously: most of all, to Dieter Gülle, Adam Lewis, and Andy Smith. I am also grateful for the friendly support of other colleagues, who helped me in many ways: Kurt Osterloh, Ivan Šteker, Prof. Uwe Ewert, Prof. Vjera Krstelj, and many others. Special thanks go to Croatian and Mozambican deminers who participated in our trials.

Finally, I would like to thank the German Federal Institute for Materials Research and Testing, the German Federal Foreign Office, and the German Ministry of Defence, for financially supporting our demining activities.

Contents

List of Figures

List of Tables

Abbreviations and Acronyms

ADP Accelerated Demining Programme (for Mozambique)

ANOVA analysis of variance

BAM Federal Institute for Materials Research and Testing, or Bundesanstalt für Materialforschung und -prüfung

CCW Convention on Certain Conventional Weapons

CEN European Committee for Standardisation

CWA 14747:2003 CEN Workshop Agreement, Humanitarian Mine Action — Test and Evaluation — Metal Detectors

EOD explosive ordnance disposal

ERW explosive remnants of war

FAR false alarm rate

GICHD Geneva International Centre for Humanitarian Demining

GPR ground penetrating radar

HCR-CTRO Croatian Mine Action Centre — Centre for Testing, Development and Training

ICBL International Campaign to Ban Landmines

IMAS International Mine Action Standards

IPPTC International Pilot Project for Technology Co-operation

ITEP International Test and Evaluation Program for Humanitarian Demining

ITOP International Test Operations Procedures

JRC Joint Research Centre of the European Commission

LIDAR light detection and ranging

LSD least signifficant difference

MAIC Mine Action Information Center of the James Madison University

MDD mine detection dogs

MDH maximum detection height

POD probability of detection

PPE personal protective equipment

R&D research and development

REST Remote Explosive Scent System

ROC receiver operating characteristic

SOPs Standard (or Standing) Operating Procedures

STEMD Standardised Testing and Evaluation of Metal Detectors

UNADP Accelerated Demining Programme

UNMAPA Mine Action Programme for Afghanistan

UNMAS United Nations Mine Action Service

UXO unexploded ordnance

WTD 52 Military Engineering Department 52 of the German Federal Armed Forces

Chapter 1

Introduction

The aim of this thesis is to establish a design of experiment for testing and evaluation of the equipment and the methods used in manual mine clearance.

The thesis was written during the author's work at BAM, Federal Institute for Materials Research and Testing (Bundesanstalt für Materialforschung und -prüfung), in the working group VIII.33 Reliability of Nondestructive Diagnostic Systems, in Berlin, Germany. It is based on several metal detector trials and a trial of manual demining methods. The core of this dissertation comprises four metal detector trials performed in Germany and Croatia in 2003 and 2005. The purpose of these trials was to investigate the feasibility of the tests described in the CWA 14747:2003 (where CWA stands for Comité Européen de Normalisation /CEN/ Workshop Agreement), the standard for testing metal detectors for humanitarian demining. The goals of the trials were: to find an appropriate design of experiment for testing metal detectors; to establish the use of ROC diagrams and POD curves in reporting the experimental results; and to gain practical experience in organising and conducting metal detector trials. A part of this thesis is devoted to a comparative trial of manual demining methods performed in Mozambique. The main goal of that trial was to compare the speeds of various manual demining methods, including the most common excavation methods. The practical outcome of all trials and of this dissertation are recommendations for the update of the standard CWA 14747:2003. These recommendations deal with the experimental design and data evaluation of metal detector trials.

The trial of manual demining methods was conducted as a part of the Study of Manual Mine Clearance by the Geneva International Centre for Humanitarian Demining. It was organised by A. Smith, an independent consultant, and executed in Mozambique, Maputo province, in November 2004. The four metal detector trials were performed within ITEP projects 2.1.1.2 and 2.1.1.8 (where ITEP stands for International Test and Evaluation Program for Humanitarian Demining) and organised by BAM, in a cooper-

ation with the Joint Research Centre of the European Commission (JRC). They took place in Oberjettenberg, Germany, in May 2003; in Benkovac, Croatia, in July 2003; in Oberjettenberg, in November 2003; and in Benkovac, in May 2005. The hosts of these trials were the Accelerated Demining Programme (ADP), Mozambique; the Military Engineering Department 52 of the German Federal Armed Forces (WTD 52), Germany; and the Croatian Mine Action Centre — Centre for Testing, Development and Training (HCR-CTRO), Croatia. The author of this thesis has participated in the organisation, design, execution, monitoring and data evaluation of all experiments described in this dissertation.

Context. Since the very beginnings of humanitarian demining, most efforts of the research and development (R&D) community were directed at improving landmine detection. Many scientists hoped to find a solution among technologies not used before in the field. However, all technical developments in that direction have so far failed to meet field needs [4, 43, 35]. Almost all improvements were a result of small investments of commercial companies, rather than large investments in R&D of new technologies [89].

One of the areas through which research efforts notably contributed to humanitarian demining is testing and evaluation of equipment, especially metal detectors [49]. Clearance organisations need reliable tests to choose the most suitable device for their needs. It is therefore necessary to test the equipment in conditions as close as possible to the actual field conditions. The tests closest to representing field conditions in demining are detection reliability tests described in the CWA 14747:2003. They include many environmental influences and, most importantly, many of the human factor influences. However, completely representative conditions cannot be achieved: each trial is a compromise between a faithful reproduction of demining conditions and the efficient use of available resources.

Structure. This dissertation is structured in three parts: a theoretical part, an experimental part, and practical recommendations. Chapters 2, 3, 4 and 5 form the theoretical part of this dissertation. This part deals with landmines, metal detectors, testing and evaluation of metal detectors and design of experiments. Chapter 2 presents the problem of landmines and how it is addressed. It describes mines and their impact, the international agreements regulating their use, the clearance methods currently employed and the possible future use of some new technologies for detection of landmines. The subject of Chapter 3 is the metal detector, the main detection tool used in humanitarian demining. This chapter gives an overview of the physical principles and the most important properties of metal detectors, describes the conditions of their use in minefields and presents the reliability model, a concept for understanding of all factors influencing the

performance of metal detectors. Chapter 4 discusses the current "state of the art" of testing metal detectors. It deals with the purpose of testing metal detectors, the existing standards for testing and it gives an overview of the trials performed up to the present. Chapter 5 is an introduction to statistical design of experiments. It presents the main principles of experimental design and gives examples of experimental design relevant to mine detection tests.

The experimental part of this work is contained in Chapters 6, 7 and 8. These chapters discuss the design of experiment and the data analysis for a comparative trial of manual demining methods, maximum detection height measurements and detection reliability tests. The trial of manual demining methods is the topic of Chapter 6. This chapter deals with the estimate of the experimental error and the problem of statistical bias. Chapter 7 discusses the measurements of the maximum detection height performed during the trial in Croatia, May 2005. The uncertainty of the measurements is estimated and discussed. Chapter 8 presents the experimental design and the results of the detection reliability tests performed during the metal detector trials in Croatia and Germany in 2003 and 2005. Detection reliability tests include more human factor influences, compared with maximum detection height measurements. This chapter clarifies the connection between the reliability test results and the maximum detection height measurements.

Chapter 9 is a practical conclusion of this work. It presents proposals and recommendations for an update of the standard for testing metal detectors CWA 14747:2003. The proposals are based on the results of the maximum detection height measurements and the detection reliability tests described in the previous chapters.

Information sources. The most important literature sources are stated in the introductory part of each chapter. A reader willing to learn more about humanitarian demining will find the following information sources useful. The web site of ITEP [61] hosts a very comprehensive data base of publications related to test and evaluation of humanitarian demining equipment. Two e-mail forums of demining specialists, the IGEOD forum [53] and the MgM forum [69], provide a lot of information about practical problems of demining and EOD. Conversations with deminers and observations of their work in minefields give an invaluable insight into the demining practice; they were also very important for the creation of this thesis.

Chapter 2

Landmines and Humanitarian Demining

This chapter briefly presents the problem of landmines and how it has been addressed to date. The first section describes mines and similar threats to civilians. The impact of landmines and some measures of success of mine clearance are discussed in the next section, with some examples. The third section deals with the instruments of international law regulating the use of landmines. The clearance methods currently employed in humanitarian demining are described in the fourth section, and the last section discusses some recent improvements and the possible future use of some new technologies.

There are numerous publications dealing with different aspects of mine action. Some web sites provide excellent short overviews, for example the International Campaign to Ban Landmines (ICBL) [60] and the Geneva International Centre for Humanitarian Demining (GICHD) [45]. A more comprehensive introduction to mine action is the "Guide to Mine Action" by GICHD [35], which has served as the main literature source for this chapter. A reader interested in more specific details will find the following sources helpful: the "Landmine Monitor" [57, 58, 59] published yearly, with the most current information on the mine situation worldwide; the "Jane's Mines and Mine Clearance" [65], with elaborate descriptions and drawings of hundreds of mines; the "Study of Manual Mine Clearance" [36], actually encompassing the broader area of humanitarian demining; the web site of the Mine Action Information Center of the James Madison University (MAIC) [70], with some databases and useful links; and finally, the Journal of Mine Action (available on the MAIC web site), which covers all topics of mine action.

2.1 Landmines

Long after their use in an armed conflict landmines continue to be a threat to the civilian population. In even greater numbers, unexploded ordnance — munitions that have been employed but which have failed to detonate as designed — plague post-conflict societies around the world.

Landmines (or simply *mines*) are explosive traps that are victim-activated, whether the intended target is a person or a vehicle [65, 78]. Designed to be exploded by the presence, proximity or contact of the victim, mines are placed under, on or near the ground. A mine comprises a varying quantity of high explosive contained within a casing (metal, plastic or wood), and a fusing mechanism to detonate the explosives. Mines are generally classified into two categories: anti-tank (or anti-vehicle) and anti-personnel. Antipersonnel mines are further divided into four categories based on their primary method of operation: blast, simple fragmentation, bounding fragmentation and directional fragmentation, based on their primary method of causing injury.

Among these types *blast mines* are by far the most common. They are activated with the victim's weight. The energy released by the explosive charge, typically 50-300 g of TNT, is the major cause of injuries, but secondary fragmentation injuries are also possible. The smaller blast mines are deliberately designed not to kill the victim, but to cause severe injuries that often lead to amputation. Minimum metal content blast mines can be very difficult to detect with a metal detector. However, fragmentation mines are responsible for more lethal demining accidents than blast mines [66, 88].

Simple fragmentation mines are installed on short wooden or metal posts and activated by tripwires. They scatter hundreds of metal fragments with the aim to kill or severely injure the victim.

Fragmentation bounding mines are also activated by tripwires, but their fuzes are often also tilt and pressure sensitive. They are partly buried in the ground and propelled in the air before exploding. The fragments cause death and severe injuries.

Directional fragmentation mines, sometimes called 'claymore' mines, direct hundreds of precut metal fragments in an arc on one side of the mine. They may be mounted on a tree or placed standing on their own folding legs. They are activated electronically by a soldier or initiated by tripwires.

Both terms, 'mine' and 'anti-personnel mine', are defined in international law in the Anti-Personnel Mine Ban Convention and the Convention on Certain Conventional Weapons (CCW) (more about this topic in Section 2.3). Anti-tank or anti-vehicle mines are often referred to as 'mines other than anti-personnel mines'.

The term *unexploded ordnance* (UXO) refers to munitions (bombs, shells, mortars, grenades and the like) that have been used but which have failed to detonate as intended, usually on impact with the ground or other hard

surface. Bomblets from cluster bombs are particularly dangerous [67]. Their failure rates vary between 1 and 40 per cent, depending on a range of factors, such as the age of the weapon, its storage conditions, the method of use and environmental conditions. Mine clearance operations include the removal of UXO and other ordnance, whether fuzed, fired or in storage.

To cover this a new term has entered the lexicon of international law: *explosive remnants of war* (ERW). Article 2 of Protocol V of the CCW defines ERW as "unexploded ordnance and abandoned explosive ordnance". The term does not include mines, booby-traps or other devices. However, some authors use it to include all these explosive devices.

2.2 Mine Action and Mine Situation

The International Mine Action Standards (IMAS) of the United Nations [55] define *mine action* as "activities which aim to reduce the social, economic and environmental impact of mines and UXO".

The term 'mine action' covers five groups of activities:

- mine risk education (also known as mine awareness);
- humanitarian demining, i.e. mine and UXO survey, mapping, marking and clearance;
- victim assistance, including rehabilitation and reintegration;
- stockpile destruction;
- advocacy against the use of anti-personnel mines.

A number of other activities support these five components of mine action, including: assessment and planning, the mobilisation and prioritisation of resources, information management, human skills development and management training, quality management and the application of effective, appropriate and safe equipment [35].

One of the most frequently asked questions regarding uncleared landmines is the question about their number. However, the size of the area known or believed to be affected by mines and its importance to the local populations are much better indicators of the impact and the clearance costs of landmines [96, 24]. Even better indicator is the required clearance time, since it has the largest influence on the clearance costs. The time needed to clear mine-affected land varies enormously depending on local conditions. The number of mines may have little effect on the speed of the demining process, because the search for mines usually takes much more time than their destruction.

The requirement of the IMAS is to clear all mines and UXO items to a specified depth. However, the same standard recognises that the term 'safe'

or 'mine-safe' (in the meaning: complete absence of risk) is not as "appropriate and accurate" as the term 'tolerable risk'. It should be accepted that complete safety cannot be assured in mine clearance. The everyday needs of local populations often force them to accept higher risk. According to the statistics of the Cambodian Mine Action Centre, about 60% of all the landmines cleared in Cambodia before 1999 were cleared by local people with no training, no funding and no equipment [25]. Many similar examples prove that people are ready to manage risk. This is why the local demining authorities sometimes actually tolerate a small reduction in the clearance standard if that speeds up the return of the valuable land to the community. The best measure of success is the greatest benefit possible to the largest number of people given the limited resources in both time and money. The socio-economic impact caused by mines and UXO is assessed through landmine impact surveys, with the goal of assisting the planning and prioritisation of mine action programmes and projects.

It is very difficult to estimate the total required funds to free the world from the impact of mines. A good indicator of the costs would be the size of the mined area, but it is not known with any certainty. The Croatian example can illustrate the extent of the problem [22, 21]. Two years after the end of the armed conflict, in 1997, 23% of the Croatian territory was considered mine suspected. By the year 2004 mine suspected areas were reduced to 2.1% of the country's area, which is 1,174 square kilometres. About 135 square kilometres of that area are known to be mined. Only a small part of this area reduction was achieved through clearance, while most of it was achieved through general survey. In that same period approximately $160 million has been expended on mine action, most of it on clearance. In 2005 the average cost of clearance of a square kilometre was about $1.68 million [59]. The costs of area reduction through general survey were much lower. The Croatian government and public companies finance about 80% of mine clearance in Croatia [97]. Most of the other countries affected by mines cannot finance their demining activities without significant foreign help, which is why their progress in clearance is much slower.

The "Landmine Monitor Report 2005" [58] states that more than 135 square kilometres were cleared worldwide in 2004. An additional 250 square kilometers were surveyed and returned to the community. According to the same source, over 190,000 mines were destroyed during the same year. According to the same source [57], 174,167 antipersonnel mines, 9,330 antivehicle mines, and 2.6 million items of UXO were found and destroyed in the year 2003. At the end of 1990's different estimates of the number of uncleared mines worldwide varied between several million and 150 million [96].

Similarly, it is difficult to assess the total number of victims with any certainty. It is only certain that landmines continue to claim human victims. The International Campaign to Ban Landmines [60] reported: "In 2002

and through June 2003, there were new landmine casualties reported in 65 countries; the majority (41) of these countries were at peace, not war. Only 15 per cent of reported casualties in 2002 were identified as military personnel." The same organisation estimates the total number of victims each year to be 15,000-20,000 [57, 58].

The indirect influence of the suspected presence of anti-personnel mines has devastating effects for the social and economic development of a country. Mined infrastructure such as transportation systems, electrical installations etc. can paralyse a country with a post-conflict infrastructure. For rural communities the loss of fertile agricultural land and safe access to water can be among the most serious consequences of the use of landmines.

The number of UXO items is even more difficult to estimate than the number of landmines, but it can be said with certainty that the total number of UXO items worldwide far exceeds the total number of landmines. In the last decade the international concern was dedicated to the impact of landmines, especially anti-personnel mines, so appearing to neglect the threat posed by UXO, but in fact demining agencies have always given the threat from ordnance the attention it deserves.

The notable decrease in the use of anti-personnel landmines throughout the world is a promising development. Since 1997, the year of the Ottawa Convention, there has been no legal trade in antipersonnel mines [58]. Already before that year the use of landmines dropped dramatically following the end of the cold war. Those changes are not just a consequence of increasingly higher public awareness, but also a result of a change in military practices; some commanders are reluctant to use mines that they would later have to clear themselves. Self-deactivating mines are being developed as a "strategic substitute" of antipersonnel landmines, but their "safety" is seriously questioned by many people involved in landmine clearance largely because of a lack of confidence that they will reliably detonate or deactivate as designed.

2.3 Instruments of International Law Addressing the Problem of Landmines

Two instruments of international law apply specifically to landmines: the *Convention on Certain Conventional Weapons* (CCW) [20] from 1980 and the *Ottawa Convention* (also known as *Mine Ban Treaty*) [79] from 1997. The CCW restricts non-detectable shrapnel, blinding lasers, depleted uranium, incendiary bombs and landmines. The Ottawa Convention prohibits the production, stockpiling, transfer and use of all anti-personnel mines. Both only apply to countries that sign up to the constraints.

The 1980 Convention on Certain Conventional Weapons (CCW) regulates the use, and in certain circumstances also the transfer, of specific

conventional weapons. In addressing landmines, booby-traps and "other devices", CCW Protocol II, adopted in 1980, reflected customary law by limiting the use of mines to military objectives. The 1996 Amended Protocol II strengthened the rules governing anti-personnel mines, though it did not include their total prohibition. The signatories are obliged not to use non-metallic antipersonnel mines, as well as those designed to be activated by metal detectors.

The Ottawa Convention was adopted on 18 September 1997 and entered into force on 1 March 1999. The purpose of the Convention is stated in the first line of its preamble: the States Parties are "determined to put an end to the suffering and casualties caused by anti-personnel mines, that kill or maim hundreds of people every week (...), obstruct economic development and reconstruction, inhibit the repatriation of refugees and internally displaced persons, and have other severe consequences for years after emplacement". By the end of 2005, about 150 states have signed the Convention, which is about three-quarters of the world's states. The Convention obliges the signatories not to use, develop, produce, stockpile or transfer anti-personnel mines, and it requires that they destroy existing stocks of anti-personnel mines, clear mined areas, and assist victims.

The definition of anti-personnel mine crucially influences the reach of the Ottawa Convention. An anti-personnel mine is defined as a subset of a mine. The Convention defines a mine as "a munition designed to be placed under, on or near the ground or other surface area and to be exploded by the presence, proximity or contact of a person or a vehicle". The anti-personnel mine is defined as a "mine designed to be exploded by the presence, proximity or contact of a person and that will incapacitate, injure or kill one or more persons". Thus the Ottawa Convention does not prohibit antitank mines, nor mines which are activated by an operator from a safe distance, although all these devices can be a threat to civilians.

Each state is obliged to clear all emplaced anti-personnel mines not later than 10 years after it becomes Party to the Convention, but an extension period of up to 10 years is possible. It is apparent that many countries will ask for an extension because of their limited ability to finance clearance operations.

A new protocol to the CCW, Protocol V, adopted in November 2003 and entered into force in November 2006, addresses the humanitarian impact of explosive remnants of war (ERW). That document defines ERW as "unexploded ordnance and abandoned explosive ordnance", and calls for "all feasible precautions" to protect civilians from their risks and effects.

2.4 Clearance and Detection Methods in Humanitarian Demining

Humanitarian demining is the core component of mine action. It covers all activities which lead to the removal of mines and UXO hazard. These include technical survey, mapping, clearance, marking, post-clearance documentation, community mine action liaison and the handover of cleared land. Among these activities, clearance operations usually carry most of the expenses. Humanitarian demining should be clearly distinguished from military demining — when the land is cleared for civilian use, the public requires complete safety.

In many clearance operations a combination of several methods is used. According to the field conditions, the following three elements can be applied: manual demining [36], mine detection dogs [33] and mechanical equipment [34]. The most frequently used method is manual clearance, although machines and dogs are playing an ever-increasing role in humanitarian demining. The chosen method needs to be cost effective, but it also needs to minimise the risk for deminers. The choice of the most appropriate combination of methods depends on many factors: mines/UXO, terrain, infrastructure, logistics, national legislation, and others. For example, in countries with higher labour costs the use of machines is more cost-effective [96], which is why even 85% of the Croatian mine clearance in 2004 was performed with the help of demining machines. Up to date there were no reported accidents behind a machine in Croatia.

Manual demining [36] comprises the use of metal detectors and/or excavation tools. Metal detectors are used for detection of mines, since all mines known to be used in military conflicts contain some metal, while excavation tools are used to uncover the finding. If the metal contamination of the ground is too high, deminers cannot use metal detectors and they are forced to prod and to excavate the soil over the entire suspect area. A deminer typically works in a one metre wide lane until he locates a suspicious object. He uncovers it or excavates it, and then, if the object is a mine or UXO, it is blown up in situ or disarmed and moved for destruction at the end of the day.

The prodder is a usually needle-shaped about 30 cm long tool used as a physical check of the presence of a mine or UXO. It is simple, cheap and often effective. Its shortcomings are that it brings the hands and the face of the operator close to the mine and it is hazardous if the orientation of the mine is different than horizontal. Furthermore, it hardly penetrates rocky soil and, since it is applied under about 30° angle to the ground surface, the depth of its penetration can not be greater than 10-15 cm. It is usually used together with a small spade, trowel or some other tool to remove the loosened spoil. In most demining organisations the prodder is the main

tool used to verify the presence of a mine after it is detected with a metal detector.

Metal detectors work on principles of induced eddy currents. The presence of metal, or any other conducting material, is indicated to the operator with the mean of an audio signal. The metal content of mines has been decreasing in line with the development of more sensitive metal detectors. In many mines it is confined to a firing pin and a tiny detonator case. Only a few mine types with no metal content have found their way into use and they were used in such small numbers that they have not caused recorded accidents [86, 87]. Their use has been mostly in Lebanon and some African countries [1] [49]. More and more sensitive metal detectors are being built to cope with the problem of small metal content of mines. As a consequence, the metal fragments present in the soil and the natural magnetisation and conductivity of the soil represent the main limitations for the use of metal detectors. Most modern metal detectors have some ability to compensate for the effects of the magnetism of the ground. The next chapter describes metal detectors and their use in more detail.

Under favourable circumstances, *mine detection dogs* (MDD) [33] can discriminate between varieties of substances, and they can be more sensitive than human made devices. Some gas chromatographs may be able to detect concentrations down to 10^{-12}, while dogs (and rats) can detect 10^{-15} or less. Even though they cannot replace manual detection, dogs can be a powerful tool when used in combination with manual and mechanical clearance methods. They are particularly useful on areas inaccessible to machines and on areas with metal contamination. The main weaknesses of MDD's are that they fail in hot climates, their reliability cannot be easily checked and that they cannot normally be used if the concentration of mines is too high.

Many *mechanical devices* have been produced to assist mine clearance by detonating or destroying mines, or typically by cutting vegetation. There is increasing empirical evidence that demining machines designed to detonate or destroy mines can be efficiently used as part of a system in which other processes are also applied in order to meet the clearance requirements of the IMAS [34, 92]. Demining machines are usually followed by manual demining and/or dog teams. The main advantage of machines is that they can speed up the demining process significantly and their main shortcoming is their high price. Their environmental effects have not yet been sufficiently studied.

The tools and methods used for UXO detection are mostly inadequate for mine searching, since they are designed to detect larger amounts of metal at larger depths. The signal of a small metal content mine would be overwritten

[1] Surprisingly, there is no available literature dealing with the problem of non-metallic mines applied in the field. However, two interesting discussions have developed on the IGEOD forum [53] and the MgM forum [69] (subject: non-metal mines, from 30 January 2006). The entire e-mail correspondence is available from both forums on request.

by other signals.

2.5 New Technologies in Humanitarian Demining

Most efforts of the research and development (R&D) community have been directed towards improving landmine detection. Many scientists hoped to find a solution among technologies that had not been used before in the field. However, all technical developments in that direction have so far failed to meet field needs [4, 43, 35, 38]. The only innovations that have found their way into widespread use in the field are technical improvements of the existing equipment: metal detectors, manual demining toolkit, personal protective equipment and especially methods of mechanical assistance [89]. Almost all improvements have been the result of small investments by commercial companies, rather than large investments in the R&D of new technologies. In recent years demining has become both cheaper and safer, but that is more the result of improvements in management and working practices than significant technological or technical advances [44, 96, 36].

The main reason for the failure of new technologies is probably the lack of international coordination of the R&D activities and the lack of interaction between researchers, field operators, and donors [1]. This was recognised already in 1997, at the workshop that accompanied the signing of the Ottawa Convention. Nevertheless, the cooperation between the actors of the mine action community is still not satisfactory. As a result, considerable effort has been invested, often with competition between national projects, in developing equipment that is inadequate for field use [4].

Contrary to the opinions of some disappointed practitioners, this is not a reason to completely abandon financing R&D. The Ottawa Convention commits the States Signatories to clear landmines until 2009, but with the current technology and management practices even a tenfold increase of funds for clearance operations would not solve the landmine problem by 2009 [4]. About $400 million is yearly spent on clearance worldwide, while only about $30 million is spent on R&D for humanitarian demining [58]. The R&D funds are not transferable to clearance operations, but even if they were, the benefit of such a transfer would be minor compared to the potential benefits of research. New technologies are obviously needed, as well as new management practices. Technologists need to invest their time to understand the end-users' needs. They need to visit mine fields and learn more about the existing methods and the working conditions. They also need to understand that detection is only one of the areas that can improve demining efficiency; others include area reduction, strategic planning using information technology tools, programme management, etc. Even more important, this needs to be understood by donors and by policy makers at the national and international level.

The detection technology closest to application in landmine detection is the ground penetrating radar (GPR) [18]. It consists of a transmitter, which sends an electromagnetic pulse or a continuous wave in the microwave region, from several hundred MHz to several GHz, and a receiver, which receives the reflected signals. It detects the difference of the permittivity or dielectric constant, which is why plastic or other non-metallic objects can be detected by GPR. In a dual sensor in conjunction with metal detectors they can reduce the false alarm rate, but may also increase the probability of missing a mine. Despite the significant improvements already achieved, such combined detectors have not yet found widespread use in humanitarian mine clearance.

The field of vapour detection also achieved some success. Rats are already in use in some organisations and the results are very encouraging. Samples of air are collected and brought to them to identify traces of explosive, and the "free-running mode" is also being investigated. The Remote Explosive Scent System (REST) should also be mentioned. For this, the air above the road surface is filtered with filters carried on a vehicle. The sampling filters are replaced at recorded intervals and later analysed by mine detection dogs. In case of a positive identification free-running dogs return to the corresponding part of the road to locate the mine.

Another method that is considered promising is infrared detection, but the resolution of current infrared cameras is insufficient to identify small mines. Its use against antivehicle mines and UXO remains a possibility. Some other detection methods that have been occasionally recognised as potentially promising are some acoustic or seismic methods, light detection and ranging (LIDAR), insects, chemical sensors (especially gas chromatography), etc. Sensor fusion was also considered (especially dual sensors of metal detectors with GPR), some vehicle mounted detectors, etc, but none of these methods have found widespread use in the field.

There were some important advances in technology other than mine detection. Most important are the improvements of personal protective equipment, information technology, and prosthetic feet [1]. One of the most important areas to which research has contributed significantly are the standards for testing and evaluation of demining equipment. The whole of Chapter 4 is devoted to that topic.

Chapter 3

Metal Detectors in Humanitarian Demining

This chapter describes the main detection tool used in humanitarian demining, that is, the metal detector. The main principles of electromagnetic induction are presented in the first section, and the application of these principles in the construction of metal detectors is elaborated in the next section. The third section describes the use of metal detectors in minefields and some problems that deminers encounter in their daily work. The last section presents a concept called 'reliability model', which is a possible approach to understanding all factors that influence the performance of metal detectors.

An excellent study of metal detectors by C. Bruschini [16] offers a detailed overview of the working principles and some technical properties of metal detectors. The "Metal Detector Handbook" by D. Guelle et al. [49] and the web site of A. Smith [91] provide a more practical view for field use, including detailed descriptions of demining procedures and many practical problems of mine clearance.

3.1 Physical Principles of Electromagnetic Induction

Various detection devices based on electromagnetic induction are used in many areas of non-destructive testing. Metal detectors used in demining are only one example of those devices. Their operating principles are described in this section and the same principles are valid for other electromagnetic induction devices.

All *metal detectors* used in demining consist of a search head, a telescopic pole or an extension rod with a handle, and an electronic unit. The search head of the metal detector contains a coil carrying a time-varying electric current. This current generates a time-varying magnetic field (according

to Ampere's law), called a primary magnetic field, which propagates in all directions [68]. In neighbouring objects it induces electromotive force (Faraday's law), which causes currents in conductive materials like metals. These currents create another magnetic field, which is called a secondary magnetic field. Another coil placed in the search head receives both magnetic fields: the primary field and the much weaker secondary field. The coil generating the primary field is termed 'primary coil' or 'transmitter coil' and the one receiving both fields is named 'secondary coil' or 'receiver coil'. The secondary field is converted into an audio signal to be easily interpreted by the deminer.

For a better understanding of these processes, it is worthwhile to discuss them more in detail. Let us investigate the case when a small conductive object is placed on the symmetry axis of a circular metal detector coil. We are interested in the magnetic field created by the coil, the so called primary field, at the position of the object. Since our object is small, the local variations of the primary field can be neglected and we approximate that the object is placed in a uniform field. From Ampere's Law it is easy to find a solution on the symmetry axes where our small object is placed. If the coil has a radius R and carries a current I, the primary field on the symmetry axes at a distance d from the coil behaves as

$$B_z\left(d\right) = \frac{\mu I}{2} \frac{R^2}{\left(R^2 + d^2\right)^{3/2}} \tag{3.1}$$

where μ is the magnetic permeability of the medium.

According to Faraday's law this field creates an electromotive force in the object,

$$EMF = -\frac{\partial \Phi}{\partial t} \tag{3.2}$$

with the magnetic flux

$$\Phi = \int_S \vec{B}\, \mathrm{d}\vec{S} \tag{3.3}$$

where S is the surface of the object and B the primary field. This electromotive force creates eddy currents, which are thus proportional to the primary field. Let us suppose that our object is a small horizontally laid circular loop of a radius R_{EC}, carrying the induced eddy current I_{EC}. This current creates another field called a secondary field, $\vec{B_2}$, which induces an electromotive force in the coil. To find that electromotive force, we need to integrate the secondary field over the area of the search head, which is why we need the distribution of the secondary magnetic field in space. An exact analytical solution to the problem of the field induced by a circular coil does not exist, however, in most cases the size of the object is much smaller than its distance to the search head, $R_{EC} << d$. Using Ampere's law, it can be shown [63] that for this case the secondary field in point $\vec{r}$ is approximated

with

$$\vec{B_2}(\vec{r}) = \frac{3\vec{n}(\vec{n}\cdot\vec{m}) - \vec{m}}{|\vec{r}|^3} \tag{3.4}$$

where $\vec{r}$ is the position vector with an origin in the middle of our small object, $\vec{n}$ is a unit vector pointing in the same direction ($\vec{n} = \frac{\vec{r}}{|\vec{r}|}$) and $\vec{m}$ is the magnetic dipole moment of the object. It is defined as

$$|\vec{m}| = \frac{\mu_0}{4\pi} \cdot I \cdot S = \frac{\mu_0}{4\pi} \cdot I \cdot R_{EC}^2 \pi \tag{3.5}$$

having a direction towards the middle of the search head.

The induced voltage in the coil, or the electromotive force, is obtained by integrating the secondary field over the area of the search head:

$$EMF_2 = -\frac{\partial}{\partial t} \int_{searchhead} \vec{B_2} \cdot \mathrm{d}\vec{S} \tag{3.6}$$

The result of the integration is

$$EMF_2 = -\frac{\mu_0 \pi}{2} \frac{\partial I_{EC}}{\partial t} \frac{R^2 R_{EC}^2}{(R^2 + d^2)^{3/2}} \tag{3.7}$$

Since the eddy currents are proportional to the time derivative of the primary field, we can write

$$EMF_2 \propto \frac{\partial^2 B_z}{\partial t^2} \frac{1}{(R^2 + d^2)^{3/2}} \tag{3.8}$$

Using Equation (3.1) we get

$$EMF_2 \propto \frac{\partial^2 I}{\partial t^2} \frac{1}{(R^2 + d^2)^3} \tag{3.9}$$

Thus we see that the induced voltage in the search coil reduces very rapidly with the distance between the search head and the object. This is why we say that metal detectors are proximity sensors. We can see from the same equation that metal detectors with larger search heads (larger R) are less sensitive on close targets, but they are more sensitive on targets on larger distances from the search head. This is why they are used for detection of UXO, which are often larger and can be found deeper than antipersonnel mines.

The response of many minimum metal mines, for example, the PMA-2, can be well approximated by the magnetic dipole model (Equations (3.4) and (3.5)). Larger objects and composite objects will produce a different secondary field [16]. An example of an elongated composite object known from experience is the GYATA-64, which can even produce alarms at two locations [90] (see Chapter 6).

Eddy currents are not the only mechanism forming a secondary field. Paramagnetic soils also create a secondary field that causes serious detection problems if the goal is to detect conductive objects. This phenomenon is responsible for the difficulties that metal detectors have with paramagnetic soils and it is tackled in Section 3.2.5 in more detail.

The secondary field depends on many factors: the conductivity and permeability of the object and of the background, the size and the shape of the object, its geometry (distance, orientation) and the temporal and spatial distribution of the primary field [23]. It can be shown that the influence of the target's dielectric properties are negligible in the frequency range used by mine searching metal detectors, which is between 1 and 100 kHz [16]. Low conductivity materials like stainless steel, which is contained in some landmines, are harder to detect. Ferromagnetic objects create a larger secondary field due to their higher relative permeability, or alternatively it could be said due to their higher magnetic susceptibility, since they are connected in the expression $\mu_r = 1 + \chi$, where μ_r is the relative permeability and χ is the magnetic susceptibility.

It is often believed that an object containing more metal will be easier to detect with a metal detector, but this is not necessarily true. The eddy currents circulate close to the surface of the object, which is why it is said that metal detectors are surface area detectors. An electromagnetic field decays in a conducting medium as $e^{-r/\delta}$, where r is the distance from the surface and δ is a characteristic depth of penetration. This depth is called *skin depth* and the whole effect is called *skin effect*. The skin depth in the frequency range of metal detectors is typically of the order of 1 mm (e.g. for aluminum at 20 kHz, $\delta = 0.60$ mm). Since metal detectors are mostly surface area detectors, the amount of metal does not necessarily influence the size of the signal. The skin effect is well described in all physics textbooks [68, 63]. A simple explanation of this phenomenon is offered by C. Bruschini [16].

This section gave some insight into the main principles of metal detectors. Actual detectors use more than one frequency or they use an electromagnetic pulse. Many detectors have only one coil, having the role of both the primary and the secondary coil. All the varieties of metal detectors are briefly described in the next section.

3.2 Technical Properties of Metal Detectors

Metal detectors used in humanitarian demining typically weigh less than 2 kg [15]. The price of the latest models is usually between US$2000 and 4000. The most common search head shapes are round, oval and rectangular. Most can detect a small metal content mine like the PMA-2 at about 10 cm distance in air and a typical antivehicle mine with a metal casing at more

than 50 cm. Magnetic soils reduce their detection capabilities. Most metal detectors use standard cell batteries that last several days. They are easy to use and the latest models are more ergonomic than their predecessors. Their output is an audio signal resulting from an extensive data processing and it helps the operator to locate a conductive object. For more specific technical details of available metal detector models the reader should consult the "Metal Detectors and PPE Catalogue 2005" [31] and the older "Metal Detectors Catalogue 2003" [32].

All the varieties of metal detectors are briefly presented in Subsections 3.2.1, 3.2.2 and 3.2.3. There are some limitation to the achievable sensitivity of metal detectors and these limitations are presented in Subsections 3.2.4 and 3.2.5, and in Section 3.3. A fundamental limitation is the noise of the electronic elements, which will not be discussed in this work. It can easily overwrite the signal coming from the eddy currents, since the secondary field produced by eddy currents is much smaller than the primary field of the transmitter coil. Other limitations include the electromagnetic interference with external sources of electromagnetic fields, the magnetic response of the soil and the eddy currents in the soil.

3.2.1 Coil Configurations

Several different coil configurations are used in the design of metal detectors for humanitarian demining. Only those that were encountered in the trials described in Chapters 7 and 8 are considered here.

Most metal detectors for landmine detection are coplanar, which means that the primary and the secondary coil are in the same plane. Their diameter is typically 20-30 cm. One of the oldest designs has two circular and concentric coils, like those of the Schiebel AN-19/2. An oval coil is also in use, for example, the Vallon VMH3CS and Ebinger EBEX® 420 detectors. Some detectors use only one coil having the function of both the primary and the secondary coil. So called *"double-D"* metal detectors use a receiver coil consisting of two halves resembling two D letters. This is the favoured solution of some manufacturers such as CEIA and Foerster. The technical specifications of the detectors mentioned in this paragraph and many other detectors can be found in the "Metal Detectors and PPE Catalogue 2005" [31]. For other possible coil designs see [16].

3.2.2 Audio Signals

The audio signals of metal detectors can be very different. The "double-D" detectors record the difference of the signals in the left and the right side of the search head. The positive and the negative difference produce two different tones, so that it appears to the deminer that each half of the search head produces a different tone. The audio signal has a sharp transition when

the search head is moved left-right just above the object. All manufacturers who decide to use other designs have a choice between the static mode and the dynamic mode. *Static mode* detectors produce a sound whenever the secondary field exceeds a detection threshold. *Dynamic mode* detectors give an alarm when they detect a change of the secondary field, that is, when the search head approaches a conductive object and when it strays from it, but no alarm when the detector is not moved relative to the target.

3.2.3 Frequency Domain and Time Domain

Metal detectors can be classified according to the time dependence of the primary field. There are frequency domain (or continuous wave, or sine wave) and time domain (or pulse induction) metal detectors [94, 16, 15]. Tests performed so far showed no systematic differences in performance between frequency domain and time domain metal detectors.

Frequency domain metal detectors have a sinusoidal primary field and they often use several frequencies in the range between 1 and 100 kHz. The amplitude and the phase of the received signal contain some information about the detected object. As an equivalent alternative we could imagine a complex change of impedance as a consequence of the proximity of the target. Frequency domain detectors can use a single coil or separate transmit and receive circuits.

Time domain metal detectors create pulses of electromagnetic field with a typical repetition rate of the order of 1 kHz and measure the exponential decay of the secondary magnetic field. This decay is slower than the decay of the primary field as a consequence of the eddy currents induced in the object. Either the secondary field itself or its time integrals in certain time windows are used as a source of information about the target. Since the receive phase follows after the transmit phase, the same coil can be applied for transmitting and receiving.

3.2.4 Electromagnetic Interference

A signal from an external source can induce a voltage in the detector coil or directly influence the electronics of the metal detector, so that it causes an audio signal without any presence of metal [49]. Possible sources of *interference* are high-voltage power lines and substations, radio transmitters, electric motors and other metal detectors. Some detectors filter radio signals and the frequency of electric power transmission, 50 Hz or 60 Hz. The interference with other metal detectors occurs if their mutual distance is smaller than a certain critical distance which varies between 1 and 20 m, depending on the detector model. The working frequency of some detector models can be adjusted so that the critical distance is very small. The interference between detectors is rarely a problem in practice, since the

safety distance between deminers is 25 m. It is an important issue after an accident, if a rescue team needs to use metal detectors to approach the victim having a switched on metal detector, and during testing, since there is no need to keep a safety distance.

3.2.5 Ground Compensation

The magnetic properties of soil are one of the most important factors influencing detection capabilities of metal detectors [19, 16, 27, 49, 64, 13, 17]. Magnetic susceptibility and to a lesser degree electric conductivity of the soil create a secondary magnetic field which makes the detection of conductive objects more difficult. The frequency dependence of the magnetic susceptibility is the most important influencing factor [9, 76]. While the metal components of mines can be very small, the soil occupies the whole space below the search head, so even though the ground electro-magnetic properties are much weaker than those of the metal piece, they can still produce a significant secondary field making detection more difficult. Soils that reduce the performance of metal detectors are termed *uncooperative* or *noisy*, while the other soils are called *cooperative* or *neutral*.

A careful examination of Equation (3.1) in Section 3.1 reveals that the soil hardly influences the primary and the secondary magnetic field coming from the detected object. All non-ferromagnetic media, which means most of the soils found in nature, have a relative magnetic permeability close to 1, i.e. their permeability μ is almost equal to the permeability of air. The source of problems for metal detectors is not the alteration of the primary or the secondary field, but the frequency dependence of the soil susceptibility.

Most modern detectors have so called *ground compensating* abilities, which allows them to reduce their sensitivity to the soil, with much smaller reductions of sensitivity to metal objects. This procedure is based on the electromagnetic differences between metal and the soil. However, no detector known to the author has achieved a ground compensation that would not reduce its sensitivity to metal. Heterogeneous soils are particularly difficult to compensate out. The ground compensation is performed in the field in a way that the detectors are adjusted to the soil in which deminers search for mines. This procedure is simple and it lasts no longer than two minutes with the latest detector models. The exact procedure varies between models of detector. Some models require the search head to be moved, while some require the search head to be brought into a specific position.

Frequency domain detectors use the information contained in the amplitude and phase of the received signal. During the ground compensation procedure they record the background signal, that is, the signal coming from the soil. During mine searching this signal is subtracted from the signal coming from the search head.

Time domain detectors make use of the difference in the decay time of

the soil and the metal [17]. A pulse detector samples only specific time windows of the received signal, or the signal itself in just a few points, so that it can be made less sensitive to soil.

3.2.6 Possible Technological Improvements

The R&D community seeking improvements in existing metal detector technology faces two major challenges. The first is to increase the sensitivity of metal detectors to smaller and deeper targets. The other challenge is to reduce the false alarm rate. There are two main sources of false alarms: metal clutter and the soil. Over recent years we have witnessed a constant improvement in the ground compensation abilities of metal detectors. Since the ground compensation reduces the noise from the soil, metal detectors can be setup to a higher sensitivity, thus improving their detection capabilities. Metal detectors are not mine detectors, they are designed to detect metal. Deminers frequently have to investigate several hundred signals caused by metal clutter before they find a mine, which can add significantly to the time taken to clear an area. The only way to deal with this problem using today's metal detectors is to reduce the sensitivity to the level at which the expected threat will be still detected with certainty and most of the clutter will not produce an alarm. This calibration procedure is sometimes called "setting up the detector to the known threat" and it certainly carries some risks.

Many R&D efforts are directed to reducing the false alarm rate by target identification and parameter estimation, like the target depth, size or type of metal [16]. Some imaging applications and some sensors other than coils are being studied. These technologies are not yet ready for field use.

3.3 Use of Metal Detectors

Daily routines of the work with metal detectors are well described by D. Guelle [49], R. Gasser [44] and A. Smith [91]. Demining drills vary locally, so that a more interested reader should study the Standard Operating Procedures (or Standing Operating Procedures, SOPs) and training materials of the specific demining group. However, actual practice may not be recorded in writing because it can take a lot of time to update the documentation to reflect current best practice. For example, many SOPs prescribe that deminers lie prone when excavating mines, but this is very rarely being followed [88, 87, 86]. The parts of the drills of interest for this work are described in the following paragraphs.

Deminers work in 1 m wide lanes and they typically mark them with sticks and tapes as they progress. In many areas, most of the time they spend removing vegetation before the actual search begins. When the threat

includes tripwires, deminers may use their detectors to detect tripwires before they start to cut the undergrowth. A visual check gives them the first information about the area to be cleared. It is common for deminers to have a wooden stick placed horizontally across the lane at the limit of the cleared area. This stick marks the so called baseline and deminers never cross it. During the search deminers move their metal detectors sideways, and such a movement is called a *sweep*. The distance between consecutive sweeps is termed *sweep advance*. It is most common that the sweep advance is a half of the search head width.

The audio signal produced by detectors in proximity to a metal piece is called a *reading*, a *(detection) alarm* or a *(detector) signal*. Every reading has to be *investigated*, which means that the deminer has to look for the cause of the signal, first visually, than with a prodder. The search head must be used to approach the signal from different angles to pinpoint the signal source. If there are two signals, the deminer pinpoints the signal closest to him and places a marker where the detector starts to signal. The excavation starts usually 20 cm before the marker. The deminer prods and excavates (for example, with a trowel), gradually getting closer to the detected object. A deminer frequently investigates hundreds of detector signals before he finds a mine.

There are numerous factors influencing the detection capabilities of metal detectors. Here are some of them: the size and the shape of the target, its orientation and placement in relation to the search head, the electromagnetic properties of the target and the soil, electromagnetic interference, detector design (in particular how the detection threshold is determined and how the ground compensation is performed), detector stability, repeatability of the setup procedure, sweeping speed, and finally, the operator's skill and interpretation of the signals.

Problems of mine clearance are not just those of detection. In most cases it is easy to locate a fragmentation mine, since it is often partly above the ground and so visible to a deminer. It contains a large amount of metal, so that it is easily detectable with a metal detector if it is not visible. However, it can still pose a great danger to deminers [30]. The stakes carrying fragmentation mines may have fallen over, and tripwires may have corroded [96]. There is no metal detector that can reliably detect tripwires, even if they are still in one piece [49]. Winds may sway a bush enough to pull a tripwire and activate a mine.

Blast mines probably cause the largest number of accidents [88]. Minimum metal blast mines are difficult to detect with metal detectors due to their low metal content. They are buried close to the surface, but they can be covered with undergrowth or floodwater sediment. It is possible that they may be pushed deeper by the flails of a demining machine. Antitank mines with a non-metal case can also be difficult to detect, since they are usually placed to larger depths.

Mines of all types may have been in place for many years, their metal parts may be corroded, their cases filled with soil, and they can behave unpredictably. Often mines were laid by untrained personnel or civilians, without a predictable pattern, which makes their detection more difficult.

3.4 Reliability Model

The overall *reliability* of a mine detection system (R) can be understood as a function of three factors: an *intrinsic capability* (IC) describing the physics and the basic technical capabilities of the device and representing an upper limit of R, *factors of application* such as specific environmental conditions in the field (AP) generally diminishing R and finally the *human factor* (HF), which lowers R. All three factors are described in a concept called *reliability formula* or a *reliability model* [73, 77]:

$$R = f\left(IC, AP, HF\right) \tag{3.10}$$

The reliability of the system is indicated as a function of all influencing factors, whose mutual correlation can be very complex. The reliability formula should be understood as a compact description of the factors influencing detection, rather than a strict quantitative statement. It emphasizes the influence of the human operator on the process of detection, an influence which is often ignored or underestimated. In humanitarian demining the influence of the last two factors has already been recognised as very important, since the conditions in the field and the behaviour of the operators have proven to have a significant impact on the overall performance.

Chapter 4

Testing and Evaluation of Metal Detectors

This chapter discusses the current "state of the art" of testing metal detectors for humanitarian demining. The first section deals with the reasons for testing metal detectors. The topic of the next section are the existing standards for testing metal detectors for humanitarian demining, i.e. the CWA 14747:2003 (where CWA stands for Comité Européen de Normalisation /CEN/ Workshop Agreement) and the International Mine Action Standard IMAS. The third section gives an overview of trials performed worldwide in the last several years.

The web site of the International Test and Evaluation Program for Humanitarian Demining (ITEP) [61] offers many documents related to test and evaluation of demining equipment. All trial reports referenced in Section 4.3 are available on that web site, as well as the standards described in Section 4.2.

4.1 Purpose of Testing Metal Detectors

The focus of our interest in testing metal detectors is the determination of their detection capabilities. Some other important factors determining the quality of a detector are: battery life, ergonomics, robustness, etc. The most important factors that influence the detection capabilities of metal detectors are the target that they need to detect, the depth of the target, and the electromagnetic properties of the soil.

The purpose of testing and evaluation is to provide an assessment of the effectiveness of mine action equipment. The International Mine Action Standard "Test and Evaluation of Mine Action Equipment" IMAS 03.40 [54, 37] defines several categories of trials: concept and technology demonstration trials, development trials, acceptance trials, and consumer report. Most metal detector trials performed up to now do not fit entirely into any of

the listed categories. The purpose of most metal detector trials was to find the best device for a given range of conditions. Such trials should constitute a separate category and they could be called *comparative trials*. They are similar to acceptance trials and to a lesser degree to consumer reports, but there are important differences. The purpose of *acceptance trials*, as defined in the IMAS, is "to provide the Sponsor with sufficient information so that a decision can be taken on the acceptability of an equipment for its intended use." As we see, they do not determine the best among the acceptable devices. *Consumer reports*, according to the IMAS, "may involve a review of previous trials, tests in laboratory conditions, and some new field trials to enable a useful summary of current systems." It is not recommended to base acquirement decisions solely on consumer reports, since the equipment should be tested against the conditions of its intended use. However, consumer reports can provide an overview of the market and provide an informed guide when selecting the devices for a comparative trial.

4.2 CEN Workshop Agreement CWA 14747:2003

Although metal detectors have been used to clear landmines since the Second World War, an international standard for their testing was proposed only in 2003. That standard is the *CEN Workshop Agreement CWA 14747:2003*, where CEN stands for the European Committee for Standardisation [26, 11, 12]. The United Nations Mine Action Service (UNMAS) has referenced the CEN Workshop Agreement in its *International Mine Action Standard*, *IMAS*. The IMAS 03.40 [54, 37] provides some general guidelines to testing and evaluation of demining equipment and it refers to the CWA 14747:2003, which is much more specific.

The results of test and evaluation are more useful if the testing conforms to standard protocols and if the results are reported in a uniform manner. It is stated in the CWA 14747:2003 [26] (Section 5.2.3 of that document) that standardised procedures and tests in controlled conditions may enable comparisons of the results of tests performed at different times. However, even if the conditions are nominally the same, there are so many factors, perhaps some of them not yet known, that influence the performance of demining equipment, that it is practically impossible to recreate them all accurately. (This problem is discussed in detail in Subsections 8.8.1 to 8.8.4.) For example, let us imagine that we want to compare two metal detectors. We might decide to use the result of an earlier test on one of these detectors and to perform a new test on the other detector. Many things change in time, most obviously the weather conditions and the operators. Since an essential requirement of a good statistical design is that like be compared with like, the two detectors should be compared under comparable conditions, and that also means roughly at the same time. Standardisation enables others to

infer how useful a device would be for their own needs, and this is certainly helpful, but this is not the main benefit of standardisation, since it does not exclude the need to test the equipment under comparable conditions. It is a simple fact that two nominally identical trials have never been performed. The reasons are constant improvements of the experimental design, execution of trials and the choice of locally specific conditions. The main benefit of standardisation is facilitated sharing of the best practices in organising trials, testing, data analysing and reporting.

It is important that the devices are tested in conditions relevant for their use: on local soil, with local deminers, on local threats. An example from practice can illustrate this problem. It has been proposed to use simple geometric targets as standard targets for metal detector tests (Section 5.6 of the CWA 14747:2003 [26]). Although such "unrealistic" targets can provide some information about the performance of metal detectors, it has been shown [10] that they can help to predict only roughly how the metal detectors will react to mines.

The CWA 14747:2003 specifies procedures for: in-air tests, tests of the immunity to environmental and operational conditions, in-soil tests, and operational performance tests. In-air tests, in-soil tests and tests of the immunity to environmental and operational conditions are all based on the maximum detection height measurements (see Section 4.2.1) and they contain sensitivity profile measurements and measurements of the repeatability of the setup. The tests of the immunity to environmental and operational conditions contain tests of the influence of the search head orientation, moisture, temperature etc. on the maximum detection height. The tests that receive the most attention by detector end users are the blind in-field tests called "detection reliability tests", which are the main subject of this thesis. They are described in Section 4.2.2. The operational performance tests include the tests of: pinpointing accuracy, resolution of adjacent targets, detection near large linear metal objects, electromagnetic interference, ergonomics, robustness, and some others.

4.2.1 Maximum Detection Height Measurements

The *maximum detection height* (MDH) is a measure of the detection capability of a metal detector. The CEN Workshop Agreement [26] gives the following definition of the maximum detection height: "The maximum height above a test target at which a metal detector at given settings produces a true alarm indication due to that target."

4.2.2 Detection Reliability Tests

In blind tests, the operators using the detectors do not know the target positions. *Detection reliability tests* are the blind tests defined in Section

8.5 of the CWA 14747:2003 [26]. Among all the tests in the CWA 14747:2003, these tests include the influence of most factors that affect the performance of metal detectors, including a large part of the human factor influence. They are also the only tests that can evaluate the ability of metal detectors to deal with false alarms. In the CWA 14747:2003 detection reliability is defined 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."

In a reliability test, targets are placed in metal free lanes at positions not known to detector operators. While searching, the operators mark the places of indications and, later, supervisors measure and record the positions of the markers. A target is considered to have been detected when a marker is dropped within a prescribed radius around the true target location. The area defined with this radius is called a *detection halo*, or in this document just a *halo*. The *halo radius*, according to CWA 14747:2003, is defined as "the half of the maximum horizontal extent of the metal components in the target plus 100 mm". An indication falling inside the halo is called a *true positive* indication. A missed mine, or an absence of a marker in the halo, is called a *false negative* indication. A marker outside of the halo is classified as a *false positive* indication, or a *false alarm*.

CWA 14747:2003 makes recommendations about the lane width, soil types, target types, numbers, depths, orientation, separation and halo size and gives some practical instructions about the lane preparation.

4.3 Metal Detector Trials Performed up to the Present

Many metal detector trials have been performed in recent years. Most of these trials had the aim of choosing the most suitable detector for a certain clearance program. Numerous trials were organised by the military forces all over the world for the purposes of military demining, but the reports of those trials are rarely publicly available. The results of some non-military trials have been distributed informally, since the publicly available reports often contain little information [48].

In the mid-1990s some trials were executed in Cambodia (results not published). The series of tests was performed under the UN umbrella starting from 1997, when the United Nations Mine Action Service (UNMAS) performed trials in Sarajevo and Mostar, towns in Bosnia and Herzegovina (without publishing any results). The following UN trials were the Mine Action Programme for Afghanistan (UNMAPA) trial in Peshawar, Jalalabad and Kabul in 1999/2000 [2], the Accelerated Demining Programme (UNADP) trials in three southern provinces of Mozambique (Inhambane, Gaza and Maputo) in 1999 and 2000 [46, 47] and another Afghan UNMAPA trial

in 2002 [3]. Among these trials, only some results of the ADP and the second Afghan trials were made public. The reports of these two trials were not very detailed and the knowledge was transferred mostly informally. Each trial was conducted differently, since there was no agreed standard methodology.

A great step towards a standardised testing procedure was the International Pilot Project for Technology Co-operation (IPPTC) launched in 1998. Five research organisations and two national mine action authorities carried out tests on 29 devices under controlled conditions in laboratories and in blind field trials in two countries. Their intention was to evaluate the performance of the devices in a wide range of conditions. The results of the study were published in a consumer report in 2000 [28]. The time needed to perform and to publish the study was its main weakness. By the time the results were made available, several new detector models had appeared on the market. However, the experiences gained during that work had a crucial influence on the shaping of the standard for testing metal detectors for humanitarian demining, the CWA 14747:2003 (described in Section 4.2). Many of the tests that later entered the CWA 14747:2003 were performed during the IPPTC.

Although most of the metal detector trials performed by the military were not made public, there are three exceptions known to the author of this work. These are the Nicaraguan trial by the US Department of Defense Humanitarian Demining R&D Program in 2001 [82], the Colombian trial organised by "Defence R&D Canada" in 2002 [95], and a trial in the Netherlands organised by the Engineers Centre of Expertise of the Royal Netherlands Army in 2004 [83]. The last of these trials partially used the CWA 14747:2003 as a guideline, which had already been published in 2003. All three reports are fully transparent, describing both the testing procedure and the results in detail.

The first trial that was carried out according to the guidelines of the CWA 14747:2003 was the trial organised by the Federal Institute for Materials Research and Testing (or BAM, standing for Bundesanstalt für Materialforschung und -prüfung) in partnership with the Joint Research Centre of the European Commission (JRC). The trial was executed in 2003 and the results were published in 2004 [76, 10], together with the results of other two trials performed in the same year by BAM. Their goal was to validate the part of the testing procedure proposed in the CWA 14747:2003 dealing with detection capabilities of metal detectors. They comprised maximum detection distance measurements and detection reliability tests, which have the aim of evaluating the detection capabilities of metal detectors. Another trial organised by BAM was performed in 2005. Two articles presenting the results of this trial were published in 2006 [41, 40] and the final report [72] is expected in 2007. A detailed description of these four trials organised by BAM is given in Chapters 7 and 8, which are the core of this PhD work.

Other trials carried out after the standard CWA 14747:2003 was already

published were the Lao and Mozambican trials as a part of the STEMD project (Standardised Testing and Evaluation of Metal Detectors) [48, 50, 51, 52] organised by JRC and published in 2005, and the Croatian STEMD trial executed in October 2006 and organised by BAM. The final report of that trial is expected in 2007. The importance of the STEMD trials is that they included many commercially available metal detectors not included in the IPPTC, and that the names of the manufacturers were made public.

A comparative trial of manual demining methods [90] performed as part of the GICHD Study of Manual Mine Clearance [36] contained only a small metal detector trial. It was executed in Mozambique in 2004 and its report published in 2005, with the participation of the author (see Chapter 6).

Chapter 5

Design of Experiment

The topic of this chapter is the statistical design of experiments. The basic principles are discussed in the first section and some examples of experimental design relevant to mine detector tests are offered in the second section.

It is expected from the reader to be familiar with the main statistical concepts such as experimental error, random sample, with some distributions (normal, t, chi-square and F), with the notion of confidence intervals and with the basics of hypothesis testing. Excellent textbooks about the design and analysis of experiments were written by G. E. P. Box, W. G. Hunter and J. S. Hunter [14], and by D. C. Montgomery [71]. The first of these books includes a broader introduction to basic statistical concepts, with many examples, while the other is more compact and it offers more about the analysis of experimental data. They are both written for engineers and require no previous statistical knowledge.

5.1 Basic Principles

The *statistical design of experiments*[1] is the process of planning the experiment so that appropriate data will be collected, enabling sound and objective conclusions. A scientific approach to planning an experiment results in the highest possible efficiency of that experiment. High efficiency means that the results are unambiguous and as little affected by experimental error as possible.

When the data are subject to experimental errors, statistical evaluation is necessary. This is why all experimental problems have two aspects: design of experiment and statistical analysis of the data. If the design of experiment

[1] Besides "statistical design of experiment(s)", the terms "design of experiment(s)" and "experimental design" are used as synonyms. Most statisticians prefer to use the first two expressions. One of the reasons to use "experimental design" is to avoid double genitive, like "methods of design of experiment". In 1999 an ISO standard has been written to bring order in the terminology of this field [62]. All definitions in this work are in compliance with that standard.

is well chosen, sometimes a very elementary analysis, maybe even a visual examination of the data, will provide an answer. On the other hand, even the most sophisticated analysis cannot save the experiment if the data contain no information, as a result of the poor design of the experiment.

Experimental design methods are widely used in industry, for improving a manufacturing process or as a part of engineering design activities. The use of experimental design can result in products that have better performance and reliability, lower production costs and shorter development time.

When a measurement is repeated under, as nearly as possible, the same conditions, the observed results are never identical. The *experimental error* or simply *error* is the variation caused by uncontrolled and generally unavoidable factors. Usually only a small part of it is caused by the measurement procedure, most of it is caused by incomplete control of the experimental environment. An adequate design of experiment reduces the effect of the experimental error, which can otherwise obscure important effects.

In a designed experiment we make a difference between the *predictor variables* of a process and the *response variables*. A predictor variable is a variable that can contribute to the explanation of the outcome of an experiment. A response variable is a variable representing the outcome of an experiment. 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 *levels*. If one of the factors is of special interest, it is called a *principal factor*. The experimental goal is to investigate the influence of factors on the response variables. Another important notion is the *treatment*. Treatment is a specific combination of factor levels that appears in the experiment.

5.1.1 Replication, Randomisation and Blocking

Three basic principles of experimental design are *replication*, *randomisation* and *blocking*. These principles are briefly described in the next paragraphs and their application is explained in Section 5.2.

Replication is the performance of an experiment more than once for a given set of predictor variables. It allows the experimenter to estimate the error, which is important for determining whether the observed differences between the data are significant. More replicates allow more precise estimates of the model parameters.

Randomisation is a fundamental part of the experimental design. It means that both the allocation of the experimental material and the order of execution of measurements are determined randomly. As a result of randomisation, errors are usually independently distributed random variables, which is important for the statistical analysis. Another consequence of randomisation can be "averaging out" the effects of extraneous factors that may be present.

A block is a portion of the experimental material that is expected to be more homogeneous than the entire set of material. Blocking enables comparisons within each block. This way the variability between blocks does not affect the experimental error, so that the precision of the experiment is higher.

5.1.2 Guidelines for Designing Experiments

All participants of an experiment need to clearly understand the goal of their study, the process of data collection and at least something about how those data will be analysed. The following list was proposed by D. Montgomery [71] as a guideline to an experimenter:

1. *Recognition of and statement of the problem.* All participants of the experiment must have a clear understanding of the problem. However obvious this point might seem, the practical experience teaches us that it is often difficult to reach common understanding and a clear statement of the experimental goals.

2. *Choice of factors and levels.* The experimenter needs to determine the factors to be varied, the ranges of variations, and the specific levels to be used in the experiment. It has to be established how these factors are to be controlled and how they will be measured. At this point, but also at points 1 and 3, the non-statistical professional knowledge of the experimenter will be crucial.

3. *Selection of the response variable and the response parameter.* The response variable needs to be adequately chosen, so that it really provides useful information. It is usual to choose a mean or a standard deviation of the response variable, or both, as a response parameter.

4. *Choice of experimental design.* This step involves defining the number of replicates, the suitable order of runs, the decision about blocking or other randomisation restrictions and the choice of treatments. Simple solutions should be preferred, since they are almost always best. Some experimental designs are presented in Section 5.2.

5. *Performing the experiment.* Errors in the execution can easily destroy experimental validity. This is why careful monitoring and detailed planning are very important to success. Logistical and similar problems are often underestimated.

6. *Data analysis.* If the experiment has been designed and performed correctly, the data analysis is usually simple. Many software packages that assist in data evaluation are present on the market.

7. *Conclusions and recommendations.* The experimenter must give practical conclusions and recommend actions. Very often these recommendations will include additional tests.

Points 2, 3 and 4 form a more coherent whole; the experimenter's knowledge in design of experiment influences his choice of factor levels and response parameters, which again influence the design of the experiment.

It is usually not recommended to design a trial of a larger scope at the beginning of the study. The first experiments help the experimenter to learn more about his problem. With this new knowledge he starts a new iteration, another experiment. This is why it is said that experimentation is an iterative process.

A common misconception is that statistical methods can prove an influence of a certain factor. They cannot do that, but, when applied properly, they can give guidelines to objective and reliable conclusions. In particular, they help us to assign a confidence level to our statements or to estimate the likely error in our conclusions, thus helping the decision-making process.

5.2 Randomised Blocks, Latin Squares and Graeco-Latin Squares

5.2.1 Randomised Complete Block Design

In some experiments the influence of only one factor on a certain response needs to be evaluated. If some variability of the output comes from a known nuisance source, blocking can be an effective way to control it. As an example, suppose we want to compare the maximum detection heights of four metal detectors on the same target in air. We have decided to perform five measurements with each device to estimate the experimental error. Our principle factor is "metal detector" having four levels, and the response variable is the maximum detection height. We know from experience that the measurements performed within a short period of time will have a lower variance. Actually, we are aware that some factors like temperature, electromagnetic surroundings, etc. influence the result, but we are not interested in measuring those influences separately. Instead, we introduce a single nuisance factor called "time", with levels "day 1", "day 2" etc. All measurements in a day will be executed within an hour, a period short enough to assume that the conditions stay constant. Our goal is to remove the variability between days from the experimental error.

An experimenter not familiar with the principles of experimental design might perform repeated measurements with each detector, each day using another detector. This approach would lead to biased[2] conclusions: the

[2]While the term 'bias' sounds pejorative, it is not necessarily used in that way in statistics.

	Block 1	Block 2	Block 3	Block 4	Block 5
Treatment 1	y_{11}	y_{12}	y_{13}	y_{14}	y_{15}
Treatment 2	y_{21}	y_{22}	y_{23}	y_{24}	y_{25}
Treatment 3	y_{31}	y_{32}	y_{33}	y_{34}	y_{35}
Treatment 4	y_{41}	y_{42}	y_{43}	y_{44}	y_{45}

Table 5.1: Randomised complete block design.

differences between the detectors would not be distinguishable from the differences between days. The two factors, "metal detector" and "time", would be *confounded*.

The appropriate design is presented in Table 5.1 and it is called *randomised complete block design*. The word 'complete' indicates that each block contains all principle factor levels. In our example, blocks are days and principal factor levels are metal detectors: each day only one measurement with each metal detector will be performed. 'Randomised' indicates that the measurements within a block are executed in a random order. A result of a single measurement y_{ij} is called an *observation*. The variance between the blocks is most welcome, but the blocks should be chosen to represent realistic conditions. In tests of demining equipment, typical blocking factors are operators and time.

After the data are collected, they need to be analysed. Suppose we have a principal factor levels and b blocks. The most frequently used statistical model for the randomised complete block design is

$$y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij} \qquad \text{with} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, b \end{cases} \tag{5.1}$$

where μ is the overall mean, τ_i is the effect of the i-th level, β_j is the effect in the j-th block, and ϵ_{ij} is the random error, normally and independently distributed with the mean 0 and a variance σ^2. The level and block effects are defined so that

$$\sum_{i=1}^{a} \tau_i = 0 \tag{5.2}$$

and

$$\sum_{j=1}^{b} \beta_j = 0 \tag{5.3}$$

The analysis described in the following paragraphs is called *analysis of variance* (ANOVA). We want to test the equality of the level means. Our hypothesis is

$$\begin{aligned} H_0: &\quad \mu_1 = \mu_2 = \cdots = \mu_a \\ H_1: &\quad \text{at least one } \mu_i \neq \mu_j \end{aligned}$$

where the statement H_0 is the null hypothesis, H_1 the alternative hypothesis, and $\mu_i = \mu + \tau_i$ is the i-th level mean. It is helpful to introduce some abbreviations. Let $y_{i.}$ be the sum of all observations of level i, similarly $y_{.j}$ is the sum of all observations of block j, and $y_{..}$ is the total sum of observations. Let us call the total number of observations $N = ab$. We can write

$$y_{i.} = \sum_{j=1}^{b} y_{ij} \qquad \text{for } i = 1, 2, \ldots, a \tag{5.4}$$

$$y_{.j} = \sum_{i=1}^{a} y_{ij} \qquad \text{for } j = 1, 2, \ldots, b \tag{5.5}$$

and

$$y_{..} = \sum_{i=1}^{a} \sum_{j=1}^{b} y_{ij} \tag{5.6}$$

The corresponding averages are marked as follows:

$$\bar{y}_{i.} = \frac{y_{i.}}{b} \qquad \bar{y}_{.j} = \frac{y_{.j}}{a} \qquad \bar{y}_{..} = \frac{y_{..}}{N} \tag{5.7}$$

The expression $\sum_{i=1}^{a} \sum_{j=1}^{b} (y_{ij} - \bar{y}_{..})^2$ is called the total corrected sum of squares and it can be expanded to three terms:

$$\sum_{i=1}^{a} \sum_{j=1}^{b} (y_{ij} - \bar{y}_{..})^2 = b \sum_{i=1}^{a} (\bar{y}_{i.} - \bar{y}_{..})^2 + a \sum_{j=1}^{b} (\bar{y}_{.j} - \bar{y}_{..})^2 + \sum_{i=1}^{a} \sum_{j=1}^{b} (y_{ij} - \bar{y}_{.j} - \bar{y}_{i.} + \bar{y}_{..})^2 \tag{5.8}$$

We can write them abbreviated,

$$SS_T = SS_{Levels} + SS_{Blocks} + SS_E \tag{5.9}$$

The sum of squares due to levels, SS_{Levels}, is the measure of the differences between the level means. The sum of squares due to blocks, SS_{Blocks}, is the measure of the differences between the blocks. SS_E is the measure of the differences between the observations within a level from the level average and it is caused by random error, which is why it is called sum of squares due to error.

To calculate the sums of squares, it is easier to use totals than averages. With some simple calculations we get

$$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{b} y_{ij}^2 - \frac{y_{..}^2}{N} \tag{5.10}$$

$$SS_{Levels} = \sum_{i=1}^{a} \frac{y_{i.}^2}{b} - \frac{y_{..}^2}{N} \tag{5.11}$$

$$SS_{Blocks} = \sum_{j=1}^{b} \frac{y_{.j}^2}{a} - \frac{y_{..}^2}{N} \tag{5.12}$$

The SS_E is found by subtraction, using Equation (5.9).

It can be shown that the quantity

$$MS_E = \frac{SS_E}{(a-1)(b-1)} \tag{5.13}$$

has an expected value σ^2, which describes the experimental error.

$$E(MS_E) = \sigma^2 \tag{5.14}$$

We call this quantity, MS_E, the mean square of the error. The mean square of the levels is defined as

$$MS_{Levels} = \frac{SS_{Levels}}{a-1} \tag{5.15}$$

and its expected value is

$$E(MS_{Levels}) = \sigma^2 + \frac{b\sum_{i=1}^{a} \tau_i^2}{a-1} \tag{5.16}$$

If the level means differ, the expectation of the MS_{Levels} is larger than the expectation of the MS_E, that is, $E(MS_{Levels}) > E(MS_E)$. The test of our hypothesis can be performed by comparing these two values of mean squares. It can be shown that they are independent and each of them follows a chi-square distribution, so that their ratio follows an F distribution if the null hypothesis is valid. Therefore our test statistic is

$$F_0 = \frac{\frac{SS_{Levels}}{a-1}}{\frac{SS_E}{(a-1)(b-1)}} = \frac{MS_{Levels}}{MS_E} \tag{5.17}$$

If the null hypothesis is true, F_0 is distributed as F with $a-1$ and $N-a$ degrees of freedom. If the null hypothesis if false, the expected value of F_0 is greater than 1. We reject the null hypothesis if

$$F_0 > F_{\alpha,a-1,(a-1)(b-1)} \tag{5.18}$$

where α is the significance level. We can test the hypothesis against a given significance level, for example $\alpha = 0.05$, but it is more informative to state the significance level at which the hypothesis is rejected, saying, for example, "H_0 is rejected with the p-value $\alpha = 0.0034$".

The whole procedure is summarised in Table 5.2 and the method is called *analysis of variance*, or by many authors *ANOVA*.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	MS is the Estimator of	F_0
Levels	$SS_{Levels} = \sum_{i=1}^{a} \frac{y_{i.}^2}{b} - \frac{y_{..}^2}{N}$	$a-1$	$\frac{SS_{Levels}}{a-1}$	$\sigma^2 + \frac{b\sum_{i=1}^{a}\tau_i^2}{a-1}$	$\frac{MS_{Levels}}{MS_E}$
Blocks	$SS_{Blocks} = \sum_{j=1}^{b} \frac{y_{.j}^2}{a} - \frac{y_{..}^2}{N}$	$b-1$	$\frac{SS_{Blocks}}{b-1}$	$\sigma^2 + \frac{a\sum_{j=1}^{b}\beta_j^2}{b-1}$	$\frac{MS_{Blocks}}{MS_E}$
Error	$SS_E = SS_T - SS_{Levels} - SS_{Blocks}$	$(a-1)(b-1)$	$\frac{SS_E}{(a-1)(b-1)}$	σ^2	
Total	$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}^2 - \frac{y_{..}^2}{N}$	$N-1$			

Table 5.2: Analysis of variance (ANOVA) for a randomised complete block design.

The test statistic

$$F_0 = \frac{MS_{Blocks}}{MS_E} \tag{5.19}$$

is also important, because it gives us feedback about the usefulness of blocking. In the case when no significant difference between blocks is detected, it would be better to design an experiment without blocks. The reason is that blocking reduces the number of degrees of freedom for the SS_E, thus increasing the error.

The model of the Equation (5.1) is linear and hence completely additive. This means that, for instance, the first block increases the observations in all levels for the same amount, namely β_1. In many cases such a model is useful, but there are situations when it is inadequate, for example, if one of the blocks strongly affects only one level, but not the others. In such a case we say there is an *interaction* between the blocks and the levels. An interaction generally inflates the experimental error and thus affects the comparison of level means. If both factors and their interactions are of interest, factorial designs have to be used.

After the experimenter finds out that there are significant differences between the levels, he or she is usually interested in multiple comparisons to discover which level means differ. There are several available methods, but the most powerful ones are the Duncan's multiple range test and the least significant difference (LSD) test [71]. The LSD test is described in the next lines.

Let us suppose we want to test $H_0 : \mu_i = \mu_j$. The pair of means μ_i and μ_j would be declared significantly different if

$$|\bar{y}_{i.} - \bar{y}_{j.}| > t_{\frac{\alpha}{2},N-a}\sqrt{\frac{2MS_E}{b}} \tag{5.20}$$

The right hand side is called the *least significant difference*. It also defines the confidence interval for this specific difference:

$$(\bar{y}_{i.} - \bar{y}_{j.}) \pm t_{\frac{\alpha}{2},N-a}\sqrt{\frac{2MS_E}{b}} \tag{5.21}$$

5.2.2 Latin Square Design

Suppose that an experimenter wants to evaluate the influence of a certain factor on a certain response variable and that he wants to include the influence of two nuisance factors in his evaluation. The appropriate design is the *Latin square design* presented in Table 5.3. The principal factor levels are denoted with Latin letters A, B, ..., while the columns and rows represent the two nuisance factors. Each column and each row can be understood as a block. Principal factor levels are orthogonal to both rows and columns, which means that each letter appears once in each row and in each column.

A	D	B	E	C
D	A	C	B	E
C	B	E	D	A
B	E	A	C	D
E	C	D	A	B

Table 5.3: An example of a 5×5 Latin square.

The statistical model for the Latin square design is

$$y_{ijk} = \mu + \alpha_i + \tau_j + \beta_k + \epsilon_{ijk} \qquad \text{with} \begin{cases} i = 1, 2, \ldots, p \\ j = 1, 2, \ldots, p \\ k = 1, 2, \ldots, p \end{cases} \tag{5.22}$$

where y_{ijk} is the observation in the i-th row, k-th column, for the j-th level, μ is the overall mean, α_i and β_k are the i-th row effect and the k-th column effect, τ_j is the j-th level effect, and ϵ_{ijk} is the random error. Any two indexes of an observation y_{ijk} determine the third one, since each level appears only once in each column and row. Therefore, the total number of observations is $N = p^2$. If all factor level combinations would be present, the total number of observations would be p^3. Like the model for the randomised complete block design (Equation (5.1) on page 51), this model is completely additive, which means that there is no interaction between the rows, columns and levels. Similarly as with the randomised complete block design described in the previous section, we perform the analysis of variance. The procedure is described in [14] and [71].

5.2.3 Graeco-Latin Square Design

In the section about the randomised complete block design we considered only one nuisance factor, while the Latin square design dealt with two nuisance factors. We continue to increase the number of nuisance factors and we ask ourselves what is the best design if there are three nuisance factors influencing our result. The solution is in superimposing two Latin squares. The procedure is illustrated in Table 5.4. The second Latin square is written in Greek letters and superimposed on the first one. If the two squares have the property that each Latin letter appears once and only once with each Greek letter, we say that the two squares are *orthogonal* and the design obtained is called a *Graeco-Latin square*. This design allows investigations of four factors, each with p levels, in only p^2 runs.

Latin square 1				Latin square 2				Graeco-Latin square			
A	B	C	D	A	B	C	D	$A\alpha$	$B\beta$	$C\gamma$	$D\delta$
B	A	D	C	D	C	B	A	$B\delta$	$A\gamma$	$D\beta$	$C\alpha$
C	D	A	B	B	A	D	C	$C\beta$	$D\alpha$	$A\delta$	$B\gamma$
D	C	B	A	C	D	A	B	$D\gamma$	$C\delta$	$B\alpha$	$A\beta$

Table 5.4: An example of a 4×4 Graeco-Latin square.

The statistical model for the Graeco-Latin square design is

$$y_{ijkl} = \mu + \alpha_i + \tau_j + \omega_k + \beta_l + \epsilon_{ijkl} \qquad \text{with} \begin{cases} i = 1, 2, \ldots, p \\ j = 1, 2, \ldots, p \\ k = 1, 2, \ldots, p \\ l = 1, 2, \ldots, p \end{cases} \tag{5.23}$$

where y_{ijkl} is the observation in the i-th row and l-th column for Latin letter j and Greek letter k, μ is the overall mean, α_i is the effect of the i-th row, τ_j is the effect of Latin letter j, ω_k is the effect of Greek letter k, β_l is the effect of column l, and ϵ_{ijkl} is the random error. Two subscripts identify an observation, e.g. a row and a column identify the Latin letter and the Greek letter. The total number of observations is only p^2. If all factor level combinations would be present, the total number of observations would be p^4. The analysis of variance of the Graeco-Latin square design is an extension of the Latin square analysis of variance and it is described in [14] and [71].

This concept of combining orthogonal Latin squares can be extended. A $p \times p$ *hypersquare* is a design with three or more superimposed $p \times p$ orthogonal Latin squares. The highest possible number of superimposed Latin squares is $p - 1$, because a design with $p - 1$ Latin squares would utilise all $p^2 - 1$ degrees of freedom, so that an independent estimate of the error variance σ^2 would be necessary.

Chapter 6

Comparative Trial of Manual Mine Clearance Methods

A trial of manual mine clearance methods [90] took place in Moamba in Maputo Province, Mozambique, at the training centre of ADP (Accelerated Demining Programme), in November 2004, as a part of the Study of Manual Mine Clearance [36] managed by the Geneva International Centre for Humanitarian Demining. This chapter deals with some aspects of the trial mostly concerning the estimate of the experimental error and the problem of bias.

The goals of the trial and the testing procedure are described in the introductory section. The next section presents the results, followed by a discussion. The chapter ends with conclusions, including some recommendations for future tests.

6.1 Introduction

The primary goal of the trial of manual demining methods performed in Mozambique in 2004 was to compare the speeds of seven manual demining methods. Three of them include a metal detector, while the other four were excavation methods during which the top layer of the soil was removed from the entire lane (see [90] for a complete description of the demining methods and the tools). Some metal fragments were deliberately placed on the lanes to simulate a difficult scenario for metal detectors. A part of the test area labelled "method 4" was kept almost free from metal to provide a benchmark. The demining procedures of methods 1 and 4 were identical, the only difference being the metal contamination of the lanes. Table 6.1 is an overview of the tested methods.

Another goal of the trial was to compare the results of a detection reliability test with the results of a test with full excavation of the targets. This test was performed on the lanes with a minimum metal contamination

Method	Description	Team
1	metal detector with standard tools	A
2	metal detector + magnet attached to a trowel	B
3	metal detector + magnet attached to a brush-rake	B
4	metal detector with standard tools (no metal fragments)	B
5	excavation with rakes	C
6	excavation with a spade	A
7	excavation with a trowel	D
8	excavation with a mattock	B

Table 6.1: List of all methods tested in Mozambique trial.

labelled "method 4". The positions of the deminers' indications were compared with the target positions and the halo radii. All indications falling inside the halo radii were counted as true positives, while the other were counted as false alarms.

Four 5-m lanes were prepared for each method. Six metal fragments per square metre were placed on random positions to depths between 0 and 1 cm. Some targets simulating mines were placed in the lanes to enable the evaluation of the thoroughness of the excavation methods. The simulated targets were GYATA-64 and Type 72. The clearance was performed by professional deminers from several demining organisations working in Mozambique. Four teams participated in the tests, each consisting of two deminers and a section leader controlling their work. Each deminer cleared 10 m, if his speed was sufficient to complete these 10 m in three days, otherwise the test was interrupted. For that reason, the test of methods 5 and 7 was not completed. Team A executed the clearance with methods 1 and 6, team B with methods 2, 3, 4 and 8, team C with method 5 and team D with method 7, as indicated in Table 6.1. The deminers wore their personal protective equipment, because it has a large influence on their comfort and consequently on their speed. It might also have had some positive influence on their concentration, since it is a part of their daily routine. The section leaders were present during the whole trial and they controlled the work of the deminers. The involvement of section leaders and the protective equipment was important because the goal was to test the whole demining system, not just the tools.

6.2 Results

For a reliable comparison between the methods, some confidence limits need to be attributed to the average speed of each method. Each lane was di-

vided to four or five areas of similar size (1-1.5 m) and these areas are called segments. The time required for each segment was measured without interrupting the work of the deminers. The time required to clear a segment of a certain area divided with the size of that area is a reciprocal of speed.[1] The measurement unit of the reciprocal of speed is $\mathrm{min/m^2}$. We use the standard deviation of the segments' reciprocal speeds to estimate the experimental error of the reciprocal speed for each method. We assume that the reciprocal values of the segment speed are roughly normally distributed to construct 95% confidence limits. The results of the experiment, i.e. the reciprocal speeds for each segment and the averages for each method are presented in Table 6.2. Figure 6.1 is a graphical representation of the overall results. (The figure was created in Microsoft Excel 2002.)

As expected, the results of the reliability test and the test with a full investigation of the signal were very similar. The only significant difference was with the GYATA-64. The signal of the original mine, as well as that of the surrogate used in this test, has two local maxima, so that it appears to the deminer that there are two point-like metallic objects under the soil surface. Two out of eight deeply buried GYATAs were narrowly missed in the reliability test because both signal maxima fell just outside the detection halo. The same deminers who closely missed these two targets in a reliability test found the targets in the full investigation test.

6.3 Discussion

When comparing the speeds or the reciprocals of speeds of the tested methods, we should keep in mind that the results are highly biased. The strongest source of bias is the confounding between the methods and the deminers (see Table 6.1). The individual differences between the deminers, although not measured, were obvious during the trial, and they certainly influenced the outcome. The methods were not tested simultaneously, so that there is confounding between the methods and time. The influencing variable "time" actually encompasses the influences of many variables not explicitly considered, like weather, experience of the personnel etc.

Nevertheless, the differences between the estimated speeds of some methods were so high, that there is no doubt about the actual difference between

[1] The reciprocals of speed are easier to analyse than the speeds. The speed of each method cannot be estimated with the average of the segment speeds, since the segments are approximately equally large. If we have N segments and if x_i and t_i are the segment area and the clearance time for the segment i, then the average segment speed is $\frac{1}{N}\sum_{i=1}^{N}\frac{x_i}{t_i}$, while the estimated speed is $\frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} t_i}$. These expressions are equal only if the times t_i are equal. However, the average reciprocal of speed, $\frac{1}{N}\sum_{i=1}^{N}\frac{t_i}{x_i}$, is approximately equal to the estimated reciprocal of speed, $\frac{\sum_{i=1}^{N} t_i}{\sum_{i=1}^{N} x_i}$. Only approximately, because the segments are only approximately equal.

		method							
		1	2	3	4	5	6	7	8
lane 1	segment 1	60	11.7	11.6	20.7	107	69	104	22
	segment 2	28.7	9	10.5	12.7	105	43	242	23
	segment 3	57.7	49	13.1	6.09	145	47	230	23
	segment 4	79.1	21	2.73	13.6	44.6	53		29
	segment 5					79.5	46		31
lane 2	segment 1	138	16.5	11.8	19.7	110	53		26
	segment 2	31.2	10	8.18	29.2	145	29		23
	segment 3	20.8	7.33	18.5	15.6		36		35
	segment 4	36.8	11.3	18.6	9.03		28		24
	segment 5						23		13
lane 3	segment 1	135	57.9	12.2	32.4	138	70	159	56
	segment 2	37.3	13.1	21.3	15.1	67	85	142	54
	segment 3	25	28.9	16.9	3.1	80.4	49	74.4	41
	segment 4	17.5	32.9	15	6.07	96.4	37	101	22
	segment 5						36	109	30
lane 4	segment 1	42.7	40	22.5	16.5		49		32
	segment 2	54.5	49.2	15	11.5		56		25
	segment 3	17.2	12.7	38.2	7.8		38		27
	segment 4	35.8	7.27	17.3	10.3		40		32
	segment 5						22		34
average		51.1	23.6	15.8	14.3	102	45.5	145	30.1
standard deviation		37.4	17.1	7.78	8.14	32.5	16.1	61.8	10.4
estimated error		19.9	9.1	4.1	4.3	22	7.6	52	4.9
quality assurance		4	2	4	4	4.5	0.4	0	0
average + quality assurance		55.1	25.6	19.8	18.3	106	45.9	145	30.1
measurement unit: $\mathrm{min/m^2}$									

Table 6.2: Trial of manual demining methods, results. Each lane was divided to four or five segments and the reciprocal of speed is given for each of these segments, measured in $\mathrm{min/m^2}$. The row "estimated error" refers to 95% confidence limits based on the assumption of normal distribution. The time spent on quality assurance performed by the section leader is added to form the total result.

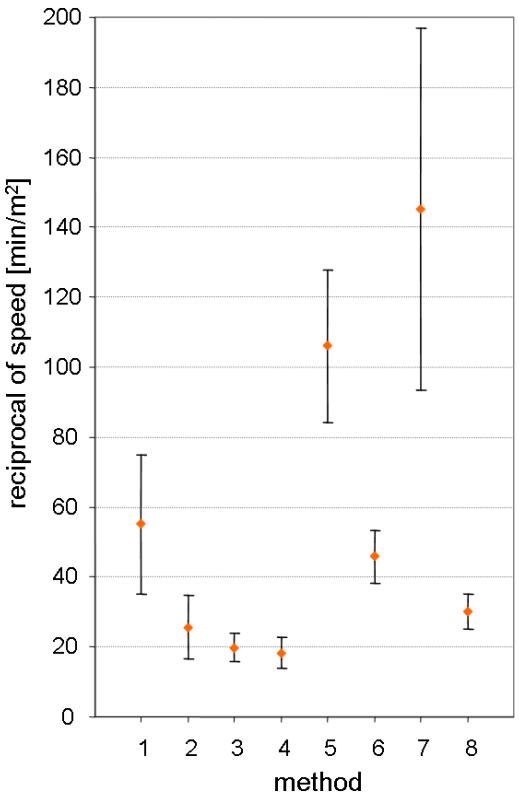

Figure 6.1: Trial of manual demining methods, results, overview. The data include the time spent on quality assurance. The error bars indicate 95% confidence limits based on the assumption of normal distribution.

some methods. A comparison of the methods tested by the same deminer team is less subject to bias. Thus we can see, for example, that the use of magnets (methods 2 and 3) significantly improved the performance of deminers, making it comparable with their performance in lanes without any metal fragments (method 4).

It is possible to perform a Duncan's multiple range test or an LSD test (see Section 5.2.1) to compare the methods, however, the comparison would be unreliable and could be misleading due to high bias present in the trial results. The bias is high because not all methods were tested with the same deminer team; in other words, because the teams and the methods were confounded (see Table 6.1). The presence of bias also justifies the approximations used in the data analysis and mentioned in the previous section.

The results of the reliability test indicate no significant difference between the overall results of the reliability test and the full investigation test. A similar conclusion follows from another test with a full investigation, performed in Oberjettenberg in November 2003 and described in Section 8.5. There is, however, a difference in the results for the GYATA-64. The problem of indications falling very close to the detection halo can be easily solved by choosing a slightly larger halo radius.

The presence of section leaders had a marked influence on the deminers. This conclusion is based on a comparison with the experiences from the Oberjettenberg trials and the Benkovac 2003 trial described in Chapter 8, when no section leaders were included in the test.

6.4 Conclusions

The results of this trial are highly biased mostly due to confounding between the tested methods and operators. Nevertheless, the large differences between the results indicate that some methods are significantly faster than others.

Future tests should be organised and planned so that all methods are tested in the same conditions. They should be tested by the same personnel, on the similar ground and under the same weather conditions. It is not possible to assure that all conditions are the same for all tested methods, but it is possible to minimise the impact of the uncontrolled factors by randomising the design. For example, selecting adjacent lanes for testing one of the methods could lead to bias, since the soil properties might be different on that area than on the rest of the testing area. The assignment of lanes to the methods should be random. All these requirements are a serious challenge for the organiser of the trial, but they have to be fulfilled in order to reach unbiased conclusions.

For testing detection abilities of metal detectors, a full investigation of

each signal with an immediate excavation of the targets is not necessary. It is very time consuming and eventually it provides the same result as a reliability test, provided that the work of the deminers in a reliability test follows a procedure similar to their standard operating procedures applied in their daily work.

Chapter 7

Maximum Detection Height Measurements

This chapter deals with the measurements of the maximum detection height[1] performed during the metal detector trial in Croatia, May 2005. These measurements include the first attempt to estimate the variability of maximum detection height measurements. A short introduction presents the notion of the maximum detection height. The topics of the next two sections are the experimental design of the maximum detection height measurements and the corresponding data analysis. The results are presented, interpreted and discussed in the next two sections, followed by a conclusion.

7.1 Introduction

Measurements of the maximum detection height are used to estimate the detection capabilities of metal detectors. The *maximum detection height* (MDH) 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 audio signals [26]. The position of the target is clearly marked and known to the operator. In all earlier investigations it has been conjectured that repeated measurements give very similar results. The experiment described in this work had been designed to check this conjecture of the stability of metal detectors and to investigate the possible sources of variability.

The main difference between the maximum detection height measurements and the blind trials is that the operators know the positions of the targets during the maximum detection height measurements. Therefore the influence of the operator, i.e. the "human factor" in the reliability model (see Section 3.4), is much smaller than in the blind trials. However, it is

[1] *Maximum detection height* is sometimes called *maximum detection depth* or *maximum detection distance*. Regarding terminology, this work follows the recommendations of the CWA 14747:2003 [26].

not entirely excluded: the operator sets up the metal detector to the required sensitivity, performs the ground compensation procedure, operates the device and decides whether the audio signal is clear enough to call it detection. In this respect metal detector models can be very different. The setup procedure of some models minimises the influence of the operator and the audio signal of some models is very clear, while some detectors require more actions and decisions from the operator.

7.2 Design of Experiment

During the trials in Benkovac in May 2005 (see Section 8.6 and [72] two series of maximum detection height (MDH) measurements were performed: the first one during the two weeks of the reliability test and the second one within a two-day period after that test. The goals of the maximum detection height measurements were:

1. To assess the variability of the maximum detection height measurements.

2. To compare the detecting capabilities of four metal detectors in two soil types separately.

3. To compare the surrogate of the PMA-2 with the real mine, using the in-air measurements.

4. If the surrogate faithfully represents the real mine, to use it for comparing the influence of the two soils used in the experiment.

The first measurement series was based on the design of the reliability test, which is described in Section 8.6.2, Table 8.11. During that series each deminer performed the MDH measurement just before each start of the reliability test, with the detector scheduled for that start, in the corresponding soil. This way each deminer tested each detector in each soil exactly once. The original intention was to use only one target type, PMA-2, in both soils. Unfortunately, only a limited number of these targets was available, so that surrogates of PMA-2 were buried in one of the soils. There were two target-soil combinations in this first series of measurements: PMA-2 in soil 1 and PMA-S in soil 2. PMA-S is the surrogate of the PMA-2 (see Figure 7.1), soil 1 is the uncooperative Obrovac soil present in lanes 1 and 2 of the test site and soil 2 the cooperative Sisak soil of lanes 3 and 4, where the numbering of the lanes follows the numbering used in the reliability trials described in Section 8.6.2. For example, operator *D* performed the MDH measurement in soil 1 with detector *delta* just before start 3 of the reliability trial, according to Table 8.11. The order of execution within a start was not random, but dictated by organisational requirements. In most cases

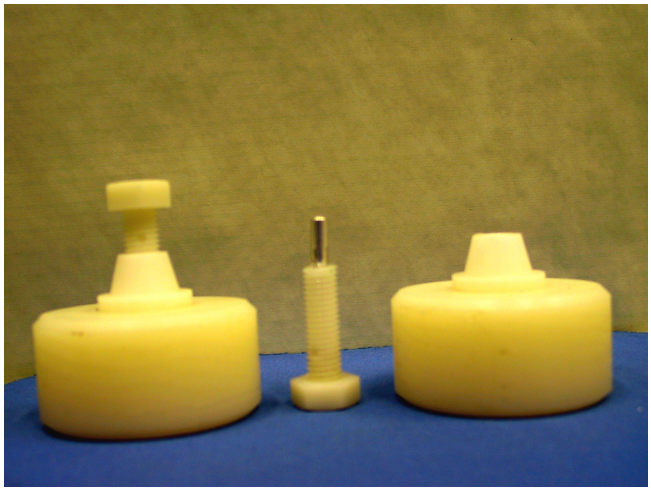

Figure 7.1: PMA-S, a surrogate of the PMA-2. The surrogate on the right side of the image is taken apart to make the metal content visible.

two deminers started earlier than the other two due to interference between detectors.

The second measurement series was performed according to the design presented in Table 7.1. All levels of all factors are combined in a *full factorial design*. This means that all combinations operator-detector are present in this design. MDH measurements were performed on the two combinations target-soil from the first series (see the last two columns of Table 7.1) and on three targets in air. The third target for the in-air measurements was the antipersonnel mine PMA-1A. A careful examination of the designs of both measurement series reveals that the same measurements were performed in both series, just in a different order. In addition, the second series included some in-air measurements. The numbering of treatments indicates the order of execution of the measurements, i.e. the measurements were performed starting with the 1st treatment and ending with the 16th. (The only exception was treatment 6. There was a mistake in the setup of the metal detector, so that this measurement was repeated after treatment 16 and only the result of the repeated measurement was counted as valid.) Within each treatment, the in-air measurements were performed before the in-soil measurements, numbers 1-3 indicating the order of execution. For example, treatment 10 was executed after treatment 9 and the measurements were performed in this order: PMA-2 in air, PMA-S in air, PMA-1A in air, PMA-2 in soil 1, and finally PMA-S in soil 2. This way the in-air measurements on the PMA-2 and the PMA-S were always performed one after the other, in a random order. As can be seen from the table, the order of treatments 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 and all operators equally. The measurements with treatments 1-8, except treatment 6, were performed 27 May 2005 11:45-12:45, while the other measurements, treatments 6 and 9-16, were performed 30 May 08:30-09:30.

The targets were placed on a top of a wooden board and buried together with the board, which eased the control of their depths during the burial and even afterwards (see Figure 7.2). There was no indication that the presence of the board influenced the performance of the metal detectors. The targets were buried to depths 3, 4, 5, ..., 11 cm in soil 1 and 6, 7, 8, ..., 15 cm in soil 2. Their mutual distance was 50-60 cm. After the burial their positions were clearly marked. The area prepared for the MDH measurements is illustrated on Figure 7.3.

For the measurements of each treatment the deminer performed the detector setup and the soil compensation. The in-soil measurements were performed after a ground compensation, while the in-air measurements were performed without ground compensation. The in-soil measurements consisted of the following procedure: the deminer checked the detectability of each target by moving the search head as close as possible to the ground without touching it. His indications ('detected' or 'not detected') were recorded for each depth. The in-air measurements were performed as indicated on Figure 7.4. The targets were repeatedly moved horizontally sideways with the help of a board, thus simulating the sweeping of the metal detector search head. The deminers made the decision whether the target is detected or not. The results were recorded to the closest centimetre at which there was a signal.

7.3 Data Analysis

7.3.1 Comparison of Detectors

The main goal of maximum detection height measurements is to detect differences between the detectors. Let us consider the measurements for each target-soil combination separately. An important assumption necessary for the analysis of variance (see subsection 5.2.1) is that the experimental error is normally and independently distributed. There are good reasons to suspect that the error of MDH measurements depends on the detector model, which is why the analysis of variance cannot be used. (The results, Figure 7.5 in Section 7.4, confirm the validity of this approach.) If the result

			In air			In soil 1	In soil 2
Treatment	Operator	Detector	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 7.1: Design of the maximum detection height measurements, Benkovac, May 2005.

Figure 7.2: Placement of targets before their burial for maximum detection height measurements. The boards help to control the depth of the targets.

Figure 7.3: Area fully prepared for the maximum detection height measurements. The target positions and their depths are clearly indicated with red markers.

Figure 7.4: In-air maximum detection height measurements. The target is being moved, while the search head is kept still.

of a single MDH measurement is denoted with y_{ijk}, it can be written

$$y_{ijk} = \mu + \tau_i + \beta_j + \rho_k + \epsilon_{ijk} \qquad \text{with} \begin{cases} i = 1, 2, \dots a \\ j = 1, 2, \dots b \\ k = 1, 2, \dots c \end{cases} \tag{7.1}$$

where μ is the overall mean, τ_i is the effect of the i-th detector, with $a = 4$, β_j is the effect in the j-th operator, $b = 4$, ρ_k is the effect of the k-th repetition, $c = 2$ for in-soil measurements and 1 for in-air measurements, and ϵ_{ijk} is the random error, normally distributed with the mean 0 and a variance σ_i^2 depending on the detector model i. We assume that all y_{ijk} for the i-th detector are distributed normally, the parameters of this distribution depending on the detector. The variance of the results for a particular detector i is

$$\frac{\Sigma_j \Sigma_k y_{ijk}^2 - \frac{{y_{i..}}^2}{bc}}{bc - 1}$$

We want to compare a pair of detectors i and l on each of the soil-target combinations separately. We want to test the hypothesis H_0: $\mu_i = \mu_l$, where μ_i and μ_l are the means of the MDH for detectors i and l. The pair of means would be declared significantly different if

$$|\bar{y}_{i..} - \bar{y}_{l..}| > t_{\frac{\alpha}{2},\nu} \sqrt{\frac{S_i^2}{n_i} + \frac{S_j^2}{n_l}} \tag{7.2}$$

with

$$\nu = \frac{\left(\frac{S_i^2}{n_i} + \frac{S_l^2}{n_l}\right)^2}{\frac{(S_i^2/n_i)^2}{n_i - 1} + \frac{(S_l^2/n_l)^2}{n_l - 1}} = (bc - 1)\frac{\left(S_i^2 + S_l^2\right)^2}{S_i^4 + S_l^4}$$

where $\bar{y}_{i..}$ and $\bar{y}_{l..}$ are the MDH averages, $t_{\frac{\alpha}{2},\nu}$ is the quantile of the Student's t-distribution, $n_i = n_l = bc$ is the sample size, $b = 4$ is the number of operators, $c = 2$ the number of repetitions, and S_i^2 and S_l^2 are the sample variances of the measurements with detectors i and l respectively. The right hand side of Equation (7.2) also defines the confidence interval for this specific difference:

$$(\bar{y}_{i..} - \bar{y}_{l..}) \pm t_{\frac{\alpha}{2},\nu}\sqrt{\frac{S_i^2}{n_i} + \frac{S_l^2}{n_l}} \tag{7.3}$$

We construct confidence intervals on the mean of each detector, to express the uncertainty of our estimates of the mean:

$$\bar{y}_{i..} \pm t_{\frac{\alpha}{2},bc-1}\frac{S_i}{\sqrt{bc}} \tag{7.4}$$

The estimated MDH for the i-th detector is $\bar{y}_{i..}$ increased by 0.5 cm, since the targets were buried in 1-cm steps:

$$\hat{y}_i = \bar{y}_{i..} + 0.5 \text{ cm} \tag{7.5}$$

7.3.2 Comparison between PMA-2 and PMA-S

Let us investigate the problem of comparing the PMA-2 with the PMA-S. The maximum detection height measurements of these two targets in air can be observed as paired measurements. Each combination operator-detector can be understood as a block (see Section 5.1.1) and the two targets can be understood as two levels of the principal factor "target". Each combination operator-detector-target represents a treatment. We want to test the null hypothesis that the difference between the maximum detection heights of the two targets is zero. If we name the measured maximum detection heights of the two targets y_{1i} and y_{2i}, where i denotes the i-th block, than the measured difference in that block equals

$$d_i = y_{1i} - y_{2i} \tag{7.6}$$

This variable follows a t-distribution with $n - 1$ degrees of freedom, where $n = ab$ is the number of blocks (in our case, $n = 16$). Therefore our test statistic is

$$t_0 = \frac{\bar{d}}{S_d/\sqrt{n}} \tag{7.7}$$

where $\bar{d} = \Sigma_{i=1}^{n} d_i/n$ is the average difference, and S_d is the sample standard deviation of the difference d_i. We reject the null hypothesis with significance level α if

$$\left|\bar{d}\right| > t_{\frac{\alpha}{2},n-1}S_d/\sqrt{n} \tag{7.8}$$

7.4 Results

This section presents the results of the two series of maximum detection height measurements described in the previous section.

It occurred during the in-soil measurements that a certain target was detected at depths, for example, 6, 7, 8 and 10 cm and not detected at 9, 11, 12 etc. Such cases were treated the same as if the target had been detected on all depths up to 10 cm, including on 9 cm.

Separate results for each target and each detector with the corresponding standard deviations are presented on Figure 7.5, containing both measurement series. Each column presented in the diagram is an average of four in-air measurements or eight in-soil measurements. The same results, but with confidence intervals instead of standard deviations, are presented on Figure 7.6 (see Equation (7.4)). The standard deviations give us some information about the variability of the MDH measurements, while the confidence intervals indicate the uncertainty of our estimates of the mean MDH. (All data analyses in this chapter were performed with Microsoft Excel 2002.)

We can compare each pair of detectors using Equation (7.2). The results are presented in Table 7.2. The numbers 1 and 0 indicate whether the difference is statistically significant with the significance level $\alpha = 0.05$. For example, for the in-air measurements with the PMA-1A there is a significant difference between detectors X and Y and also between Y and Z, while all other differences are not significant.

	PMA-2 soil 1			PMA-S soil 2			PMA-1A in air			PMA-2 in air			PMA-S in air		
	X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z
U	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
X		1	1		1	1		1	0		0	0		0	0
Y			0			0			1			1			1

Table 7.2: Results of the multiple comparisons of the maximum detection heights. Number 1 indicates a significant difference at significance level $\alpha = 0.05$.

To learn more about the possible sources of variability, we perform the analysis of variance for each detector-soil combination separately. There are two factors: operator, with four levels, and repetition, with two levels. The results, Table 7.3, indicate that the difference between the operators is greater than the difference between the two repetitions.

The targets PMA-2 and PMA-S are compared using the in-air measurements as described in Section 7.3.2, Equation (7.8). The results indicate that the maximum detection height of the PMA-2 is (9.4±6.6) mm larger

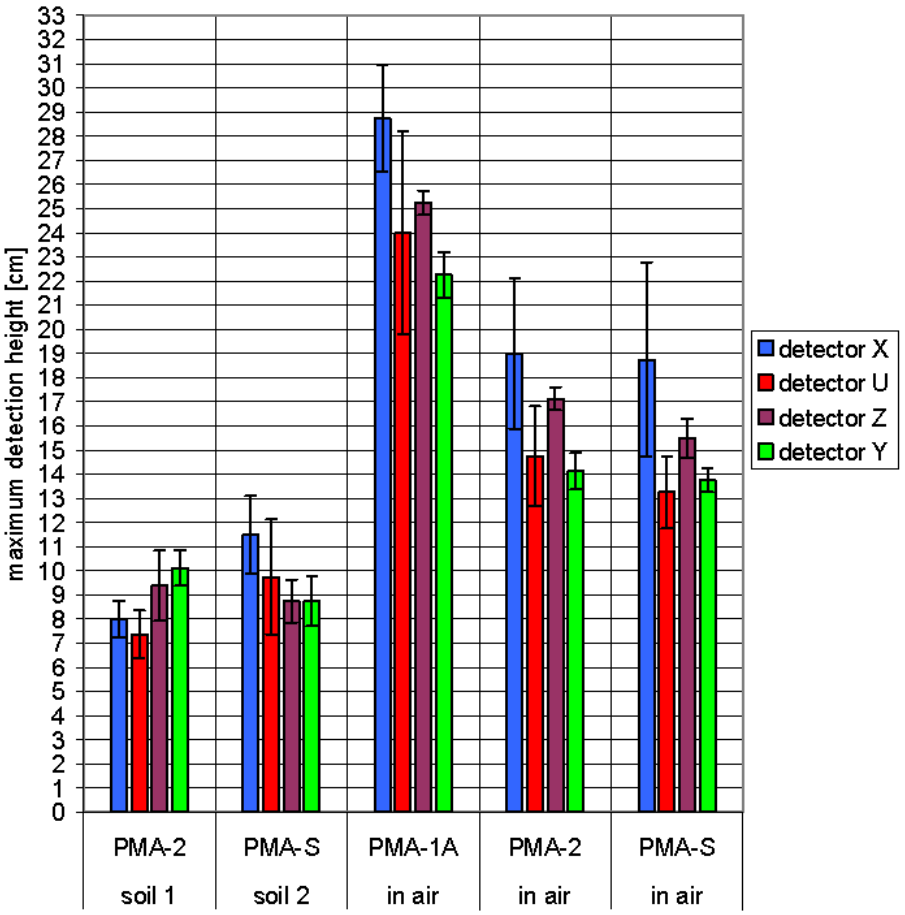

Figure 7.5: Results of the maximum detection height measurements with standard deviations.

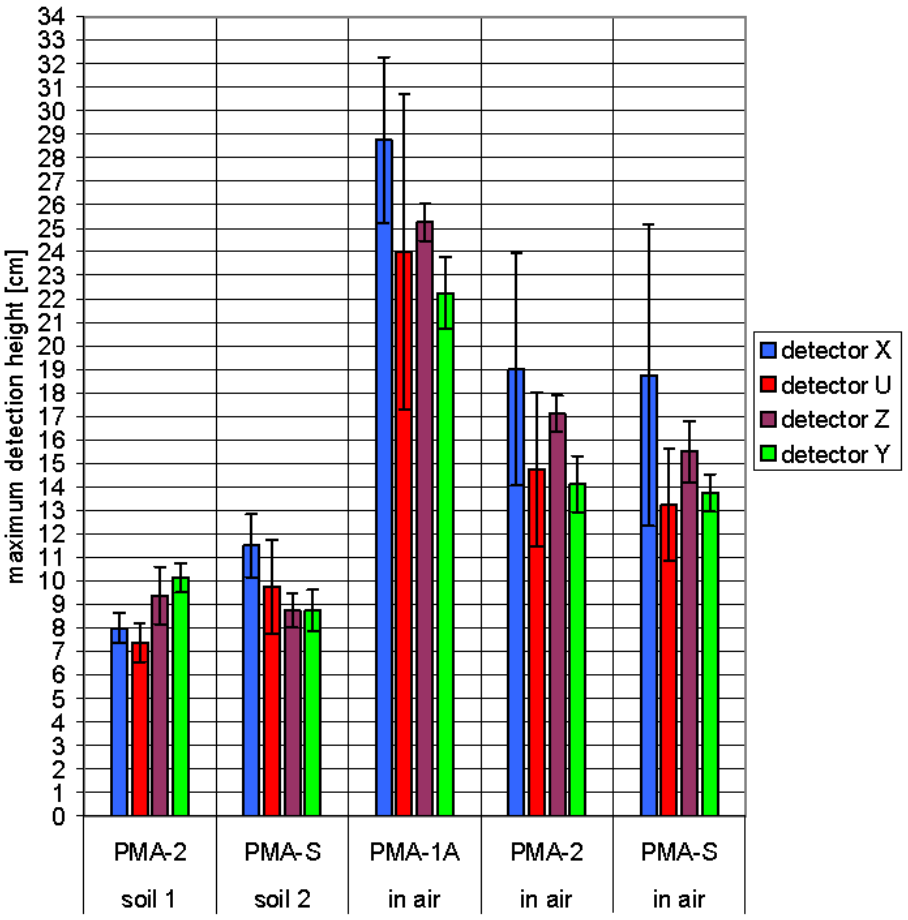

Figure 7.6: Results of the maximum detection height measurements with 95% confidence intervals.

PMA-2 in soil 1:

	detector U	detector X	detector Y	detector Z
MS(operators)	1.792	1	1.125	2.125
MS(repetitions)	1.125	0.5	0.125	1.125
MS(residual)	0.125	0.167	0.125	2.458
$\frac{\text{MS(operators)}}{\text{MS(repetitions)}}$	1.593	2	9	1.889

PMA-S in soil 2:

	detector U	detector X	detector Y	detector Z
MS(operators)	2.458	3	0.5	0.833
MS(repetitions)	1.125	0.5	0.5	0.5
MS(residual)	3.125	2.833	1.833	0.833
$\frac{\text{MS(operators)}}{\text{MS(repetitions)}}$	2.185	6	1	1.667

Table 7.3: Results of the analysis of variance for the in-soil maximum detection height measurements. The mean squares of the operators, repetitions and residuals.

than that of its surrogate, where the numbers mark the 95% confidence limits. The null hypothesis that the two targets are identical is rejected with the p-value $\alpha = 0.008$.

7.5 Discussion

7.5.1 Variability of Maximum Detection Height Measurements

The results (Figure 7.5) clearly show that comparing two single measurements of MDH can not give a reliable answer about the difference between two detector models. MDH measurements have a high variance that cannot be ignored. The causes of such a high variability are discussed in this subsection.

The ground compensation procedure includes some actions of the operator. Let us examine if the soil compensation is the main source of variability. The in-air measurements were performed without ground compensation. The in-air standard deviations for two of the detectors, U and X, are much larger than the in-soil standard deviations, while the standard variations for the other two detectors are similar in air and in soil. Therefore, it can be concluded that the soil compensation procedure is not the main source of variability.

The analysis of variance (see Table 7.3) showed no important differences

between the averages of the first and the second series of measurements (the first and the second repetition). The main source of the variability in this experiment was therefore not the repetition.

The variability was caused by the differences between the operators, by the setup, and by the remaining sources of the experimental error, which are the changing subjectivity of the operators, the short-term instability of the hardware of the devices, and the uncertainty of the measurements of the distance between the search head and the target. From the results of this experiment we cannot conclude which of these influences was dominant. For example, it is not possible to separate the error coming from the subjectiveness of the operator from the instability of the devices, because the devices were always setup and operated by human beings. Some tests with automated scanning and automated ground compensation could give more insight into the main cause of the experimental error.

The only earlier publication with some tests of the setup repeatability is the final report of the IPPTC [28] (see Section 4.3). The measurements performed during those trials indicated that the metal detector setup is an important source of variability. However, all influencing variables have been examined separately, so that the influence of the setup variability to all other results could not be evaluated. Some of the reported differences between the detector models might have easily been caused by the variability of the setup. The influence of the differences between the operators was not studied during the IPPTC.

7.5.2 In-Air Measurement Procedure

During the in-air measurements the targets were moved by hand and the position and the orientation of the search head was not strictly controlled (see Section 7.2, Figure 7.4). This introduced a source of error not present in the in-soil measurements. However, this error might be a few millimetres (this is a subjective estimate of the double standard deviation) and it is for most targets and most detectors much smaller than the estimated standard deviation.

The main advantage of the in-air measurement procedure applied in these trials is the speed. Some measurement devices were built with the aim to improve the control of the distance between the target and the search head (as described in the CWA 14747:2003 [26]). However, the use of more precise equipment would be much more time consuming. It would also add little value, since it would decrease only a small part of the experimental error: the operators and the setup introduce a much higher error than the distance measurements. There might be cases when such a high precision is necessary, but in most trials a simple measurement procedure as described in this work will be sufficient.

7.5.3 Comparison between PMA-2 and PMA-S

It has been shown that the PMA-2 and its surrogate PMA-S give slightly different signals in air. It is important to note that the measured difference of (9.4±6.6) mm is valid only in air and for the four detectors used in the experiment. Even though the PMA-S is slightly more difficult to detect, in some cases it can still be used in metal detector trials as a surrogate of the PMA-2. However, the difference between the two targets is too large to enable us to make a valid comparison between the results in two soils.

7.5.4 Choosing the Most Appropriate Detector

If the choice between detectors has to be made only based on the MDH measurements, and if there is a statistically significant difference at level $\alpha = 0.05$ between two detectors, the obvious choice would be the detector with the higher average MDH. If the difference is not statistically significant, the choice is not that easy. Both the average and the variance have to be taken into account. Let us look at a fictional example. If detector 1 has slightly higher average MDH than detector 2, but also much higher variance, then it would miss shallowly buried targets more frequently that detector 2. Since most of the mines are buried shallowly, we cannot say that detector 1 would be a better choice. This problem is discussed in more detail in Section 8.7, where the in-soil MDH measurements are connected with the probability of detection.

7.6 Conclusions

The maximum detection height measurements performed during the metal detector trial in Benkovac, Croatia, May 2005, have shown that the sensitivity of metal detectors has a high variability that has to be taken into account. The variability was caused by the differences between the operators, by the subjectivity of the operators, by the setup, by the instability of the hardware of the devices, and by the uncertainty of the measurements of the distance between the search head and the target.

Repeated measurements with several operators are necessary to estimate the experimental error and to attribute some confidence to our statements about the performance of metal detectors. Planning an experiment according to the principles of experimental design enables an unbiased estimate of different influences and minimises the experimental error.

The simple in-air measurement procedure proposed in this work is sufficiently precise for most purposes. The use of more complicated equipment (for example, as proposed in CWA 14747:2003 [26]) would be more time-consuming, and it would be justified only if higher precision in the distance measurement is necessary.

Chapter 8

Detection Reliability Tests

This chapter describes the detection reliability tests, which were a part of metal detector trials in Oberjettenberg, Germany, and Benkovac, Croatia in 2003 and 2005. The design of experiment, the data evaluation and the test results are presented. Each trial included the experience and lessons learned from the earlier trials. Consequently, all sections of this chapter are connected, each section often referring to earlier sections. To understand the connection between the sections and to follow the discussions, the reader will need to be familiar with Section 8.1 (overview of the tests) and Section 8.2 (design of experiment and data analysis).

Sections from 8.3 to 8.6 contain the information specific for each trial, including the test results. Section 8.7 explains the connection between the maximum detection height measurements and the probabilities of detection from the reliability tests. A discussion in Section 8.8 connects all test results, including the maximum detection height measurements. It also contains some recommendations regarding future testing. A short conclusion closes this chapter.

8.1 Overview of Detection Reliability Tests Performed in Oberjettenberg and Benkovac

A series of metal detector trials was conducted in accordance with the CWA 14747:2003, the standard for testing metal detectors for humanitarian demining [26] (see Section 4.2). The purpose of these trials was to investigate the feasibility of the tests described in the CWA 14747:2003. The trials were performed within the frame of ITEP projects 2.1.1.2 and 2.1.1.8. The aims of the projects were to find an appropriate design of experiment for testing metal detectors, meaning the appropriate choice of factors, levels and treatments; to establish the use of ROC diagrams and POD curves (explained in Subsection 8.2.2) in reporting the experimental results; and to gain practical experience in organising and conducting metal detector trials. The

trials comprised detection reliability tests and maximum detection height measurements in different soil types and with different targets, including real mines. Some results of these trials have been published in two reports [76, 72] and presented at several conferences [42, 39, 41, 40, 75, 74, 5, 6]. Many laboratory measurements, mostly maximum detection height measurements, were performed at the Joint Research Centre of the European Commission and a separate technical note describing those measurements is available [10]. Table 8.1 provides an overview of all trials.

The goal of the first trial, Oberjettenberg May 2003, was to find the most appropriate metal detector for the choice of factor levels — soils, targets and target depths — selected for that trial. This choice of factor levels was representative for a rather difficult scenario. In later trials the goal was to find the most adequate device for each combination soil-target-depth. To achieve this goal, the number of investigated factor levels had to be reduced and the design of the experiment had to be modified.

Two trials were conducted at the test sites of the German Federal Armed Forces at the Military Engineering Department 52 (WTD 52) in Oberjettenberg, near Bad Reichenhall, Germany. Another two trials were conducted at the test sites of the Croatian Mine Action Centre – Centre for Testing, Development and Training (HCR-CTRO) in Benkovacko Selo near Benkovac, Croatia.

The metal detector models tested in the trials were European-manufactured models currently in use in clearance operations worldwide. They are 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 labeled with Latin letters U, W, X, Y and Z throughout all publications, including this thesis. In many discussions of the experimental design throughout this thesis, the names alpha, beta gamma and delta are used, to emphasise that the design is based on the Graeco-Latin square and to avoid confusion with the operators. Detector models U, X, Y and Z were tested in all trials, while detector W was tested only in the Oberjettenberg November 2003 trial. This was a prototype of a new model and only one specimen was provided for testing. The detector models 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. The design of detector U allows the user to chose between the static mode and the dynamic mode. All of the models had some ground compensating abilities. Two specimens of each model were used in all trials in 2003 and only one specimen in the trial in 2005. The trial in Oberjettenberg in November 2003 was an exception; three detector models were present with two specimens and another two models with one specimen each. The tested specimens were chosen by the manufacturers. An overview is presented in Table 8.2.

Location and Time	Oberjettenberg May 2003	Benkovac July 2003	Oberjettenberg November 2003	Benkovac May 2005
Lanes	4	7	7	4
Lane size	20 m^2	30 m^2	20 m^2	29 m^2
Soils	4	3	7	2
Detector models	4	4	5	4
Specimens of each detector model	2	2	2 or 1	1
Sensitivities	2	2	1	1
Operators	8	8	8	4 + section leader
Targets in a lane	24-28 (12-15 types)	32 (10 types)	22-28 (11-14 types)	30 (2 types)

Table 8.1: Overview of all metal detector trials described in this dissertation.

Detector	Search Head Coil	Mode	Electromagnetic Wave
U (beta)	single	static or dynamic	time domain
W (alpha-1)	single	dynamic	time domain
X (alpha-2)	single	dynamic	time domain
Y (delta)	double-D	static	frequency domain
Z (gamma)	double-D	static	frequency domain

Table 8.2: Detectors tested in the trials.

The tests were performed on several different soil types. The lanes were prepared according to CWA 14747:2003 [26]. The first tests in Oberjettenberg, May 2003, were performed on four lanes, each containing different soil (Table 8.3). One of them was covered with a 2-cm layer of blast furnace slag to simulate uncooperative soil. The three soil types used in Benkovac trials, in July 2003, represent some mined regions in Croatia (Table 8.4). Two of these were highly uncooperative. In the next trial in Oberjettenberg, November 2003, the same lanes were used as in the first trial, with the addition of other three lanes, one of which was highly uncooperative. The last Benkovac trial, May 2005, included tests on only two of the soil types used in the previous Benkovac trial. All lanes were cleared of metal with the aid of detectors. The soils are described in detail in [76].

The targets used in the trials were real mines modified to be safe, ITOP inserts (defined in [80] and described in [26]) and a chromium steel ball. The mines used in the Oberjettenberg trials can be found in minefields worldwide, while those used in the Benkovac trials are representative of South-Eastern Europe.

In both Oberjettenberg trials, the operators were soldiers of the German army with no experience in demining. In the Benkovac trials, the operators were professional deminers. In the first two trials the operators had undergone a two-day training for four detector models. In the second two trials the training was twice as long, that is, one day per detector model.

In the first two trials, two sensitivity levels of metal detectors were used. The aim was to study the influence of the sensitivity on the probability of detection and on the false alarm rate. Professional deminers sometimes use lower sensitivities to avoid detecting metal clutter. This lower sensitivity is usually calibrated so that the mine representing the local threat is buried to a specific depth and the sensitivity of the metal detector is set up so that this mine can still be detected. The calibration target for the trials was a chromium steel ball.

Soil Types in Oberjettenberg Trials	χ (958 Hz) [10^{-5}]	χ (465 Hz) – χ (4650 Hz) [10^{-5}]
Lane 1 artificially uncooperative soil	244 ± 64	6,1
Lane 2 cement gravel	0 ± 1	– 0,2
Lane 3 clay	2 ± 1	– 0,5
Lane 4 concrete gravel	6 ± 1	– 0,5
Lane 5 magnetite mixed with sand	3000 ± 500	6 ± 7
Lane 7 cement gravel	– 1,0 ± 0,2	– 0,1 ± 0,2
Lane 8 concrete gravel	7 ± 1	– 0,1 ± 0,1

Table 8.3: Soil types in the Oberjettenberg trials. In the first trial, May 2003, only lanes 1, 2, 3 and 4 were used. Magnetic susceptibility measurements were performed with a Bartington MS2 magnetometer, MS2D sensor at 958 Hz and MS2B sensor at 465 and 4650 Hz. The sensor of the circular loop probe operating at 958 Hz (MS2D) is calibrated to read 0.5 χ on rough soils and will give about 0.75 χ on smooth surfaces [7]. The MS2B sensor works with 10 cm^3 samples.

Soil Types in Benkovac Trials	χ (958 Hz) [10^{-5}]	χ (465 Hz) – χ (4650 Hz) [10^{-5}]
Lanes 1, 5 (2003) Lanes 1, 2 (2005) bauxite	154 ± 13	25,5
Lanes 2, 6 (2003) Lanes 3, 4 (2005) clay	13 ± 2	0,6
Lanes 3, 4, 7, 8 (2003) bauxite with limestone	190 ± 36	35,4

Table 8.4: Soil types in the Benkovac trials. In 2005 the labeling of the lanes was different than in 2003. Lanes 1 and 5 from 2003 trials were named lanes 1 and 2 in 2005. Lanes 2 and 6 were renamed to lanes 3 and 4. Lanes 3, 4, 7 and 8 from 2003 trial were not used in the 2005 trial. Magnetic susceptibility measurements were performed with a Bartington MS2 magnetometer, MS2D sensor at 958 Hz and MS2B sensor at 465 and 4650 Hz. The sensor of the circular loop probe operating at 958 Hz (MS2D) is calibrated to read 0.5 χ on rough soils and will give about 0.75 χ on smooth surfaces [7]. The MS2B sensor works with 10 cm^3 samples.

8.2 Design of Experiment and Data Analysis

8.2.1 Design of Experiment

This section describes some common properties of the experimental designs of the four detection reliability tests. Detection reliability tests in general are described in Section 4.2.2.

The outcome of the experiment is described with two response variables: variable 'detected' with two levels, 1 ("detected") and 0 ("not detected"), defined for each pass over a target, and variable 'signals' with levels 0, 1, 2, 3, ..., which is the number of false alarms in a run. A *run* is a single pass of an operator with a detector through a lane.

The factors appearing in the tests were: 'detector model', 'detector specimen', 'detector sensitivity', 'lane', 'operator', 'target', and 'target depth'. Only one of the tests (Benkovac, July 2003) included all of these factors, while the others included only a selection of them. Each of the four tests had a different design of experiment, each of them introducing some improvements. They are described in detail in Subsections 8.3.2, 8.4.2, 8.5.2 and 8.6.2. All designs are based on the idea described in the following lines.

A factorial design including all factor level combinations would certainly solve our experimental problem. However, such a test would require a lot of

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

Table 8.5: Design of experiment for a simplified problem of testing four metal detectors. A, B, C and D are operators and *alpha*, *beta*, *gamma* and *delta* are detectors.

time and therefore an unacceptably high budget. 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. The choice of treatments (i.e. factor level combinations) is determined with a Graeco-Latin square, thus making an orthogonal design. A solution to a simplified problem with only four factors is presented in Table 8.5. This design was the basis for the design of all trials. An additional factor called 'start' had been introduced to solve the problem of the order of execution. Four runs of a single start were planned to be executed simultaneously, thus saving time. The levels of the variable 'start' indicate their order of execution. (Actually, in some tests the runs of the same start were not performed simultaneously because of electromagnetic interference between detectors, but they were performed consecutively or two at a time.)

8.2.2 Data Analysis

Section 8.5.7 of the standard CWA 14747:2003 [26], dealing with reliability tests, gives a recommendation to report about the numbers of true positive, false positive and false negative indications (see their definitions in Section 4.2.2 of this dissertation). The standard does not specify how this information should be presented, neither a method of attributing a confidence level to the results.

The estimated *probability of detection* for a particular factor level combination is the ratio of the number of detected targets and the total number of opportunities to detect a target. The estimated *false alarm rate* is defined as the number of false alarms counted on an area divided by the size of that area, or the average number of false alarms per square metre. The area is calculated as the area of the test lane minus the area of all detection halos. If we assume a binomial distribution for the number of true positive indications, we can find the 95% confidence limits for the probability of detection. Similarly, if we assume Poisson distribution for the false alarms, we can construct the confidence limits for the false alarm rate [98, 14, 84, 85].

For the probability of detection and the false alarm rate, we will use the usual abreviations POD and FAR respectively, and their estimated values will be marked with a circumflex: $\widehat{POD}, \widehat{FAR}$.

The upper and the lower confidence limits of the POD and those of the FAR can be computed with many commercially available computer programmes or calculated with the help of some statistical tables ([98, 14]). Before the widespread use of computers some approximative procedures had been developed. However, their use today seems to be hardly justified, since even the most common spreadsheet programmes can deal with the functions necessary for computing the confidence limits of a binomial and a Poisson distributed variable. We will nevertheless mention two extreme approximations, since their use requires very little calculations, thus making them suitable for a quick assessment of the size of the confidence interval.

Let us call the number of opportunities to detect a target n, and the number of detections y. We introduce two more abbreviations: $p = POD$ and $q = 1 - p$. The number of detections is binomially distributed with the parameter p. The two sided $1 - \alpha$ confidence limits[1] [98, 84] are

$$\begin{aligned} POD_{lower} &= \frac{y}{y+(n-y+1)F_{1-\alpha/2,\, f_1,\, f_2}} \quad \text{with } f_1 = 2(n-y+1),\ f_2 = 2y \\ POD_{upper} &= \frac{(y+1)F_{1-\alpha/2,\, f_1,\, f_2}}{n-y+(y+1)F_{1-\alpha/2,\, f_1,\, f_2}} \quad \text{with } f_1 = 2(y+1),\ f_2 = 2(n-y) \end{aligned} \tag{8.1}$$

The quantities $F_{1-\alpha/2,\, f_1,\, f_2}$ are F-quantiles (also called percentage points) of the F distribution. In the special case when $y = 0$ the two sided confidence limits are

$$\begin{aligned} POD_{lower} &= 0 \\ POD_{upper} &= 1 - \sqrt[n]{\alpha/2} \end{aligned} \tag{8.2}$$

When $y = n$,

$$\begin{aligned} POD_{lower} &= \sqrt[n]{\alpha/2} \\ POD_{upper} &= 1 \end{aligned} \tag{8.3}$$

It has been proposed [14] that a normal approximation of a binomial distribution can be used if $n > 5$ and

$$\frac{\left|\sqrt{\frac{p}{q}} - \sqrt{\frac{q}{p}}\right|}{\sqrt{n}} = \frac{|p-q|}{\sqrt{npq}} < 0.3 \tag{8.4}$$

This condition means that n is sufficiently large and both p and q are sufficiently far from 1 and 0. For the confidence level $1 - \alpha = 95\%$ and with some additional approximations (described by K. M. Simonson [84]) we get the following relation:

$$POD_{upper/lower} = p \pm 2\sqrt{\frac{pq}{n-1}} \tag{8.5}$$

[1] It can be shown from the relation $F_{\alpha,a,b} = 1/F_{1-\alpha,b,a}$ that the confidence limits given in [84] are identical to those from [98].

If p is close to 0.5, further approximation is possible:

$$POD_{upper/lower} = p \pm \frac{1}{\sqrt{n}} \tag{8.6}$$

The number of false alarms in a single scan over a lane follows a Poisson distribution. A variable which is a sum of Poisson distributed variables also follows a Poisson distribution. Consequently, the total number of false alarms x in N repeated scans over an area of size A also follows a Poisson distribution. The estimated false alarm rate is $\widehat{FAR} = \frac{x}{N \cdot A}$. The two sided confidence limits are [98, 85]

$$\begin{aligned} FAR_{lower} &= \tfrac{1}{2\,N \cdot A} \chi^2_{\alpha/2,\,f} \quad \text{with } f = 2x \\ FAR_{upper} &= \tfrac{1}{2\,N \cdot A} \chi^2_{1-\alpha/2,\,f} \quad \text{with } f = 2(x+1) \end{aligned} \tag{8.7}$$

where $\chi^2_{\alpha/2,\,f}$ and $\chi^2_{1-\alpha/2,\,f}$ are called quantiles or probability points of the χ^2-distribution. In the special case when $x = 0$, the confidence limits are

$$\begin{aligned} FAR_{lower} &= 0 \\ FAR_{upper} &= ln(2/\alpha) \end{aligned} \tag{8.8}$$

When x exceeds 15, the Poisson distribution can be approximated by a normal distribution [85]. Setting the variance of that normal distribution to be equal to the variance of the Poisson distribution, we get approximate 95% confidence limits for the FAR:

$$FAR_{upper/lower} = \widehat{FAR} \pm 2\sqrt{\frac{\widehat{FAR}}{N \cdot A}} \tag{8.9}$$

Because of their simplicity, equations (8.6) and (8.9) can be helpful in the preparation phase of the experiment. Normally equations (8.1) and (8.7) should be used, since F-quantiles and χ^2-quantiles can be easily calculated with most modern spreadsheet programmes.

The confidence limits to a binomially distributed and to a Poisson distributed variable have been known much before their application in tests of demining equipment. Their first correct use in demining related papers was in 1998 in two articles by K. M. Simonson [84, 85].

In this work, the $\widehat{POD}$ and the $\widehat{FAR}$ are combined in a diagram called an *ROC diagram*, where ROC stands for *receiver operating characteristic*. In earlier applications, ROC diagrams are diagrams of POD versus probability of false alarm. Receiver operating characteristic is an analytical procedure based in statistical decision theory and was developed in the context of electronic signal detection [93]. It has been applied to many fields, including human perception and decision making, diagnostic systems in clinical medicine, non-destructive testing etc. Already in 1998 K. M. Simonson

[84, 85] proposed to use the FAR instead of the probability of false alarm in tests of demining equipment, and she argued that a point on an ROC diagram describes a so called ROC curve if the threshold is varied.

Another kind of diagram taken over from the field of non-destructive testing and applied in demining is the *POD curve* [29, 8]. It describes the dependency of the POD on a parameter of the target. In non-destructive testing, that parameter is the flaw size , while in demining it is the target depth. Some other target parameters have been considered, like the mass or the volume of the metal part, but they are not convenient for this kind of diagrams. The constitutive parts of each mine type have different shapes and sizes, and they are made of different materials. The POD depends on all these factors, which is why, for example, a diagram of POD versus mass would not be very informative.

The detection of a target is modelled as a Bernoulli experiment where the binary random variable Y takes its value $y = 1$ ("detected") with the probability p and its value $y = 0$ ("not detected") with the probability $1 - p$. The parameter p is specific for each treatment and it depends on the influence variables characterising that treatment. We cannot relate p linearly with the influence variables, since p is limited to $0 \leq p \leq 1$. Therefore, the parameter p of the Bernoulli distribution is transformed into the parameter η:

$$\eta = ln\left(\frac{p}{1-p}\right) \tag{8.10}$$

This transformation is called *logistic* (or *logit*) *transformation* and the inverse function

$$p = \frac{1}{1+e^{-\eta}} \tag{8.11}$$

is called the *logistic function*. It is a monotonically increasing S-shaped curve starting with $p(-\infty) = 0$ and ending with $p(\infty) = 1$. The parameter η, which is between $-\infty$ and $+\infty$, is linearly related to the influence variables:

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q = \sum_{j=0}^{q} \beta_j x_j \tag{8.12}$$

where one of the x_j's stands for the depth of the target and the other x_j's are indicator variables indicating the presence of a particular level of a qualitative factor. This model is called a *generalised linear model* [56]. The unknown parameters β_j of the generalised linear model are estimated by maximum likelihood estimation. The result is a curve of POD versus target depth for each combination of other factor levels. It is possible to create confidence bounds to POD curves. The procedure is described in [72].

A simpler analysis of POD dependence on depth has been proposed in [42] and [76] and used later in the STEMD trials [50]. This analysis does not use the generalised linear model described in Subsection 8.2.2. The

detections for each depth are simply counted and expressed as $\widehat{POD}$ with the appropriate confidence limits calculated from Equations (8.1). Such a method has the obvious advantage that it needs fewer calculations and that it does not depend on any assumptions about the relationship between the POD and the depth; the only assumption is that the detections on a certain depth are binomially distributed. Its disadvantage is that it produces larger than necessary confidence intervals and that it cannot give information about the depths not present in the test. An example, Figure 8.15, can be found in Subsection 8.5.3.

The POD curves in this thesis, as well as some of the ROC diagrams, were created with a programme written in R 2.0.1 [81] by Prof. P.-T. Wilrich. All other diagrams were created in Microsoft Excel 2002.

8.3 Reliability Tests, Oberjettenberg, May 2003

8.3.1 Introduction

The goal of the detection reliability tests performed in Oberjettenberg in May 2003 was to compare four metal detector models in conditions represented in the tests. These conditions are considered representative for a certain scenario and the total results reflect the abilities of the detectors to deal with this scenario. The four detector models were detectors U, X, Y and Z from Table 8.2. There were two specimens of each model, marked with numbers 1 and 2. The tests were performed with two sensitivity settings, called low and high. The high sensitivity was the maximum sensitivity of a metal detector. The low sensitivity was calibrated so that a detector could just detect a 16 mm steel ball (100Cr6 steel) buried to a specified depth. This depth was 15 cm in the cooperative soils in lanes 2, 3 and 4, and 12 cm in the uncooperative soil in lane 1.

There were four lanes in the experiment, each containing an other soil type. Lane 1 was covered with a 2-cm layer of blast furnace slag to simulate uncooperative soil. The other lanes were magnetically neutral (see Table 8.3). All lanes were about 20 m long and 1 m wide. Each lane contained different targets buried to different depths. Their number varied between 24 and 28. The target positions and depths were determined before the planning of these trials and it was decided not to change them, since the targets were already in the ground. They are listed in Table 1 in the Appendix.

Eight deminers tested the devices, four in the first week and four in the second one. They were soldiers of the German army without any previous demining experience. All operators were introduced to all four detector models in a two-day training lead by the representatives of the manufacturers.

8.3.2 Design of Experiment

The solution to the experimental problem described in the previous section is given in Table 8.6. This design is a combination of two Graeco-Latin squares. The factor 'specimen' is a nested variable, nested in the factor 'detector', meaning that the level of the factor 'specimen' is meaningful only within a level of the factor 'detector'. For each specimen, a separate Graeco-Latin square was constructed and they were combined together to form this design. The organiser of the experiment did not have the opportunity to chose the targets, neither their depths, but they had to be accepted with the test site. This is why the mine types and the depths cannot be considered factors in this test. The detector sensitivity is not fully integrated in the design for two reasons. The first is that the frequent change of the sensitivity would cause significant additional stress for the operators as well as the trial monitors. The other reason is that the effect of the sensitivity was expected to be much higher than a possible effect of time, so that the possibility of a bias is negligible.

Eight starts were planned to be performed each day. However, that was not possible to achieve, due to weather conditions and the slow advance on the first day of the trial. This is why the factor "days" should be understood solely as a label and it does not exactly correspond to actual days, however, it faithfully represents the order of execution of the measurements. Day 2 is the exact repetition of day 1. All starts within days 1 and 2 were performed with high sensitivity. Days 3 and 4 are identical to days 1 and 2, except that the measurements were performed with low sensitivity. Days 5, 6, 7 and 8 are a repetition of the days 1, 2, 3 and 4, but with other operators: A, B, C, D were replaced by E, F, G, H.

8.3.3 Results

Figure 8.1 is a ROC diagram for the complete set of data, containing all factor levels (all targets, target depths, lanes and operators). In this diagram we can compare the overall performance of the four detectors tested in the trial. There are obvious differences between the detectors, two of them having a higher POD than the other two.

An ROC diagram based on the same data set, but comparing four lanes, is on Figure 8.2. This diagram should be interpreted as a comparison of the lanes, and not of the soils. Namely, the lanes contained different targets on different depths. In other words, the targets, their depths and the soil types were all confounded. Nevertheless, the results for the lanes 2, 3 and 4 are very similar. The slightly increased FAR in lanes 3 and 4 might had been caused by some minor metal contamination.

The same kind of diagram can be used to examine the differences between the operators: Figure 8.3. The points lie on what could be an ROC curve:

Days 1, 2, 3, and 4:

	Start 1	Start 2	Start 3	Start 4	Start 5	Start 6	Start 7	Start 8
Lane 1	*A alpha*-1	*C gamma*-2	*B beta*-1	*D delta*-2	*C gamma*-1	*A alpha*-2	*D delta*-1	*B beta*-2
Lane 2	*B gamma*-1	*D alpha*-2	*A delta*-1	*C beta*-2	*D alpha*-1	*B gamma*-2	*C beta*-1	*A delta*-2
Lane 3	*C delta*-1	*A beta*-2	*D gamma*-1	*B alpha*-2	*A beta*-1	*C delta*-2	*B alpha*-1	*D gamma*-2
Lane 4	*D beta*-1	*B delta*-2	*C alpha*-1	*A gamma*-2	*B delta*-1	*D beta*-2	*A gamma*-1	*C alpha*-2

Days 5, 6, 7 and 8:

	Start 1	Start 2	Start 3	Start 4	Start 5	Start 6	Start 7	Start 8
Lane 1	*E alpha*-1	*G gamma*-2	*F beta*-1	*H delta*-2	*G gamma*-1	*E alpha*-2	*H delta*-1	*F beta*-2
Lane 2	*F gamma*-1	*H alpha*-2	*E delta*-1	*G beta*-2	*H alpha*-1	*F gamma*-2	*G beta*-1	*E delta*-2
Lane 3	*G delta*-1	*E beta*-2	*H gamma*-1	*F alpha*-2	*E beta*-1	*G delta*-2	*F alpha*-1	*H gamma*-2
Lane 4	*H beta*-1	*F delta*-2	*G alpha*-1	*E gamma*-2	*F delta*-1	*H beta*-2	*E gamma*-1	*G alpha*-2

Days 1, 2, 5 and 6: low sensitivity. Days 3, 4, 7 and 8: high sensitivity.

Table 8.6: Design of the reliability test, Oberjettenberg, May 2003.

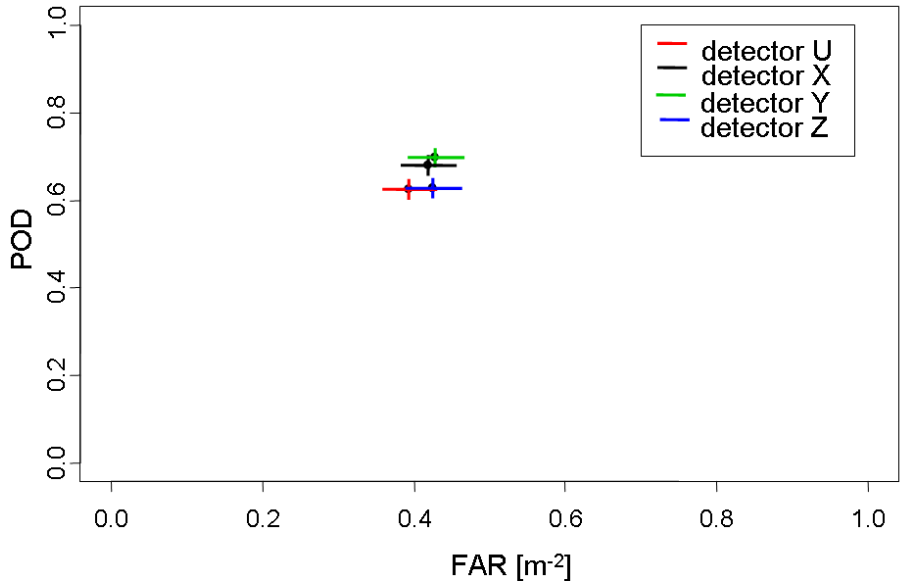

Figure 8.1: Oberjettenberg May 2003, an ROC diagram for the complete data set, a comparison of detectors. The crosses indicate 95% confidence limits.

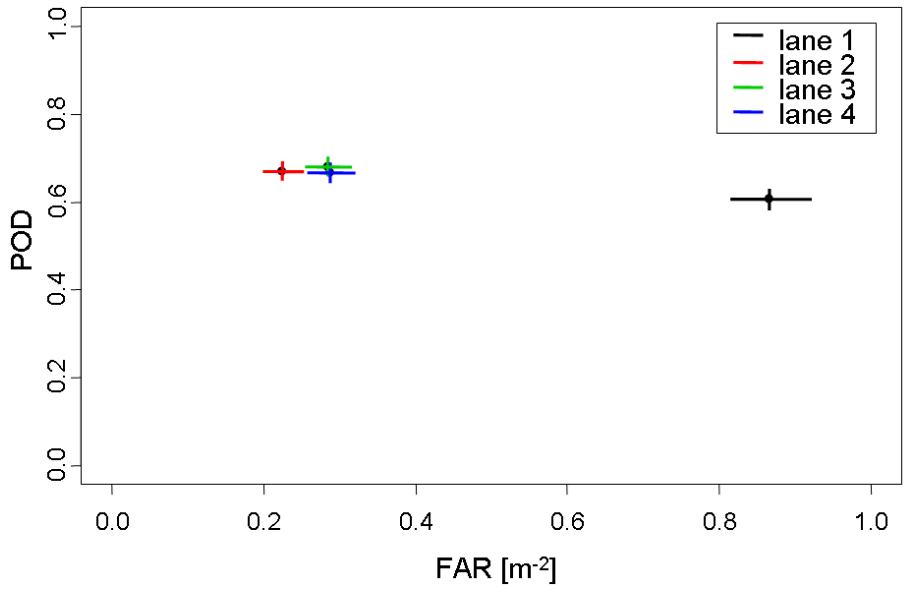

Figure 8.2: Oberjettenberg May 2003, an ROC diagram for the complete data set, a comparison of lanes. The crosses indicate 95% confidence limits.

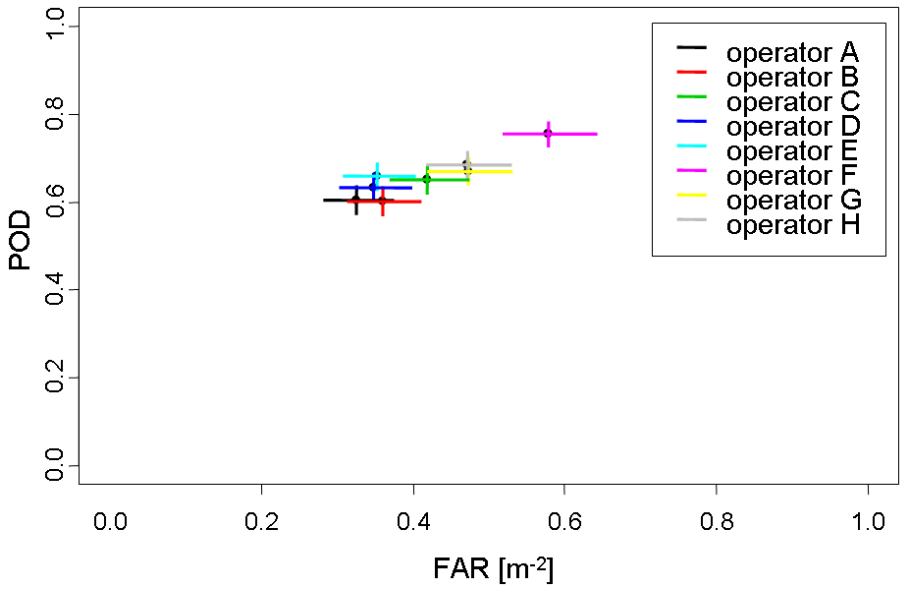

Figure 8.3: Oberjettenberg May 2003, an ROC diagram for the complete data set, a comparison of operators. The crosses indicate 95% confidence limits.

the operators who detected more targets also had more false alarms. It might be that they were more careful and more thorough and that their indications counted as false alarms are actually caused by metal clutter. The problem of metal clutter causing alarms is discussed in Section 8.8.2.

The next diagram provides a comparison between the results with the high and the low sensitivity. On Figure 8.4 we see that the difference between the two results is very small, not even statistically significant at the level $\alpha = 0.05$.

It is possible to create POD curves from the data of this trial, however, that would not be very informative. The reason is the unfavorable choice of depths. None of the target types was buried to a wide range of depths, so that the estimated POD curve would have a very wide confidence region. The estimated POD would be reliable only around the depths at which the targets were buried.

A comparison of the two specimens of each model revealed no significant differences.

8.3.4 Discussion

The probability of detection in this trial was much lower than many experts expected. However, all earlier trials had similarly low detection rates. The reasons for low PODs in this trial are:

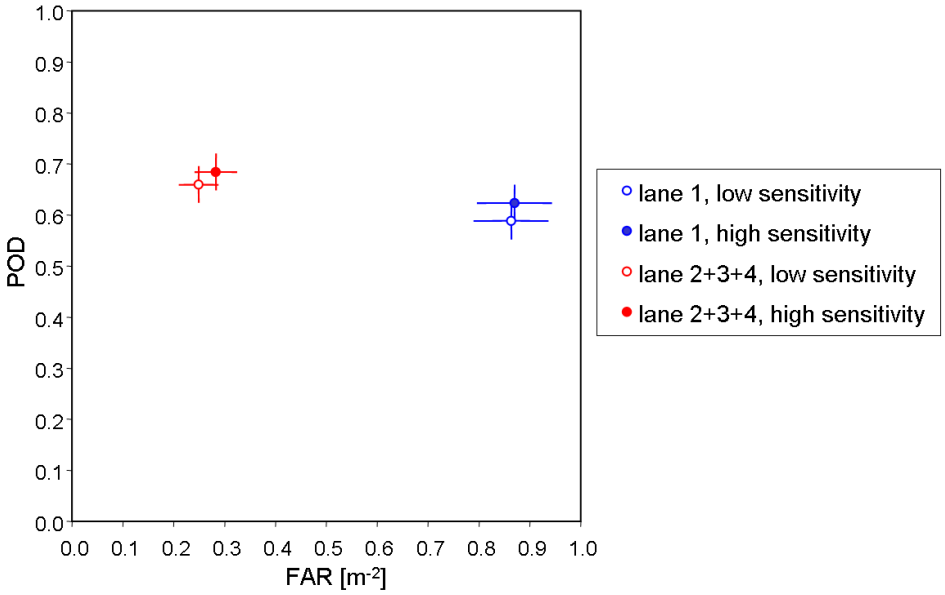

Figure 8.4: Oberjettenberg May 2003, an ROC diagram for the complete data set, a comparison of high and low sensitivity measurements. The crosses indicate 95% confidence limits.

1. The test was purposely designed to be difficult. The targets were placed to depths that do not represent a typical realistic scenario. The goal of the test was to discriminate the detector models, and not to estimate the actual detection rates in a minefield.

2. The time schedule and the obligation to complete a large number of runs in a limited time created time pressure on the operators, which caused a faster progression and lower concentration than in a minefield.

3. The absence of danger decreased the alertness of the operators.

4. The training with the new detector models was shorter than in practice. The operators had to master four detector models in two days.

5. The operators changed the detector models very often, with little time to adjust to the next device.

As a result, the most difficult targets were often missed and the pinpointing was less precise than it would be in a real minefield. Poor pinpointing is the reason why some mines with a significant metal content were missed, even if they were buried shallowly. The results of these and other trials certainly do not reflect the actual clearance success. If they would faithfully represent the actual performance of metal detectors, there would be many

more demining accidents. Most accidents are caused by a violation of the prescribed clearance procedures, and not by the failure of the metal detector [86, 86, 88].

The improvements of the experimental design and organisation in the following trials reduced the influence of points 2, 4 and to some extent point 5 from the upper list. The problems numbered 1 and 3 remained, since they cannot be solved, as it is discussed in Section 8.8.

The choice of target depths in this trial was not suitable for constructing POD curves (POD versus target depth), because the target depths were not systematically determined. Since the targets were placed in each lane to different depths, it was not possible to evaluate the influence of soils.

The use of blast furnace slag in lane 1 has proven to be inappropriate, since it contained metal fragments. Only the largest ones could be found and removed during the preparation of the trial. The fragments remaining in the lane created signals that were sometimes stronger than the signals coming from the nearby lying targets, which is why some targets were missed.

The calibration for the so called low sensitivity produced an insignificant change compared with the so called high sensitivity (see Figure 8.4). This is why the results of this trial provided no insight into the nature of the ROC curve. If the calibration target had been buried closer to the surface, the difference between the two sensitivities would have been larger. Namely, the calibration target would have been detected with a lower sensitivity, so that the "low sensitivity" setting would have been lower. An improved calibration procedure was applied in the next trial, see Section 8.4.

The design of experiment applied in this trial allows us to draw conclusions about the performance of metal detectors in conditions which are faithfully represented by the choice of factor levels present in the trials. The design does not allow an unbiased comparison of detector models in each soil separately. Let us, for example, select the data obtained from lane 2. We can produce an ROC diagram with four points, with the intention to compare the four detectors. Each of these points would be the average $\widehat{POD}$ versus the average $\widehat{FAR}$ of all starts performed in lane 2 with one of the four detectors. By referring to Table 8.6 presenting the design of experiment, we read that each detector was always used by the same two operators. For example, detector model gamma was used only by operators B and F. If operators B and F (or only one of them) were much different from the others, then the differences between the four points would be caused by the differences between the operators and not by the differences between the detectors.

8.4 Reliability Tests, Benkovac, July 2003

8.4.1 Introduction

The goal of these tests was to evaluate the performance of four detector models in each soil type separately. Detectors U, X, Y and Z (see Table 8.2) were tested. There were two specimens of each model, marked with numbers 1 and 2, and all measurements were performed with two sensitivity settings called high and low. The high sensitivity was the maximum sensitivity of a metal detector. The low sensitivity was calibrated so that a detector could just detect a 16 mm steel ball (100Cr6 steel) buried to a specified depth. This depth was 11 cm in Obrovac soil (lanes 1 and 5), 12 cm in Benkovac soil (lanes 3, 4, 7 and 8) and 13 cm in Sisak soil (lanes 2 and 6). The calibration of the low sensitivity was such that the difference between the high and the low sensitivity was higher than in the Oberjettenberg May trials.

Eight operators tested the devices, four deminers in the first week and four in the second one. They were all employees of demining organisations working in Croatia, three of them being currently active as deminers, while the others were former deminers. They were introduced to the four detector models in a two-day training led by manufacturer representatives.

There were eight lanes in the experiment. Lanes 1 and 2 contained soil from the area around Obrovac, Croatia, lanes 3 and 4 contained soil from the surroundings of Sisak, Croatia, and lanes 5, 6, 7 and 8 contained the original local soil from Benkovac (see Table 8.4). All lanes were 30 m long and 1 m wide. Each lane contained the same 32 targets buried to the same depths, as presented in Table 2 in the Appendix. For example, PMA-2 was buried to 0, 5, 10, 13 and 20 cm depth in each lane, the same as PMA-1A and PMA-3. The antitank mines TMA-3 and TMA-4 were treated as the same target, since they have exactly the same metal content. They both contain three detonators of the same type that is used in PMA-2 and PMA-3 antipersonnel mines. The antitank mines TMRP-6 and TMM-1 both contain large amounts of metal, so that they were considered to be the same target [65].

Let us compare these trials with the Oberjettenberg May trials described in Subsection 8.3.1. The main improvement was the inclusion of the targets and their depths into the design of experiment. In Benkovac trials, the targets were systematically distributed over a wide range of depths (0-20 cm). This enabled an evaluation of the dependence of POD on depth.

Eight lanes were used instead of four and they were longer. Since the working hours were the same, the workload was higher than in the Oberjettenberg tests. The density of the targets was smaller, 32 targets per 30 m^2 compared with 24 to 28 targets per 20 m^2 in Oberjettenberg, which allowed for some empty space in the lanes. Fewer soil types were present, only three. The operators were experienced deminers, compared with inexperi-

enced soldiers in the Oberjettenberg trials. The low sensitivity was lower than in the Oberjettenberg trials. The improvements of the experimental design are discussed in Subsection 8.4.2.

During the preparation of the trials, the lanes had to be cleared of metal debris. Two teams of two persons worked several days on clearing the lanes with the help of metal detectors. Lane 8 was so contaminated with metal fragments, that the teams decided to give up after the first five metres. It was decided not to use lane 8 in the tests, but instead to use lane 4, which contained the same soil type. To keep the data analysis, the description of the design and of the results as simple as possible, the label 'lane 8' was kept, but the reader should know that the starts planned for lane 8 were performed in lane 4.

8.4.2 Design of Experiment

The design of experiment, Table 8.7, was very similar to the design of the Oberjettenberg May tests described in Subsection 8.3.2, Table 8.6. The starts of day 1 of the two tests are identical. Day 2 of the Benkovac test can be understood as a repetition of day 1, with changed lanes and operators. Instead of lanes 1, 2, 3, 4, we have lanes 5, 6, 7, 8; lane 5 having the same soil type as lane 1, lane 6 the same as lane 2, etc. The same operators performed the measurements on day 2, but they were permuted. The change can be symbolically expressed as (A, B, C, D) $\leftrightarrow$ (C, D, A, B), where letters represent the operators. Days 3 and 4 are low sensitivity measurements with the same design as days 1 and 2. Days 5, 6, 7 and 8 are identical to days 1, 2, 3 and 4, except that these measurements were performed with four new operators labeled E, F, G and H. The main difference to the Oberjettenberg tests was the choice of targets and their depths (see Subsection 8.4.1 and the tables in the Appendix). The targets and their depths can be considered factors, since they were systematically chosen.

The motivation to permute the operators in the repeated measurements (days 2, 4, 6 and 8) was to get four instead of two operator-detector-lane combinations and thus get closer to a full factorial design. For example, in Obrovac soil, lanes 1 and 5, detector alpha was used by four operators: A, C, E and G, while in the Oberjettenberg May test, we had only two operators using detector alpha in the soil type contained in lane 1. This way the comparison of detectors in only one soil type is less influenced by the differences between the operators. The factors 'detector' and 'operator' are still confounded, but much less than in the Oberjettenberg May tests.

8.4.3 Results

Unless stated otherwise, only the high sensitivity measurements are presented in this subsection. Namely, the calibration procedure for the low

Days 1 and 3:

	Start 1	Start 2	Start 3	Start 4	Start 5	Start 6	Start 7	Start 8
Lane 1	*A alpha-1*	*C gamma-2*	*B beta-1*	*D delta-2*	*C gamma-1*	*A alpha-2*	*D delta-1*	*B beta-2*
Lane 2	*B gamma-1*	*D alpha-2*	*A delta-1*	*C beta-2*	*D alpha-1*	*B gamma-2*	*C beta-1*	*A delta-2*
Lane 3	*C delta-1*	*A beta-2*	*D gamma-1*	*B alpha-2*	*A beta-1*	*C delta-2*	*B alpha-1*	*D gamma-2*
Lane 4	*D beta-1*	*B delta-2*	*C alpha-1*	*A gamma-2*	*B delta-1*	*D beta-2*	*A gamma-1*	*C alpha-2*

Days 2 and 4:

	Start 1	Start 2	Start 3	Start 4	Start 5	Start 6	Start 7	Start 8
Lane 5	*C alpha-1*	*A gamma-2*	*D beta-1*	*B delta-2*	*A gamma-1*	*C alpha-2*	*B delta-1*	*D beta-2*
Lane 6	*D gamma-1*	*B alpha-2*	*C delta-1*	*A beta-2*	*B alpha-1*	*D gamma-2*	*A beta-1*	*C delta-2*
Lane 7	*A delta-1*	*C beta-2*	*B gamma-1*	*D alpha-2*	*C beta-1*	*A delta-2*	*D alpha-1*	*B gamma-2*
Lane 8	*B beta-1*	*D delta-2*	*A alpha-1*	*C gamma-2*	*D delta-1*	*B beta-2*	*C gamma-1*	*A alpha-2*

Days 5 and 7:

	Start 1	Start 2	Start 3	Start 4	Start 5	Start 6	Start 7	Start 8
Lane 1	*E alpha-1*	*G gamma-2*	*F beta-1*	*H delta-2*	*G gamma-1*	*E alpha-2*	*H delta-1*	*F beta-2*
Lane 2	*F gamma-1*	*H alpha-2*	*E delta-1*	*G beta-2*	*H alpha-1*	*F gamma-2*	*G beta-1*	*E delta-2*
Lane 3	*G delta-1*	*E beta-2*	*H gamma-1*	*F alpha-2*	*E beta-1*	*G delta-2*	*F alpha-1*	*H gamma-2*
Lane 4	*H beta-1*	*F delta-2*	*G alpha-1*	*E gamma-2*	*F delta-1*	*H beta-2*	*E gamma-1*	*G alpha-2*

Days 6 and 8:

	Start 1	Start 2	Start 3	Start 4	Start 5	Start 6	Start 7	Start 8
Lane 5	*G alpha-1*	*E gamma-2*	*H beta-1*	*F delta-2*	*E gamma-1*	*G alpha-2*	*F delta-1*	*H beta-2*
Lane 6	*H gamma-1*	*F alpha-2*	*G delta-1*	*E beta-2*	*F alpha-1*	*H gamma-2*	*E beta-1*	*G delta-2*
Lane 7	*E delta-1*	*G beta-2*	*F gamma-1*	*H alpha-2*	*G beta-1*	*E delta-2*	*H alpha-1*	*F gamma-2*
Lane 8	*F beta-1*	*H delta-2*	*E alpha-1*	*G gamma-2*	*H delta-1*	*F beta-2*	*G gamma-1*	*E alpha-2*

Days 1, 2, 5 and 6: low sensitivity. Days 3, 4, 7 and 8: high sensitivity.

Table 8.7: Design of the reliability test, Benkovac, July 2003.

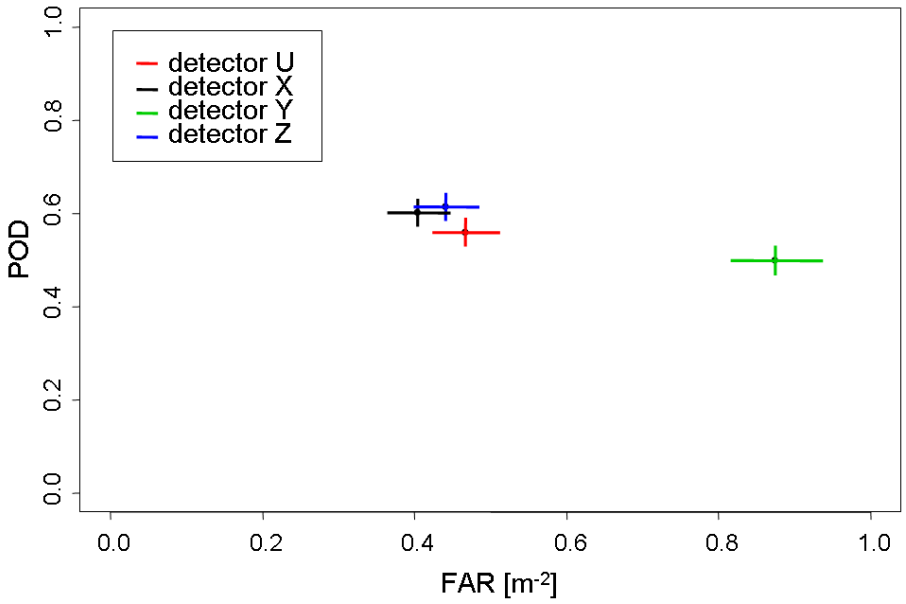

Figure 8.5: Benkovac July 2003, an ROC diagram for the high sensitivity, a comparison of detectors. The crosses indicate 95% confidence limits.

sensitivity has the effect that the PODs of all detectors become more similar when using the low sensitivity. The reason for this is that all detectors are calibrated to detect the same target on the same depth, i.e. they are calibrated to the same sensitivity. This is why it is better to select only the high sensitivity data. The same argument holds for the Oberjettenberg May tests (Section 8.3), but in those tests the low sensitivity was almost identical to the high sensitivity, so that the selection of high sensitivity was not necessary.

ROC diagrams for the high sensitivity data are in Figures 8.5, 8.6 and 8.7. Except for the sensitivity, all levels of all factors have been selected. The first of these diagrams compares the four detectors. We see clearly that one of the detectors achieved much lower detection rates and a much higher false alarm rate. A separate analysis of three soil types present in the tests showed that detector Y is comparable with other detectors in the more cooperative soil from Sisak (lanes 2 and 6), but it has serious difficulties in coping with the uncooperative soils.

The next figure, 8.6, provides a comparison between the soils. There is no significant difference between the results in the two uncooperative soils from Obrovac and Benkovac, although the Benkovac soil contains large magnetically neutral limestones (see Table 8.3 with the magnetic properties of the soils). The results in the cooperative Sisak soil are clearly higher.

The operators are compared in Figure 8.7. Operators A, B and D have

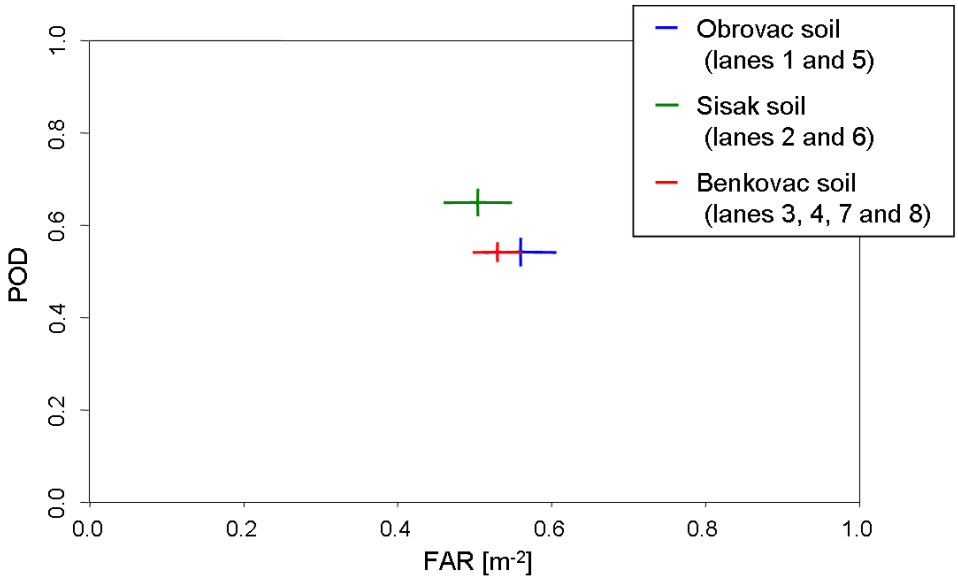

Figure 8.6: Benkovac July 2003, an ROC diagram for the high sensitivity, a comparison of soils. The crosses indicate 95% confidence limits.

achieved results obviously different from the other operators. These three operators were the only currently active deminers among the test participants, the others work in quality assurance or in more senior posts.

It should be noticed that the operators A, B, C, and D worked in the first week of the tests, and the operators E, F, G and H in the second week: the operators and the weeks are confounded. It is possible that the observed difference between the operators, or a part of that difference, is caused by the difference between the two weeks. However, the organisers and the participants of the trials could not notice any differences between the two weeks of the trials.

The design of experiment applied in these tests enables us to study the performance of metal detectors for a selected combination of a target, its depth and a soil type. Let us take the same example mentioned in Subsection 8.4.2: PMA-2, Obrovac soil (lanes 1 and 5), all depths. With an ROC diagram we can compare the PODs and the FARs of the four detectors, see Figure 8.8. Detector Y achieved much lower POD and much higher FAR than the other detectors, which are similar. We should keep in mind that the results are still not entirely free from confounding effects, as explained in Subsection 8.4.2.

We might wish to compare the eight operators for the same selection of factor levels: PMA-2 and the Obrovac soil (lanes 1 and 5). The resulting ROC diagram is Figure 8.9. However, the confounding is too strong to

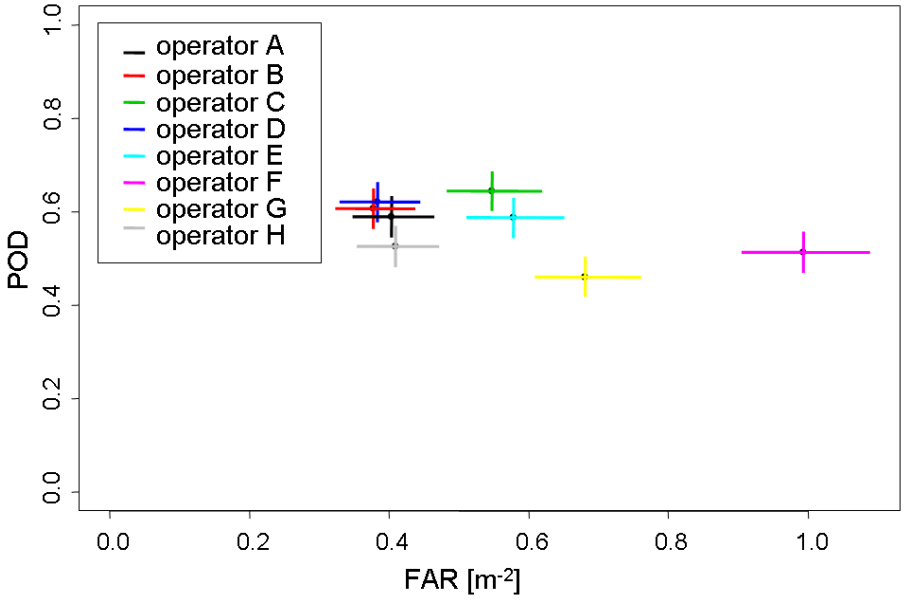

Figure 8.7: Benkovac July 2003, an ROC diagram for the high sensitivity, a comparison of operators. The crosses indicate 95% confidence limits.

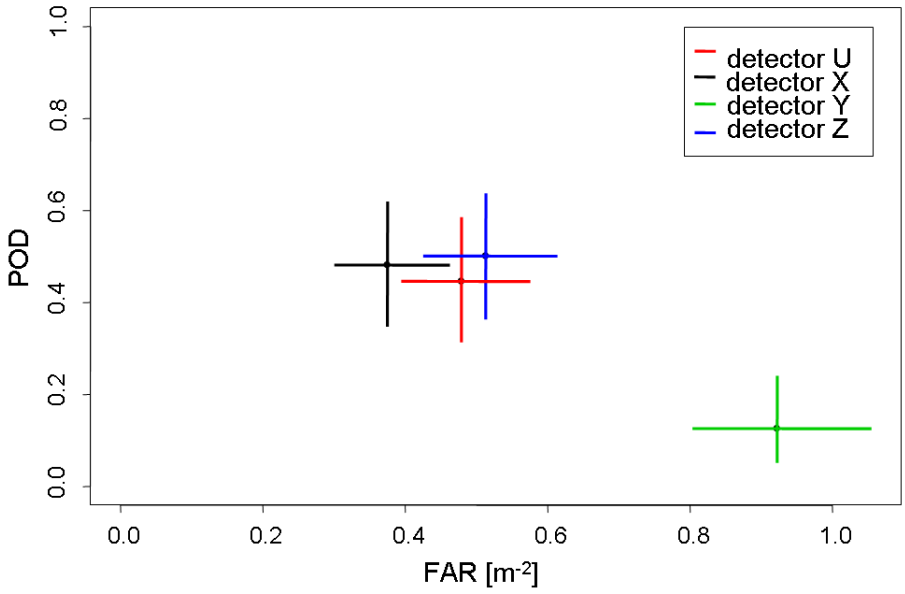

Figure 8.8: Benkovac July 2003, an ROC diagram for the high sensitivity, target PMA-2 in Obrovac soil (lanes 1 and 5), a comparison of detectors. The crosses indicate 95% confidence limits.

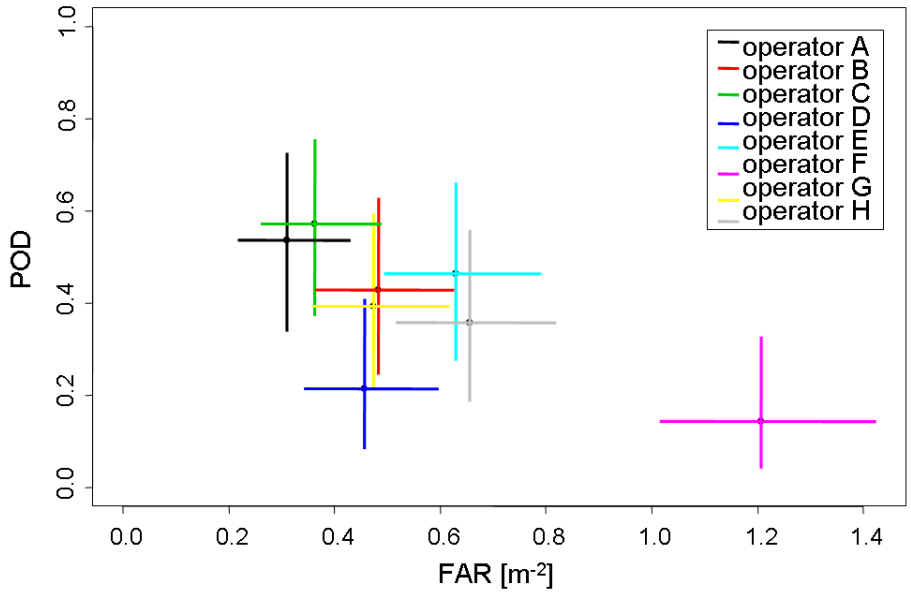

Figure 8.9: Benkovac July 2003, an ROC diagram for the high sensitivity, target PMA-2 in Obrovac soil (lanes 1 and 5), a comparison of operators. The crosses indicate 95% confidence limits. This diagram should be interpreted carefully, since it leads easily to biased conclusions.

allow us to draw conclusions about the differences between the deminers. By looking at the design of experiment, Table 8.7, we see that each operator used only two detectors. For instance, the low result of operator F might have easily been caused by the low performance of the two detectors he used in this soil, namely, beta and delta.

The systematic choice of targets and their depths enabled the creation of POD curves (they are described in Section 8.2.2). The curves on Figure 8.10 are a result of the same selection of factor levels as for the previous two diagrams, PMA-2 in Obrovac soil. We see again a clear difference between detector Y and the other three detectors. We also see that the POD falls rapidly with depth. At about 7 cm depth the POD of detectors U, X and Z falls to 0.5, and for detector Y it is below 0.5 at all depths.

The next diagram, Figure 8.11, provides a comparison between the results with the high and the low sensitivity. The difference between the two results is much more pronounced than in the Oberjettenberg May trials, see Figure 8.4 for a comparison.

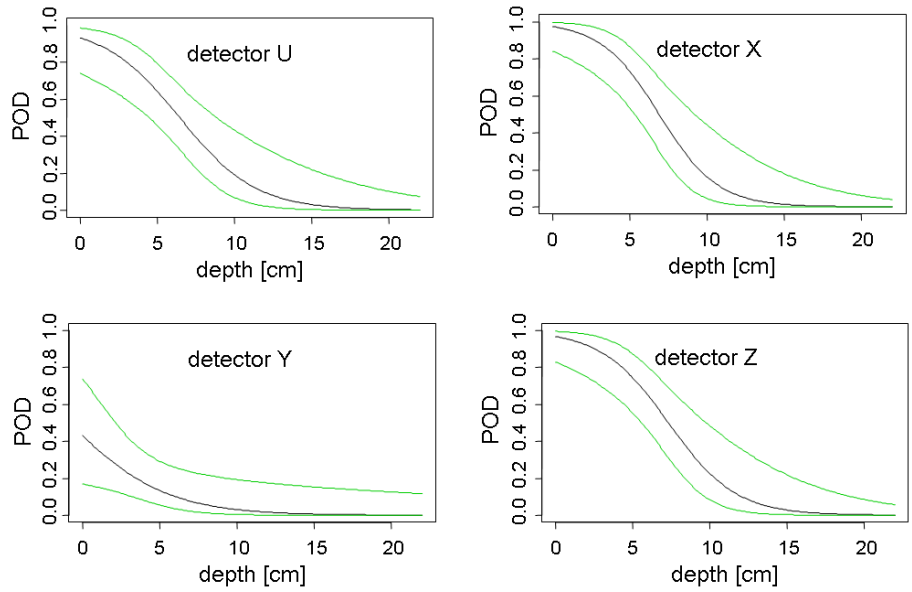

Figure 8.10: Benkovac July 2003, POD curves for the high sensitivity, target PMA-2 in Obrovac soil (lanes 1 and 5), a comparison of detectors. The crosses indicate 95% confidence limits.

8.4.4 Discussion

Since the working procedures and the training were similar to the trials in Oberjettenberg in May, the PODs are again lower than they would be in a real minefield. The reasons are discussed in the section about the Oberjettenberg May tests, Subsection 8.3.4.

The calibration procedure of the low sensitivity caused a statistically significant difference both between the PODs and between the FARs of the high and the low sensitivity measurements. The detectors should be compared using only the high sensitivity measurements, since the calibration of the low sensitivity creates similar sensitivities of the four metal detectors. It has been shown that the two points on an ROC diagram representing the high and the low sensitivity results lie on an ROC curve, but two points are not sufficient to allow further conclusions about the shape of the ROC curves.

The analysis of a selection of some factor levels was made possible due to the improved design of experiment. The changes aimed for a design closer to a full factorial design to reduce the operator-detector confounding. The bias due to this confounding is larger if the differences between the operators are more pronounced. Fortunately, it has been shown that the variability among the experienced and well-trained currently active deminers is much smaller than the variability between persons who do not work day

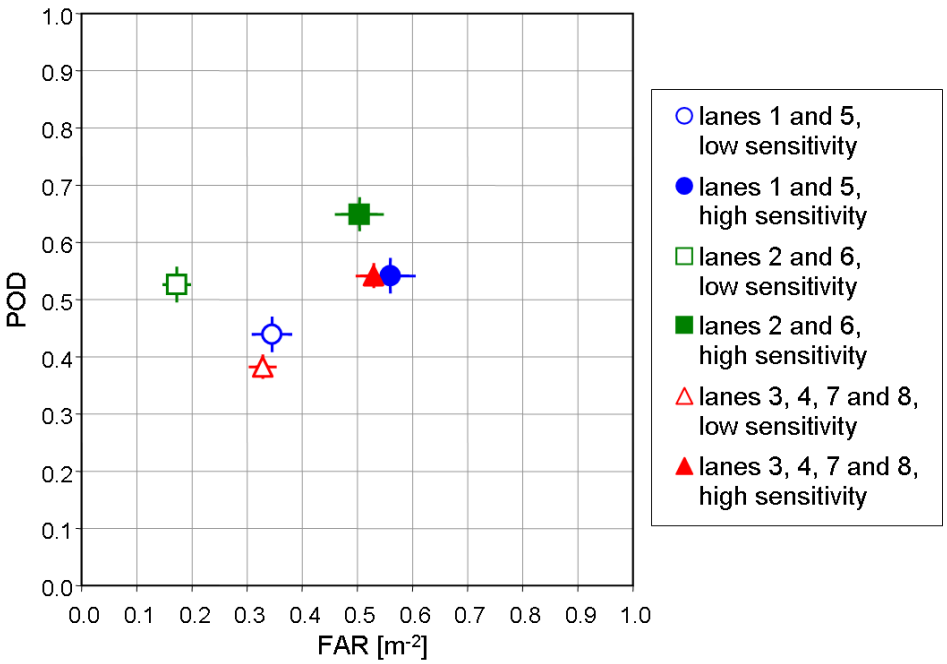

Figure 8.11: Benkovac July 2003, an ROC diagram for the complete data set, a comparison of high and low sensitivity measurements. The green lines indicate 95% confidence limits.

to day with a metal detector. (This experience is also known from some tests in non-destructive testing [77].) This is an important finding, since it is recommended to perform tests with currently active deminers who will actually perform the clearance operations. As expected, currently active deminers also achieved better results.

When some levels of some factors are selected from the full data set, the total number of opportunities to detect a target is smaller. Consequently, the confidence intervals are wider than they would be with the full data set. We see that clearly by comparing the results for the PMA-2 in Obrovac soil, measured with high sensitivity, Figure 8.8, with the overall high sensitivity results, Figure 8.5. The number of opportunities to detect a target is about 18.3 times smaller for the selection of Obrovac soil and PMA-2 (see the design of experiment, Table 8.7; and the list of targets, Table 2 in the Appendix). Equation (8.6) roughly implies that the confidence intervals are about $\sqrt{18.3} = 4.3$ times wider.

Due to metal contamination, the runs planned for lane 8 were performed in lane 4 (see Subsection 8.4.1). The negative consequence was that a source of variance was thus abolished. There were concerns that the operators might remember the positions of the targets, since they would search lane 4 more times than the other lanes. However, the results did not indicate that the operators remembered the target positions.

According to the standard CWA 14747:2003 [26] (see Section 4.2), the distance between each pair of targets should be at least 50 cm. During the preparation of the test lanes it became evident that that condition is not sufficient. Some smaller targets were not detectable when placed closer than 70 cm from a large metal content mine like TMRP-6. This problem can be solved in one of the future updates of the standard. There should be an additional requirement that all targets are distinguishable and that none of the signals is overwritten by a signal of a neighbouring target. The resolution of adjacent targets is important, but it is a subject of a separate test described in CWA 14747:2003.

8.5 Reliability Tests, Oberjettenberg, November 2003

8.5.1 Introduction

The main goal of these tests was to compare the performance of metal detectors in each soil and with each target separately. The design of experiment had been altered to meet this goal. Other important modifications compared to previous tests were the reduction of the workload on the operators and a longer training. The influence of these changes was evaluated by comparing the results of this test with the Oberjettenberg May test, using the factor

	detectors: alpha, beta	detectors: gamma, delta
Week 1	operators: A, B, C, D	operators: E, F, G, H
Week 2	operators: E, F, G, H	operators: A, B, C, D

Table 8.8: Training and testing scheme, Oberjettenberg, November 2003.

levels that were in common for both tests.

A new training scheme was applied, with a longer training then in the previous trials. The training for two groups of detectors was separated, as well as the tests (see Table 8.8). Eight operators participated in the tests and they were all soldiers of the German army with no previous experience in demining. In the first week, operators A, B, C and D were trained for the time domain detectors alpha and beta and they used only these detectors during the first week of the tests (see the design of experiment, Table 8.9). The other four operators were simultaneously trained for the other detectors, so that the tests were performed at the same time with all eight persons. In the second week the deminers switched the devices. The training took a day for a detector model, which is twice as long as in previous tests. Consequently, the operators had enough time to practice on hidden targets. The number of starts performed per day was reduced to 6, compared to the average of 8 starts per day in the previous two tests. This was a significant reduction of the work load for the operators.

Five detector models were tested: the same four models as in the previous two tests and an additional one, detector W (see Table 8.2). That detector was a new prototype and only one specimen was provided. Also the other model of the same manufacturer, detector X, was represented with only one specimen. Only the high sensitivity was used in the test.

It was planned to use eight lanes. Four of them were the same lanes as in the Oberjettenberg May trials, with most of the targets still in the ground. Four new lanes were prepared, lanes 5, 6, 7 and 8. However, lane 6 was so contaminated with metal debris, that it could not be cleared in a reasonable time and it was abandoned for the tests. The experiment has been designed for eight lanes, so that the operator scheduled for lane 6 made a break. One of the new lanes, lane 5, contained a very uncooperative soil, a mixture of magnetite and coarse sand. The soils in lanes 7 and 8 were magnetically neutral. The summary of the soil properties is provided in Table 8.3 on page 83.

Most of the targets in lanes 1 to 4 were kept from the first tests in May 2003. In the new lanes 5, 7 and 8, the targets were buried according to the scheme used in the Benkovac July 2003 tests (see Table 3 in the Appendix): the targets Maus, PMN (MS3) and PMA-S were buried to 0, 5, 10, 13 and

20 cm depth. The target PMA-S is a surrogate of the PMA-2 (see Figure 7.1 on page 67). In the analysis, the targets PMN and MS3 were treated as the same target, since they have exactly the same metal content and almost identical shapes.

8.5.2 Design of Experiment

The experimental design of the Oberjettenberg November 2003 tests, Table 8.9, is also based on the Graeco-Latin square, but it is more complex than the previous two tests. Seven operators worked at the same time. (Actually, they could not all work simultaneously, because of electromagnetic interference between their detectors, but the runs of the same start were executed shortly one after the other.) The starts of the first two quarters were executed in the first week and those of the second two quarters in the second week. It is important that each operator used only two or three detector models each week. In the second week they have switched the detector models (see Table 8.8). By some authors, such an approach is called *crossover design* [71].

This design is a full factorial design if we count the operators, the detectors and the lanes as the only factors and leave out the starts. In other words, all factor level combinations of the factors 'operator', 'detector' and 'lane' are present in the test. Let us take the example of detector alpha in lane 1. We can read from Table 8.9 that each of the eight operators used detector alpha in lane 1. As a consequence, there is no more confounding between the factors 'operator', 'detector' and 'lane'. Consequently it is possible to compare the detectors in each lane separately.

8.5.3 Results

An ROC diagram for the complete set of data, Figure 8.12, contains all factor levels, i.e. all targets, lanes and operators. The confidence intervals are wider for detectors X and W, since only one specimen of these detector models had been tested (see Subsection 8.5.4).

Figure 8.13 is a ROC diagram based on the same data set, but comparing four lanes. The same as Figure 8.2 in the Oberjettenberg May tests, this diagram should be interpreted as a comparison of the lanes, and not of the soils, since the lanes contained different targets. However, the comparison of the results in lanes 5, 7 and 8 gives information about the differences between soils, since these lanes contained the same targets buried to the same depths. There is a clear difference between the three soils: the POD is the lowest in lane 5, which contained magnetite.

The differences between the operators can be read from the diagram on Figure 8.14. There are differences between the achieved false alarm rates. The POD of operator F is clearly lower than the POD of all other operators.

Quarter 1:

	Start 1	Start 2	Start 3	Start 4	Start 5	Start 6	Start 7	Start 8
Lane 1	*A alpha-1*	*H delta-2*	*C beta-1*	*F gamma-2*	*B alpha-2*	*E gamma-1*	*D beta-2*	*G delta-1*
Lane 2	*D beta-1*	*E gamma-2*	*B alpha-1*	*G delta-2*	*C beta-2*	*H delta-1*	*A alpha-2*	*F gamma-1*
Lane 3	*C alpha-2*	*F delta-1*	*A beta-2*	*H gamma-1*	*D alpha-1*	*G gamma-2*	*B beta-1*	*E delta-2*
Lane 4	*B beta-2*	*G gamma-1*	*D alpha-2*	*E delta-1*	*A beta-1*	*F delta-2*	*C alpha-1*	*H gamma-2*
Lane 5	*G delta-1*	*B alpha-2*	*E gamma-1*	*D beta-2*	*H delta-2*	*C beta-1*	*F gamma-2*	*A alpha-1*
Lane 7	*E delta-2*	*D alpha-1*	*G gamma-2*	*B beta-1*	*F delta-1*	*A beta-2*	*H gamma-1*	*C alpha-2*
Lane 8	*H gamma-2*	*A beta-1*	*F delta-2*	*C alpha-1*	*G gamma-1*	*D alpha-2*	*E delta-1*	*B beta-2*

Quarter 2:

	Start 1	Start 2	Start 3	Start 4	Start 5	Start 6	Start 7	Start 8
Lane 1	*C alpha-1*	*F delta-2*	*A beta-1*	*H gamma-2*	*D alpha-2*	*G gamma-1*	*B beta-2*	*E delta-1*
Lane 2	*B beta-1*	*G gamma-2*	*D alpha-1*	*E delta-2*	*A beta-2*	*F delta-1*	*C alpha-2*	*H gamma-1*
Lane 3	*A alpha-2*	*H delta-1*	*C beta-2*	*F gamma-1*	*B alpha-1*	*E gamma-2*	*D beta-1*	*G delta-2*
Lane 4	*D beta-2*	*E gamma-1*	*B alpha-2*	*G delta-1*	*C beta-1*	*H delta-2*	*A alpha-1*	*F gamma-2*
Lane 5	*E delta-1*	*D alpha-2*	*G gamma-1*	*B beta-2*	*F delta-2*	*A beta-1*	*H gamma-2*	*C alpha-1*
Lane 7	*G delta-2*	*B alpha-1*	*E gamma-2*	*D beta-1*	*H delta-1*	*C beta-2*	*F gamma-1*	*A alpha-2*
Lane 8	*F gamma-2*	*C beta-1*	*H delta-2*	*A alpha-1*	*E gamma-1*	*B alpha-2*	*G delta-1*	*D beta-2*

Quarter 3:

	Start 1	Start 2	Start 3	Start 4	Start 5	Start 6	Start 7	Start 8
Lane 1	*E alpha-1*	*D delta-2*	*G beta-1*	*B gamma-2*	*F alpha-2*	*A gamma-1*	*H beta-2*	*C delta-1*
Lane 2	*H beta-1*	*A gamma-2*	*F alpha-1*	*C delta-2*	*G beta-2*	*D delta-1*	*E alpha-2*	*B gamma-1*
Lane 3	*G alpha-2*	*B delta-1*	*E beta-2*	*D gamma-1*	*H alpha-1*	*C gamma-2*	*F beta-1*	*A delta-2*
Lane 4	*F beta-2*	*C gamma-1*	*H alpha-2*	*A delta-1*	*E beta-1*	*B delta-2*	*G alpha-1*	*D gamma-2*
Lane 5	*C delta-1*	*F alpha-2*	*A gamma-1*	*H beta-2*	*D delta-2*	*G beta-1*	*B gamma-2*	*E alpha-1*
Lane 7	*A delta-2*	*H alpha-1*	*C gamma-2*	*F beta-1*	*B delta-1*	*E beta-2*	*D gamma-1*	*G alpha-2*
Lane 8	*D gamma-2*	*E beta-1*	*B delta-2*	*G alpha-1*	*C gamma-1*	*H alpha-2*	*A delta-1*	*F beta-2*

Quarter 4:

	Start 1	Start 2	Start 3	Start 4	Start 5	Start 6	Start 7	Start 8
Lane 1	*G alpha-1*	*B delta-2*	*E beta-1*	*D gamma-2*	*H alpha-2*	*C gamma-1*	*F beta-2*	*A delta-1*
Lane 2	*F beta-1*	*C gamma-2*	*H alpha-1*	*A delta-2*	*E beta-2*	*B delta-1*	*G alpha-2*	*D gamma-1*
Lane 3	*E alpha-2*	*D delta-1*	*G beta-2*	*B gamma-1*	*F alpha-1*	*A gamma-2*	*H beta-1*	*C delta-2*
Lane 4	*H beta-2*	*A gamma-1*	*F alpha-2*	*C delta-1*	*G beta-1*	*D delta-2*	*E alpha-1*	*B gamma-2*
Lane 5	*A delta-1*	*H alpha-2*	*C gamma-1*	*F beta-2*	*B delta-2*	*E beta-1*	*D gamma-2*	*G alpha-1*
Lane 7	*C delta-2*	*F alpha-1*	*A gamma-2*	*H beta-1*	*D delta-1*	*G beta-2*	*B gamma-1*	*E alpha-2*
Lane 8	*B gamma-2*	*G beta-1*	*D delta-2*	*E alpha-1*	*A gamma-1*	*F alpha-2*	*C delta-1*	*H beta-2*

Table 8.9: Design of the reliability test, Oberjettenberg, November 2003.

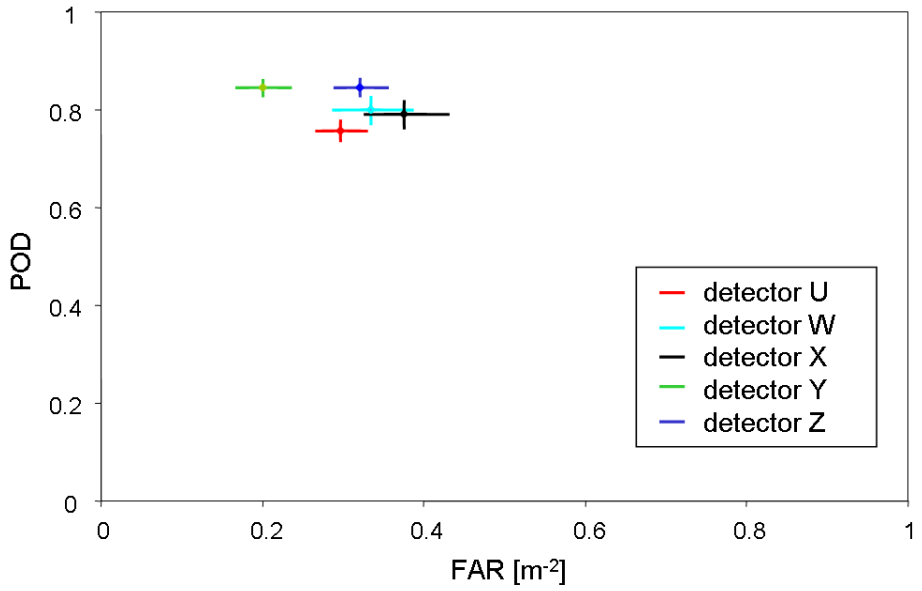

Figure 8.12: Oberjettenberg November 2003, an ROC diagram for the complete data set, a comparison of detectors. The crosses indicate 95% confidence limits.

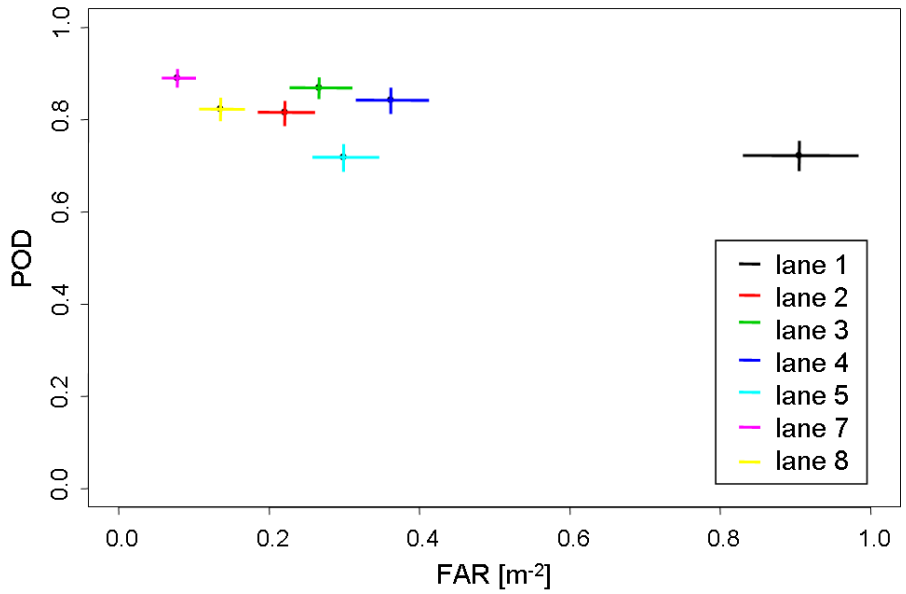

Figure 8.13: Oberjettenberg November 2003, an ROC diagram for the complete data set, a comparison of lanes. The crosses indicate 95% confidence limits.

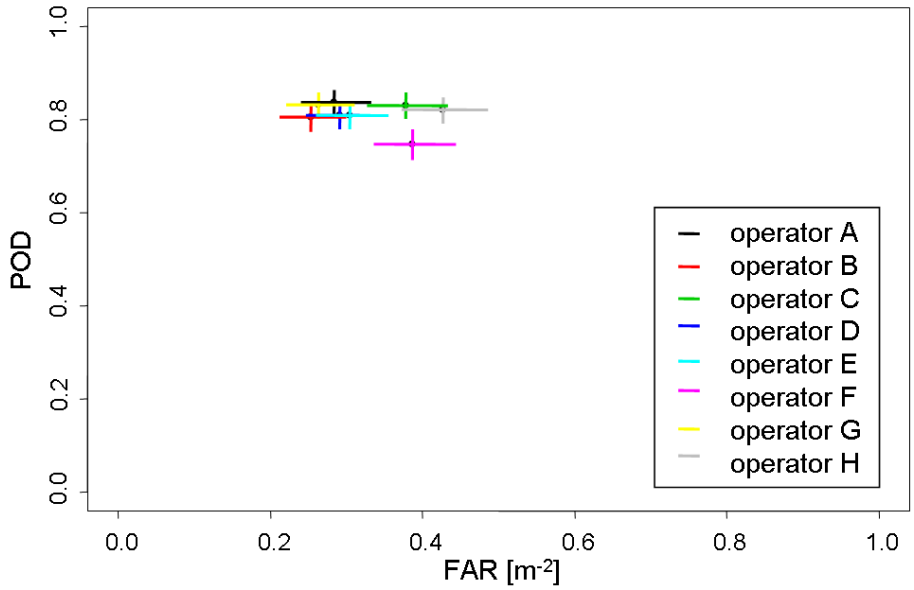

Figure 8.14: Oberjettenberg November 2003, an ROC diagram for the complete data set, a comparison of operators. The crosses indicate 95% confidence limits.

One of the easier targets in these tests was the antipersonnel mine PMN. A single measurement of the maximum detection height in air was performed with each detector model and the results were in the range between 43 and 63 cm. The antipersonnel mine MS3 has the same metal content, so it was treated as the same target. For the entire range of depths used in the reliability test (0-20 cm), the POD was very close to 1. For the selection of this target, the logistic regression model does not describe the test results adequately. To evaluate the dependence of POD on depth, the simpler method of analysis is used, described in Subsection 8.2.2 (first proposed in [42] and [76]). Each depth is analysed separately and the confidence limits are those of a binomial and a Poisson distribution, for the POD and the FAR respectively. The diagram on Figure 8.15 presents the results in lane 5, containing the uncooperative magnetite. Despite the difficult soil type, the PODs are very high and the detectors are indistinguishable. Even if all depths are counted together, the differences between the PODs are not very pronounced, as can be seen on Figure 8.16. However, this ROC diagram reveals an important difference in FAR between detector X and the others.

To evaluate the influence of the reduced workload and an improved training, we compare the ROC points of the two Oberjettenberg tests. For a valid comparison, only the factor levels common to both tests have to be chosen. Figure 8.17 shows there is a clear difference between the two test results.

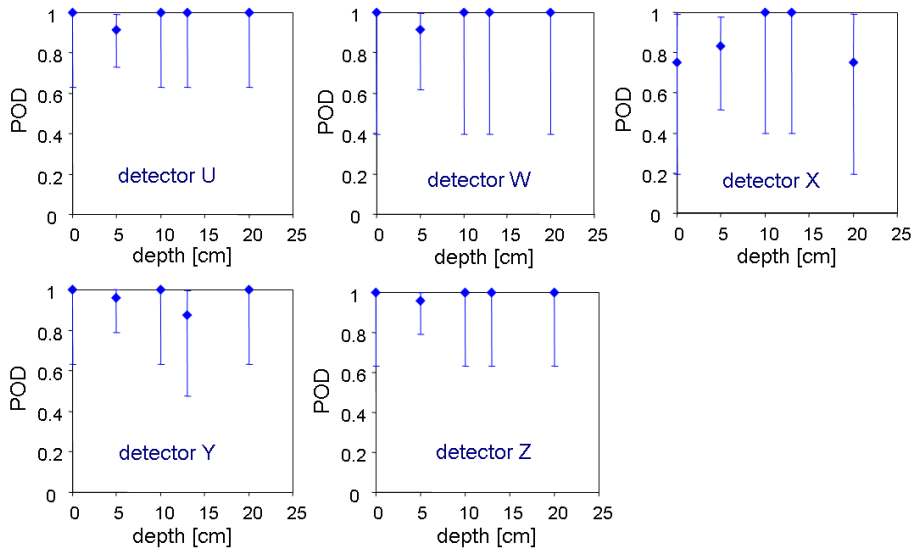

Figure 8.15: Oberjettenberg November 2003, diagrams of POD versus depth for lane 5 containing magnetite mixed with sand, targets PMN and MS3. 95% confidence limits are indicated.

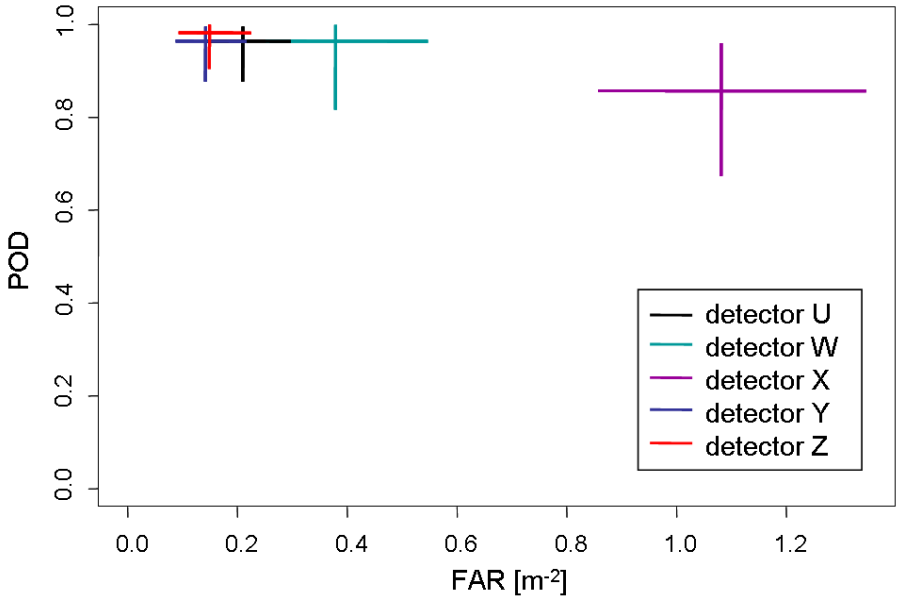

Figure 8.16: Oberjettenberg November 2003, an ROC diagram for lane 5, targets PMN and MS3, a comparison of detectors. The crosses indicate 95% confidence limits.

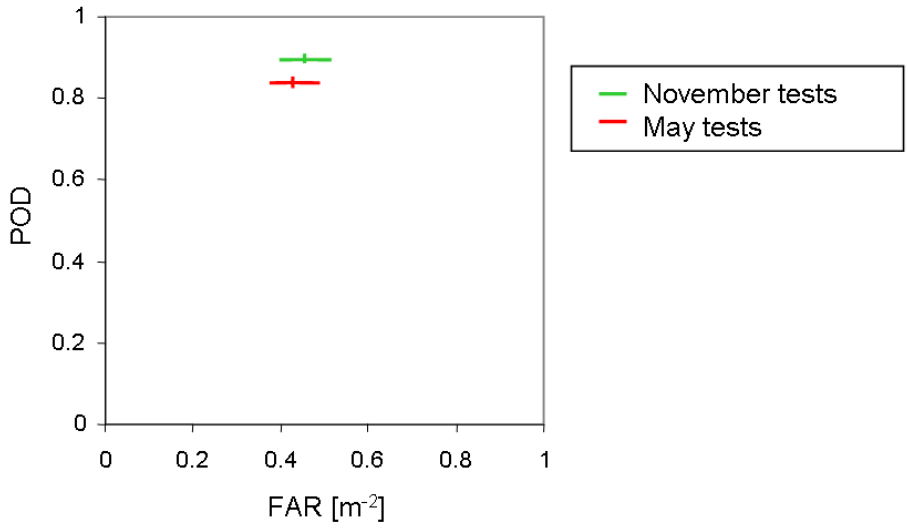

Figure 8.17: A comparison of the two Oberjettenberg tests, an ROC diagram. Only the levels common to the two tests were selected. The crosses indicate 95% confidence limits.

The POD in the May tests was 0.84, and in the November tests 0.90, while the width of the confidence interval (the difference between the higher and the lower confidence limit) is only about 0.03.

After the completion of the test, an additional test was performed. The operators indicated the positions of all alarms, they were recorded, and the operators investigated each signal until they found its source: a mine or a metal fragment. Seven operators completed only one run each, using one detector, thus excavating the targets from all seven lanes, so that the test could not be repeated. There were some opinions that such a test would produce much higher PODs because the procedure is more similar to the one applied in a minefield. However, the results did not confirm that belief. The procedure is so slow, that it produced a very small amount of data resulting in very wide confidence intervals. For example, for $POD = 0.5$ and $n = 25$ targets, the confidence intervals for the POD are approximately $2/\sqrt{N} = 0.4$ (follows straight from Equation (8.6)). It is therefore not surprising that the results in some of the lanes were lower than in the reliability test and in some lanes higher. Second, such a procedure is very much subject to confounding: the runs on the same lane cannot be repeated with other persons. For these reasons, testing with full investigation of each signal is not recommended. The results of the Mozambique trial described in Chapter 6 support this conclusion.

However, this test with the excavation of the targets produced some interesting insights. In all lanes except lane 5, almost all signals were caused

by metal fragments, which were overlooked during the preparation of the tests. This means that the vast majority of the false alarms in the detection reliability test performed earlier was actually caused by metal fragments, and not by the soil. The way to handle the problem of these false alarms is discussed in Section 8.8.2.

8.5.4 Discussion

The improvements of the experimental design enabled an unbiased comparison of detectors in each lane separately, without confounding between operators, lanes and detectors. The results were compared with the results of the May Oberjettenberg tests. The improved training scheme and the reduced workload on the operators resulted in significantly higher PODs and in no changes in FAR.

Since four lanes were kept from the Oberjettenberg May trials, many remarks regarding those trials are also valid for the November trials. As discussed in Subsection 8.3.4, the choice of target depths for the May tests was not the most suitable for the evaluation of the influence of depth. Nevertheless, those lanes were kept to provide a comparison of the two training schemes and to evaluate the influence of the workload reduction.

The confidence intervals of detectors X and W are wider than the confidence intervals of the other detectors, since only one specimen of these detector models had been tested. We can see from the design of experiment (Table 8.9) and from the number of targets in each lane (Table 3 in the Appendix) that each of these two detector models had a total of $N = 708$ opportunities to detect a target, while N for the other three detectors was $2 \cdot 708 = 1416$. Consequently, the approximate width of the confidence interval, according to Equation (8.6), is $POD_{upper} - POD_{lower} = 2/\sqrt{N} = 0.075$ for detectors W and X, and only 0.053 for the other detectors.

The testing procedure with the full investigation of the audio signals and excavation of all targets is unnecessary, since it is very time consuming and produces unreliable and biased results.

8.6 Reliability Tests, Benkovac, May 2005

8.6.1 Introduction

As in the previous two trials, the goal of this trial was to compare the metal detectors in each soil and with each target separately. The major improvement compared with the previous trials was in the treatment of the human factor. The time pressure on the operators was much smaller than in all previous trials and the operators followed a procedure which is similar to their standard operating procedures. A section leader was included in the test to supervise their work and they wore their personal protective

	detectors: alpha, beta	detectors: gamma, delta
Week 1	operators: A, B	operators: C, D
Week 2	operators: C, D	operators: A, B

Table 8.10: Training and testing scheme, Benkovac, May 2005.

equipment. The design of experiment was simplified, so that the number of factor levels was drastically reduced. The effect of all these changes was investigated by comparing the results of these tests with the results of the Benkovac 2003 tests.

Four experienced deminers operated the detectors and an additional person played the role of a section leader. The section leader supervised the work of all four deminers. After a deminer finished a run, the section leader chose randomly about 3 m of the lane and searched that area with the same metal detector the deminer was using. If he estimated that the deminer did not do his job properly, he had the authority to order him to repeat his run. The deminers wore their personal protective equipment to create a sense of work in a real mine field, thus increasing their attention. A similar training scheme was applied as in the previous trials. Each person was trained for a day for each detector model. In the first week, operators A and B were trained for the time domain detectors alpha and beta. The other two deminers were trained for the other two models. The two days of training were followed by three days of blind tests. In the next week the operators switched the detectors. This training scheme is summarised in Table 8.10. There were only four starts in each week of the tests. The number of starts performed per day did not exceed two, which is much less than in the earlier trials. Consequently, the workload was much smaller and the deminers had sufficient time for careful pinpointing of the targets.

Four detector models were tested. Detectors Z and U were the same as in the first Benkovac trials, and detectors Y and X were new models. Detector Y was a double-D frequency domain detector working in a static mode and detector X a single coil time domain one, working in a dynamic mode, so that Table 8.2 is valid also for this trial. Only one specimen of each model was tested, since no significant differences were noted between the specimens in the earlier trials. All detectors were operated at their highest achievable sensitivity.

Only four lanes were used: two lanes with the Obrovac soil and two with the Sisak soil. In the Benkovac July 2003 trials, these lanes were labeled 1, 5, 2 and 6 respectively. For simplicity they were renamed in these trials into 1, 2, 3 and 4, so that lanes 1 and 2 contained the Obrovac soil and lanes 3 and 4 the Sisak soil. Table 8.4 with an overview of soils is given on page 84.

Week 1:

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

Week 2:

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

Table 8.11: Design of the reliability test, Benkovac, May 2005.

The number of target types was also reduced. Only PMA-2 and PMA-1A were used, 15 pieces of each type in each lane, i.e. 30 targets per lane. Five pieces of each target type were buried to each of the three depths: PMA-2's were buried to 0, 5 and 10 cm (0 cm meaning just below the surface so that they are not visible), while PMA-1A's were buried to 5, 10 and 15 cm depth (see Table 4 in the Appendix).

8.6.2 Design of Experiment

The design of experiment, Table 8.11, was much simpler than the design of the earlier tests, since the number of investigated factors and factor levels was much smaller. This design is a crossover design, like the design of the Oberjettenberg November 2003 tests. Recalling that lanes 1 and 2 contained the same soil type, as well as lanes 3 and 4, we see that each detector was used by each person in each soil type. All 32 detector-soil-operator combinations are present in the test. It is thus possible to compare the performance of the four detectors in any of the soil types without bias, since there is no confounding between the operators, detectors and the soils. Strictly speaking, Table 8.11 does not represent completely the design of experiment, since the targets and their depths are also factors and are a part of the design.

8.6.3 Results

The following three diagrams provide an overview of the overall test results. The full data set is included in the analysis: both mine types, both soil

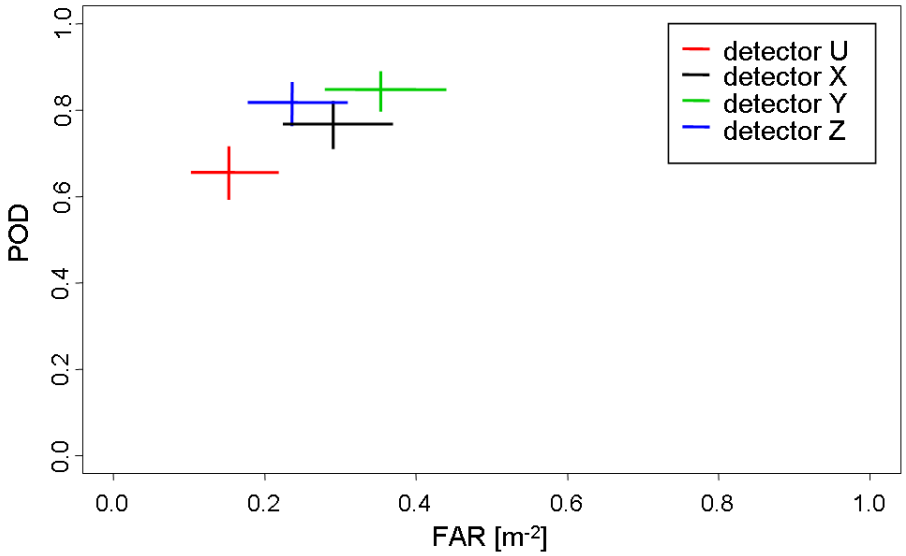

Figure 8.18: Benkovac May 2005, an ROC diagram for the complete data set, a comparison of detectors. The crosses indicate 95% confidence limits.

types, all detectors and all operators. The first diagram, Figure 8.18, is an ROC diagram showing the difference between the detector models.

The difference between the lanes is the subject of the next diagram, Figure 8.19. The false alarm rate is higher in the uncooperative Obrovac soil contained in lanes 1 and 2. The difference between the false alarm rates is much less pronounced.

No significant differences were detected between the deminers, Figure 8.20.

The next diagram, Figure 8.21, is an ROC diagram for the selection of PMA-2 and lanes 1 and 2, that is, the Obrovac soil.

Figure 8.22 is the POD curve for the same selection of factor levels. This diagram shows that the application of the nonlinear regression model described in Subsection 8.2.2 is not the most appropriate to deal with this data. It has been shown in this trial (see Section 7) that the maximum detection height has some variability, which is one of the reasons why it is likely that some of the mines buried to 0 cm depth are missed and some of those buried to 10 cm are found. However, it can easily occur that all at 0 cm depth are found and all at 10 cm are missed, as actually happened in the case of detector U in Obrovac soil and with PMA-2. In the cases when there is only one depth with $\widehat{POD}$ different from 0 or 1, the regression model will produce extremely wide confidence intervals and no conclusions about the shape of the POD curve will be possible.

Figure 8.23 presents the results for the same selection of factor levels,

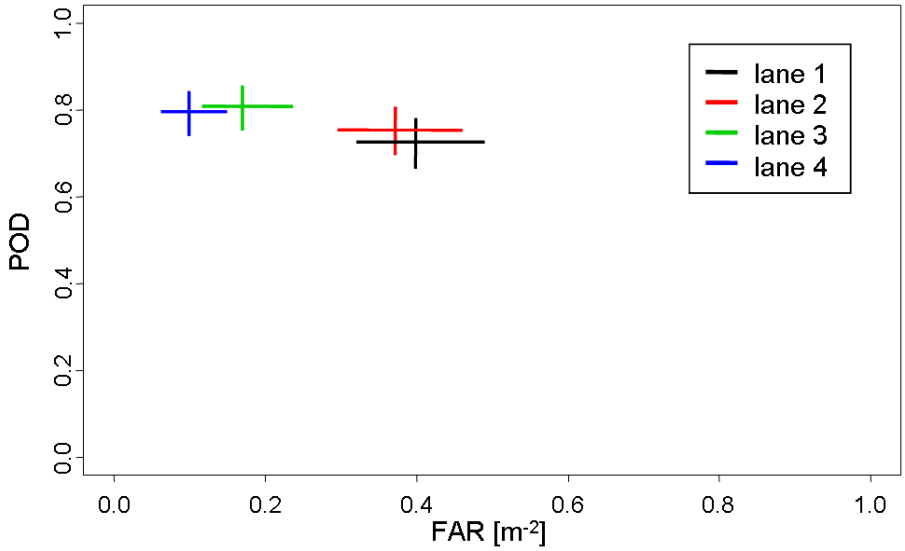

Figure 8.19: Benkovac May 2005, an ROC diagram for the complete data set, a comparison of lanes. Lanes 1 and 2 contained Obrovac soil, while lanes 3 and 4 contained Sisak soil. The crosses indicate 95% confidence limits.

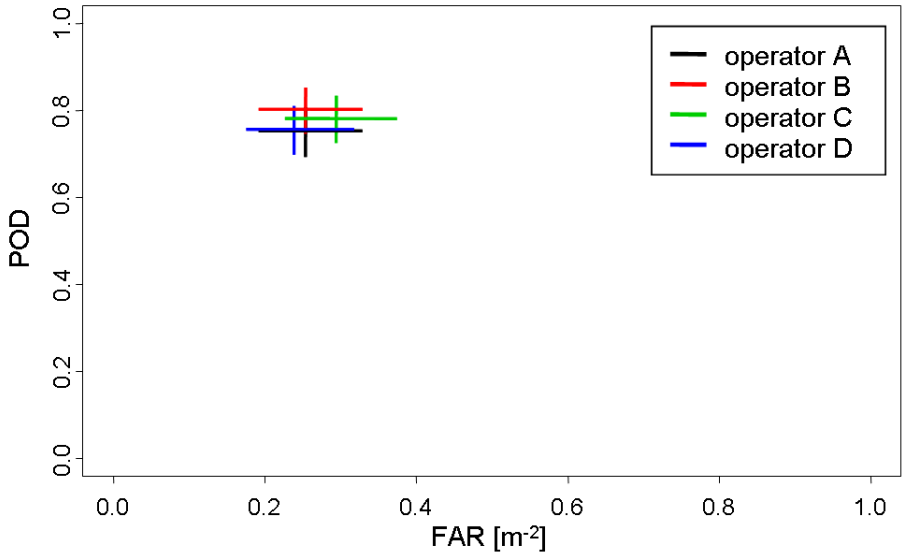

Figure 8.20: Benkovac May 2005, an ROC diagram for the complete data set, a comparison of operators. The crosses indicate 95% confidence limits.

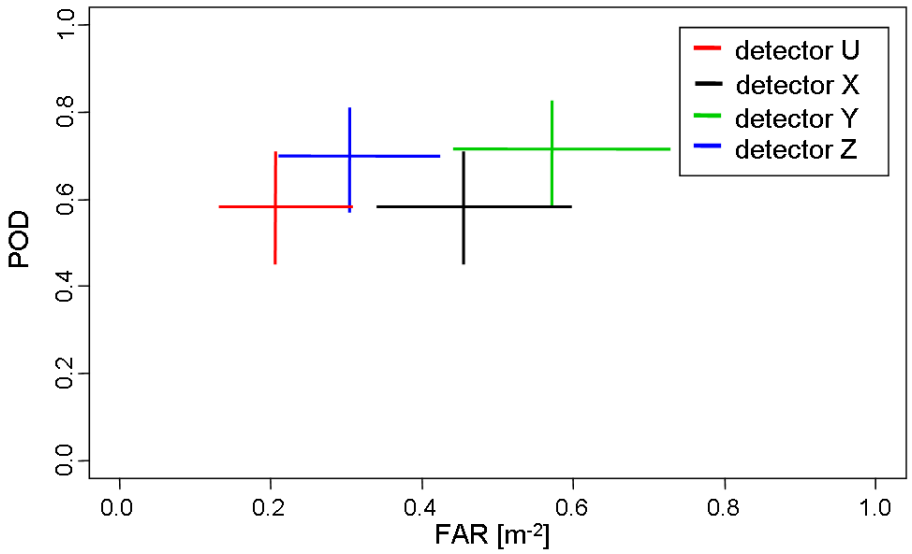

Figure 8.21: Benkovac May 2005, an ROC diagram for the PMA-2 in Obrovac soil, a comparison of detectors. The crosses indicate 95% confidence limits.

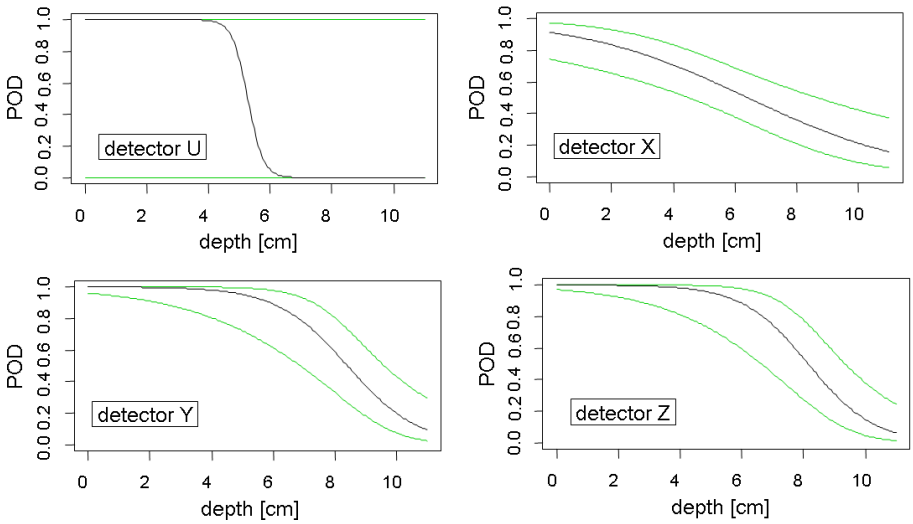

Figure 8.22: Benkovac May 2005, POD curves for the PMA-2 in Obrovac soil, a comparison of detectors. The green lines indicate 95% confidence bounds.

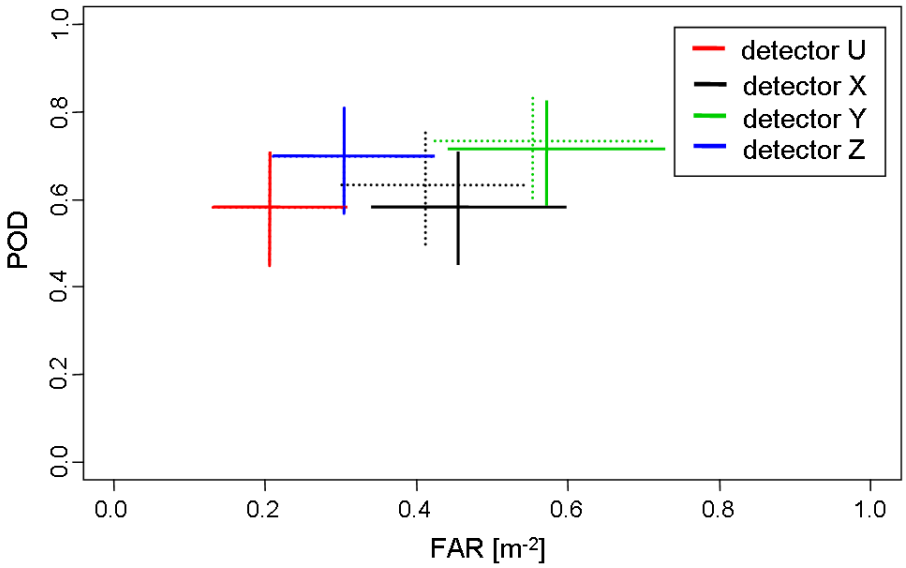

Figure 8.23: Benkovac May 2005, ROC diagrams for the PMA-2 in Obrovac soil with different halo radii. The full line and the dashed line indicate the results with the halo radius 10 cm and 13 cm respectively. The results of detectors U and Z have not changed with the increase of the halo radius. The crosses indicate 95% confidence limits.

but for two different halo radii: 10 cm and 13 cm.

The next two diagrams, Figures 8.24 and 8.25, compare the results of this test with the previous Benkovac test, that took place in July 2003 (Subsection 8.4). Only the factor levels common to both tests are selected: detectors U and Z, Obrovac and Sisak soil and PMA-2 on depths 0, 5 and 10 cm. The first of these figures contains an ROC diagram, while the other contains POD curves. There is an obvious increase in performance, both in terms of POD and FAR, especially for the shallowly buried targets.

8.6.4 Discussion

The results of the Benkovac 2005 test are noticeably better than the results of the previous tests. The most important outcome is that the detectors are easier to distinguish. This is a result of a better choice of factors and factor levels, and an improved human factor. There were very few targets with $\widehat{POD}$ close to 0 or 1, which made the detectors more distinguishable. The number of starts performed per day was reduced to one or two, which is much less than in the earlier trials, when it was sometimes higher than eight. The operators did not feel any time pressure and therefore had better pinpointing than in previous trials. The presence of the section leader improved their

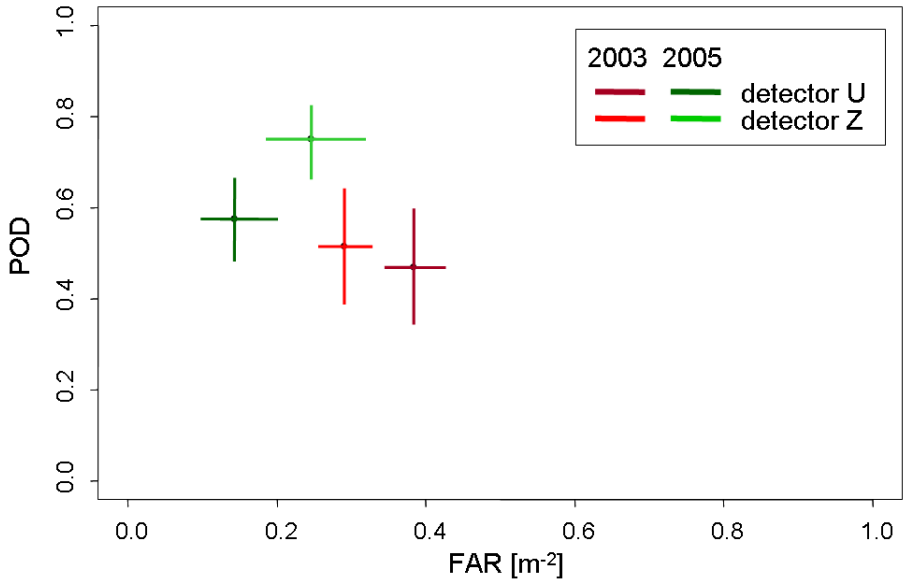

Figure 8.24: A comparison of the Benkovac July 2003 and the Benkovac May 2005 test results, an ROC diagram. Only the factor levels common for the two trials are selected: PMA-2 at depths 0, 5 and 10 cm in Obrovac soil and Sisak soil, detectors U and Z. The red crosses indicate the 2003 results, while the green ones the 2005 result. The size of the crosses indicates 95% confidence intervals.

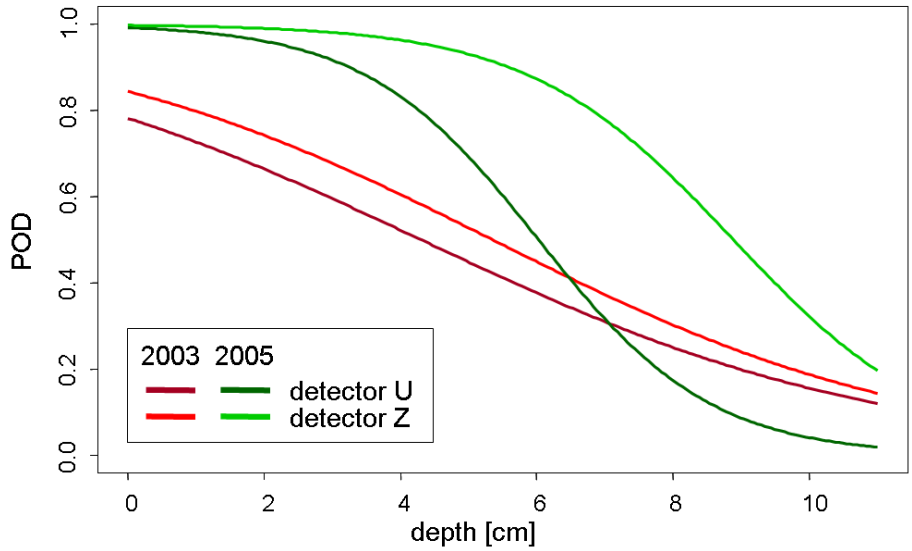

Figure 8.25: A comparison of the Benkovac July 2003 and the Benkovac May 2005 test results, POD curves. Only the factor levels common for the two trials are selected: PMA-2 at depths 0, 5 and 10 cm in Obrovac soil and in Sisak soil.

concentration.

Improved pinpointing reduces the scatter of the deminers' indications around the target position. With an improved pinpointing, more markers fall inside the halo and the POD rises. The increase of the POD will depend, among other things, on the number of markers around the target, that is, on the POD. The larger the POD, the larger its increase will be. Since shallowly buried targets have a larger POD, the increase of the POD due to better pinpointing will be larger for shallowly buried targets. This is clearly seen on Figure 8.25.

There were no significant differences between the results of the operators. This was a consequence of the choice of skilled and currently active deminers. A good training and the application of some elements of the local standard operating procedures have also contributed to the decrease of the variability between the operators. The same was observed for all selections of targets and soils.

The regression model with the logistic function was not adequate for all data selections of this test. The data can be analysed as in the example from Subsection 8.5.3, Figure 8.15, that is, each of the three depths, 0, 5, and 10 cm, can be analysed separately. To avoid this problem in the future, the targets should be buried to depths in smaller steps, only one or two targets on each depth, for example, to 0, 1, 2, 3 cm, etc. The largest depth

should be the depth at which the $\widehat{POD}$ is expected to fall between 0 and 0.5. The design would be less efficient if some targets would be buried to depths at which no detector could detect them. If some maximum detection height measurement results are available, they can guide the experimenter to choose the appropriate maximum depth.

A possibility to choose a larger halo radius than the one prescribed in the CWA 14747:2003 should be considered. The current definition of the halo size was a result of an estimate and it was not preceded by any scientific investigations. It has been shown (Figure 8.23) that a small increase of the halo radius leads to some changes in the results. The results differ due to pinpointing errors, but also due to errors of the measurements of the marker positions. Further investigations are necessary to determine the criterion for the appropriate size of the halo.

The Benkovac 2005 test results were compared with the Benkovac July 2003 results using the factor levels common to both tests. As shown on Figures 8.24 and 8.25, there is a significant improvement in the performance. For detector U only the FAR improved significantly, while for detector Z only the POD increased. The POD curves from 2005 are closer to detection rates expected to be found in actual minefields. Especially the performance at smaller depths improved. Since the same factor levels were chosen for this analysis, we conclude that the difference between the results from 2003 and 2005 is caused only by the difference of the human factor. Compared to the trials in 2003, the operators had undergone a longer training, they had much shorter working hours similar to those in the field and a section leader supervised their work.

8.7 Connection between Maximum Detection Height Measurements and Reliability Tests

There is clearly a connection between the maximum detection height (MDH) and the probability of detection in a detection reliability tests. This section describes that connection, taking the example of the MDH measurements from the Benkovac 2005 trial elaborated in Chapter 7.

Let us consider the maximum detection height measurements with a specific detector-target-soil combination. MDH measurements are subject to experimental error and they follow a certain distribution $p(MDH)$. We assume here that the MDH does not change during an MDH measurement. The targets of the same type are buried to different depths (as described in Section 7.2). When a deminer approaches one of the targets, knowing its position, he detects it with a probability that can be found from the distribution $p(MDH)$. The probability of detecting a target at depth h is actually the probability that the MDH is larger or equal to that depth. That probability is the integral of the distribution $p(MDH)$ over the depths larger

than h:

$$POD(h) = P(MDH > h) = \int_h^\infty p(MDH)\,\mathrm{d}MDH \tag{8.13}$$

This way we can construct POD curves, that is, curves of the functional dependency of POD on depth, $POD(h)$. If the distribution $p(MDH)$ is assumed to be normal, it is completely defined by two parameters: the mean and the standard deviation. These parameters completely define the POD curve too, which has the shape of the cumulative normal distribution:

$$\widehat{POD}(h) = \Phi\left(\frac{\overline{MDH} - h}{\hat{\sigma}}\right) \tag{8.14}$$

where Φ is the cumulative standard normal distribution, $\overline{MDH}$ is the average of the MDH measurements, and $\hat{\sigma}$ is the estimated standard deviation of the MDH measurements. The estimated POD has the value 0.5 at the depth equal to $\overline{MDH}$: $\widehat{POD}(\overline{MDH}) = 0.5$. At the depth $h = \overline{MDH} - \hat{\sigma}$ the $\widehat{POD}$ reaches approximately 0.84. The connection between POD curves and MDH measurements is illustrated on Figure 8.26.

We expect that these POD curves will be different from the POD curves obtained as a result of a blind detection reliability test. The searched depth will not be equal to the MDH, because the targets will not always be directly below the center of the search head. The area around the search head in which a particular target causes the detector to alarm is called a *footprint*[2] This area has an approximate paraboloid shape. The operators involved in a blind test do not know the positions of the targets, which is why it is more difficult for them to detect them. There are additional sources of variation involved with blind trials. The depths of the targets cannot be controlled as well as in MDH measurements. Each target is at a different location, with different local soil properties and a different configuration of the surface. Pinpointing is not always accurate. All these influences introduce a higher experimental error, both for the POD and for the target depths. This is why the POD curves obtained from a reliability trial will have a smaller slope than those obtained from MDH measurements as described above. In other words, the POD will not fall as abruptly as expected from the MDH measurements.

Let us apply Equation (8.14) to the maximum detection height measurements performed during the Benkovac May 2005 trials and presented in Chapter 7. Figure 8.27 refers to the PMA-2 in the Obrovac soil. The red curves are calculated from the MDH measurements according to equation (8.14), while the black curves with green confidence bounds are the results

[2]The standard CWA 14747:2003 defines the footprint differently. However, clause 6.7.2 of that standard describes a measurement procedure and calls it a footprint measurement (method 2), although it does not correspond to the footprint definition given in the standard. The definition used in this thesis adequately describes the measurements of the so called method 2.

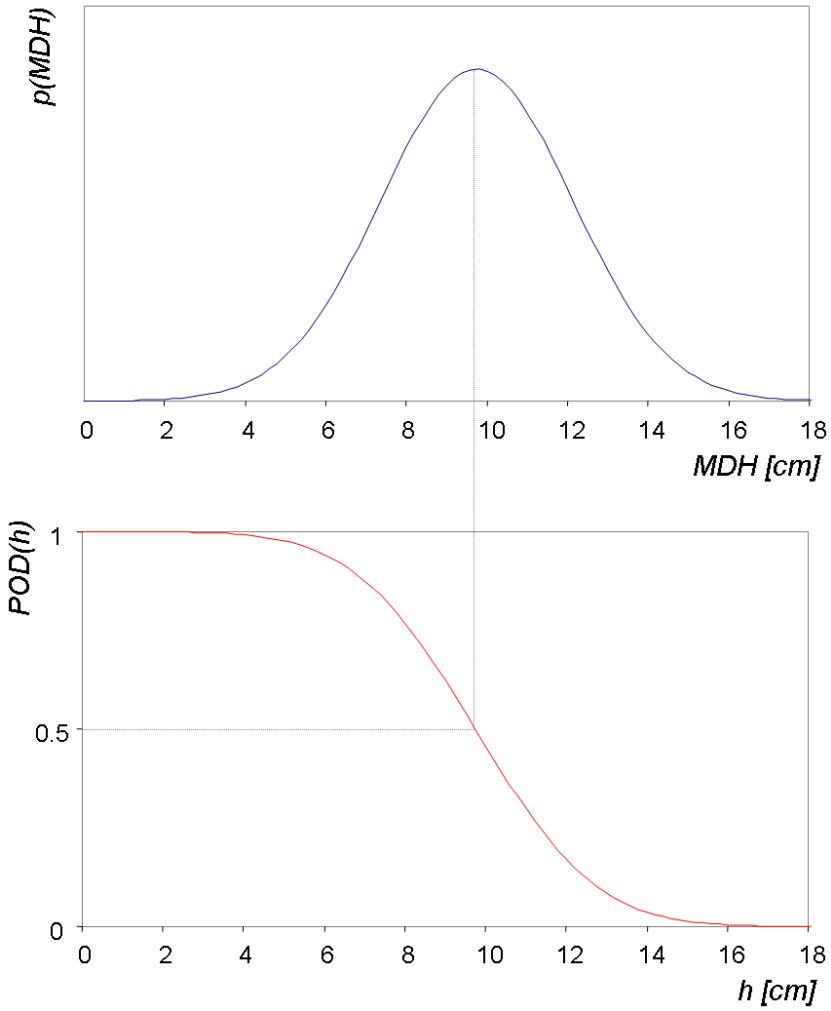

Figure 8.26: The connection between POD curves and MDH measurements. $p(MDH)$ is the distribution of the maximum detection height, and h is the target depth.

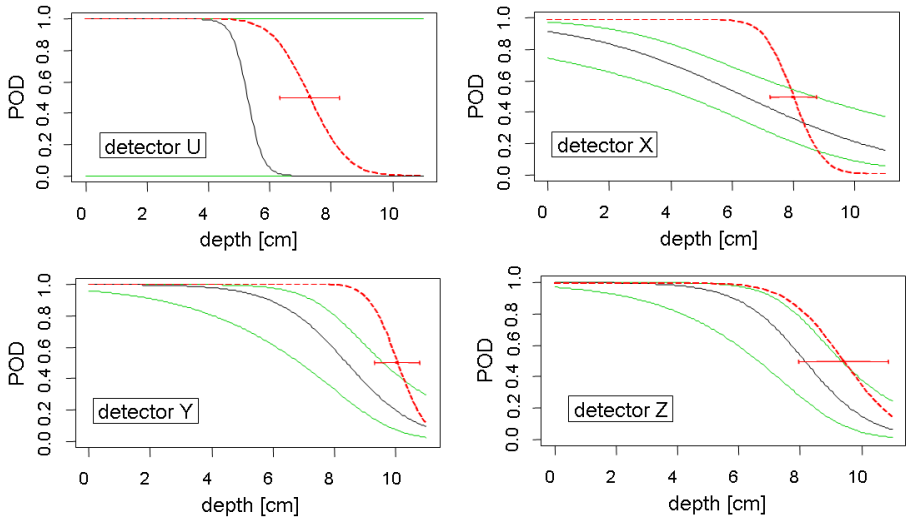

Figure 8.27: The POD curves of the reliability test compared with the POD curves calculated from the MDH measurements, PMA-2 in Obrovac soil. The black curves are the POD curves of the reliability tests, the green curves are their corresponding 95% confidence bounds, while the red curves are calculated from the MDH measurements using Equation (8.14). The red error bars indicate the average of the MDH measurements and the corresponding standard deviations.

of the reliability test. There is an excellent correspondence between the two results. As expected, the depth at which $\widehat{POD}(h) = 0.5$ in the reliability test is for all four detectors lower than the average MDH. The slope of the curves of the reliability test is higher, as predicted.

It is more difficult to compare the results in the Sisak soil, since the MDH measurements were performed on PMA-S, the surrogate of the PMA-2. It has been shown in Section 7.4 and Subsection 7.5.3 that the surrogate is slightly more difficult to detect. The diagrams for detectors U and X on Figure 8.28 are similar to the diagrams in the Obrovac soil, Figure 8.27 : the depth $h_{0.5}$ at which the reliability test gives $\widehat{POD}(h_{0.5}) = 0.5$ is smaller than the MDH, as expected. However, in the case of the other two detectors, Y and Z, this depth is larger than the MDH. This can be explained as a possible consequence of several influences. The most important one is the difference between the PMA-2 and the PMA-S: if the MDH measurements had been performed with PMA-2 instead of PMA-S, the MDH would probably have been higher. The other influence is a possible systematic measurement error of the target depths for the reliability test. Another possible influence is the local variation of the soil magnetic properties. Rather than speculating any further, we recommend that any future measurements are performed with

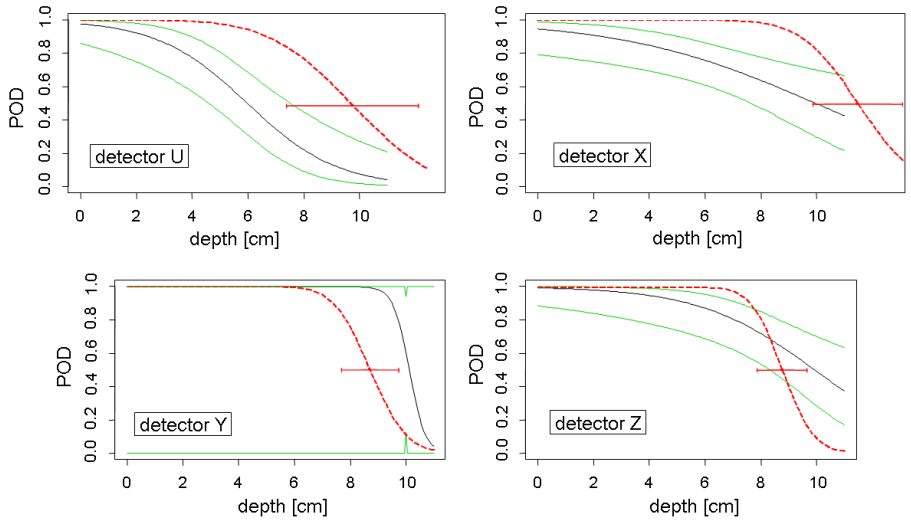

Figure 8.28: The POD curves of the reliability test compared with the POD curves calculated from the MDH measurements, PMA-2 and PMA-S in Sisak soil. The reliability tests were performed with PMA-2, and the MDH measurements with PMA-S. The black curves are the POD curves of the reliability tests, the green curves are their corresponding 95% confidence bounds, while the red curves are calculated from the MDH measurements. The red error bars indicate the average of the MDH measurements and the corresponding standard deviations.

real targets instead of surrogates, or with faithful copies of original targets.

We can conclude that the MDH measurements give us the information about the maximum possible performance in the reliability test.

It has been explained at the beginning of Section 7.4 that the MDH is defined as the largest depth at which the target has been detected. It occurred during some MDH measurements that the target was detected at depths, for example, 6, 7, 8 and 10 cm, but not on 9 cm. In such cases, the lack of detection (also called false negative indication) at 9 cm depth was ignored. Thus we obtained the results presented in Section 7.4 and used again in this section. Let us investigate if ignoring of the false negative indications had a large influence on the estimated MDH.

An MDH measurement as described in Section 7.2 can be understood as a series of Bernoulli experiments: for each target placed on a certain depth, a binary variable takes its value $y = 1$ ("detected") with the probability p and its value $y = 0$ ("not detected") with the probability $1 - p$. The same analysis applied to the reliability test results can be applied to the MDH measurements. By applying the generalised linear model described in Section 8.2.2, equations (8.10) to (8.12), we get the estimated curve of

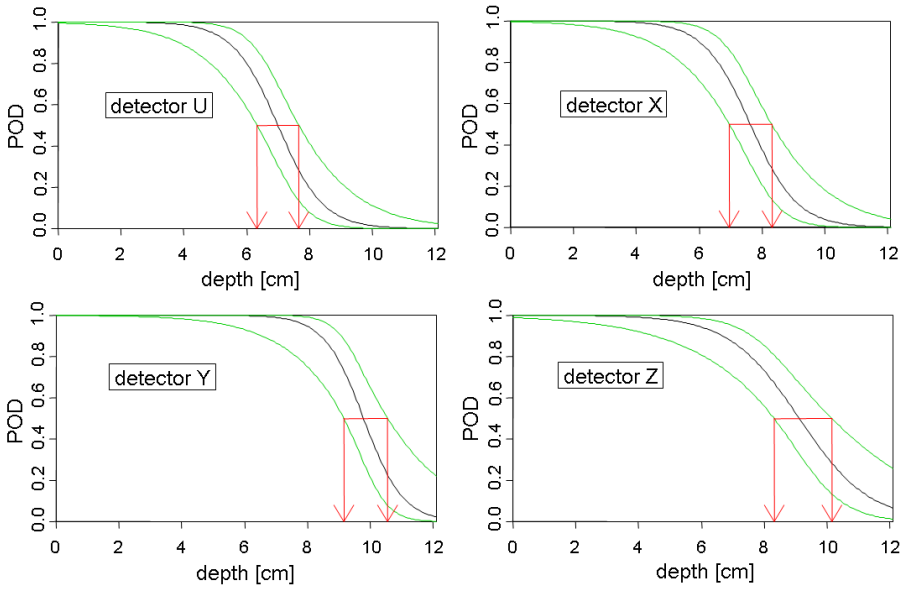

Figure 8.29: The POD curves obtained from the MDH experiment using the generalised linear model. The measurements were performed in the Obrovac soil on the target PMA-2 during the Benkovac May 2005 trials. The black curves are the POD curves, while the green curves are their corresponding 95% confidence bounds.

POD versus target depth. The POD will be 0.5 if the target depth equals the MDH. The intersection of the estimated POD curve with the straight $POD = 0.5$ defines the estimated MDH, while the intersections of the 95% confidence bounds with the straight $POD = 0.5$ define the 95% confidence bounds of the estimated MDH. The procedure is illustrated on Figure 8.29.

Let us compare the MDHs estimated with the generalised linear model with those obtained with a simpler method presented in Chapter 7. A comparison is given in Table 8.12 and Figure 8.30. The two methods give very similar estimates of the MDH. We therefore conclude that the simpler method described in Chapter 7 is adequate to describe the measurements presented in that chapter. However, in soils with less homogeneous magnetic properties, the difference between the two methods might be larger. The simpler method with the normality assumption requires that some missed targets are treated as if they were detected. The number of such targets would be higher in a heterogeneous soil. The generalised linear model overcomes that problem, since it treats each measurement with a target buried to a specific depth as a separate experiment. It is therefore recommended to use the generalised linear model whenever the discrepancy between the MDH results obtained with the two methods is too large.

	U	X	Y	Z
normal distribution	7.4 ± 0.8	8.0 ± 0.6	10.0 ± 0.6	9.4 ± 1.2
GLM	7.0 ± 0.7	$7.6^{+0.7}_{-0.6}$	9.8 ± 0.6	$9.1^{+1.1}_{-0.8}$

Table 8.12: A comparison of two methods for estimating MDH. The numbers in the table are the estimated MDHs in centimetres with 95% confidence limits. The first row contains the results obtained with the method presented in Chapter 7, which relied on the assumption that the MDH is normally distributed. The second row contains the results obtained with the generalised linear model.

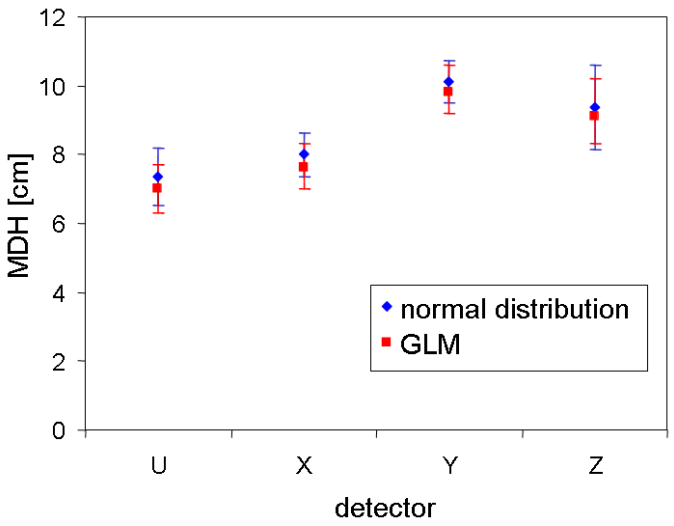

Figure 8.30: A comparison of two methods for estimating MDH. The blue points indicate the results obtained with the method presented in Chapter 7, which relied on the assumption that the MDH is normally distributed. The red points indicate the results of the generalised linear model. The error bars indicate 95% confidence limits.

8.8 Discussion

8.8.1 Representative Conditions in Tests

The end user of the demining equipment needs to know which device is the best suitable for his specific application conditions. Ideally, a test of demining equipment would be performed in conditions representative of their intended use. However, creating those representative conditions is not an easy task and many compromises have to be made to minimise the costs of the trial. An experimenter aiming for fully representative testing conditions would have to take into account the following three elements:

1. The variety of soils and mine types. The test would need to include the soils and the targets present in the area where the equipment is intended to be used.

2. The depths at which the mines are typically found. In most minefields, mines are mostly found close to the surface. A test would include a large number of shallowly buried mines and a smaller number of deeply buried mines.

3. The influence of the human factor. Recreating the same behaviour as in a real minefield is the most difficult task of all. If people know that they are participating in a test, they behave differently than in a real minefield.

The topic of this section is how to deal with these three elements with an appropriate design and planning of the test. Only the tests of detecting abilities will be discussed, other properties of metal detectors are not a subject of this section. It has been recognised in the demining community that the blind tests called detection reliability tests are the closest to the ideal of reproducing representative testing conditions. In addition, they are the only tests devised to evaluate the false alarm rate. Our discussion will therefore refer to detection reliability tests.

8.8.2 Soils and Targets

The first of the three elements that have to be considered, the choice of the soils and the mine types, is relatively straightforward and simple, compared with the other two. The end user is usually interested in the performance of the demining equipment in each soil and with each target separately. To achieve a sufficient number of runs for each target-soil combination, the experimenter has to choose a small number of soils and target types, and a large number of targets and starts. If it is known that a certain target is very easy or almost impossible to detect, it should not be included in the test.

The standard for testing metal detectors CWA 14747:2003 [26] gives a recommendation to use some ITOP inserts (defined in [80] and described in the CWA 14747:2003) to provide a reference to other trials. This is why some of these targets were included in the tests described in this chapter. However, the usefulness of such a procedure is limited. It has been shown in this chapter that the test results strongly depend on the conditions in which the tests are performed, especially on the human factor, to name just a few: working hours, training, supervision, etc. It is therefore very difficult to compare the results of two different tests. There was an idea to compare the results of the Oberjettenberg tests with those of the Benkovac tests, but the number of common targets (mostly ITOP inserts) was too small for any reliable conclusions. Consequently, the ITOP inserts were not used in the 2005 trial. The ITOP inserts still have their place in test and evaluation: they can be used for maximum detection height measurements, which are much faster than reliability tests and less subject to human factor influences, so that a comparison between trials is possible. However, it should be kept in mind that the final choice of the metal detector has to be based on measurements on the targets which represent the local threat.

If one of the goals of the experiment is to compare the influence of soils, than it is recommended to have all lanes with different soils on a single location, to allow simultaneous testing, and thus to avoid the unwanted influence of the changes in the environmental conditions, for example, rain. It is also important that all lanes contain the same targets buried to the same depths.

In all metal detector trials much effort has been put to clearing the lanes from metal clutter before the trial. The lanes were cleared with the aid of metal detectors. The experiences from all trials show that such a procedure can never result in a completely metal-free lane. After each test the positions of the deminers' indications were examined. Places with a higher concentration of false alarms were investigated. If a metal fragment was found, the indications in the proximity of that fragment were ignored; they were not counted as false alarms, neither as true positive indications. However, it is still likely that some smaller metal fragments stayed undiscovered. These fragments led to higher false alarm rates of more sensitive detectors. To reduce this problem to a minimum, two measures are necessary: before the trial, the lanes need to be thoroughly searched and cleared from metal fragments; after the trial, the deminers' indications caused by the remaining metal fragments have to be ignored.

8.8.3 Target Depths

The second element important for the design of experiment is the choice of the target depths. If we investigate Equation (8.5) from Section 8.2.2, we see that the confidence interval for the POD is the widest around $p = \widehat{POD} = 0.5$

and the narrowest close to $p = 0$ and 1. Some have argued [84] that, for this reason, the experimenter trying to detect differences between detectors should aim for an experiment with the probability of detection p close to 0 or 1, simulating a very "easy" or a very "difficult" scenario. However, this argument overlooks the fact that the differences between the detectors would in that case also be smaller. The results of Oberjettenberg November 2003 tests confirm this statement (see Figure 8.15 on page 111). The easiest way to controll the difficulty of the scenario is with the choice of target depths. The slope of the POD curve is the highest around $POD = 0.5$. According to the logistic regression model introduced in Section 8.2.2, it is exactly at $POD = 0.5$. The model based on maximum detection height measurements and introduced in Section 8.7 has the same property. If the targets are buried to the depths covering the maximum detection heights of all detectors and a bit smaller, then the POD will be around 0.5, as explained in 8.7. Since the slope of the POD curve is the highest at $POD = 0.5$, detectors with different detection capabilities will have very different PODs in a reliability test with such a choice of depths. If the same number of targets is buried shallowly (or very deeply), most of them will be detected (or missed) by all detectors and there would be no statistically significant difference between them. A larger total number of mines would be necessary to allow more reliable conclusions, which also means a longer test, with considerably higher costs. It is therefore not recommendable to choose only small depths in a test.

However, choosing only depths producing $POD = 0.5$ would also not be the best choice. The slope of the POD curve of some detectors might be very low and the POD at depth zero might not approach 1. Less stable detectors with a high variance of the MDH measurements would have that property. We would lose all knowledge about the detector performance at small depths if we did not choose smaller depths for the test. Most detectors can detect many shallowly buried targets easily, but the possible exceptions to this rule are important, since most mines in minefields are found at very small depths.

The most appropriate solution is to place the targets to depths between zero and a depth at which $\widehat{POD}$ is expected to be smaller than 0.5 for all detectors in the test (or in other words for the most sensitive detector). The MDH measurements of the most sensitive detector in the test can serve as a guidance in choosing the largest depth. If the MDH is not known during the test planning, it needs to be estimated. As discussed in Section 8.7, the $\widehat{POD}$ will be about 0.5 or below for depths equal to the MDH. It follows from Equation (8.14) that the $\widehat{POD}$ will fall bellow 0.16 at depths equal to MDH increased by the estimated standard deviation of the MDH measurements. It is therefore pointless to place some targets in a reliability test beyond that depth, since the design would be less efficient if some targets would be buried to depths at which they could hardly be detected.

For the construction of POD curves, it is recommended to bury the

targets at equally spaced depths, as, for example, in the Benkovac 2003 tests or in lanes 5, 7 and 8 of the Oberjettenberg November 2003 tests. To avoid difficulties like those in the Benkovac 2005 tests, when the regression model could not be used, the targets should be buried in smaller steps, for example, to 1, 2, 3, ... cm. Choosing fewer depths has the advantage that a simpler statistical evaluation is possible, that is, the results for each depth can be analysed separately, as done in Oberjettenberg November 2003 tests with PMN in lane 5 (see Figure 8.15 on page 111 and the corresponding discussion).

8.8.4 Human Factor

The third element guiding the design of experiment is the human factor. It has been shown in some investigations in the area of non-destructive testing that the persons involved in the test can have a large influence on the test results (see Section 3.4 and [73]). The metal detector trials described in this thesis lead to the same conclusion. The influences of the human factor are the most difficult to reproduce in a test. It follows from the discussion in Subsection 8.3.4 that most reasons for PODs lower than in reality can be attributed to the human factor. Most of all, the absence of danger decreases the alertness of the operators. The time pressure causes a faster progression through the lanes with a lower concentration than in a minefield. The training is necessarily shorter than in practice. During the test, about one day per detector model is the maximum affordable, due to limited financial resources, while it is customary to train professional deminers at least a week for a new detector model before they use it in a minefield. In a well designed experiment the operators change the detector models very often, with little time to get accustomed to the next device. All these effects need to be taken into account and all efforts have to be made to reduce them by careful planning and execution of the trials.

It is probably impossible to achieve that the operators have the same concentration as in a minefield, since the strongest motivator for thoroughness, a life threatening danger, is not present in a trial. Other ways have to be found to enhance motivation, like competitiveness, curiosity, sense of duty and importance of the work, etc.

The time pressure on the operators has to be reduced by careful planning of the trial. The number of runs per day and per person should not be too high, three or four runs on 30-m lanes are recommended.

It is very important to choose currently active deminers for tests, since they will actually perform the clearance operations. The deminers chosen for the test should be randomly chosen from the group of deminers who would later operate the detectors in real minefields.

The training should be as long as possible. However, one day of training for each detector model seems to be sufficient for the operators to feel confi-

dence in a detector. The training should be adapted to the requirements of the test. The operators do not need to master the detectors as thoroughly as they would for their work in a minefield. They should feel that they can reliably detect the targets used in the trial buried in the soils selected for the trial. Additional short "refreshment trainings" can be organised during the trial and a competent person should be available to help the operators in some problem situations and to supervise and correct their work. Further investigations are needed to evaluate the influence of training on the test results.

Most changes of the experimental design, from the first trial in May 2003 to the last one in 2005, had the aim of improving the influence of the human factor. The workload was lower in the second Oberjettenberg trial and even lower in the last Benkovac trial. Professional deminers were used in both Benkovac trials. In the first two trials, the operators had on average 8 starts per day; in the Oberjettenberg November 2003 trials, 4-6 starts; and in the Benkovac May 2005 trials, maximum 2 starts a day. Some elements of the local standard operating procedures (supervision of the section leader, personal protective equipment) were introduced to increase the attention of the operators. With a crossover design applied in Oberjettenberg November 2003 and Benkovac May 2005 trials, the operators had to switch between fewer metal detectors. That reduced the stress, since they could concentrate on fewer detector models.

All these changes are visible in the results. The two Oberjettenberg test results were compared, as well as the two Benkovac test results. The improvements of the experimental design and organisation of the trials have lead to:

1. higher PODs,

2. better distinction between the detectors,

3. smaller variability between deminers.

The PODs were especially higher at smaller depths, because pinpointing was more precise. It is equally important that the detectors are better distinguishable in the last trials. Finally, in the Benkovac 2005 trial, there were no significant differences between the results of the operators.

The finding that the currently active and skilled operators performed similarly is important. It has the consequence that any confounding between operators and any other factor is relatively small. This can help in the planning of future tests.

Some blind tests were performed in actual minefields [46, 47]. For obvious reasons detection capabilities cannot be tested in such trials; only detectors with a proved ability to detect the expected threat can be applied in a minefield. Some other properties of metal detectors were tested in this

way. The major problem of such tests is the lack of control over the testing conditions. It is very difficult or sometimes impossible to distinguish the influence of different factors, for example, the influence of the vegetation from the influence of the differences between detector models. An evaluation based on such tests is liable to many subjective estimates, and transparent reporting is very difficult. Apart from all these shortcomings, an important benefit of such trials is the opinion of the deminers who participate in the trials and who would work with the tested equipment.

There were some ideas to perform a detection reliability test in a continuation of a live minefield [6], or in a simulated minefield, so that the deminers would believe they are in a minefield and they would not know that their detectors are being tested. The behaviour of deminers would thus not change. Real mines rendered safe would be used. The conditions of the test would be better controlled than in a real minefield. The detection capabilities of the detectors would be evaluated in such a trial. This is hardly achievable for many reasons.

1. Sample size. Most of the mines would have to be placed shallowly, as they are in real minefields. It has been shown in Subsection 8.8.3 that a very large area and a long time would be needed to achieve statistical significance. That would cause high costs.

2. Keeping the secrecy. Deminers often know the area and know the locations of minefields. They have often a good contact with the local population. Monitoring would be difficult, the deminers would notice that 'something unusual' is happening.

3. Safety issues after the tests. At latest after the trial, the deminers would find out that they had participated in a test. They would never be sure in future, whether they are in a real minefield, or they participate in a test. This could be dangerous for their future safety.

For the safety of the end user of the land, it is of high interest to know how many mines are actually found and how many missed. Some believe that it is possible to organise tests that would answer this question. The author of these lines very strongly believes that no tests can answer that question, even if they are designed for that purpose. All the problems discussed previously in this section speak in favour of this conclusion.

A better way to estimate the actual detection rates is to investigate accidents on areas proclaimed cleared and in minefields during clearance. Even then, the estimates of the number of missed mines will be only rough speculations, but still more reliable than speculations based on a much smaller data sample collected in a limited range of conditions reproduced in a test.

8.8.5 Improvements of Experimental Design

The former discussion was about the human factor influences and the appropriate choice of factor levels. A few words should also be said about how these factor levels are combined in an experiment, or about the choice of treatments.

The basic approach proposed in this thesis is that of the Graeco-Latin square. The modifications of the experimental design explained and discussed in this chapter enabled the analysis of a selection of some factor levels, that is, an unbiased comparison of detectors in each soil type with each target separately. The design was changed to be closer to a full factorial design with the aim to reduce the operator-detector confounding when the results are analysed in each soil type separately.

If there is some confounding between the operators and the detectors, it is smaller if the differences between the operators are less pronounced. It has been shown, especially in Benkovac July 2003 tests, that the variability among experienced well trained currently active deminers is much smaller than the variability between persons who do not work daily with a metal detector. This is an additional reason why currently active deminers should be chosen for a test. The first and the more important reason was mentioned earlier in this section: the operators should represent the persons who would use the detectors in minefields.

When some factor levels are selected from the full data set, the total number of opportunities to detect a target is smaller than for the full data set. Consequently the confidence intervals are wider. This is why the number of soils and target types should be as small as possible. However, the number of operators should be as large as possible, since operator is a nuisance factor. The targets should be buried to depths in smaller steps, as discussed earlier in this section.

8.8.6 Maximum Detection Height and Reliability Tests

The maximum detection height (MDH) measurements were compared with the reliability test results in Section 8.7. The MDH measurements provide information about the maximum possible performance of a metal detector in a reliability test. MDH measurements can therefore be used for a preselection of metal detectors before the reliability test. MDH measurements take considerably less time and their preparation requires fewer resources. However, detection reliability tests include more of the human factor influences and thus more closely represent the realistic demining conditions. It is also important that the false alarm rate can be evaluated only in reliability tests. This is why it is highly recommended that the final decision about the acquisition of metal detectors is based on the results of reliability tests.

If there is no possibility for organising a detection reliability test, then

the metal detector has to be chosen based on MDH measurements. If there is a statistically significant difference between the average MDHs of two detectors, the detector with the higher average will be chosen. If the difference is not significant, the choice has to be done based both on the averages and on the estimated standard deviations. A detector with a higher MDH might not be the best choice. Namely, a detector with the slightly higher MDH, but with a much higher standard deviation of MDH would miss more of the shallowly buried mines.

The criterion for the comparison of the tested metal detectors needs to be defined prior to the test. That criterion could be the target depth $h_{0.95}$ at which the POD equals 0.95, $POD(h_{0.95}) = 0.95$. This POD is estimated from the MDH measurements as described in 8.7. The estimated depth $\hat{h}_{0.95}$ equals

$$\hat{h}_{0.95} = \overline{MDH} - Z_{0.95} \cdot \hat{\sigma} \tag{8.15}$$

where $\hat{\sigma}$ is the estimated standard deviation of the MDH measurements and the $Z_{0.95}$ is the upper 0.95 percentage point of the normal distribution.

Another possible approach would be to specify a depth and to compare $\widehat{POD}$'s at that depth using Equation (8.14).

8.9 Conclusions

Detection reliability tests come closest to representing the real field conditions in humanitarian demining. Most important, they include a large part of the human factor influences. Each test design is a compromise between the fully representative conditions and cost effectiveness.

A reliable estimate of the detector performance is possible only with a scientifically planned experiment. Statistical design of experiment leads to smaller experimental errors and reduces bias. In this thesis, fractional factorial designs based on the Graeco-Latin square have been proposed as a solution to the experimental problem. The subsequent changes to the design enabled an unbiased comparison of detectors in each soil and with each target model separately. The results were reported in the form of ROC diagrams and POD curves, with estimated uncertainties. With the help of these tools, it was possible to distinguish the metal detectors and to evaluate their performance in dependence on target depth.

An important part of the experiment planning is the choice of factor levels. The number of target types and soil types should be as small as possible to have a larger data set for each target-soil combination. The soils and the targets have to be representative of the regions where the metal detectors would be eventually used. The target depths have to be systematically chosen to enable an evaluation of the dependency of POD on depth. The targets can be buried to depths in smaller steps and the dependency of POD on depth can be evaluated with a regression model.

The depths should be chosen so that the estimated PODs are approximately between 0.5 and 1, depending on the depth.

The human factor has a large influence on the test results and it is the most difficult to reproduce in a test. Most of all, the absence of danger decreases the alertness of the operators, and this problem is inevitably present in all tests of demining equipment. All other influences, however, can be reduced. The planned workload on the operators should not be too high, to avoid time pressure. The training needs to be sufficiently long; one day per detector model is probably sufficient. It is essential to choose currently active and skilled deminers for a test, since they would later use the tested equipment in a minefield. The application of some elements of the local standard operating procedures may have a positive influence on the attention of the operators.

The design and the organisation of the experiments described in this chapter had been improved to meet all these requirements. The improvements have caused an increase of the PODs, a better distinction of the metal detectors and a smaller variability between the operators.

The maximum detection height measurements provide the information about the best possible performance of a metal detector in a reliability test. Both the variance and the mean are important indicators of the detection capabilities of metal detectors. The variability is mostly caused by the operators and by the electronics of the devices. Only repeated measurements in a scientifically planned experiment can provide an estimate of the variance and provide reliable results and an unbiased comparison of the equipment and the personnel under test.

Chapter 9

Proposals for Update of CWA 14747:2003

This chapter contains proposals for changes of the CWA 14747:2003, the standard for testing metal detectors for humanitarian demining [26]. The proposals are based on the experiences gained during the metal detector trials described in Chapters 6, 7 and 8. These experiences include the experiences with the statistical design of experiments, data evaluation and reporting, and practical experiences from the field. The first section of this chapter deals with maximum detection height measurements, since most of the tests described in the CWA 14747:2003 are based on the maximum detection height. The topic of the second section are the detection reliability tests. The chapter is closed with conclusions, containing a summary of all recommended changes.

9.1 Maximum Detection Height Measurements

9.1.1 Uncertainty of Maximum Detection Height Measurements

Almost all tests of the detection abilities of metal detectors described in the CWA 14747:2003 have the maximum detection height (MDH) as a response variable. The CWA 14747:2003 defines the MDH as "the maximum height above a test target at which a metal detector at given settings produces a true alarm indication due to that target."

All laboratory tests of metal detectors performed up to the present and having the MDH as a response variable were one-factor tests, meaning that the variable of interest was varied, while all other predictor variables were kept constant. These tests usually included only one measurement for each factor level, so that the variability of the results could not be estimated. However, repeating the measurements is essential for MDH measurements

because they have a large variability. An obvious advantage of laboratory measurements is that they are performed in controlled conditions. However, if the experimenter disregards some predictor variables (most importantly, the detector setup and the operator), than his experimental conditions will only apparently be controlled and his conclusions will not be valid. A simple advice to repeat the measurements is not sufficient for an experimenter to design an experiment. He needs to take into account all variables influencing the MDH to design an unbiased experiment and to reach reliable conclusions.

The predictor variables associated with the MDH as a response variable and included in the CWA 14747:2003 are:

1. sweep speed,
2. setup of the detector,
3. time after the detector is adjusted for use (the effect is called sensitivity drift),
4. orientation of the search head,
5. shaft extension,
6. moisture on the search head,
7. temperature extremes (0 °C and 60 °C),
8. temperature shock,
9. battery life,
10. soil electromagnetic properties,
11. electromagnetic properties of media other than soil (magnetic stones, bricks, pottery, etc.), and
12. ground compensation performed on strongly magnetic media (e.g. magnetic rocks) influencing the MDH measurements in less magnetic media (e.g. cooperative soils).

Here we do not mention the metal detector and the target, since they are included in the definition of the MDH. Remembering the experimental results presented in Chapters 7 and 8, we see that an important predictor variable is missing in the standard: the operator. The differences between the operators have been proven to have an important influence on the results of MDH measurements (see Subsection 7.5.1). We have solid grounds to believe that the differences between operators who are not professional deminers are even larger, as it is certainly the case with the reliability tests (see Subsection 8.8.4).

Another predictor variable deserving our special attention is the setup. The setup cannot be controlled like all other predictor variables. If the experimental goal is to compare the maximum detection heights of different metal detectors, it is essential that the measurements with the same detector are performed with repeated setups. If measurements with only one setup were performed, the setup would cause a systematic error and the experimenter would come to biased conclusions.

To compare the MDHs of different detector models, the experimenter does not need to include all predictor variables in his experiment as factors. (As defined in Section 5.1, predictor variables are called factors if they are deliberately varied during the test with the intention of measuring their influence on the response variable.) Some of the predictor variables have a more important influence than the others. These are the soil (or other medium), the setup and the operator. If the experimental goal is to compare metal detectors, these variables have to be included in the experiment as factors. Their influence is more relevant because they always affect the performance of metal detectors, while the other predictor variables in most realistic situations are constant most of the time. There are possible exceptions to this rule, in which case the variable with some relevant influence should be included in the experiment.

When the predictor variables described here as less relevant are not constant, they can change the performance of metal detectors significantly. It is recommended to evaluate their influence in separate one-factor experiments, or in multiple-factor experiments, with a fixed setup, operator and medium, which can be the air. The setup and the operator would thus form a block and the measurements would be repeated with several blocks. Crossover design should be applied in all one-factor experiments, meaning that the order of execution of the measurements with two levels of the principal factor should be chosen randomly. There are some obvious exceptions to the rules exposed in this paragraph, for example, the influence of temperature extremes cannot be evaluated with the setup as a block. However, it is beyond the scope of this work to go into details of the evaluation of each factor separately.

An example of an experiment with the setup as a block are the MDH measurements performed in Benkovac in 2005 and described in Chapter 7. For the purposes of comparison of PMA-S and PMA-2, the in-air measurements in each row of Table 7.1 were performed with the same setup, the same detector and by the same person.

The setup cannot be used as a block over the levels of the factor "detector", since it is a nested factor, nested in the levels of the factor "detector". In other words, "setup no. 1" over several detector types is meaningless.

A question arises, whether it is possible to evaluate the influence of the setup independently from the experimental error. If every measurement is performed with a new setup, the influence of the setup will not

be distinguishable from other causes of the experimental error. In such an experiment, if all predictor variables listed in this subsection including the operator are included in the design of experiment, the experimental error would be caused by:

- the setup (including the operator's influence on the setup),
- the uncertainty of the distance measurements,
- the subjectiveness of the operator (without the operator's influence on the setup),
- the fluctuations of the sensitivity due to the electronics of the device, due to some unknown influences of the surroundings, or due to any other unknown influences.

If the experimenter would wish to evaluate the influence of the setup, he would have to plan a number of measurements with the same setup and the setup would need to be varied. He would thus be able to separate the influence of the setup from the remaining experimental error. The measurements with the same setup do not need to be performed with other factor levels fixed, but there need to be several measurements with the same setup. However, such an experiment has never been executed, since it would be difficult, or even impossible, to conduct it in practice. In tests, metal detectors are used with their highest achievable sensitivity. When they are set to their highest sensitivity, they can be very instable, so that they need a frequent recalibration, i.e. a repeated setup. In practice, the subjective estimate of the operator determines when a new setup is needed. When used on their highest sensitivity, metal detectors can work only a limited time with one setup and only a limited number of measurements can be performed. This time depends on many unknown factors and it can generally not be predicted.

It is therefore recommended to perform repeated measurements each time with a new setup, if the experimental goal is to compare the MDHs of several detector models. Thus the influence of the setup will be included in the experimental error (as it was done in Benkovac May 2005 trial, see Chapter 7). The influence of the setup on the experimental error is not just an obstacle that needs to be avoided, but it is a necessary part of the experiment.

The influence of the setup could be evaluated using another response variable, for example, the output voltage. Such experiments, however, lie beyond the interest of this dissertation.

From our discussions of the experimental error we see that it is pointless to reduce the uncertainty of the distance measurements beyond the level when the uncertainty is dominated by other contributions. This is why a simple in-air measurement procedure as described in Chapter 7 is adequate in most cases.

9.1.2 Layout of Test Area and Execution of Measurements

The procedures for the in-soil MDH measurements on disarmed mines are not specified in the CWA 14747:2003, although they are most important among all MDH measurements. Measurements with small targets like standard test targets simulating metal mine components are described in clause 8.2.2 of the CWA 14747:2003. It is prescribed to use a small diameter tube to place and move the targets. Measurements on mines would require larger tubes. However, the use of larger tubes is not recommended, since some detectors might produce an alarm tone as a reaction to the large cavity in an uncooperative soil. The experimental results of the Benkovac July 2003 trial (presented in the project final report [76]) have shown that a repeated burial of a target to increasing depths displaces some soil volume and thus changes the soil electromagnetic properties, consequently influencing the results. Clause 8.4.3 of the CWA 14747:2003 describes a so called "fixed-depths detection test" with targets buried to depths of 0, 5, 10, 15 and 20 cm.

If the targets are buried to depths in smaller steps, than the maximum detection height can be measured, provided that the targets have identical electromagnetic properties and that the soil is electromagnetically homogeneous. That is how the MDH measurements in the Oberjenttenberg November 2003 trials were performed (with only one measurement per each detector model) and also the Benkovac May 2005 measurements described in Chapter 7 of this dissertation.

Clause 8.4.3 of the CWA 14747:2003 describing the fixed-depths detection tests prescribes that the search head is swept over the test area at a height of 3 cm. This condition is not necessary and can be difficult to follow in practice, which is why it was never followed in any of the tests mentioned in this dissertation.

It is recommended to bury the targets in 1 cm steps, in an increasing order, placed on a board, as described in Chapter 7. The experimenter needs to make sure that the board does not change the signal of the tested metal detectors. The detector search head shall be swept over the test area at a height as close as possible to zero.

9.2 Detection Reliability Tests

9.2.1 Statistical Design of Experiment

The experimenter needs to decide how to combine the factor levels in a reliability test. Unfortunately, it is not possible to give simple prescriptions applicable to every experimental problem. Each trial is different: both the available resources and the experimental goal differ from trial to trial. The experimenter should be familiar with the principles of experimental design

and approach every trial as a new experimental problem and a new challenge. However, some general recommendations can be made.

It is recommended to use fractional factorial designs based on the Graeco-Latin square. The experimenter usually wants to compare metal detectors in each soil type and with each target separately. In order to meet this goal, the design needs to be unconfounded with respect to soil type, detector model, operator, target and target depth.

The number of soils and target types should be as small as possible, so that the number of targets in each target-soil combination can be as large as possible. The number of operators should be as large as possible, since operator can be understood as a nuisance factor, in the sense that the experimenter is not interested in the results of any particular operator. If some earlier measurements (e.g. maximum detection height measurements) imply that there is very little variation between specimens of the same metal detector model, than only one specimen can be used in a reliability test. The recommendation for the choice of target depths can be found in Subsection 9.2.5.

By applying a crossover design, it is possible to achieve that the operators work with fewer detector models at a time. This is strongly recommended, since it reduces stress on the operators.

9.2.2 Data Analysis and Reporting

It is specified in clause 8.5.7 of the CWA 14747:2003 that the number and the location of true indications, of missed targets and of false indications shall be reported. It is not specified how these numbers will be presented and there is no mention of the uncertainty of the result.

It is recommended in this thesis that the results are presented in the form of ROC diagrams and POD curves, with the corresponding measures of uncertainty, as described in Subsection 8.2.2.

9.2.3 Choice of Targets

Targets which are certainly known to be detectable with all detector models should not be included in the test. The same applies to extremely difficult targets for which the expected POD is very close to 0 for all tested detectors.

The CWA 14747:2003, clause 8.5.3, gives a recommendation to use some ITOP inserts (defined in [80] and described in the CWA 14747:2003) as standard targets, to enable a comparison between the results of different trials. However, the usefulness of such a procedure is quite limited. It is very difficult to compare the results of two different tests because the test results strongly depend on the conditions in which the tests are performed. The ITOP inserts still have their place in test and evaluation: they can be used for maximum detection height measurements, which are less subject to

human factor influences, so that a comparison between trials is possible. The in-air measurements of the maximum detection height with ITOP targets can be used to evaluate the influence of many factors, e.g. moisture on the search head, search head orientation, temperature extremes, battery life, etc. However, it should be kept in mind that the final choice of the metal detector has to be based on the measurements on the targets which represent the local threat.

9.2.4 Target Layout

It is specified in the CWA 14747:2003 that the targets shall be buried to random locations within a 1 m wide stripe placed in the middle of a lane, which is 1.5 to 2 m wide. However, this condition is not sufficient: the targets should be placed so that their entire detection halo lies within the 1-m stripe. If some halos partially lie outside of the 1-m stripe, there are two problems. The first are indications falling in a halo, but outside of the 1-m stripe. We cannot be sure if an indication falling in the halo but outside of the lane is caused by the target or by a metal fragment from the surroundings of the lane. This problem becomes more important if the lanes are narrower than prescribed in the standard. The other problem is that the operators might decide not to indicate the signal, since they are instructed to look for targets inside the 1-m stripe. If the entire halo of each target lies within this stripe, both problems are solved.

After the experimenter has chosen the target types for his test, one of his first tasks is to determine the number of targets in a lane. The size of the lane is usually known prior to the test. The location of each target is random, respecting the limitations described in the previous paragraphs. Due to these limitations, the number of targets per lane will be limited. It is, of course, desirable to have as many targets as possible. However, if the number of targets per square metre exceeds approximately 1.2, the target positions start to form some patterns, what should be avoided. It is recommendable to leave some parts of the lane empty, but empty areas are very likely to occur if the targets are placed to random positions and if their number is close to 1 per square metre. It is therefore recommended that the average number of targets per square metre is between 1 and 1.2.

9.2.5 Target Depths

The targets should be placed to depths between zero and a depth at which $\widehat{POD}$ is expected to be smaller than 0.5 for the most sensitive detector in the test. As discussed in Section 8.7, the $\widehat{POD}$ will be about 0.5 or lower for depths equal to the maximum detection height.

For the construction of POD curves, it is recommended to bury the targets to depths in equal steps. If the experimenter uses a regression model

containing an assumption about the dependency of POD on depth, than it is recommended to bury the targets in smaller steps, for example, to 1, 2, 3, ... cm. However, choosing fewer depths has the advantage that a simpler statistical evaluation is possible, without any assumptions about the dependency of POD on depth. In that case, the results for each depth will be analysed separately, based on the assumption of a binomial distribution. (an example is given in Subsection 8.5.3, Figure 8.15, page 111).

9.2.6 Operators

It is recommended in the CWA 14747:2003 that the operators "should be representative of the operators that would use the detector in the field". This recommendation should be stronger: it is essential to choose currently active deminers for tests, since they represent the persons who will actually perform the clearance operations. The deminers chosen for the test should be randomly chosen from the group of deminers who would later operate the detectors in minefields.

The operators should feel no time pressure. The number of runs per day and per person should not be too high, about three or four runs per day on 30-m lanes are recommended. If the lanes are shorter, the number of runs can be greater.

Time taken for training should be as long as possible. However, one day of training for each detector model seems to be sufficient for the operators to feel confidence in a detector.

Some elements of the local standard operating procedures could be applied to improve the concentration of the operators and to make their work similar to their daily routine. These elements could be the presence of a section leader performing quality assurance, or the wearing of personal protective equipment.

9.3 Conclusions

The standard for testing metal detectors in humanitarian demining CWA 14747:2003 needs a thorough revision. A simple specification of some details of the tests incompletely described in the standard would not adequately reflect the current knowledge. In this chapter we discussed the measurements based on the maximum detection height and the detection reliability tests.

Maximum Detection Height. Two important predictor variables are neglected in the standard CWA 14747:2003: the operator and the setup. We need to make a difference between two kinds of experiments. The first has the experimental goal of comparing the in-soil maximum detection heights of several detectors. In this kind of experiment, it is essential

that the measurements with the same detector are performed with repeated setups and with several operators. It is necessary to design the experiment so that the setup and the operators are included in the experimental design as factors. It is recommended that repeated measurements be made each time with a new setup. Thus the setup cannot be evaluated independently from the experimental error.

The second kind of experiment has the goal of evaluating the influence of variables other than the medium, setup, or the operator. These variables are: the sweep speed, the time after the detector is adjusted for use, the orientation of the search head, the shaft extension, the moisture on the search head, temperature extremes, and the temperature shock. If the influence of one of these variables proves to be important, that variable can be included in the first kind of experiments as a factor. In the second kind of experiments, it is recommended to perform one-factor or multiple-factor in-air measurements with the operators and the setup as a block, whenever possible. Crossover design should be applied in all one-factor experiments: the order of execution of the measurements with two levels of the principal factor should be chosen randomly. The first kind of experiment is more relevant than the second one, since it evaluates the influences which are always present in the field.

For the in-soil measurements with larger targets like mines, it is recommended that the targets be buried at depths varied in 1-cm steps. It is recommended that repeated measurements be conducted according to a carefully prepared design of experiment, in which the setup and the operators are varied in accordance with the principles of experimental design. The reported results should include at least the average MDHs, the estimated errors and the detailed design of the experiment.

Detection Reliability Tests. It is recommended to use fractional factorial design based on the Graeco-Latin square. The operator has to be included in the design as one of the factors. The experimenter usually wants to compare metal detectors in each soil type and with each target separately. In order to meet this goal, the design needs to be unconfounded with respect to soil type, detector model, target and target depth. This can be achieved by permuting the operators or the detectors. The application of a crossover design is recommended, since it allows the operators to work with fewer detector models at a time.

The number of soils and target types should be kept as small as possible, so that the number of targets within each target-soil combination is as large as possible. The number of operators should be as large as possible. If earlier experiments confirm that the specimens of the same detector model are very similar, than only one needs to be used in a reliability test.

When choosing targets, preference should be given to disarmed mines

representing the local threat, rather than to standard test targets simulating mine components. The targets should be placed with their entire halo areas within the 1 metre wide stripe placed in the middle of the lane. The number of targets in the lane should be between 1 and 1.2 per m^2. They should be buried to depths between zero and the maximum detection height of the most sensitive detector in the test.

It should be emphasized more strongly that professional deminers should operate the metal detectors in the trial. One day of training per detector model is probably sufficient for them to gain confidence in the detectors. The recommended daily number of runs per person is 3 to 4 if the lanes are 30 metre long. Some elements of the local operating procedures can be applied to ease the concentration of the operators.

It is recommended to report the results of reliability tests in the form of ROC diagrams and POD curves, with the corresponding measures of uncertainty. They are described in Section 8.2.2.

Chapter 10

Conclusions

The aim of this work was to set up a design of experiment for testing and evaluation of the equipment and methods used in manual mine clearance. Most of the work deals with metal detector tests, since the metal detector is the most common detection tool in humanitarian demining. Other demining methods considered in this thesis are manual excavation methods. Detectors and excavation methods need to be compared reliably, regarding their conditions of application, so that the most suitable device or demining method can be selected for given conditions of use. A reliable estimate of the detector performance is possible only with a scientifically planned experiment. Statistical design of experiment leads to smaller experimental errors and reduces bias.

Comparative Trial of Manual Mine Clearance Methods. The trial was performed in November 2004, in Mozambique, as a part of the Study of Manual Mine Clearance [36] managed by the Geneva International Centre for Humanitarian Demining. The results of the trial were highly biased, because not all methods were tested by the same personnel. In future tests, more care should be taken that the methods are tested in similar conditions. Most importantly, they should be tested by the same operators. However, these tests gave valuable lessons about the treatment of the human factor in reliability tests. The work of the deminers in a reliability test should follow a procedure similar to their standard operating procedures applied in their daily work.

Maximum Detection Height Measurements. Maximum detection height (MDH) measurements were performed during the trials in Oberjettenberg, Germany, and in Benkovac, Croatia, in 2003 and 2005. The measurements in Croatia in 2005 were the first in which the uncertainty was estimated from the measurement results. These measurements have shown that the MDH has a high variability that has to be taken into account in

all experiments. That variability was caused by the differences between the operators, by the setup, and by the remaining sources of the experimental error, which are the changing subjectivity of the operators, the instability of the hardware between two setups, and the uncertainty of the measurements of the distance between the search head and the target. A comparison with the results of reliability tests have shown that the MDH provides the information about the best possible performance of a metal detector in a reliability test.

Many experiments described in the standard for testing metal detectors CWA 14747:2003 are experiments with the MDH as a response variable. The influences of the setup and the operator are not adequately treated in the standard. It is recommended that two kinds of experiments are defined in the next update of CWA 14747:2003. The first kind should include the setup, the soil (or an other medium) and the operator as factors in the design of experiment. In these experiments, the experimental goal is to compare the in-soil maximum detection heights of several detectors in each soil type separately. The measurements with the same detector should be performed with repeated setups and with several operators. Thus the setup could not be evaluated independent of the experimental error. The experiment performed in Croatia in 2005 and described in Chapter 7 belongs to this group of experiments.

The second kind of experiments should be experiments evaluating the influence of other predictor variables than the medium, setup or the operator. These predictor variables are: the sweep speed, the time after the detector is adjusted for use, the orientation of the search head, the shaft extension, the moisture on the search head, temperature extremes, and the temperature shock. To evaluate the influence of these predictor variables, it is recommended to perform one-factor or multiple-factor in-air measurements with the operators and the setup as a block, if possible. Crossover design needs to be applied in all one-factor experiments, i.e. the order of execution of the measurements with two levels of the principal factor should be chosen randomly. If the influence of some of these variables is both relevant and statistically significant, than it can be included in the design of experiment of the first kind.

Detection Reliability Tests. The reliability tests described in Chapter 8 of this thesis were performed in Oberjettenberg, Germany, and in Benkovac, Croatia, in 2003 and 2005.

Detection reliability tests come closest to representing the real field conditions in demining. However, each test design is a compromise between fully representative conditions and cost effectiveness. In this thesis, a fractional factorial design based on the Graeco-Latin square has been proposed as a solution to the experimental problem. The subsequent changes to the

design enabled an unbiased comparison of detectors in each soil and with each target model separately. The crossover design allowed the operators to work with fewer detector models at a time. The results were reported in the form of ROC diagrams and POD curves, with the corresponding measures of uncertainty. It is recommended that the design of experiment, the data analysis and the reporting proposed in this thesis be included in the update of the standard for testing metal detectors CWA 14747:2003.

It is not possible to give exact prescriptions for a design of experiment applicable to every experimental problem. Each trial is different: both the available resources and the experimental goal differ from trial to trial. The experimenter should be familiar with the principles of experimental design and approach every trial as a new problem and a new challenge.

Appendix: Target Positions

	target	halo radius	Lane 1			Lane 2			Lane 3			Lane 4		
			x	y	h	x	y	h	x	y	h	x	y	h
1.	PPM-2	16	102	128	4	29	606	2	33	628	3	130	137	2
2.	PPM-2	16	90	627	5	31	1099	2	89	1215	3	130	928	2
3.	PPM-2	16	94	1539	4	102	1440	2	25	1658	3	129	1253	3
4.	PMN	16	35	272	5	54	1542	3	14	119	3	64	83	2
5.	PMN	16	42	1698	5	110	1877	2	86	931	2	134	390	3
6.	PMN	16	65	1442	17	98	101	15	98	745	15	45	623	15
7.	Maus	14	53	469	5	50	45	3	83	500	2	77	172	2
8.	Maus	14	88	1315	4	29	910	1	32	959	2	60	1376	2
9.	Maus	14	101	1651	5	55	1050	3	82	1422	1	80	1575	1
10.	Maus	14	29	1847	5	103	1976	2	31	1927	2	108	1901	1
11.	SchAMi DM 31	15	44	1279	7	28	1256	5	86	654	5	127	216	6
12.	TM-46	25	42	42	8	32	1800	6	80	397	5	64	289	5
13.	PT-Mi-Ba-III	11	43	178	8	96	1370	5	55	259	7			
14.	PT-Mi-Ba-III	11	75	1212	8	36	1947	6	55	1881	5			
15.	TM-62 P2	16				56	489	5	35	553	5	67	1003	6
16.	TM-62 P3	16	35	1057	7	63	814	5	73	55	6	110	729	5
17.	TM-62 P3	16	90	1919	7	32	1365	6	76	1830	5			
18.	TM-62 M	26	38	1512	7	74	678	15	33	1078	5	94	1171	15
19.	TM-62 M	26	36	555	17	35	1637	6	39	1513	16	104	1484	6
20.	C0	10	24	367	7	87	1178	5	90	162	5	121	33	5
21.	E0	10							15	211	10	78	443	10
22.	G0	10	34	1115	7	23	1885	5	56	1598	5			
23.	I0	10				24	210	10	19	678	10	43	224	10
24.	K0	10	55	899	7	46	306	5	23	1280	5	46	498	5
25.	K0	10				64	728	10	98	1665	10	45	1645	10
26.	G0	10	87	745	12	74	156	10	68	806	10	38	330	10
27.	I0	10				107	1003	5	19	1367	5	112	568	5
28.	K0	10	91	1800	12	87	1678	10				123	1657	10
29.	100Cr6 ball	11	73	85	22	87	423	20				46	812	20

Table 1: List of targets, Oberjettenberg, May 2003. The coordinates x and y mark the positions in the lane, while h denotes the depth measured from the soil surface to the top of the target. All measures in centimetres.

	target	halo radius	h	Lane 1		Lane 2		Lane 3		Lane 4		Lane 5		Lane 6		Lane 7		Lane 8
				x	y	x	y	x	y	x	y	x	y	x	y	x	y	
1.	100Cr6 ball	11	10	10	90	326	86	2230	78	1038	56	139	85	1241	72	1852	14	757	All passes were performed in Lane 4.
2.	E0	10	5	5	27	60	58	491	36	699	85	2550	62	1461	20	1055	47	498	
3.	G0	10	5	5	87	683	33	1308	70	131	69	169	74	534	31	1146	19	2285	
4.	K0	10	5	5	62	1070	93	2750	79	2175	78	1238	12	2754	90	68	21	300	
5.	PMA-1A	13	0	0	76	120	68	1900	28	2522	50	68	80	1161	23	776	77	2167	
6.	PMA-1A	13	5	5	33	1585	80	1562	24	241	73	1099	15	1104	31	359	79	848	
7.	PMA-1A	13	5	5	40	2520	23	1590	20	535	29	2273	65	1697	78	705	31	1614	
8.	PMA-1A	13	5	5	24	2798	19	2905	88	2713	19	2843	19	1965	23	1702	23	2790	
9.	PMA-1A	13	10	10	75	1154	71	628	50	1640	21	576	52	2919	30	110	23	575	
10.	PMA-1A	13	13	13	16	1006	30	976	40	2770	60	961	20	66	77	1998	29	1984	
11.	PMA-1A	13	20	20	87	405	84	130	22	1941	65	1520	30	1531	76	1194	23	426	
12.	PMA-2	11	0	0	60	2698	22	1705	88	2821	44	2128	23	2350	53	1264	22	39	
13.	PMA-2	11	5	5	24	624	29	228	43	470	81	389	35	1877	87	394	31	1166	
14.	PMA-2	11	5	5	22	2120	30	709	52	628	30	1180	45	2050	17	1657	47	1868	
15.	PMA-2	11	5	5	88	2612	29	1199	20	1350	58	1624	80	2407	35	2788	78	2841	
16.	PMA-2	11	10	10	81	2788	25	2268	72	1154	51	2468	64	1304	59	913	43	1056	
17.	PMA-2	11	13	13	38	939	88	2078	79	282	30	332	59	140	32	1943	29	2980	
18.	PMA-2	11	20	20	86	863	51	2551	78	1230	57	2753	22	1401	38	254	82	70	
19.	PMA-3	10	0	0	56	2878	20	430	80	1390	57	519	30	616	17	2524	56	1303	
20.	PMA-3	10	5	5	18	1374	78	380	68	389	60	261	69	689	72	1395	69	1548	
21.	PMA-3	10	5	5	83	1695	66	2812	38	760	55	1355	86	1633	71	1691	78	2346	
22.	PMA-3	10	5	5	76	2222	85	2933	74	1550	59	2024	20	2532	15	2126	63	2637	
23.	PMA-3	10	10	10	38	510	50	2353	17	1490	18	1865	48	2851	61	1588	44	1465	
24.	PMA-3	10	13	13	23	1221	82	1029	20	344	58	1715	41	809	74	998	27	2579	
25.	PMA-3	10	20	20	76	2144	80	1130	64	2645	71	2219	49	961	55	2606	44	2441	
26.	PROM-1	14	0	0	15	1795	78	1463	79	1706	62	666	25	492	65	2275	57	947	
27.	PROM-1	14	0	0	15	2936	28	2428	66	1790	53	1437	77	1826	81	2732	20	2702	
28.	PROM-1	14	5	5	47	786	58	1796	21	65	58	1819	20	213	65	2193	61	369	
29.	PROM-1	14	5	5	61	1450	67	2152	50	2240	41	1934	76	452	45	2415	46	1743	
30.	PROM-1	14	5	5	81	2447	86	2470	58	2403	68	2915	24	2626	84	2928	85	2528	
31.	TMA-3	22	10	10	68	1843									56	545	36	165	
32.	TMA-4	23	10	10			84	273	40	2920	49	2653	81	2197					
33.	TMRP-6	19	10	10	40	1510	22	1968	57	961			54	290					
34.	TMM-1	26	10	10							60	851			30	1517	32	675	

Table 2: List of targets, Benkovac, July 2003. The coordinates x and y mark the positions in the lane, while h denotes the depth measured from the soil surface to the top of the target. All measures in centimetres.

	target	halo radius	Lane 1			Lane 2			Lane 3			Lane 4			Lane 5			Lane 7			Lane 8		
			h	x	y	h	x	y	h	x	y	h	x	y	h	x	y	h	x	y	h	x	y
1.	100Cr6 ball	11	10	102	74	10	96	425	10	98	1660	10	89	815	10	38	960	10	82	1300	10	55	1905
2.	C0	10	5	94	240	5	96	1180	5	88	162	5	124	31	5	81	358	5	34	820	5	48	810
3.	E0	10	5	34	1608	5	96	1770	5	20	1370	5	119	1657	5	96	1391	5	96	1925	5	81	1376
4.	G0	10	5	98	1025	5	23	1878	5	64	1598	5	81	440	5	60	1311	5	47	745	5	93	285
5.	I0	10	5	37	652	5	77	182	5	68	808	5	41	345	5	64	1520	5	27	78	5	34	875
6.	K0	10	5	71	953	5	58	306	5	27	1285	5	46	498	5	31	298	5	95	1790	5	45	225
7.	Maus	14	5	56	470	3	50	49	2	83	500	2	72	177	20	42	210	20	93	283	20	86	1003
8.	Maus	14	4	89	1315	1	28	910	2	31	960	2	60	1376	0	106	260	0	37	495	0	81	1838
9.	Maus	14	5	107	1653	3	55	1060	1	87	1427	1	79	1585	5	82	625	5	91	872	5	35	730
10.	Maus	14	5	32	1847	2	104	1972	2	31	1927	1	105	1906	13	33	1140	13	86	1423	13	90	396
11.	Maus	14													5	38	1440	5	75	1530	5	67	1295
12.	Maus	14													10	95	1759	10	53	1857	10	29	1107
13.	Maus	14													5	93	1850	5	32	1732	5	31	1632
14.	PMN	16	5	35	270	15	96	105	3	14	119	2	60	86	5	38	533	5	104	646	5	64	660
15.	PMN	16	17	70	1440	3	56	1556	15	91	750	3	128	392	0	90	750	0	85	225	0	37	1186
16.	PMN	16	5	46	1698	2	110	1875	2	84	934	15	49	630	5	92	1244	5	37	678	5	50	1770
17.	MS3	16													10	39	671	10	45	610	10	71	133
18.	MS3	16													13	90	840	13	92	1599	13	96	1135
19.	MS3	16													20	94	995	20	91	450	20	89	1966
20.	MS3	16													5	33	1680	5	35	1135	5	45	1495
21.	PPM-2	16	4	103	123	2	29	607	3	37	633	2	121	140									
22.	PPM-2	16	5	88	624	2	31	1099	3	89	1215	2	130	928									
23.	PPM-2	16	4	98	1540	2	103	1438	3	28	1658	3	129	1253									
24.	PT-Mi-Ba-III	11	8	41	178	5	91	1368	7	55	259												
25.	PT-Mi-Ba-III	11	8	75	1211	6	38	1945	5	57	1876												
26.	SchAMI DM 31	15	7	47	1278	5	28	1256	5	86	654	6	120	216									
27.	TM-46	25	8	42	38	6	35	1811	5	78	397	5	62	301									
28.	TM-62 M	26	17	44	459	15	72	675	5	33	1079	15	94	1171	10	67	480	10	83	371	10	86	1703
29.	TM-62 M	26	7	42	1506	6	36	1639	16	36	1510	6	98	1490									
30.	PMA-S	11	5	58	491	5	36	558							5	37	37	5	91	525	5	27	63
31.	PMA-S	11													0	82	111	0	94	1084	0	84	1240
32.	PMA-S	11													13	96	170	13	70	1220	13	88	1572
33.	PMA-S	11													20	23	800	20	78	1017	20	45	560
34.	PMA-S	11	7	40	1054	5	61	816	6	75	56				5	52	885	5	91	1671	5	98	768
35.	PMA-S	11				6	36	1365	5	77	1828				10	54	1070	10	100	118	10	33	1423
36.	PMA-S	11													5	79	1620	5	58	1962	5	50	955

Table 3: List of targets, Oberjettenberg, November 2003. The coordinates x and y mark the positions in the lane, while h denotes the depth measured from the soil surface to the top of the target. All measures in centimetres.

				Lane 1		Lane 2		Lane 3		Lane 4	
	mine type	**halo r.**	**depth**	x	y	x	y	x	y	x	y
1.	PMA-1A	13	5	31	90	35	343	33	362	71	196
2.	PMA-1A	13	5	85	704	63	496	20	947	18	233
3.	PMA-1A	13	5	74	1147	76	1978	48	1553	53	910
4.	PMA-1A	13	5	80	1736	16	2490	54	1721	25	2000
5.	PMA-1A	13	5	72	2368	43	2862	78	2072	60	2482
6.	PMA-1A	13	10	70	50	79	288	30	479	21	121
7.	PMA-1A	13	10	82	265	32	878	27	660	24	1243
8.	PMA-1A	13	10	57	433	72	1833	17	1155	83	2023
9.	PMA-1A	13	10	41	820	71	2183	72	1607	14	2538
10.	PMA-1A	13	10	66	2567	74	2393	73	2521	41	2747
11.	PMA-1A	13	15	36	718	62	401	72	524	40	511
12.	PMA-1A	13	15	25	1550	34	787	53	890	74	1113
13.	PMA-1A	13	15	65	2010	82	1349	64	1863	84	1272
14.	PMA-1A	13	15	19	2256	20	1654	65	1940	75	1695
15.	PMA-1A	13	15	84	2634	75	2617	16	2740	77	2546
16.	PMA-2	10	0	40	1215	69	50	64	87	71	449
17.	PMA-2	10	0	85	1450	38	996	59	1009	47	1366
18.	PMA-2	10	0	25	1745	48	1582	16	2114	85	2266
19.	PMA-2	10	0	70	2776	31	1915	74	2340	30	2626
20.	PMA-2	10	0	12	2820	64	2108	86	2443	55	2847
21.	PMA-2	10	5	56	912	17	80	28	2012	77	115
22.	PMA-2	10	5	26	1116	19	151	40	2218	21	680
23.	PMA-2	10	5	83	1340	73	1078	42	2396	63	800
24.	PMA-2	10	5	42	1618	22	2319	23	2548	29	1589
25.	PMA-2	10	5	62	1952	45	2544	67	2790	23	2288
26.	PMA-2	10	10	27	548	16	1990	24	283	33	31
27.	PMA-2	10	10	50	633	52	2259	82	1089	28	372
28.	PMA-2	10	10	34	1362	20	2685	20	1261	61	637
29.	PMA-2	10	10	52	2443	70	2732	47	1808	67	2212
30.	PMA-2	10	10	27	2625	82	2819	64	2668	77	2689

Table 4: List of targets, Benkovac, May 2005. The coordinates x and y mark the positions in the lane, while h denotes the depth measured from the soil surface to the top of the target. All measures in centimetres.

Bibliography

[1] M. Acheroy. Mine Action Technologies: Problems and Recommendations. *Journal of Mine Action*, 7(3), Dec 2003. Available at `http://maic.jmu.edu/`.

[2] Mine Detector Trial Report. Technical report, UN Mine Action Programme for Afghanistan, 2000. Conducted in Pakistan and Afghanistan Sep-Oct 1999 and Feb-Mar 2000. Not publicly available. Distribution of this report to be authorised by Mine Action Programme for Afghanistan.

[3] Summary of Metal Detector Trial Report. Technical report, UN Mine Action Programme for Afghanistan, 2002. Available at `http://www.itep.ws/`.

[4] H. Bach. Mine Action Technology Now and In The Future: Is it Realistic to Expect Great Leaps Forward in Technology? *Journal of Mine Action*, 6(1), Dec 2002. Available at `http://maic.jmu.edu/`.

[5] Workshop on Reliability Tests for Demining, Dec 16-17 2003. Presentations from the Workshop and a summary of the Breakout Sessions are available at `http://www.kb.bam.de/ITEP-workshop-03/`.

[6] Round Table Discussion on Reliability Tests in Demining, Dec 15-16 2005. Presentations from the Workshop are available at `http://www.kb.bam.de/GICHD-BAM-workshop-05/`.

[7] Bartington Instruments LTD. *Operation Manual for* MS2 *Magnetic Susceptibility System.* Bartington Instruments LTD, 10 Thorney Leys Business Park, Witney, Oxford, OX28 4GG, England.

[8] A. P. Berens. NDE Reliability Data Analysis. *Metals Handbook, ninth edition*, 17:689–701, 1988.

[9] S. D. Billings, L. R. Pasion, D. W. Oldenburg, and J. Foley. The Influence of Magnetic Viscosity on Electromagnetic Sensors. In *EUDEM2-SCOT 2003, International Conference on Requirements and Technologies for the Detection, Removal and Neutralization of Landmines and*

UXO, Brussels, Belgium, Sep 15-18, 2003. Available at `http://www.eudem.vub.ac.be/`.

[10] T. J. Bloodworth. Development of Tests for Measuring the Detection Capabilities of Metal Detectors. Technical Note I.03.168, JRC, Joint Research Centre of the European Commission, Nov 2003. Available at `http://www.itep.ws/`.

[11] T. J. Bloodworth and A. J. Sieber. Standardized Test and Evaluation of Metal Detectors. In *EUDEM2-SCOT 2003, International Conference on Requirements and Technologies for the Detection, Removal and Neutralization of Landmines and UXO*, Brussels, Belgium, Sep 15-18, 2003. Available at `http://www.eudem.vub.ac.be/`.

[12] T. J. Bloodworth and A. J. Sieber. Standardized Testing of Metal Detectors. *Journal of Mine Action*, 7(3), Dec 2003. Available at `http://maic.jmu.edu/`.

[13] F. Borry, D. Guelle, and A. Lewis. Soil Characterization for Evaluation of Metal Detector Performance. In *EUDEM2-SCOT 2003, International Conference on Requirements and Technologies for the Detection, Removal and Neutralization of Landmines and UXO*, Brussels, Belgium, Sep 15-18, 2003. Available at `http://www.eudem.vub.ac.be/`.

[14] G. E. P. Box, W. G. Hunter, and J. S. Hunter. *Statistics for Experimenters*. John Wiley & Sons, 1978.

[15] C. Bruschini. Metal Detectors for Humanitarian Demining: from Basic Principles to Modern Tools and Advanced Developments. In *MINE'99, Mine Identification Novelties Euroconference*, Florence, Italy, Oct 1999. Available at `http://demining.jrc.it/aris/events/mine99/`.

[16] C. Bruschini. *A Multidisciplinary Analysis of Frequency Domain Metal Detectors for Humanitarian Demining*. PhD thesis, Vrije Universiteit Brussel, Sep 2002. Available at `http://www.eudem.vub.ac.be/publications/`.

[17] C. Bruschini. Metal Detectors for Humanitarian Demining: a Patent Search and Analysis (Reference Study), Nov 2003. Available at `http://www.eudem.vub.ac.be/`.

[18] C. Bruschini, K. De Bruyn, H. Sahli, and J. Cornelis. *EUDEM: The EU in humanitarian DEMining, Final Report*. Brussels, 1999. Available at `http://www.eudem.vub.ac.be/`.

[19] A. Carruthers. "Problem Soils" and Metal Detector Performance. *Mine Action Technology Newsletter*, (2), Apr 2005. Available at `http://www.gichd.ch/` and at `http://www.itep.ws/`.

[20] Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects. Available at `http://www.un.org/`.

[21] Croatian Mine Action Centre. Brochure "Mine Action in the Republic of Croatia", Mar 2004.

[22] Croatian Mine Action Centre. *`http://www.hcr.hr/`*, Aug 2006.

[23] R. H. Chesney, Y. Das, J. E. McFee, and M. R. Ito. Identification of Metallic Spheroids by Classification of Their Electromagnetic Induction Response. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):809–820, Nov 1984.

[24] C. J. Coffin. A Call for Standardized Data: the Demining 2010 Initiative Conference as an Opportunity for Consensus. *The Journal of Humanitarian Demining of the James Madison University*, 2(2), Jun 1998. The journal has been renamed to "Journal of Mine Action" starting from Vol. 3 No. 1. Available under the new name at `http://maic.jmu.edu/`.

[25] A. Craib. Standards and Measures of Success. In *World-Wide Humanitarian Demining Conference, Panel 3: Measures of Success in Mine Clearance*, Oct 1998. Available at `http://www.maic.jmu.edu/`.

[26] CWA 14747:2003, CEN Workshop Agreement, Humanitarian Mine Action — Test and Evaluation — Metal Detectors, Jun 2003. Available at `http://www.itep.ws/`.

[27] Y. Das. A Preliminary Investigation of the Effects of Soil Electromagnetic Properties on Metal Detectors. In *SPIE Defense and Security Symposium, Detection and Remediation Technologies for Mines and Minelike Targets IX*, Orlando, Florida, USA, Apr 12-16, 2004. Available at `http://spie.org/` and at `http://www.itep.ws/`.

[28] Y. Das, J.T. Dean, D. Lewis, J. H. J. Roosenboom, and G. Zahaczewsky. International Pilot Project for Technology Co-operation. Final report, 2001. Available at `http://serac.jrc.it/` and at `http://www.itep.ws/`.

[29] Department of Defense, United States of America. *Nondestructive Evaluation System, Reliability Assessment (Handbook MIL-HDBK-1823)*, Apr 1999.

[30] M. Fernandez, A. Lewis, and F. Littmann. PROM 1 Anti-personnel Landmines; Possibility of Activation by Physical Contact with a Metal Detector. Technical note S.P.I.01.29, Joint Research Centre of the European Commission, Humanitarian Security Unit, Ispra, Italy, Mar

2001. Available at http://demining.jrc.it/aris/ and at http://www.itep.ws/.

[31] Geneva International Centre for Humanitarian Demining. *Metal Detectors and PPE Catalogue 2005*. Available at http://www.gichd.ch/.

[32] Geneva International Centre for Humanitarian Demining. *Metal Detectors Catalogue 2003*. Available at http://www.gichd.ch/.

[33] Geneva International Centre for Humanitarian Demining. *Mine Detection Dogs: Training, Operations and Odour Detection*, 2003. Available at http://www.gichd.ch/.

[34] Geneva International Centre for Humanitarian Demining. *A Study of Mechanical Application in Demining*, 2004. Available at http://www.gichd.ch/.

[35] Geneva International Centre for Humanitarian Demining. *A Guide to Mine Action*, 2005. Available at http://www.gichd.ch/.

[36] Geneva International Centre for Humanitarian Demining. *A Study of Manual Mine Clearance*, 2005. Available at http://www.gichd.ch/.

[37] Geneva International Centre for Humanitarian Demining. *A Guide to the International Mine Action Standards*, Apr 2006. Available at http://www.gichd.ch/.

[38] Geneva International Centre for Humanitarian Demining. *Guidebook on Detection Technologies and Systems for Humanitarian Demining*, Mar 2006. Available at http://www.gichd.ch/.

[39] M. Gaal, S. Baer, T. J. Bloodworth, D. Guelle, A. M. Lewis, C. Mueller, and M. Scharmach. Optimising Detector Trials for Humanitarian Demining. In *SPIE Defense and Security Symposium, Detection and Remediation Technologies for Mines and Minelike Targets IX*, Orlando, Florida, USA, Apr 12-16 2004. Available at http://spie.org/ and at http://www.itep.ws/.

[40] M. Gaal and C. Mueller. Statistical Design of Experiments Applied to Tests of Metal Detectors Used for Mine Detection. In BAM, editor, *4th International Probabilistic Workshop*, Berlin, Oct 12-13, 2006.

[41] M. Gaal, C. Mueller, U. Ewert, P.-T. Wilrich, and W. Spyra. Trial Design for Testing and Evaluation of Metal Detectors Used in Humanitarian Landmine Clearance. In *ECNDT, 9th European Conference on Non-Destructive Testing*, Berlin, Sep 25-29, 2006. Available also at http://www.itep.ws/.

[42] M. Gaal, C. Mueller, K. Osterloh, M. Scharmach, S. Baer, U. Ewert, A. Lewis, and D. Guelle. Metal Detectors Test Trials in Germany and Croatia — ITEP Project 2.1.1.2. In *Conference MATEST 2003*, Brijuni, Croatia, Sep 28-30, 2003.

[43] Land Mine Detection. DOD's Research Program Needs a Comprehensive Evaluation Strategy. Report to the Chairman, Subcommittee on Military Research and Development, Committee on Armed Services, House of Representatives GAO-01.239, GAO — United States General Accounting Office, Apr 2001. Available at `http://www.gao.gov/`.

[44] R. Gasser. *Technology for Humanitarian Landmine Clearance*. PhD thesis, University of Warwick, Sep 2000. Available at `http://www.eudem.vub.ac.be/publications/`.

[45] Geneva International Centre for Humanitarian Demining. *http://www.gichd.ch/*, Aug 2006.

[46] D. Guelle. International Detector Test. Preliminary results, UNADP, Dec 1999. Available at `http://www.itep.ws/`.

[47] D. Guelle. International Detector Test. Final report, UNADP, Mozambique, Dec 2000. Available at `http://www.itep.ws/`.

[48] D. Guelle and A. Lewis. Systematic Test and Evaluation of Metal Detectors: The EC's STEMD Project. *Journal of Mine Action*, 9(1), Jul 2005. Available at `http://maic.jmu.edu/`.

[49] D. Guelle, A. Smith, A. Lewis, and T. Bloodworth. *Metal Detector Handbook for Humanitarian Demining*. European Communities, Italy, 2003. Available at `http://www.itep.ws/`.

[50] D. M. Guelle, A. M. Lewis, M. A. Pike, A. Carruthers, and S. M. Bowen. Systematic Test and Evaluation of Metal Detectors (STEMD). Interim Report Field Trial Lao, 27th September - 5th November 2004. Technical report, JRC, Joint Research Centre of the European Commission, Mar 2005. Available at `http://www.itep.ws/`.

[51] D. M. Guelle, A. M. Lewis, M. A. Pike, and C. Craill. Systematic Test and Evaluation of Metal Detectors (STEMD). Interim Report Field Trial Mozambique, 12th April - 5th May 2005. Technical report, JRC, Joint Research Centre of the European Commission, Nov 2005. Available at `http://www.itep.ws/`.

[52] D. M. Guelle, A. M. Lewis, and P. Ripka. Metal Detector Trials — Detector Test Results and Their Interpretation. Technical report, JRC, Joint Research Centre of the European Commission, 2006. Available at `http://www.itep.ws/`.

[53] IGEOD e-mail forum. Supported by IGEOD, Inter-Galactic EOD and Demining Foundation, http://www.igeod.org/. The old messages are available on request.

[54] *IMAS 03.40, Test and Evaluation of Mine Action Equipment*, second edition, Jan 2003. Available at http://www.mineactionstandards.org/imas.htm/.

[55] *IMAS 04.10, Glossary of Mine Action Terms, Definitions and Abbreviations*, second edition, Jan 2003. Available at http://www.mineactionstandards.org/imas.htm/.

[56] StatSoft Inc. *Electronic Statistics Textbook*. Tulsa, OK, USA, 2006. http://www.statsoft.com/textbook/stathome.html.

[57] International Campaign to Ban Landmines. *Landmine Monitor Report 2004*. Available at http://www.icbl.org/lm/.

[58] International Campaign to Ban Landmines. *Landmine Monitor Report 2005*. Available at http://www.icbl.org/lm/.

[59] International Campaign to Ban Landmines. *Landmine Monitor Report 2006*. Available at http://www.icbl.org/lm/.

[60] International Campaign to Ban Landmines. *http://www.icbl.org/*, Aug 2006.

[61] International Test and Evaluation Program for Humanitarian Demining. *http://www.itep.ws/*, Aug 2006.

[62] International Standard ISO 3534-3:1999(E/F), Statistics — Vocabulary and Symbols — Part 3: Design of Experiments, 1999. Available at http://www.iso.org/.

[63] J. D. Jackson. *Classical Electrodynamics*. Wiley, third edition, 1998.

[64] EC Joint Research Centre, editor. *Discussion Day on Soil Electromagnetic Characteristics and Metal Detector Performance*, Dec 12, 2002. Available at http://www.itep.ws/.

[65] C. King. *Jane's Mines and Mine Clearance 2003-2004*. Bath Press, eighth edition.

[66] C. King. Demining: Enhancing the Process. In *Second International Conference on the Detection of Abandoned Land Mines*, Edinburg, UK, Oct 12-14, 1998. Available at http://www.iee.org/.

[67] T. Küchenmeister. *Cluster Bombs and Cluster Munitions, a Danger to Life*. Aktionsbündnis Landmine.de, 2005. Available at http://www.landmine.de/.

[68] L. D. Landau and J. M. Lifschitz. *Lehrbuch der theoretischen Physik, Band 8, Elektrodynamik der Kontinua.* Akademie Verlag, Berlin, fifth edition, 1991. Written in German.

[69] MgM's Demining Network, e-mail forum. Supported by Menschen gegen Minen, `http://www.mgm.org/`. The old messages are available on request.

[70] Mine Action Information Center of the James Madison University. *`http://maic.jmu.edu/`*, Aug 2006.

[71] D. C. Montgomery. *Design and Analysis of Experiments.* John Wiley & Sons, third edition, 1991.

[72] C. Mueller et al. Reliability Model for Test and Evaluation of Metal Detectors — Part 2. Final report of the ITEP Project 2.1.1.8, Federal Institute for Materials Research and Testing (BAM), Berlin, Germany. To be published at `http://www.itep.ws/`.

[73] 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.

[74] C. Mueller, M. Gaal, M. Pavlovic, M. Scharmach, and P.-T. Wilrich. Reliability Tests for Demining. In *International Symposium Humanitarian Demining*, Sibenik, Croatia, Apr 25-28 2005.

[75] C. Mueller, M. Gaal, M. Scharmach, S. Baer, A. Lewis, T. Bloodworth, D. Guelle, and P.-T. Wilrich. ITEP Test Trials for Detection Reliability Assessment of Metal Detectors. *Journal of Mine Action*, 8(2), Nov 2004. Available at `http://maic.jmu.edu/`.

[76] 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, Sep 2004. Available at `http://www.itep.ws/`.

[77] 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.

[78] K. Osterloh, C. Mueller, and U. Ewert. Bedrohung durch Minen — können zerstörungsfreie Prüfmethoden zur Beseitigung beitragen? *ZfP-Zeitung*, (82), Dec 2002. Written in German.

[79] Convention On The Prohibition Of The Use, Stockpiling, Production And Transfer of Anti-Personnel Mines And On Their Destruction. Available at http://www.un.org/.

[80] F. B. Paca, C. D. Hoover, and R. M. Ess. Simulant Mines (SIMs). Scientific and technical report, Oct 1998. Available at http://www.uxocoe.brtrc.com/techlibrary/TechRpts/misc1.asp.

[81] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. ISBN 3-900051-07-0.

[82] D. M. Reidy. Handheld Metal Detectors: Nicaraguan Field Test Report. Technical report, Department of Defense Humanitarian Demining Research and Development Program, Oct 2001. Available at http://www.itep.ws/.

[83] A. J. Schoolderman and J. H. J. Roosenboom. Detection Performance Assessment of Hand-held Mine Detection Systems in a Procurement Process: Test Set-up for MDs and MD/GPRs. In *SPIE Defense and Security Symposium, Detection and Remediation Technologies for Mines and Minelike Targets X*, Apr 2005. Available at http://spie.org/ and at http://www.itep.ws/.

[84] K. M. Simonson. Statistical Considerations in Designing Tests of Mine Detection Systems: I — Measures Related to the Probability of Detection. Technical report, Sandia National Laboratories, Aug 1998. Available at http://www.itep.ws/.

[85] K. M. Simonson. Statistical Considerations in Designing Tests of Mine Detection Systems: II — Measures Related to the False Alarm Rate. Technical report, Sandia National Laboratories, Aug 1998. Available at http://www.itep.ws/.

[86] A. Smith. Database of Demining Incidents and Victims. Last updated: August 2006. Available online at http://ddasonline.com/.

[87] A. Smith. Database of Demining Accidents, May 2002. Maintained by Geneva International Centre for Humanitarian Demining. Available as a CD at http://www.gichd.ch/. Currently out of print.

[88] A. Smith. What Use is a Database of Demining Accidents? *Journal of Mine Action*, 6(2), Aug 2002. Available at http://maic.jmu.edu/.

[89] A. Smith. Myths, Mines and Ground Clearance. *Journal of Mine Action*, 7(2), Aug 2003. Available at http://maic.jmu.edu/.

[90] A. Smith. Comparative Trials of Manual Mine Clearance Techniques, Mozambique, 2004. Technical report, AVS Mine Action Consultants, Dec 2004. Available at `http://www.itep.ws/`.

[91] A. Smith. *`http://www.nolandmines.com/`*. AVS Mine Action Consultants, Aug 2006.

[92] I. Steker. Testing and Use of Demining Machines in the Republic of Croatia. *Journal of Mine Action*, 7(3), Dec 2003. Available at `http://maic.jmu.edu/`.

[93] J. A. Swets and R. M. Pickett. *Evaluation of Diagnostic Systems*. Academic Press, 1982.

[94] P. Szyngiera. A Method of Metal Object Identification by Electromagnetic Means. In *MINE'99, Mine Identification Novelties Euroconference*, Florence, Italy, Oct 1999. Available at `http://demining.jrc.it/aris/events/mine99/`.

[95] J. D. Toews and W. Sirovyak. Metal Detector Trial – Colombia, Results from 2002. Technical report, Defence R&D Canada – Suffield, Jul 2003. Available at `http://www.itep.ws/`.

[96] J. Trevelyan. Landmines: A Humanitarian Demining Approach. *The Asia Pacific Magazine*, (11), Jun 1998. Available at `http://coombs.anu.edu.au/asia-pacific-magazine`.

[97] Vlada Republike Hrvatske. *Plan humanitarnog razminiranja za 2006. godinu*, Apr 2006. Written in Croatian. Available (Nov 2006) at `http://www.hcr.hr`.

[98] P.-T. Wilrich and H.-J. Henning. *Formeln und Tabellen der angewandten mathematischen Statistik*. Springer-Verlag, Berlin, third edition, 1987. Written in German.

Index