Data mining to improve management and reduce costs of environmental remediation

Dara M. Farrell, Barbara S. Minsker, David Tcheng, Duane Searsmith, Jane Bohn and Dennis Beckman

ABSTRACT

In this paper, data from 105 soil and groundwater remediation projects at BP gasoline service stations located in the state of Illinois were mined for lessons to reduce cost and improve management of remediation sites. Data mining software called D2K was used to train decision tree, stepwise linear regression and instance-based weighting models that relate hydrogeologic, sociopolitical, temporal and remedial factors in the site closure reports to remediation cost. The most important factors influencing cost were found to be the amount of soil excavated and the number of groundwater monitoring wells installed, suggesting that better management of excavation and well placement could result in significant cost savings. The best model for predicting cost classes (low, medium and high cost) was the decision tree, which had a prediction accuracy of approximately 73%. The misclassification of approximately 27% of the sites by even the best model suggests that remediation costs at service stations are influenced by other site-specific factors that may be difficult to accurately predict in advance.

Key words | D2K, data extraction, data mining, decision trees, excavation, remediation

Dara M. Farrell
The Power Generation Company of Trinidad and Tobago,
6A Queen's Park West, Port of Spain,
Trinidad
Tel.: +1 868 720 4929
E-mail: *d.m.farrell@gmail.com*

Barbara S. Minsker (corresponding author)
Department of Civil and Environmental Engineering,
University of Illinois,
3230 Newmark Lab, MC-250,
205 N. Mathews Avenue, Urbana IL 61801,
USA
Tel.:+1 217 333 9017
Fax: +1 217 333 6968
E-mail: *minsker@uiuc.edu*

David Tcheng
Duane Searsmith
National Center for Supercomputing Applications (NCSA),
University of Illinois at Urbana-Champaign,
MC-476, 605 E. Springfield Avenue
Champaign IL 61820,
USA

Jane Bohn
Atlantic Richfield Company (a BP affiliated company),
28100 Torch Parkway, Warrenville IL 60555,
USA

Dennis Beckman
Remediation Engineering and Technology,
BP Corp North America Inc,
501 Westlake Park Boulevard, Houston TX 77079,
USA

INTRODUCTION

Federal and state regulations in the United States require that leaks from underground storage tanks (USTs) be reported and remediated. The federal rules were promulgated in 1984 under Subtitle I of the Hazardous and Solid Waste Amendments (HSWA) to the Resource Conservation and Recovery Act (RCRA - 40 CFR Part 280). In general, these rules are administered under state programs. Oil companies can incur sizable penalties as a result of violation of environmental regulations; emissions from leaking underground storage tanks into the atmosphere and the release of hydrocarbons into lakes and rivers can result in fines being levied against companies. Additional expense is incurred if settlements are necessary because of migration onto non-company-owned properties. In 2003, the US-based branch of BP (formerly British Petroleum) reported an expenditure of $5.6 million in fines due to alleged underground storage tank and waste

doi: 10.2166/hydro.2007.004

management violations as well as settlements to governmental organizations and members of the public.

This study investigates the use of data mining, specifically text mining, for improving management and reducing costs associated with remediating gasoline station sites in the US. Better management of clean-up procedures and reducing the costs associated with liabilities allows businesses to be more competitive on a worldwide scale and potentially be more proactive in addressing environmental problems associated with their operations. An understanding of which factors affect cost is critical if this is to be achieved.

Data mining is one step in the process of knowledge discovery in databases (KDD). KDD involves:

- data cleaning (where inconsistent data are removed),
- data integration (where multiple data sources may be combined),
- data selection (the retrieval of relevant data),
- data transformation (where data are converted into a form suitable for mining),
- data mining (the application of statistical methods in order to discover patterns in the data),
- pattern evaluation (the identification of interesting patterns), and
- knowledge presentation (the use of various visualization and knowledge presentation techniques).

In a broader sense, however, data mining can be defined as the process of discovering interesting patterns from large amounts of data (Han & Kamber 2001).

Data mining is a powerful approach for the analysis of trends when the quantity of data is large; with hundreds of BP service station sites in Illinois (and thousands more globally), the amount of data for analysis can become overwhelming. In this study, a manageable subset of the available data is used to investigate whether data mining can be used to recognize patterns and interesting phenomena that may lead to better management of these sites in the future, focusing particularly on factors that influence the remediation cost of sites. Future work can then extend the study to a broader set of service stations.

To the authors' knowledge, no previous work has been done on this particular topic, although the use of data mining for other environmental applications has been explored. Michael *et al.* (2005) evaluated the use of different data mining methods to more effectively combine different types of data and models for predicting hydraulic heads. Su *et al.* (2002) used a data mining approach to investigate the relationship between environmental factors and the distribution pattern of living organisms. More recently, Bessler *et al.* (2003), as well as Anderton *et al.* (2004), applied decision trees to water resources problems.

METHODOLOGY

The first step in the methodology involved working with BP staff and consultants from Delta Environmental Consultants Inc. (BP's environmental management consultants who managed the sites in this study) to identify which features and sites should be investigated. This process was iterative, as shown in Figure 1, with these experts also providing ongoing interpretation and evaluation of initial results and identifying additional factors affecting remediation cost ("features") that should be included in the analysis. The next section discusses this aspect of the study.

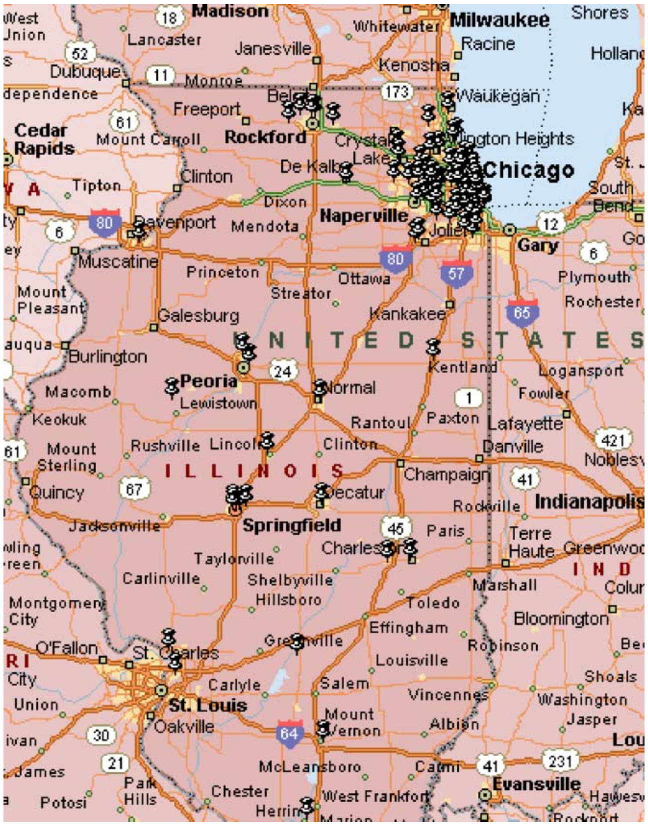

Figure 1 | Site locations.

Once appropriate features and sites were identified, data mining software called D2K was used to select, transform, and analyze the data. Three different models were considered: decision trees, stepwise linear regression and instance-based weighting. A later section gives an overview of D2K and the following subsections discuss the models within D2K that were used for this work.

Data collection and site selection

Table 1 shows the features (attributes) that were chosen for analysis following consultation with BP management and Delta Environmental consultants. Features have been grouped into categories of cost, time, hydrogeologic characteristics, contaminant characterization, remediation approach and political/social/legal characteristics.

The reader may be familiar with most of the features listed; the less familiar terms are explained in the following paragraphs. Once all program requirements and remediation objectives have been satisfied, a “no further remediation” letter is issued by the state Environmental Protection Agency (EPA) and no further corrective action is required by the company. The feature, “Time until the EPA granted no further remediation (NFR) status” is the time until the “no further remediation” letter is received by BP. The groundwater classification of a site is a factor in determining the level of remediation required for a site. The “Hydrogeologic characteristics” category takes this into account with the feature “class of groundwater”. The three classes of groundwater, as defined by Illinois regulation, are:

Class I – potable resource groundwater,
Class II – general resource groundwater,
Class III – special resource groundwater.

The features in the “Political/legal/social characteristics” category attempt to capture factors related to the sociopolitical status of the site, such as what types of institutional controls (ICs) are granted and through which body they are obtained. Institutional controls are legal and administrative means of controlling human exposure to residual site contaminants, such as the posting of warning signs and notices as well as the implementation of zoning restrictions. If there are pre-existing agreements for the granting of institutional controls, then there may be cost

Table 1 | List of attributes

Attributes
Cost
Total cost of site remediation from initiation to case closure
Time
Time until EPA granted “no further remediation” (NFR) status
Assessment time
Year of closure
Hydrogeologic characteristics
Was groundwater encountered?
Hydraulic gradient
Hydraulic conductivity
Porosity
Class of groundwater
Contaminant characterization
Was BTEX (benzene, toluene, ethylbenzene, xylenes) a site contaminant?
Were PNA’s (polynuclear aromatics) site contaminants?
Were metals site contaminants?
Was free product documented at the site?
Offsite migration?
Remediation approach
Were remediation technologies used?
Did natural attenuation occur?
Was there excavation?
Amount of soil excavated
Were tanks removed from the site?
Number of tanks removed
Number of geoprobes/borings installed
Were wells installed (remediation and monitoring)?
Number of wells (remediation and monitoring)
Political/cocial/legal characteristics
Municipal/non-municipal
Was an agreement in place between the company and other governing agencies?
Was the site owned by company originally?
Were institutional controls (ICs) applied to groundwater?
Were ICs applied to soil?
Classification of site location (either mixed or commercial)

savings because closure can be obtained more quickly. There may be some differences in the savings achieved if the ICs are granted by a municipal or non-municipal body, such as different requirements in the ICs, and this is captured by the "Municipal/non-municipal" attribute.

This study focuses on BP-owned sites in the US, since company records show that the US branch of its operations contributes heavily to the overall expenditure of the international organization. All sites selected for analysis are service stations located in the state of Illinois whose remediation projects have been closed under consent orders. A consent order is a legally binding agreement between a state enforcement agency (in this case the Illinois EPA) and a company; it lists the remediation activities to be performed while specifying a timeframe for their completion. Sites matching these criteria were randomly selected and information on each site was gathered from closure reports submitted to the Illinois Environmental Protection Agency (IEPA). Some additional data that were recorded regularly as a part of the site documentation were also included. These data included whether or not offsite migration of the plume occurred, all of the time characteristics given in Table 1 and all of the political/legal/social characteristics included in Table 1 except for the classification of site location.

The general distribution of the 105 selected sites is shown in Figure 1. It can be seen that most of the sites are located around the Chicago metropolitan area. This is to be expected because most of BP's service station sites are located in the urban and suburban areas of Chicago. Additionally, because Chicago has an ordinance restricting the installation of water wells, it is easier to get the institutional controls required to obtain closure in this area and it should be expected that a high percentage of the closed sites in Illinois would be located in this area.

D2K

Data selection, transformation, and analysis were completed using D2K – a Java-based machine learning system created by the National Center for Supercomputing Applications (NCSA) (Welge *et al.* 2003). D2K combines analytical data mining methods of prediction, discovery and deviation detection with data and information visualization in a powerful computational infrastructure which enables the use of distributed computing and the running of processor-intensive and data-intensive applications. Its visual programming environment increases its user-friendliness; the user is allowed to create other programs through a graphical environment. Users connect software components ("modules") together to produce "itineraries" (programs) that perform the desired task.

While D2K can be used as a stand-alone application (as in this study), developers can also utilize the D2K infrastructure modules to build applications (employing D2K functionality in the background and allowing for the use of specialized user interfaces).

Users will appreciate the low overhead (time, cost and resources) associated with module execution and the functionality and portability of this Java-based system. Additional information on D2K can be found by referring to Welge *et al.* (2003).

D2K offers many modules that can aid the knowledge discovery process, including modules for naive Bayesian, neural networks, decision trees, and cleaning and transforming of data. This study utilized feature extraction and selection methods, and model fitting and testing modules with decision trees, stepwise linear regression and instance-based weighting modules. The following subsections describe each of these data mining approaches in turn.

Feature extraction and selection

Because of the large volume of data in the closure reports to be processed, D2K's text extraction itinerary (for details see Farrell (2004)) is used to provide semi-automation of the data extraction process. Figure 2 summarizes the steps involved in the data extraction itinerary.

In this algorithm, keywords are specified in one input file and this list, along with the closure reports, was fed into the itinerary. Each closure report is "parsed" or separated into "tokens" (words, or symbols such as commas and parentheses) within the documents. These tokens are then matched with the keywords specified in the input file. The output consists of all occurrences of the keywords along with the five closest words on either side of the occurrence of the keywords; this is called the context and allows the

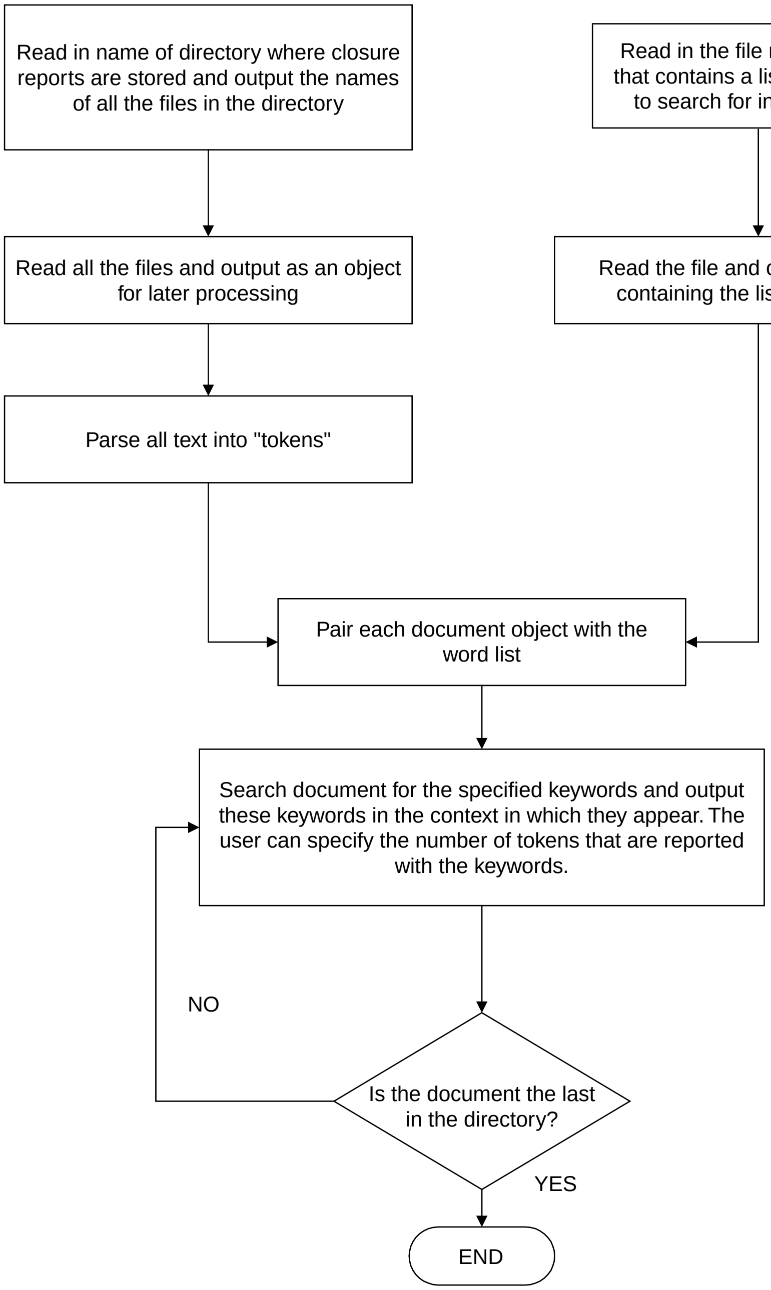
Read in name of directory where closure reports are stored and output the names of all the files in the directory
Read in the file name of a file that contains a list of keywords to search for in the reports
Read all the files and output as an object for later processing
Read the file and output an object containing the list of keywords
Parse all text into "tokens"
Pair each document object with the word list
Search document for the specified keywords and output these keywords in the context in which they appear. The user can specify the number of tokens that are reported with the keywords.
NO
Is the document the last in the directory?
YES
END

Figure 2 | Summary of data extraction process.

user to determine if these keyword occurrences contain valuable information. The user then reads the program's output (the keywords in their context) and uses his/her judgment to determine what data should be input into the training set.

Consider an example: in order to determine the amount of soil excavated at a site, the keyword "excavated" could be entered in the input file. Because the amount of soil could be represented in text form in a number of different ways, the algorithm would search through all the closure reports and display all occurrences of the word "excavated" along with the five closest words on either side. The user would then look at the program's output and identify the amount of soil excavated from the extracted text.

Although this extraction process is only semi-automated, the user still saves time by quickly and simultaneously isolating text sections of interest in a large number of documents.

Features with non-numerical values need to be transformed in order to be analyzed with the data-driven models in D2K. Table 2 shows how the values for each feature are encoded. In the cases where a "yes/no" response is given, this is encoded in binary form, with 1 meaning "yes" and "0" meaning "no". In other cases (for example, for the feature "classification of site location"), the numbers represent a classification.

The training data are then read into an "input file creation itinerary" in D2K (see Farrell (2004) for itinerary details); the user specifies which columns of the training dataset are input (or independent) variables and which are output (or dependent) variables. The data are then stored as a serializable object to facilitate ready access to the data using object-oriented programming.

Model fitting and testing

Three types of models are considered: decision trees (Quinlan 1986), stepwise linear regression and instance-based weighting. The optimal parameter settings for each model were found using a model-fitting itinerary created in D2K. For each type of learning machine, twenty cross-validation experiments are performed (automatically in the D2K itinerary using random search within a user-defined range) to identify good model parameters for the learning machines (i.e. the ones that yield the lowest cross-validation error). The specific model parameters selected vary for each experiment, but ranges are given in the following subsections.

In the cross-validation experiments, 15 of the 105 sites are reserved for testing. The learning machine is then trained to predict cost from the attributes describing the remaining data (90 sites) and tested in its predictions of site remediation cost using the testing set. The cross-validation error is calculated to be the mean of the absolute difference between the predicted values and the actual values in the testing set.

Once the optimal parameters are identified, the entire training set and its optimized learning model parameters are fed into a model builder in the "optimal model itinerary" (see Farrell (2004) for details) and the resulting model is used to predict site costs. The resubstitution error (the sum of the differences between the actual and predicted costs when the model is applied to *all* of the data) is also calculated.

Model predictions are also grouped into classes to test the ability of the models to predict broad classes of cost (high, medium and low) rather than exact costs. The class boundaries were selected based on recommendations from BP staff. It should be noted that this method of class prediction is not the same as training and testing directly on classification error, where the model predicts the probability of each class. Direct class prediction is not currently possible within the data mining itineraries in D2K, so this approach was not investigated here. Classification accuracy is defined here as the percentage of correctly classified sites.

To assess the performance of each model, the resubstitution error is compared to the resubstitution error from the simplest possible model, where the mean cost of all the sites (found by summing the costs for all sites and dividing by the total number of sites) is used as the predicted cost for a given site. The cross-validation errors are also quoted for comparison between methods. To test whether the predictions made by the models are significantly different from simply using the mean as the cost estimation, a t-test is used to test the significance of the predictions.

Decision trees

A decision tree may be described as a flowchart-like tree structure, where each node denotes a test on an attribute and each branch represents the outcome of that test (Han & Kamber 2001). Generally the form can be described as a series of IF–THEN statements, hence the rules obtained by the use of decision trees are often readily understood. Additionally the use of decision trees is advantageous because the learning and classification steps are generally quite fast.

Decision trees are constructed by recursively selecting the most predictive features ("attributes") and splitting the training sets into subsets. Splitting continues until the information in the inputs is exhausted and the terminal nodes are the classification of the final instances (Matheus 1990). In the decision tree each node represents an input and each branch a possible value of that input. The end nodes (or leaves) on the

Table 2 | Data transformation

Attributes	**Encoding**
Cost	
Total cost of site remediation	Numerical value
Time	
Time until EPA granted “no further remediation” (NFR) status	Numerical value
Assessment time	Numerical value
Year of closure	Numerical value
Hydrogeologic characteristics	
GW encountered	Binary
Hydraulic gradient	Numerical value
Hydraulic conductivity	Numerical value
Porosity	Numerical value
Class of groundwater	Numerical value (coded 1,2,3)
Characterization of contamination	
Was BTEX (benzene, toluene, ethylbenzene, xylenes) a site contaminant?	Binary
Were PNA’s (polynuclear aromatics) site contaminants?	Binary
Were metals site contaminants?	Binary
Was free product documented at the site?	Binary
Offsite migration?	Binary
Remediation approach	
Were remediation technologies used?	Binary
Did natural attenuation occur?	Binary
Was there excavation?	Binary
Amount of soil excavated	Numerical value
Were tanks removed from the site?	Binary
Number of tanks removed	Numerical value
Number of geoprobes/borings installed	Numerical value
Were wells installed (remediation and monitoring)?	Binary
Number of wells (remediation and monitoring)	Numerical value
Political/social/legal characteristics	
Municipal/non-municipal	Binary
Was an agreement in place between the company and other governing agencies?	Binary
Was the site owned by company originally?	Binary
Were institutional controls (ICs) applied to groundwater?	Binary
Were ICs applied to soil?	Binary
Classification of site location (either mixed or commercial)	Binary

tree specify the output value for the combination of input values that prescribe the path to that end node.

The decision tree algorithm within D2K (which is of the type first introduced by Tcheng *et al.* (1989)) is summarized in Figure 3. This process was discussed in Michael *et al.* (2005).

A single node is first used to represent all training samples. Then, the predicted output (in this case, remediation cost) is designated as the mean output across all of the training data, $\bar{x}_0$. The initial error, E_0, is calculated using

$$E_0 = \sum_{i=1}^{n} |x_i - \bar{x}_0| \qquad (1)$$

where:

n = total number of training data points,
x_i = actual output value for training data point i, and
$\bar{x}_0$ = predicted output (mean of the training output).

Using the mean of the attribute as the splitting criterion, the data are split into two candidate groups for each of the k attributes being analyzed. The average error associated with each prospective new split, k, is calculated thus:

$$E_k = \left(\frac{n_k}{n}\right)\sum_{i=1}^{n_i} |x_i - \bar{x}_{n_k}| + \left(\frac{n-n_k}{n}\right)\sum_{i=n_k}^{n} |x_i - \bar{x}_{n-n_k}| \qquad (2)$$

where:

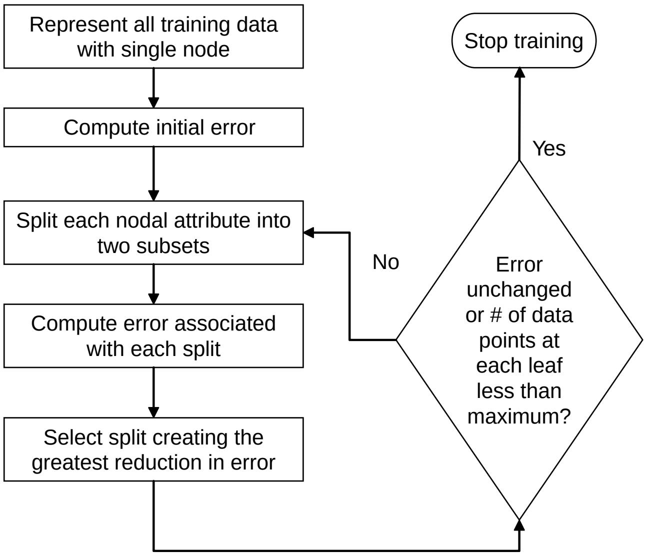
Represent all training data with single node
Compute initial error
Split each nodal attribute into two subsets
Compute error associated with each split
Select split creating the greatest reduction in error
Error unchanged or # of data points at each leaf less than maximum?
No
Yes
Stop training

Figure 3 | Decision tree training process (from Michael *et al.* 2005).

n_k = number of training points in the first group associated with split k,
$\bar{x}_{n_k}$ = mean output for the first group associated with split k, and
$\bar{x}_{n-n_k}$ = mean output for the second group associated with split k.

In Equation (2), the group errors are weighted by the fraction of the population in each group. The algorithm uses the calculated errors for each split, k, to choose the split with the greatest error reduction when compared to the previous error. For the first split, the reuction in error for each split, ΔE_k, is

$$\Delta E_k = E_0 - E_k. \qquad (3)$$

The split with the greatest reduction in error is then chosen for that node. This process is repeated at each leaf (end node) until there is no further reduction in error or when splitting the node would result in less than the minimum number of examples per leaf specified by the user (Michael *et al.* 2005). This parameter (minimum number of training data at the end of each leaf) is selected from a range between 1 and 50 during the training experiments.

The decision tree algorithm described above is a greedy algorithm: each split is chosen based solely on the greatest reduction in error for only that split – a “standard practice” for decision tree formulation. The algorithm does not seek to ensure that the overall prediction error is minimized; however, this method does allow for rapid formation of the decision tree (Michael *et al.* 2005).

Stepwise linear regression

Stepwise linear regression is a modelling technique to develop an optimal linear equation for the prediction of a dependent variable (in this case remediation cost) from several independent variables. The basic procedure involves iteratively adding or removing attributes to an initial model (“stepping”), in accordance with the user’s “stepping” criteria.

In this study step-up regression was used: features were iteratively added to a model containing only one feature. Models were allowed to range in size from one feature (a linear equation with only one independent variable and one dependent variable – cost) to a maximum of eight features (eight independent variables and one dependent variable – cost).

Instance-based weighting

An instance-based weighting model is trained on the input data and then predicts based on the feature values of these stored cases. When asked to make a prediction for a test case (in this case, a particular site), the values of the features in the test case are compared with the stored values from the training set and the degree of match is computed. The value of the new prediction is a weighted average of the "nearest neighbours" (the closest points in the dataset based on the distance metric chosen by the user). This is known as instance-based learning.

A weighting coefficient (factor) is used to represent the importance of each feature to the match. Inverse distance weighting was chosen because it makes sense that features closer to the input should have more impact on the prediction. The weighting factors are calculated as

$$w_i = \frac{1}{d_i} \tag{4}$$

where the distance d_i is the Euclidean distance between the desired point and the ith closest stored point. Instance-based weighting is quite fast and can learn complex functions. The Euclidean distance is calculated as

$$d_i = \left(\frac{\Delta x_1^2 + \Delta x_2^2 + \cdots + \Delta x_m^2}{m} \right)^{1/2} \tag{5}$$

where m is the number of inputs and Δx_m is the difference between the values of the mth input and its nearest neighbour. The number of neighbours, n, is allowed to range from a minimum of 1 to a maximum of 100 and is selected by the learning machine to minimize the cross-validation error on the training data.

RESULTS

This section presents the results from the conducted experiments; discussion of the findings can be found in the next section. A summary of performance of the three methods is given below. The next two subections then present the optimal models found using decision trees and stepwise linear regression, respectively, and discuss the implications of the features included in those models. Instance-based models do not have explicit representations of features. As such, they are not as useful for providing insights on the factors that influence site cost and will not be discussed in detail in the later subsections.

Experiments were performed for two feature sets. Feature set 1 contained all of the attributes documented in Table 1, while feature set 2 omitted the "Time" attributes. These attributes, such as year of closure, would not be known in advance of completion of the remediation. Hence, the models made using feature set 2 would be most useful for making cost predictions for estimating future liabilities, while those made with feature set 1 represent the best possible case of full knowledge of all information in the closure reports.

The cross-validation error (the mean of the absolute difference between the predicted values and the actual values in the testing set) results for each model are summarized in Figures 4 and 5. For all models (the decision tree, linear regression and instance-based models), the cross-validation error is always lower than the simplest prediction model (i.e. when the mean is used as the predicted cost), indicating that all of the models provide more accurate estimates than simply using the mean cost per site. The decision tree and stepwise linear regression models have the lowest cross-validation errors.

Figures 6 and 7 show the resubstitution error (the sum of the differences between the actual and predicted costs when the model is applied to all of the data, expressed as a percentage of the actual cost) for each model tested. These charts show the percentile errors (line graphs), including

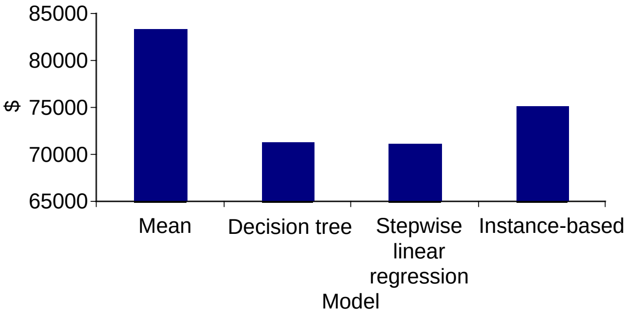

Figure 4 | Cross-validation errors for feature set 1.

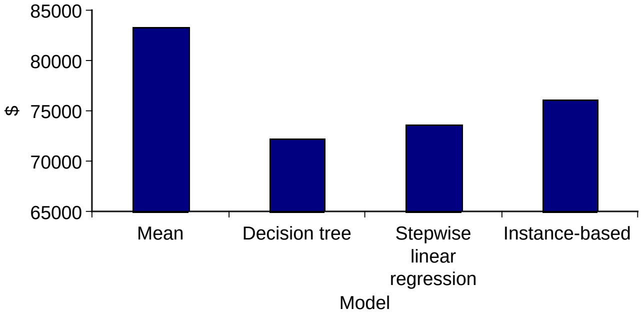

Figure 5 | Cross-validation errors for feature set 2 (no time attributes).

the 5th, 50th and 95th percentiles, as well as the mean errors (bars). The decision tree model performs the best, but still has relatively high mean and 95th percentile errors given that the costs for the sites can range from approximately $13 000 to more than $650 000. For a low cost site (around the $13 000 range), an error of 41.8% (the average percentage error for the decision tree model under feature set 2 and the lowest average error of all the models used in this study) represents a large margin of error.

The models were able to achieve more success in predicting cost classes (given in Table 3) rather than absolute costs. The classification accuracies of the models are compared for feature set 1 and feature set 2 in Figure 8; the decision tree and the stepwise linear regression models had the highest classification accuracies. Since the results from class prediction are better, details in the following sections will focus primarily on the class prediction models.

For the decision tree, the cross-validation and resubstitution errors obtained when using feature set 2 are lower and the classification accuracy is higher than when using feature set 1; the opposite is observed for both the stepwise linear regression and the instance-based models. Since the models are fit using greedy approaches, global optimality is not guaranteed and it is possible that a better model may exist for feature set 1. It is also possible that the time attributes included in feature set 1 were redundant relative to other features. Additional experiments were performed to determine whether removing other features that were less correlated to cost (based on correlations found using one attribute at a time) improved predictions further; no improved trees were found.

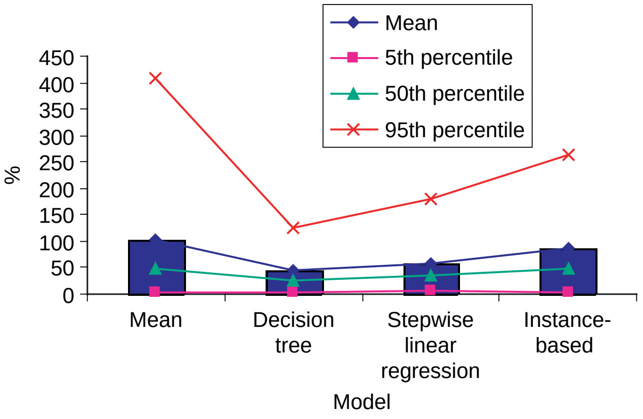

Figure 6 | Summary of cost prediction results using feature set 1.

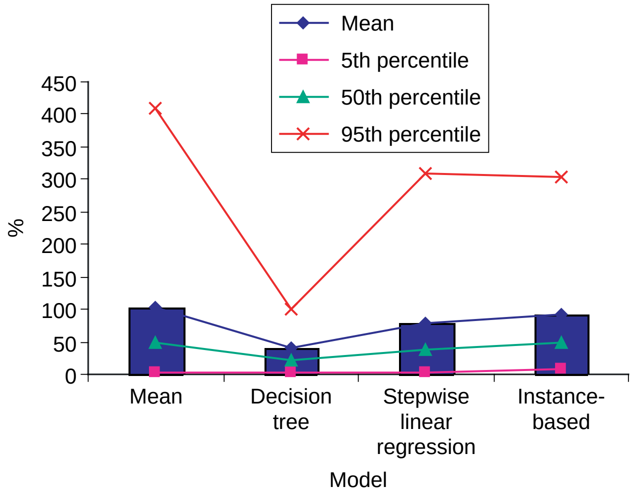

Figure 7 | Summary of cost prediction results using feature set 2 (no time attributes).

The results from the *t*-tests, which show the confidence levels at which the predictions of each model are significantly different from the simplest model (the mean), are shown in Table 4. The *t*-test experiments showed that, in all cases, the predictions from each model were significantly different from the mean at confidence levels over 95%.

Table 3 | Cost classes

Symbol	Classification
⊗	High cost (>$250 000)
✙	Medium cost (>$100 000 and <$250 000)
☆	Low cost (<$100 000)

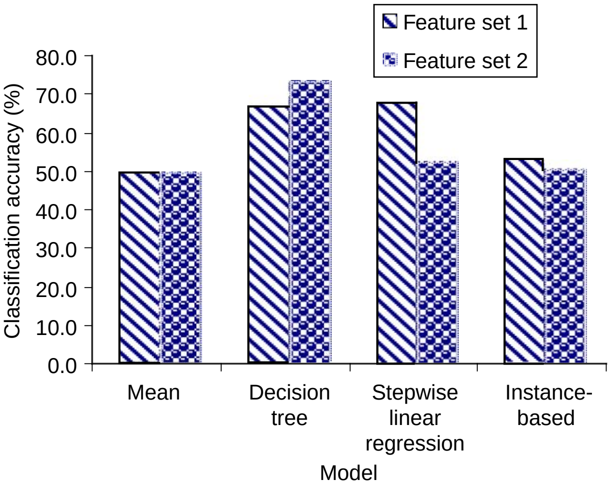

Figure 8 | Classification accuracies for all models and both feature sets.

Decision tree results

Figure 9 shows the decision tree obtained when all of the attributes were included (feature set 1). As discussed previously, the branch level on the decision tree indicates the importance or influence of the feature – the higher the branch level, the more important the attribute is to the model. Hence, the most important attributes are the amount of soil excavated and the time until no further remediation (NFR) is required. Also important are whether remediation technology is applied to the site and the classification of the site location (i.e. commercial or mixed). In Figure 9, the classification of cost is noted by the symbols given in Table 3.

The decision tree given in Figure 9 represents the best possible case, where all information in the closure reports is known. However, at the beginning of remediation, when estimates of cost liabilities are most needed, it is difficult to know when no further remediation will be achieved and it is for this reason that this attribute and others in the Time category (as shown in Table 1) were omitted from feature set 2. The tree that resulted when these features were removed is shown in Figure 10.

Table 4 | Confidence levels from *t*-test experiments

	Decision tree	Stepwise linear regression	Instance-based
Feature set 1	99.1%	97.8%	99.96%
Feature set 2 (no time attributes)	99.1%	97.1%	99.96%

The attribute "the amount of soil excavated" remains very prominent in the tree, appearing not only at the top branch level, but also at the second, third and fourth levels. However the number of wells (both remediation and monitoring) has replaced the attribute "Time until NFR". Whether or not remediation technology is used at the site remains an important factor in the model.

Stepwise linear regression results

As discussed previously, in step-up linear regression features are successively added to an initial model to create an optimal linear equation to predict a dependent variable (in this case, cost) from several independent variables (for example, the amount of soil excavated, the number of geoprobes, the number of wells, etc.). Up to four features, shown in Table 5, were chosen by the models before the prediction error became worse than the error obtained when the mean was used for cost estimation.

The first data set (feature set 1) included "Time" attributes and the optimal model contained two dependent variables: the time until NFR and the amount of soil excavated. The optimal equation was

$$\begin{aligned}\text{Cost} = {} & 54.5 \times \text{Time until NFR (d)} \\ & + 29.20 \times \text{Amount excavated} - 54400. \end{aligned} \tag{6}$$

As previously, the time attributes were removed in feature set 2 and another model was developed with the remaining features. This model contains only one feature, the amount of soil excavated.

The optimal equation was

$$\text{Cost} = 33.8 \times \text{Amount excavated} + 127000. \tag{7}$$

DISCUSSION

In this section the implications of the results are discussed in greater detail and recommendations are made for improving cost management of the type of sites studied.

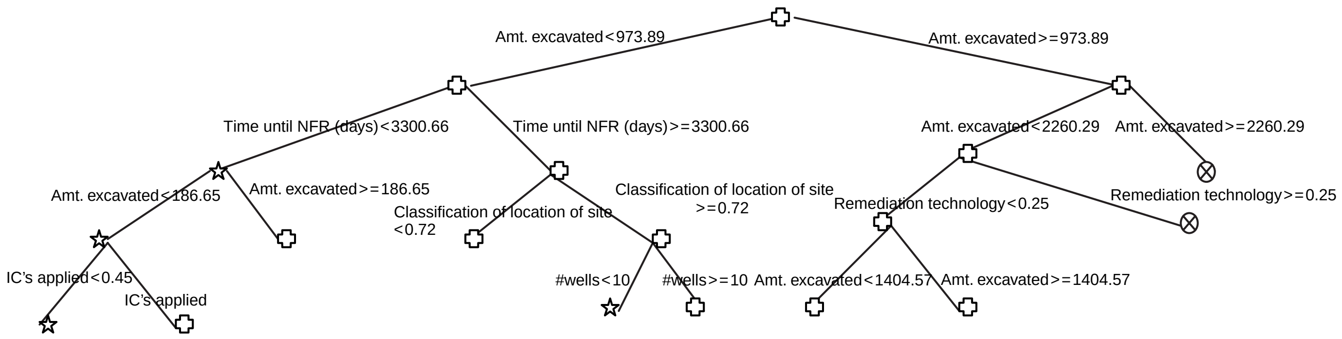

Figure 9 | Decision tree when all of the attributes are included (feature set 1).

Given that it appears at the top tier of the decision trees and was also selected in both regression models, the amount of soil excavated is the most influential attribute in cost prediction at these sites. Additionally, according to the decision tree models, the number of wells installed and whether remediation technology is applied are also quite important. These three attributes can be used to predict whether a site will be high cost, low cost or medium cost. The time until no further remediation is attained, though not particularly useful as a predictive feature, is correlated with cost because the longer the site remains "open", the more costs accrue because of sampling and labour costs. Therefore its inclusion in the models where the time attributes were included is reasonable.

Discussions with the site consultants at Delta Environmental revealed that the amount of soil excavated can vary widely depending on legal drivers, the time frame in the project that the excavation occurred (for example, whether excavation was done immediately or later in the project timeline), the date of excavation and current environmental regulations. Though the potential high cost of excavation is well known, especially when large amounts of soil must be removed, it is often still selected because there is complete removal of the source of contamination, which is thought to facilitate relatively rapid cleanup (Wood 1997; Lambert *et al.* 2003). To test this hypothesis, a decision tree model that included only the time attributes and the attribute "was there excavation" was developed. The results from this model show that, if excavation is done, then the time until NFR is actually *longer*. This decision tree is shown below (Figure 11). It can be seen from the "Time until NFR" branch that, when excavation was done, the average time until NFR was approximately 200 d longer than if excavation was not done. However, 68% of the sites investigated in this study that were excavated also had existing groundwater contamination. In only 7% of the sites was

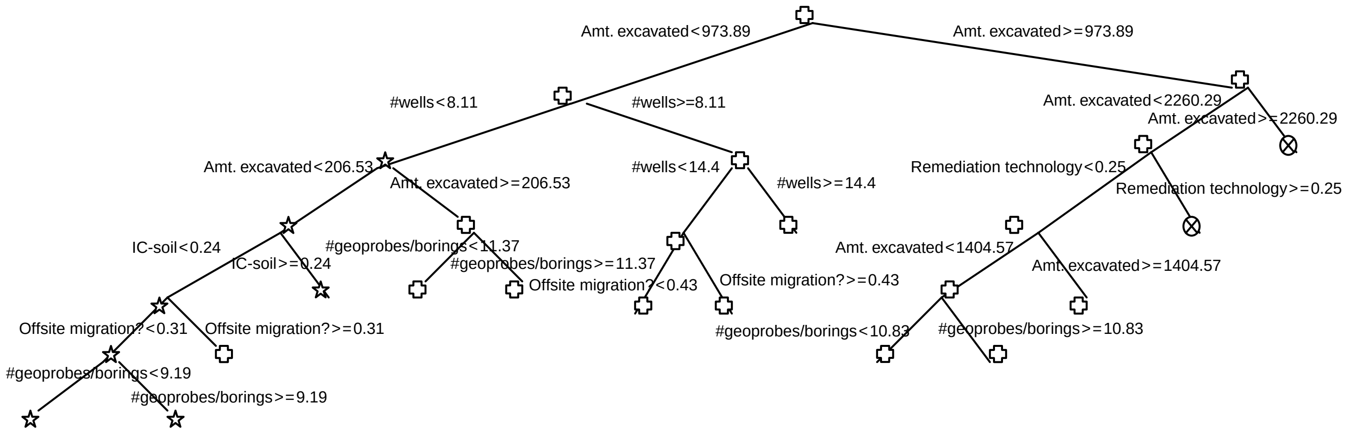

Figure 10 | Optimal decision tree without remediation time attributes (feature set 2).

Table 5 | Attributes included in each regression model (both feature sets)

Round	Attributes included in regression models
1	Amount of soil excavated
2	Amount of soil excavated, no. ofgeoprobes/borings, no. of wells, presence of free product, use of remediation technology
3	Amount of soil excavated, no. of geoprobes/borings, no. of wells, presence of free product, use of remediation technology, hydraulic conductivity, whether there was offsite migration
4	Amount of soil excavated, no. of geoprobes/borings, no. of wells, presence of free product, use of remediation technology, hydraulic conductivity, whether there was offsite migration, presence of PNA's, no. of tanks removed

excavation carried out in the absence of groundwater contamination. This suggests that excavation is more likely to occur when there is contaminated groundwater and/or later in the lifecycle of the site. Since groundwater treatment can be a lengthy process, this may have contributed to the extended time until NFR. Nonetheless, given the high cost of excavation, its benefits in terms of reducing cleanup times should be carefully considered in light of these results. Perhaps there are cases where over-excavation could be avoided in favor of *in situ* treatment alternatives, particularly when groundwater is already contaminated and hence a lengthy remediation is likely.

The second attribute that was shown to be important in affecting cost – the number and placement of wells and borings during site delineation – is influenced by the site lithology. Typically sites with clay-type geologies (low hydraulic conductivities) do not have extensive offsite migration and fewer wells and borings may be needed. Care should be taken to minimize the number of wells while still obtaining adequate information for plume delineation.

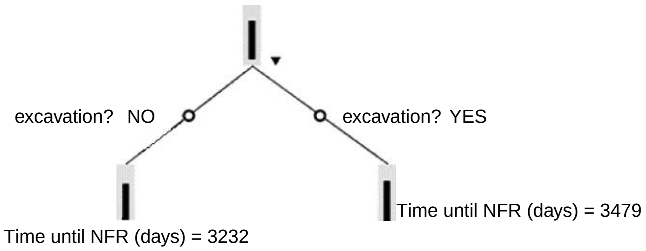

Figure 11 | Decision tree testing relationship between features "Time until NFR" and "Was there excavation".

The third attribute that was shown to be highly correlated with cost was the use of remediation technologies. Remediation technologies can be quite costly and, if used, the final site remediation cost will be higher. From the second tier in the decision tree shown in Figure 9, it can be seen that, if remediation technology is applied when significant excavation has already been done, the site cost classification changes from "medium" to "high". However, in the data set used for this study, only about 19% of the sites required both excavation and the use of remediation technology. This indicates that the use of both excavation and remediation technology is a relatively rare, and expensive, occurrence.

Although the amount of soil excavated, the number of wells installed and whether remediation technology was used are clearly the most important features for predicting cost, the inclusion of the other features identified in the decision tree and regression models can also be rationalized. The features "number of geoprobes/borings" and whether offsite migration occurred have been selected by both the decision tree and the stepwise linear regression models as being indicative of cost. The number of geoprobes/borings has an effect on cost for much the same reason that the number of installed wells is important – for sites that require further sampling and delineation, more borings must be completed and additional cost is incurred. Offsite migration could necessitate the placement of more wells and more borings, or settlements may need to be paid to affected parties, driving cost up.

Using decision trees, the attributes included in this study were able to predict the level of cost in three classes with up to 73.3% accuracy, a substantial improvement over the default mean value that is only 50% accurate. Stepwise linear regression and nearest-neighbour approaches were less successful at making reasonable predictions. However, even decision trees, the most successful approach, predicted the wrong cost class nearly 30% of the time. Moreover, none of these approaches were able to give good predictions of the actual site cost. These difficulties are likely to occur because the approach to closure and hence the amount of money spent on each site can be very site-dependent, with factors such as those listed below having influence.

- *Differences in professional opinion*. Each site is unique and remediation professionals are often required to use their professional judgment on a case by case basis.
- *Resampling*. Sites where the period of assessment and closure covers an extended period of time or active remediation has occurred are often re-sampled to confirm current conditions at the time of closure to accommodate natural attenuation and remedial activities. The number of samples that must be analyzed varies from site to site. This may increase site cost.
- *Type of remediation technology employed*. In this study we only considered whether remediation technology was used and not the type of technology. There are cost differences in the type of remediation technology employed and this may affect the overall site remediation cost.
- *Institutional controls*. The total costs of institutional controls necessary to bring a site to closure can vary widely and can be subject to the individual parties involved. Timeframes involved in reaching successful resolution to negotiations with third parties can delay closure and increase site cost.

The results of this study suggest that costs can be controlled at sites such as those investigated here if more stringent and consistent policies can be adopted for deciding how much soil to excavate and where wells should be placed. Perhaps more effort could be put into optimizing well placement for plume delineation.

CONCLUSIONS

This study has shown that text mining of closure reports can be useful in identifying features that influence cost at remediation sites. Decision trees are particularly useful because the model is not only produced quickly and is easily understood, but it is able to predict costs more accurately than the other approaches. The best decision tree was 73% accurate in predicting the cost in three categories, while the best regression model was only 67% accurate and the instance-based model was only 53% accurate. These results can be compared with the simplest possible model, using the mean cost for all sites under consideration, which was approximately 50% accurate.

Further research is needed to investigate whether other features that were not included in the closure reports would provide additional insights. One such feature could be the number of extensions granted. At times, companies may need additional time to comply with regulatory requirements; however, more extensions may mean that more cost is incurred. Other useful features could be created from more detailed cost information related to wells, which were found to be a major cost driver. With more details on well installation and mobilization costs (for drillers, geologists, sampling crews and analyses), the most important features could be identified that cause increased costs with increasing numbers of wells. This could lead to insights on how to better manage these costs.

To increase the models' prediction accuracy and usefulness, more sites could also be included in the analysis. However, including more sites will be tedious and time-consuming unless the data extraction process can be fully automated. Much work is being done both in the US and internationally in the field of text extraction that may be useful for increasing the automation of data extraction. As more sites move to electronic storage of information in databases, perhaps less emphasis can be placed on data stored in closure reports and other paper documents and more use can be made of data already stored electronically in databases. In this study, features related to political/social/legal aspects of the site (see Table 1) were collected from database records, making incorporation of these attributes quite simple. As the remediation industry moves towards more extensive and efficient data manage4132#-ment methods, the data extraction process will be aided greatly.

Finally, additional work could be done to investigate the extent to which the models and methods utilized in this study are applicable to other gasoline station sites in the US and globally. Data could be collected to determine if there are lessons to be learned from other countries and other states on how to manage similar sites. The models utilized here could be useful in answering questions such as whether the differences in liability costs between the US and other branches of BP worldwide are due to differences in the regulatory environments or other factors such as cost of labour. Any patterns in costs associated with using different consulting firms could also be assessed in such a broader study. This feature was not addressed in this study, since all of the sites were managed by a single contractor.

REFERENCES

Anderton, S. P., White, S. M. & Alvera, B. 2004 Evaluation of spatial variability of snow water equivalent in a high mountain catchment. *Hydrol. Process* **18** (4), 435–453.

Bessler, F. T., Savic, D. A. & Walters, G. A. 2003 Water reservoir control with data mining. *J. Wat. Res. Plann. Mngmnt.* **129** (1), 26–34.

Farrell, D. 2004 *Data Mining to Improve Management and Reduce Costs Associated with Environmental Remediation*. MS thesis, University of Illinois, Urbana-Champaign, IL.

Han, J. & Kamber, M. 2001 *Data Mining- Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco, CA, USA.

Lambert, M., Levin, B. A. and Green, R. M. (2003). *New Methods of Cleaning Up Heavy Metal in Soils and Water*. Available at: http://www.envirotools.org/factsheets/Remediation/cleanup_heavymetals.shtml.

Matheus, C. J. 1990 *Feature Construction: Analytical Framework and an Application to Decision Trees*. PhD dissertation, University of Illinois, Urbana, IL.

Michael, W. J., Minsker, B. S., Tcheng, D., Valocchi, A. J. & Quinn, J. J. 2005 Integrating data sources to improve hydraulic head predictions: a hierarchical machine learning approach. *Wat. Res. Res.* doi: 101029/2003WR002802.

Quinlan, J. R. 1986 Induction of decision trees. *Machine Learning* **1**, 81–106.

Su, F., Zhou, C., Lyne, V., Du, Y. & Shi, W. 2002 A data-mining approach to determine the spatio-temporal relationship between environmental factors and fish distribution. *Ecol. Modell.* **174** (4), 421–431.

Tcheng, D., Lambert, B., Lu, S. C.-Y. & Rendell, L. 1989 Building robust learning systems by computing induction and optimization. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*. Morgan Kaufmann Publishers, Detroit, pp. 806–812.

Welge, M., Auvil, L., Shirk, A., Bushell, C., Bajcsy, P., Cai, D., Redman, T., Clutter, D., Aydt, R. and Tcheng, D. (2003). *Data to Knowledge (D2K). An Automated Learning Group Report.* National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign. Available at: http://alg.ncsa.uiuc.edu.

Wood, P. 1997 Remediation methods for contaminated sites. In *Contaminated Land and its Reclamation* (ed. R. E. Hester & R. M. Harrison). The Royal Society of Chemistry, London, pp. 45–71.