THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Decision support for complex planning challenges

Combining expert systems, engineering-oriented modeling, machine learning, information theory, and optimization technology

LARRY M. DESCHAINE

CHALMERS

Physical Resource Theory
Department of Energy and Environment
Chalmers University of Technology
Göteborg, Sweden 2014

Decision support for complex planning challenges – combining expert systems, engineering-oriented modeling, machine learning, information theory, and optimization technology

LARRY M. DESCHAINE

To Blessed John Paul II

Doktorsavhandlingar vid Chalmers Tekniska Högskola
Ny serie nr: 3661
ISBN: 978-91-7385-980-6
ISSN: 0346-718X

Physical Resource Theory
Department of Energy and Environment
Chalmers University of Technology
SE-412 96 Göteborg
Sweden

Telephone: +46 (0) 31-772 10 00
Email: Larry.M.Deschaine@alum.mit.edu
URL: http://www.chalmers.se/ee/EN/

Printed by Arkitektkopia AB
Göteborg, Sweden 2014

Abstract

This thesis develops an approach for addressing complex industrial planning challenges. The approach provides advice to select and blend modeling techniques that produce implementable optimal solutions. Industrial applications demonstrate its effectiveness. Industries have a need for advanced analytic techniques that encompass and reconcile the full range of information available regarding a planning problem. The goal is to craft the best possible decision in the time allotted. The pertinent information can include subject matter expertise, physical processes simulated in models, and observational data. The approach described in this paper assesses the decision challenge in two ways: first according to the available knowledge profile which includes the type, amount, and quality of information available of the problem; and second, according to the analysis and decision-support techniques most appropriate to each profile. We use model-mixing techniques such as machine learning and Kalman Filtering to combine analysis methods from various disciplines that include expert systems, engineering-oriented numerical and symbolic modeling, and machine learning in a graded, principled manner. A suite of global and local optimization methods handle the range of optimization tasks arising in the demonstrated engineering projects. The methods used include the global and local nonlinear optimization algorithms. The thesis consists of four appended papers. *Paper I* uses subject matter expertise modelling to provide decision analysis regarding the environmental issue of mercury retirement. *Paper II* provides the framework for developing optimal remediation designs for subsurface groundwater monitoring and contamination mitigation using numerical models based on physical understanding. *Paper III* provides the results of a machine learning study using the Compiling Genetic Programming System (CGPS) on multiple industrial data sets. This study resulted in a breakthrough for identifying underground unexploded ordnance (UXO) and munitions and explosives of concern (MEC) from inert buried objects. *Paper IV* develops and uses the model mixing and optimization approach to expound on understanding the MEC identification technique. It uses the methods in the first three papers along with additional technology. Each thesis paper includes complimentary citations and web links to selected publications that further demonstrate the value of this approach; either via industrial application or inclusion in US government guidance documents.

Keywords: Decision analysis, model blending, model mixing, data modelling engineering-oriented modelling, energy, environmental, optimization, analytic hierarchy processes, machine learning, UXO, and MEC.

List of Papers

The following appended papers provide the thesis basis:

I. **Application of the Analytic Hierarchy Process to Compare Alternatives for the Long-Term Management of Surplus Mercury.** Randall PJ, Brown LA, Deschaine LM, DiMarzio J, Kaiser G, and Vierow J. *Journal of Environmental Management*, 71 (2004) 35-43, Elsevier Press.

II. **Simulation and Optimization of Large Scale Subsurface Environmental Impacts; Investigations, Remedial Design and Long Term Monitoring**. Deschaine LM. *Journal of Mathematical Machines and Systems*, Kiev. Number 3, 4. (2003) pages 201-218. *Thesis contains an updated and extended version of this paper.*[1]

III. **Extending the Boundaries of Design Optimization by Integrating Fast Optimization Techniques with Machine-Code-Based Linear Genetic Programming.** Francone FD and Deschaine LM. *Information Sciences Journal*, Elsevier Press, Vol. 161(3-4), pages 99–120, (2004), Amsterdam, the Netherlands.

IV. **A Computational Geometric / Information Theoretic Method to Invert Physics-Based MEC Models Attributes for MEC Discrimination.** Deschaine LM, Nordin JP, and Pinter JD. *Journal of Mathematical Machines and Systems*, Kiev. Number 2, (2011), (pages 50-61). *Thesis contains an extended version of this paper due to MMS page limits.*[2]

[1] Original abstract: http://www.immsp.kiev.ua/publications/eng/2003_3_4/index.html
[2] Original article: http://www.immsp.kiev.ua/publications/articles/2011/2011_2/02_2011_Deschaine.pdf

Related papers, author's contribution, and validity threat assessment

Each paper in the thesis is a culmination of several precedent works. Citations of works prepared after publication of these papers illuminate current advancements. The Ph.D. candidate has authored over 100 papers and conference presentations in this area since 1985.[3] The below list mitigates validity threat by demonstrating that the approaches presented in this thesis as adopted and published in US Government guidance documents and used by industry.

Paper I:

- United States Environmental Protection Agency (USEPA) (2002). Guidance Document: Preliminary Analysis of Alternatives for the Long Term Management of Excess Mercury. EPA/600/R-03/048, August, 65 pages. http://nepis.epa.gov/Adobe/PDF/P100E50I.pdf. *This report cites the candidate as key decision support analyst and cites his previous work which served as the basis of the analysis (Deschaine, et. al., 2001).*

- USEPA (2005). *Guidance Document: Economic and Environmental Analyses of Technologies to Treat Mercury and Dispose in a Waste Containment Facility*, EPA/600/R-05/157, April, 100 pages.http://nepis.epa.gov/Adobe/PDF/P1009CC3.pdf.

Paper II:

- Interstate Technology and Regulatory Council (ITRC) (2007). *In-Situ Bioremediation of Chlorinated Ethene DNAPL Source Zones: Case Studies.* Bioremediation of Dense Non-Aqueous Phase Liquids (Bio DNAPL) Team, Washington, D.C. pp. 128-147. http://www.itrcweb.org/Guidance/GetDocument?documentID=11.

- Deschaine, Larry M., Theodore P. Lillys, and János D. Pintér. "Groundwater remediation design using physics-based flow, transport, and optimization technologies." *Environmental Systems Research* 2, no. 1 (2013): 6. http://www.environmentalsystemsresearch.com/content/2/1/6/[4]

[3] Full list of candidate's publications is available upon request.

[4] Animations that illustrate the application and effectiveness of these environmental optimization on project examples are provided at: (http://www.hglsoftware.com/cleanup.cfm)

Paper III:

- Ott A (2010). "Increasing Market and Planning Efficiency through Improved Software and Hardware-Enhanced Optimal Power Flow Models (Washington, DC, USA; free webcast), FERC: Development of Enhanced Generation/Demand Response Control Algorithm, PJM Interconnection. June 23. 23 slides. Candidate developed the Adaptive Generator Model (AGM) discussed in this work which is available at: https://www.ferc.gov/EventCalendar/Files/20100623161840-Ott,%20PJM%206-23-10.pdf

- Deschaine LM, Hoover RA, Skibinski JN, Patel JJ, Francone FD, Nordin JP, and Ades MJ (2002). Using Machine Learning to Complement and Extend the Accuracy of UXO Discrimination Beyond the Best Reported Results of the Jefferson Proving Ground. Technology Demonstration, pages 46-52. Society for Modeling and Simulation International Advanced Technology Simulation Conference, San Diego, CA, USA, April.

  http://www0.cs.ucl.ac.uk/staff/ucacbbl/ftp/papers/deschaine/ASTC_2002_UXOFinder_Invention_Paper.pdf.

Paper IV:

- Deschaine LM and Patel J (2010). MEC-ID: The Evolution of International MEC Identification Research and Development Breaks through into an Efficient, Effective and Defensible Field-Ready Production Tool. NDIA: Environment, Energy, & Sustainability Symposium & Exhibition. Colorado Convention Center Denver, CO, USA, June 14 – 17.

  http://e2s2.ndia.org/pastmeetings/2010/tracks/Documents/9835.pdf.

- United States Department of Defense (DoD) Environmental Security Technology Certification Program (ESTCP) Certification (2009) of the UXO discrimination approach (email: from Dr. Herb Nelson, OSD, Munitions Management Program Manager, SERDP/ESTCP, 901 N. Stuart St., Suite 303, Arlington, VA 22203 – to Larry M. Deschaine, PE), dated May 21, 2009.

Summary of Author Contribution to the Papers

The ideas presented in the thesis papers comprise the candidate's original ideas and designs. Each paper represents a snapshot of development, which occurred over multiple decades. The papers properly cite and provide credits for algorithms developed by others and incorporated or extended. Table 1 summarizes the candidate's thesis paper contributions.

Table 1. Summary of candidate's contribution to the thesis papers

Paper	Idea	Design	Implementation	Analysis	Writing
I	**P**	**P**	**S**	**S**	**S**
II	**P**	**P**	**P**	**P**	**P**
III	**S**	**S**	**S**	**S**	**S**
IV	**P**	**P**	**S**	**P**	**P**

*Note: The symbol "***P***" indicates primary responsibility. The symbol "***S***" indicates shared responsibility.*

Paper I: The candidate developed the idea, approach and design and mentored the project team. He shared the implementation, analysis and writing. A project team collected the data used for the analysis. The candidate first developed computer based analysis for environmental decision in 1985 (Deschaine, *et. al*., 1985). The candidate previously developed and implemented the algorithms as demonstrated in precedent works on the AHP algorithm, and its extensions (Deschaine *et al*., 1997, 1998a and b, 2001). Regarding those works, he performed all functions except the data collection.

Paper II: The candidate developed the site-wide optimization idea and designed and coded the algorithmic extensions. He performed the implementation, analysis and wrote the paper. The candidate first developed computer based analysis for environmental decision optimization as part of his graduate thesis (Deschaine, 1992). The paper uses properly cited references and credits for algorithms developed by others and adapted them for use in this work. He performed all functions except the data collection used in the model calibration.

Paper III: Paper III was a requested article spawned from the candidate's previous published works in this area since 2000, with (Deschaine, *et. al*, 2002) providing the stimulus for the invited paper. The PhD candidate, his advisor (Peter Nordin) and the co-author shared developing the idea for the extended paper, including ideas for extending the CGPS algorithm, and shared the design, implementation, analysis, and article writing. The candidate and the co-author performed the model runs.

Paper IV: The idea, design, implementation, analysis, extensions and writing to document using the CGPS machine learning technology for high accuracy UXO discrimination using non-destructive field data, date back to the candidate's work and original concept (Deschaine *et al*., 2002). In Paper IV, the PhD candidate developed the idea, design, implementation, analysis and performed all the writing. Others collected the field data as part of the cited government study. The MEC features used for discrimination are a blend; developed by the candidate and several parties including participants of US Government sponsored research programs over a 10-year period. A project geophysicist extracted the features.

Validity Threats:

Techniques developed and used in this thesis is both stakeholder and peer reviewed. They have been tested and proven successful on industrial applications. The PhD candidate has published over 100 works regarding the development and application of these algorithms to industrial challenges since 1985. He has been incrementally improving, expanding and refining the mixed model approach as the challenges became increasingly complicated, applications expanded in breadth, and computer power became available to solve them. The thesis provides the overarching algorithm for the decision support approach that leverages the published examples. Two papers, Papers I, and II, are summaries of work now published in United States Agency documents released for unlimited distribution by USEPA and ITRC, respectively. The ITRC published the algorithm in Paper II as applicable for general purpose, mitigating that external validation threat of non-generality. In addition, industry has accepted and deployed the solutions developed by using this approach. The approaches and solutions have been subject to peer review by stakeholders, USEPA, United States Department of Energy (DOE), and DoD.

Table of Contents

Acronyms and Definitions

3-D	three-dimensional
AHP	analytic hierarchy process
AI	artificial intelligence
ANN	artificial neural network
CDF	cumulative distribution function
CGPS	compiling genetic programming system
COM	component object model
DLL	dynamic link library
DNAPL	dense non-aqueous phase liquid
DoD	United States Department of Defense
DOE	United States Department of Energy
DOE-NETL	U.S. Department of Energy National Energy Technology Laboratory
DSS	dynamic subset selection
EMI	electromagnetic instrument
EO	engineering-oriented
ES	Evolutionary Strategies
ESTCP	Environmental Security Technology Certification Program
FPU	floating-point unit
HE	high explosive
ITRC	Interstate Technology and Regulatory Council
LGO©	Lipschitz Global Optimization
LGP	linear genetic programming
LNAPL	light non aqueous phase liquid
LTM	long term monitoring
MAG	magnetometer
MEC	munitions and explosives of concern

MINLP	Mixed Integer Nonlinear Programming
MP	mathematical programming
MWR	Method of Weighted Residuals
NAPL	non aqueous phase liquid
NRC	National Research Council
OSD	U.S. Office of the Secretary of Defense
PBC	Pontifical Biblical Commission
R&D	research and development
RDX	Royal Demolition Explosive
ROD	Record of Decision
ROI	Return on Investment
SERDP	Strategic Environmental Research and Development Program
SME	subject matter expert
SRS	DOE Savannah River Site, Aiken, SC, USA
SWMU	solid waste management unit
TAG	Technical Advisory Group
TCE	trichloroethene
USEPA	United States Environmental Protection Agency
UXO	unexploded ordnance
VOC	volatile organic compound
VV&A	verification, validation, and accreditation
WSD	worth-of-sample-data

1. Introduction

Effective and efficient decision making within environmental systems is critical to promoting growth and supporting development of sustainable societies. The interrelationship between anthropogenic activity and environment quality is complex. Anthropogenic activity provides the infrastructure and systems that enable life quality whilst the environment sustains life. Tension can arise when anthropogenic activity intended to increase the quality of life impairs the environment, and thereby potentially compromises the planet's long-term sustainability potential. Examples of this include management and retirement of industrial chemicals, groundwater, and surface water contamination from chemical releases, inefficient or ineffective industrial processes, and degradation of useable lands due to military action. Simulation modeling and optimization are methods to deploy to support better use and preservation of the earth's natural resources. Modeling provides a framework for understanding the cause and effect of actions on the environmental and industrial systems. Optimization guides decision makers to the best decisions regarding use of limited resources. This thesis concerns developing decision model formulation, model blending using machine learning, implementation, and optimization to support solution development for complicated and natural resource intensive industrial problems.

1.1 Background

Modeling of industrial and environmental systems uses a variety of methods, and we explore the use and hybrid integration of three of the widely used methods: *a priori* human expert knowledge, in-depth physics models, and automatic modeling based on real world empirical data obtained from case studies. In this paper, we demonstrate that *blending* the information content from human expert knowledge (aka subject matter expert (SME) knowledge), physically based (i.e., using in-depth physically based knowledge) modeling approaches, and data-driven (aka machine learning), provides higher accuracy, improves stakeholder acceptance, and reduces computational requirements. This hybrid approach to the development of solution algorithms is applicable to a variety of problem areas. The availability of empirical data, engineering-oriented models, and SME information affects the solution methods applicable to a problem. This thesis describes algorithms developed by the candidate or, when pre-existing algorithms were available, extended by the candidate. It details testing the

approach on real-world industrial problems by applying the various algorithms to a variety of knowledge spaces and application domains. The need for optimal decision making by project teams, management, stakeholders, or regulators, to support the cost-effective use of limited resources on industrial problems motivated this work.

Reliable, optimal knowledge and models concerning an environmental or industrial system is paramount for efficient and effective decision-making. This includes understanding both the current and potential future states of the system under consideration. Specifically, estimation of industrial or environmental systems relies on a combination of *a priori* subject matter expertise, comprehensive 'in-depth' physics-based models, and data. Unfortunately, a complete, descriptive set or subset of the problem space and information necessary to produce a unique and complete optimal solution is often unavailable. This occurs for complicated processes, such as the movement of chemicals and fluids in the subsurface. It can also occur during experimental development such as occurs in industrial production process invention. A hierarchical approach for integrated decision model development, implementation, and optimization address the issue of the lack of deterministic closed form solutions. Each modeling approach, or combination of modeling approaches, has benefits and limitations. The field of optimal sequential decision-making (Hutter, 2005) and multi-sensor fusion (Luo and Kay, 1989) provide insight into the analysis of optimal use of scarce resources. For example, see the work of Hernández *et al.*, (2012), which includes approaches and demonstrations of blending model approaches and optimization technology to support decision-making.

1.1.1 Subject matter expert models

Subject Matter Experts (SMEs) develop models via formal knowledge capture and processing with expert system algorithms and techniques. The cognitive models developed by human experts provide valuable insight, but have a basic inherent constraint: humans often cannot easily process large volumes of data. These types of models are useful at an early stage for screening alternatives and guiding policy. The process of soliciting experience-based knowledge from SMEs, and then numerically coding the findings into programs to construct the model is time consuming. The experts needed for the model building are often in high demand and have competing priorities, which affects their availability for participation or constrains the schedule for model development. Further, analysis of expert opinions can reveal inherent uncertainties in interpretation and application of experiential knowledge. Human

expertise also contributes to both data-derived and engineering-oriented model development and algorithms designed to integrate knowledge from SMEs to capture both insight and experience. As such, SME models or expertise can be fused with, or incorporated directly into, other types of models. Section 4 summarizes SME models. Paper I and associated references discusses these models in detail.

1.1.2 Engineering-oriented models

Numerical or symbolic models based on the solution of the physically based governing equations (aka "engineering-oriented" in our work) models range from simplified heuristics to highly comprehensive system representations. [5] The benefit of deep physics-based models is they are fully transparent and understandable to the expert and, therefore, are reviewable by expert stakeholders. The benefit of a well-constructed physics-based model is the high confidence that the decision is based on solid foundation, including comprehensive information and principled analysis procedures and protocols.

The main limitation of these types of models is the often extensive time required to build, test, verify, validate and pass credibility and accreditation acceptance. While a simplified model can be developed in days or weeks, the development of comprehensive models can take decades. Engineering-oriented models regardless of comprehensiveness have limits regarding system representation fidelity and the ability to support the input data requirements. The increased fidelity and system resolution provided by some physics-based model implementations can require prolonged run times, often days or even weeks. The practical impact associated with prolonged run times, in terms of both equipment and personnel resources, complicates efforts during their use to optimize industrial and environmental systems.

The impacts to businesses and society resulting from extended development time of engineering-oriented models are quantified as lost opportunity—a postponed decision. Conversely, a less than optimal decision may be made based on reduced fidelity physics-based models. This poses a trade-off of model accuracy versus the time value of a decision. Engineering-oriented models are discussed in Section 5 and Papers II, IV and associated references.

[5] The engineering-oriented model is sometimes referred to as a "physics" or "physically based" model in this thesis. These models are characterized by symbolic or numerical mathematical formulations that capture the physics-based processes under investigation for the purposes of simulation and prediction.

1.1.3 Data-driven models

Data-driven modeling techniques are also known as “equation-writers”, and the resulting models may be fully reviewable and uncover the underlying physically based processes. Data-driven modeling algorithms (also referred to as data-mining) can process vast amounts of data to identify relationships, but relationships identified through algorithms do have limitations in that they do not guarantee causality determination. They produce functions that are statistically true to the information content in the data; therefore, it is imperative the data cover the range of information for which a model is sought. In data-driven models, fewer physical processes may be represented inside the model, relying more on the inputs and outputs to inform the solution or demonstrate insolvability. Conversely, the resulting model may not be fully reviewable (meaning it may not be represented in closed form) and the dimensions of the input and outputs may not be consistent. The data-driven models may appear more as “black-boxes”, meaning the construct inside the black box can be unknown to the modeler. This can make it difficult to understand the solution. Because of this limitation, data-driven models may not readily garner stakeholder acceptance. Since data-driven models are generated automatically through induction, they can have varying generalization capabilities on unseen data. Hence, blind testing of unseen data is used to ensure that useful models are generated. Machine learning discussion is provided in Section 6 and solutions to industrial problems provided in Papers III and IV.

1.1.4 Integrated models

In cases where the SME, engineering-oriented, or data-driven modeling approaches used individually do not produce results acceptable to decision makers, an integrated approach can have enormous utility. Integrated models leverage the information resources of two or more information dimensions. An overview of integrated models is provided in Section 7 via industrial examples. The use and value of this modeling approach on the industrial problem of developing a high accuracy non-destructive, non-invasive UXO discrimination technique is detailed in Paper IV.

1.1.5 Parsing by problem type

The question, therefore, is to determine to what extent are SME, engineering-oriented, or data-driven models appropriate for a problem, and when is integrated modeling the best course of action? The answer sought in the problem-specific tradeoff exists as a

trade-off. The availability of information and data; the engineering and physics; and the human, computer, and monetary resources available within the required time frame to solve the problem to the degree of accuracy and fidelity needed to meet the decision objective and manage uncertainty.

One view is to categorize the different industrial process problem challenges into quadrants. The view suggested in this work is to use four quadrants to partition problems based on the amount of data availability and process knowledge.

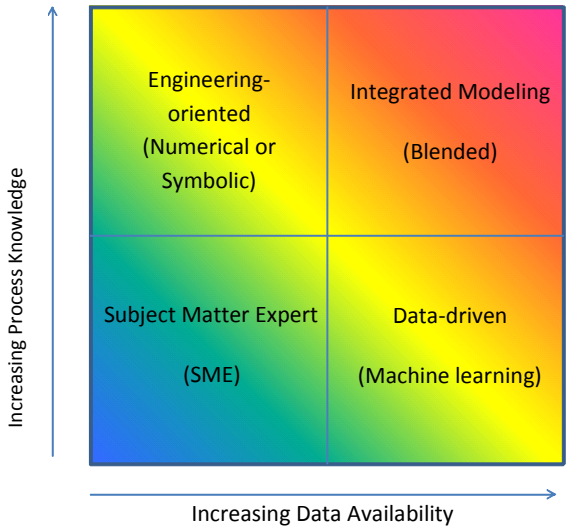

Figure 1. Process knowledge versus observed data availability. Industrial problems are categorized and applicable solution techniques applied by parsing the problems into these quadrants. Inspired by (Findler, 1991).

The categorization provides an approach to parse a problem by the amount of knowledge known and data available about the process under consideration. Once a problem is parsed, and the objective of the analysis quantified, various methods can be used to generate viable solutions. Analysis methods that solve problems in the four quadrants of this parsing structure differ along several dimensions. Development of a successful solution stems from knowing how to identify which category the problem resides in, and choosing the appropriate solution method, or combination of methods. This thesis is concerned with using engineering and physics, data and/or human expert insight synergistically to develop high fidelity models that have development periods and computational requirements that are practical for their intended use while providing the level of analysis warranted for solving the problem. Based on this outcome, decision makers or stakeholders can evaluate whether the proposed model meets the needs of the analysis, whether the system is amenable for optimization analysis, or whether

additional research is needed for the model or solution to be deemed “stakeholder acceptable”.

1.1.6 Conceptual solution approach

In developing a solution approach, the first step is to identify what constitutes a *solution to the problem*. To frame the answer to this basic question, a requirements document is developed that specifies the problem objective, identifies constraints, and provides an equation or model representing the information content and utility of the data derived from either experiments or simulation (or both). The second step is to specify how to produce the desirable solution. As discussed above, the method, or the combination of methods, will be different depending on the level of understanding of the physics of the processes in the problem, and the available data to describe it uniquely. The steps enumerated below, along with these new or extended algorithm developments, provide techniques to solve problems in the energy, environmental and industrial processing fields within the framework of the four knowledge quadrants.

To conceptualize this approach, one starts the solution process at the SME level (the lower left corner of Figure 1) and increases analysis complexity only where warranted. Figure 2 conceptualizes this overall generalized approach.

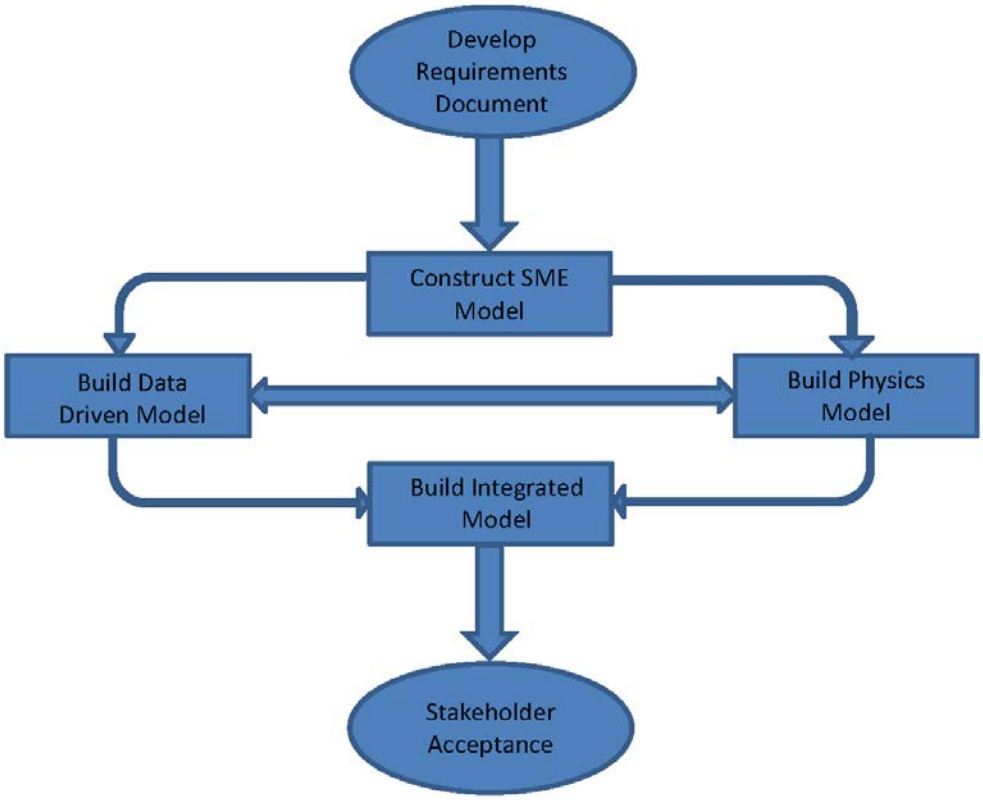

Figure 2. Conceptual overview of model building sequence. The model building process hierarchy is described below. The analysis starts at the SME stage and may have different paths and stopping points as the final analysis can stop at any stage when an acceptable solution is generated (See MIL-STD-882D for acceptance approaches). These include the value, risk, and consequences of decision-making. The modeling process can range from solely using an SME model to using various models that are fully integrated and optimized. After the models are built and acceptance gained, what-if scenarios are explored, uncertainty assessed through simulation. Optimal decisions are developed by linking these models with optimization as discussed in Paper II and IV and (Deschaine, et. al, 2013).

1. Initiate SME Decision Model: Initiate the problem solution using only information from the SME's insight and easily obtained data. If an acceptable model exists, and the solution is obvious, use it. If not, build a formal SME model. If the problem is solved at this level of analysis, stop here; if not, go to Step #2.
2. Decide where performance and information improvements are needed. Building the *a priori* SME models will uncover where the strengths and weaknesses of the analysis reside. It will also guide the decision-maker regarding uncertainty reduction. The next step is deciding if more empirical data and analysis are needed (i.e. data-driven model), or whether building an engineering orientated model based on the mathematics which describe the process physics would advance the decision making process. To formulate this decision to ascertain a feasible path forward, one must assess if a physically based modelling code for

the challenge is available or whether sufficient data exists to use the data mining technique to generate one. The time frame for problem solution is a key factor in this determination because developing a physically based modeling code can require many months or even years (BWXT and SAIC, 2002). For some cases, the decision path is clear as discussed in Paper II. For the environmental challenges evaluated in the case study for Paper II, engineering oriented models already existed, and existing data was too sparse to develop a model via machine learning. Therefore, the clear path forward was to use the existing physically based model code. The eight industrial examples provided in Paper III document the conditions when machine learning was the appropriate next evolution along this path. The general decision is made on a problem-specific basis, considering cost, time, accuracy and available personnel resources. Also, a problem-specific Analytic Hierarchy Process (AHP) model can support the framing the problem and assisting with this decision. Next step: Implement decision path and evaluate the results. If a final decision can be made, stop; otherwise, go to Step #3.

3. Decide if more improvement of available information is needed. This step addresses residual uncertainties by implementing the path not chosen in Step #2. If under Step #2, more data was collected and a data-driven model was constructed, then construct or update a physics-based model. If under Step #2 the numerical engineering-oriented model was selected, then data gaps could be filled. If machine learning first implemented, then physically based input parameters could be included. If an engineering-oriented model was first used then additional data can be obtained for model calibration and validation. Assess if a final decision can be made with the desired accuracy. If not, go to Step #4.
4. Construct Integrated Data/Engineering/SME Models. If none of the simpler precedent approaches led to stakeholder acceptance due to insufficient technical performance, then develop an integrated data/physics/SME model. The basic approach is to use a machine learning technique to fuse the information content and utility of raw data, processed data, physics model information (by components) and SME information into an optimal decision policy. This hybrid method can be quite complex to implement and apply; however, it can provide a near optimal solution in effort/utility space according to optimal sequential decision making theory (Hutter, 2005). Based on experience, models that can be

successfully integrated using machine learning are quantitative and have a definite structure and well defined scenarios with known outcomes (See Paper IV).

When implementing this approach, the decision maker or stakeholders can advise whether the proposed solution meets the needs (utility) of the analysis or whether additional research is needed for the solution to be deemed "stakeholder acceptable". This decision is often determined on a project-specific basis. This process is successfully implemented and achieved unanimous stakeholder acceptance (BWXT and SAIC, 2002). Upon model acceptance, optimal solution analysis commences (See references by Deschaine concerning optimization cited in Paper I and Paper II).

1.2 Objective and scope of this thesis

This thesis consists of four appended papers and these introductory remarks which explain the integration of the approach. The objective of the thesis is to provide a singular or hybrid approach when evaluating applicable problem domain-independent solution techniques using a combination of SME heuristics, in-depth physically based numerical models, and data-driven (machine constructed) models from empirical data. We demonstrate that blending the *non-redundant* information content from SME, data-driven, and physics-based modeling approaches using machine learning can provide higher accuracy models. As an industrial application PhD thesis, all methods are tested and demonstrated using real-world applications to promote acceptance by industry.

Developing new algorithms and the extending existing algorithms requires, at a minimum, proof of concept testing and, preferably, application to industrial data. Hence, the industrial case studies solved with each algorithm (new, extended, existing or combined) as evidenced by references to published works. This provides the confidence that the individual methods, and the overall approach, meet the objectives of the thesis. In each problem determination of the goals, constraints and model quality objectives are highly problem specific. The stakeholder group often decides how to assess the tradeoffs between the reliability, cost, and uncertainty of the solution.

During the research and development of this thesis, algorithms where developed and others extended by the candidate en route to achieving this objective. The cohesive structure of the overall solution approach is provided in these papers starting with: SME

models (Paper I); physically based numerical models (Paper II); machine learning (Paper III), and; the integrated (blended) model approach (Paper IV).

Paper I: Subject Matter Expert Models. This paper discusses the SME models used to solve problems characterized by the lower left corner of Figure 1. These problems are characterized by having both a low amount of available data and physical knowledge about the process under consideration, but the process is understood well enough that SME's can provide knowledge directed opinions. Developing an SME model is the first step in the conceptualized approach illustrated in Figure 2. Paper I documents an application is to provide decision support for the environmental management of global stores of mercury, an industrially useful but highly toxic material. The analysis described in Paper I provides the decision foundation for the United States Environmental Protection Agency (USEPA) guidance documents released for unlimited public distribution (USEPA, 2002 and 2005).

Paper II: Engineering-oriented Modeling. This paper discusses numerical models of subsurface physically based systems, those characterized by the upper left corner of Figure 1 where the physical aspects of the process are known well enough to construct either representative symbolic or numerical models. As illustrated on Figure 2, this is one of the two optional paths that may be chosen based on the success of the initial SME model. It represents a common problem in earth sciences, where physics-based understanding is documented through a conceptual site model, but the available data is sparse due to the prohibitive cost of collecting the amount of subsurface data needed to support the development of a data-driven model. The approach described in Paper II has been applied by the candidate to one of the most complex subsurface pollution modeling projects ever successfully completed in the United States (BWXT and SAIC, 2002 and USEPA, 2010), the DOE Pantex Plant environmental evaluation in Amarillo Texas. This application of the candidate's approach received written unanimous stakeholder acceptance and approval from all stakeholders. An extended version of Paper II authored by the candidate is published in the Interstate Technology and Regulatory Council (ITRC) BioNAPL guidance document released for unlimited public distribution (ITRC, 2007). This thesis contains an extended version of the original journal article and the ITRC document that includes additional information (the journal has a page length limit) as well as to include new results and findings since the initial

publications in 2003 and 2007. The groundwater remediation optimal design algorithm approach is made industrial strength and serves to substitute for the optimization methods cited in Paper II (Deschaine, *et. al.,* 2013).

Paper III: Data-driven (aka Machine Learning) Modeling. This paper describes solving problems characterized by the conditions of the lower right corner of Figure 1. It uses machine learning to accomplish that goal. As illustrated on Figure 2, this is one of the two optional paths that may be chosen based on the success of the initial SME model and in cases where physics-based understanding is not well known. For this reason, when considering problems within this category, it is imperative to ensure that solutions are representative on unseen data, with a minimum of over-fitting (aka memorization). Paper III discusses an eight-problem study using the machine learning technique, compiling genetic programming system (CGPS), as part of developing and enhancing the CGPS technology and tools.[6] This paper documents issues encountered when developing the machine learning algorithm, and the solution to increase its robustness. It provides discussion on the automated *discovery* of Darcy's Law (a physically based laws for groundwater flow) from data and also shows the solution to UXO discrimination first successfully solved in 2001 by (Deschaine *et al*., 2002) via CGPS data-mining. These two findings propelled the blended SME-physics-data modeling approach research and development initiative for high accuracy MEC items discrimination that culminated with Paper IV.

Paper IV: Blended Modeling and Optimization. Paper IV discusses integrating SME, physically based and data-driven models, problems characterized by the upper right quadrant of Figure 1. This technique incorporates information theory, inverted engineering-oriented models, computational geometry with optimization and machine learning on both data and physically based parameters developed by the SME approach. The result is a blended model of the MEC discrimination system that leverages the best of both approaches, empirical data and in-depth physics, to provide a comprehensive problem solution. This approach was accepted by the United State government (Environmental Security Technology Certification Program [ESTCP], 2009). Using this approach, the UXO discrimination challenge which was first solved by the candidate using data mining in 2001, is now solved using an understandable, principled

[6] Linear Genetic Programming (LGP), a term used in some of the literature, uses CGPS as its foundation.

approach. Paper IV extends the version of the journal published paper. It includes additional information (the journal has a page length limit) as well as to include new results and findings.

2. Method

All research and development described in thesis includes testing and demonstrating the applications for industrial problem solving.

2.1 Decision analysis and solution techniques

This section presents a brief overview of decision analysis and solution techniques useful in the challenge of developing an optimal decision support approach. The goal of optimal decision support is to develop a knowledge acquisition approach, develop a model that describes the process, and then optimize the decision. Optimization provides the best acceptable solution. Both heuristic and formal optimization approaches are used in this thesis. We proceed along the lines of (Buchanan, 1985) for segregating the problem into extracting information and knowledge content from SMEs, physics-based representations, or large datasets. This extracted information is represented in the form of an equation (aka function or model) which can be represented by an expert model, a numerical or symbolic model, a data-driven model, or an integrated (or hybrid) model. Given the general nature of the breadth of potential applications and various solution options discussed below, one challenge is to determine a practical and suitable set of applicable techniques for developing an acceptable solution. Acceptability to stakeholders defines a successful solution.

The algorithms researched and developed for this decision support approach are demonstrated on industrial problems. Approach generality is promoted by allowing other algorithms for the SME, engineering-oriented and data driven class of approaches be substituted for the ones used in this thesis. This is both in recognition that each decision area is a field unto itself, and the use of one technique over another for demonstration purposes does not constitute an endorsement. The general approach is shown in Figure 1 which provides a classification framework for problem types and generalized solution methods. Figure 2 provides the decision logic for the approach application. The various SME, physically based, and data-driven approaches are presented in Sections 3, 4, and 5 in more detail, respectively. Section 6 provides the integrated approach.

2.1.1 Computational Intelligence

Decision modeling is the assembly of encoding processes (or "intelligence") into an algorithm. Knowledge capture used in this thesis is in the form of computer programs. There is much debate on what constitutes "intelligence", the subject of artificial intelligence (AI), and its subset machine learning. This thesis reviews those threads but does not attempt to solve the definition problem. Rather, it takes the view that an important purpose of a computer system is to capture or emulate the information from an SME or in the data or physics basis of the process under investigation for use in optimal industrial decision making support.

Early approaches towards understanding AI are documented in "Steps Towards Artificial Intelligence" (Minsky, 1960) and the "Society of Mind" (Minsky, 1986). In the first reference, Minsky divides AI into optimization, pattern recognition, learning systems, problem solving, and planning. The latter reference attempts to explain how minds may work. The premise is that each mind is comprised of many smaller processes called "agents". Each agent can only do small, simple tasks, but the assembly of these agents in a society results in true intelligence forming. The application of agent systems is discussed in (Deschaine *et al*., 2000; Weiss, 1999; and Caglayan and Harrison, 1997). These concepts are brought to the group level in (Dawkins, 1989; Brodie, 1996; and Blackmore, 1999) through the concept of a "meme," (Dawkins, 1989) which is essentially a thought "gene" that is "interested" in survival and replication as would be an evolutionary biological entity's genetics. For optimal decision support and control, emergent intelligence on the distributed processor level is feasible and possibly necessary (Sapaty, 1999).

One hypothesis regarding the universe is that it is inherently mathematical (Schmidhuber, 1997 and Zuse, 1967). To expand Edward Deming's, "*In God we Trust, all others bring data*" (Deming, 1943), the approach presented in this work is conveyed by: "*In God we trust, all others bring the mathematics, physics, data and methods to support optimal decisions useful in industry*." This philosophy does present a challenge; namely, what to do when a in-depth mathematical theory is not available for understanding the process being investigated. This is quite often the case when working with complex systems (Miller and Page, 2007) or with new, experimentally developed

discoveries, where the physics may not yet be well understood and that some processes may not be solvable by direct computation of closed form formal mathematics.

AI techniques are surveyed in (Bundy, 1996) and (Luger, 2002). These references provide a broad introduction to AI, including the formal problem definition, state-space search, knowledge representation, reasoning in uncertain situations, machine learning, automated reasoning, and understanding of natural language. These references also provide a summary of machine learning, including what a definition of a well-posed learning problem. They also discuss the design of learning systems, concept learning, decision-tree learning, artificial neural networks, Bayesian learning, computational learning theory, instance-based learning, learning sets of rules, genetic algorithms, analytical learning (and combined with inductive learning), and reinforcement learning.

The mind can be thought of as continuous rather than Boolean discrete concepts; an aggregate rather than monolithic system, enabled by a multitude of disparate mechanisms (Franklin, 1999). Essentially, the major goal of the mind is to produce the next action by a reconstructive process of past experiences rather than through retrieval, and it can create information (simulate) for its own uses. Reinforcement learning is a method that lets an agent maximize its return based on actions and reinforcement feedback while in an uncertain, complex environment (Sutton and Barto, 2000). A perspective on reinforcement learning is given in (Mainzer, 1997), which discusses complex thinking paradigms and the path from Turing machines to knowledge-based systems and from artificial intelligence to artificial life.

A set of computational intelligence tools for the PC (Eberhart *et al*., 1996) explains in detail the use, application, and implementation of techniques for neural networks; evolutionary computation including genetic algorithms, genetic programs, evolution strategies, and particle swarms; fuzzy systems, expert systems, hybrid systems and the like. Additional detail regarding candidate decision approach characterization, quantification and applicable solution techniques that were reviewed when researching and developing this approach are found in (Beyer, 1998; Carrol, 2001; Chambers, 1999; Chambers, 2001; Coley, 1999; DeJong, 1992; Fogel, 2000; Goldberg, 1989; Jacob, 2001; Konar, 2000; Koza, 1992 and 2009; Man *et al*., 1999; Michalewicz and Fogel, 2004; Mitchell, 1997; Nordin, 1997; Schwefel, 1995; Siha and Gupta, 2000).

2.1.2 Solvability, timeliness and usefulness

In an industrial setting, the meaning of solvability is expressed as two-fold questions: 1) is the problem solvable? and; 2) is a dependable answer available when it is needed for optimal business use?

2.1.2.1 Solvability: Solvability in this context concerns the theory of computability and non-computability (aka the theory of recursive functions). It is focused on the existence of purely algorithmic procedures for solving various problems in certain complexity classes. This branch of mathematics has certain philosophical significance (Davis, 1982), such as the existence of unsolvable problems and Gödel's theorem (Nagel and Newman, 1986). An essential tool for the analysis of complexity is the universal Turing Machine. A negative proof of the existence of non-recursive, recursively enumerable sets and the impossibility of constructing a program to test whether an arbitrary program is free from "loops," is known as *the halting problem.* Expositions on thought and scientific revolution deal with scientific tradition and change (Kuhn, 1977) and the structure of scientific revolutions (Kuhn, 1996), these provides insight into this question. Hofstadter (Hofstadter, 1995) discusses creativity and computer models of thought mechanisms.

2.1.2.2 Timeliness and usefulness: As important as the theoretical ability to solve an industrial problem is the timeliness and usefulness of the solution generated. Important classes of formal problems while theoretically solvable are intractable (Garey and Johnson, 1979). In some cases, a model run on a complex problem can take hours, days, and even weeks. The power grid example (Ott, 2010) is solved by the candidate using the techniques described in this thesis. A reliable answer of future electrical power (MW 20-minutes out) production which includes both the estimate and uncertainty are needed for real-time one minute production cycles. Further complicating developing a usable analysis approach is when optimization is desired. The optimal solvability is correlated with the central processing unit (CPU) time needed to complete one function evaluation *and* the number of function calls needed by the optimization algorithm. Hence, fast running models (or approximations of models) and efficient optimization algorithms are required for timeliness and usefulness. Also, there may be a fundamental difference between mathematics and digital computation capabilities. Feynman presents in detail the aspects and limits of computation (Feynman, 1996). The collection of

papers in Laplante, 1996 reviews much of the chronology of foundational complexity analysis in computer science.

2.2 Modeling approaches

There are three principal divisions of AI activity (Findler, 1991). At one end of the spectrum is the engineering-oriented approach using numerical or symbolic models, where solutions are desired for repetitive tasks (such as handwriting and speech recognition) or even somewhat creative tasks such as composing music and computer programs. At the other end of the spectrum are approaches involving the simulation of cognitive behavior where the desire is to explain, understand, and simulate these processes. This thesis uses this framework to represent the spectrum of modeling approaches techniques as SME, engineering-oriented modeling, and data-mining. By combining this with the thoughts of the decision process as discrete then transitioning continuous (Franklin, 1999) processes, the integrated modeling approach is developed.

2.2.1 Heuristic / Subject Matter Expert (SME)

These represent approaches whereby the decision support algorithm developer attempts to determine how humans solve the problem, and then develops a model in code (manually or via automated induction) that approximates this behavior. There are multiple approaches for capturing human expert knowledge. Boole (Boole, 1854) provides a seminal treatise on the laws of thought that link mathematical theories and probabilities. When an expert has quantitative understanding of the process, an industrial implementation of logical expert systems can be developed (Giarratano and Riley, 1998). An overview of non-monotonic aspects inference can be found in the following references: (Minsker, 1993) or fuzzy (Klir and Yuan, 1995; Altrock, 1995; Altrock, 1997 and Cox, 1995). The CLIPS tool (Giarratano and Riley, 1998) has been extended to include fuzzy logic (Orchard, 1998) to accommodate this paradigm. Other prioritization methods include the Delphi method, developed at RAND Corporation in the 1960s, utility theory, the 100-dollar method, and the like. In the 100-dollar method, one is given 100 dollars and asked to prioritize allocation of resources. This technique is effective for developing heuristic solutions to a wide range of management challenges when the problem is tractable in one's mind. For complicated decisions, or ones that require decision documentation for stakeholder review and acceptance, formal decision supports tools are employed (Hernández, *et. al*., 2012).

The AHP, a technique for using expert opinion to rank options and resource allocation, is discussed in Saaty, 1980. This AHP algorithm was extended by the candidate. It now includes probabilistic analysis and stochastic optimization (Deschaine, et. al, 2001). This algorithm extension builds on the previous works cited whilst providing for a higher degree of quantification and justification, paramount when justifying budget expenditures in governmental or industrial settings.

2.2.2 Engineering-oriented models

Engineering-oriented (EO) models solve or simulate a set of governing equations based on the describing process knowledge using numerical approximation or symbolic representations of systems which are natural, physically based. This thesis is concerned primarily with the flow and quality of water both on the surface and below the surface in groundwater aquifers. Because these models are highly specialized for applications, relevant examples of these models are discussed in Section 4 and Paper II. Developing practical optimal solutions using computationally expensive simulation models of physical systems can be a challenge to code. Further complications in obtaining optimal answers are due to the often extended simulation times, which can range from hours to weeks. Computationally expensive models provide the motivation or developing high fidelity approximate models (see Paper III) and optimization algorithms capable of leveraging parallel simulations (Deschaine, *et. al.,* 2013).

2.2.3 Data-driven models

Data-driven models are developed using automated induction; they replace the information content contained in large datasets with equations and hence the techniques are sometimes referred to as “equation-writers”. This technique is used as a stand-alone analysis technique on data sets, or as the glue that combines the information content from several modeling techniques into an overall integrated model. Data mining is closely related to the machine learning field and uses these techniques as well as statistical tools. A practical guide to the assembling data for use in machine learning or data mining applications is covered in detail in Pyle (1999), and the process of building models is covered in Pyle (2003). Practical aspects of data mining from the business aspect are covered extensively. Other useful books on practical industrial data mining are the works of (Weiss and Indurkhya, 1998) and (Witten and Frank, 2000). When data is limited, bootstrapping methods can be used (Davison and Hinkley, 1997 and

Chernick, 1999). Redundant information in datasets is addressed using the independent or principal component analysis (Hyvarinen *et al*., 2001; and Peng *et al*., 2005) to reduce the problem's dimensions. The artificial neural network approach (ANN) is described in Fausett, 1994; Masters, 1995; Rao and Rao, 1995; Bigus, 1996; Swingler, 1996 and Skapura, 1996. Various algorithm repositories are available on the internet; Library of Efficient Data Types and Algorithms (LEDA, 2009), The Net Library (Netlib, 2009), the Guide to Available Mathematical Software (GAMS, 2009), the Collected Algorithms of the ACM (CALGO, 2009), The Stanford Graphbase (Stanford, 2009), the Mathematica library (Combinatorica, 2009), and the like.

2.3 Model integration techniques

This section discusses using mutual information theory as it relates to computation of model input relevancy and redundancy. This supports the assessment of the information value of modeling inputs from various sources.

2.3.1 Model integration input value assessment

In many cases, the information derived from the various modeling techniques can be thought of an information dimension. Maximizing the value of the model information dimensions is one approach to ensure minimum redundancy of information input. This concept can be used to design successful solution approaches and hence obtain acceptable solutions. It also helps avoid analysis that simply replicates information (Paper IV). Various techniques can be used to assess information value. The seminal work of Eckschlager and Stepanek (1979) provides methods for optimizing decision making in the context of analytical chemistry analysis and production planning, including optimizing the number and accuracy of analytical equipment types, optimal number of complementary methods of analysis for determining concentrations and optimal production planning from an information theory perspective. Eckschlager's concepts provided the inspiration for this hierarchical modeling approach – the challenges of when to add analyses and which analysis to select next – are similar in problem structure.

For example, the AHP (Saaty, 1996) approach can be used with SME knowledge in a comparative sense, but at some point, additional information from the expert or team of experts will not improve model quality (Paper I) or the ultimate decision. Similarly,

with engineering-orientated models, Kalman filtering can be used to determine the optimal system estimate and computes the worth of collecting sample data (Paper II). At some level of analysis either the engineering model will be accurate enough, or the data will be sufficient enough. Collecting additional data or further calibrating the engineering model will no longer affect the optimal decision. When conducting data mining analysis, at some point there will be enough columns and rows of data to develop accurate predictions, and collecting additional data or different types of data will not affect the decision. In the case of the integrated modeling approach, where machine learning is the integrating technique, the information between two or more data sources can be quantified by assessing their mutual information and can advise when the information derived from different modeling approaches is complementary and valuable, or redundant and unnecessary.

2.3.2 Mutual information analysis

The mutual information technique is regularly used in signal processing for assessing the information transmitted between a transmitter and a receiver. The method described here is applicable when quantifying the mutual information between multiple inputs in a single dimension, or when multiple information dimensions are present. In this application, the "transmitter" is the information contained in one or more of the SME, data-driven or physically based approaches; the "receiver" is the answer or the solution to the problem. In the Venn diagram on Figure 3, each circle represents the total amount of information in each approach, and the overlap between the approaches is the mutual information, information that is in common (that is, redundant).

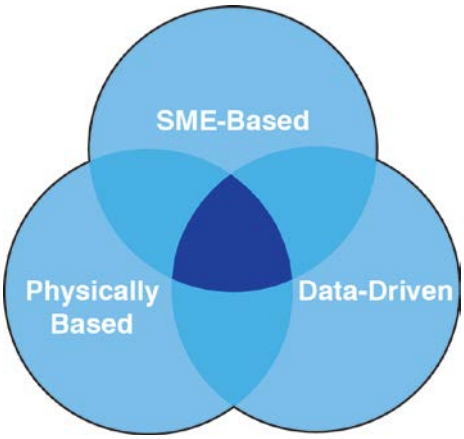

Figure 3. Mutual information Venn diagram. This figure conceptually shows the redundant information as the overlap in the information quantification calculations.

Problem characterization captures the breadth of available information. Broad knowledge capture encompasses all three mutually exclusive areas of information space such that the three circles would have a large area with minimal overlap. This would be a situation of low redundancy whereby each approach contributes information that the others do not. However, totally mutually exclusive information spaces are rare. For example, SMEs contribute to building engineering-oriented and data-driven approaches and data observations are used in developing engineering-oriented approaches. Maximizing the non-redundant information shows certain efficiencies in data collection, storage and processing. However, low redundancy *in a set* of information sources might be an indicator of fragile system design if the non-redundant information is critical for decision making and is not reliably available. This has more importance when designing real-time systems such as electrical power grid operations. These systems must regularly and automatically provide solutions in a real-time production environment. Other analysis may be more static, performed on the desktop and presented in a regularly scheduled planning meeting.

When building a single or integrated solution approach using machine learning, each of the information dimensions of SME, data-driven and engineering-oriented models often contain one or more vectors. These vectors are illustrated conceptually as input columns in Table 2.

Table 2. Information dimension assembly for information computations and machine learning dataset design

Case	SME	Data Driven	Physically Based	Solution
e_1	$SME_1(i=1…a)$	$DD_1(i=1…b)$	$EO_1(i=1…c)$	L_1
e_2	$SME_2(i=1…a)$	$DD_2(i=1…b)$	$EO_2(i=1…c)$	L_2
e_3	$SME_3(i=1…a)$	$DD_3(i=1…b)$	$EO_3(i=1…c)$	L_3
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
e_n	$SME_n(i=1…a)$	$DD_n(i=1…b)$	$EO_n(i=1…c)$	L_n

Each information dimension is separated into its components and placed in a flat file. Information theory and machine learning help understand the value and relative contribution of the information and develop functional relationships. See Paper IV for a worked example.

The input vectors consist of the problem information set and include the information contained in SME, data-driven or physically based models. These inputs include items necessary to making the corresponding decision (L for “label”). SME_n refers to the set of (integer features) *inputs* that contain the information obtainable from the SME dimension. DD_n and EO_n are defined similarly. The case example (e_n) is a complete

problem definition with known inputs and a known solution. It is developed either via algorithmic computation (in which the mapping function is known) or from observations but without the mapping function. In other words, the mapping function *f(x) = {SME, DD, EO}* that maps the inputs to the solution label [*L*] may be unknown. In this scenario, machine learning can be used to construct the mapping functions (the mapping function is sometimes referred to as a "model" in the machine learning field).

The amount of mutual information between the information sources, which are sub-systems, can be conceptually assessed or quantifiably computed. The mutual information is defined as follows (MacKay 2003): Given two random variables (*X, Y*), the mutual information *I(X;Y)* is defined according to their marginal and joint probability density functions *p(x), p(y),* and *p(x, y)*:

$$I(X;Y) = \iint p(x,y) log \frac{p(x,y)}{p(x)p(y)} dxdy \qquad (2.1)$$

Visually, *I(X;Y)* is represented by the areas of overlap in Figure 3. The variables *(X, Y)* represent any two of the three dimensions. Note that if mutual information *I(X; Y)* is high (represented by a large overlap) then most of the information in "*Y*" is already contained in "*X*". This indicates "*Y*" is redundant with respect to "*X*": therefore, it is likely that conducting an analysis with the information set $X \cup Y$ will not provide a much better answer than using *X* or *Y* individually (cf. Deschaine *et al*., 2011; noting that equation 2.1 is implemented as a discrete summation). Hence, one can use this knowledge to spend fiscal resources most effectively by generating the most informative information at the best cost. The analysis also holds when extending it to all three information dimensions *(X, Y, Z).*

The amount of information per added input is computed using the discrete form of the mutual information equation and the units are "bits." By combining the number of bits with the cost to collect the information, the unit cost per information bit is obtained. By including the reliability of the information sources, a risk adjusted information value can be calculated.

There are trade-offs between risk and reliability. Conceptually, for example, imagine two SMEs exist in an organization that provides exactly the same answer for the same reasons. One is redundant and may be deemed unnecessary, but if the input is critical, a

decision maker may choose to keep both SMEs informed to ensure reliability (the certainty) of being able to obtain the critical input when needed. This assessment of information value and reliability is vital to understand when designing a solution to a problem.

The area of the Venn diagram (the area outlined in bold) is given by the joint entropy (*H*) minus the mutual information (*I*), which represents the overlapping areas. To maximize this area:

$$\max H(X, Y, Z) = \max [H(X) + H(Y) + H(Z) - I(X;Y;Z)] \quad (2.2)$$

When a large number of inputs are available, one may desire to "downsize" the number of the inputs (Deschaine *et al*., 2011) to focus on the most salient ones. This focused input selection technique helps to understand the value of the input, and to promote more accurate machine learning results (Peng *et al*., 2005). Minimization of the information overlaps (the mutual information) between two of the dimensions is expressed in Peng *et al*., (2005) as

Minimize the internal redundancy (*R*) of the input vectors set *S*:

$$\min(R) \quad (2.3)$$

where

$$R = \frac{1}{|S|^2} \sum_{x_i x_j \in S} I(x_i; x_j) \quad (2.4)$$

R(S) is the redundancy computation of a pair of inputs (x_i; x_j) in information theoretic units of "bits," and S is the set of inputs from the information dimension(s). The goal is to select from among various problem solution approaches (more specifically, their respective input vectors). The selection of the set S of m features $\{x_i, i=1\ldots m\}$, where *m* is any of the input vectors, which jointly have the minimum redundancy with respect to each other. Minimization is used to force the information overlap to be small. The technique provides an especially useful way to quantify the value of the inputs as it allows for the optimal solution design and either reduces unnecessary inputs, and when combined with cost and reliability of the inputs can design a system that uses less expensive inputs, inputs of equal informational value or greater reliability (Paper IV). These types of designs can save both time and fiscal resources.

At the same time, one wants to have a high degree of mutual information between the selected *set* of inputs and the solution ("label") when using supervised machine learning techniques. Here, the goal is selecting from among various problem solution approaches the selection of the set *S* of m features $\{x_i, i=1\ldots m\}$ which jointly have the largest relevance with respect to the solution's answer while having minimal redundancy. The The label "*L*" represents the answer to the solution known. This is accomplished using the formulation:

$$\max D(S,L) \tag{2.5}$$

here

$$D = \frac{1}{|S|}\sum_{x_i \in S} I(x_i; L) \tag{2.6}$$

Conceptually, the mutual information shared internally within the *set* of informational dimension inputs (which consist of m total features) should overlap with the answer *L* to a large extent. The label "*L"* is a known answer to a given problem. It is essential as it provides the means to validate the solution process. For example, a uniform input, or a randomly distributed one, would have low mutual information with the solution label, and hence is not valuable for contributing to the problem solution.

To optimize the solution approach design, the *D* (dependency) and *R* (redundancy) is determined simultaneously using an objective function formulation of either their difference [(maxΦ(*D,R*), Φ=*D-R*] or alternatively their quotient [maxΦ(*D,R*), Φ=*D/R*]. In some real-time applications, a specified level of redundancy may be desired. The optimization is currently conducted using a (heuristic) greedy algorithm: the dataset constructed by adding one input at a time. For example, when the algorithm is scanning through the inputs, the vector with the largest [*D-R* for difference] or [*D/R* for quotient] value would be added to the set *S* of m inputs. The current mRMR algorithm uses a greedy optimization method. This approach can be extended to include the cost and reliability of the inputs in the formulation. A non-greedy optimization algorithm can be used to ensure the optimal solution is determined. This extension was not necessary for research in this thesis (Paper IV). A non-greedy algorithm was needed for the design optimization task (Deschaine, *et. al*, 2013). The mRMR approach focuses on determining the set of relevant features whose aggregated information content is

designed for improving the accuracy and performance of a machine learning algorithm. The economy of a set of information sources as well as their reliability is a vital consideration in system design. The *specific information price* (Eckschlager and Stepanek, 1979) is defined by:

$$C(p, p_o) = \frac{\tau}{M(p, p_o)} \tag{2.7}$$

Where τ is the cost of carrying out the analysis and M the real amount of information obtained from analyses. The information price c is enumerated in ratios of currency units per information units (e.g., $/bit). When analyzing problems of unknown complexity, characterization and solution, we proceed along the hierarchal approach shown on Figure 2. The precision of the analysis is often correlated with cost, and at some point, additional analysis will not provide meaningful differences in the decision. This trade-off is addressed in (Eckschlager and Stepanek, 1979), including a discussion of quantifying optimal levels of information content in concert with analytical chemistry production planning including staff labor and equipment cost considerations. By adjusting the measures and confidence of signal accuracy, the equations for conducting an assessment of an analysis including risk-based considerations are obtainable.

2.4 Optimization

Once a team-acceptable model is developed, the objective and constraints are defined. Then some level of the feasibility (or infeasibility) is assessed, and a candidate solution is developed. The assessment of developing an optimal solution is discussed in (Deschaine, et. al, 2013). An optimal solution (or a set of such solutions) is generated by linking the model(s) with an optimization algorithm(s). Optimization algorithms can be either deterministic or stochastic, can solve for single- or multi-objective functions and the model equations and constraints can be linear, mildly nonlinear, or highly nonlinear. Furthermore, the characteristic of the model state equations can transition between linear and nonlinear, which can complicate the optimization process. Solving these situations is discussed in (Deschaine, *et. al*, 2013).

In the industrial decision making, the objective is to determine how to allocate limited resources optimally, in order to achieve a certain objective under the constraints. Such

decision problems can be formally modelled by corresponding constrained optimization (mathematical programming [MP]) models.

While the decision model formulation is evidently always problem specific, for illustrative purposes we provide a description of a constrained optimization problem. The objective is to minimize the total cost of an environmental or industrial management program expressed by the net present value of all related (construction and/or operational) cost components. The decision is subject to the following requirements: the solution has to meet all considered physical, engineering, environmental, and stakeholder constraints.

The goal of this section is to summarize the state-of-the art of algorithms for solving global optimization problems and selecting a set of algorithms to use to support making optimized decisions. Reflecting the needs of the industrial applications the present exposition will be focused on global nonlinear optimization.

The general MP model is defined by the following ingredients:

x	decision vector, an element of the real n-space $\boldsymbol{R}^n$;
$f(x)$	continuous objective function, f: $\boldsymbol{R}^n \rightarrow \boldsymbol{R}$, ($\boldsymbol{R}=\boldsymbol{R}^1$);
D	non-empty set of admissible decisions, a subset of $\boldsymbol{R}^n$.

More specifically, the set D is defined by:

l, u	explicit, finite n-vector bounds of x (a "box") in $\boldsymbol{R}^n$;
$g(x)$	m-vector of additional continuous constraint functions, g: $\boldsymbol{R}^n \rightarrow \boldsymbol{R}^m$

Applying these notations, the (continuous) MP model is stated as follows: we want to minimize the objective function $f(x)$ under the assumption that x belongs to the feasible set D. Applying standard notation, this is concisely expressed as

$$\min f(x)\ x \in D,\ D := \{l \le x \le u,\ g(x) \le 0\} \subset \boldsymbol{R}^n. \qquad (2.8)$$

In the definition of set D, all vector inequalities are interpreted component-wise (since l, x, and u, $l<u$ are n-vectors), and the zero in the relation $g(x) \le 0$ denotes an m-vector. The components of x are denoted by $x_1, x_2,\ldots,x_n$; the components of the vector function g are functions $g_1, g_2,\ldots,g_m$.

The model formulation is generalized as follows:

- Maximization problems can be deduced to the general form by using $-f$ as the objective function.

- Similarly, = and ≥ constraint relations and/or explicit lower and upper bounds regarding the constraint function values can be simply deduced to the model form (2.8).

- If the set of additional constraint functions g is empty (m=0), the formulation is a box-constrained optimization model.

Also, combinatorial optimization problems with discrete variables and thus mixed integer-continuous optimization problems can be, at least in a formal sense, directly transformed into continuous GO models (Pintér, 2002).

Next, we introduce the key concepts of *local vs. global* optimality. The point $x_l^* \in D$ is a local solution of (2.8) if $f(x_l^*) \leq f(x)$ holds for all points $x \in D$ located within a certain "neighborhood" of x_l^*. In the real n-vector space, the concept of a neighborhood can be defined by some norm function. (For concreteness, we can think of the standard Euclidean norm.) The point $x^* \in D$ is a global solution of (2.8) if $f(x^*) \leq f(x)$ holds for all points $x \in D$. The entire set of global solutions will be denoted by X^*.

The basic analytical assumptions stated above guarantee that the optimal solution set X^* of the MP model is non-empty. [7]

The above technical remarks imply that (2.8) covers a general class of optimization models and useful for general industrial optimization problems discussed in this thesis. Consequently, this class includes as provided in the examples, difficult model instances for which traditional (local) optimization methods will typically fail. Local scope search methods, as a rule, find only local solutions depending on the starting point ("initial solution guess") of the search algorithm. A significant class of nonlinear models for which local scope optimization suffices is the minimization of a convex function over a convex set. For completeness, we include definitions of convexity. The set $D \subset \boldsymbol{R}^n$ is convex if for each pair of points from D, the entire line segment connecting these points also belongs to D. A function f is convex over D if its level sets $D_c := \{x \in D: f(x) \leq c\}$ are convex, for all real values of c. If the decision model does not meet (essentially) these convexity requirements, then in general solving the model calls for global scope algorithms. This optimization scheme is implemented for environmental groundwater

[7] This key result directly follows by the classical theorem of Weierstrass that establishes the existence of the minimizing point set of a continuous function over a non-empty, closed and bounded set.

remediation design which demonstrates an industrial strength tool that supports the site-wide optimization approach discussed in Paper II (Deschaine, *et al.,* 2013). Rio Tinto Iron Ore (RTIO) mining company computed a cost savings of 9.1M (AUD) from the application of this design optimization approach in the development of a dewatering strategy for Nammuldi Mine, Pilbara region, Western Australia (Lim, *et al*., 2012).

2.5 Individual analysis techniques investigated for integrated modeling and optimization analysis

The broad array of options associated with the aforementioned solutions techniques can be overwhelming to the decision support analyst. First, examination of the existing analysis options occurred before developing and extending algorithms to solve the specific industrial problems. The industrial applications provided in this thesis discuss and exemplify specific algorithms for testing and demonstrating the optimal decision support approach. Since the approach is general, the specific method within the class of SME, engineering-oriented, and data-driven approaches can be substituted for the ones used in this work (listed in Table 3). Specific solution techniques are interchangeable. Their viability for application in this integrated approach can and likely will be problem specific. Their inclusion into the decision process can be evaluated for information value by using the techniques developed in this thesis.

Table 3. List of pre-existing primary analysis components deployed in thesis

ANALYSIS COMPONENT CATEGORY	TECHNIQUES
Subject Matter Expert	AHP
Data Mining	CGPS, WEKA, PCA
Numerical Models	PTC, BioFT3D™, MODHMS™, UTCHEM™
Data Fusion	CGPS, Kalman Filter w/GSLIB™, WEKA
Optimization	LP, SLP, SLA, SQP, GA, OA, LGO, MINLP
Information Theory	Eckschlager, K., Stepanek, V., (1979), mRMR

Note: AHP = Analytic Hierarchy Process; BioFT3D = Biological Fate and Transport in 3 Dimensions; MODHMS= Surface and Subsurface flow and transport code; UTCHEM=Subsurface multiphase flow and transport code; CGPS = Compiling Genetic Programming System; PCA= Principal Components Analysis; PTC = Princeton Transport Code; LP=Linear Programming; SLP = Sequential Linear Programming; SLA = Sequential Linear Approximation; SQP = Sequential Quadratic Programming; GA = Genetic Algorithm; OA = Outer Approximation; LGO = Lipschitz Global Optimization; MINLP = Mixed Integer Nonlinear Programming; mRMR = minimum Redundancy Maximum Relevance.

Table 4 provides a summary of key contributions by the Ph.D. candidate.

Table 4. Summary of principal state-of-the-art technology extensions in modeling techniques

KNOWLEDGE QUADRANT	CONTRIBUTIONS
Subject Matter Expert	• Extended AHP to include optimization, including probability distribution functions of expert opinions, constraints and resources to enable a solution to stochastic optimization challenge. • Applied to: o Superfund regulation modeling o Environmental health & safety optimization o Mercury retirement o Optimal technology section for Low CO_2 power plant design
Engineering-orientated	• Extended the Kalman Filtering with data fusion approach. Optimizes the worth of sample data. Considers the complete monitoring system including monitoring data interactions with model predictions. • Applied to: o Two applications of optimal groundwater contamination monitoring for MGD drinking water resource protection.
Data Mining	• Developed and validated adaptive power generator modeling approach for probabilistic response technique for economic power dispatch predictive modeling. • Uncovered the need for use of high mutation rates for effective program development in CGPS. • Applied to: o Power utility grid operations o Eight industrial problems
Model Blending	• Developed high accuracy munitions of explosive concern identification (MEC-ID) algorithm. • Linked multiple and mixed models with optimization. • Applied to: o Geotechnical property estimation o Properties impacted with subsurface explosive devices.

The remainder of the thesis discusses these algorithms, methods, approaches and examples in more detail, including references to published work by the candidate which has demonstrated beneficial results in several industrial settings.

3. Subject Matter Expert Modeling using AHP

An example of SME modeling technique using AHP is presented in Paper I. Decision analysis is initiated by creating a depiction of the problem and solution; its attributes, information, and goal. Once the problem domain and solution space are qualitatively understood, a decision analysis approach may be selected. A comprehensive SME algorithm was developed and deployed for application to a statewide environmental impact priority ranking algorithm for 114 sites from the Connecticut Department of Transportation operations (Deschaine *et al*., 1985). While successful, this comprehensive decision support algorithm was missing the ability for including varied degrees of opinions from multiple experts of different knowledge pedigrees.

In situations where both the amount of data and knowledge concerning the physical processes is difficult to quantify, the AHP is an effective algorithm candidate to use. AHP, developed at the Wharton School of Business (Saaty, 1980), allows decision makers to model a complex problem in a hierarchical structure showing the relationships of the goal, objectives (criteria), sub-objectives, and alternatives as shown in Figure 3. This is a well-understood and widely used algorithm. We developed a non-trivial extension of the AHP algorithm with the formal inclusion of uncertainty in expert opinion (as a replacement for the AHP single-point averaging approach), and provide for optimization even when constraints are non-deterministic. Specifically, we extended AHP to include random variables, stochastic simulation, and optimization under uncertain information and future conditions. These developments were extensions beyond the work described in (WSRC, 1996; Deschaine *et al.*, 1998a and Deschaine *et al*., 1999). The approach cited as a leading application in the environmental restoration decision support field by the U.S. Army Engineer Research and Development Center (Kiker *et al*., 2005) and by distinguished leaders in the decision analysis field (Forman and Gass, 2001). A similar work (Ghodsypour and O'Brien, 1998) also uses the AHP method and linear programming as the optimizer, which is similar to the candidate's work (WSRC, 1996 and Deschaine *et al*., 1997) but theirs was applied to the supplier selection problem.

This AHP discussion approach is demonstrated here through case study examples. Appendix A of (USEPA, 2002) provides the mathematical details with example computations. An example of the deterministic AHP method is provided via a

spreadsheet model (Ragsdale, 2003). We use a modified spreadsheet implementation as one of the platforms to extend the AHP capability. Advantages of this platform choice include ease of implementation and enabling stakeholder involvement. Understanding the analysis and solution and being involved is critical for acceptability. This also enables exploring what-if scenarios and thus iteratively adjusting the problem and optimizing the solution as appropriate.

3.1 Analytic hierarchy process

Figure 4 depicts an analytical, hierarchal decision process, and specifically the AHP process overview. It is a structured decision tree.

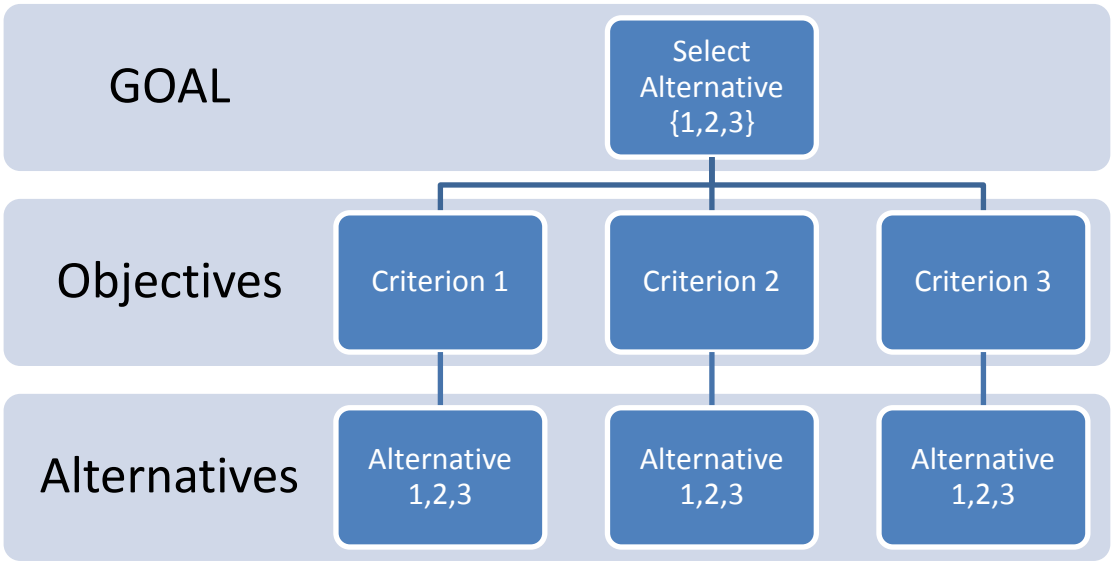

Figure 4. AHP Decision Hierarchy. In the AHP process, SMEs are queried to build both the model structure as well as work through the pair-wise comparisons, and different experts may be involved at the various stages of model building and analysis. We extended AHP to include random variables, stochastic simulation, and optimization under uncertain information and future conditions.

AHP provides SMEs a formal method to apply and describe their expertise to a problem. Specifically, the application of data, experience, insight, and intuition is processed in a logical and well-defined approach. The AHP enables decision makers to make judgment calls concerning the pair-wise comparison of alternatives. These judgments are used to derive ratio-scale priorities or weights as opposed to arbitrarily assigning them on some linear scale. In doing so, AHP not only assists decision makers by enabling them to structure complicated interrelated systems and exercise judgment, it also allows them to incorporate both objective and subjective considerations in the decision process.

The steps in applying AHP to a decision problem include:

Step 1: Problem identification and research

a) Problem identification.

b) Identify objectives and alternatives. A list of the pros and cons of each alternative is often helpful in identifying the objectives.

c) Research the alternatives.

Step 2: Eliminate infeasible alternatives

a) Determine the "musts."

b) Eliminate alternatives that do not meet the "musts" criteria.

Step 3: Develop decision structure

a) Structure a decision model in the form of a hierarchy to include goal, objectives (and subobjectives), and alternatives.

b) Add other relevant factors (such as scenarios) as required.

Step 4: Evaluate the factors in the model by making pair-wise relative comparisons.

a) Use a sufficient amount of available factual data. Interpret the data as it relates to satisfying the objectives (that is, do not assume a linear utility curve without thinking about whether it is a reasonable assumption).

b) Use knowledge, experience, and intuition for these qualitative aspects of the problem or when no hard data is available.

Step 5: Best alternative computation

a) Synthesize to identify the "best" alternative.

b) Once judgments are entered for each part of the model, the information is synthesized to achieve an overall preference. The synthesis ranks the alternatives in relation to the goal.

Step 6: Examine and verify decision. Iterate as required.

a) Examine the solution and perform sensitivity analyses. If the solution is sensitive to factors in the model for which accurate data are not available, consider spending the resources to collect the necessary data and iterate back to Step 4.

b) Optimization. Deterministic AHP can be optimized using linear programming (when the equations are all linear), or with global and stochastic optimization techniques.

c) Check the decision against intuition. If they do not agree, ask why intuition suggests that a different alternative is best. See if the reason is already in the model. If not, revise the model (and or judgments). Iterate as required as both the model and intuition may change as more information about the problem becomes available.

Step 7: Documentation

a) Document the decision for justification, quality assurance, and quality control and for decision forensics.

3.1.1 Stochastic AHP with optimal decision capability

The above deterministic AHP is demonstrated as a powerful and successfully applied algorithm; however, it is deterministic. It *averages* the insight of the SMEs and does not fully capture nor transmit the value contained in the *uncertainty* of the SMEs or group of experts. We extended the basic AHP algorithm, and the following is implemented:

1) Each expert vote is recorded as in deterministic AHP.
2) Assemble the votes and process as a collection of knowledge – a discrete distribution.
3) Assign weights to the SME's vote (e.g., uniform or weighted based on experience).
4) Using Monte-Carlo simulation techniques evaluate AHP stochastically.
5) Optimize the decision using stochastic optimization.

3.2 Successful examples of AHP analyses

Successful applications of both the deterministic and stochastic AHP algorithm are presented below. The AHP was developed by (Saaty, 1980) while we developed all the algorithmic extensions. Several precedent applications occurred prior to the development of Paper I, wherein new algorithms were developed and applied to the following real-world scenarios. This work in optimal decision support analyzes multiple expert opinions, uses these opinions as probability distributions (point averages or probabilistic distributions may be used), and uses optimization technology to provide a risk-based optimal answer that included uncertainty in SME knowledge of technology success and funding levels.

3.2.1 Superfund legislation review

An AHP model was designed to simulate the effect of proposed legislation in the U.S. Congress on the cost and effectiveness of the Superfund Cleanup program. This proposed bill (HR 2500) considered using a combination of passive and active cleanup approaches to best respond to environmental contamination whilst providing a measure cost effectiveness. The analysis involved coding the proposed legislation in AHP and analyzing 50 Superfund Records of Decision (ROD). The cleanup decisions from the model of the proposed legislation were compared to the baseline decisions of the documented decisions in the 50 RODs. The results (Deschaine *et al*., 1999) showed that a 35 percent cost savings ($9.1B of the expected $26B cost for the Superfund program) is achievable through enactment of the proposed legislation. The Business Roundtable^{SM} conducted an independent analysis and developed similar cost saving numbers. This legislation was not enacted. However, the analyses were valuable in contributing to discourse towards discussing environmental remediation solutions in the U.S. that included the expanded use of institutional controls and long term monitoring.

3.2.2 Mercury retirement

In Paper I, the options for the long term management of surplus elemental mercury in the U.S. were assessed using deterministic AHP. A limited scope, multi-criteria, decision analysis was performed. Two general types of treatment technologies were evaluated (stabilization/amalgamation and selenide) and combined with four disposal options as follows: (a) hazardous waste landfill, (b) hazardous waste monofill, (c) engineered below-ground structure, and (d) mined cavity. In addition, the following

three storage options for elemental mercury were considered: (a) aboveground structure, (b) hardened structure, and (c) mined cavity. Alternatives were evaluated against criteria that included costs, environmental performance, and compliance with current regulations, implementation considerations, and technology maturity, potential risks to the public and workers, and public perception. Considering non-cost criteria only, the three storage options rank most favorably. If both cost and other criteria are considered, then landfill options are preferred (because they are the least expensive). Storage options rank unfavorably on cost. This is because even relatively small annual costs add up over time. Also, storage is a temporary solution. Eventually, a treatment and disposal technology must be adopted, which adds to the cost. However, the analysis supports continued storage for a short period (up to a few decades) followed by permanent retirement when treatment technologies have matured. In the models developed for mercury retirement options, consistency ratios of 0 to 6 percent were achieved. This work described herein was accepted by the USEPA and institutionalized in EPA Guidance Documents (USEPA, 2002 and 2005).

3.2.3 Worker health and safety improvements

DOE desired an increase in health, safety, and general well-being of workers performing various activities working in and near Cold War-generated residual radiologic environments at its Savannah River Site (SRS) in South Carolina, USA. In achieving this goal, DOE faced the competing issues of continued safe production and maintaining plant operations and requested a generalized prioritization algorithm with optimization capabilities be developed for optimizing portfolios with over 100 radioactively impacted areas for United States Department of Energy (DOE). An AHP model was developed linked with linear programming and deployed using a test case of 113 radiological impacted areas at SRS that required on-going human activities be conducted in them. The optimal analysis proposed a programmatic approach whereby an estimated $18M per year could be saved while improving worker safety (Deschaine, 1998a, b). The U.S. Government invested $2M in a pilot program to test the approach and cost savings of $20M was realized (Coffield and Guy, 1998). This resulted in an $18M savings predicted in the first year, the actual savings “tracked to the penny”.

3.2.4 New electrical power plant technology (source selection) analysis

The DOE's National Energy Technology Laboratory (DOE-NETL) is tasked with developing energy technology for supplying efficient and uninterrupted power supplies to the United States in the 21st Century. DOE requested an algorithm be developed to increase the effectiveness and efficiency of program management by assisting program and project managers that support research and development in selecting best value projects. This resulted in an SME-based optimization algorithm for deciding which new technologies to fund for low CO_2 power plant design (Deschaine *et al.*, 2001). Because each of the selected projects must fit into the mission of the DOE program as a whole, this is an interconnected matter because each project has an uncertain cost and benefit which can propagate throughout the plant design and achievable performance. The challenge is further complicated by the uncertainty associated with funding availability for the program as a whole and year-to-year. The algorithm assists in the areas of program design, balanced project source selection, quantitative program and project cost, time and technical achievability risk assessment, decision support, and optimization. Continuing to fund a project that is not producing the desired outcomes is referred to among program managers as "escalation of commitment to a failing course of action"; on average 30 to 40 percent of innovative technology projects exhibit escalation (Brockner, 1992). Knowing when to deselect, or terminate a project, is as weighty as selecting it in the first place. The key to breaking the cycle of spending good money after bad early on is:

- Develop performance metrics that allow decision makers to recognize problems early on;
- Re-examine the course of action;
- Search for an alternative course of action; and
- Implement an exit strategy as warranted.

The tool kit developed as part of this work provides the quantitative tools to implement such a cost- and resource-saving approach. This includes, for example, comparing a project risk-adjusted baseline with actual project performance and assessing trends. This is accomplished by inserting the actual costs and task durations as the project proceeds, and updating the prediction and statistics of the expected final project costs and duration with uncertainty ranges. If the project begins to exceed a threshold, a

project warning can be issued. Additional thresholds can be set for project termination, and remaining resources can be reallocated to enhance existing projects or fund alternative projects. See Deschaine, *et al*., (2001) for further details.

3.3 Summary

This section discusses determination of the performance of the AHP solution approach to be applied when both data availability is low and, specific engineering-oriented models are unavailable. The AHP solution was examined, and when limitations detected, we extended that algorithm to enhance applicability and performance. The extended algorithm was demonstrated as successful using real-world case studies which included stakeholder acceptance. The practical impact of deploying this AHP technique has: assisted in reducing the cost of environmental remediation while protecting human and environmental receptors; provided guidance and direction on a long term solution for the mercury retirement issue; provided safer work environments in radiological areas, and; facilitated design choices for low-carbon emitting electrical power plants in the United States.

4. Engineering-oriented modeling

4.1 Engineering-oriented modeling

Optimal decision making in this knowledge quadrant requires the description of the process being optimized be in some form of physically based engineering-oriented model. This is a model that captures the specific process being investigated. It is also used to test the performance of various alternative actions. For the purpose of this demonstration, we tackled the challenge of identifying the extent of subsurface contaminant plumes and optimal remedial design. Subsurface contamination conditions can render drinking water non-potable. Contamination can originate from a variety of sources such as chemical spills, leaking underground storage tanks, subsurface chemical disposal, or carbon sequestration. Investigation of the subsurface impacts requires boring into the ground and collecting samples of water and soils. The processes of water flow, material transport, and eventual fate of the chemicals in the subsurface are simulated by building numerical models. These models are physically based. They are calibrated and validated on site-specific information. Obtaining observational data is expensive. This is due to the high cost of collecting samples from material at depth below the ground surface.

Subsurface simulators represent in mathematical form the natural physical processes of flow and transport in the aqueous phase and soil matrices. These processes occur in porous or fractured media. The flow system can consist of saturated or variably saturated conditions. The fluids that flow in the media can be either single phase (water) or multiphase including light and dense non-aqueous phase liquids (LNAPL/DNAPL). Simulation of additional phases requires using a water-LNAPL/DNAPL-gas multi-phase formulation. Transport processes can be represented as simple advective particles, as a single non-reactive solute that undergoes dispersion and diffusion, and as a multi-component system in which constituents interact with each other and the surrounding environment and thus are subject to hysteresis. The equations that govern these processes are Lipschitz-continuous. The mathematical characteristics of the state-space equations include linear and non-linear, elliptic, parabolic or nearly hyperbolic when dominated by advection. When no/low diffusion/dispersion is present relative to advection, steep fronts or shocks occur in the solutions. In practice, diffusion and dispersion are almost always designated as having non-zero values. The sheer size

of the problems considered in this work is a result of the large physical domain (area and/or depth) and the number of linear/nonlinear equations actively solved for multiple phases and multiple chemical species simulated. Excellent references for the general governing equations are found in Bear, (1979); Huyakorn and Pinder (1983); Abriola (1983); Baehr (1984); de Marsily (1986); Bear *et al* (1993); and Pinder (2002). Biogeochemical process descriptions, including natural attenuation of fuels and solvents in subsurface systems, with discussions on policy and complex chemistry, are described in NRC (2000); Wiedemeier *et al*. (1999); and ITRC (2007). Sequential solution methods for these equations of science and engineering are found in Lapidus *et al* (1999). Parallel solutions are discussed in Dongarra *et al*. (2003) and Grama *et al*. (2003). The user manuals for modeling code describe how the general equations are assembled into a unique numerical simulation tool. The candidate developed a process guide for DOE for selecting groundwater modeling codes (BWXT and SAIC, 2002).

4.2 State estimation techniques

4.2.1 Optimal estimation techniques

The goal of optimal estimation is to be able to develop an estimate of the subsurface conditions with respect to flow and transport for the least cost. This estimate then becomes the state of the system for optimization and decision making under uncertainty. Because the costs associated with collecting information about the subsurface is high, this industrial reality results in often only sparse data-derived knowledge is available for interpretation and analysis. Hence, the engineering-oriented models are used to estimate subsurface conditions between the measurement locations.

This work provides an example of providing a near optimal estimate of subsurface conditions via in-depth physically based models, sparse data observations, and Kalman filtering. The algorithm developed extended is based on a previous algorithm (McGrath and Pinder (1996). The extension provides the computation of the maximal knowledge gained from an additional monitoring point(s) relative to the entire monitoring system. The previous algorithm provided the point estimate value only.

Typical model calibration uses the sequential paradigm of collecting data, calibrating the model and using it for making predictions. Kalman filtering provides an optimal method for estimating the state of the system. It fuses the information content of both

the model and the data simultaneously, not sequentially. This is in concert with the multi-sensor fusion approaches discussed in Luo and Kay (1989). The signal is derived by using the physically based model as a transmitter. The noise is provided by the geologic uncertainty. Samples collected from groundwater monitoring wells provide the observations at specific (discrete) times and locations (including depth). The uniqueness being instead of gathering data, calibrating a model, and then using just the model for future predictions, one uses both the model and the recent observed data to provide the optimal state estimate update, which is conceptually similar to approaches such as reinforcement learning or other on-line learning algorithms.

To generate the signal, geostatistical simulation provides an estimate of the subsurface variability of a system property, such as hydraulic conductivity. This approach generates an ensemble of representative realizations, such as various representations of the conductivity field. These realizations become the inputs to the computational fluid dynamic models, which produces analysis results as a distribution of subsurface conditions as opposed to a single-valued estimate. Predictions of the subsurface condition are available from both the models and the data from field surveys. Common practice involves using data to calibrate a model. The calibrated model provides a tool to generate predictions. The future observations are unused when developing the predictions. These models become out of date and must be recalibrated periodically. However, this thesis approach assembles and integrates all the current information. Hence, fusing the model and the current field monitoring data provides the optimal state estimate. This type of problem falls into the general field of optimum filtering and the stochastic signal extraction from noisy data (Reza, 1994). The Kalman filter provides the data-model fusing function. This approach provides both the optimal state estimate and the locations where the uncertainty in the state estimate is minimized if a data value is available. This is the ideal location for data collection via a soil boring or groundwater monitoring well. Stochastic simulation generates a probability distribution function for the unknown (e.g., concentration of contaminant in groundwater) within the bounds of the observed data and quantifies uncertainty.

Common parameter estimation in the geosciences groundwater modeling community consists primarily of Bayesian estimators, co-kriging estimators, geostatistical inverse methods, Kalman filtering, least squares methods, maximum likelihood methods, and pilot point techniques (McLaughlin and Townley, 1996). They showed that all these

methods are special cases of the Gaussian maximum *a posteriori* estimator. Additionally, it is shown that using equivalent assumptions, the Kalman filter is equivalent to the least squares estimate, maximum likelihood estimate, and the maximum *a posteriori* estimate (Gelb, 1974; Jazwinski, 1970; Lewis, 1986 and Stengel, 1994). Comprehensive references provide the mathematics of state-space models and linear estimation including Kalman filtering (Kalith *et al*., 2000 and Govindaraju, 2002).

The first references found using Kalman filtering in groundwater investigations appeared in 1990s. Techniques have since been developed to integrate the information content from both the predictive models and the observed measurements. The technique used in this work uses a computational fluid dynamic model with a Kalman filter. This approach has been demonstrated to provide the best unbiased estimate of the subsurface conditions integrating the uncertainty in the simulator and field data (Herrera *et al.*, 1998a, b; and Zhang, 2002).

4.2.2 Extended Kalman filtering

The extended Kalman filter is a method to combine the information from samples with the model predictions. Kalman Filtering fuses the data (available at a discreet time and spatial location) with the predictions of the subsurface simulator to provide the minimum error estimate of subsurface conditions.

For extended Kalman filtering to be effective, a stochastic representation of the aquifer is necessary. Stochastic aquifer realizations were discussed above and obtained via geostatistical analysis. These approaches often use variograms to generate aquifer realizations; the set of these realizations is called the "ensemble." The ensemble can be generated automatically and can represent different conceptual models of the subsurface created by various experts. This is because an accurate deterministic representation is difficult to obtain, so one is always dealing in stochastic nature and uncertainty when developing predictions of subsurface behavior.

The Kalman filter composition consists of essentially of two components:

1) *The propagation component* that specifies how the conditional moments (i.e. hydraulic head, contaminant distribution, flow velocity fields) evolve between

times information is available (via sensor measurements). This component performs what a subsurface flow and transport simulator typically perform.

2) *The updating component* incorporates the new information and specifies how the propagated moments are modified. This component performs the activity typical of a parameter estimation algorithm.

The key benefit that a Kalman filter provides is a formal way to integrate information from the physical simulator and the field data. Rather than analyze the information separately, the Kalman filter updates both the mean and the covariance of the model state and associated parameters. Since the conditional statistics are used as the uncertainty measure and not the spatial variability, the assumption of ergodicity is not required. "Ergodicity" refers to a stationary random function and its ability to tend towards the stationary mean of its cumulative distribution function (CDF). This concept is used widely in geostatistical analysis, and this is a salient point. At the scales that are of interest in most flow and transport studies, the conditional hydrobiogeochemical moments are most likely non-stationary and, hence, nonergodic. It should be noted that the updated estimates need not be mass conservative, but the best representation of the mass available given the uncertainty of the information available about the system and its performance.

The Kalman filter is a recursive algorithm. It is a convenient way to integrate the predictions between a subsurface simulator and field data. It estimates the state variables in a linear system by optimally combining the information content of the model and data, incorporating uncertainty. In linear systems, the Kalman filter estimate is the true conditional mean, the truly optimal (minimum variance) estimate. However, the Kalman filter must be extended to handle nonlinear systems, such as most groundwater flow and transport challenges. Linearizing the state equation around the latest parameter estimates to approximate the conditional mean does this. Essentially, this formulation is like a series of linear batch filters. Practice has shown that even with this reduced dimensionality and linearization, the extended Kalman filter will provide a useful estimate that is close enough to the conditional mean and mode.

This work builds on the foundational algorithm for developing optimal monitoring well networks (McGrath and Pinder, 1996). The algorithm is extended by computing and minimizing the uncertainty surface using the entire volume of uncertainty over the

entire area of groundwater impact. This initial algorithm only minimized using a local maximum point on the uncertainty surface. This approach provides a better quantification of the total uncertainty. To test the extended algorithm, we applied it two industrial case studies as discussed below.

4.3 Examples

4.3.1 Optimal monitoring well design – DOE Pantex Plant

The DOE Pantex Plant in Amarillo, TX, is a nuclear material plant covering 9,100 acres. It was established in 1942. Its mission was to build conventional munitions and high-explosive (HE) compounds in support of WWII. Currently, the plant is used to develop, test, and fabricate HE components; for nuclear weapons assembly and disassembly; for interim storage of plutonium and weapon components; and for component surveillance. Historical waste practices at the facility have resulted in 140 Solid Waste Management Units (SWMUs) containing metals, radionuclides, inorganics, perchlorate, various explosives such as Royal Demolition Explosives (RDX) and volatile organic compounds (VOCs), and semi-volatile compounds. Plant wastewater discharges have created a large mound (16 billion gallons) of contaminated groundwater at a depth of 250 to 300 feet below ground surface. The impacted groundwater is perched about 150 feet above Ogallala aquifer, which is the principal source of groundwater for the City of Amarillo and agriculture in the region.

The analysis objectives: (1) Develop a stakeholder-acceptable project approach; (2) Find and define the trichloroethene (TCE) and RDX plumes, and design a risk-acceptable remedial action; (3) Optimize long term monitoring to provide stakeholders with assurances that impacts are monitored properly; and; (4) Develop a contingent remedial design should conditions change (e.g. mitigate migration of contaminants from the perched water mound to the Ogallala aquifer should it occur).

Comprehensive analysis for environmental restoration is usually needed when the stakes are high, such as assessing the risk to a city's sole source of drinking water. In every stage of the analyses, calculation transparency ensures that partners and stakeholders understand and concur with the analysis regardless of the level of sophistication of the analysis or proposed response action. This ensures interested parties can be kept apprised of site investigation processes and progress and are fully prepared to review

and implement the best, all-round remedies. At the Pantex Plant, the transparency of the approach helped transform what had been an adversarial situation into a highly productive partnership between stakeholders and the government.

In response to a potential off-site groundwater contamination issue, a Pantex Plant Technical Advisory Group (TAG) was convened by the stakeholders. It was composed of plant personnel; representatives from universities, national laboratories, government centers of excellence, and state and federal regulatory authorities; SMEs (our involvement); and community stakeholders. As the TAG team conducted an analysis of the situation, it requested field tests and obtained guidance from leading SMEs on the groundwater flow and transport simulation and optimization tools that could simulate the subsurface processes operating within the complex geological, hydrological, biological, and chemical subsurface environmental systems. One example of the proactive approach taken at this site, a remedial action contingency plan was prepared to provide stakeholders with the confidence that timely, effective action would occur should a release to the sole source aquifer be detected. As part of the contingency plan, a remedial design tool was developed that consisted of linking the flow and transport model with the outer approximation optimization method to facilitate rapid, optimal designs for strategies to mitigation migration of contaminated groundwater.

Throughout the process, Pantex stakeholders were kept fully informed of the TAG activities and findings, which models and tools were being recommended for selection, and the rationale for selection. Stakeholder involvement also was facilitated through regular technical meetings and training courses, so once the analysis progressed to developing the simulation models and optimization systems, Pantex stakeholders understood and trusted the analysis, visualizations, and the three-dimensional (3-D) physical model that were used to translate the complex subsurface processes, and the associated models and physical observations into results.

As a result, the Pantex TAG reached an informed decision about the path forward and unanimously endorsed the simulation/optimization approach that was proposed to resolve the challenges at the plant (BWXT & SAIC, 2002). Further, the TAG continued to stay involved reviewing the work as it was implemented over a 10-year period.

Not only did the analysis reduce uncertainty and help gain stakeholder acceptance, it did so without conducting unnecessary work; as a result, plans could be implemented

efficiently and effectively (USEPA, 2010). Savings captured from avoiding unnecessary work were considerably, over $2M, and Pantex management invested the released funds in more productive activities. While cost savings from the optimized processes continue to increase, the bigger story was that the process facilitated the best operational designs and investments with unanimous, written stakeholder acceptance.

The approach is using a groundwater flow and transport model integrated with the Kalman filtering approach addressed the issue of assessing the potential for off-site chemical migration. The goal was to assess the residual uncertainty regarding potential TCE plumes in the Ogallala aquifer north of the Pantex plant using data from existing early warning monitoring well networks developed by the SMEs, the physically based modeling results, and Kalman filtering (Paper II). This integration approach used SME-determined groundwater monitoring well locations to identify locations where maximum uncertainty exists. We determined that the existing SME-developed monitoring well network for the TCE issue reduced uncertainty regarding the behavior of the plume by 93.4 percent and that only an additional 1 or 2 optimally located wells would be required to obtain the optimal results (a 98 percent reduction in uncertainty). The original plan to expand the monitoring well network consisted of installation of an additional 6 to 12 deep (~ 600 ft.) wells at an additional construction cost of $2M. All analyses and conclusions were accepted by DOE, USEPA, and state regulators, which indicated a high confidence in the analyses. DOE management reinvested the $2M savings into extending the accuracy of the site-wide, complex and comprehensive flow and transport model, which produced better solutions and additional savings, which were reinvested in additional analysis. The lesson learned from this application pertains to both the importance of developing optimization analysis techniques and the ability for these to be transparent and explained understandably. For some analyses that concern environmental contamination and potential degradation of natural resources, it is paramount that stakeholders accept both the tool and the analysis.

4.3.2 Optimal monitoring well design – DoD Anniston Army Depot

We conducted a second test of the extended algorithm at the Anniston Army Depot, a U.S. military (combat tank and armored vehicle maintenance depot) site that covers 15,000 acres in Anniston, AL. It has been contaminated with dense non-aqueous phase liquid (DNAPL) and dissolved TCE and breakdown products. The hydrogeological

environment consists of karstic and fractured rock. The objectives of the analysis were to define the extent of the dissolved TCE plume and assess the adequacy of the SME-developed monitoring well network. The extent of the TCE plume and in situ processes (e.g. stability analysis of redox zones and biodegradation) were analyzed. The analysis indicated that the plume extent was adequately characterized, and no additional downgradient wells were needed.

Specifically, based on the outcomes of the optimization through heuristic judgment, SMEs were able to reduce the long term monitoring program, from 200 wells sampled per year to 120 wells sampled per year. Optimization consisted of integration of stochastic physically based and data-driven models using the extension of the Kalman filtering (Deschaine *et al.*, 2010). The result indicated that only 40 wells needed annual sampling to monitor plume behavior. The stakeholders accepted the analysis and recommendations were implemented in 2004. The implementation realized a cost savings of $2.4M over the first five years (2004-2009). The savings continue to accumulate over time at that average annual rate to this day.

The plume identification approach using groundwater flow and transport model with sparse datasets when combined with remedial design optimization (Deschaine, et. al, 2013) is valuable when allocating resources among the various total costs of an environmental contamination challenge. The technique helps balance investigation with remediation and long term monitoring. It can be used as input to the programmatic optimization toolkit. The analysis also quantifies the level of certainty between investigations or various source areas on a site.

4.4 Summary

The engineering-oriented modeling approach was deployed at the DOE-Pantex site, which is located northeast of Amarillo, TX. The approach and the solution unanimously by the stakeholder group (BWXT & SAIC, 2002). This approach was applied and accepted at the Anniston Army Depot, a US Army base in Alabama. The applications resulted in gaining a significant understanding of the processes. Additional benefits include increases in operational efficiency and reduced costs. The ITRC requested the approach documented in Paper II be extended for use with DNAPL sites and documented in their guidance document. This document was for open distribution in 2007 (ITRC, 2007).

5. DATA-DRIVEN MODELING

5.1 Data-mining / machine-learning

The use of data mining on datasets with vast quantities of information is well documented and a field in and of itself with free, open-source tools (Hall *et al*., 2009). Machine-code-based, CGPS represents the direct evolution of binary machine code through GP techniques (Nordin *et al*., 1994, 1995a, 1995b, 1996, 1998a, 1999a and 1999b). Thus, an evolved CGPS program is a sequence of binary machine instructions. After completing a machine-coded CGPS project, the CGPS software decompiles the best evolved models from machine code into Java, ANSI C, or Intel Assembler programs. Paper III documents the testing and enhancements to the CGPS approach to increase the methods robustness including the finding of needing to use high mutation rates and also adding multiple restarts to avoid getting stuck in low fitness local minima. CGPS algorithm discussed in Paper III and was used in part in Paper IV. The candidate further developed an algorithm to assist PJM Interconnection with economic electric power dispatch challenges using machine learning (Ott, 2010).

Hence, this section is not concerned with whether the CGPS form of machine learning is an effective approach as this is demonstrated in (Paper III), but rather with expounding on some of those results, specifically:

1) Derive physically based laws from noisy data, and;
2) Develop a high-fidelity representation of a slow executing, physics-based model with faster execution times.

5.2 Examples

5.2.1 Deriving a physical law

CGPS evolves a program irrespective of dimensional analysis consideration. While the program may fit the data, it may not necessarily be consistent regarding the dimensional units unless the inputs and outputs are dimensionless.[8] We tested to assess whether CGPS could develop a physically based law in closed form and understandable to a SME. Darcy's Law was chosen as the test function as it describes the flow of water

[8] For a treatise on a GP paradigm that incorporates correct dimensional analysis, see (Kiejzer, 2001).

through soils which often the researcher has data with a high degree of noise in both the input and output data. Darcy's Law describes the flow of water through porous media. The equation in closed form is:

$$Q = K * I * A$$

Where:

Q = flow [L^3/T],

K = hydraulic conductivity [L/T],

I = gradient [L/L], and;

A = area [L^2].

To test whether CGPS can uncover the true physics, we generated a realistic input set of data points and then used Darcy's Law to produce outputs. We then added 10 percent random variation (white noise) independently to the inputs and outputs and applied the CGPS software on these data.

The best solution derived by the CGPS software from these data was a four-instruction program that is precisely Darcy's Law, represented in American National Standards Institute (ANSI) for the C programming language (ANSI C) as:

```
Q   =  0.0
Q +  =  I
Q *  =  K
Q *  =  A
```

In this CGPS-evolved program, *Q* is an accumulator variable that is also the final output of the evolved program. CGPS reproduced the closed form of Darcy's Law. This example demonstrates the plausibility of using CGPS to evolve understandable programs that conform to known physically based processes.

This program model of Darcy's Law was derived as follows. First, program evolution occurred by CGPS. The "raw" CGPS solution was accurate, though somewhat unintelligible. Application of intron removal (Nordin, 1996) and evolutionary strategies unveiled the form of Darcy's Law. This process is coded in the CGPS software; we used the "Interactive Evaluator" module, which links to the "Intron Removal," automatic

"Simplification," and "Optimization" functions. These functions combine heuristics and Evolutionary Strategies (ES) optimization to derive simpler versions of the programs that CGPS evolves.

5.2.2 Approximate function development of complicated production model

As discussed in Section 4, SMEs use their knowledge to develop engineering-oriented models that represent physical processes of environmental, energy or industrial significance. These codes are handwritten and undergo VV&A processes prior to use to solve a specified problem. There are considerable advantages to this approach, including complete understanding of the underlying physics, numerical methods and defined solution limits. Major disadvantages include the time needed to code and test the solutions, which can require years for multiphase finite elements. Also, some of the computationally intensive engineering-oriented models can have extended execution times, providing a challenge for use in simulation and optimization. For example, the solution to the DOE-Pantex subsurface environmental model includes simulation of a variably saturated flow system consisting of 45 square miles and a 1,000 foot depth required between 9 to 21 days on an AMD 2.4 GHz 64-bit processor. The case study discussed below, which uses a physically based industrial production simulator, demonstrates how machine learning can be used to reduce CPU execution times from 16 to 24 hours per run to mere milliseconds. The motivation for conducting this experiment of replacing the physically based model with a small footprint, fast running high fidelity approximation was that running the complete physically based model required more time than was available for the answer (Deschaine, *et al*., 2005), and efforts to speed up the code were not successful.

This test was used a dataset that contained both the input (five production-related variables) and the output from a complex and slow running process simulator (Deschaine, *et al*., 2005). While the physics of the process were well known to the external developers, the model is proprietary, and no information was provided regarding the process. Hence, the goal is to create a fast-executing model that is based solely on the data provided, and compare the associated run times and accuracy. No information concerning the production line process physics was transferred to us; this

represents a controlled test case for assessing whether machine learning can reproduce complicated numerical industrial model.

The dataset consisted of 7,547 solutions generated from various values from the five input variables common to making production decisions. The dataset took the engineers at Kodak months of CPU time to develop. The CGPS analysis was designed to capture the structure of the underlying dataset. The data was randomly divided into three subsets:

- Training: 2506 data points
- Testing: 2520 data points
- Blind Validation: 2521 data points

The learning occurred on the training dataset. The best-evolved programs were selected using the training and validation dataset. The applied dataset is used to assess if the "true" structure of the solution was captured. In other words, the applied data (also referred to as testing data) played no part in training or as part of the best program selection. Accordingly, the results on the applied data measure how well the evolved solution generalizes to unseen data. The results are provided in Table 5.

Table 5. Results of machine learning on Kodak process data.

Fitness [R^2]	Single Solution	Team Solution
Training	0.9934	0.9975
Validation	0.9893	0.9939
Applied	0.9783	0.9889

The CGPS machine learning algorithm produced these results. Multiple runs conducted by randomizing the parameter settings. This included values for population size, maximum program size, maximum number of the floating-point unit (FPU) registers, dynamic subset selection (DSS) subset size, percent by difficulty, crossover rate, homologous crossover rate, and mutation rate. The reported results represent the progress at the end of the 218^{th} run (64,290 generation equivalents).

To demonstrate repeatability, the analysis was conducted a second time and achieved R^2 fitness on the applied dataset of 0.989. The fitness on the training and validation datasets is 0.997. For comparison, using a statistical regression approach on this dataset yielded an R^2 fitness of 0.80.

CGPS is demonstrated to be able to represent with high fidelity a complex and proprietary simulator. The CGPS solution computed in milliseconds, whereas the engineering oriented model required hours. The solution produced results a high degree of accuracy [R^2 =0.989 to 0.997]. This CGPS solution executes on a standard personal computer in less than a second, compared to the hours required by the original physics-based process simulator. The resulting decompiled code may be linked to the optimizer and compiled, or it may compile into a dynamic link library (DLL) or a common object module (COM) object which can be called from the optimization routines and placed in a web page on the production floor or in a real-time system (Regmi, Deschaine and Regmi, 2004).

5.3 Summary

The machine learning tools CGPS, WEKA™ and Salford™ were used to perform the analyses. This research discovered the need to use high mutation rates and multiple restarts to increase the robustness of CGPS. With these enhancements, we demonstrated a valuable attribute of the CGPS machine learning system, namely its ability to derive closed form physically based laws and to produce a fast and accurate representation of a simulator. This enables production decisions be made in real-time in the case where quick, but not perfectly accurate solutions, have high value. CGPS enables development of these high-fidelity approximate solutions, which because they are fast and small, are easy to deploy in a web page or real-time system where fast execution (milliseconds) response is desired, such as economic dispatch on electrical power grids (Ott, 2010).

6. Integrated SME, engineering-oriented, and data-driven modeling

6.1 Problem statement

The final knowledge quadrant in the modeling approach decision space is concerned with the difficult challenges for which high accuracy solutions are needed, but have been elusive. This occurs in the areas of earth sciences and UXO discrimination, for example. In the situation where both data is plentiful, and physics is well understood, optimal decision-making in this quadrant requires that the process being optimized have some form of in-depth physical *understanding* for which to test the performance of various alternatives. However, it is not necessary to have a complete engineering-oriented model.

In this section, we detail how to develop accurate SME/engineering-oriented/data-derived models. This fused SME/physics-based/data-derived approach was inspired by Professor Donald F. Harlemen (MIT, 1983) with whom the candidate worked on the solar salt pond project for the Qattara Depression in Egypt and also from a statewide environmental portfolio optimization project (Deschaine, *et al.,* 1985). Further insight was gained from the research and development on intelligent agents (Deschaine *et al.*, 2000; Deschaine *et al*., 2001) and the works of Hutter (2001) and Schmidhuber (2002). See also the works of (Riuz-Mier, 1987) and (Giaglis, 2001). A review of the candidate's previous work and that of these researchers' work reveal a strong synergy with multiple points of conceptual contact with some of the approaches used in this thesis.

The approach for developing a fused system using supervised machine learning as the "glue" is implemented as follows:

- Step 1: Assemble the SMEs' opinions and information developed from the models constructed in Section 3 (SME, cf. Paper I).
- Step 2: Assemble the engineering-oriented model inputs gathered as part of the model development discussed in Section 4 (engineering-oriented models, cf. Paper II).

- Step 3: Assemble the raw data sources and the SME data expansions developed during Section 6 (Machine Learning, cf. Paper III).

- Step 4: Compute the mutual information content, accounting for maximum relevance and minimum redundancy, between the various inputs to develop the optimal combination of inputs. Inspect and understand these results (cf. Paper IV).

- Step 5: Apply supervised machine learning using the combined SME, data-derived, and physics-based data training dataset to build the fused model. Predict performance, and optimize if merited (Paper IV, Deschaine, *et al.*, 2013).

6.2 Demonstrations of integrated models

Two representative examples demonstrate the effectiveness of this process. These are from Paper III and Paper IV, respectively.

6.2.1 Example 1: soil analysis: percent fines (cone penetrometer)

Identifying the flow and transport of water and chemicals through soils is essential in the environmental field. Many decisions pertaining to risk assessment and remediation depend on the soil properties. The percentage of fine material in soils is a fundamental parameter for predicting the rate of flow and transport of water and chemicals in the subsurface. Because collecting soil samples directly is expensive, one promising technique we assess is whether a truck mounted cone penetrometer mobile sampling device can be used to make subsurface soil sampling more cost-effective. This instrument works by pushing a rod into the ground and measuring the resistance (via its sleeve and tip pressures). We investigate whether this information can reliably be used to infer the value of percent fines in soils with depth at each location (Paper III). If successful, we would be able to generate a 3-D representation of the estimated hydraulic conductivity from this information. There exist physics-based equations that describe this relationship; however, for the dataset provided, we analyzed the fitness (R^2) of these physically based equations ranged from 0.26 to 0.40, which is well below the project requirements of 0.70. The machine learning approach CGPS was applied to the raw data (not using the physically based relationships), and an R^2 fitness value of 0.60 was achieved, which though better, was still below project requirements.

The physics equations were then separated into their mathematical terms and added to the raw data inputs and CGPS system we applied. The value in doing this is that when in-depth physically based knowledge about a process is known, it should be used because it provides the machine learning algorithms with information about relationships between the variables it otherwise would need to discover. Hence, the CPU time used by machine learning is applied to learn the unknown aspects of a problem. Using this approach, the CGPS system evolved a solution with R^2 fitness of 0.72, which met project requirements.

To the best of knowledge, this is the first time CGPS has been used to develop an optimal estimate of a system from the components of an engineering-oriented model based on partial physics based descriptions and combined with raw data. This approach opens up many possibilities of linking physically based and data-derived models for optimal estimation as there are no assumptions about any of the properties of either the physics simulator or the statistical distribution of the data. The approach also demonstrates that when SME knowledge is available regarding observed data relationships, it is wise to use it to develop robust, accurate models. Without an understanding of the underlying physics, we were unable to derive the project required accuracy solely from the raw data inputs. This illustrates the value of an in-depth (physics-based) and broad (data-driven) search technique for finding high-fitness algorithms. The machine learning CGPS approach increases predictive algorithmic performance by extending the accuracy of the physics-based model. Using integrated site and process specific data provides a more comprehensive formulation.

6.2.2 Example 2: UXO discrimination and certification

The objective of this project was to demonstrate a computer-assisted process to discriminate subsurface MEC items from clutter based on electromagnetic induction (EMI) and magnetometer (MAG) data. The algorithm development has been ongoing since calendar year 2000 (cf. Paper IV). The algorithmic test reported herein was conducted using data from the Camp Sibert test site (Keiswetter, 2008) and peer reviewed by USDoD using blind data.

The MEC discrimination process uses geophysical signals to provide a data description of a buried anomaly. This data is then processed using geophysical software and exposed to a variety of supervised machine-learning techniques (Paper IV and

Deschaine *et al.*, 2002). The purpose of the analysis is to ascertain whether a subsurface anomaly is a MEC item or not. This is accomplished by uncovering relationships in the data and understanding their respective power in contributing to a prediction of probability of the presence of a MEC item. First, we developed a minimum ellipse model to produce features from the raw geophysical sensor data. This mathematical structure produces hundreds of site-specific features. These features consist of information from the raw ungridded data, along with the physics and statistical topology of the channel-specific response signals. Application of information theory is applied to determine the most relevant features. Then, supervised machine learning is used to develop a predictive model that predicts MEC presence. These techniques are combined to form a successful computer-based search strategy (Paper IV).

This approach was inspired by the inverse modeling work conducted in numerous other Strategic Environmental Research and Development Program (SERDP) and ESTCP projects along with optimal estimation work done for the DoD and DOE. The machine learning method automatically develops the function that maps the feature information to the decision of whether an anomaly is MEC or is not a MEC item. Teams of solutions are generated by solving the discrimination problem using various algorithms. This assesses the certainty of the MEC prediction and provides fail-safe protection from potential errors that could occur and go undetected if only one solution method was used. Error correcting code increases the discriminative power when solving sites with multiple MEC types existing in a single dataset. By inspecting and performing functional analysis on the relevant inputs and the developed computer code, insights of the value of principal predictors are gained.

The test data is from the former Camp Sibert site in Alabama, USA. It was provided by USDoD. The analysis uses both physically based and data-extracted features information from two geophysical instruments: EMI and MAG. Information theory and principle components were used to reduce the size of the dataset. The analysis using information theory compared the discrimination value of the inputs (computational geometry [data] versus physics). We found that the computational geometric approach contained all the information the inverse physics modeling approach provided for the EMI sensor. The integrated modeling technique with information content assessment identifies the type and level of analysis needed, so it facilitates resource planning for these types of projects. This MEC discrimination approach was successful and received

official certification from the U.S. Office of the Secretary of Defense (OSD) in 2009 (OSD email to Larry Deschaine, 2009). Predictability of an algorithmic performance is as valuable as the ability to produce accurate results. Predictions of how well the approach performed were confirmed upon receipt of the independent validation test results from OSD.

6.3 Summary

The value of the decision support approach presented is two-fold. First, by using both the data- and the combined engineering-oriented/SME-based features in solving these difficult problems, the machine learning solution algorithms have demonstrated higher accuracy than when the non-integrated features are used separately. This is a result of combining the depth (physics and SME) and breadth (data-driven) approaches and by providing expected relationships between inputs to learning algorithms, the search space is better focused on achieving higher accuracy solutions. This was demonstrated for both MEC identification and also predictions of percent fines in soils using only cone penetrometer (friction and pressure) measurements and physical understanding.

Also, the analyses discussed in Papers IV identify when data from various sources is synergistic or redundant. Reduction of redundant data results in increased performance of machine learning algorithms. Secondly, by including physically based and data-based attributes the machine learned solution becomes an extension of the engineering-orientated model, calibrated to the specific site data. This approach of using machine learning for UXO/MEC discrimination was first proposed and successfully demonstrated as discussed in Deschaine *et al*. (2002). Banks *et al*. (2005) independently verified the use of the genetic programming method for MEC discrimination. It has now become part of the industry standard process for conducting UXO/MEC discrimination projects as evidenced by multiple works documented on the SERDP and ESTCP web pages.

7. Discussion and conclusions

A structured decision support approach is developed and tested. This includes algorithm development and extensions of existing works for solving a wide variety of challenges in the energy, industrial and the environmental fields. Solution of specific challenges is based on the availability, depth and breadth of SME knowledge, engineering-oriented models and data. The approach implements the graded use of models and model types. Model integration or simplification, which can consist of any or all of the model types, uses machine learning as the integrative "glue". The approach is tested first by demonstrating that each model type implemented independently can provide value for solving problems in the environmental and energy fields. Then, the integrated approach is evaluated solving two problems not solvable to the required accuracy standards by a singular model type; 1) the percent fines estimation using cone penetrometer information, and; 2) the MEC identification problems using non-destructive investigative techniques. Table 6 demonstrates the diversity of verified applicability of this approach and summarizes the technical contributions.

The principles of this integrated decision support approach are that problems can be categorized into four quadrants depending on the relevance of knowledge from data, engineering-oriented models, human expertise sources, and the ability to integrate the various sources and models. Fortunately, for many problems in the environmental, industrial and energy fields, some form of solution attempt has been developed for increasing productivity, efficiency and effectiveness. When appropriate, its use is advocated. Unfortunately, there exist a class of problems where the available and current state-of-the-art solution techniques fall short of developing the optimal solution. Since these types of problems are cost intensive, even a modest increase in productivity can translate into significant savings. Hence, from an industrial perspective, it makes sound business sense to invest in an analysis technique that provides desirable return on investment (ROI). The problems discussed above have provided both favorable ROI and increased the quality of the natural and work environments as well as guided optimal operations of the electrical power grid.

Table 6. Summary of the candidate's contributions of the optimal decision support approach application

Industry	Example	Relevant Thesis Paper	Contribution
National Association of Manufacturers	Regulatory Analysis (Superfund Reform; HR2500)	Paper I	Provided analysis of new regulations that supported environmental remediation decision making and strategy development (Deschaine *et al*., 1999). Analysis characterized cost savings of Superfund program at 35%.
USEPA	Mercury Retirement	Paper I	Provided analysis of options for the long-term planning of mercury retirement; a toxic metal with known serious health effects when exposed. Published in USEPA Guidance documents and (Paper I).
DOE-SRS	Health & Safety	Paper I	Provided input for a program to improve the health and safety of workers in radiological environments (Deschaine *et a*l., 1998).
DOE-NETL	Low CO_2 Power Plant Design	Paper I	Provided a portfolio investment tool. Tool supported optimal research and development (R&D) program design. Included optimal allocation funding for promising technologies for lowering greenhouse gas emissions that contribute to global warming (Deschaine *et al*., 2001).
DOE-Pantex	Human Health Risk Assessment, Corrective Measures Study, and Groundwater resource protection	Paper II	Provided input for an environmental monitoring, compliance and remediation program. The program concerned chemical releases to groundwater. The subsurface environment includes a sole-source drinking water aquifer. (BWXT & SAIC, 2002).
RTIO Nammuldi Mine	Mine Dewatering	Paper II	Provided optimal design of mine dewatering to minimize impacts to the environment while enabling below ground iron ore extraction resulting in 9.1M (AUD) savings over planned design (Lim, *et al*., 2012).
DOE-Pantex and DoD-Anniston Army Depot (ANAD)	Optimal monitoring	Paper II	Provided an optimal monitoring plan for tracking environmental contamination such that containment of plumes can be monitored effectively and safely and groundwater resource protected (Paper II).
DoD-Umatilla Army Depot	Optimal Environmental Remedial Design	Paper II	Optimal design of groundwater remediation systems saves both time and money, reduces greenhouse gas emissions (Deschaine, *et. al.* 2013).
PJM Interconnection	Economic Power Grid Dispatch	Paper III	Provided *a priori* knowledge of electrical power generator's response to an economic dispatch signal; thereby enabling more efficient and reliable power grid operations (Ott, 2010).
Kodak	Industrial Process	Paper III	Developed high fidelity fast executing industrial process model to enable real-time assessment and control (Paper III).
DOE-SRS	Geologic material characterization (percent fines)	Paper IV	Developed a geophysical predictive model with increased accuracy thereby reducing the level of effort in characterizing the subsurface (Paper III).
DoD-ESTCP (UXO/MEC multiple sites)	MEC Identification	Paper IV	Demonstrated the value of using machine learning and optimization for non-destructive identification of underground explosive objects with less effort (Paper IV).

Note: The focus on the technology development, extensions and societal benefits are significant. Benefits are not always easy to quantify in dollars. Industry has enjoyed the results of applying this approach, both through reduced costs and better solutions. Documented savings include $18,000,000 in the first year (DoE-SRS) and repeatedly in excess of $1,000,000 (RTIO 9.1M AUD, DoD ANAD $5.52M, DoE Pantex $2M)); from the benchmark conditions. Cost savings figures are provided for the purpose of industrial thesis value determination.

The most information comprehensive modeling quadrant is where the SME, engineering-oriented, and data-derived dataset provide different information. To leverage this for better problem solving, a new approach using the CGPS genetic programming paradigm as the integration method is proposed and demonstrated. This approach ensures that all the information content from all sources is available for use in the final solution. This approach changes the common paradigm in environmental model building that consists of sequentially using the SME knowledge and available data to form the basis for building an engineering-orientated model which can then be optimized (Deschaine, *et al*., 2013). The proposed integrated modeling approach retains as separate inputs the SME, data, and engineering-oriented models. This approach reduces information loss. Improvement in the predictive models results from the fusion of the information from various sources using machine learning modeling techniques as the "glue". The integration methodology fuses the engineering-oriented modeling capability and the SME knowledge with the unadulterated information content in the raw (observed) data. To keep the problems tractable, information theory analysis identifies the relevant and redundant information. Machine learning modeling integrates the relevant information from the various sources and input, and provides a solution, which has been empirically shown equal to or superior than using solely the individual information sources (e.g. data, engineering-oriented, or SMEs).

Modeling and optimization was conducted using a broad and in-depth understanding of the problem for developing the solution via integrated modeling. Essentially, when used as the glue in this integrated approach, the machine learning can be thought of as completing the SME physically based engineering-oriented modeling equations of understanding in a manner that honors the observed data.

Acknowledgements

Many people supported this work through ideas, conversations and criticism. A significant number of colleagues have been co-authors on published papers and conference presentations, over 100 in all, over a 30 year time span. Many independent peer reviewers from journals, industries, conferences, and government agencies provided excellent blind feedback. All the input is appreciated. A few key people stand out regarding development of this thesis. The first is Kristian Lindgren, who constantly promoted and advised this work; fostering and encouraging deeper and deeper understanding and promoting clarity of presentation. Jan Peter Nordin provided significant technical advices throughout the approach development, and was instrumental is translating computational concepts of genetic programming into useable computer algorithms. Thomas L. Saaty supported the SME expert system technology and the success at DoE-SRS. Peter S. Huyakorn, and Ashok K. Katyal supplied valuable technical comments and support regarding subsurface modeling (engineering-oriented models). Frank Francone supported the CGPS technology (machine learning) enhancements used in part of that work. The Pontifical Biblical Commission (PBC) provided the approach (integrated modeling) of developing a comprehensive understanding of an over-arching message contained in various independent or related information sources by keeping them as independent inputs, including assessing their relevance and redundancy, and integrating them into a coherent understanding. This is a model for developing understanding with breadth and depth. This hierarchical and integrated modeling thesis approach shares many points of contact with PBC guidance; the challenge was translating those concepts in a manner that enabled development of this mathematical based decision support approach for industrial application. Janos D. Pinter, George F. Pinder, and David P. Mark have provided excellent instruction over the years regarding optimization. Optimization is the necessary technology required to compute the best solution regardless whether the industrial model is developed using a stand-alone approach (SME, EO, DD), or is a model based on partially or fully integrated approaches.

REFERENCES

Abriola LM (1983). Mathematical Modeling of the Multiphase Migration of Organic Compounds in a Porous Medium, Ph.D. Dissertation, Princeton University, Princeton, NJ, USA.

Altrock CV (1995). Fuzzy Logic and Neurofuzzy Applications Explained, Prentice Hall, 350 pages.

Altrock CV (1997). Fuzzy Logic and Neurofuzzy Applications in Business and Finance, Prentice Hall, 375 pages

Baehr AL (1984). Immiscible Contaminant Transport in Soils with an Emphasis on Gasoline Hydrocarbons, Ph.D. Thesis, University of Delaware, Newark, DE, USA.

Bear J (1979). *Hydraulics of Groundwater*, McGraw-Hill.

Bear J, Tsang CF, and de Marsily G. (1993). *Flow and Contaminant Transport in Fractured Rock*, Academic Press.

Banks ER, Núñez E, Agarwal P, Owens C, McBride M and Liedel R (2005). Genetic Programming for Discrimination of Buried Unexploded Ordnance (UXO), GECCO, 2005. http://www.cs.bham.ac.uk/~wbl/biblio/gecco2005lbp/papers/66-banks.pdf.

Beyer HG (1998). The Theory of Evolution Strategies, Springer, 380 pages.

Bigus JP (1996). Data Mining with Neural Networks, McGraw-Hill, 220 pages.

Blackmore S (1999). The Meme Machine, Oxford Press, 264 pages.

Boole G (1854). An Investigation of The Laws of Thought on which are Founded the Mathematical Theories of Logic and Probabilities. Reprinted by Dover Publications, 1958, 424 pages.

Brockner, J. (1992). The escalation of commitment to a failing course of action: Toward theoretical progress. Academy of Management Review, 17(1), 39-61.

Brodie R (1996). Virus of the Mind: The new science of the Meme. Integral Press, Seattle Washington, 251 pages.

Buchanan BG (1985). Some Approaches to Knowledge Acquisition, Knowledge Systems Laboratory, Stanford University, Report No. KSL-85-38, October.

Bundy (1996). Artificial Intelligence Techniques – A Comprehensive Catalogue, Springer, Fourth edition, Berlin, 248 pages.

BWXT and SAIC (2002). Recommendation of the 2001 Groundwater Modeling Technical Advisory Group for the DOE Pantex Plant, 125 pages.

Caglayan A and Harrision C (1997). Agent Source Book, John Wiley & Sons, Inc. Publishers, 349 pages.

CALGO (2009). Collected Algorithms of the ACM. Internet site: http://www.acm.org/calgo.

Combinatoria (2009). Mathematica collection of Algorithms for combinatorics, and graph theory. Internet; anonymous ftp: ftp.cs.sunysb.edu in the pub/Combinatorica directory.

Carrol DL (2001). Fortran Genetic Algorithm (GA) Driver, CU Aerospace, Urbana, ILL, ftp://cuaerospace6.cuaerospace.com/pub.

Chambers LD (1999). Practical Handbook of Genetic Algorithms: Complex Coding Systems, Volume III, CRC Press, 572 pages.

Chambers LD (2001). Practical Handbook of Genetic Algorithms: Applications CRC Press, 501 pages.

Chernick MR (1999). Bootstrap Methods: A Practitioners Guide. Wiley Interscience, 264 pages.

Coffield T and Guy G (1998). “Pollution Prevention through Contaminated Area Rollbacks – Deploying Technologies and Industry Best Practices to Reduce Waste”. DOE-Savannah River Site, WSRC FY98, PBI#8 Status review, Westinghouse Savannah River Company, Slide 9, September 29.

Coley DA (1999). An Introduction to Genetic Algorithms for Scientists and Engineers, World Scientific Press, 227 pages.

Cox E (1995). The Fuzzy Systems Handbook, 2nd Ed., Academic Press, 716 pages.

Davidson AC and Hinkley DV (1997). Bootstrap methods and their Applications, Cambridge University Press, 582 pages.

Davis M (1982). Computability and Unsolvability. Dover Press, 248 pages.

Dawkins R (1989). The Selfish Gene (2 ed.), Oxford University Press, p. 192

DeJong K (1992). Are Genetic Algorithms Function Optimizers? In Manner & Mendrick, pages 3–13.

De Marsily G (1986). *Quantitative Hydrology: Groundwater Hydrology for Engineers*, Academic Press.

Deming EW (1943). Statistical Adjustment of Data, Dover Publishing, 261 pages.

Department Of Defense (2000) Standard Practice For System Safety MIL-STD-882D 10 February 31pp.

Deschaine LM and Singarella PN (1985). "Priority Ranking of 114 Salt Storage and Maintenance Facilities with a Lotus Spreadsheet Model," Prepared for the Connecticut Department of Transportation in concert with a Department-wide environmental assessment and remediation program. Presented at the MIT Microcomputer Short Course, Cambridge, MA, June.

Deschaine LM, Breslau B, Selg RA and Ades MJ (1997). Optimal Allocation of Multi-Project Remediation Resources, Presented at the Joint Chemical Safety Workshop, Co-sponsored by the Chemical Manufacturers Association and Department of Energy, Washington, DC, July 23 and 24.

Deschaine LM, Breslau B, Ades MJ, Selg RA and Saaty TL (1998a). Decision Support Software to Optimize Resource Allocation − Theory and Case History, in Society for Computer Simulation's Advanced Simulation Technology Conference − Waste Management and Environmental Sciences Section, Boston MA, USA, April. Simulators International XV, Simulation Series Vol. 30, No. 3, ISBN: 1-56555-144-3, pages. 139-144.

Deschaine LM (1998b). Enhanced Work Process (EWP) Guidance Document, US Department of Energy, Savannah River Site, Aiken, SC.

Deschaine LM, Zafaran F, Forman EH and Kluck V (1999). Simulation of Proposed Superfund Legislation using Expert Choice. Presented at the National Conference on Environmental Decision Making, Knoxville, TN, USA, April 11-14.

Deschaine LM, Brice RS, and Nodine MH (2000). Use of InfoSleuth to Coordinate Information Acquisition and Analysis in Complex Applications. (Agent-based). Society for Computer Simulation's Advanced Simulation Technology Conference, Washington, DC, USA, April. ISBN 1-56555-199-0, pp. 13-18.

Deschaine LM, Rawls P, Manfredo L, and Patel JJ (2001). The DOE NETL Program and Project Source Selection, Risk Quantifier, Management Support and Optimization Tool-Kit The Society for Modeling and Simulation International: Advanced Simulation Technology Conference, Seattle, WA, USA, April. ISBN: 1-56555-238-5, pages 165-174.

Deschaine LM, Hoover RA, Skibinski JN, Patel JJ, Francone FD, Nordin P and Ades MJ (2002). Using Machine Learning to Compliment and Extend the Accuracy of UXO Discrimination Beyond the Best Reported Results of the Jefferson Proving Ground. Technology Demonstration, pages 46-52. Society for Modeling and Simulation International Advanced Technology Simulation Conference, San Diego, CA, USA, April.

Deschaine, L. M., and Francone, F. D. (2002). Design Optimization Integrating the Outer Approximation Method with Process Simulators and Linear Genetic Programming, Joint Conference on Information Science, March, Research Triangle Park, NC, ISBN 0-9707890-1-7, pages 618-621.

Deschaine LM, Huyakorn P, Guvanasen V and Lillys T (2010). FTMO: The Comprehensive Flow, Transport, and Management Optimization Tool-kit. NDIA: Environment, Energy, & Sustainability Symposium & Exhibition, Colorado Convention Center Denver, CO, USA, June 14-17.

Deschaine, Larry M., Theodore P. Lillys, and János D. Pintér (2013). "Groundwater remediation design using physics-based flow, transport, and optimization technologies." Environmental Systems Research 2, no. 1: 6.

Dongarra J, Foster I, Fox G, Gropp W, Kennedy K, Torzon L, White A (2003). Sourcebook of Parallel Computing, Morgan Kaufmann.

Eberhart R, Simpson P, and Dobbins R (1996). Computational Intelligence PC Tools – An Indispensable Resource for the Latest in: Fuzzy Logic, Neural Networks, Evolutionary Computing. Academic Press, Inc., 464 pages.

Eckschlager K and Stepanek V (1979). Information Theory as Applied to Chemical Analysis, John Wiley and Sons, Inc.

Fausett L (1994). Fundamentals of Neural Networks – Applications, Algorithms, and Applications. Prentice Hall, 461 pages.

Feynman (1996). Feynman Lectures on Computation, Westview Press, England, 303 pages.

Findler NV (1991). An Artificial Intelligence Technique for Information and Fact Retrieval - An application in Medical Knowledge Processing, MIT Press, Cambridge, MA, USA, 155 pages.

Franklin S (1999). Artificial Minds, MIT Press, Cambridge, MA, 449 pages.

Fogel DB (2000). Evolutionary Computation: Towards a New Philosophy of Machine Intelligence. 2nd ed. IEEE Press, 270 pages.

Forman EH and Gass SI (2001). The Analytic Hierarchy Process – An Exposition, undated, Operations Research, 49(4), pages 469 486.

GAMS (2009). Guide to Available Mathematical Software, internet site: http://gams.nist.gov.

Garey MR and Johnson DS (1979). Computers and Intractability: A Guide to the Theory of NP-Completeness, Freeman Press, 338 pages.

Gelb A (1974). Applied Optimal Estimation. The MIT Press, 382 pages.

Ghodsypour SH and O'Brien C (1998). A decision support system for supplier selection using an integrated analytical hierarchy process and linear programming. International Journal of Production Economics 56-67: 199-212.

Giaglis GM (2001). A taxonomy of business process modeling and information systems modeling techniques. International Journal of Flexible Manufacturing Systems; April.

Giarratao J and Riley G (1998). Expert Systems: Principles and Programming, 3rd Ed. PWS Publishing Company, 597 pages.

Goldberg DE (1989). Genetic Algorithms: In Search, Optimization, and Machine Learning, Addison-Wesley, Publishers, 412 pages.

Govindaraju RS (2002). Stochastic Methods in Subsurface Contaminant Hydrology. American Society of Civil Engineers. 410 pages.

Grama A, Gupta A, Karypis G and Kumar V (2003). Introduction to Parallel Computing, Addison Wesley Publishers, 2nd ed.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, and Witten I (2009). The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

Hernández JE, Zarate P, Dargam F, Delibašić B, Liu S and Ribeiro R (eds.) (2012). Decision Support Systems – Collaborative Models and Approaches in Real Environments, Euro Working Group Workshops, EWG-DSS 2011, London, UK, June 23-24, 2011, and Paris, France, November 30 - December 1, 2011, Revised Selected and Extended Papers Series: Lecture Notes in Business Information Processing, Vol. 121, 2012, XIV, 167 pages.

Herrera GS and Pinder GF (1998a). Cost-Effective Groundwater Quality Sampling Network Design. In: Proceedings Computational Methods in Water Resources: pages 51–58, Crete, Greece.

Herrera GS, McGrath WA and Pinder GF (1998b). Computer-Aided Risk Assessment in Problems of Groundwater Contamination. In: Proceedings of the First International Conference on Computer Simulation in Risk Analysis and Hazard Mitigation. WIT Press, Computational Mechanics Publications: 51–60, Southampton, UK.

Hofstadter D (1995). Fluid Concepts and Creative Technologies, Perseus Books, LLC (BasicBooks), 518 pages.

Hutter M (2001). Proceedings of the 12th European Conference on Machine Learning (ECML-2001) pages 226-238.

Hutter M (2005). Universal Artificial Intelligence, Sequential Decisions based on Algorithmic Probability, Springer, 300 pages.

Huyakorn, PS and Pinder, GF (1983). Computational Methods in Subsurface Flow, Academic Press, 473 pages.

Hyvarinen A, Karhunen J and Oja E (2001). Independent Component Analysis, Wiley Interscience, 481 pages.

Interstate Technology and Regulatory Council (ITRC) (2007). Simulation and Optimization of Subsurface Environmental Impacts; Investigations, Remedial Design and Long Term Monitoring of BioNAPL Remediation Systems. Chapter 9, in In-Situ Bioremediation of Chlorinated Ethene DNAPL Source Zones: Case Studies, Prepared by The (ITRC), Bioremediation of Dense Non-Aqueous Phase Liquids (Bio DNAPL) Team.

Jacob C (2001). Illustration Evolutionary Computation with Mathematica, Morgan Kaufmann, 578 pages.

Jazwinski AH (1970). Stochastic Processes and Filtering Theory, Academic Press, San Diego.

Kailath T, Sayed AH and Hassibi B (2000). Linear Estimation. Prentice Hall, 854 pages.

Keijzer, M. (2001). Scientific discovery using genetic programming. IMM, Informatik og Matematisk Modellering, Danmarks Tekniske. 173 pages.

Keiswetter D (2008). SAIC Analysis of Survey Data Acquired at Camp Sibert, Interim Report, ESTCP Project MM-0210, July, 2008. 112 pages.

Kiker GA, Bridges TS, Varghese A, Seager TP and Linkov I (2005). Application of Multicriteria Decision Analysis in Environmental Decision Making, Integrated Environmental Assessment and Management –Volume 1, Number 2 – pages 95–108

Klir GJ and Yuan B (1995). Fuzzy Sets and Fuzzy Logic: Theory and Applications, Prentice Hall, Inc. 574 pages.

Konar A (2000). Artificial Intelligence and Soft Computing – Behavior and Cognitive Modeling of the Human Brain. CRC Press, 786 pages.

Koza JR (1992). Genetic Programming: On the Programming of Computers by Means of Natural Selection, MIT Press.

Koza JR (2009). Web page titled: Seven Differences Between Genetic Programming and Other Approaches to Machine Learning and Artificial Intelligence http://www.genetic-programming.com/sevendiffs.html and web page titled: What is genetic Programming, http://www.genetic-programming.com/gpanimatedtutorial.html

Koza JR, Bennett FH, Andre D and Keane MA (1999). Genetic Programming III, Darwinian Invention and Problem Solving. Morgan Kaufmann, San Francisco, 1154 pages.

Kuhn TS (1977). The Essential Tension: Selected Studies in Scientific Tradition and Change. University of Chicago Press. 366 pages.

Kuhn TS (1996). The Structure of Scientific Revolutions, 3rd Ed. The University of Chicago Press, 212 pages.

Kurzweil R (1992). The Age of Intelligent Machines, MIT Press, 579 pages.

Lapidus L and Pinder GF (1999). Numerical Solution of Partial Differential Equations in Science and Engineering, Wiley Interscience.

Laplante P (1996). Great Papers in Computer Science, IEEE Press, 717 pages.

LEDA (2009). The Library of Efficient Data types and Algorithms, web site; http://www.mpi-sb.mpg.de/LEDA/leda.html.

Lewis (1986). Optimal Estimation with an Introduction to Stochastic Control Theory. Wiley-Interscience, 368 pages.

Luo R and Kay M (1989). Multisensor integration and fusion in intelligent systems, IEEE Transactions on Systems, Man and Cybernetics 19 (5) (1989) pages 901–931.

Luger GF (2002). Artificial Intelligence - Structures and Strategies for Complex Problem Solving, Addison-Wesley publishers, 856 pages.

Mainzer K (1997). Thinking in Complexity: The Complex Dynamics of Matter, Mind and Mankind, 3rd ed. Springer, 361 pages.

Man KF, Tang KS and Kwong S (1999). Genetic Algorithms, Springer, 344 pages.

Masters T (1995). Neural, Novel & Hybrid Algorithms for Time Series Predictions, Wiley Publishers, 514 pages.

McGrath WA and Pinder GF (1996). Sampling Network Design for Delineating Groundwater Contaminant Plumes. In: Proceedings: Computational Methods in Water Resources: pages 185–192, Cancun, Mexico.

McGrath WA (1997). Sampling Network Design to Delineate Groundwater Contaminant Plumes. Research Center for Groundwater Remediation Design, October. University of Vermont, Burlington, VT, USA.

McGrath WA, Pinder GF, Olivaries J and Dougherty, DE. (1997). *MOCUS Users Guide, Version 1.0*. Research Center for Groundwater Remediation Design, University of Vermont, Burlington, VT, USA.

McLaughlin DB and Townley (1996). A Reassessment of the Groundwater Inverse Problem. In: *Water Resources Research*, 32(5): pages 1131–61.

MIT (1983). Personal communication with Professor Donald F. Harlemen, Supervisor of MIT undergraduate research opportunities solar salt ponds project.

Michalewicz Z and Fogel DB (2004). How to Solve It: Modern Heuristics, Springer, 467 pages.

Miller JH and Page SE (2007). Complex Adaptive Systems: An Introduction to Computational Models of Social Life, Princeton University Press, 284 pages.

Minker J (1993). An overview of Non-Monotonic Reasoning and Logic Programming, Journal of Logic Programming, Special Issue, 17.

Minsky M (1960). Steps Towards Artificial Intelligence, Institute of Radio Engineers, October 24.

Minsky M (1986). Society of Mind, Simon and Schuster Publishers, NY, NY, USA, 339 pages. Book and associated CD.

Mitchell M (1996). An Introduction to Genetic Algorithms. MIT Press, 209 pages.

Mitchell TM (1997). Machine Learning, Co-published with the MIT Press and WBC/McGraw-Hill, 414 pages.

Nagel E and Newman JR (1986). Gödel's Proof, New York University Press, 118 pages.

National Research Council (2000). Natural Attenuation for Groundwater Remediation, National Academy Press.

Nelson H (2009). Office of Secretary of Defense. Email to Larry M. Deschaine, PI for ESTCP Discrimination project MM-0811 (http://www.estcp.org/Technology/MM-0811-FS.cfm), email dated May 21, 2009.

Netlib (2009). Library of Algorithms on the world wide web; http://www.netlib.org/.

Nordin JP (1994). A Compiling Genetic Programming System that Directly Manipulates the Machine Code. In: *Advances in Genetic Programming*, K. Kinnear, Jr. (ed.), MIT Press, Cambridge, MA, USA:

Nordin JP and Banzhaf W (1995a). Complexity Compression and Evolution. In Proceedings of Sixth International Conference of Genetic Algorithms, Morgan Kaufmann Publishers, Inc.

Nordin JP and Banzhaf W (1995b). Evolving Turing Complete Programs for a Register Machine with Self Modifying Code. In: Proceedings of Sixth International Conference of Genetic Algorithms, Morgan Kaufmann Publishers, Inc.

Nordin JP, Francone F and Banzhaf W (1996). Explicitly Defined Introns and Destructive Crossover in Genetic Programming. Advances in *Genetic Programming 2*, K. Kinnear, Jr. (ed.), MIT Press, Cambridge, MA, USA.

Nordin JP (1997). Evolutionary Program Induction of Binary Machine Code and its Applications, Krehl Verlag, 290 pages.

Nordin JP, Francone F and Banzhaf W (1998). Efficient Evolution of Machine Code for CISC Architectures Using Blocks and Homologous Crossover. In Advances in Genetic Programming 3, MIT Press, Cambridge, MA, USA.

Nordin JP, Banzhaf W and Francone FD (1999a). Advances in Genetic Programming: Volume III, Chapter 12. MIT Press.

Nordin JP (1999b). Evolutionary Program Induction of Binary Machine Code and Its Applications, Krehl Verlag.

Orchard RA (1998). ERB-1054 FuzzyCLIPS Version 6.04A User's Guide, Integrated Reasoning Institute for Information Technology, National Research Council Canada, 88 pages.

Ott A (2010). Market and Planning Efficiency Through Improved Software and Hardware-Enhanced optimal power flow models (Washington, DC; free webcast), FERC: Development of Enhanced Generation/Demand Response Control Algorithm, PJM Interconnection. June 23. 23 slides. http://www.ferc.gov/EventCalendar/Files/20100623161840-Ott,%20PJM%206-23-10.pdf.

Peng H, Long F and Ding C (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, in IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8, pages 1226-1238.

Pinder, GF (2002). Groundwater modeling using geographic information systems, John Wiley & Sons; ISBN: 0471084980; 1st edition, 224 pages.

Pintér JD (2002). Global optimization: software, test problems, and applications. In Pardalos, P. M. and Romeijn, H. E., eds. Handbook of Global Optimization, Volume 2, pp. 515-569. Kluwer Academic Publishers, Dordrecht.

Pyle D (1999). Data Preparation for Data Mining, Morgan Kaufmann, 540 pages.

Pyle D (2003). Business Modeling and Data Mining, Morgan Kaufmann, 692 pages.

Ragsdale C (2003). Spreadsheet Modeling and Decision Analysis, South-Western College Pub; 4th edition, 864 pages.

Rao VB and Rao (1995). C++ Neural Networks & Fuzzy Logic, 2nd ed., MIS Press, NY, NY, 551 pages.

Regmi S, Deschaine LM, and Regmi SR (2004). High Fidelity Approximation of Slow Simulators Using Machine Learning for Real-time Simulation/Optimization. Society for Modeling and Simulation – International Advanced Technology Simulation Conference, Washington, DC, April.

Reza FM (1994). An Introduction to Information Theory, page 464. 528 pages.

Ruiz-Mier S and Talavage J (1987). A hybrid paradigm for modeling of complex systems Journal of Simulation, Novak B., Society for Computer Simulation International San Diego, CA, USA, Volume 48, Issue 4, April 1, 1987, pages 135 – 141.

Saaty TL (1980). The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation, McGraw-Hill, 287 pages.

Saaty TL (1996). Decision Making for Leaders; The Analytic Hierarchy Process for Decisions in a Complex World. RWS Publications, Pittsburgh, PA.

Sapaty P (1999). Mobile Processing in Distributed and Open Environments, Wiley Interscience, John Wiley & Sons, Inc., 410 pages.

Schmidhuber J (1997). A Computer Scientist's View of Life, the Universe, and Everything, in Foundations of Computer Science: Potential- Theory-Cognition, Lecture Notes in Computer Science, C. Freksa, ed., (Springer: Berlin).

Schmidhuber J (2002). Speed Prior and Optimal Simulation of the Future. In M. Ades and L. M. Deschaine, editors, Proceedings of the Business and Industry Symposium, 2002 Advanced Simulation Technologies Conference, San Diego, CA, USA. Simulation Series, vol. 34:4, p. 40-45, (invited).

Schwefel HP (1995). Evolution and Optimum Seeking, Wiley Interscience, 444 pages.

Sinha NK and Gupta MM (2000). Soft Computing & Intelligent Systems – Theory and Applications, Academic Press, London, 639 pages.

Skapura DM (1996). Building Neural Networks, Addison Wesley, NY, NY, USA, 286 pages.

Stanford (2009). The Stanford GraphBase, internet site, anonymous ftp from labrea.stanford.edu, directory pub/sgb.

Stengel (1994). Optimal Control and Estimation. Dover Publications, 639 pages.

Sutton RS and Barto AG (2000). Reinforcement Learning: An Introduction. MIT Press, 322 pages.

Swingler K (1996). Applying Neural Networks, - A Practical Guide. Academic Press, London, 303 pages.

United States Environmental Protection Agency (USEPA) (2002). Preliminary analysis of alternatives for the long term management of excess mercury, Report 600/R03/048, 128 pages.

USEPA (2005). Economic and Environmental Analyses of Technologies to Treat Mercury and Dispose in a Waste Containment Facility, EPA/600/R-05/157, April, 100 pages.

USEPA (2010). PANTEX PLANT (USDOE) in Carson County, Texas Site Status Summary, USEPA Region 6. http://www.epa.gov/region6/6sf/pdffiles/0604060.pdf.

USEPA-MOFAT (1991). Version 2.0.a - May 1991. http://www.epa.gov/ada/csmos/models/mofat.html

van Genuchten MTh and Wierenga PJ (1976). Mass transfer studies in sorbing porous media: I. Analytical solutions. Soil Sci. Soc. Am. J., 40(4): 473-480.

Wiedemeier TH, Rifai HS, Newell CJ and Wilson JT (1999). Natural Attenuation of Fuels and Chlorinated Solvents in the Subsurface, John Wiley & Sons.

Weiss G (1999). Multi-agent Systems: A Modern Approach to Distributed Artificial Intelligence, The MIT Press, 619 pages.

Weiss SM and Indurkhya N (1998), Predictive Data Mining – A Practical Guide. Morgan Kaufmann Publishers, 228 pages.

Westinghouse Savannah River Corporation (WSRC) (1996). Enhanced Work Planning Rollback Handbook (U), Environmental Health Technical Assistance Program, US Department of Energy. EWP Facilitator, (Released for unlimited distribution), Document No. WSRC-IM-96-0166. December. 88 pages.

Witten IH and Frank E (2000). Data Mining – Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 369 pages.

Zhang, Y. (2002). Optimal design of Groundwater Quality Networks, Ph.D. Dissertation, May, University of Vermont, Burlington, VT, USA.

Zuse K (1967). Rechnender Raum, Elektronische Datenverarbeitung, Vol. 8, pages 336-344 (via MIT translation into English, 1970, 98 pages).

ELSEVIER

Journal of Environmental Management 71 (2004) 35–43

Journal of Environmental Management

www.elsevier.com/locate/jenvman

Application of the analytic hierarchy process to compare alternatives for the long-term management of surplus mercury

Paul Randall[a,*], Linda Brown[b], Larry Deschaine[b], John Dimarzio[b], Geoffrey Kaiser[b], John Vierow[b]

[a]*Land Remediation and Pollution Control Division, National Risk Management Research Laboratory, Office of Research and Development, US Environmental Protection Agency, 26W Martin Luther King Drive, Cincinnati, OH 45268, USA*

[b]*Science Applications International Corporation, 20201 Century Boulevard Germantown, MD 20874 USA*

Received 1 July 2003; revised 30 December 2003; accepted 6 January 2004

Abstract

This paper describes a systematic method for comparing options for the long-term management of surplus elemental mercury in the US, using the analytic hierarchy process as embodied in commercially available Expert Choice software. A limited scope multi-criteria decision analysis was performed. Two (2) general types of treatment technologies were evaluated (stabilization/amalgamation and selenide), combined with four (4) disposal options: (a) hazardous waste landfill; (b) hazardous waste monofill; (c) engineered below-ground structure; and (d) mined cavity. In addition, three storage options for elemental mercury were considered: (a) aboveground structure; (b) hardened structure; and (c) mined cavity. Alternatives were evaluated against criteria that included costs, environmental performance, compliance with current regulations, implementation considerations, technology maturity, potential risks to the public and workers, and public perception.

Considering non-cost criteria only, the three storage options rank most favorably. If both cost and other criteria are considered, then landfill options are preferred, because they are the least expensive ones. Storage options ranked unfavorably on cost because: (a) even relatively small per annum costs will add up over time; and (b) storage is a temporary solution and, sooner or later, a treatment and disposal technology will be adopted, which adds to the cost. However, the analysis supports continued storage for a short period (up to a few decades) followed by permanent retirement when treatment technologies have matured.

Suggestions for future work include: (a) involving additional stakeholders in the process, (b) evaluating alternatives for mercury-containing wastes rather than for elemental mercury only, (c) revisiting the analysis periodically to determine if changes are required, (d) conducting uncertainty analyses utilizing Monte Carlo-based techniques.

Published by Elsevier Ltd.

Keywords: Mercury; Mercury management; Analytic hierarchy process; Environmental decision analysis

1. Introduction

Over the past decade, the US Environmental Protection Agency (EPA) has promoted the use of alternatives to mercury because it is a persistent, bio-accumulative, and toxic (PBT) chemical. The Agency's long-term goal for mercury is the elimination of mercury released to the air, water, and land from anthropogenic sources. The use of mercury in products and processes has decreased. The US Department of Defense (DoD) and the US Department of Energy (DOE) have excess mercury stockpiles that are no longer needed. Mercury cell chlor-alkali plants, although still the largest worldwide users of mercury, are discontinuing the use of mercury in favor of alternative technologies. The amount of surplus elemental mercury is continuing to increase. Therefore, there is a need to consider alternatives for the long-term management of excess mercury. This paper is intended to describe the use of a systematic method for comparing options for the retirement of excess mercury, based on an application of the analytic hierarchy process (AHP).

* Corresponding author. Tel.: +1-513-569-7673; fax: +1-513-569-7620.
E-mail address: randall.paul@epa.gov (P. Randall).

0301-4797/$ - see front matter Published by Elsevier Ltd.
doi:10.1016/j.jenvman.2004.01.004

2. Approach

The approach chosen for the present work is AHP as embodied in the Expert Choice software.[1] AHP is a highly regarded and widely used multi-criteria decision-making tool (Saaty, 2000). AHP engages decision-makers in breaking down a decision into smaller parts, proceeding from the goal to criteria to sub-criteria down to the alternative courses of action. Decision-makers then make simple pairwise comparison judgments throughout the hierarchy to arrive at overall priorities for the alternatives. AHP is a methodology based on relative instead of absolute ratings. AHP gives a structure and a mathematical base upon which many problem domains can be modeled. The decision problem may involve social, political, technical, and economic factors. AHP helps people cope with the intuitive, the rational and the irrational, and with risk and uncertainty in complex situations. It can be used to: (a) predict likely outcomes; (b) plan projected and desired futures; (c) facilitate group decision making; (d) exercise control over changes in the decision making system; (e) allocate resources; (f) select alternatives; and (g) perform cost/benefit comparisons.

3. Sources of information

The principal sources of information that were consulted to obtain data for this study are as follows.

3.1. Canadian study

A draft report (Senes, 2001) was prepared for Environment Canada on the development of retirement and long-term storage options for mercury. The report provides comprehensive identification of the range of technologies that are potentially available for mercury storage or retirement, together with a wealth of references.

3.2. Mercury management environmental impact statement

The Defense Logistics Agency (DLA) has prepared a draft Mercury Management Environmental Impact Statement (MMEIS). The Draft MMEIS analyzes three mercury management alternatives: (1) consolidate storage of the mercury stockpile at one site, (2) sale of the stockpile and (3) no-action, i.e. leave the mercury at the existing storage locations. The DLA's preferred alternative is consolidated storage. To the extent that this information is non-proprietary, it has been used in the present study. In addition, the MMEIS project has assembled a long list of references on mercury treatment. The Draft MMEIS will help decide the best way to manage the elemental mercury in the National Defense Stockpile over the long term (DLA, 2003).

3.3. Mercury workshops

EPA has prepared the proceedings of the mercury workshop that was held in March 2000 in Baltimore, Maryland (US EPA, 2001). This workshop covered: (a) the state of the science of treatment options for mercury waste; and (b) the state of the science of disposal options for mercury waste, such as landfill disposal, sub-seabed emplacement, stabilization, and surface and deep geological repositories for mercury waste storage.

In May 2002, EPA's National Risk Management Research Laboratory (NRMRL) sponsored a conference regarding issues related to the long-term management of mercury. This conference focused on the policies, technologies and techniques to address environmentally sound management and treatment of surplus mercury supplies and stockpiles, and mercury-bearing wastes.

3.4. Other US EPA and US DOE activities

For several years, both EPA and DOE have been evaluating the performance and feasibility of mercury treatment technologies. DOE has published various Innovative Technology Summary Reports that evaluate the treatment technologies applicable to mercury containing mixed wastes (i.e. wastes that are both hazardous and radioactive). The reports include environmental performance testing, cost information, and other operations information. (US DOE, 1999a–e)

In addition, EPA has conducted performance testing of mercury-containing wastes processed by various treatment technologies (US EPA, 2002a,b). Performance testing in these studies has involved both comprehensive analytical testing and standard toxicity characteristic leaching procedure (TCLP) tests.

4. Limitation of Scope

The resources available for this study were limited, therefore certain ground rules and simplifications were developed:

- Industry-specific technologies were excluded on the grounds that they can only manage a small fraction of the total mercury problem and in any case should be regarded as an integral part of that specific industry's waste management practices.
- The study focuses on options for retirement of surplus bulk elemental mercury on the grounds that: (a) elemental mercury addresses a large fraction of the problem; and (b) that it would provide an adequate

[1] Information on the Expert Choice software can be found at www.expertchoice.com

demonstration of the decision-making technique that can readily be expanded in the future. For example, the treatment of wastewater streams was excluded.
- The chemical treatment options are limited and are chosen to be representative of major classes of treatment options, such as metal amalgams, sulfides, or selenides. The choice was to some extent driven by available information. If the decision analysis favors any one class of options, then in principal it would be possible later to focus on individual technologies within that class and perform a further decision analysis to choose between individual technologies.
- Only technologies that can in principal treat contaminated media as well as elemental mercury were considered. This compensates to some extent for the decision to focus on elemental mercury.
- Retorting was excluded as merely being a well-established prior step for producing elemental mercury, some of which may end up in the pool of surplus mercury.
- Deep-sea disposal was excluded because obtaining the necessary modifications to international laws and treaties was regarded as too onerous a task.
- Storage in pipelines was excluded because the project team could not find information about this option.

As a result of the above-described ground rules and simplifications, two types of treatment technologies were evaluated: sulfide/amalgamation (S/A) techniques and the mercury selenide treatment process. The S/A techniques were represented by: (a) DeHg® amalgamation; (b) the sulfur Polymer Solidification/Stabilization (SPSS) process; and (c) the Permafix sulfide process. These were grouped as a single class because they have very similar characteristics when compared against the criteria defined by the team (comprised of EPA and SAIC staff) and modeled in Expert Choice. Therefore, only these two general types of treatment technologies were evaluated. These were combined with four disposal options: (a) disposal in a RCRA-permitted landfill; (b) disposal in a RCRA-permitted monofill; (c) disposal in an engineered belowground structure; and (d) disposal in a mined cavity. In addition, there are three storage options for elemental mercury: (a) storage in an aboveground RCRA-permitted facility; (b) storage in a hardened RCRA-permitted structure; and (c) storage in a mined cavity. Altogether, 11 options were chosen for examination with the multi-criteria decision-making tool:

- Storage of bulk elemental mercury in a standard RCRA-permitted storage building
- Storage of bulk elemental mercury in a hardened RCRA-permitted storage structure
- Storage of bulk elemental mercury in a mined cavity
- Stabilization/amalgamation followed by disposal in a RCRA-permitted landfill
- Stabilization/amalgamation followed by disposal in a RCRA-permitted monofill
- Stabilization/amalgamation followed by disposal in an earth-mounded concrete bunker
- Stabilization/amalgamation followed by disposal in a mined cavity
- Selenide treatment followed by disposal in a RCRA-permitted landfill
- Selenide treatment followed by disposal in a RCRA-permitted monofill
- Selenide treatment followed by disposal in an earth-mounded concrete bunker
- Selenide treatment followed by disposal in a mined cavity

Several of the more critical assumptions made in compiling these options include the following:

1. The project team considered storage to be temporary. As a result, costs were considered as those associated with storage itself (e.g. initial costs and operating costs), as well as projected costs for subsequent treatment and disposal when storage is terminated. As is demonstrated in the sensitivity analyses in Table 1, this is an assumption that has an important effect on the ranking of the storage options.
2. Storage, treatment, or disposal of the mercury was assumed to require RCRA-permitting. There is uncertainty as to whether local and federal environmental authorities would require such permitting for all management steps; this is a conservative assumption.
3. No distinction is made between individual stabilization and amalgamation technologies. As a result, the model is intended to identify the performance of this management technique relative to other options rather than assessing the performance of individual treatment technologies.

5. Goals, criteria and intensities

Expert Choice requires the definition of a goal, criteria, and intensities. The goal in this case is simple, namely to 'Select the best alternatives for mercury retirement'. The team developed two first-level criteria, benefits and costs. Initially, equal weights were assigned to them. This is a simple example of the pairwise comparison that is performed at every level in the hierarchy of criteria developed as input to Expert Choice.

Under costs, two-second level criteria were developed, implementation costs and operating costs. For each retirement option, the team then asked, whether the implementing costs would be low, medium, or high, and whether the operating costs would be low, medium, or high. These assignments of low, medium, or high are examples of intensities.

Table 1
Summary of results for 11 evaluated alternatives

Alternative	Ranking (as fraction of 1000[a])					
	Overall		Non-costs only		Costs only[b]	
	Score	Rank	Score	Rank	Score	Rank
Stabilization/amalgamation followed by disposal in a RCRA-permitted landfill	137	1	99	5	217	1
Selenide treatment followed by disposal in a RCRA-permitted landfill	123	2	66	9	217	1
Storage of elemental mercury in a standard RCRA-permitted storage building	110	3	152	2	126	5
Stabilization/amalgamation followed by disposal in a RCRA-permitted monofill	103	4	92	7	135	3
Storage of elemental mercury in a hardened RCRA-permitted storage structure	95	5	173	1	44	6
Selenide treatment followed by disposal in a RCRA-permitted monofill	94	6	74	8	135	3
Storage in a mine	81	7	140	3	44	6
Stabilization/amalgamation followed by disposal in an earth-mounded concrete bunker	70	8	108	4	42	8
Stabilization/amalgamation followed by disposal in a mined cavity	63	9	97	6	42	8
Selenide treatment followed by disposal in an earth-mounded concrete bunker	62	10	–[c]	–[c]	–[c]	–[c]
Selenide treatment followed by disposal in a mined cavity	61	11	–[c]	–[c]	–[c]	–[c]
Number of alternatives evaluated	11	–	9	–	9	–
Total	1000	–	1000	–	1000	–
Average score (total divided by number of alternatives, either 9 or 11)	91	–	111	–	111	–

[a] Scores normalized to total 1000.
[b] Costs for storage options include both the storage costs as well as end-of-storage costs for subsequent treatment and disposal.
[c] These options were evaluated for the overall goal but were not evaluated at the lower levels of cost and non-cost items separately, due to the low score from the overall evaluation.

Six second-level criteria were developed under the heading of benefits. Some of the second-level benefits were further split into third-level criteria. The identification of these criteria was conducted in a brainstorming session using criteria identified as important by other researchers (Senes, 2001), (US DOE, 1999a–e). Intensities were then assigned to each of the lowest-level criteria. The six second-level criteria and associated sub-criteria are listed below. The criteria were ranked and compared with one another by importance during the brainstorming activity. The figures in parentheses give the weights assigned to each of the criteria and sub-criteria using the process of pairwise comparison, which is at the core of AHP. Thus, it can be seen that, of the six second-level criteria, the analysts judged that environmental performance (0.336) and risks (0.312) are the most important. At the second level, the weights add to one. At each sub-criterion level, the weights are determined independently and also add to one.

- Compliance with Current Laws and Regulations (0.045)
- Implementation Considerations (0.154)
  - Volume of waste (0.143)
  - Engineering requirements (0.857)
- Maturity of the Technology (0.047)
  - State of maturity of the treatment technology (0.500)
  - Expected reliability of the treatment technology (0.500)
- Risks (0.312)
  - Public risk (0.157)
  - Worker risk (0.594)
  - Susceptibility to terrorism/sabotage (0.249)
- Environmental Performance (0.336)
  - Discharges during treatment (0.064)
  - Degree of performance testing of the treatment technology (0.122)
  - Stability of conditions in the long term (0.544)
  - Ability to monitor (0.271)
- Public Perception (0.107)

There is great uncertainty regarding the long-term fate of mercury in a disposal environment. The above sub-criteria of environmental performance are intended to address various issues impacting long-term environmental performance. Most data reflect leaching performance rather than mercury volatilization.

As noted above, intensities were then assigned to each of these criteria and sub-criteria. For example, three intensities were assigned to the sub-criterion ‘State of maturity of the treatment technology’: (a) experience with full-scale

operation; (b) pilot treatment technology with full-scale operation of disposal option; and (c) pilot treatment technology with untested disposal. Brainstorming about the relative importance of each pair of these three intensities ('pairwise comparison') leads to the following relative ranking of the importance of these intensities: 0.731, 0.188, and 0.081, respectively. These are numerical weights that factor into the final AHP calculations.

The intensities for each criterion are then established for each alternative, based on data and information collected. The application of this portion of the AHP process is illustrated in Table 2. Table 2 identifies each of the 11 alternatives considered and a summary of the data available to develop intensities for the two cost criteria (i.e. implementation costs and operating costs). Based on this information, an intensity of low, medium or high was selected for each criterion. This process is repeated for each criterion identified above. Typically, the quality and availability of data differs greatly among the various alternatives, as illustrated in Table 2 for cost criteria.

6. Results and discussion

Table 1 summarizes the results of the base-case analysis together with variations on the results assuming that only benefits (non-costs) or only costs are important. The ranking from the base-case analysis appears in the second column ('overall') and shows that the landfill options are preferred independent of the treatment technology. The storage options rank next, followed by the treatment technologies combined with monofills, bunkers, or mined cavities.

The reasons why the landfill options are preferred become apparent when costs are considered. The third column of results shows the rankings if only cost is taken into account. The landfill options are cheapest and this clearly outweighs the relatively unfavorable rankings that result from a focus on the benefits. However, if the costs are not an important factor, then the three storage options occupy the first three places in the 'non-costs only' ranking.

The last column of Table 1 shows unfavorable rankings for the operating costs of the storage options. This arises for two reasons: (a) if storage continues for a long period, even relatively small per annum costs will add up; and (b) storage is not a means for permanent retirement of bulk elemental mercury and the analysts assumed that, sooner or later, a treatment and disposal technology will be adopted, which adds to the cost. This is enough to drive the storage options out of first place in the base-case rankings. However, the analysis would support continued storage for a short period (up to a few decades) followed by a permanent retirement option. This would allow time for the treatment technologies to mature.

Table 3 displays a sensitivity study for non-cost criteria only.[2] These sensitivity studies show that, if cost is not a concern, then storage in a hardened, RCRA-permitted structure performs favorably against all the criteria. By contrast, the landfill options do not perform as well, with public perception and environmental performance being among the criteria for which these options receive relatively low rankings.

The standard storage option ranks least favorably of all against risks (public, worker, and susceptibility to terrorism). Although the analysts consider that none of the options has a high risk, the fact that the standard storage option would have large quantities of elemental mercury in a non-hardened, aboveground structure suggested to the team that the risks are somewhat higher than those for other options.

The options that include selenium treatment also rank less favorably with respect to risk because they were assigned a higher worker risk than were the other retirement options due to the relatively high temperature of operation and the presence of an additional toxic substance (selenium). They also (unsurprisingly) perform relatively unfavorably with respect to technological maturity.

The last row of Table 3 shows the ratio between the scores for the alternatives that are ranked highest and lowest. Table 3 shows that, if high importance is assigned to them, compliance with laws and regulations (ratio 7.1), implementation considerations (ratio 6.8) and the maturity of the technology (ratio 5.0) are the most significant discriminators between the retirement options. By contrast, the ratio for sensitivity to risks is only 1.6. This is because the analysts concluded that none of the retirement options has a high risk and that any variations are between low and very low risk.

Finally, a limited number of analyses were performed to address uncertainties in the assignment of the retirement options to each intensity. Examples include increasing implementation costs for storage in a mine from medium to high, decreasing operating costs for storage of elemental mercury in a hardened, RCRA-permitted structure from high to low, and looking forward to when selenide treatment followed by storage in a mined cavity can be considered as a fully mature technology. Altogether, 12 such analyses were performed by changing just one intensity assignment from the base case. These analyses showed expected trends, with scores and rankings improving if a more favorable assignment was made and decreasing if a less favorable assignment was made. In no case did the score increase or decrease by more than 40% and in most cases the change was less than 10%. These analyses are only uncertainty

[2] The sensitivity studies were performed by adjusting weights so that the individual criterion receives 90% of the weighting, while the rest receive only 10% altogether while maintaining the relative weightings from the base case. The exceptions are columns 2 and 3 of the results in Table 1 where only benefits or only costs were considered, respectively.

Table 2
Developing ranking intensities for cost criteria based on available data

Alternative	Cost data	Ranking intensity for cost components
Storage of elemental mercury in a standard RCRA-permitted storage building	Implementation: ~$4 million for new structure or zero for existing structure; Operating: Requires further 'end of life' management	Implementation: low; Operating: high
Storage of elemental mercury in a hardened RCRA-permitted storage structure	Implementation: up to $10 to $20 million; Operating: requires further 'end of life' management	Implementation: medium; Operating: high
Storage in a mine	Implementation: Not known, but expected to be similar to hardened storage case; Operating: requires further 'end of life' management	Implementation: medium; Operating: high
Stabilization/amalgamation followed by disposal in a RCRA-permitted landfill	Implementation[a]: zero (existing disposal structure); Operating[a]: $0.1 to $0.6 million/1500 ton mercury	Implementation: low; Operating: low
Selenide treatment followed by disposal in a RCRA-permitted landfill	Implementation[a]: zero (existing disposal structure); Operating[a]: $0.1 to $0.6 million/1500 ton mercury	Implementation: low; Operating: low
Stabilization/amalgamation followed by disposal in a RCRA-permitted monofill	Implementation[a]: not available; new construction required; Operating[a]: similar to commercial landfill above	Implementation: medium; Operating: low
Selenide treatment followed by disposal in a RCRA-permitted monofill	Implementation[a]: not available; new construction required; Operating[a]: similar to commercial landfill above	Implementation: medium; Operating: low
Stabilization/amalgamation followed by disposal in an earth-mounded concrete bunker	Implementation[a]: unknown; as high as $240 million; Operating[a]: $1.6 to $3.2 million/1500 ton mercury	Implementation: high; Operating: medium
Selenide treatment followed by disposal in an earth-mounded concrete bunker	Implementation[a]: unknown; as high as $240 million; Operating[a]: $1.6 to $3.2 million/1500 ton mercury	Implementation: high; Operating: medium
Stabilization/amalgamation followed by disposal in a mined cavity	Implementation[a]: unknown; may be part of operating costs; Operating[a]: $30 to $90 million/1500 ton mercury	Implementation: high; Operating: medium
Selenide treatment followed by disposal in a mined cavity	Implementation[a]: unknown; may be part of operating costs; Operating[a]: $30 to $90 million/1500 ton mercury	Implementation: high; Operating: medium

Costs listed are best estimates based on management of similar materials.

[a] Costs are comprised of treatment plus disposal costs, but data for treatment costs are extremely variable and are not included. Only final management (disposal) costs are included. Therefore cost deviations among treatment options are not made.

Table 3
Sensitivity analysis of non-cost criteria

Alternative	Ranking (as fraction of 1000[a]; average score 111)													
	Non-cost baseline		Sensitivity: env. perf.		Sensitivity: risks		Sensitivity: implement		Sensitivity: public		Sensitivity: maturity		Sensitivity: Compliance	
	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank
Storage of elemental mercury in a hardened RCRA-permitted structure	173	1	176	1	142	1	172	2	197	1	226	1	263	1
Storage of elemental mercury in a standard RCRA-permitted building	152	2	173	2	87	9	259	1	52	5	224	2	261	2
Storage in a mine	140	3	145	3	101	5	168	3	193	2	223	3	78	3
Stabilization/amalgamation followed by disposal in an earth-mounded concrete bunker	108	4	94	5	132	2	57	5	190	3	52	6	74	4
Stabilization/amalgamation followed by disposal in a RCRA- permitted landfill	99	5	71	8	131	3	146	4	46	6	67	4	73	5
Stabilization/amalgamation followed by disposal in a mined cavity	97	6	110	4	95	6	38	9	189	4	51	7	37	9
Stabilization/amalgamation followed by disposal in a RCRA-permitted monofill	92	7	92	6	130	4	55	6	46	6	66	5	73	5
Selenide treatment followed by disposal in a RCRA-permitted monofill	74	8	81	7	92	7	53	7	44	8	46	8	71	7
Selenide treatment followed by disposal in a RCRA-permitted landfill	66	9	58	9	91	8	52	8	43	9	45	9	70	8
Total	1000	–	1000	–	1000	–	1000	–	1000	–	1000	–	1000	–
Range: highest to lowest alternative	2.6 times		3.0 times		1.6 times		6.8 times		4.6 times		5.0 times		7.1 times	

In the sensitivity analysis for each criterion, the importance of the criterion is set at 90%. The five other criteria comprise the remaining 10%, proportional to their original contributions. Two options were not evaluated for the sensitivity analysis: selenide treatment followed by disposal in a mined cavity, and selenide treatment followed by disposal in an earth-mounded concrete bunker. This is because of the low score from the overall evaluation and because the version of Expert Choice used for this analysis only allowed the use of nine alternatives for the sensitivity analysis.

[a] Scores normalized to total 1000.

analyses in a very limited sense and only one parameter at a time could be varied. A future study could potentially perform a true uncertainty analysis using Monte Carlo techniques.

7. Conclusions

A limited scope decision analysis has been performed to compare alternatives for the long-term management of surplus mercury. The analysis has demonstrated that such a study can provide useful insights for decision-makers. Some observations and conclusions are as follows:

- The analysis was useful in identifying the advantages and disadvantages of applying AHP to the decision-making process for this complex problem. AHP is a decision support tool and should not be the only input to a decision-making process. Results provided environmental decision makers with useful pieces of information among many others, such as budget constraints, political pressures, legal mandates, stakeholder opinions, and so forth.
- An important consideration in using AHP is that it forces the prioritization of various factors influencing a decision. For example, the principal criteria identified above (e.g. costs, environmental performance, compliance with current laws and regulations) were all deemed important enough to be considered in decision-making. Ideally, an alternative will perform well in all these areas. As shown in Tables 1 and 3, different results were obtained as the emphasis placed on each criterion is changed.
- Considering non-cost criteria only, the three storage options rank most favorably. If both cost and other criteria are considered, then landfill options are preferred, because they are the least expensive ones. In this analysis, costs were considered somewhat separately from other criteria. There is a tendency in some environmental management problems to identify the best alternative regardless of costs (i.e. no importance is given to cost). For this reason, cost considerations were made flexible as shown in Table 1, where cost criteria were evaluated as completely unimportant, equally important, or overwhelmingly important (i.e. importance of 0, 50 and 100% relative to the other non-cost criteria). In this manner, the results are useful in identifying a cost effective alternative particularly if the alternatives score similarly on non-cost criteria. The three storage alternatives occupied the three highest rankings when costs are not considered although one of these alternatives (storage in a standard RCRA-permitted storage building) has a much lower cost than the other two alternatives.
- The analysis supports continued storage for a short period (up to a few decades) followed by permanent retirement when treatment technologies have matured.

Future work could include:

- Involving additional experts or stakeholders in the process of assigning weights to the various criteria. This would ensure that a wider range of expertise and interests is incorporated into the analysis. For example, working groups involving a cross-section of EPA offices would provide additional perspectives. Other examples would involve the inclusion of other Federal agencies, States, non-governmental organizations, foreign governments, industry, and academia. Such participation could be performed in stages. As discussed above, differences in the importance of the criteria relative to one another can change the results.
- The alternatives considered in this paper were limited to the long-term management of elemental mercury. Additional alternatives could be considered for mercury-containing wastes.
- Additional Expert Choice analyses could be conducted in which certain alternatives are optimized. For example, within the general alternative of stabilization/amalgamation treatment followed by landfill disposal are potential sub-alternatives addressing individual treatment technologies or landfill locations.
- Revisit the available information periodically to determine if changes in criteria, or changes in intensities, are required. For example, some candidate criteria were not considered because insufficient information was available. One example is volatilization of mercury during long-term management. Very little data are available at this time to adequately address this as a possible criterion.
- Consider performing a formal uncertainty analysis utilizing Monte Carlo-based techniques.

Acknowledgements

The US Environmental Protection Agency through its Office of Research and Development funded the research described here under contract number EPA-CI-4, GSA Environmental Services Project GS-10F-0076J. It has not been subjected to the Agency's review and therefore does not necessarily reflect the views of the Agency, and no official endorsement should be inferred. This paper is a summary of a more detailed EPA report titled 'Preliminary Analysis of Alternatives for the Long Term Management of Excess Mercury'.

References

Defense Logistics Agency (2003). Draft Mercury Management Environmental Impact Statement, http://mercuryeis.com/eis.htm#draft%20-Mercury%20 maangement%20 eis

Saaty, T.L., 2000. Decision Making for Leaders: The Analytic Hierarchy Process for Decisions in a Complex World: 1999/2000 Edition, 3rd Revision Edition (December 1999), RWS Publications, 315 p.

Senes (2001). Senes Consultants Ltd. The Development of Retirement and Long Term Storage Options of Mercury. Draft Final Report Prepared for National Office of Pollution Prevention, Environment Canada, June 2001.

US DOE (1999a). Mercury Contamination—Amalgamate (Contract with NFS and ADA) Demonstration of DeHg^{SM} Process. Mixed Waste Focus Area. Innovative Technology Summary Report, DOE/EM-0471. Prepared for Office of Environmental Management, Office of Science and Technology. http://apps.em.doe.gov/ost/itsrtmwfa.html

US DOE (1999b). Mercury Contamination—Amalgamate (Contract with NFS and ADA) Stabilize Elemental Mercury Wastes. Mixed Waste Focus Area. Innovative Technology Summary Report, DOE/EM-0472. Prepared for Office of Environmental Management, Office of Science and Technology. http://apps.em.doe.gov/ost/itsrtmwfa.html

US DOE (1999c). Demonstration of ATG Process for Stabilizing Mercury (<260 ppm) Contaminated Mixed Waste. Mixed Waste Focus Area. Innovative Technology Summary Report, DOE/EM-0479. Prepared for Office of Environmental Management, Office of Science and Technology. http://apps.em.doe.gov/ost/itsrtmwfa.html

US DOE (1999d). Demonstration of GTS Duratek Process for Stabilizing Mercury (<260 ppm) Contaminated Mixed Waste. Mixed Waste Focus Area. Innovative Technology Summary Report, DOE/EM-0487. Prepared for Office of Environmental Management, Office of Science and Technology. http://apps.em.doe.gov/ost/itsrtmwfa.html

US DOE (1999e). Demonstration of NFS DeHg Process for Stabilizing Mercury (<260 ppm) Contaminated Mixed Waste. Mixed Waste Focus Area. Innovative Technology Summary Report, DOE/EM-0468. Prepared for Office of Environmental Management, Office of Science and Technology. http://apps.em.doe.gov/ost/itsrtmwfa.html

US EPA (2001). Proceedings and Summary Report. Workshop on Mercury in Products, Processes, Waste and the Environment: Eliminating, Reducing and Managing Risks from Non-Combustion Sources. March 22–23, 2000. Baltimore. EPA/625/R-00/014.

US EPA (2002a). Technical Background Document: Mercury Wastes. Evaluation of Treatment of Mercury Surrogate Waste. Final Report.

US EPA (2002b). Technical Background Document: Mercury Wastes. Evaluation of Treatment of Bulk Elemental Mercury. Final Report.

SIMULATION AND OPTIMIZATION OF LARGE SCALE SUBSURFACE ENVIRONMENTAL IMPACTS; INVESTIGATIONS, REMEDIAL DESIGN AND LONG TERM MONITORING

Larry M. Deschaine, PE
Chalmers University of Technology
Department of Physical Resources
Complex Systems Group
Goteborg, Sweden

ldeschaine@alum.mit.edu

Journal of Mathematical Machines and Systems
Kiev, No 3, 4. 2003. Pages 201-218
(Extended Chalmers Version)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMNS AND ABBREVIATIONS

3-D	three-dimensional
AHP	Analytical Hierarchy Process
CoC	contaminant of concern
CV	coefficient of variation
DMSO	U.S. Defense Modelling and Simulation Office
DNAPL	dense non-aqueous phase liquid
DoD	U.S. Department of Defense
DOE	U.S. Department of Energy
ESTCP	Environmental Security Technology Certification Program
GAO	U.S. General Accounting Office
GSLIB	Geostatistical Software Library
GUI	Graphical User Interface
HORDA	Heuristic Optimal Remedial Design Algorithm
LGO©	Lipchitz Global Optimization
LNAPL	light non-aqueous phase liquid
MCL	Maximum Contaminant Level
MODOFC	MODFLOW Optimal Flow Control
SERDP	Strategic Environmental Research and Development Program
SGSIM	SGSIM aquifer conductivity generation package (GSLIB)
UXO	unexploded ordnance

SIMULATION AND OPTIMIZATION OF LARGE SCALE SUBSURFACE ENVIRONMENTAL IMPACTS; INVESTIGATIONS, REMEDIAL DESIGN AND LONG TERM MONITORING

ABSTRACT

The global impact to human health and the environment from large scale chemical / radionuclide releases is well documented. Examples are the wide spread release of radionuclides from the Chernobyl nuclear reactors, the mobilization of arsenic in Bangladesh, the formation of Environmental Protection Agencies in the United States, Canada and Europe, and the like. The fiscal costs of addressing and remediating these issues on a global scale are astronomical, but then so are the fiscal and human health costs of ignoring them.

An integrated methodology for optimizing the response(s) to these issues is needed. This work addresses development of optimal policy design for large scale, complex, environmental issues. It discusses the development, capabilities, and application of a hybrid system of algorithms that optimizes the environmental response. It is important to note that “optimization” does not singularly refer to cost minimization, but to the effective and efficient balance of cost, performance, risk, management, and societal priorities along with uncertainty analysis. This tool integrates all of these elements into a single decision framework. It provides a

consistent approach to designing optimal solutions that are tractable, traceable, and defensible.

The system is modular and scalable. It can be applied either as individual components or in total. By developing the approach in a complex systems framework, a solution methodology represents a significant improvement over the non-optimal "trial and error" approach to environmental response(s).

Subsurface environmental processes are represented by linear and non-linear, elliptic and parabolic equations. The state equations solved using numerical methods include multi-phase flow (water, soil gas, non-aqueous phase liquid [NAPL]), and multicomponent transport (radionuclides, heavy metals, volatile organics, explosives, etc.). Genetic programming is used to generate the simulators either when simulation models do not exist, or to extend the accuracy of them. The uncertainty and sparse nature of information in earth science simulations necessitate stochastic representations. For discussion purposes, the solution to these site-wide challenges is divided into three sub-components; plume finding, long term monitoring, and site-wide remediation.

Plume finding is the optimal estimation of the plume fringe(s) at a specified time. It is optimized by fusing geo-stochastic flow and transport simulations with the information content of data using a Kalman filter. The result is an optimal monitoring sensor network; the decision variable is location(s) of sensor in three dimensions. Long term monitoring extends this approach concept, and integrates the spatial-time

correlations to optimize the decision variables of where to sample and when to sample over the project life cycle. Optimization of location and timing of samples to meet the desired accuracy of temporal plume movement is accomplished using enumeration or genetic algorithms.

The remediation optimization solves the multi-component, multiphase system of equations and incorporates constraints on life-cycle costs, maximum annual costs, maximum allowable annual discharge (for assessing the monitored natural attenuation solution) and constraints on where remedial system component(s) can be located, including management overrides to force certain solutions to be chosen are incorporated for solution design. It uses a suite of optimization techniques, including the outer approximation method, Lipchitz Global Optimization (LGO©), genetic algorithms, and the like. The automated optimal remedial design algorithm requires a stable simulator be available for the simulated process. This is commonly the case for all above specifications sans true three-dimensional multiphase flow. Much work is currently being conducted in the industry to develop stable three-dimensional (3-D), three-phase simulators. If needed, an interim heuristic algorithm is available to get close to optimal for these conditions.

This system process provides the full capability to optimize multi-source, multiphase, and multicomponent sites. The results of applying just components of these algorithms have produced predicted savings of as much as $90,000,000(US),

when compared to alternative solutions. Investment in a pilot program to test the model saved 100% of the $20,000,000 predicted for the smaller test implementation.

1.0 INTRODUCTION

The cost to clean up environmentally impacted sites is high, both for businesses and for taxpayers. For many locations, environmental cleanup is not the core function of the facility. While a very important activity, the cleanup efforts remove dollars from operations budgets. This expense restricts productive capability in the short term and reduces investment core mission improvement in the long term. The methods of remediating aquifers consist of primarily in-situ treatment in which the contaminants are treated in place (via natural attenuation and/or engineered solutions), ex-situ treatment in which the contaminants are removed from the aquifer, and institutional controls through which the sources of contaminants are managed rather than treated. These costs are significant. The U.S. Department of Energy (DOE) estimated its environmental remediation costs at more than $150 billion (DOE 1997). About 15 years ago, the General Accounting Office (GAO) identified this issue by estimating the cost of remediating just the sites on the Superfund list at $26 billion (GAO 1993). These are only two examples of remediation cost estimates, and do not include all of the sites, such as gasoline stations, dry cleaners, manufacturing facilities, landfills, and the like, requiring remediation. In response, research into the optimization of site investigation programs and remediation designs has been intense. Optimization of solutions for these sites is both mathematically challenging and complex. Aquifers can vary from simple, single-layer geologic units to multiunit perched systems with variably saturated flow. Multiphase impacts may result,

including dense, non-aqueous phase liquids (DNAPLs) such as chlorinated solvents; light, non-aqueous phase liquids (LNAPLs) such as petroleum products; dissolved-phase contamination in which residual material migration is driven by groundwater flow; vapor-phase contamination; and residual soil contamination. Plume delineation is expensive. Solution options for contaminant migration management are many, ranging from hydraulic pump-and-treat or air sparging systems to optimization of in situ biogeochemical oxidation-reduction potential (Redox) zone design of monitored natural attenuation or engineered systems.

Since the mid-1980s to the early 1990s, optimization of groundwater remediation design for groundwater extraction/reinjection systems has resulted in significant cost reductions. Formal optimization of the remedial design problem has been examined by many researchers who directly couple saturated groundwater flow and transport simulation models with optimization algorithms (Ahlfeld 1986, Ahlfeld and Heidari 1994, and Wagner 1995). In the early 1990s, work was started on an algorithm to optimize remedial designs by integrating advanced multiphase, multicomponent flow and transport simulators, formal optimization algorithms, heuristic optimization algorithms, and economic functions into a multisource, multi-plume, site-wide risk-based optimization tool (Deschaine 1992, Karatzas and Pinder 1993, and Karatzas and Pinder 1996). Along the way the top three deployments of these tools have saved $100M, $90M, and $72.6M(US).

2.0 ALGORITHM DESCRIPTION

This formal site wide environmental impact remedial design optimization system is a system of algorithms that fills the gap in available aquifer optimization tools. It forms a bridge between the "trial and error" and the saturated groundwater pump-and-treat-only approaches to remedial system design by providing a formal, structured approach with which to find a least-cost solution, even when the costs and the flow and transport aquifer properties are uncertain. Figure 2-1 provides an overview of the tool.

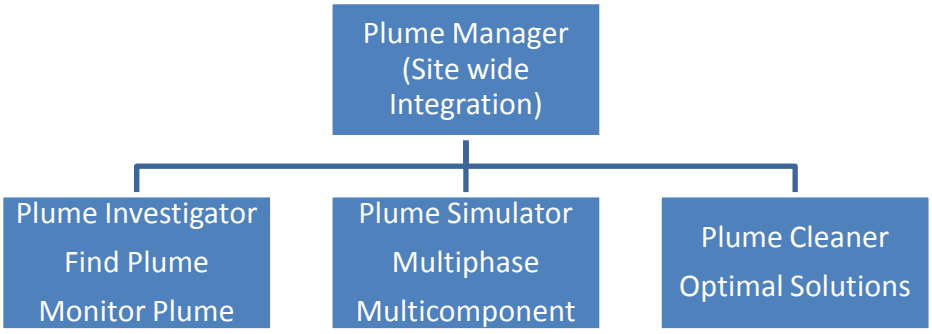

Figure 2-1. Overview of the Site-wide Optimization Tool

- The Plume Manager module controls the overall solution. It integrates information on cost functions and spend-out constraints with environmental simulations to develop the best solution for the challenge at hand for a site-wide integrated solution.

- The Plume Investigator module integrates the flow and transport simulations with information-content modeling (deterministic or stochastic) to optimally design sample programs that define the extent of the impacts and monitor them long term.
- The Plume Simulation module consists of a large number of subsurface flow and transport simulators used to evaluate baseline conditions as well as evaluate potential remedial responses.
- The Plume Cleaner module integrates optimization algorithms with the Plume Simulation algorithms to design least-cost solutions to various remediation options, including monitored natural attenuation and active remediation.
- Information needed to maximize the effective deployment of this tool include the following:
    - A mathematically correct statement of the flow properties of the aquifer. This model can be deterministic or stochastic.
    - A mathematically correct statement of the transport properties of the aquifer, including the biogeochemical processes affecting them in space and time. This model can be deterministic or stochastic.

  - An annual and project life cycle cost function that represents both the capital and operational costs of the various remedial options under consideration.

- The constraints on the solution include the maximum desired annual costs, the limits on contaminant discharge or concentrations at point(s) of compliance, and the constraints as to where the remedial action can be physically located and where it cannot. Management overrides—specifying or prohibiting a remedial option at a specific source—are important and are accommodated by the tool.

The integrated optimization algorithm reads the above information and provides feasible and optimal or near-optimal solutions when all things are considered. The main technologies consist of Lipchitz Global Optimization (LGO©), genetic and other evolutionary computation algorithms, Tabu search algorithms, and Monte Carlo and Latin Hypercube simulation algorithms.

3.0 SITE-WIDE RISK-BASED OPTIMAL MANAGEMENT MODULE

The site-wide risk-based optimizer integrates the physics of flow and transport with the economics of project management. Acceptance is gained using the verification, validation, accreditation and credibility guidance of the U. S. Defense Modeling and Simulation Office (DMSO). The optimizer uses the subsurface flow and transport models as subroutines, so it is physics model independent. The most suitable model can be chosen for the site / question, expanding the flexibility, adaptability, and solution correctness of the optimizer. This approach leverages advances in subsurface simulation code, which have been substantial over the tool's 29-year development cycle, beginning with a 114 site state-wide optimization program (Deschaine *et al.*, 1985) depicted in Figure 3-1.

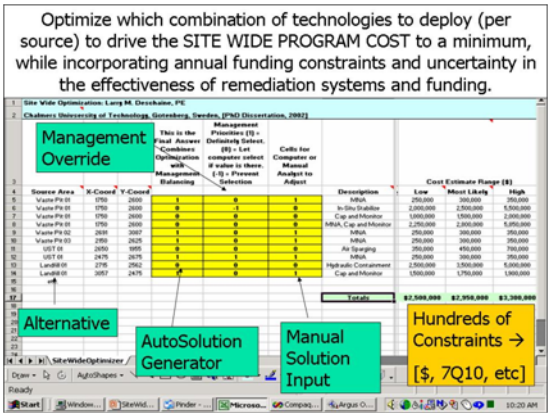

Figure 3-1. Site-wide Optimization Tool's Spreadsheet GUI

The stochastic constraints (discharge limits for each contaminants of concern [CoC], annual funding, etc.) are located to the right of the figure and represent literally hundreds of constraints on the optimal solution.

The algorithm links the science and economics—the feasible solutions with the business, regulatory, physical and social constraints. Specifically, these business algorithms provide the functionality to optimize the integrated site-wide remediation decision. Because each algorithm must, by necessity, be customized for a specific site, only a general description of the algorithm is possible. Such a description is provided below.

1. Using a suitable subsurface flow and transport model, develop a mathematically correct statement of the source-specific flow and transport system as is and predict future impacts if the contaminant is allowed to migrate unabated. This output can be the result of deterministic simulations (i.e., simulate one “best” representative aquifer) or of stochastic simulation of equally likely aquifers (i.e., use GSLIB to simulate the realizations of equally likely aquifers). Store these results.

2. Screen remedial options using standard European or U.S. Environmental Protection Agency’s guidance, for example, and select the most likely options per source area if multiple sources are present. Perform value engineering on new or existing solutions, as appropriate. Assess whether the treatment

effectiveness is known (deterministic) or estimated. If estimated, build an option-specific stochastic function. Store these results.

3. Assemble the cost constraints, which can be per source area, per year, or per life cycle. Assess whether the cost constraints are known or estimated. If estimated, build an option-specific stochastic function. Store these results.

4. Assemble the environmental and physical constraints. These constraints can be the point of compliance, cleanup level, discharge mass per year, or treatment system component location and can be per source area, per year, or per life cycle. Assess whether the environmental constraints are known or estimated. If estimated, build an option-specific stochastic function. Store these results.

5. Assemble the management override constraints. These constraints are typically to force a certain remediation at one location or prohibit one at another and often relate to land use. For example, force capping on a land parcel for sale or prohibiting pumping and treating using vertical wells on an airport runway.

The current implementation of the site-wide optimizer uses a combination of optimization techniques (discussed below) and Monte Carlo and Latin Hypercube simulation to optimize the solution under uncertainty. It designs and answers the following questions:

1. What is the least-cost solution to the site remediation? How does this solution vary if annual funding is constrained or changed?
2. How will the solution change if cleanup levels are relaxed, points of compliance are changed, or time to clean up varies? How will it change if various remedial alternatives are selected or deselected?
3. When multiple projects or opportunities for savings are present, which ones should be implemented first?
4. Given a desire to be 95% confident that a site-wide solution strategy will be protective of the environment and all constraints will be met, which course of action is best? How will this change if desired confidence is 90% or 50%?

A simplified version of this business process optimization algorithm was used to project savings of $90M over a 5-year period [Deschaine et al. 1998b]. An extended version of this algorithm was developed for energy systems and technologies research and development program design for DOE's Vision 21 Energy Development Program [Deschaine et al. 2001b].

4.0 PSUEDO-CODE FOR THE SITE-WIDE OPTIMIZER

The generalized pseudo-code for the optimization tool that incorporates the above is as follows:

- Minimize Total Economic and Social Cost of Plume Presence
    - Total Cost = f(investigation, remediation, monitoring reporting, administration, increase in health services, loss of recreational and other use, loss of natural resources, etc.).
- Subject to:
- The risks to human health and the environment are within acceptable levels within the specified time frame
    - ($C_{i,t^*}<C_{i,max}$), Concentrations $C_{i,t}$ for all spatial locations (i) within exposure time (t).
- The mass flux into receptors are less than assimilative capacity (natural attenuation) of the receiving system, such as a receiving surface water body or aquifer.
    - ($M_{i,t^*}<M_{i,max}$), Mass fluxes $M_{i,t}$ for all spatial locations (i) within exposure time (t).

- The annual costs are below a specific value, with certainty "p", for all time periods (t) for example 95% certain that costs will be below budget in any given year.
    - (95^{th} percentile of $D_{t^*}<D_{t,max}$), Dollar requirements for all annual times (t).
- Expert consensus expectation of success greater than minimum value, with probability P_{se} that success not less than S_{min}. This constraint balances the "risky" but potentially high reward options with the tried and true methods, similar to balancing a portfolio of investments while minimizing the downside risk, for example constrain 25^{th} percentile of selected portfolio of projects to some absolute downside value. It captures the uncertainty of the experts to know a priory the success of the various site management policies.
    - ($[P_{se}] > P_{semin1}$ and 25^{th} percentile $P_{se} > P_{semin2}$)

Modifying a well-known and widely used decision support algorithm, the analytical hierarchy process to accept stochastic inputs, captures the expert's uncertainty. The Analytical Hierarchy Process (AHP) algorithm is presented in (Saaty 1996). Other methods reviewed, and viable options to this approach to represent human behavior include fuzzy logic, non-monotonic reasoning, expert systems, reinforcement learning and the like. The specific human behavior method is flexible in this tool, so long as the results pass the DSMO requirements for validation of human behavior in simulations.

4.1 LIFE-CYCLE COST FUNCTION

The cost of the remediation project can essentially be broken down into two components: the capital cost and the variable or annual cost. In the cost function discussed above, there are costs that the polluter pays, and costs borne by society. The cost function is developed by an assembly of stakeholders who decide or negotiate which cost components are included in the analysis and optimization of the response policy. This may not be a trivial exercise in many cases. This work allows for the flexible and comprehensive complete or partial analysis of cost functions.

The capital portion of a response policy consists of the cost of the infrastructure, such as the treatment building(s), permits, ancillary components, and the like. The variable cost portion (cost of operations and maintenance over the project life cycle) consists of items such as the recovery wells and pump systems; the pipe and trenching; and the cost of the cleanup operation, including system operation, maintenance, and environmental sampling. The total project life-cycle cost, which consists of both components, is the cost function that is optimized within the annual funding and other project constraints discussed above. The “balance” question is what capital improvements to plume management to make when, recognizing the action and inaction have economic and social costs when the total cost of plume existence is considered. The total cost of ownership is developed in the value engineering phase. In this phase, the stakeholders are assembled and each aspect of the problem and potential solution is quantified to the best of the group’s ability – recognizing and

cataloging uncertain aspects. Solving for the optimal solution, based on this quantification and acknowledged uncertainties, is discussed below.

4.2 VALUE ENGINEERING

Value engineering is used to specifically define the cost function(s) of the various site management options, including scope of investigation, remediation alternatives and long term monitoring. This consists of essential meetings with the various stakeholders to develop the goals of the management policy. The stakeholder group(s) heuristically optimize, to the extent practical, the decision components. This step is the precursor to the formal optimization, and supplies the accreditation and facilitates acceptance of any optimal solutions developed. The investigation and remedial options are evaluated, and using a combination of costs in the project databases or actual operation costs, a very detailed cost matrix is developed. The cost function is developed such that changes in one aspect of a solution policy under consideration propagate throughout the integrated site-wide optimization system.

5.0 PLUME SIMULATION MODULE

Many groundwater flow and transport codes currently exist that employ either the finite difference or finite element solution techniques. They can solve either single- or multi-phase flow and single- or multi-component transport. The choice to model a site using finite differences or finite elements is site-specific—it depends greatly on the ability of the technique to represent the subsurface flow system with a mathematically correct statement. Guidance is available on model selection criteria and solutions for many of the state-of-the-art models available at this time (BWXT and SAIC 2002). Examples of the modelling codes used, along with a brief summary description of their capabilities, are presented below.

5.1 FINITE ELEMENT AND FINITE VOLUME METHODS

- Princeton Transport Code (PTC) with Plume-Finding Technology: a saturated/unsaturated flow and single-component transport finite element code, efficient and robust, excels when many calls are needed to evaluate options.

- BioF&T3D and SA_MAPS: a variably saturated, multiphase, multicomponent flow and transport finite element code that includes biochemistry and dual-porosity formulations such as air sparging and groundwater/surface water interactions.

- FRAC3DVS and FEHMN: Variably saturated flow and multicomponent flow and transport finite element codes that include varying degrees of biogeochemistry and fractured/porous media. FEHM also allows for multiphase and dual-porosity/dual-permeability formulations.
- COMPFLO: Multiphase Multi Compositional finite volume code that simulates three-phase flow and multi-component transport in fractured/porous media.

Documentation reports of the above mentioned codes are provided in Table 5-1.

Table 5-1. Finite element and finite volume methods

Code Name	Reference
PTC	Babu, et al., 1987
BioF&T3D and SA_MAPS	Katyal, A., 2013
FRAC3DVS	Therrien, R., et. al., 1994
FEHMN	Zyvoloski, G. A., et. al., 1996
COMPFLO (Bio)	Unger, A.J.A, *et al.*, 1995

5.2 FINITE DIFFERENCE METHOD

- MODFLOW: a saturated flow model from which several transport and optimization codes are linked.
- MT3D and MT3DMS: single-component and multi-component transport codes linked with MODFLOW.
- MODFLOW-SURFACT/MODHMS™: A comprehensive set of groundwater and groundwater/surface water flow and transport codes.
- RT3D: biologic transport code linked with MODFLOW.
- TOUGH2v2: multiphase simulators with various transport capabilities.

Documentation reports of the above mentioned codes are provided in Table 5-1.

Table 5-2. Finite difference method

Code Name	Reference
MT3D	Zheng, C., 1990
MT3DMS	Zheng, C., 1999
MODFLOW-SURFACT/MODHMS™	HGL, 2014
RT3D	Clement, TP 1997
TOUGH2v2	Zhang, K., 2003

It is important to note that for various remediation technologies, various simulation codes may be used. No one code can always meet the needs of the analysis,

whether they be complexity of flow or transport or solution time. For example, a fully three-dimensional, multi-phase, multi-component simulator would not usually be chosen to solve a saturated regional flow system with a nonreactive contaminant. Not only would use of such a simulator be overkill, but it would also be inefficient and require much longer run times. Conversely, bioairsparging of an LNAPL or surfactant flushing of a DNAPL would not be solved with MODFLOW. It would be an inappropriate use of the simulator.

6.0 PLUME INVESTIGATION MODULE

Determining the nature and extent of a plume requires finding the fringe in three dimensions. The plume investigation module of this site-wide optimizer uses a combination of tools to accomplish this effectively and efficiently (McGrath and Pinder 1996, McGrath 1997). The plume-finding module is a program for guiding the location of sensors (i.e., soil borings, cone-penetrometer probes, or monitoring wells) in groundwater investigations of contaminated aquifers. The objective of the investigation is to provide the groundwater professional and stakeholder assistance in determining the number of samples and locations needed to delineate the boundary of a contaminant plume in a 3-D aquifer. The "boundary" of the plume is defined by the aqueous-phase contaminant concentration standard, such as a maximum contaminant level (MCL) or other risk-based standard.

The goal is to quantify the information that an existing monitoring well network provides in establishing knowledge of the location of the groundwater plume and to identify (prior to installing a sensor) the next sampling location(s) in three-dimensional space that, when sampled, minimizes the uncertainty of the plume boundary location. This will provide a prioritization of the sensor installation activity with each new proposed sampling location at the location where it will provide the maximum amount of information for solving the plume location challenge. It will also quantify the reduction in incremental knowledge gained by each new sensor; thereby providing a stopping point when the sensor network is adequate. Adequacy of plume delineation is

defined by the stakeholder specified data quality objectives. The following describes the major elements in the plume finding process:

1. Determine the plume fringe by assessing where the fringe of the CoCs isocontour(s) lines are located. Generating multiple realizations of the aquifer and plotting the expected concentrations and distributions at each node provide identification of this plume line. The fringe is added by identifying which nodes contain the iso-contour with 95% certainty, for example. This is straightforward statistical analysis.
2. Assess the reduction in the uncertainty of the knowledge of the plume location given an existing monitoring well network. This is accomplished by combining the physical subsurface flow and transport model and the Kalman filtering approach (*Note: The Kalman filter is discussed below in the long term monitoring section. The major difference is that here we specify the time and solve for the best location to have monitoring information. The long term monitoring solves for the best space and time for the information being collected*.)
3. Assess the best location(s) to add new sensors (monitoring wells, soil gas data collection, etc.) to reduce the uncertainty in plume location definition by the maximal amount.

The procedure below is used to implement the plume-finding algorithm. This involves construction of a preliminary subsurface flow and transport model as

discussed above. The initial model can be quite simple and does not require detailed site knowledge to be effective. The procedure consists of four steps:

1. Apply the plume-finding technology to generate a rank-ordered list of sample locations of monitoring points/wells.
2. Collect the data and add this information to the observation database.
3. Update the plume finder statistical information.
4. Repeat until the confidence in the knowledge of the plume location meets project requirements.

By maximizing the benefit of each sampling point, optimal sampling programs (maximum information for the least cost) can be designed to best meet project goals, and the design can be defended. An optimal network can be designed by selecting which location(s) for new monitoring well(s) reduce the volume under the uncertainty surface by the greatest amount.

The plume finding example, shown in Figure 6-1, is based on field data from the DoE Pantex Plant. The analysis was conducted using PTC as the subsurface simulator, the SGSIM aquifer conductivity realization generation package from GSLIB, and the Kalman Filter as implemented by (McGrath and Pinder 1996). A general description of the problem follows.

The goal of the study was to assess the placement of between 6 and 12 new monitoring wells to assess the effectiveness of the existing monitoring well network on delineating the plume fringe. Approximately a dozen monitoring wells exist at the

site. The aquifer of concern is 500 feet below the ground surface. The cost of each well is $150,000 to install, with additional life cycle costs resulting from the periodic sampling and reporting of results. An existing site model was available, as well as a regional variogram. The first step was to assess the maximum uncertainty of the plume without any monitoring wells installed. In Figure 6-1, the direction of flow is along the longer x-axis (0 to 24,000 feet).

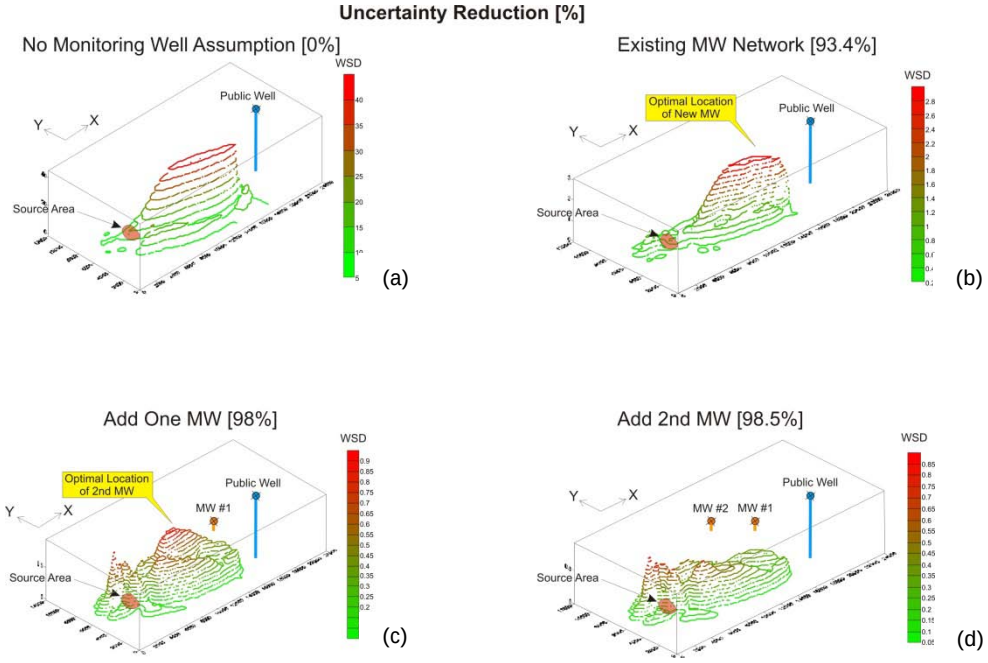

Figure 6-1. Summary of Plume Finder Analysis Value of Additional Wells

[Results of 4000 flow and transport simulations; linkage with the Kalman Filter and Observed monitoring well data. Note: vertical scale adjusted to see results]

The y-axis extends from 0 to 12,000 feet. The worth of sample data values is plotted on the z-axis. The highest value represents the location where the most value is obtained by collecting information at the peak of that surface. Without a monitoring well network, the maximum uncertainty was determined to be 40 “worth of sample data” units. This condition is shown on Figure 6-1a.

The next part of the analysis was to assess the robustness of the existing monitoring well network on meeting the project objectives of plume delineation. About 12 wells are present. The value these well provide on reducing the uncertainty is substantial. The maximum uncertainty measure has been reduced to less than three, and occurs primarily downgradient of the existing monitoring well network.

If further reduction in uncertainty is desired, then the optimal location to install one additional well is at the top of the surface shown in Figure 6-1b. The addition of a well at that location will reduce the uncertainty at that point to almost zero (the sampling error). The added information will also propagate in space around that location. The resulting new surface is shown on Figure 6-1c.

Note how the surface in Figure 6-1b was “pulled down” resulting in Figure 6-1c when the well was hypothetically added to the existing monitoring well network at the peak in Figure 6-1b. This shows how the information obtained from a monitoring well not only provides information at the sampling point, but aerially as well. This quantification of spatial information from a point estimate is a direct result of the linkage of flow and transport simulator, multiple aquifer realizations, and the Kalman

filter. If another far field well is desired, the optimal location is the location of the peak in Figure 6-1c. Adding information at this location results in the uncertainty surface depicted in Figure 6-1d. The percentages shown in Figure 6-1 relate to the uncertainty reduction after a monitoring well is installed.

At this point, the uncertainty in plume location is more focused near the source area. This indicates that the efforts to characterize the source will have more value that adding monitoring wells. Should constraints on well locations be present, the technique allows the placement of a well anywhere and will quantify the reduced benefit of not being able to locate a well in the precise optimal location.

The key finding is that knowing the concentration value in a far field well reduces all the uncertainty between the well and the source area. It also propagates this information in the perpendicular direction to that line. Hence, the benefit of adding distal wells, and the information they provide is very substantial. This is in contrast to approaches where the space-time correlation is not fully exploited, and a well only reduces uncertainty near the sampling point.

The plume finding technique is very valuable when allocating resources among the various total costs of an environmental contamination challenge. Specifically, this enables balancing investigation with remediation and long term monitoring. It also quantifies the level of certainty between investigations or various source areas on a site. This balances the site-wide investigations so that one area is not over assessed at the cost of large uncertainty in another part of the site.

6.1 LONG TERM PLUME MONITORING MODULE

Once the solution to the plume fringe finding has been designed and accepted, the long-term operations and monitoring often represent an expensive and significant portion of the investigation portion of the life-cycle cost. Solutions to this challenge are found in Herrera, et al. 1998, Herrera and Pinder 1998, Herrera 1998, Zhang and Pinder 2000, Zhang and Pinder 2002, and Minsker 2003. The long term monitoring program is optimized through an integration of plume simulation, statistical regression, Kalman filter, and a genetic algorithm. The objective of this analysis is to determine the location and timing of groundwater quality sampling events (in existing or proposed sensor locations) so as to achieve specified target goals of statistical accuracy over the specified period of analysis at the minimum cost. In this case, the target goal is the coefficient of variation (CV) divided by the mean concentration at the required time, and the location should be less than a certain number (e.g., 20%).

Essentially, this technique can be thought of as an extension of the plume finding algorithm. Whereas the plume finding algorithm uses an estimate of site conditions from a physical simulator and the field data to estimate plume location and maximum uncertainty in that knowledge, the long term monitoring algorithm incorporates the time dimension as well. In this case, the information content of a sample degrades as time elapses. One can visualize in Figure 6-2 a concentration graph as below, with a predicted smooth concentration transitions and sample point values proximate to it. This results in error bars on the time-concentration curve, so

that when a sample is taken the error is at a minimum. As time passes, the error bars will increase until it reaches a preset limit that triggers the next sampling event, reducing the error bars to a minimum value again. By optimizing the maximum error as a system, the least cost accuracy constrained long term monitoring policy is developed.

The procedure below is used to implement the long term monitoring optimization algorithm.

Objective Function:

$$\text{Minimize} \sum_j \left(\sum_i c \cdot \omega_{i,j} + c_0 \right) \cdot x_j \tag{6.1}$$

subject to:

$$\omega_{i,j} \le x_j \qquad (\forall i \in I,\ j \in J)$$

$$CV_{i,k} < CV_0 \qquad (\forall i \in I,\ k \in K)$$

$$x_j = \begin{cases} 1 & \text{if a well is installed at location j} \\ 0 & \text{if no well is installed at location j} \end{cases} \qquad \forall j \in J$$

$$\omega_{i,j} = \begin{cases} 1 & \text{if a sample is taken at time i, location j} \\ 0 & \text{if no sample is taken at time i, at location j} \end{cases} \qquad \forall i \in I,\ j \in J$$

Where J represents the candidate sets of well locations and i the candidate sets of sampling periods. The results are illustrated in Figure 6-2.

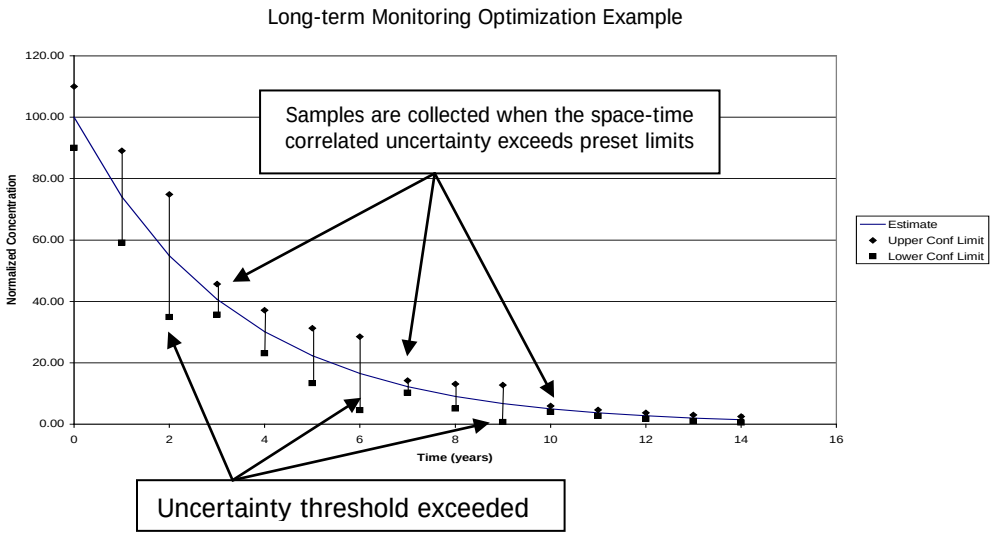

Figure 6-2. Long Term Monitoring Optimization Algorithm Concept.
When the uncertainty becomes excessive, a signal to collect is sample is triggered. Note: In this formulation, the uncertainty is formally computed as a correlated function of both space and time.

The Kalman filter is an optimal estimator that combines the measurements and system information to achieve a minimum error estimate. It is used to update the statistics of the concentration field after a sample or samples are taken. By combining the information from the physical models with the field measured data sets, the site knowledge and analysis techniques are used to a maximal extent.

In our case, we have a "measurement model" as follows:

$$Z = H x + \upsilon \tag{6.2}$$

where

- Z is a vector of m-noise-corrupted measurements;
- H is the measurement matrix, the dimension is m x n; and in this case, H is the sampling matrix of the form {0,0,1,1,0,1…..0,0,1,0,1}, where the values of [0,1] indication to take a sample at position (x,y,z) at time (t).
- x is a vector of dimension n which is the computed value of the desired quantity;
- υ is a vector of normal random noise, υ~N (0, R).

The Kalman filter is incorporated as follows:

State estimate update:

$$\hat{x}(+) = \hat{x}(-) + K[Z - H\hat{x}(-)] \tag{6.3}$$

$$P(+) = [I - HK]P(-) \tag{6.4}$$

where, (+) and (-) signs are estimates immediately after and before a measurement is taken, respectively;

- P is the error covariance matrix, $P = E[(\hat{x} - x)(\hat{x} - x)^T]$
- K is the Kalman gain matrix, obtained by minimizing a weighted scalar sum of the diagonal elements of P(+); $K = P(-)H^T[HP(-)H^T + R]^{-1}$

To implement this approach to long term monitoring design:

- Generate realizations using a Latin Hypercube sampling technique assuming:
    - hydraulic conductivity is a random field, and
    - contaminant source is a stochastic process.

- Simulate groundwater flow and transport to generate contaminant-concentration realizations.
- Compute a space-time correlation matrix of concentration.
- Use a genetic algorithm to select target locations and times to reduce most of the CV at the monitoring locations. A Kalman filter is used in this step to update the space-time correlation matrix after a sample is taken.
- Continue the selection process until the project goal is achieved.

Long term monitoring is the process where over time information on the location and characteristics of a subsurface plume in monitored. For further details, see (Zhang and Pinder 2002) as extended by (Zhang and Deschaine 2004). Specifically, the confidence of the plume knowledge is specified within a preset level, and the least cost, multi-period, long-term solution that satisfies the accuracy constraints is developed. To develop the long term monitoring program that best suits that needs of a project, several preset levels can be analyzed, and a cost versus accuracy curve generated. The stakeholders can then decide what level of monitoring will be necessary given the quantitative information on how various costs and accuracy are interrelated. The resulting optimization will allow for the development of a monitoring program using optimization algorithms to determine which monitoring wells to sample, the frequency to sample them, and which analytes to sample. The solution will identify which wells are key for the plume monitoring and which have minimal value. Depending on the results, minimal or no value wells can be removed from

consideration by the algorithm and the optimization rerun to see if any degradation in accuracy occurs. If not, the stakeholders could consider these wells for abandonment. If so, the monitoring wellhead protection plans can be reassessed to ensure the monitoring sensor assets are adequately protected.

7.0 EXTENSIONS TO DNAPL SOURCE AND UXO FINDING ALGORITHMNS

A DNAPL source algorithm has been deployed using the above plume and long term monitoring techniques, and is significantly extended to include new formulations regarding geologic uncertainty, incorporation of expert knowledge and the like. Information is available from the SERDP.ORG web page (Pinder *et al.*, 2008). In a similar manner, an unexploded ordnance (UXO) finding algorithm is developed (Deschaine *et al.*, 2002c), and deployed (Francone *et al.* 2007) on the field scale, and funded for additional testing by DoD as Environmental Security Technology Certification Program (ESTCP) project number MM-0811 (Deschaine *et al.*, 2008). To wit, these techniques – and the ability to add to and extend them in other disciplines - have proven quite beneficial.

8.0 ACTIVE PLUME REMEDIATION MODULE

The constrained engineering design problem is quite complex and well-studied. The site-wide optimizer needs to have the option of using multiple optimization algorithms, depending on the site-specific flow and transport system and the source-specific remedial technology. Aquifer remediation methods consist of primarily in situ treatment in which the contaminants are treated in place (via natural attenuation and/or engineered solutions), ex situ treatment in which the contaminants are removed from the aquifer, and institutional controls through which the sources of contaminants are managed rather than treated. The objective function is concave, with the minimum occurring on the lower rim of the state-space. Constraints on allowable residual contamination levels are present, as are physical constraints of the aquifer ability to be remediated as well as locations available for remediation infrastructure.

The general formulation for this challenge is:

$$\text{Minimize } f(x) = \sum_{i=1}^{N} (\delta_i, \alpha_i, q_i) \tag{8.1}$$

subject to:

$$\sum_{i=1}^{N} q_i \geq Q^* \forall i \in I$$
$$q_i \leq q_i^* \forall i \in I$$
$$c_j \leq c_j^* \forall j \in J$$
$$h_k \geq h_k^* \forall \in K$$
$$t \leq T_{\max}$$

In the above formulation, the minimization function f(x) is a simplified version of life cycle cost. This example cost formulation is approximated by the flow from a recovery well, q_i times a unit cost to treat the water, α_i times a (0,1) δ_i multiplier if a well installation is needed or is preexisting. This simplified version of the cost function is for illustrative purposes. Some solutions to this challenge include quite extensive implementation of a cost function, including using industrial remediation cost estimating software. Formulations exist for other remediation technologies (see Deschaine, 2007 for example), including bioremediation but are beyond the scope of this paper to formally explain. For pump-and-treat applications, the $q > Q$ constraint requires that the solution pump at least some minimum amount of water. The $q < q^*$ constraint sets upper limits on a flow at a well. The $c<c^*$ constraints ensure the aquifer is cleaned up to a certain acceptable residual level. The $h>h^*$ constraints set maximum allowable drawdown in the aquifer. These can be expanded to include constraints for differential settling from dewatering by adding constraints on the maximum slope of the water table. The $t<T$ constraint limits the allowable time for the activity to achieve the specified goals. Different remedial technologies have various and different physics models. Also, the general constraints - like the cost function - can be modified and adapted as specific project needs dictate.

The goal of the constrained global optimization algorithm for this challenge is to find at least one point that x^* within the feasible region that satisfies $f(x^*) \leq f(x)$ for all x, or show that such a point does not exist. Showing that a point does not exist

is instrumental to this challenge, as cases have occurred when the number and type of constraints added precluded a feasible region to the problem (Pinder 2003 – personal communication). This requirement screens out the many of the blind search methods, such as genetic algorithms, as these methods would iterate ad infinitum without providing either a solution or proof that one did not exist.

From a purely mathematical point of view, concave minimization belongs to the hardest class of mathematical programming problems (Pintér 1996, 2002). Outer approximation and Lipchitz Global Optimization (LGO) are proven techniques for this challenge. On the other hand, it is this characteristic that makes many of the interior point methods, such as evolutionary computation methods, not well suited to solve this type of problem from a theoretical standpoint. Other researchers have tried simulated annealing, genetic algorithms, and the like, without much success.

8.1 OUTER APPROXIMATION METHOD

This optimization algorithm, the Outer Approximation Method (a.k.a., the Cutting Plane Technique) is described by Karatzas and Pinder 1993; Karatzas and Pinder 1996; Spiliotopoulos, et al. 2000; and Deschaine, et al. 2001a. The method is a global minimization technique that uses a cutting plane approach to determine the optimal solution. The algorithm starts by determining a polytope defined by a set of vertices that encloses the feasible region. The feasible region is determined as the space in which all of the constraints are satisfied. It should be noted that a robust simulator of the subsurface physics is necessary for automatic optimization techniques

to be successful. Some of the problems, such as multiphase flow of water, NAPL and gas where all three phases are active are inherently difficult to get to converge, are expected to create a challenge during automated optimization.

The objective function, the function to be minimized, is formulated as a concave function. The objective function is also not differentiable at the intersection of the capital and operational costs. This algorithm takes advantage of the fact that a basic property of a concave function (f) is that the minimum of the function over a compact set of constraints is always obtained in at least one extreme point of the set. It does not rely on derivatives and, hence, the discontinuity of the objective function is not a performance issue. It also handles integer programming, as either a well is selected for use or not, but a half well has no physical meaning.

The outer approximation algorithm (Karatzas 2001) is implemented as follows:

Step 1. Define the enclosing polytope.

Step 2. Define the vertex that will minimize the objective function.

Step 3. Calculate the constraints and determine the most violated constraint.

Step 4. Generate a cutting hyperplane at the vertex of the most violated constraint and determine the new set of vertices without cutting through the feasible region.

Step 5. Iterate until optimal solution found or infeasibility (zero vertices) declared.

The outer approximation algorithm determines the vertex of the enclosing polytope that minimizes the objective function. Next, it examines if the selected vertex is feasible. If all constraints are satisfied, it declares this vertex as the optimal solution. Otherwise, a cutting plane is introduced that eliminates this vertex and its surroundings, creating a new enclosing polytope that is a better approximation of the feasible region, and the process is repeated. The goal of this process is not to determine the best approximation of the feasible region but rather to determine the most extreme point of the feasible region without eliminating any part of it.

If the cutting plane technique cuts the solution space until there is nothing left, then an "infeasible solution" will be declared. Infeasibility in the operations research world translates into technical impracticability in the remediation world. That a solution is impracticable with have mathematical basis, as opposed to having an analyst search for a solution and when one is not found, having the decision makers ask them to "try harder." Once infeasibility is declared, it will not matter how long an analyst attempts to find a solution, because none will exist. To find a feasible solution in this case, the constraint relaxation is necessary. That is, either the point of compliance and/or the concentration at the point of compliance needs to be made less strict. The relaxation of the constraints can be found by the algorithm. See [Deschaine, et. al. 2001a] for details of the algorithm and an applied solution technique. In one trial this method solved a problem about 15 times faster than the MINOS algorithm. While it is already a very efficient, effective approach, solution speed can be increased even

more if the simulation model can be accelerated. Between 20 and 100 calls to the simulator are common using this method. In one case, it was predicted that application of this technique would result in a savings of $72.6 million through the optimization of extraction and injection rates of a currently installed pump-and-treat system when compared with the current operational configuration [Karatzas 2001].

8.2 LIPCHITZ GLOBAL OPTIMIZATION

Lipchitz Global Optimization (LGO) is described in (Pinter, 1996 and Deschaine, *et al.,* 2006). The application of LGO to optimal groundwater remediation design is discussed in (Deschaine, *et al.,* 2013). The optimization procedure includes direct integration of the aboveground treatment system cost function with the below ground remediation solution and allows for minimization of cost, minimization of remediation time, and maximum mass removal for single and multiple management period scenarios. For example, it can automatically balance whether it is better to extract a lot of water at a low concentration or to extract less water at a higher concentration. This optimization includes the variable treatment cost for various flow rates and concentrations and is contaminant specific.

9.0 HEURISTIC OPTIMAL REMEDIAL DESIGN ALGORITHM

The Heuristic Optimal Remedial Design Algorithm (HORDA) is an algorithm that provides a consistent, traceable, and logical method for determining a least-cost solution to a remediation challenge. The original algorithm was first developed for saturated flow systems (Deschaine 1992) in the early 1990s and since then has been extended (and is continually being enhanced) for multi-phase/multi-component, variably saturated flow and biogeochemically reactive systems. In its simple form, it consists of a set of criteria that systematically directs the answer to the least-cost solution. The basic criteria are listed below:

1. Minimize the travel time of the chemical from its initial location to the in situ or ex situ treatment location. The in situ treatment location can move in space and time, such as in Redox zone optimization. The ex-situ treatment point is often an extraction well.
2. Transport the subsurface chemical material from locations of lower contamination through areas of higher contamination in route to the treatment location. Natural gradients, concentration field manipulation, and/or extraction/injection wells or galleries can accomplish this.
3. Achieve transport migration control to prevent further spread of the area of impact. This control can be accomplished by air sparging curtains, natural or

engineered in situ Redox zone optimization, or extraction and injection wells, for example.

4. Minimize the amount of clean water that is treated.
5. Achieve remediation of each location in the area of impact at the same time.

Each of the above criteria will have site-specific importance (i.e., weights) as it relates to the site remedial technology, source, and remediation-cost function. One application of the HORDA algorithm to a multiphase, single-component, chlorinated organic remediation project resulted in a remedial solution that took less than half the time (8 years versus 29 years) and saved more than 40% of the cost ($730,000 versus $420,000) compared to solutions proposed by others [Deschaine et al. 1998a and Deschaine 2007]. This is a simple method that remedial design practitioners without extensive optimization training can use to quickly find good – though not provably globally optimal – solutions.

10.0 MODFLOW OPTIMAL FLOW CONTROL

Developed as a tool for hydraulic optimization of groundwater extraction well system, MODFLOW Optimal Flow Control (MODOFC) couples the U.S. Geological Survey MODFLOW simulation program with an optimization algorithm (Ahlfeld and Mulligan 2000). MODOFC accommodates the linear pumping costs of wells (dollars per amount pumped or injected); installation of each well; and bounds on hydraulic head and head difference, individual and net pumping rates, and total number of wells. MODOFC converts the groundwater flow control problem into an optimization problem and requires MODFLOW to evaluate the fitness of the proposed solution. It is excellent at solving complex plume stabilization challenges in saturated aquifers with hydraulic control. It has saved $100M at a Superfund site when compared to alternative solutions (Ahlfeld 1998).

11.0 SIMULATOR REPLACEMENT – FIDELITY OPTIMIZATION

Integrated simulation and optimization typically requires a sequence of "expensive" function calls. While extremely valuable in concept, when the computation cost of simulations functions is high (hours/days) and or the optimization paradigm is inefficient (thousands of function calls), real-time or timely optimal solutions are elusive. The use of machine learning has developed high fidelity model of a process simulator that executes quickly (milliseconds), as opposed to hours. This function is then optimized using the LGO© solver, thus enabling optimization in real-time (Deschaine and Pintér 2003). Developing general functions that approximate simulators using artificial intelligence/machine learning to accelerate optimization is described using artificial neural networks (Peralta 2000) and linear genetic programming (Deschaine and Francone 2002a; Deschaine and Francone 2002b).

12.0 INFORMATION INTEGRATION AND SCALABILITY

Scalability is an important issue if we are ever to develop a system then extends beyond “site-wide” to state, country, or global optimal solution. Yet, precisely this capability is needed to respond to the global sustainability challenge. Philosophically, I define global sustainability generally as making decisions today that ensure future generations a comparable Earth.

A start of the process of integrating information globally is discussed in Deschaine *et al.,* 2000. This document describes an agent-based system that was used to link environmental databases in the United States and Europe to behave coherently as one. By linking these project files together electronically, updates that are made to the individual projects automatically update the program file. This speeds information transfer, provides a strong configuration management system, and reduces error. Linkage can be accomplished using innovative and ever changing Web-based database connectivity tools.

But actually, I am not convinced that this is the correct approach to use in all situations. I see this working on static systems at scales where I can go out, query many distributed databases, retrieve the information, assemble it into a matrix, and use it to make decisions. This is fine for slow-moving subsurface systems; however, some of the environmental analysis and protection applications are better represented as dynamic distributed systems in the scale of a country or worldwide. I am currently

investigating the Wave-WP paradigm. It may be the system of choice for optimized reaction on non-local disasters, where there are plenty of highly dynamic data at many locations and time. The Wave-WP will allow effective real-time, non-local solutions over a fully distributed data and or knowledge. These challenges include watersheds for public drinking water supplies. The watershed often has many point and non-point sources of contamination, and drinking water can consist of a mixture of surface and groundwater. A watershed is subject to rainfall events, which result in highly variable flow rates and water qualities. The watershed may be subject to instantaneous and intentional contamination. This problem specification requires spatial solutions over dynamic data, wherever the data resides. While the agent system discussed above will work fine for slow-moving groundwater systems, not handle this dynamic condition where optimal solutions are needed in real time. The solution methodology needs to keep the highest integrity as a system, react to rapid changes in the distributed environment, but without the need for central control. I conjecture that the Wave-WP paradigm is the best approach for these large-scale solutions where dynamics are important. The author of Wave-WP concurs that this is a valid conclusion and excellent route to pursue (Sapaty 1999).

13.0 CONCLUSIONS

A formal site-wide environmental impact remedial design optimization system has been developed. It is important to note that “optimization” does not singularly refer to cost minimization, but to the effective and efficient balance of cost, performance, risk, management, and societal priorities and uncertainty. This tool integrates all of these elements into a single decision framework. It provides a consistent approach to designing optimal remediation systems that are tractable, traceable, and defensible. The system is modular and scalable. It can be applied either as individual components or in total. Component deployment has already produced savings between $100M and $72.6M and timesaving of more than half of original expectations, while not sacrificing effective responses to safety of human health or the environment. These results do not indicate that the initial attempts at developing optimal site-wide response policies by others was defective, but that optimization of complex environmental challenges is very hard to accomplish without the proper integrated value engineering, simulation, and optimization tools.

14.0 REFERENCES

Ahlfeld DP (1986). "Designing Contaminated Groundwater Remediation Systems Using Numerical Simulation and Non-Linear Systems." Ph.D. dissertation, Princeton University, 1986.

Ahlfeld DP (1998). "Results of Applying the MODOFC Algorithm to a Groundwater Remediation Project in New Jersey." Presented as part of the *RCGRD Short Course on Risk-Based Groundwater Remediation Design* (University of Vermont, January), 1998.

Ahlfeld DP and Heidari M (1994). "Applications of Optimal Hydraulic Control to Ground-Water Systems." *Journal of Water Resources Planning and Management* 120, No. 3: 350–365, 1994.

Ahlfeld DP and Mulligan A (2000). *Optimal Management of Flow in Groundwater Systems*, Academic Press, San Diego, CA, 2000.

Babu, D. K., Pinder, G. F., Niemi, A., Ahlfeld, D. P., & Stothoff, S. A. (1987). Chemical transport by three dimensional groundwater flows. Department of Civil Engineering, School of Engineering & Applied Science, Princeton University.

BWXT and SAIC (2002). "Recommendation of the 2001Groundwater Modeling Technical Advisory Group for the DOE Pantex Plant", 125 pp., 2002.

Clement, T. P. (1997). A Modular Computer Code for Simulating Reactive Multispecies Transport in 3–Dimensional Groundwater Aquifers. Pacific Northwest National Laboratory, Richland WA, USA. PNNL-11720. Found online at: http://bioprocess. pnl. gov/rt3d. htm.

Deschaine LM (1992). "Cost Evaluation and Optimization of Ground Water Pump and Treat Programs," M.S. thesis, University of Connecticut, 1992.

Deschaine LM (2007). Simulation and Optimization of Subsurface Environmental Impacts; Investigations, Remedial Design and Long Term Monitoring of BioNAPL Remediation Systems, Chapter 9, in In-Situ Bioremediation of Chlorinated Ethene DNAPL Source Zones: Case Studies, Prepared by The Interstate Technology and Regulatory Council, Bioremediation of Dense Non-Aqueous Phase Liquids (Bio DNAPL) Team, 2007. www.itrcweb.org/Documents/bioDNPL_Docs/BioDNAPL-2.pdf

Deschaine LM and Francone FD (2002a). "Design Optimization Integrating the Outer Approximation Method with Process Simulators and Linear Genetic Programming." Paper presented at The Fourth International Workshop Frontiers in Evolutionary Algorithms (Research Triangle Park, NC, March 13), 2002.

Deschaine LM and Francone FD (2002b). "Extending the Boundaries of Design Optimization by Integrating Fast Optimization Techniques with Machine-Code-Based, Linear Genetic Programming." In Press as Chapter 2 of Information Processing with Evolutionary Algorithms, Advanced Information Knowledge Processing, Springer-Verlag, Heidelberg, Germany, 2002.

Deschaine LM and Pintér J (2003). Developing High Fidelity Approximations to Expensive Simulation Models for Expedited Optimization, presented at the INFORMS Multi-conference, October, Atlanta, GA, 2003.

Deschaine LM and Singarella PN (1985). “Priority Ranking of 114 Salt Storage and Maintenance Facilities with a Lotus Spreadsheet Model”. Prepared for the Connecticut Department of Transportation in concert with a Department-wide environmental assessment and remediation program. Presented at the MIT Microcomputer Short Course, Cambridge, MA, June, 1985.

Deschaine LM, Ahlfeld DP, Ades MJ, and O’Brien D (1998a). “An Optimization Algorithm to Minimize the Life Cycle Cost of Implementing an Aquifer Remediation Project—Theory and Case History.” *Simulators International XV, Simulation Series* 30, No. 3: 53–58, 1998.

Deschaine LM, Breslau B, Ades MJ, Selg RA, and Saaty TL (1998b). “Decision Support Software to Optimize Resource Allocation—Theory and Case History.” *Simulators International XV, Simulation Series* 30, No. 3: 139–144, 1998.

Deschaine LM, Brice RS, and Nodine MH (2000). Use of InfoSleuth to Coordinate Information Acquisition and Analysis in Complex Applications. Society for Computer Simulation’s Advanced Simulation Technology Conference, Washington, DC, USA April 2000. ISBN 1-56555-199-0, pp. 13-18, 2000.

Deschaine LM, Regmi S, Patel JJ, Fox TA, Ades MJ, and Katyal A (2001a). “Design Optimization of Groundwater Quality Management Challenges Using the Outer Approximation Method.” In *The Society for Modeling and Simulation International: Advanced Simulation Technology Conference* (Seattle, WA, April): 88–93, 2001.

Deschaine LM, Rawls P, Manfredo L, and Patel JJ (2001b). “The DOE NETL Program and Project Source Selection, Risk Quantifier, Management Support and Optimization Tool-Kit.” In *The Society for Modeling and Simulation International: Advanced Simulation Technology Conference* (Seattle, WA, April): 165–174, 2001.

Deschaine LM, Hoover RA, Skibinski JN, Patel JJ, Francone FD, Nordin P, and Ades MJ (2002c). Using Machine Learning to Compliment and Extend the Accuracy of UXO Discrimination Beyond the Best Reported Results of the Jefferson Proving Ground. Technology Demonstration, pages 46-52. Society for Modeling and Simulation International’s Advanced Technology Simulation Conference, San Diego, CA April 2002.

Deschaine LM, McKay M, Blanchard S, Pintér JD, and Francone F (2006). Finding and Identifying Objects Based on Noisy Data: A Global Optimization Approach – Part 1: & Part 2: Theoretical Approach and Applicability with Deployment Examples. EURO XXI, July, 2006, Reykjavik, Iceland.

Deschaine LM, Francone FD and Keiswetter D (2008). Advanced MEC Discrimination Study on Standardized Test-Site Data using LGP Discrimination and Residual Risk Analysis. DOD-ESTCP Research Grant Number MM1-017.

Deschaine, Larry M., Theodore P. Lillys, and János D. Pintér. (2013). "Groundwater remediation design using physics-based flow, transport, and optimization technologies." Environmental Systems Research 2, no. 1: 6. http://www.environmentalsystemsresearch.com/content/2/1/6/

Francone FD, Deschaine LM, and Warren JJ (2007). Discrimination of Munitions and Explosives of Concern at F.E.Warren AFB, GECCO, 2007 (London)

Herrera GS (1998). “Cost-Effective Groundwater Quality Sampling Network Design.” Research Center for Groundwater Remediation Design, University of Vermont, May, 1998.

Herrera GS, McGrath WA, and Pinder GF (1998). "Computer-Aided Risk Assessment in Problems of Groundwater Contamination." In *Proceedings of the First International Conference on Computer Simulation in Risk Analysis and Hazard Mitigation*. WIT Press, Computational Mechanics Publications, Southampton, UK: 51–60, 1998.

Herrera GS and Pinder GF (1998). "Cost-Effective Groundwater Quality Sampling Network Design." In *Proceedings: Computational Methods in Water Resources* (Crete, Greece): 51–58, 1998.

HGL (2014). MFSF/MODHMS: A comprehensive MODFLOW-based hydrologic modelling system. Code documentation and user's guide, Hydrogeologic Inc., Reston, VA USA

Karatzas GP (2001). "Results of Applying the Outer Approximation Method to a Groundwater Remediation Project in Tucson, Arizona." Presented as part of the *RCGRD Short Course on Risk-Based Groundwater Remediation Design*, (University of Vermont, January), 2001.

Karatzas GP and Pinder GF (1993). "Groundwater Management Using Numerical Simulation and the Outer Approximation Method for Global Optimization." *Water Resources Research* 29, No. 10, October: 3371–3378, 1993.

Karatzas GP and Pinder GF (1996). "The Solution of Groundwater Quality Management Problems with a Nonconvex Feasible Region Using a Cutting Plane Optimization Technique." *Water Resources Research*. 32, No. 4, April: 1091–1100, 1996.

Katyal, A., (2013) Users Manuals for BioFT3D and SA_MAPS, available from Resources & Systems International, Inc. 309 Cherokee Dr. Blacksburg, VA, 24060, USA

McGrath WA (1997). PhD Dissertation, "Sampling Network Design to Delineate Groundwater Contaminant Plumes." Research Center for Groundwater Remediation Design, University of Vermont, October, 1997.

McGrath WA and Pinder GF (1996). "Sampling Network Design for Delineating Groundwater Contaminant Plumes." In *Proceedings: Computational Methods in Water Resources* (Cancun, Mexico): 185–192, 1996.

Minsker B (2003). Long-Term Groundwater Monitoring; The State of the Art, ASCE/EWRI Task Committee Report. American Society of Civil Engineers, Reston, VA, USA, 2003.

Peralta RC (2000). *REMAXIM Users Manual, Version 1.4.* Utah State University, 2000.

Pinder GF, Dokou Z, and Deschaine LM (2008) Optimal Search Strategy for the Definition of a DNAPL Source Environmental Restoration. DOD-SERDP Research Grant Number ER-1347. Pinder, G. F., 2003 Personal communication.

Pintér JD (1996) Global Optimization in Action. Continuous and Lipschitz Optimization: Algorithms, Implementation and Applications. Kluwer Academic, Dordrecht.

Pintér JD (2002) Global optimization: software, test problems, and applications. In: Pardalos, P. M. and Romeijn, H. E., Eds. Handbook of Global Optimization, Volume 2, pp. 515-569. Kluwer Academic Publishers, Dordrecht.

Saaty TL (1996). Decision Making for Leaders; The Analytic Hierarchy Process for Decisions in a Complex World, RWS Publications, Pittsburgh, PA, 315 pp, 1996.

Sapaty, P (1999). Mobile Processing in Distributed and Open Environments, Wiley Interscience, John Wiley & Sons, Inc., 410 pages.

Spiliotopoulos AA, Karatzas GP, and Pinder GF (2000). "A Biconcave-Decomposition Method for the Optimal Design of Pump-and-Treat Remediation Systems Including the Treatment Plant." In *Computational Methods in Water Resources XIII* (Calgary, Canada): 547–554, 2000.

Therrien, R., Sudicky, E. A., & McLaren, R. G. (1994). User's Guide for FRAC3DVS: An Efficient Simulator for Three-dimensional, Saturated-Unsaturated Groundwater Flow and Chain Decay Solute Transport in Porous or Discretely Fractured Porous Formations. University of Waterloo, Ontario, Canada.

Unger, A.J.A, E. Sudicky, and P. Forsyth (1995). Mechanisms Controlling Vacuum Extraction Coupled with Air Sparging for Remediation of Heterogeneous Formations Contaminated by Dense Nonaqueous Phase Liquids, Water Resour. Res., 31, 1913–1925.

U.S. Department of Energy (DOE) (1997). *Draft National Plan 2006.* DOE Headquarters, Washington, DC, November, 1997.

U.S. General Accounting Office (GAO) (1993). *Clean-Ups Nearing Completion Indicate Future Challenges.* GAO, Washington, DC, September, 1993.

Wagner BJ (1995). "Recent Advances in Simulation-Optimization Groundwater Management Modeling." *Reviews of Geophysics*, Supplement: 1021–1028, 1995.

Zhang, K (2003). User's manual for TOUGH2 MP version 1.0. Technical report, Lawrence Berkeley National Laboratory.

Zhang Y and Pinder GF (2000). "A Latin-Hypercube Method for Evaluation of Hydraulic Conductivity Random Fields." In *Computational Methods in Water Resources XIII* (Calgary, Canada): 779–784, 2000.

Zhang Y and Pinder GF (2002). *Design of Optimal Groundwater Quality Monitoring Networks via Computer Assisted Analysis*. Research Center for Groundwater Remediation Design, University of Vermont, 2002.

Zhang Y and Deschaine, LM (2004) Optimization of Large Scale Subsurface Environmental Impacts: Investigations and Long Term Monitoring. Presented at the Conference on Accelerating Site Closeout, Improving Performance, and Reducing Costs through Optimization, June 15th – 17th, 2004, Dallas Texas.

Zheng, C. (1990). MT3D Users Manual. A Modular Three-dimensional Transport Model for Simulation of Advection, Dispersion and Chemical Reactions of Contaminants in Groundwater Systems.

Zheng, C & Wang, PP (1999). MT3DMS: A modular three-dimensional multispecies transport model for simulation of advection, dispersion, and chemical reactions of contaminants in groundwater systems; documentation and user's guide. ALABAMA UNIV UNIVERSITY.

Zyvoloski, G. A., Robinson, B. A., Dash, Z. V., & Trease, L. L. (1996). Users manual for the FEHMN application (No. LA-UR--94-3788-Rev. 1). Los Alamos National Lab., NM (United States).

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Sciences 161 (2004) 99–120
INFORMATION SCIENCES
AN INTERNATIONAL JOURNAL
www.elsevier.com/locate/ins

Extending the boundaries of design optimization by integrating fast optimization techniques with machine-code-based, linear genetic programming

F.D. Francone [a], L.M. Deschaine [a,b,*]

[a] *Chalmers University of Technology and RML Technologies, Inc., 11757, Ken Caryl Ave., #F-512, Littleton, CO 80127 USA*
[b] *Science Applications International Corporation, 360 Bay Street, Suite 200, Augusta, GA 30901, USA*

Received 10 March 2002; accepted 19 May 2003

Abstract

Optimized models of complex physical systems are difficult to create and time consuming to optimize. The physical and business processes are often not well understood and are therefore difficult to model. The models of often too complex to be well optimized with available computational resources. Too often approximate, less than optimal models result. This work presents an approach to this problem that blends three well-tested components. First: We apply Linear Genetic Programming (LGP) to those portions of the system that are not well understood—for example, modeling data sets, such the control settings for industrial or chemical processes, geotechnical property prediction or UXO detection. LGP builds models inductively from known data about the physical system. The LGP approach we highlight is extremely fast and builds rapid to execute, high-precision models of a wide range of physical systems. Yet it requires few parameter adjustments and is very robust against overfitting. Second: We simulate those portions of the system—for example, the cost model for the processes—these are well

* Corresponding author.
E-mail addresses: ffrancone@aimlearning.com (F.D. Francone), larry.m.deschaine@alum.mit.edu (L.M. Deschaine).

0020-0255/$ - see front matter
doi:10.1016/j.ins.2003.05.006

understood with human built models. Finally: We optimize the resulting meta-model using Evolution Strategies (ES). ES is a fast, general-purpose optimizer that requires little pre-existing domain knowledge. We have developed this approach over a several years period and present results and examples that highlight where this approach can greatly improve the development and optimization of complex physical systems.

1. Introduction

Engineers frequently encounter problems that require them to estimate control or response settings for industrial or business processes that optimize one or more goals. Most optimization problems include two distinct parts: (1) A model of the process to be optimized; and (2) An optimizer that varies the control parameters of the model to derive optimal settings for those parameters.

For example, one of the research and development (R&D) case studies included here involves the control of an incinerator plant to achieve a high probability of environmental compliance and minimal cost. This required predictive models of the incinerator process, environmental regulations, and operating costs. It also required an optimizer that could combine the underlying models to calculate a real-time optimal response that satisfied the underlying constraints. Fig. 1 shows the relationship of the optimizer and the underlying models for this problem.

The incinerator example discussed above and the other case studies below did not yield to a simple constrained optimization approach or a well-designed neural network approach. The underlying physics of the problem were not well understood; so this problem was best solved by decomposing it into its constituent parts—the three underlying models (Fig. 1) and the optimizer.

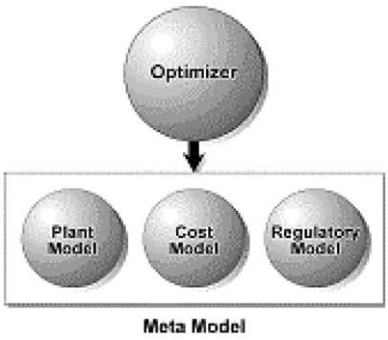

Fig. 1. How the optimizer and the various models operate together for the incinerator solution.

This work is, therefore, concerned with complex optimization problems characterized by either of the following situations.

First: Engineers often understand the underlying processes quite well, but the software simulator they create for the process is slow. Deriving optimal settings for a slow simulator requires many calls to the simulator. This makes optimization inconvenient or completely impractical.

Our solution in this situation was to reverse engineer the existing software simulator using Linear Genetic Programming (LGP)—in effect, we simulated the simulator. Such "second-order" LGP simulations are frequently very accurate and almost always orders of magnitude faster than the hand-coded simulator. For example, for the Kodak Simulator, described below, LGP reverse engineered that simulator, reducing the time per simulation from hours to less than a second. As a result, an optimizer may be applied to the LGP-derived simulation quickly and conveniently.

Second: In the incinerator example given above, the cost and regulatory models were well understood, but the physics of the incinerator plant were not. However good quality plant operating data existed. This example highlights the second situation in which our approach consistently yields excellent results. LGP built a model of plant operation directly from the plant operation data. Combined with the cost and regulatory models to form a meta-model, the LGP model permits real-time optimization to achieve regulatory and cost goals.

For both of the above types of problems, the optimization and modeling tools should possess certain clearly definable characteristics:

- the optimizer should make as few calls to the process model as possible, consistent with producing high-quality solutions;
- the modeling tool should consistently produce high-precision models that execute quickly when called by the optimizer;
- both the modeling and optimizing tools should be general-purpose tools. That is, they should be applicable to most problem domains with minimal customization and capable of producing good to excellent results across the whole range of problems that might be encountered; and

by integrating tools with the above characteristics, we have been able to improve problem-solving capabilities very significantly for both problem types above.

This work is organized as follows. We begin by introducing the Evolution Strategies with Completely Derandomized Self-Adaptation (ES-CDSA) algorithm as our optimization algorithm of choice. Next, we describe machine-code-based, LGP in detail and detail a three-year study from which we have concluded that machine-code-based, LGP is our modeling tool of choice for these types of applications. Finally, we suggest ways in which the integrated

optimization and modeling strategy may be applied to design optimization problems.

2. Evolution strategies optimization

Evolution strategies (ES) was first developed in Germany in the 1960s. It is a very powerful, general-purpose, parameter optimization technique [25–27]. Although we refer in this work to ES, it is closely related to Fogel's Evolutionary Programming (EP) [1,7]. Our discussion here applies equally to ES and EP. For ease of reference, we will use the term "ES" to refer to both approaches.

ES uses a population-based learning algorithm. Each generation of possible solutions is formed by mutating and recombining the best members of the previous generation. ES pioneered the use of evolvable "strategy parameters". Strategy parameters control the learning process. Thus, ES evolves both the parameters to be optimized and the parameters that control the optimization [2].

ES has the following desirable characteristics for the uses in our methodology:

- ES can optimize the parameters of arbitrary functions. It does not need to be able to calculate derivatives of the function to be optimized, nor does the researcher need to assume differentiability and numerical accuracy. Instead, ES gathers gradient information about the function by sampling [12].
- Substantial literature over many years demonstrates that ES can solve a very wide range of optimization problems with minimal customization [12,25–27].

Although very powerful and not prone to getting stuck in local optima, typical ES systems can be very time consuming for significant optimization problems. Thus, canonical ES often fails the requirement of efficient optimization.

But in the past five years, ES has been extended using the ES-CDSA technique [12]. ES-CDSA allows a much more efficient evolution of the strategy parameters and cumulates gradient information over many generations, rather than single generation as used in traditional ES.

As a rule of thumb, where n is the number of parameters to be optimized, users should allow between 100 and $200(n+3)2$ function evaluations to get optimal use from this algorithm [12]. While this is a large improvement over previous ES approaches, it can still require many calls by the optimizer to the model to be optimized to produce results. As a result, it is still very important

to couple ES-CDSA with fast-executing models. And that is where LGP becomes important.

3. Linear genetic programming

Genetic Programming (GP) is the automatic, computerized creation of computer programs to perform a selected task using Darwinian natural selection. GP developers give their computers examples of how they want the computer to perform a task. GP software then writes a computer program that performs the task described by the examples.

GP is a robust, dynamic, and quickly growing discipline. It has been applied to diverse problems with great success—equaling or exceeding the best human-created solutions to many difficult problems [2–4,14].

This paper presents approximately three years of analysis of machine-code-based, LGP. To perform the analyses, we used Versions 1 through 3 of an off-the-shelf commercial software package called Discipulus™ [22]. Discipulus is a LGP system that operates directly on machine code.

3.1. The genetic programming algorithm

Good, detailed treatments of Genetic Programming may be found in [2,14]. In brief summary, the LGP algorithm in Discipulus is surprisingly simple. It starts with a population of randomly generated computer programs. These programs are the "primordial soup" on which computerized evolution operates. Then, GP conducts a "tournament" by selecting four programs from the population—also at random—and measures how well each of the four programs performs the task designated by the GP developer. The two programs that perform the task best "win" the tournament.

The GP algorithm then copies the two winner programs and transforms these copies into two new programs via crossover and mutation transformation operators—in short, the winners have "children". These two new child programs are then inserted into the population of programs, replacing the two loser programs from the tournament. GP repeats these simple steps over and over until it has written a program that performs the selected task.

GP creates its "child" programs by transforming the tournament winning programs. The transformations used are inspired by biology. For example, the GP mutation operator transforms a tournament winner by changing it randomly—the mutation operator might change an addition instruction in a tournament winner to a multiplication instruction. Likewise, the GP crossover operator causes instructions from the two tournament winning programs to be swapped—in essence, an exchange of genetic material between the winners. GP

crossover is inspired by the exchange of genetic material that occurs in sexual reproduction in biology.

3.2. Linear genetic programming using direct manipulation of binary machine code

Machine-code-based, LGP is the direct evolution of binary machine code through GP techniques [15–18,20]. Thus, an evolved LGP program is a sequence of binary machine instructions. For example, an evolved LGP program might be comprised of a sequence of four, 32-bit machine instructions. When executed, those four instructions would cause the central processing unit (CPU) to perform operations on the CPU's hardware registers. Here is an example of a simple, four-instruction LGP program that uses three hardware registers:

```
register 2 = register 1 + register 2
register 3 = register 1 - 64
register 3 = register 2 * register 3
register 3 = register 2 / register 3
```

While LGP programs are apparently very simple, it is actually possible to evolve functions of great complexity using only simple arithmetic functions on a register machine [18,20].

After completing a machine-code LGP project, the LGP software decompiles the best evolved models from machine code into Java, ANSI C, or Intel Assembler programs [22]. The resulting decompiled code may be linked to the optimizer and compiled or it may compiled into a DLL or COM object and called from the optimization routines.

The linear machine code approach to GP has been documented to be between 60 and 200 times faster than comparable interpreting systems [10,15,20]. As will be developed in more detail in the next section, this enhanced speed may be used to conduct a more intensive search of the solution space by performing more and longer runs.

4. Why machine-code-based, linear genetic programming?

At first glance, it is not at all obvious that machine-code, LGP is a strong candidate for the modeling algorithm of choice for the types of complex, high-dimensional problems at issue here. But over the past three years, a series of tests were performed on both synthetic and industrial data sets—many of them data sets on which other modeling tools had failed. The purpose of these tests

was to assess machine-code, LGP's performance as a general-purpose modeling tool.

In brief summary, the machine-code-based LGP software [22] has become our modeling tool of choice for complex problems like the ones described in this work for several reasons:

- its speed permits the engineer to conduct many runs in realistic time frames on a desktop computer. This results in consistent, high-precision models with little customization;
- it is well-designed to prevent overfitting and to produce robust solutions; and
- the models produced by the LGP software execute very quickly when called by an optimizer.

We will first discuss the use of multiple LGP runs as a key ingredient of this technique. Then we will discuss our investigation of machine-code, LGP over the past three years.

5. Multiple linear genetic programming runs

GP is a stochastic algorithm. Accordingly, running it over and over with the same inputs usually produces a wide range of results, ranging from very bad to very good. For example, Fig. 2 shows the distribution of the results from 30 runs of LGP on the incinerator plant modeling problem mentioned in the introduction—the R^2 value is used to measure the quality of the solution. The solutions ranged from a very poor R^2 of 0.05 to an excellent R^2 of 0.95.

Our investigation to date strongly suggests the typical LGP distribution of results from multiple LGP runs includes a distributional tail of excellent

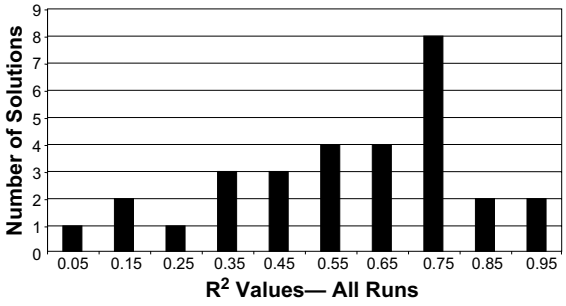

Fig. 2. Incinerator control data. histogram of results for 30 LGP runs.

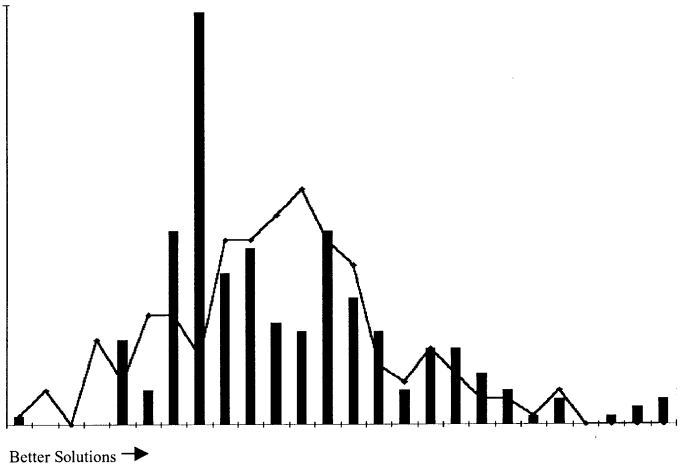

Fig. 3. Typical comparative histograms of the quality of solutions produced by LGP runs (bars) and Neural Network runs (lines). Discussed in detail in [8].

solutions that is not always duplicated by other learning algorithms. For example, for three separate problem domains, an LGP system produced a long tail of outstanding solutions, even though the average LGP solution was not necessarily very good. By way of contrast, and in that same study, the distribution of many neural networks runs on the same problems often produced a good average solution, but did not produce a tail of outstanding solutions like LGP [4,8].

Fig. 3 shows a comparative histogram of LGP results versus neural network results derived from 720 runs of each algorithm on the same problem. Better solutions appear to the right of the chart. Note the tail of good LGP solutions (the bars) that is not duplicated by a comparable tail of good neural network solutions. This same pattern may be found in other problem domains [id].

To locate the tail of best solutions on the right of Fig. 3, it is *essential* to perform many runs, regardless whether the researcher is using neural networks or LGP. This is one of the most important reasons why a machine-code approach to GP is preferable to other approaches. It is so much faster than other approaches, that it is possible to complete many runs in realistic time frames on a desktop computer. That makes it more capable of finding the programs in the good tail of the distribution.

6. Configuration issues in performing multiple LGP runs

Our investigation into exploiting the multiple run capability of machine-code-based LGP had two phases—largely defined by software versioning.

Early versions of the Discipulus LGP software permitted multiple runs, but only with user-predefined parameter settings.

As a result, our early multiple run efforts (described below as our Phase I investigation) just chose a range of reasonable values for key parameters, estimated an appropriate termination criterion for the runs, and conducted a series of runs at those selected parameter settings. For example, the chart of the LGP results on the incinerator CO2 data sets (Fig. 2) was the result of doing 30 runs using different settings for the mutation parameter.

By way of contrast, the second phase of our investigation was enabled by four, key new capabilities introduced into later versions of the LGP software. Those capabilities were:

- the ability to perform multiple runs with randomized parameter settings from run to run;
- the ability to conduct hillclimbing through LGP parameter space based on the results of previous runs;
- the ability to automatically assemble teams of models during a project that, in general, perform better than individual models; and
- the ability to determine an appropriate termination criterion for runs, for a particular problem domain, by starting a project with short runs and automatically increasing the length of the runs until longer runs stop yielding better results.

Accordingly, the results reported below as part of our Phase II investigation are based on utilizing these additional four capabilities.

7. Investigation of machine-code-based, linear genetic programming—Phase I

We tested Versions 1.0 and 2.0 of the Discipulus LGP software on a number of problem domains during this first phase of our investigation. This Phase I investigation covered about two years and is reported in the next three sections.

7.1. Deriving physical laws

Science Applications International Corporation's (SAIC's) interest in LGP was initially based on its potential ability to model physical relationships. So the first test for LGP to see if it could model the well-known (to environmental engineers, at least) Darcy's Law. Darcy's Law describes the flow of water through porous media. The equation is:

$$Q = K * I * A, \tag{1}$$

where Q = flow $[L3/T]$, K = hydraulicconductivity $[L/T]$, I = gradient $[L/L]$, and A = area $[L2]$.

To test LGP, we generated a realistic input set and then used Darcy's law to produce outputs. We then added 10% random variation to the inputs and outputs, and ran the LGP software on these data. After completing our runs, we examined the best program it produced.

The best solution derived by the LGP software from these data was a four-instruction program that is precisely Darcy's Law, represented in ANSI C as:

```
Q = 0.0
Q += I
Q *= K
Q *= A
```

In this LGP evolved program, Q is an accumulator variable that is also the final output of the evolved program.

This program model of Darcy's Law was derived as follows. First, it was evolved by LGP. The "raw" LGP solution was accurate though somewhat unintelligible. By using intron removal [19] with heuristics and evolutionary strategies the specific form of Darcy's Law was evolved. This process is coded in the LGP software; we used the "Interactive Evaluator" module, which links to the "Intron Removal" and automatic "Simplification" and "Optimization" functions. These functions combine heuristics and ES optimization to derive simpler versions of the programs that LGP evolves [22].

7.2. Incinerator process simulation

The second LGP test SAIC performed was the prediction of CO2 concentrations in the secondary combustion chamber of an incinerator plant from process measurements from plant operation. The inputs were various process parameters (e.g., fuel oil flow, liquid waste flow, etc.) and the plant control settings. The ability to make this prediction is important because the CO2 concentration strongly affects regulatory compliance.

This problem was chosen because it had been investigated using neural networks. Great difficulty was encountered in deriving any useful neural network models for this problem during a well-conducted study [6].

The incinerator to be modeled processed a variety of solid and aqueous waste, using a combination of a rotary kiln, a secondary combustion chamber, and an off-gas scrubber. The process is complex and consists of variable fuel and waste inputs, high temperatures of combustion, and high velocity off-gas emissions.

To set up the data, a zero and one hour off-set for the data was used to construct the training and validation instance sets. This resulted in a total of 44

input variables. We conducted 30 LGP runs for a period of 20 h each, using 10 different random seeds for each of three mutation rates (0.10, 0.50, 0.95) [3]. The stopping criterion for all simulations was 20 h. All 30 runs together took 600 h to run.

Two of the LGP runs produced excellent results. The best run showed a validation data set R2 fitness of 0.961 and an R2 fitness of 0.979 across the entire data set.

The two important results here were: (1) LGP produced a solution that could not be obtained using Neural Networks; and (2) Only two of the 30 runs produced good solutions (see Fig. 2), so we would expect to have to conduct all 30 runs to solve the problem again.

7.3. Data memorization test

The third test SAIC performed was to see whether the LGP algorithm was memorizing data, or actually learning relationships.

SAIC constructed a known, chaotic time series based on the combination of drops of colored water making their way through a cylinder of mineral oil. The time series used was constructed via a physical process experimental technique discussed in [24].

The point of constructing these data was an attempt to deceive the LGP software into predicting an unpredictable relationship, that is, the information content of the preceding values from the drop experiment are not sufficient to predict the next value. Accordingly, if the LGP technique found a relationship on this chaotic series, it would have found a false relationship and its ability to generalize relationships from data would be suspect.

The LGP was configured to train on a data set as follows:

- the inputs were comprised of eight consecutive values from the drop data; and
- the target output was the next-in-sequence value of the drop data.

Various attempts were tried to trick the LGP technique, including varying parameters such as the instructions that were available for evolving the solution.

The results of this memorization test are shown on Fig. 4. The "step" function shown in Fig. 4 represents the measured drop data, sorted by value. The noisy data series is the output of the best LGP model of the drop data.

It is clear that the LGP algorithm was not fooled by this data set. It evolved a program that was approximately a linear representation of the average value of the data set. But it did not memorize or fit the noise.

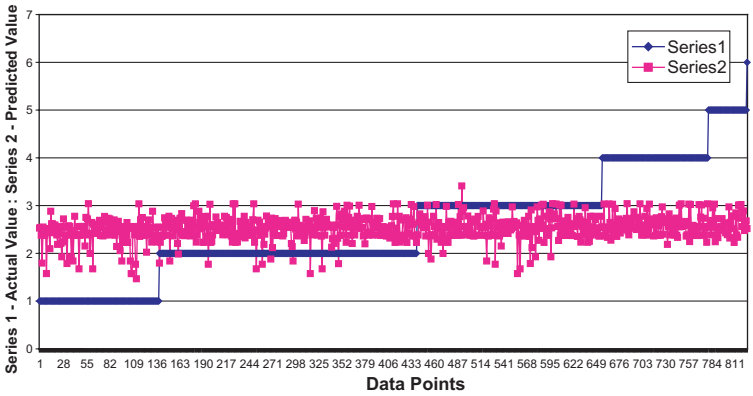

Fig. 4. Attempt to model a chaotic time series with linear genetic programming.

8. Investigation of machine-code-based, linear genetic programming—Phase II

Phase II of our investigation started when we began using Version 3.0 of the LGP software [22]. As noted above, this new version automated many aspects of conducting multiple runs, including automatically randomizing run parameters, hillclimbing to optimize run parameters, automatic determination of the appropriate termination criterion for LGP for a particular problem domain, and automatic creation of team solutions.

8.1. Incinerator problem, Phase II

SAIC used the new software version and re-ran the R&D problem involving CO2 level prediction for the incinerator plant problem (described above). A total of 901,983,797 programs were evaluated to produce the distribution of best 30 program results shown in Fig. 5.

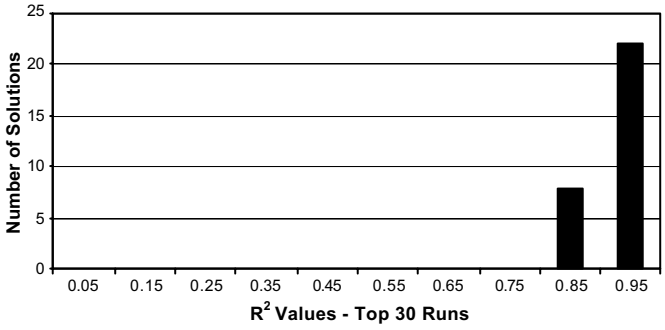

Fig. 5. Distribution of 30 best LGP runs using randomized run parameters for 300 runs on incinerator problem.

The enhanced LGP algorithm modeled the Incinerator plant CO2 levels with better accuracy and much more rapidly than earlier versions. The validation-data-set, seven-team, R^2 fitness was 0.985 as opposed to 0.961 previously achieved by multiple single runs. The CPU time for the new algorithm was 67 h (using a PIII-800 MHz/100 MHz FSB machine), as opposed to 600 h (using a PIII 533 MHz/133 FSB machine) that was needed in Phase I. It is important to note that the team solution approach was important in developing a better solution in less time.

8.2. UXO discrimination

The preceding examples are regression problems. The enhanced LGP algorithm was also tested during Phase II on a difficult classification challenge—the determination of the presence of subsurface unexploded ordnance (UXO).

The Department of Defense has been responsible for conducting UXO investigations at many locations around the world. These investigations have resulted in the collection of extraordinary amounts of geophysical data with the goal of identifying buried UXO.

Evaluation of UXO/non-UXO data is time consuming and costly. The standard outcome of these types of evaluations is maps showing the location of geophysical anomalies. In general, what these anomalies may be (i.e., UXO, non-UXO, boulders, etc.) cannot be determined without excavation at the location of the anomaly.

Fig. 6 shows the performance of ten published industrial-strength, discrimination algorithms on the Jefferson Proving Grounds Phase IV UXO data—which consisted of 160 targets [13]. The targets were either empty

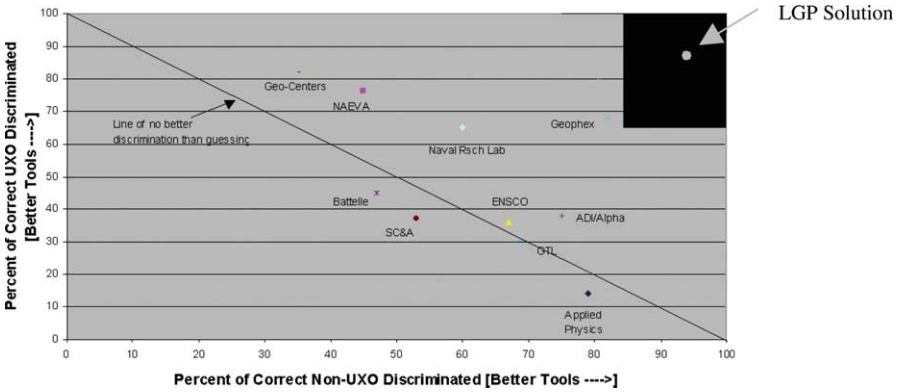

Fig. 6. LGP and ten other algorithms applied to the UXO discrimination data [13].

ground or buried scrap with characteristics similar to UXO's. Geophysical data from a ground survey was taken for each of the 160 targets.

The horizontal axis in Fig. 6 shows the performance of all ten algorithms in correctly identifying points that *did not* contain buried UXO. The vertical axis shows the performance of each algorithm in correctly identifying points that *did* contain buried UXO. The angled line in Fig. 6 represents what would be expected from random guessing.

Fig. 6 points out the difficulty of modeling these data. Most algorithms did little better than random guessing; however, the LGP algorithm derived a best-know model for correctly identifying UXO's and for correctly rejecting non-UXO's using various data set configurations [5,13]. The grey dot in the upper right hand corner of Fig. 6 shows the LGP solution on unseen data. Because the number of data points was small, we used resampling techniques to estimate the 95% confidence interval for the Discipulus solution. The black box in Fig. 6 shows that 95% confidence interval.

8.3. Eight-problem comparative study

In 2001, we concluded Phase II of our LGP study with a comparative study using machine-code-based, Linear Genetic Programming, back-propagation neural networks, Vapnick Statistical Regression [28], and C5.0 [21] on a suite of real-world, modeling problems.

The test suite included six regression problems and two classification problems. LGP and Vapnick Statistical Regression were used on all problems. In addition, on regression problems, Neural Networks were used and on classification problems, C5.0 was used.

Space constraints prevent us from detailing the experimental protocol in detail. That detail may be obtained in [9]. In summary, each algorithm was trained on the same data as the others and was also tested on the same held-out data as the others. The figures reported below are the performance on the *held-out, testing data*. Each algorithm was run so as to maximize its performance, except that the LGP system was run at its default parameters in each case.

8.4. Classification data sets results

Table 1 reports the comparative classification error rates of the best LGP, Vapnick Regression, and C5.0 results on the classification suite of problems on the held-out, testing data.

8.5. Regression data sets results

Table 2 summarizes the R^2 performance of the three modeling systems across the suite of regression problems on the held-out testing data.

Table 1
Comparison of error rates of best LGP, C5.0, and Vapnick Regression results on unseen data for two industrial classification data sets

Problem	Linear genetic programming	C5.0 decision tree	Vapnick Regression
Company H Spam filter	3.2%	8.5%	9.1%
Predict income from census data	14%	14.5%	15.4%

Table 2
Comparison of LGP, neural networks and Vapnick Regression on six industrial regression problems

Problem	Linear genetic programming	Neural network	Vapnick Regression
Department of Energy, Cone Penetremeter,	0.72	0.618	0.68
Kodak, software simulator	0.99	0.9509	0.80
Company D, chemical batch process control	0.72	0.63	0.72
Laser output prediction	0.99	0.96	0.41
Tokamak 1	0.99	0.55	N/A
Tokamak 2	0.44	0.00	0.12

Value shown is the R^2 value on unseen data showing correlation between the target function and the model's predictions. Higher values are better.

8.6. Two examples from the eight-problem study

This section will discuss two examples of results from the eight-problem comparison study—the Laser Output prediction data and the Kodak Simulator data.

8.6.1. Laser output problem

This data set comprises about 19,000 data points with 25 inputs. This is sequential data so the last 2500 data points were held-out for testing. The problem is to predict the output level of a ruby laser, using only previously measured outputs.

This is an easy data set to do well upon; but it is very difficult to model the phase with precision. Most modeling tools pick up the strong periodic element but have a difficult time matching the phase and/or frequency components—they generate their solutions by lagging the actual output by several cycles. Figs. 7 and 8 show the output of Vapnick Regression and LGP, respectively, plotted against a portion of the unseen laser testing data.

Fig. 7 is the result from the Vapnick tool. It picks up the strong periodic element but critically, the predicted output lags behind the actual output by a

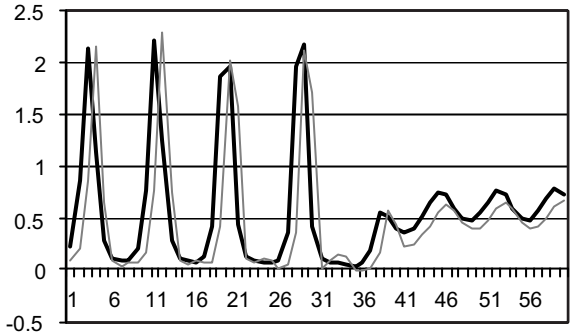

Fig. 7. Best Vapnick Regression model on laser problem (light gray line) compared to target output (heavy line) on held-out, data.

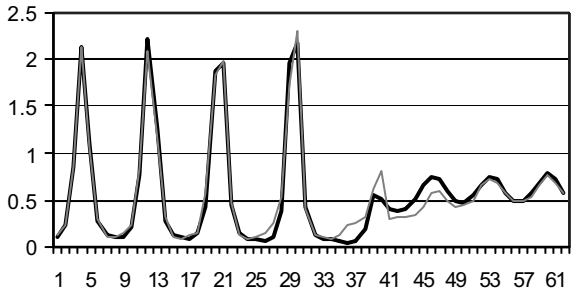

Fig. 8. Best LGP model (light gray line) on laser problem compared to target output (dark line) on held-out, testing data.

few cycles. By way of contrast, Fig. 8 shows the results from LGP modeling. Note the almost perfect phase coherence of the LGP solution and the actual output of the laser both before and after the phase change. The phase-accuracy of the LGP models is what resulted in such a high R^2 for the LGP models, compared to the others.

8.6.2. Simulating a simulator

In the Kodak Simulator problem, the task was to use LGP to simulate an existing software simulator. Past runs of the existing simulator provided many matched pairs of inputs (five production related variables) and the output from [23]. The data set consisted of 7547 simulations of the output of a chemical batch process, given the five input variables common to making production decisions. Of these data points, 2521 were held-out of training for testing the model.

The results on the testing or held-out data for LGP, Vapnick Regression, and neural networks are reported in Table 2. Figs. 9 and 10 graph the LGP and Vapnick Models against the target data.

The LGP solution (Fig. 10) so closely models the Target Output that the predictions completely obscure the target output line. In fact, for all but six of the 2521 data points, the agreement between the LGP prediction and the actual value is very good. The R^2 fitness on the applied data set for the best team solution was 0.9889. (A second batch of 232 LGP runs achieved a similar R^2 fitness on the applied data set of 0.9814, using a team of seven programs. The range of R^2 for the top 30 programs of this second batch was 0.9707–0.9585. This demonstrates analysis repeatability using LGP).

The Vapnick (Fig. 9) and Neural Network solutions were not nearly so close—the R^2 for the Vapnick Model was only 0.80, for example.

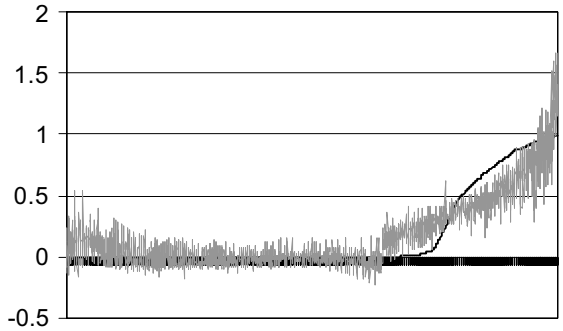

Fig. 9. Best Vapnick Predictions of Kodak simulator data (light gray series) vs. the target data (dark line) on held-out data.

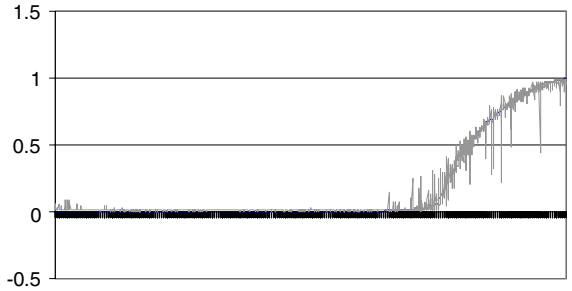

Fig. 10. Best LGP model of company K simulator problem (light gray series) vs target data (dark series) on the held-out data.

9. Conclusion regarding empirical studies

The key results of the two phases of our empirical studies of the LGP algorithm are as follows.

First: The LGP software we used consistently produces excellent results on difficult, industrial modeling problems with little customization of the learning algorithm. Note: LGP did not always produce *better* results than all other algorithms studied. However, on every problem studied, LGP produced a model that was as good as, or better than, any other algorithm.

The performance of other learning algorithms was decidedly up-and-down. For example, Vapnick Regression did quite well on the Cone Penetrometer and Company D data but quite poorly on the laser and Company K problems. Neural Networks did quite well on the laser and Company K problems but not so well on the Tokamak and incinerator CO2 data sets. C5.0 did well on the census problem but not well on the spam filter problem.

We speculate that one important reason behind the consistently good performance of LGP is that it performs, by default, many runs. Accordingly, it locates the tail of good performing solutions discussed above. Our comfort level that LGP will arrive at a good solution to most problems without customization or 'tweaking'' is one of the principal reasons we have settled on LGP as our modeling algorithm of choice for complex and difficult modeling problems.

Second: LGP produces robust models compared to other learning algorithms. Much less attention had to be paid to overfitting problems with the LGP software than with other algorithms. This is not to say that LGP will never overfit data. Give the right data set, it will. But it does so less frequently than the Neural Network, Vapnick Regression, and C5.0 alternatives we studied.

The LGP system identifies which are the important inputs, and which are not. For example, we screened a wastewater treatment plant with 54 inputs and identified 7 important ones. This reduces the number of inputs to monitor, allows assessment of what will happen if an input goes off-line (for security and contingency planning), and enhances accelerated optimization by reducing the number of decision variables, as discussed below.

10. Integrated system analysis

This work is concerned with the building of a system comprised of integrated modeling and optimization tools. The integrated tool suite, comprised of (1) Machine-code-based LGP for creating predictive models, and (2) ES-CDSA, is expected to ably handle a wide range of the complex problems with which we are concerned.

The remainder of this paper is devoted to discussing two application areas for the integration of these tools using two of the problems mentioned above—the incinerator R&D problem and the UXO discrimination problem.

10.1. Optimizing the incinerator model

The incinerator project was conceived from the beginning as a real-time control project. The models built with LGP predicted CO2 levels in the secondary combustion chamber as a function of: (1) The measured state of the plant at several previous time iterations; and (2) The plant control settings at previous time iterations.

Because plant control settings are part of the evolved LGP models, they may be optimized in response to each new state of the plant. That optimization may optimize for lowest cost operation, operation with a high probability of regulatory compliance, or both. Space limitations prevent a detailed description of the optimizer operation. However, details, including screenshots of the optimizer application, may be found in [4].

In terms of speed, optimization of these fast LGP models is practical and useful. The control programs evolved by LGP contain no more than 200 instructions so they will execute on a modern Pentium computer in far less than a millisecond. So, during optimization, each call to the optimizer occupies less than a millisecond. According to the formula given above for ES-CDSA optimization, $200 * (n+3)^2$ should suffice to derive an optimal setting for a new plant state. So, to optimize five parameters would take no more than 1.3 s—easily enough time to determine a new group of five control settings for the plant (the LGP model predicts an hour in advance).

11. Optimizing the LGP-derived UXO models

The problem of UXO or land mines affects millions of acres world-wide and includes both training areas and former battlefields. The estimated cost for remediating the U.S. training ranges alone is at least $14 billion, and this number is likely understated [11]. The very real cost of clean-up (or non-clean-up) is the injury or death to people.

Currently, too large a portion of the resources available for responding to UXO challenges is expended by digging up sites where UXO's are predicted, but which turn out to be false alarms—that is, false positives. This, in turn, limits funding available for remediating genuine UXO's.

Machine-code-based, LGP has derived the most accurate UXO discriminator among published results to date [13] by a wide margin. This LGP UXO/non-UXO identification success opens up the assessment and optimization of response to the UXO issue on both the program and the project level:

- The presence or absence of UXO can be assessed using remote, non-destructive technology such as land- or air-based sensors, including geophysics and various wave length sensors. Developed to a high degree of accuracy, wide areas can be screened and analyzed to reduce the footprint of areas needing further investigation. This will help manage the sheer size of this challenge.
- Areas of interest identified as requiring further investigation can be prioritized and ranked using good information on the probability or absence of UXO. This ranking would integrate the LGP UXO solution with multi-criteria/multi-objective decision support models; and
- site-specific remedial action plans, known as "dig sheets," can be optimally designed to focus efforts on high probability, UXO-containing areas. When the decreased predicted likelihood of UXO presence and the field verified absence of UXO are demonstrated, a stopping point for remedial activities, based on scientific principals and field validation, is provided.

12. Summary and conclusions

We are in the early stages of building a comprehensive, integrated Optimization and Modeling system to handle complex industrial problems. We believe a combination of machine-code-based, LGP (for modeling) and ES-CDSA (for optimization) together provide the best combination of available tools and algorithms for this task.

By conceiving of design optimization projects as integrated modeling and optimization problems from the outset, we anticipate that engineers and researchers will be able to extend the range of problems that are solvable, given today's technology.

Acknowledgements

The results presented in this work are part of a three-plus year collaborative effort between SAIC and Register Machine Learning Technologies to advance the state of the-art of evolutionary computation as applied to complex systems. Specifically thanked for funding and latitude from SAIC are Joseph W. Craver, John Aucella, and Janardan J. Patel. Dr. Gregory Flach, Dr. Frank Syms, and Mr. Robert Walton are gratefully acknowledged for providing the input data sets used in some of the work—as are the researchers who need to keep their identity confidential for business reasons. Christopher R. Wellington—Ad Fontes Academy, Centreville, Virginia conducted the chaotic drop experiment. All computations were performed by, and responsibility for their accuracy lies with, the authors.

References

[1] T. Bläck, H.P. Schwefel, An overview of evolutionary algorithms for parameter optimization, Evolutionary Computation 1 (1) (1993) 1–23.
[2] W. Banzhaf, P. Nordin, R. Keller, F. Francone, Genetic Programming, an Introduction, Morgan Kauffman Publishers, Inc., San Francisco, CA, 1998.
[3] L.M. Deschaine, Tackling real-world environmental challenges with linear genetic programming, PCAI Magazine 15 (5) (2000) 35–37.
[4] L.M. Deschaine, J.J. Patel, R.G. Guthrie, J.T. Grumski, M.J. Ades, Using linear genetic programming to develop a C/C++ simulation model of a waste incinerator, in: The Society for Modeling and Simulation International: Advanced Simulation Technology Conference, Seattle, WA, USA April 2001, pp. 41–48 (ISBN: 1-56555-238-5).
[5] L.M. Deschaine, R.A. Hoover, J.N. Skibinski, J.J. Patel, F.D. Francone, P. Nordin, M.J. Ades, Using machine learning to compliment and extend the accuracy of UXO discrimination beyond the best reported results of the Jefferson Proving Ground Technology Demonstration. Society for Modeling and Simulation International's Advanced Technology Simulation Conference, San Diego, CA, USA, April 2002.
[6] L.V. Fausett, A neural network approach to modeling a waste incinerator facility, in: Society for Computer Simulation's Advanced Simulation Technology Conference, Washington, DC, USA, April 2000.
[7] D.B. Fogel, Evolving artificial intelligence, PhD thesis, University of California, San Diego, CA, 1992.
[8] F. Francone, P. Nordin, W. Banzhaf, Benchmarking the generalization capabilities of a compiling genetic programming system using Sparse data sets, in: Koza et al. (Eds.), Proceedings of the First Annual Conference on Genetic Programming, Stanford, CA, 1996.
[9] F. Francone, Comparison of Discipulus™ Genetic Programming Software with Alternative Modeling Tools, 2002. Available from www.aimlearning.com.
[10] A. Fukunaga, A. Stechert, D. Mutz, A genome compiler for high performance genetic programming, in: Proceedings of the Third Annual Genetic Programming Conference, Jet Propulsion Laboratories, California Institute of Technology Pasadena, CA, Morgan Kaufman Publishers, 1998, pp. 86–94.
[11] Government Accounting Office, DOD Training Range Clean-up Cost Estimates are Likely Understated, Report to House of Representatives on Environmental Liabilities, USA General Accounting Office, April, Report no. GAO 01 479, 2001.
[12] N. Hansen, A. Ostermeier, Completely derandomized self-adaptation in evolution strategies, Evolutionary Computation 9 (2) (2001) 159–195.
[13] Jefferson Proving Grounds, Jefferson Proving Grounds Phase IV Report: Graph ES-1, May, Report No: SFIM-AEC-ET-CR-99051, 1999.
[14] J. Koza, F. Bennet, D. Andre, M. Keane, Genetic Programming III, Morgan Kaufman, San Francisco, CA, 1999.
[15] J.P. Nordin, A compiling genetic programming system that directly manipulates the machine code, in: K. Kinnear Jr. (Ed.), Advances in Genetic Programming, MIT Press, Cambridge, MA, 1994.
[16] J.P. Nordin, Evolutionary Program Induction of Binary Machine Code and its Applications, Krehl Verlag, 1999.
[17] J.P. Nordin, W. Banzhaf, Complexity compression and evolution, in: Proceedings of Sixth International Conference of Genetic Algorithms, Morgan Kaufmann Publishers, Inc., 1995.
[18] J.P. Nordin, W. Banzhaf, Evolving turing complete programs for a register machine with self modifying code, in: Proceedings of Sixth International Conference of Genetic Algorithms, Morgan Kaufmann Publishers, Inc., 1995.

[19] J.P. Nordin, F. Francone, W. Banzhaf, Explicitly defined introns and destructive crossover in genetic programming, in: K. Kinnear Jr. (Ed.), Advances in Genetic Programming 2, MIT Press, Cambridge, MA, 1996.
[20] J.P. Nordin, F. Francone, W. Banzhaf, Efficient evolution of machine code for CISC architectures using blocks and homologous crossover, in: Advances in Genetic Programming 3, MIT Press, Cambridge, MA, 1998.
[21] R. Quinlan, Data Mining Tools See5 and C5.0., Technical report, RuleQuest Research, 1998.
[22] Register Machine Learning Technologies, Inc., Discipulus Users Manual, Version 3.0. (2002) Available from www.aimlearning.com.
[23] B.S. Rice, R.L. Walton, Eastman Kodak Company, Industrial Production Data Set.
[24] Scientific American, Drop Experiment to Demonstrate a Chaotic Time Series, November 1999.
[25] I. Rechenberg, Evolutionsstrategie '93, Fromann Verlag, Stuttgart, Germany, 1994.
[26] H.P. Schwefel, Evolution and Optimum Seeking Sixth-Generation Computer Technology series, John Wiley & Sons, New York, 1995.
[27] H.P. Schwefel, G. Rudolph, Contemporary evolution strategies, in: Advances in Artificial Life, Springer-Verlag, Berlin, 1995, pp. 893–907.
[28] V. Vapnick, The Nature of Statistical Learning Theory, Wiley-Interscience Publishing, 1998.

A COMPUTATIONAL GEOMETRIC / INFORMATION THEORETIC METHOD TO INVERT PHYSICS-BASED MEC MODELS ATTRIBUTES FOR MEC DISCRIMINATION

© Larry M. Deschaine[1, 2], Peter Nordin[1] and János D. Pintér[3]

[(1)]Physical Resource Theory, Department of Energy and Environment: Chalmers University of Technology, Göteborg, Sweden; [(2)]HydroGeoLogic, Reston, VA, USA, and; [(3)]Pintér Consulting Services, Canada and Özyeğin University, Istanbul, Turkey

Journal of Mathematical Machines and Systems
(2010), Kiev No 2, (Chalmers Expanded Version)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS AND ABBREVIATIONS

3-D	three-dimensional
AUC	area under curve
EMI	electromagnetic instrument
ESTCP	Environmental Security Technology Certification Program
CG	computational geometric
CGPS	compiling genetic programming system
DGM	digital geophysical mapping
DoD	United States Department of Defense
GECCO	Genetic and Evolutionary Computation Conference
GPS	global positioning system
JPG	Jefferson Proving Ground
LGO™	Lipschitz Global Optimization
LGP	linear genetic programming
mm	millimeter
MAG	magnetic
MARS	Multivariate Adaptive Regression Splines
MEC	munitions and explosives of concern
MID	Mutual Information Difference
ML	machine-learning
mRMR	maximum-dependency, minimum redundancy
ROC	receiver operating characteristic

S/N	signal to noise
SME	subject matter expert
TOI	Target of Interest
UXO	unexploded ordnance
Warren	F.E. Warren Air Force Base

A COMPUTATIONAL GEOMETRIC / INFORMATION THEORETIC METHOD TO INVERT PHYSICS-BASED MEC MODELS ATTRIBUTES FOR MEC DISCRIMINATION

ABSTRACT

The presence of residual subsurface *munitions and explosives of concern* (MEC) is a significant issue worldwide. We are concerned with the military projectile items that were designed to explode, have been fired, did not explode and are now below ground posing life threatening risk if left unabated. To investigate the presence (or absence) of MEC in the field, non-invasive, non-destructive geophysical methods are used for data collection. To assess if an item is MEC, a wide-spread approach currently used is to examine an area using geophysical instruments, invert MEC-specific physically based models to fit the observed data; the inverted attributes (model parameters) form the basis for MEC discrimination. However, MEC discrimination via physics-based model inversion has significant difficulties succeeding in noisy environments (signal to noise (S/N) ratios below 100), is subject to non-unique solutions and the instrument location must be known to centimeter resolution. Our empirical findings demonstrate that our computational geometric method delivers an information-rich set of attributes that not only *recreates* the physical model inverted attributes but also provides valuable *additional* information useful for MEC discrimination *not obtainable using the inverse physics modeling approach.* It has successfully performed in the S/N ratio region of 10. In this work, we evaluate MEC discrimination using both methods; independently and combined. We argue for the broad applicability of the computational geometric method to develop robust MEC-discrimination attributes

to *extend the accuracy* of MEC discrimination; either as an independent analysis or combined with the physics model inversion technique.

1.0 INTRODUCTION

Solving munitions and explosives of concern (MEC) discrimination decision problems requires an in-depth understanding of the underlying science of geophysics. Our overall goal is to demonstrate the enhanced accuracy and performance possible from using machine learning function development to fuse the information content obtained from MEC feature attributes derived from both *data-driven functions* [using a blend of computational geometry, topology and algorithms (Edelsbrunner and Harer 2009)] and *physics-based models.* We describe the techniques, and how the machine-learning independent information-theoretic approach can be used to assess the contribution from each feature source (computational geometry or the fitted physics models) in MEC discrimination challenge. The physics-based governing equations provide the relevant scientific problem space of MEC item responses to geophysical interrogation. Computational geometry provides attributes for MEC and non-MEC (e.g. clutter, shrapnel). Hence, a key objective of this work is to merge and extend the techniques, effectively fusing both *a priori physics-based and automatic machine learning function-based* components to extend the maximum total discrimination/classification accuracy beyond that achievable by either method used independently. A related and equally important objective is to quantify the relative value of each component of the information sources in relationship to accuracy.

2.0 OVERVIEW OF MEC DISCRIMINATION

MEC discrimination presents one of the toughest and most challenging problems in the genre of subsurface identification tasks. A MEC item can, for instance, be unexploded ordnance (UXO) of various sizes and be buried below ground (See Figure 2-1). MEC can retain their ability to detonate; they pose a continuing risk. The United States Department of Defense (DoD) has invested heavily in basic research and development to address this challenge, but because typical MEC targets are small and surrounded by clutter (e.g., shrapnel or non-MEC items); accurate and reliable discrimination has been a challenge. Hence, while progress is being made, safe, efficient and cost-effective solutions have so far proven elusive.

Figure 2-1. Typical MEC and Non-MEC Items

(Image: US Army Environmental Command: Standardized Target Specifications: Technology Demonstration Sites)

Initially, MEC discrimination research focused on two primary approaches to evaluate a Target of Interest (TOI): the first, a physics-based approach (Bell et al. 2001), relied on mathematical models whereby model parameters were fitted to field data by solving the inverse modeling problem. A

second approach (Deschaine et al. 2002), which used automated machine-learning function development and multidisciplinary computational geometry (Welzl 1991) insights to derive features from the field data, clearly outperformed the other methods in use at that time to discriminate MEC from non-MEC. Both approaches are described below.

2.1 INVERSE (FITTED) PHYSICS-BASED MODELS

This section explains the inverse physics-based modeling approach for discriminating MEC items using electromagnetic (EMI)-based and magnetic (MAG) instruments.

One method to investigate the presence of MEC items is by conducting non-destructive geophysical surveys. This approach has value only if the resulting information is useable for locating anomalies and discriminating between MEC and non-MEC items. Since the MEC objects are not observable (being primarily below ground), the location, depth, and orientation of the MEC item are unknown. These model parameters are solved for by inverse modeling and are used to assess whether a TOI is a MEC item or not.

EMI uses induction theory and leverages the hypothesis that the distributions of the eigenvalues of magnetic polarizability provide an understandable basis for MEC versus non-MEC discrimination. This hypothesis is based on the observation that a MEC item can be approximated by an axisymetric cylindrical (as illustrated on Figure 2-1) and, therefore, has only two unique eigenvalues, one that represents the length of the object and the other two that represent the axial symmetry. Irregular objects (e.g., clutter), however, exhibit

three distinct eigenvalues (that is, different responses in three orthogonal directions). The inverse modeling solves for the best-fit eigenvalues (β1, β2, β3), the values of which are used for MEC discrimination. A MAG survey response is described by a simple dipole model. A tool that provides the best fit estimate for both EMI and MAG data (UX-Analyze) has been developed by Environmental Security Technology Certification Program (ESTCP) to facilitate these calculations (ESTCP 2009). Figure 2-2 illustrates the results of an inverse model fit for an anomaly investigated using both the EMI and MAG geophysical techniques.

The inverse modeling outputs seven EMI-fitted model parameters, these are the depth of the object (Depth), its size (Size), the eigenvalues (β1, β2, β3), the Coherence and the best-fit value (chi^2). Inverse physics modeling for the MAG sensor provides as outputs depth, size, declination, inclination, solid angle, and the magnetic moment. These parameters are then used as inputs for MEC identification function development using machine learning.

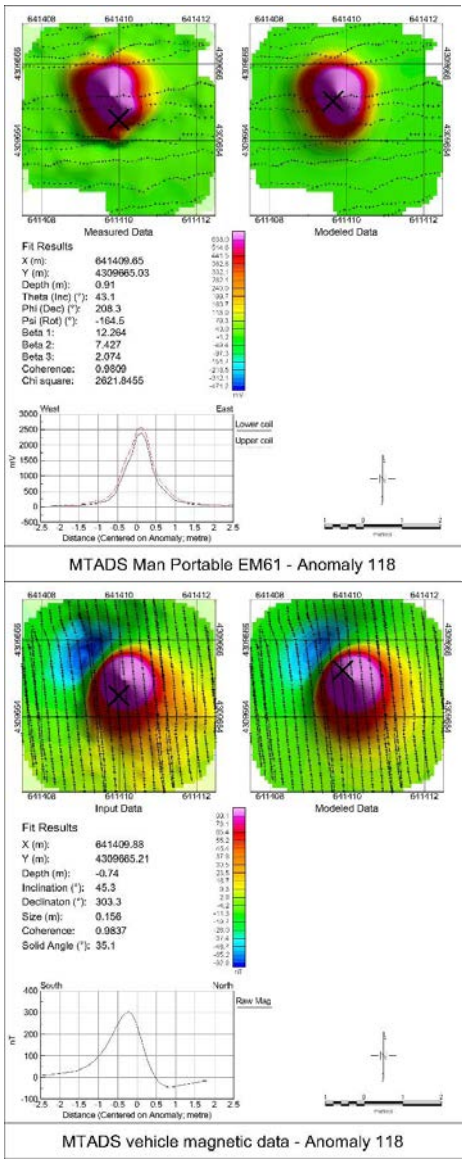

Figure 2-2. Inverse Modeling Analysis of EMI (above) and MAG (below) for One Anomaly using UX-Analyze

(From Figure 2-7 in ESTCP, 2009).

This approach provides fitted model parameters that are listed under the "fit results" output summary. MEC discrimination insight is gained from data

collected later in the decay curve which captures the anomaly metal thickness. The core concept regarding the EMI inverse model technique is that the polarizability will have one large (β1) and two small (β2, β3) and equivalent values to describe the conical MEC-shaped item. MAG relies on the shape and amplitude aspects. Hence, both shape (cylindrical versus fragments) and metal thickness (casings versus sheet metal) are also useful MEC discrimination information.

While theoretically sound, significant practical challenges to this method include: the need to overcome data collection positioning error (requires resolution *on the centimeter scale*); the need for non-noisy environments (signal-to-noise ratio [S/N] must be on the order of 100 to be successful), and; non-unique solutions of the eigenvalues. The inverse model parameters used in this work were developed by (Keiswetter 2008).

2.2 COMPUTATIONAL GEOMETRIC APPROACH

We have been incrementally developing the computational geometric machine learning approach to MEC identification since 2001 as a response to the documented difficultly the MEC discrimination industry was having in discriminating MEC items from clutter (Robitaille et al. 1999). This industry-wide difficulty in high accuracy MEC discrimination was evidenced by the field test conducted at the Jefferson Proving Ground (JPG) Phase IV. The JPG-IV MEC discrimination prove-out test site was developed by the DoD as part of the MEC (aka “UXO”) identification research program. Items were buried in a field, and vendors of MEC discrimination services were invited to identify them (the items identification was unknown to all but DoD). Each vendor submitted their analysis

to DoD for analysis. That well conducted study documents the state of the practice at that time - most of the discrimination attempts by the ten practitioners were no better than random guessing (Figure 2-3). The computational geometric approach has been tested this and subsequent data sets, and since it has evolved – and is currently evolving - we first present the applications, and then detail the current approach.

2.2.1 Jefferson Proving Ground – Phase IV Data

We first developed and tested the approach in the fall of 2001. We used the publicly available information and data sets for MEC (then called "UXO" for unexploded ordnance) discrimination from the Jefferson Proving Ground – Phase IV (Robitaille et al. 1999). The results are provided as Figure 2-3, and our approach outperformed all the techniques in use at that time (Deschaine et al. 2002). The data we used were collected by others using a Protem-47, time domain geophysical unit that provided 20 time gates of signal information, and our work was conducted after the conclusion and reporting of the 1999 study. The compiling genetic programming system (CGPS), a machine-learning technique developed by Nordin (Nordin 1994), was used as the classification algorithm (we later coined the phrase "linear genetic programming" [LGP] to differentiate it from other genetic programming algorithms). The results of this study are summarized in Deschaine (Deschaine et al. 2002) and are shown in Figure 2-3.

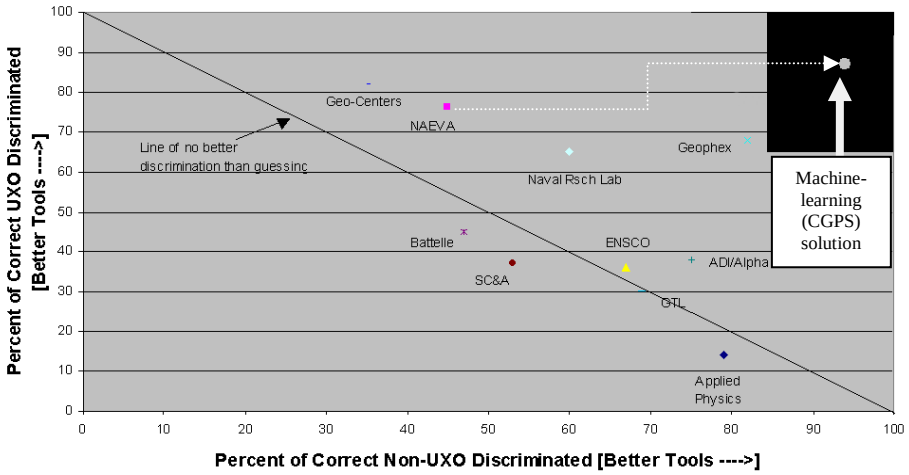

Figure 2-3. MEC Discrimination Solution Compared to Results from the JPG Phase IV UXO (MEC) Discrimination Project

Adapted from (Robitaille 1999). Note that our analysis was performed after this study was published; it was this poor performance by the MEC discrimination industry at that time that provided us the motivation to solve this challenge. The dashed white line shows the data set we used. The difference in predictive accuracy is the difference in the algorithm used.

Figure 2-3 shows the performance of the published results from 10 analyses conducted by vendors who provided MEC discrimination services as part of the JPG Phase IV project. The horizontal axis shows the performance of each method in correctly identifying anomalies that did not contain buried MEC; whereas the vertical axis shows the performance of each method in correctly identifying anomalies that did contain buried MEC. The angled line in the figure represents what could be expected from random guessing.

The difficulties of predicting MEC versus non-MEC using techniques available at the time are evident: most methods performed little better than random guessing would. The machine-learning based *computational geometric approach* using the CGPS algorithm provided the best-known approach at the time for

correctly identifying MEC and for correctly rejecting non-MEC using various data set configurations on blind data (Deschaine et al. 2002). The dashed line from the NAVEA solution in Figure 2-3 indicates which existing publically available data set for the machine-learning algorithm was used. The data we used was from a well conducted study, yet the analysis method used by others only produced results slightly above average. We selected this data because of its information value of multiple (20) time gates. Note that we intentionally did not use the data set labeled Geophex, even though it had the best performance of the group as analyzed by others, because we concluded that the NAVEA data had more information for high accuracy MEC discrimination — the team doing the original analysis just were not able to exploit it. The gray dot in the upper right-hand corner of the figure shows the CGPS solution on unseen data. Because the number of data points was small, we estimated the 95% confidence interval on this solution as shown on the black rectangle in Figure 2-3. CGPS – combined with computational geometric approach – produced by far the most accurate discrimination results. We then repeated our success of high accuracy MEC discrimination on the next round of testing at JPG, the data set from the JPG-V test set. These were published in (Francone et al. 2005).

2.2.2 Independent Validation of JPG-IV Results by DoD (US Army)

At Genetic and Evolutionary Computation Conference (GECCO) in 2004, we presented our findings (Francone et al. 2004) as a late breaking paper titled “Discrimination of Unexploded Ordnance from Clutter using Linear Genetic Programming” which discusses both the MEC discrimination successes at both

JPG-IV and JPG-V. Both sets of results bested all previously published performance results conducted by others. DoD to independently validated the use of genetic programming for high accuracy MEC discrimination (Banks et al. 2005).

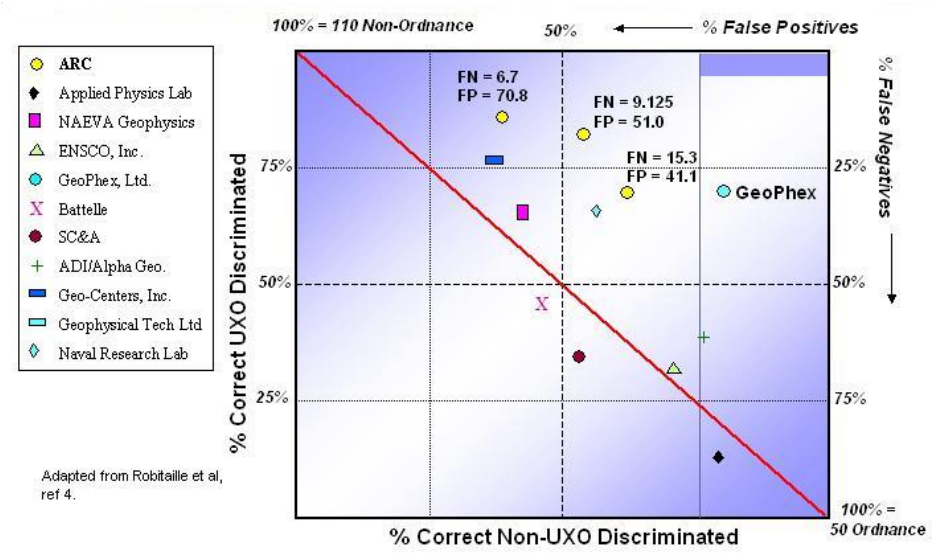

Figure 2-4. Independent Validation of MEC Discrimination Using the Genetic Programming Technique by DoD

(From Banks et al. 2005) and compared to results from the JPG Phase IV UXO (MEC) Discrimination Project (reproduced figure is from best available copy). In this well conducted study, an independent test demonstrates the value of using machine learning for high accuracy MEC identification.

The results of their well conducted study is provided on Figure 2-4, and while they did not achieve the same level of accuracy we did (they used different genetic programming tools and different feature extraction procedures) they did validate our findings that high accuracy MEC discrimination was – in fact – achievable.

2.2.3 Field-scale Test: F.E. Warren Air Force Base, Wyoming, USA

F. E. Warren Air Force Base ("Warren") is located near Cheyenne, Wyoming. In the past, a portion of that base served as a practice range, primarily for 75 millimeter (mm) and 37mm projectiles. We analyzed digital geophysical mapping (DGM) data comprising over 60 million data points in four channels of data from 600 acres of Warren (the "Site"). Each data point consisted of four channels of information gathered from a Geonics EM61 MK2 configured with three time-decay channels on the lower-coil and one upper-coil channel. These data were integrated with a differential global positioning system (GPS). These data contained almost 30,000 targets of interest identified by geophysicists, including three-hundred thirty-two 75mm projectiles (75mm) and 37mm projectiles (37mm). A little under one-third of the ground truth was held back by the independent judges for blind-testing. Our task was to discriminate intact 37mm's and 75mm's from the clutter by ordering the targets from most-likely to be MEC to least-likely to be MEC in what is referred to as a "prioritized dig-list" (Francone et al. 2007). The signal to noise ratio for the targets were on the order of 10, much lower than the order of 100 needed for the physically based inverse modeling approach to be successful. We identified all 75mm's by 28.2% of the way through our prioritized dig-list and all 37mm's by 64.2% of the way through the prioritized dig-list. Thus, depending on ordnance type, we reduced the number of targets that had to be excavated (false alarms) to clear the entire site by between 35% and 72%.

In all three tests, JPG-IV, JPG-V and FE Warren, while we were very successful, we could not articulate *why* we were so, from a physical understanding basis. The research in this paper shows that the computational geometric approach contains the information content that is obtained from applying the physically based inversion modeling approach. Hence, our approach complements and extends the information extracted from the field data with the physically based inversion approach; which is the original intent of our algorithm as noted by the title of our 2002 paper (Deschaine et al. 2002), a complementary and accuracy extension algorithm.

2.2.4 Computational Geometric Approach Details

The approach we found that is robust, practical, flexible, and effective formulation of computational geometry and topology. Because there are essentially an infinite number of features that can be derived using this approach, it presents a particular challenge for any machine-learning approach, namely that of input attribute explosion. For example, the approach used to generate the results cited below that used the EMI field instrument with four time gates generates 633 attributes. The geometric attributes extracted are based on finding an optimized ellipsoid that is constructed either automatically using the Lipschitz Global Optimization (LGO™) technique (Pintér 1996) or by an expert geophysicist who draws a polygon around the target of interest. This is effective at enclosing the geophysical data since, for a set of non-collinear points [non-planar in three-dimension (3-D)], the bounding ellipsoid exists and is unique (Welzl 1991). To generate the features (aka attributes), the ellipsoid is divided into slices and the

features are computed as a whole geometric shape, within quadrants and within the segmentations. Figure 2-5 illustrates the result of defining the minimum enclosing ellipse using computational geometry and Figure 2-6 illustrates the computational geometric process of segmenting the ellipsoid for attribute derivation.

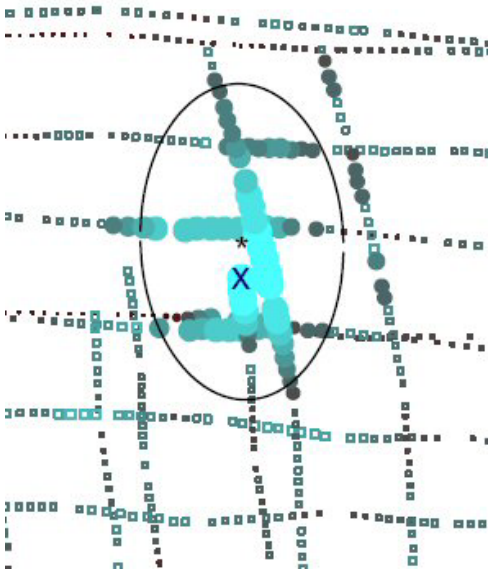

Figure 2-5. Defining a Target of Interest Area Using the Minimum Enclosing Ellipse Approach

The path the geophysical survey traversed is depicted by the survey lines. The signal associated with the TOI is segregated from the geophysical survey data set. The size of the circles represents signal strength. Computational geometry and computational topology (Novikovand Fomenko 1990) is used to find the minimum enclosing representative ellipse - which is similar in concept to finding the convex hull but that the shape of the enclosure is specified – and for feature extraction. Data points interior to the ellipse are used for feature development (see Figure 2-6). The survey lines shown here and on Figure 2-2 illustrate the difficulty in maintaining straight lines and rectangular surveys paths during field data collection.

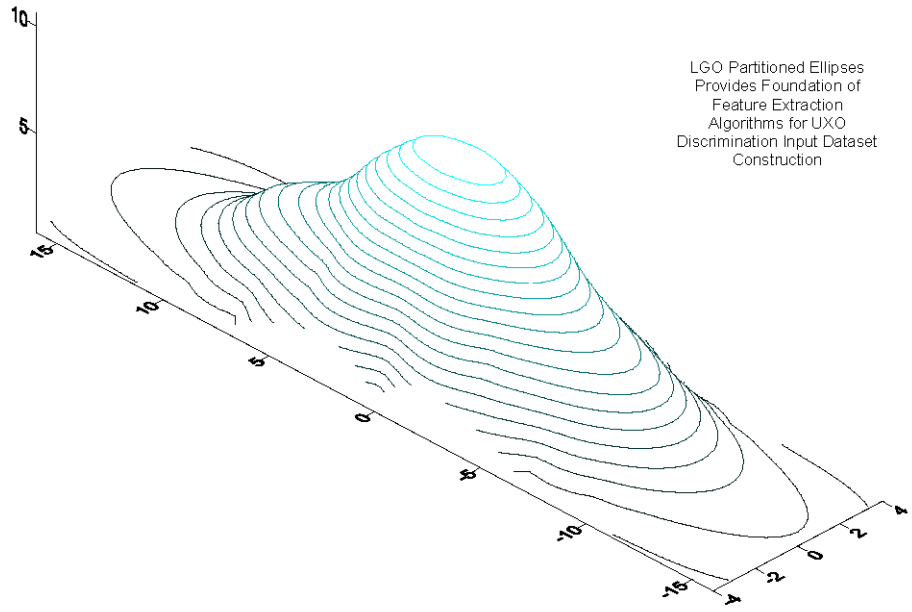

Figure 2-6. Development of Computational Geometrically Derived Attributes Using a Globally Optimized Ellipsoid

Over 500 candidate MEC discrimination features are derived from the total and segmented of the ellipsoid into regions of interest. The over 500 features developed during this process comprise of detailed statistical topology, ratio between time gates and coils, median values, decay curves, amplitudes and the like. Coordinates are local (meters).

3.0 ATTRIBUTE ANALYSIS: COMPUTATIONAL GEOMETRY AND INVERSE PHYSICS MODELS

Our hypothesis is that when the attributes from computational geometry and fitted inverse physics-based modeling approach are combined, the resulting integrated function generated with machine learning will perform at least equal to or better than either approach used alone since combining them should contain more information from which to make a MEC discrimination decision. The Venn diagram shown in Figure 3-1 illustrates the concept of MEC identification information space and their areas of overlap for subject matter expert (SME), physically based, and computational geometry (data-driven) solution approaches.

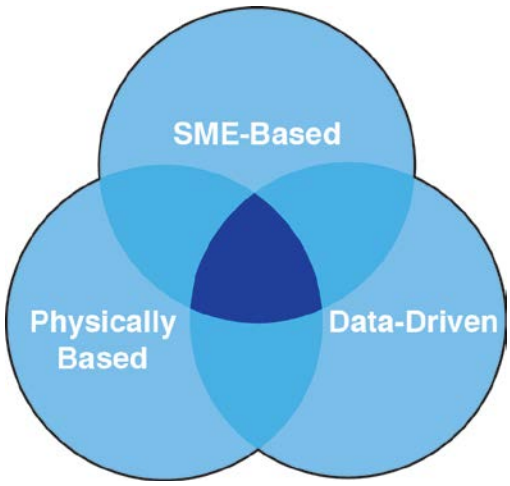

Figure 3-1. Venn Diagram Illustrating Independent and Redundant (Overlapping) Information Areas Valuable to MEC Discrimination

The equally weighted depiction is an idealized situation for illustration purposes only. The overlapping areas are the mutual, or shared, information (information that is common to both sources).

The maximal information space available for MEC identification is the total area outlined in black. It is computed as the sum of the areas of the

information circles minus any areas of overlap (aka redundancy). This overlap is called the mutual (or shared) information. The information-theoretic analysis characterizes this maximal, non-redundant problem approach mathematically. The area of the Venn diagram (the areas outline in bold) is given by the joint entropy (H) minus the mutual information (I), which represents the overlapping areas.

To maximize this area:

$$\max H(X, Y, Z) = \max [H(X) + H(Y) + H(Z) - I(X;Y;Z)] \quad (1)$$

Where X, Y, Z represent the sources of information (i.e., the information dimensions) of SME, the physically based inversion models, and data-driven (geometric and topologic) approaches, respectively. To test our hypothesis, we must quantify the relative contribution of each of the three entropy spaces both in respect to achieving the accurate MEC discrimination **and** with respect to one another. We will now test this hypothesis first theoretically using information theory, and then empirically, using machine learning.

3.1 MUTUAL INFORMATION ANALYSIS

Understandability of the individual attributes and relationships used for MEC classification analysis is important. While the computational geometric approach has been shown to be a viable approach, we have yet been able to fully understand, and explain, why. The amount of attributes and complexity of the evolved solutions can make the solution difficult to understand, and understanding a functional solution is important for having confidence in it. Furthermore, methods for feature reduction such as principal components analysis, while quite valuable for reducing the number of inputs in a data set used for machine learning,

require complex computations to be performed that combine many attributes into a single input vector. Machine learning algorithms provide rankings of variable importance, but these can vary based upon the specific algorithm. This obfuscates solution understandability. In the section below, we describe and test an approach to reduce the attributes required for MEC discrimination modeling using mutual information because if offers the advantage of preserving the individual attribute identity and is independent of the machine learning algorithm (Peng et al. 2005).

To test the mutual information approach on both attribute reduction and relevancy and redundancy assessment, the data sets from the ESTCP Camp Sibert project (Deschaine et al. 2009 and Keiswetter 2008) are combined so they contain attributes from both the fitted *physics-based model parameters* and the *computational geometric approach*; the MEC identity is a binary label (1 for MEC, 0 for non-MEC). The data was collected by others as part of the ESTCP project and provided to us for this analysis. The EMI data set consists of 174 instances (rows), of which 67 are MEC and 107 are non-MEC. There are seven attributes for the fitted physics-based model and 551 for the computational geometric based function. The MAG data set consists of 182 instances (rows) of which 56 are MEC and 126 are non-MEC. There are six attributes for the fitted physics-based model and 82 for the computational geometric-based function.

When building a solution approach, each of the information dimensions of SME, data-driven and physically based contain one or more vectors. These vectors are organized into a flat file as illustrated in Table 3-1.

Table 3-1. General Design of Information Dimension Assembly for Value Quantification

Case	SME	Computational Geometry (Data Driven)	Physically Based	Solution
e_1	$SME_1(i=1…a)$	$DD_1(i=1…b)$	$PB_1(i=1…c)$	L_1
e_2	$SME_2(i=1…a)$	$DD_2(i=1…b)$	$PB_2(i=1…c)$	L_2
e^3	$SME_3(i=1…a)$	$DD_3(i=1…b)$	$PB_3(i=1…c)$	L_3
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
e_n	$SME_n(i=1…a)$	$DD_n(i=1…b)$	$PB_n(i=1…c)$	L_n

The case example (e_n) is an example for a TOI, and the label (L) refers to whether that item is a MEC item or not. SME_n refers to the set of (integer "a" features) inputs that contain the information obtainable from SME dimension. Similarly, DD_n and PB_n are defined.

The challenge of identifying MEC from innocuous subsurface objects is analyzed using this information approach. For this Camp Sibert example, the dataset consists of geophysical signals obtained from surveying the ground with EMI over an area where MEC items (n=174) are known to exist. For the EMI instrument, a total of 558 features were computed to describe this data; seven of them derived via SME/physically based inverse modeling (a=0; b=7; a=0 since it was consumed by the physically based inverse model approach) and the rest (c=551) are from data-driven feature extraction algorithms based on the minimum enclosing ellipse approach from computational geometry.

Mutual information content has long been used for assessing the importance of attributes for function building (Varmuza 1974). The method used here is based on mutual information, using *a maximum-dependency, minimum-redundancy framework* **(mRMR)** as developed by (Peng et al. 2005). This technique provides the necessary theoretical engine to select the best candidate

features independent of a machine-learning classifier. The computations are based on the following set of equations:

Given two random variables (*X.Y*), their mutual information I(X;Y) is defined in terms of their marginal and joint probability density functions *p(x)*, *p(y)*, and *p(x,y)*:

$$I(X;Y) = \sum_{x} \sum_{y} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)} \tag{2}$$

In terms of designing a focused MEC discrimination feature *set*, the mutual information between the feature and the label (MEC identifier [0,1]) is our random variables X and Y when considering the relationships between the features and the label should be high. However, between the features, $I(X_i,X_j)$ needs be low so that redundant features are not used. Hence, the goal of feature selection is to develop the set S of m features $\{x_i, i=1\ldots m\}$ which jointly have the largest dependency (or in this case relevance) on the target class, that is the classification of MEC (aka UXO) while at the same time minimize the redundancy within the feature set S_m, hence we minimize the internal redundancy (R) of the input vectors set S:

$$\min(R) \tag{3}$$

Where:

$$R = \frac{1}{|S|^2} \sum_{x_i x_j \in S} I(x_i; x_j) \tag{4}$$

Where R is the redundancy computation of a pair of inputs (x_i; x_j) in information theoretic units of "bits," and S is the set of inputs from the m

information dimension(s). The goal of this optimization is to select from among various problem solution approaches (more specifically, their respective input vectors) which jointly have the minimum redundancy with respect to each other. Visually, minimization is used to force the information overlap in the Venn diagram of Figure 3-1 to be small. The technique provides an especially useful way to quantify the value of the inputs as it allows for optimal solution design and either reduces unnecessary inputs, and when combined with cost and reliability of the inputs can design a system that uses less expensive inputs, inputs of equal informational value or of greater reliability. These types of designs can save both time and fiscal resources.

At the same time, one desires a high degree of mutual information between the selected *set* of inputs and the solution (aka "label") when using supervised machine learning techniques. Here, the goal is selecting from among various problem solution approaches the selection of the set S of m features $\{x_i, i=1\ldots m\}$ which jointly have the largest relevance with respect to the solution's answer, known as a label "L" while having minimal redundancy. This is accomplished using the formulation:

$$\max D(S,L) \tag{5}$$

Where:

$$D = \frac{1}{|S|}\sum_{x_i \in S} I(x_i; L) \tag{6}$$

Conceptually, the mutual information shared internally within the *set* of informational dimension inputs (which consist of m total features) should overlap

with the answer L to a large extent. The label L – a known answer to a given problem - is important as it provides the means verify and validate the solution process. For example, a uniform input, or a randomly distributed one, would have low mutual information with the solution label, and hence is not be valuable for contributing to the problem solution.

To optimize the solution approach design, the D (dependency) and R (redundancy) is determined simultaneously using an objective function formulation of either a difference [(maxΦ(D,R); Φ=D-R] or the quotient [maxΦ(D,R); Φ=D/R]. The optimization is conducted using a greedy algorithm; the data set constructed by adding one input at a time. For example, when the algorithm is scanning through the inputs, the vector with the largest [D-R for difference] or [D/R for quotient] value would be added to the set S of m inputs.

The goal of this mRMR approach (Peng et al. 2005) is to reduce the number of attributes while still covering a maximal amount of information space. Using a smaller input data set (with the same information content) will result in faster running as well as higher accuracy of machine-learned functions. We use it to assess the *relative importance/redundancy* between the *fitted attributes from the physics models* and the *computational geometry attributes* on each of the subsets of the EMI and MAG data.

3.2 APPLICATION OF mRMR

The first step in applying mRMR to the feature value assessment for the MEC discrimination challenge is preparing the data set. The TOI is discrete; each case is labeled either as a 1 for MEC or as a 0 for not-MEC. The computational

geometric approach generates features that are represented as continuous variables. Mutual information of discrete variables was used and the variables discretized by using two thresholds: the mean (+/-) alpha*standard deviation as discussed in (Peng et al. 2005). mRMR is available as open source, as a web-based application, C and Matlab code. The parameter settings are alpha = 1.0, number of variable states = 3 and number of variables retained = 50 and the feature selection scheme was Mutual Information Difference (MID).

3.3 mRMR RESULTS

The mRMR ranking produces a rank-ordered list of features, with the top 50 of the candidate inverse physics or computational geometric EMI features being retained. To understand how to use the mRMR results, consider an example of the top three variables, V1, V2 and V3. This ranking means that if a single variable is desired, then variable V1 should be used. If two variables are desired, then the combination of V1 and V2 is better than the combination of V1 and V3, or V2 and V3.

The results discussed below indicate that information contained by developing a feature data set using the computational geometric approach for the EMI data set contains essentially all the relevant information needed for MEC discrimination that is contained in the inverse physics modeling. However, the results of mRMR analysis on the MAG data set clearly show the synergy possible when both methods are used. This finding is reinforced by the empirical testing we conducted via machine learning.

3.4 INVESTIGATION OF mRMR RESULTS

Since the mRMR analysis resulted in minimal to no selection of the inverse physics model attributes. Specifically, the very attributes (e.g. β1, β2, β3) that the industry relies on for MEC discrimination, were deemed not relevant for MEC discrimination when in the presence of the computational geometric feature set. We conducted to further understand this unexpected finding.

3.4.1 EMI mRMR Results Analysis

Analysis of the combined EMI data set revealed that the only inverse fitted physics-derived variable in the top 50 rank-ordered set was "Chi^2," which was ranked 45th out of 50 for variable importance. Forty-nine of the features in the top 50 were computational geometric (CG) features. The eigenvalue attributes as described by the inverse physics modeling (β1, β2, β3) – did not appear in the list of top 50 features.

Since the physically based eigenvalue attributes were not ranked with higher priority, we tested whether or not the computational geometric attributes and the eigenvalues were *information content redundant*. The results of our analysis show that they are, in fact, redundant. To assess to the extent of the redundancy, we developed a function using a common set of eight attributes from the computational geometric data set that explains more than 99% of the variation in each of the eigenvalues β1, β2, β3. *Hence, the features developed as part of the computational geometric attribute formulation contain all the information that the eigenvalues have to offer.* This is demonstrated via a regression analysis using the Multivariate Adaptive Regression Splines (MARS) algorithm (Salford 2011) with 10 times cross-validation Thus, the need to develop attributes by fitting physics

models to the field data is unnecessary, at least in this example. Since the computational geometric approach performs at lower S/N than the inverse physics modeling (10 vs. 100, respectively), more TOI can be discriminated using this method. Moreover, the eight features that form the inputs to the regression functions are those that one would expect such as peak values, ratios between the channels and parameters of power law fits. The results of the computational geometric attribute data set's ability to reproduce all of the fitted physics-based derived attributes are shown in Table 3-2: R^2 denotes the correlation coefficient, which explains the percentage of the variation in the variable captured by the function.

Table 3-2. Reproducibility of Inversion Physics-Based EMI Features using CG-EMI Features

Physics-based parameters obtained by inversion	R^2 obtained using 10 times cross-validation
β1	0.993
β2	0.993
β3	0.993
Chi^2	0.897
Size	0.981
Depth	0.994
Coherence	0.766

In hindsight, it is not surprising that the computational geometric approach includes all of the information that could be available by fitting physics models to the data. After all, we developed the computational geometric approach with the discrimination physics in mind. However, this is the first formal analysis that indicates that this information inclusivity is indeed the case. Moreover, these results show that the computational geometric approach can be used to develop a physics-based representation from the EMI Data. Interestingly, the one attribute

that did appear in the top 50 features, Chi^2 is a solid indicator of how well the inverse physics model is expected to fit the data (Chi^2 a measure of fitness of the inverse fitted-physics model).

It is also important to note that the computation geometric approach was able capture 89.7% of the variation in the expected fitness of the inverse modeling. This ability to predict *a priori* how an inverse modeling task should perform is extremely valuable for quality assurance/quality control purposes. Hence, from an information-theoretic point of view, the SME/physically based inputs are redundant in this case. Referring back to the Venn diagram of Figure 3.1, two of the three informational dimension circles (SME and physically based models) would be a subset of the information content represented by only the data-driven circle (visually, they would be completely *inside* the data-driven circle).

3.4.2 MAG mRMR Results Analysis

Analysis of the MAG data revealed three of the inverse physics-derived variables in the top 50 of the rank-ordered set; these are *Fit_size (rank #1)* which is the size of the TOI item , *Fit_inc (rank #6)* which is the TOI inclination, *and Fit_Depth (rank #48) which is the depth of the TOI below ground.* The remaining 47 of the features in the top 50 were computational geometric attributes. A test of the ability to produce the fitted physics-based attributes from the computational geometric attributes was conducted, this time with very different results as shown in Table 3-3.

Table 3-3. Reproducibility of Fitted Physics-based MAG Features Using Computational Geometric MAG Attributes

Physics-based parameters obtained by model inversion (# is the parameter ranking)	Function fitness (R^2)obtained using 10 times cross-validation
Depth (#48)	0.67
Size (#1)	0.73
Dec	0.21
Inc (#6)	0.49
Solid Angle	0.30
Magnetic Moment	0.49

Referring to Figure 3.1, the physically based and data driven circles have overlap, but not virtually complete overlap as was evidenced by the analysis of the EMI data. Clearly, the less well-developed computational geometric approach for MAG sensors is currently not as effective as the EMI approach in capturing the information content from the fitted physics-based inversion model; therefore, further work in this area is warranted.

3.5 mRMR ANALYSIS SUMMARY

This mRMR approach is particularly valuable because it provides the ability to screen important features and reject ones of lesser value or that are redundant to make classification predictions *without the need to run classification algorithms*. This means that important variables can be *identified in minutes* as opposed to hours or days of computation time. Thus the benefits associated with the *machine-learning, algorithm-independent analysis* of feature contribution made possible with the mRMR approach are multifold. Not only is it fast and cost-efficient, it guides when easily computed data-driven features should replace more complex ones to obtain features such as those arrived at via fitted physics-

based inversion. Additionally, it provides a very fast and efficient screening mechanism to rank the value of new or proposed features, especially when compared to existing feature sets. Additionally, these characteristics of the information-theoretic mRMR approach, when corroborated with results from machine-learning algorithms, effectively streamline the understanding of attribute importance and help to focus new research into less well-understood areas. This benefit is discussed in more detail below.

Given the prospect that the next generation of geophysical instruments will produce even more data and resultant features, the industry would benefit from an efficient and reproducible site-specific feature reduction methodology—which is precisely the role the information-theoretic approach mRMR would serve.

4.0 MACHINE LEARNING ANALYSIS AND RESULTS

Machine-learning (ML) techniques are tools that interrogate the information content in the data set and then replace that content with a representative relation(s). The machine-developed representations are known both as “functions” or “models”. That representation can then be used to make predictions relative to unseen instances: in this case sensor data returned from a geophysical investigation in which subsurface MEC items may be present.

Based on the information-theoretic mRMR analysis outlined and demonstrated above, we can anticipate and expect certain outcomes when building functions from the data sets using machine-learning algorithms and various combinations of attributes. For example, functions produced using the EMI data set should rank as:

- Best: Combined geometric and fitted physics model attributes;
- Second: Geometric attributes, and;
- Third: Fitted physics-attributes.

This ranking reflects the fact that the computational geometric approach *replicated the information content in the inverted fitted physics models.* The machine-learned functions based on the combined geometric-fitted physics attribute data may be slightly better (or tie with) the geometric attribute approach, since only one physics attribute appeared in the top 50 features (the measure of the inverse physics model fitness) and then at a very low rank (#45). The data based on the fitted physics models will rank as third accurate, the loss of accuracy being

representative of the extent that the inverse physics approach does not contain the information content that the geometric data set provides.

Discrimination results produced using the MAG data set are a different story. Clearly, the geometric attributes present valuable information, as do the fitted physics-inversion attributes. One can only conclude, therefore, that the combined CG-physics data set will produce a more accurate discrimination function than either data or physics alone.

4.1 EMPIRICAL TESTING USING MACHINE LEARNING

The discrimination functions were constructed from the EMI and MAG data sets (fitted physics, geometric, geometric-fitted physics). In the initial paper (Deschaine et al. 2011), all functions were developed using 10 times cross-validation, and all used the designated technique subset (not just the subset of the top 50 features identified above). The tool used was TreeNET (Salford et al. 2011) and was used with default settings, except the number of trees was set to 2,000. The results matched expectations. To test whether the results may have been influenced by the machine learning method, in this paper we use a different machine learning method – Rotation Forrest. Rotation Forrest is a combination of principal component analysis and the J48 decision tree algorithm as available in the WEKA open source machine learning software (Hall et al. 2009). Ten times cross validation was also used in the classification function development, and the default algorithm parameters used.

The results are presented both in the form of a receiver operating characteristic (ROC) chart. A ROC chart is one that plots the accuracy of a

classifier over the data set; true positive rate of detection on the y-axis and the false positive rate on the x-axis (Hanley J and McNeil B 1982). The ROC area under curve (AUC) is used as a measure of quality of a probabilistic classifier, with an area of 1.0 being best achievable, a 45 degree angled line is no better than random guessing. The graph shows the order of MEC removal, progressing from left to right, with the final excavation occurring at the right-most section of the graph that intersects with the y-axis equal to one. Since the last remaining MEC item is removed when the value of the y-axis is 1.0 there are two important characteristics when evaluating MEC discrimination results: the AUC is important for general classifier testing (higher is better) but the number of ranked data points it takes to reach the maximum of the y-axis relates specifically to the efficiency of the MEC remediation. The faster the top of the y-axis is reached, the more effective the algorithm. For MEC removal projects, additional MEC would be removed beyond the last known MEC item as a means of validation that the MEC has been removed.

4.1.1 EMI Machine Learning Results

The results using the EMI inverse physics model data set is provided in Figure 4-1 and show respectable MEC discrimination (ROC > 0.95), however the last MEC is not identified until 30% through the TOIs.

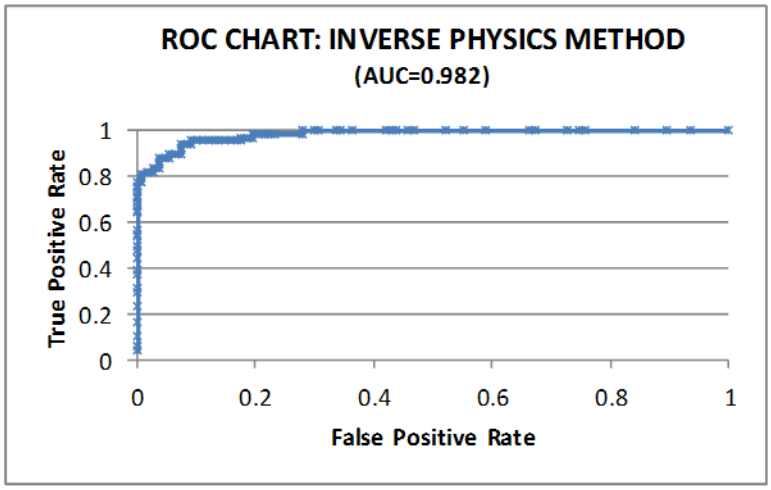

Figure 4-1. EMI: Inverse Physics Model; ROC (AUC) =0.982

The discrimination function results using the EMI geometric data set (provided in Figure 4-2) also show respectable MEC discrimination (ROC > 0.95) with a much better time/speed curve for identifying MEC of about 5% through the TOIs.

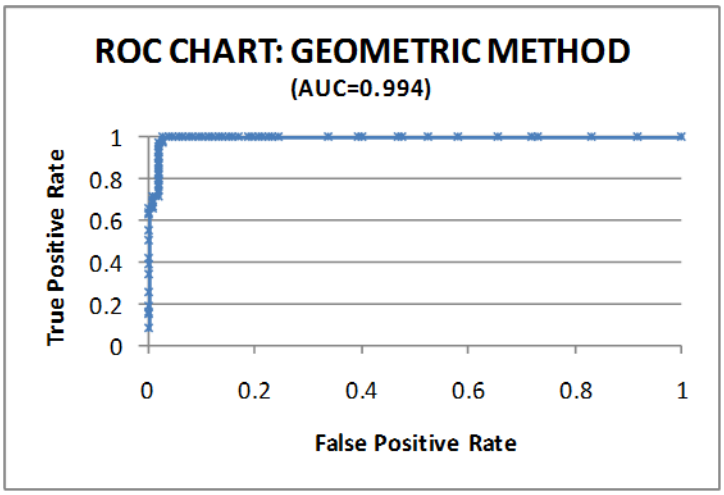

Figure 4-2. EMI: Computational Geometry; ROC (AUC) =0.994

The discrimination function results using the combined EMI-geometric and inverse physics model data set provided in Figure 4-3 also show respectable MEC discrimination (ROC > 0.95), again with a much better time/speed curve for identifying MEC of about 5% through the TOIs.

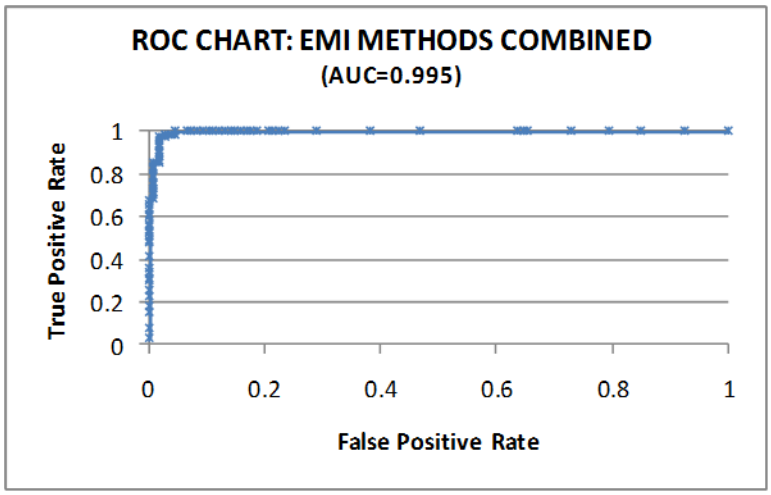

Figure 4-3. EMI: Combined; ROC (Area AUC) =0.995

The expectation of the classifier performance is in concert with the understanding gained from the information-theoretic mRMR analysis. The computational geometric approach performed better than the inverse physics model, because it replicates basically all the important information content of the fitted physics model and also generated additional information essential for higher accuracy MEC classification. The combined geometric-fitted physics discrimination function slightly outperformed the geometric-only discrimination function.

4.1.2 MAG Machine Learning Results

The discrimination function results using the MAG fitted physics data set, provided in Figure 4-4, show respectable MEC discrimination (ROC > 0.95); however the last MEC is not identified until 35% through the TOIs.

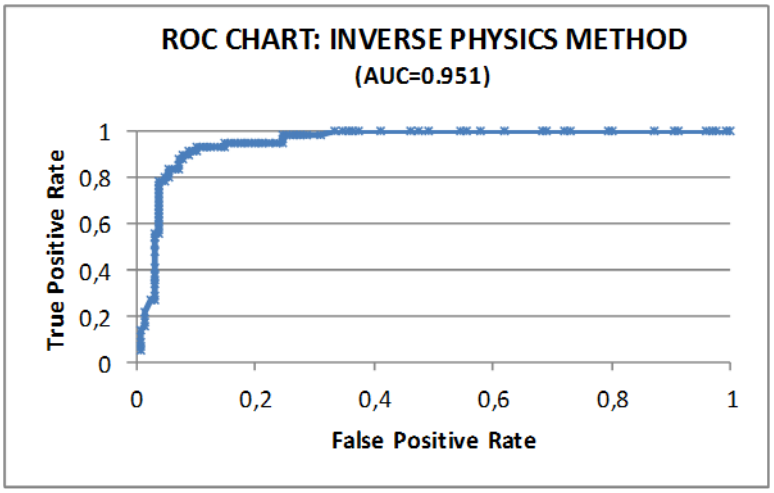

Figure 4-4. MAG: Inverse Physics Model; ROC (AUC) =0.951

The discrimination function results using the MAG geometric data set provided in Figure 4-5 also show respectable MEC discrimination (ROC > 0.95); again, however the last MEC is not identified until 35% through the TOIs.

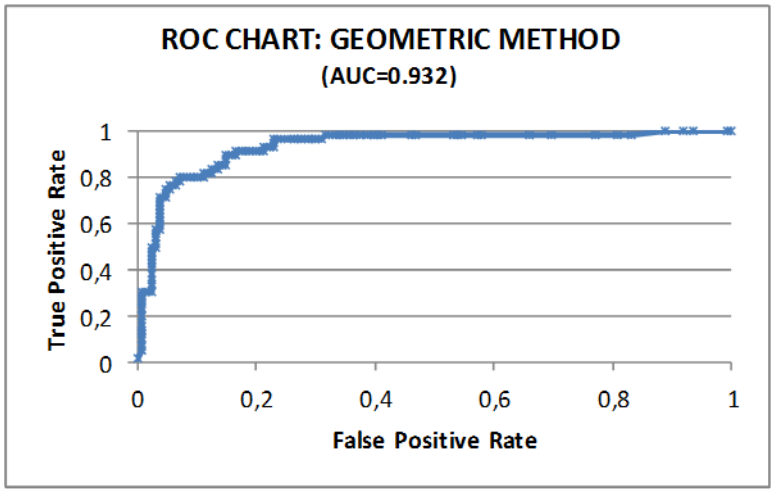

Figure 4-5. MAG: Computational Geometry, ROC (AUC) =0.932

The discrimination function results using the combined MAG geometric and inverse physics model data set (shown in Figure 4-6) also demonstrate respectable MEC discrimination (ROC > 0.95), with a higher AUC ROC value with a slightly better curve than the individual MAG data set analyses, but even so it indicates a slower identification of the final MEC found when compared to the EMI instrument results.

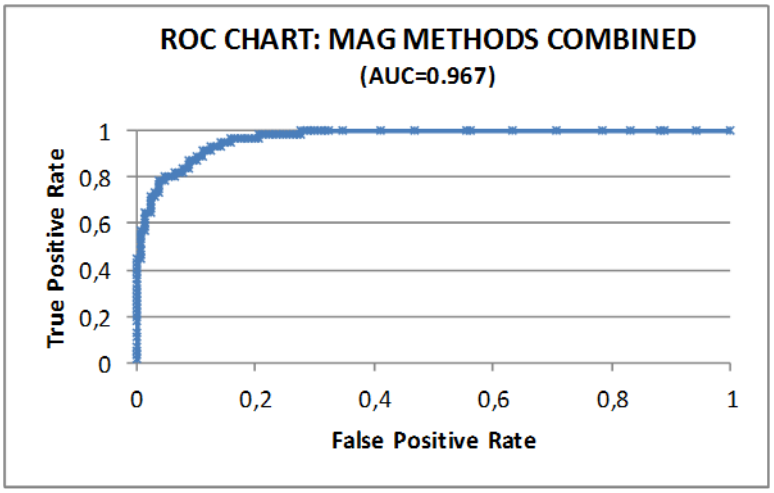

Figure 4-6. MAG: Computational Geometry and Inverse Physics Model, ROC (AUC) =0.967

The expectations of the classifier performance are in concert with the information-theoretic mRMR analysis in terms of overall performance (a higher AUC ROC value was obtained using the combined data-physics data sets). However, the overall identification of that last MEC was slower; hence this solution would require more holes to be dug (and non-MEC items excavated) than would be needed using the EMI geophysical instrument.

5.0 SUMMARY AND RESULTS

We demonstrate the value and understandability of the computational geometric MEC discrimination approach, and developed an understanding of the value of MEC features by applying information theory. We used machine learning to fuse the information content of attributes derived from both the machine-learning computational geometric approach and from fitted physics-based models.

While analyzing the data sets for information value via mRMR, we found that the SME/physically based information content was completely represented by the computational geometric approach. Further examination revealed that a common set of eight of the data-driven features contained all the information in the SME/physically based information. Hence, from an information-theoretic point of view, the SME/physically based inputs were redundant in this case. Referring back to the Venn diagram of Figure 3-1, two of the three informational dimension circles (SME and physically based models) would be a subset of the information content represented by only the geometric (data-driven) circle (visually, they would be completely *inside* the data-driven circle). As validation – using machine learning as the integrating approach - acceptable discrimination results were produced by implementing the MEC versus not-MEC classification analysis using only the data-driven information content. Combining the SME/physically based and data-driven information did not change the MEC classification accuracy (to three decimal places). In contrast, using only the physically based approach would have resulted in about 25% of the MEC

anomalies being excavated, while the data driven approach would require less than 5%, a factor if five different.

We believe that the inverse physics modeling, while providing great insight, over compressed the information available in the geophysical signals into too few variables in this case and hence impose an artificial limit on that methods accuracy. However, the computational geometric approach is intended to extend – not replace - this deep physics-based understanding by supplementing the discrimination information with factors the inverse physics modeling approach cannot capture. For the Camp Sibert test case, the computational geometric approach was found to contain all the information (in eight common variables) contained in the EMI data set developed by physics model inversion, but not the MAG data set. Of additional note, while the MEC discrimination industry standard geophysical instrument uses 4 time gates (EMI), several vendors are developing instruments with 40 time gates (ESTCP 2011). Preliminary results indicate achievement of as high discrimination accuracy as we achieved using the 20-time gate Protem 47 from the JPG-IV analysis (Figure 2-3). These newer instruments, combined with the continual advancement in discrimination algorithms, so great promise in obtaining the goal of high accuracy MEC discrimination using non-destructive electromagnetic geophysical sensing.

The empirical tests conducted using machine-learning are consistent with their performance predicted using information theory. We further identified the information overlap between our computational geometric approach and the fitted physics model approach developed and by others: complete overlap for the EMI

sensor at this site – indicating a rational physical basis for the method - and a partial overlap for the MAG sensors.

6.0 REFERENCES

Banks RE, Núñez E, Agarwal P, Owens C, McBride M, and Liedel R (2005). Genetic Programming for Discrimination of Buried Unexploded Ordnance (UXO). Late-breaking paper at The Genetic and Evolutionary Computation Conference (GECCO-2005)

Bell TH, Barrow BJ, and Miller JT (2001). Subsurface Discrimination Using Electromagnetic Sensors, IEEE Transactions on Geoscience and Remote Sensing, Vol. 39, No. 6, June, 2001.

Deschaine LM Keiswetter D and Francone FD (2009). Advanced MEC Discrimination Comparative Study on Standardized Test-Site Data Using Linear Genetic Programming (LGP) Discrimination (MM-0811 Fact Sheet) completed 2009.

Deschaine LM, Hoover RA, Skibinski JN, Patel JJ, Francone FD, Nordin P, and Ades MJ (2002). Using Machine Learning to Compliment and Extend the Accuracy of UXO Discrimination Beyond the Best Reported Results of the Jefferson Proving Ground. Technology Demonstration, pages 46-52. Society for Modeling and Simulation International's Advanced Technology Simulation Conference, San Diego, CA April 2002.

Deschaine L, Nordin JP, and Pintér JD (2011). A Computational Geometric / Information Theoretic Method To Invert Physics-Based MEC Models Attributes For MEC Discrimination, Journal of Mathematical Machines and Systems, National Academy of Sciences of Ukraine, Kiev. No 2, Pages 50-61.

Edelsbrunner H and Harer JL (2009). Computational Topology, American Mathematical Society (2009), 241 pages.

Environmental Security Technology Certification Program (ESTCP) (2009). Technical Report Description and Features of UX-Analyze ESTCP Project MM-0210, 2009, 42pp.

ESTCP (2011). Hand-on Course On Classification Methods Applied To Munitions Response. January 19-20, 2011, Washington, DC.

Francone FD, Deschaine LM, Battenhouse T and Warren JJ (2004). Discrimination of Unexploded Ordnance from Clutter Using Linear Genetic Programming, Proceedings of the Genetic and Evolutionary Computation Conference, Late Breaking Papers, 2004, Seattle, WA, USA.

Francone FD, Deschaine LD, Battenhouse T and Warren JJ (2005). Discrimination of Unexploded Ordnance from Clutter using Linear Genetic Programming. In Tina Yu and Rick L. Riolo and Bill Worzel editors, Genetic Programming Theory and Practice III, Volume 9 of Genetic Programming, Chapter 4, pages 49-64. Ann Arbor, 2005.

Francone FD, Deschaine LM, and Warren JJ (2007). Discrimination of Munitions and Explosives of Concern at F.E.Warren AFB, GECCO, 2007 (London)

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, and Witten IH (2009). The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

Hanley J and McNeil B (1982). The Meaning and Use of the Area under a Receiver Operator Characteristic (ROC) Curve," Radiology, Vol. 143, pp. 29-36, 1982.

Keiswetter D (2008). SAIC Analysis of Survey Data Acquired at Camp Sibert, Interim Report, ESTCP Project MM-0210, July, 2008. 112 pages.

Nordin JP (1994). A Compiling Genetic Programming System that Directly Manipulates the Machine Code. In: Advances in Genetic Programming, K. Kinnear, Jr. (ed.), MIT Press, Cambridge MA. Pages 311-331.

Novikov SP and Fomenko AT (1990). Basic Elements of Differential Geometry and Topology (Mathematics and its Applications) Springer, 1990 (reprinted 2010), 504 pages.

Peng H, Long F, and Ding C (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8, pp.1226-1238, 2005.

Pintér JD (1996). Global Optimization in Action. Kluwer Academic Publishers, Dordrecht, 1996. Now distributed by Springer Science and Business Media, New York. 512 pages.

Salford Systems Inc. (2011). Salford Data Miner User's Manual: CART Version 6.4, TreeNET Version 2.0, MARS Version 3.0, and Random Forrest Version 1.0, 2011, San Diego, CA.

Robitaille G, Adams J, O'Donnell C, and Burr P (1999). Jefferson Proving Ground Technology Demonstration Program Summary.

Welzl E (1991). Smallest enclosing disks (balls and ellipsoids) in New Results and New Trends in Computer Science, Lecture Notes in Computer Science, Vol. 555 (1991), 359-370

Varmuza K, (1974). Monatsh. Chem. 105, 1.