Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Uncertainty quantification and integration of machine learning techniques for predicting acid rock drainage chemistry: A probability bounds approach

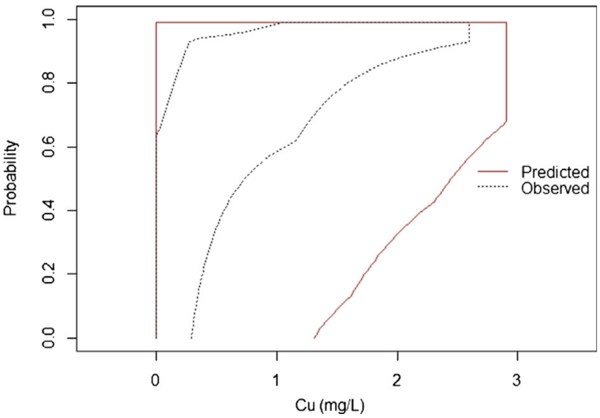
Getnet D. Betrie [a,*], Rehan Sadiq [a], Kevin A. Morin [b], Solomon Tesfamariam [a]

[a] School of Engineering, UBC, Kelowna, BC, Canada
[b] Minesite Drainage Assessment Group, Surrey, BC, Canada

HIGHLIGHTS

- A method to quantify the predictive uncertainty of machine learning was developed.
- Two machine learning techniques were integrated to improve their predictions.
- The sources of uncertainty in model prediction were identified.
- A possible way for reducing prediction uncertainty was suggested.
- A better technique to evaluate the performance of models is found and recommended.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 15 February 2014
Received in revised form 19 April 2014
Accepted 29 April 2014
Available online xxxx

Editor: F.M. Tack

Keywords:
Acid rock drainage
Machine learning
Artificial neural network
Support vector machine
Uncertainty analysis

ABSTRACT

Acid rock drainage (ARD) is a major pollution problem globally that has adversely impacted the environment. Identification and quantification of uncertainties are integral parts of ARD assessment and risk mitigation, however previous studies on predicting ARD drainage chemistry have not fully addressed issues of uncertainties. In this study, artificial neural networks (ANN) and support vector machine (SVM) are used for the prediction of ARD drainage chemistry and their predictive uncertainties are quantified using probability bounds analysis. Furthermore, the predictions of ANN and SVM are integrated using four aggregation methods to improve their individual predictions. The results of this study showed that ANN performed better than SVM in enveloping the observed concentrations. In addition, integrating the prediction of ANN and SVM using the aggregation methods improved the predictions of individual techniques.

* Corresponding author.
E-mail address: getnet.betrie@ubc.ca (G.D. Betrie).

http://dx.doi.org/10.1016/j.scitotenv.2014.04.125
0048-9697/© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Globally acid rock drainage (ARD) is a major pollution problem that poses severe adverse risks to the environment (Gray, 1996, 1998; Azapagic, 2004). The probable global area covered with mine waste is in the order of 100 million ha that contain several hundred thousand million tonnes of mine wastes, and 20,000–25,000 million tonnes of solid waste is added every year (Lottermoser, 2010). The associated liability costs of potentially acid-generating wastes at minesites are estimated to be US$ 1.2–20.6 billion in USA, US$ 1.3–3.3 billion in Canada, and US$ 530 million in Australia (Miller et al., 2006). However, the authors suspect that the estimate for Canada should be at least an order of magnitude higher.

ARD is generated when a sulfide-bearing material is reacted with oxygen and water during mining activities (Morin & Hutt, 2001; Price, 2009). The reaction results in oxidation and other weathering processes, which changes relatively insoluble chemical species in sulfide minerals into more easily dissolved free ionic species (e.g., Cu, Cd and Zn) or secondary minerals (e.g., sulfate, carbonates and oxyhydroxides). Moreover, the oxidation of some sulfide minerals produces acid that may lower the drainage pH. A lower drainage pH could increase the rate of sulfide oxidation, solubility of many products of sulfide oxidation, and rate of weathering for other minerals.

Predicting the future drainage chemistry is important to assess potential environmental risks of ARD and implement appropriate mitigation measures that reduce adverse environmental risks (Betrie et al., 2012). Predictive models are one of the approaches used to predict the future drainage chemistry of minesites. These models are classified as process-based and empirical (data-driven) models (Maest et al., 2005; Price, 2009). Process-based models describe the ARD system in terms of chemical and/or physical processes that are believed to control ARD generation (Betrie et al., 2012). Nevertheless, the physical/chemical processes that govern generation of ARD are not fully understood (Price, 2009). Subsequently, uncertainty is introduced in the prediction of drainage chemistry due to poor representation of the ARD system. In addition, process-based models introduce uncertainty due to data because they use database information (e.g., solubility product) that might not match a given site (Price, 2009). On the other hand, data-driven models (e.g., machine learning, soft-computing, computational intelligence) describe the time-dependent behavior of one or more variables of the ARD system in terms of observed data trends obtained from years of monitoring at a minesite (Betrie et al., 2012). Therefore, these models are prone to uncertainties in the data that arise due to epistemic (e.g., measurement errors and limited sample size) and aleatory (e.g., temporal and spatial variations) uncertainties, where these uncertainties arise due to incomplete knowledge and natural stochasticity, respectively (Sentz & Ferson, 2002).

The literature review shows that machine learning techniques (e.g., ANN and SVM) have been used to predict the ARD drainage chemistry. Khandelwal & Singh (2005) compared ANN and multiple regression analysis (MRA) to predict chemical parameters (sulfate, chloride, total dissolved solid (TDS) and others) as function of physical parameters (pH, temperature, and hardness). They reported that ANN provided acceptable results compared to MRA. Rooki et al. (2011) evaluated two types of ANN (back propagation neural network (BPNN) and general regression neural network (GRNN)) and MRA to predict heavy metals (Cu, Fe, Mn, Zn) as function of physical/chemical parameters (pH, sulfate, and Mg) in the Shur River near Sarcheshmeh Copper mine, Iran. They reported that the predictive accuracy of BPNN is the best followed by GRNN and MRA. For the Shur River and the same input–output variables, Aryafar et al. (2012) applied SVM and compared to their GRNN model results. The results showed that the predictive accuracy of SVM was slightly better than ANN.

Betrie et al. (2012) evaluated the predictive accuracy and uncertainty of four machine learning techniques (ANN, SVM, mode trees, and K-nearest neighbors) to predict copper concentration as a function of physical/chemical parameters and their time lags. The authors reported that SVM performed best followed by ANN, model trees and K-nearest neighbors both in terms of predictive accuracy and uncertainty. The prediction accuracy refers to the difference between observed and predicted values, whereas the predictive uncertainty refers to the variability of the overall error around the mean error (Betrie et al., 2012).

Although identification and quantification of uncertainties are integral parts of ARD assessment and risk mitigation (Price, 2009), previous studies have not addressed uncertainty issues except a minor attempt by Betrie et al. (2012). In this paper, predictive uncertainties of ANN and SVM due to input data are quantified using the probability bounds approach. The probability bounds approach is an uncertainty analysis method that combines probability theory and interval arithmetic to produce probability boxes, which allow the comprehensive propagation of both variability and uncertainty rigorously (Tucker & Ferson, 2003). Furthermore, predictions of ANN and SVM are integrated using four aggregation methods in order to improve the prediction of the individual technique. Aggregation methods are used to combine information obtained from various sources in order to improve the reliability of information (Sentz & Ferson, 2002).

The remainder of this paper is structured as follows. The next section presents the descriptions of ANN and SVM techniques and the method used for data preprocessing including treating missing and outlier values, defining modeling variables and conducting uncertainty analysis. The Results and discussion section presents the main findings of this study and discusses these findings, respectively. The Summary and conclusions section of this study completes this paper.

2. Material and methods

The methodology followed in this study is depicted in Fig. 1. It shows that the methodology consists of five blocks. In the first block, data pre-processing is done that includes filling missing values and outlier analysis. In the second block, variables that control drainage chemistry of ARD are identified and used to develop model using the machine learning techniques. In the third block, the dataset is divided into training and testing sets using ten-fold cross-validation technique. The training dataset is used to optimize parameters of the models, whereas the test dataset is used for predicting drainage chemistry. In the fourth block, first the predictive accuracy of training models is evaluated using four statistical techniques. If the results of training are not acceptable based on the obtained statistics, the modeling process would be reinitiated from the second block. However, the predictive accuracy and uncertainty for test models would be initiated if the training models provide acceptable results. In the last block, the uncertainties due to data and model are quantified using probability bounds approach. Also, predictions from ANN and SVM are integrated using four aggregation methods to reduce the predictive uncertainty of individual models.

2.1. Machine learning techniques and uncertainty analysis

Machine learning is an algorithm that estimates an unknown dependency between mine waste geochemical system inputs and its outputs from the available data (Betrie et al., 2012). In this study, ANN and SVM techniques are used since they performed well in our previous studies. These two techniques are implemented using WEKA 3.6.4 Software (Bouckaert et al., 2010). The concept of machine learning and the detailed evaluation of various machine learning techniques can be seen in Betrie et al. (2012). The description of ANN and SVM techniques is described in detail, consistent with Betrie et al. (2012), in the following subsections.

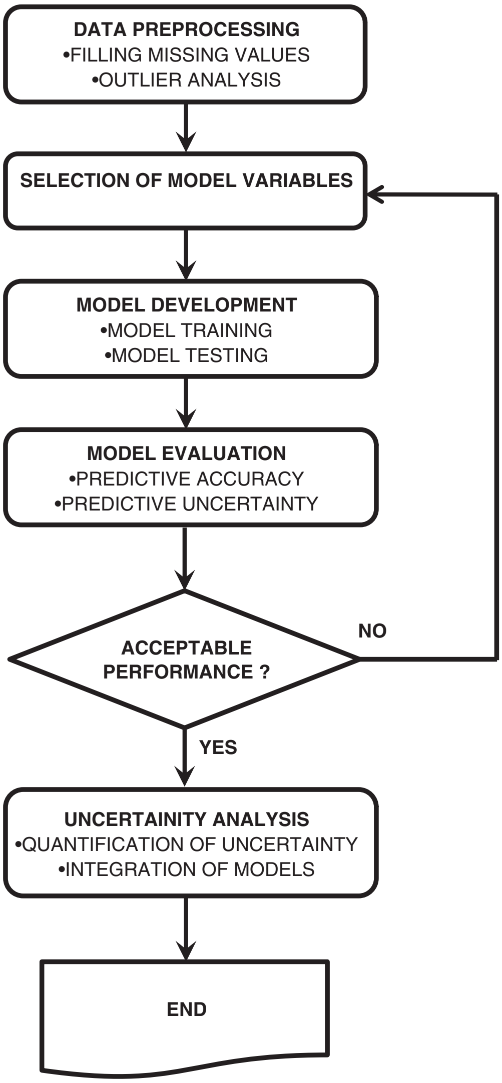


Fig. 1. A schematic representation of the methodology used in this study.

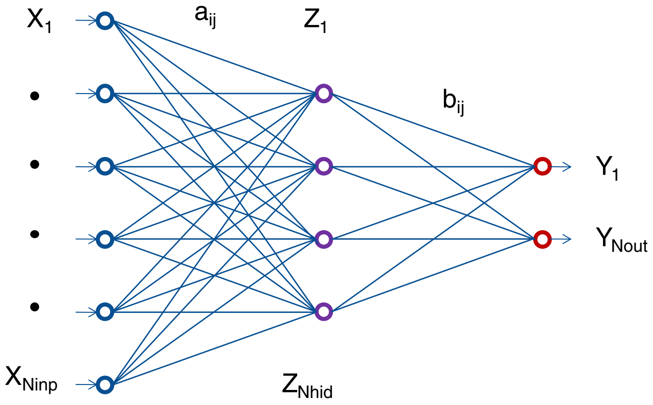


Fig. 2. Multilayer perceptron neural networks.

2.1.1. Artificial neural network (ANN)

Artificial neural network (ANN) is one of machine learning techniques that consist of neurons with massively weighted interconnections (Mitchell, 1997). These neurons are arranged as input layer, hidden layer and output layer as displayed in Fig. 2. The task of input layer is only to send the input signals to the hidden layer without performing any operations. The hidden and output layers multiply the input signals by set of weights and either linearly or nonlinearly transform results into output values. These weights are optimized during ANN training process to obtain reasonable predictive accuracy.

In this study, the multilayer perceptron is used although there are various types of ANN algorithms (Mitchell, 1997). Multilayer perceptron is a feedforward neural network, where signals always travel in the direction of the output layer. Typical multilayer perceptrons with one hidden layer can be mathematically expressed in Eqs. (1)–(4). The outputs of hidden layer (Z_j) is obtained as (i) summing the products of the inputs (X_i) and weight vectors (a_{ij}) and a hidden layer's bias term (a_{0j}), and (ii) transforming this sum using transfer function g as shown in Eqs. (1) and (2), respectively. The most widely used transfer functions are logistic and hyperbolic tangent. Similarly, the outputs of the output layer (Y_k) are obtained by (i) summing the products of hidden layer's outputs (Z_j) and weight vectors (b_{jk}) and output layer's bias term (b_{0k}), and (ii) transforming this sum using transfer function gas shown in Eqs. (3) and (4), respectively.

$$u_j = \sum_{i=1}^{N_{inp}} X_i a_{ij} + a_{0j} \tag{1}$$

$$Z_j = g\left(u_j\right) \tag{2}$$

$$v_k = \sum_{j=1}^{N_{hid}} Z_j b_{jk} + b_{0k} \tag{3}$$

$$Y_k = g(v_k) \tag{4}$$

2.1.2. Support vector machine (SVM)

The support vector machine was mainly developed by Vapnik and co-workers (Vapnik, 1998; Cherkassky and Mulier, 2007). Its principle is based on the Structural Risk Minimization that overcomes the limitation of the traditional Empirical Risk Minimization technique under limited training data. The Structural Risk Minimization aims at minimizing a bound on the generalization error of a model instead of minimizing the error on the training dataset. The SVM algorithm was first developed for classification problems and then adapted to address regression problems. In this study, the basic idea of SVM regression is illustrated since a regression problem is solved.

The complete description of SVM regression is well presented by Smola and Scholkopf (2003) and summary of it presented in this study. Given a training dataset (x_i, y_i), where x_i is the i-th input pattern and y_i is the corresponding target value $y_i \in \mathbb{R}$. The goal of SVM regression is to find a function $f(x)$ that has at the most ε deviation from actually obtained targets y_i for all training data, and at the same time, is as flat as possible (Vapnik, 2000). The function f is represented using a linear function in the feature space

$$f(x) = \langle w, x \rangle + b \text{ with } \quad w \in X, b \in \mathbb{R} \tag{5}$$

where $\langle .,. \rangle$ denotes the dot product in X. In this case, the flatness means seeking a small w. This can be ensured by minimizing the norm (i.e., $||w||^2 = \langle w.\, w \rangle$) if the assumption that a function f is known a priori to approximate all pairs (x_i, y_i) with precision. If such function is not known a priori, it is possible to introduce slack variables ξ_i, ξ_i^* and allow for some errors. This minimization problem can be mathematically expressed as

$$\begin{aligned}&\text{minimize}\ \frac{1}{2}\|w\|^2 + C\sum_i(\xi_i+\xi_i^*)\\ &\text{Subject to}\begin{cases} y_i - \langle w, x_i\rangle - b \le \varepsilon + \xi_i\\ \langle w, x_i\rangle + b - yi \le \varepsilon + \xi_i^* \ .\\ \xi_i, \xi_i^* \ge 0\end{cases}\end{aligned} \tag{6}$$

The constant $C > 0$ determines the tradeoff between the flatness of f and the amount up to which deviations larger than ε are tolerated. The constrained optimization problem is converted into unconstrained optimization by introducing Lagrange function. The Lagrange function is constructed from the objective function and the corresponding constraints by introducing a dual set of variables as follows:

$$\begin{aligned}L = &\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i+\xi_i^*) - \sum_{i=1}^{l}\alpha_i(\varepsilon + \xi_i - y_i) + \langle w, x_i\rangle + b) -\\ &\sum_{i=1}^{l}\alpha_i^*(\varepsilon + \xi_i^* + y_i - \langle w, x_i\rangle - b) - \sum_{i=1}^{l}(\eta_i\xi_i + \eta_i^*\xi_i^*).\end{aligned} \tag{7}$$

It follows from the saddle point condition that the partial derivatives of L with respect to the primal variables (w, b, ξ_i, ξ_i^*) have to vanish for optimality. Substituting the results of this derivation into Eq. (7) yields the dual optimization problem.

$$\begin{aligned}&\text{Maximize}:\ \frac{1}{2}\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)\left(\alpha_j-\alpha_j^*\right)\left\langle x_i, x_j\right\rangle - \varepsilon\sum_{i=1}^{l}(\alpha_i+\alpha_i^*) + \sum_{i=1}^{l}y_i(\alpha_i-\alpha_i^*)\\ &\qquad\qquad\text{Subject to}:\ \sum_{i=1}^{l}(\alpha_i-\alpha_i^*) = 0\,\text{and}\,\alpha_i, \alpha_i^* \in [0, C].\end{aligned} \tag{8}$$

Once the coefficients α_i and α_i^* are determined from Eq. (8), the desired vectors can be written as follows:

$$w = \sum_{i=1}^{l}(\alpha_i-\alpha_i^*)x_i, \text{and therefore}\, f(x) = \sum_{i=1}^{l}(\alpha_i-\alpha_i^*)\langle x_i, x\rangle + b. \tag{9}$$

Nonlinear regression problems are very common in most engineering applications. In such case, a nonlinear mapping kernel K is used to map the data into a higher-dimensional feature space or hyperplane by the function Φ. The kernel function, $K(x_i, x) = \langle\Phi(x_i), \Phi(x)\rangle$ can assume any form. In this study, the polynomial (SVM-Poly) kernel is used, which is shown in equation

$$K(x_i, x) = (\gamma\langle x_i, x\rangle + \tau)^d, \gamma > 0 \tag{10}$$

where γ, τ, and, d are kernel parameters.

2.2. Uncertainty analysis

Uncertainties such as epistemic and aleatory from data, parameters and model conceptualization propagate into prediction results. In this subsection, the probability bounds approach is used to quantify the prediction uncertainties and aggregation methods are used to improve model uncertainties.

2.2.1. Probability bounds

Probability bounds (*P*-boxes) is a method used to represent imprecise probability. Imprecise probability is a generalization of probability theory when one is not able to define a precise probability function P for an event x, which is element of the universal set X (Walley, 1991). An imprecise probability function $P(x)$ is characterized by its lower probability $\underline{P}(x)$ and upper probability $\overline{P}(x)$. Lower probability and upper probability functions map an event $x \in X$ in interval values between zero and one (Ferson et al., 2003). The lower and upper bounds on $P(x)$ are the probabilities that a random variable X is smaller and greater than x, respectively. In this study, the predicted and observed *P*-boxes are constructed to compare the uncertainties and this is implemented using Risk Cal 4.0 software (Ferson, 2000).

2.2.2. Aggregation methods

Aggregation methods are used to combine information obtained from various sources in order to improve the reliability of information for better decision-making (Sentz & Ferson, 2002). In this study, four aggregation methods are used to integrate the prediction results of ANN and SVM. These methods are intersection, envelope, mixture, and averaging.

The intersection method gives the smallest region that all predictions agree with high degree of confidence (Ferson et al., 2003). This method is appropriate to use when a modeler strongly believes that the prediction of each model encloses the distribution of the observed data. Suppose $P_1 = \left[\underline{P_1}, \overline{P_1}\right], P_2 = \left[\underline{P_2}, \overline{P_2}\right], \dots, P_\mathrm{n} = \left[\underline{P_\mathrm{n}}, \overline{P_\mathrm{n}}\right]$ are prediction *P*-boxes of many models, the intersection method can be mathematically expressed using the equation below

$$P_1 * P_2 * \dots * P_\mathrm{n} = \left[\max\left(\underline{P_1}, \underline{P_2}, \dots, \underline{P_\mathrm{n}}\right), \min\left(\overline{P_1}, \overline{P_2}, \dots, \overline{P_\mathrm{n}}\right)\right]. \tag{11}$$

The envelope method gives the biggest region that each prediction has a certain degree of confidence. It is appropriate when a modeler believes that at least one of the predictions encloses the observed distribution (Ferson et al., 2003). For n prediction *P*-boxes mentioned above, the envelope method can be mathematically expressed as follows

$$P_1 * P_2 * \dots * P_\mathrm{n} = \left[\min\left(\underline{P_1}, \underline{P_2}, \dots, \underline{P_\mathrm{n}}\right), \max\left(\overline{P_1}, \overline{P_2}, \dots, \overline{P_\mathrm{n}}\right)\right]. \tag{12}$$

The mixture method treats disagreements between the prediction *P*-boxes from each model and gives a condensed *P*-box without erasing disagreements (Ferson et al., 2003). For n prediction *P*-boxes mentioned above, the mixture method can be mathematically expressed by the equation below

$$\begin{aligned}\underline{P} &= \left(w_1\underline{P_1} + w_2\underline{P_2} + \dots + w_\mathrm{n}\underline{P_3}\right)/\sum w_i\\ \overline{P} &= \left(w_1\overline{P_1} + w_2\overline{P_2} + \dots + w_\mathrm{n}\overline{P_\mathrm{n}}\right)/\sum w_i\end{aligned} \tag{13}$$

where w_1, w_2, …, w_n are weights of the *P*-boxes.

The average method simply horizontally averages the edges of the *P*-box by finding the inversion of the lower and upper bounds (Ferson et al., 2003). For n-boxes the average method can be mathematically expressed by the following equation

$$\begin{aligned}(\underline{P}*)^{-1} &= (1/\mathrm{n})\left(\underline{P}_1^{-1} + \underline{P}_2^{-1} + \dots + \underline{P}_\mathrm{n}^{-1}\right)\\ (\overline{P}*)^{-1} &= (1/\mathrm{n})\left(\overline{P}_1^{-1} + \overline{P}_2^{-1} + \overline{P}_\mathrm{n}^{-1}\right)\end{aligned} \tag{14}$$

where −1 represents the inverse function.

2.3. Case study

2.3.1. Data preparation

The data used for this study was obtained from a copper–molybdenum–gold–silver–rhenium (CMGSR) minesite located in British Columbia, Canada. The dataset was collected for over 25 years and it consists of 5000 values and 13 variables. These variables are pH, conductivity (μs/cm), acidity ($CaCO_3$ mg/L), alkalinity (mg/L), sulfate (mg/L), flow (mg/L), copper (mg/L), cadmium (mg/L), zinc (mg/L), calcium (mg/L), magnesium (mg/L), and aluminum (mg/L). Conductivity, pH, and flow were measured in situ at monitoring stations, whereas the concentrations of alkalinity, acidity and dissolved metals were measured at CMGSR environmental laboratory. However, this dataset contains numerous missing values and outliers (Betrie et al., 2012; Morin et al., 2012).

2.3.1.1. Missing values. Missing data present challenges in data driven modeling that includes machine learning, soft-computing and data mining (Cherkassky and Mulier, 2006; Betrie et al., submitted for publication). The challenges associated with missing data include the use and interpretation of partially collected data (Betrie et al., 2012) and accuracy of learning algorithms (Bello, 1995; Acuna & Rodriguez, 2004).

The missing values in this database were estimated using the iterative robust model-based imputation (IRMI) algorithm. The IRMI algorithm is a model-based imputation method where missing values are estimated using sequence of regression models (Templ et al., 2011). The IRMI algorithm, used to estimate the missing values, is implemented in RStudio (2012) and a summary of this algorithm is presented in Betrie et al. (submitted for publication). The number of missing values estimated for various parameters were: 434 for pH, 790 for conductivity, 3412 for alkalinity, 3725 for acidity, 1657 for sulfate, 1253 for calcium (Ca), 1257 for magnesium (Mg), 1543 for aluminum (Al), 4217 for flow, 124 for copper (Cu), 134 for cadmium (Cd), and 70 for zinc (Zn).

2.3.1.2. Outlier analysis. In data-driven modeling, while extreme values are very important for learning algorithms to make an accurate prediction under extreme conditions, outlier values could lead the learning algorithm to provide false prediction under extreme conditions. According to Reimann et al. (2005), extreme values are those values that belong to the same distribution of data, but far away from the center, whereas outlier values belong to different distributions. From environmental risk point of view, accurate predictions of extreme values are necessary in order to make a conservative decision. In this study, therefore, multivariate outlier analysis is conducted using the adaptive outlier detection algorithm in RStudio to remove outlier values from the data. In order to identify outlier values, this function compares the observed and empirical chi-square distributions, where the latter is computed from a robust square distance on the basis of the minimum covariance determinant estimator (Filzmoser et al., 2005). The 99% percentile of observed and empirical distributions was used to differentiate outliers from extreme values. As a result, 951 outliers were found and removed from this dataset.

2.3.2. Selection of variables for drainage chemistry

The input variables to machine learning techniques should consist of all relevant variables that influence the ARD generation (Betrie et al., 2012). However, overlapping information of input variables is avoided to simplify the task of the training algorithms. For this reason, a nonlinear correlation analysis was conducted using the maximum criterion information (MIC) (Reshef et al., 2011) algorithm implemented in RStudio to identify correlation between input and output variables. Unlike other correlation analysis methods (e.g., Spearman, Pearson, and Kendall), the MIC detects not only a linear relationship but also other nonlinear relationships such as cubic, exponential, categorical, periodic, hyperbolic and various sinusoidal types. Its value ranges between zero and one, where zero and one indicate no and perfect relationships, respectively. The correlation between input and output variables is shown in Table 1. A value greater than or equal to 0.3 was used from the values presented in Table 1 as criterion to select input variable. It shows that the copper concentration correlated to pH, alkalinity, sulfate, acidity, and Al. The Cd concentration is correlated to pH and Al. The Zn concentration is correlated to pH and Al. Therefore, heavy metals (Cu, Cd, and Zn) are predicted as function of pH, alkalinity, sulfate, acidity, and flow. Although flow has a weak correlation with other variables, it is included as input variable because we believe that the machine learning techniques can extract more complex nonlinear relationships than the MIC.

2.3.3. Model development and evaluation

The dataset was divided into training and testing sets following the *k*-fold cross-validation method (Mitchell, 1997). In this method, the dataset is subdivided into *k* subsets preferably of equal size. Next, the $k-1$ subsets are used to train the machine learning algorithms and the remaining one subset is used for testing the models. In this study, each subset has the size of 475 values and ten-fold cross-validation is used.

The predictive accuracy of each machine learning technique was evaluated using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Root Relative Squared Error (RRSE), Relative Absolute Error (RAE), where the smaller value indicates a better technique (Betrie et al., 2012). The predictive accuracy helps to evaluate the overall match between observed and predicted values for each machine learning technique. The equations of the error estimates are given in Eqs. (15)–(18):

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{\mathrm{n}} \left(Y_o - Y_p\right)^2}{\mathrm{n}}} \tag{15}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{\mathrm{n}} \left|Y_o - Y_p\right|}{\mathrm{n}} \tag{16}$$

$$\mathrm{RRSE} = \sqrt{\frac{\sum_{i=1}^{\mathrm{n}} \left(Y_o - Y_p\right)^2}{\sum_{i=1}^{\mathrm{n}} \left(Y_o - \overline{Y_o}\right)^2}} \tag{17}$$

Table 1
Correlation between input and output variables using the MIC.

	pH	Conductivity	Sulfate	Alkalinity	Acidity	Flow	Ca	Mg	Al	Cu	Cd	Zn
pH	1.0											
Conductivity	0.3	1.0										
Sulfate	0.3	0.5	1.0									
Alkalinity	0.4	0.2	0.1	1.0								
Acidity	0.4	0.4	0.5	0.2	1.0							
Flow	0.2	0.2	0.2	0.1	0.2	1.0						
Ca	0.1	0.4	0.5	0.1	0.2	0.2	1.0					
Mg	0.3	0.5	0.5	0.1	0.3	0.2	0.6	1.0				
Al	0.5	0.3	0.4	0.2	0.5	0.2	0.3	0.4	1.0			
Cu	0.8	0.3	0.4	0.3	0.5	0.1	0.2	0.3	0.7	1.0		
Cd	0.5	0.1	0.2	0.2	0.3	0.1	0.3	0.2	0.5	0.5	1.0	
Zn	0.7	0.2	0.3	0.3	0.3	0.1	0.2	0.2	0.5	0.7	0.6	1.0

$$\mathrm{RAE} = \frac{\sum_{i=1}^{n}\left|Y_o - Y_p\right|}{\sum_{i=1}^{n}\left|Y_o - \overline{Y_o}\right|} \tag{18}$$

where Y_o and Y_p represent the observed and predicted outputs, $\overline{y}_p$ represents the mean the predicted output and n represents the number of examples presented to the learning algorithms. On the other hand, the predictive uncertainty of each machine learning technique is evaluated by comparing the observed and predicted *P*-boxes.

3. Results and discussion

Performances of the ANN and SVM models in predicting the heavy metals in terms of the four evaluation methods are presented in Table 2. For the prediction of Cu using ANN, it shows that Cu-5 and Cu-4 models are the best and the worst predictions, respectively. For the prediction of Cu using SVM, Cu-2 and Cu-1 models are the best and worst predictions, respectively. The range of the performance indicates that the ANN model has a higher prediction uncertainty than the SVM model. For the prediction of Cd, both models have the same uncertainty band and the performances of the models are not good as shown by the RRSE and RAE measures. For the prediction of Zn, Zn-6 and Zn-2 models of ANN are the best and worst, whereas Zn-8 and Zn-2 of SVM are the best and the worst models, respectively. The range of the performance measures indicates that the prediction of ANN is higher than that of SVM.

The predicted and observed *P*-boxes are shown in Fig. 3. This figure shows that the predicted Cu concentration using ANN very well enveloped the observed Cu distribution, whereas the prediction of SVM has not enveloped the observed Cu distribution completely. As indicated by the performance measure above, the upper bound of ANN prediction is higher than the upper bound of SVM prediction. The observed of Cd concentration is not enveloped by both the ANN and SVM models prediction. This is attributed to the majority of Cd data (over 90%) below 0.05 mg/L, which is a value of analytical detection limit. The data above 0.05 mg/L do not have information to identify an empirical model. Nevertheless, it is interesting to note that 90% of the data are well enveloped by ANN and SVM. The majority of observed Zn distribution is well enveloped by the prediction of ANN except a small portion of the upper and lower bounds. On the other hand, the prediction of SVM has not enveloped some parts of the observed Zn.

Table 2
Performance of ANN and SVM models in terms of four evaluation measures.

	ANN				SVM			
Model	MAE	RMSE	RRSE	RAE	MAE	RMSE	RRSE	RAE
Cu-1	0.11	0.20	34.20	23.23	0.16	0.25	41.73	32.81
Cu-2	0.12	0.20	37.65	27.89	0.13	0.20	39.17	30.98
Cu-3	0.13	0.18	32.21	26.67	0.15	0.21	38.66	31.99
Cu-4	0.29	0.32	57.32	60.94	0.15	0.22	39.08	31.75
Cu-5	0.09	0.16	28.20	20.00	0.14	0.21	37.06	30.21
Cu-6	0.13	0.18	31.32	26.86	0.14	0.20	34.97	27.76
Cu-7	0.13	0.21	34.83	26.07	0.15	0.22	36.77	29.76
Cu-8	0.13	0.20	36.54	26.91	0.15	0.22	39.58	32.06
Cu-9	0.11	0.17	30.86	22.19	0.14	0.20	36.31	29.21
Cu-10	0.11	0.17	33.28	24.56	0.14	0.21	40.24	32.03
Cd-1	0.01	0.02	73.89	60.39	0.01	0.02	73.89	60.39
Cd-2	0.01	0.02	77.82	64.53	0.01	0.02	77.82	64.53
Cd-3	0.01	0.02	75.19	61.14	0.01	0.02	75.19	61.14
Cd-4	0.01	0.01	71.89	63.86	0.01	0.01	71.89	63.86
Cd-5	0.01	0.01	71.62	61.78	0.01	0.01	71.62	61.78
Cd-6	0.01	0.01	71.15	59.63	0.01	0.01	71.15	59.63
Cd-7	0.01	0.02	74.43	60.73	0.01	0.02	74.43	60.73
Cd-8	0.01	0.01	71.54	58.65	0.01	0.01	71.54	58.65
Cd-9	0.01	0.02	72.73	59.23	0.01	0.02	72.73	59.23
Cd-10	0.01	0.02	79.26	62.12	0.01	0.02	79.26	62.12
Zn-1	1.34	2.41	57.68	42.30	1.48	2.55	60.95	46.91
Zn-2	1.51	2.52	59.56	46.19	1.58	2.69	63.66	48.30
Zn-3	1.48	2.53	59.28	43.93	1.58	2.69	62.81	46.95
Zn-4	1.25	1.83	49.11	40.94	1.42	2.04	54.65	46.35
Zn-5	1.19	1.94	50.09	39.84	1.36	2.15	55.58	45.62
Zn-6	1.19	1.86	48.67	38.06	1.38	2.12	55.27	44.21
Zn-7	1.34	2.33	56.95	41.87	1.50	2.50	61.05	46.90
Zn-8	1.16	2.06	52.37	36.52	1.34	2.19	55.82	42.12
Zn-9	1.32	2.25	54.88	41.75	1.45	2.47	60.29	45.70
Zn-10	1.33	2.68	62.81	42.93	1.49	2.79	65.38	48.05

The comparison between the integrated (ANN–SVM) prediction of Cu and observed *P*-boxes is shown in Fig. 4. This figure shows that the integrated prediction using the envelope and intersection methods gives the individual prediction of ANN and SVM, respectively. The integrated prediction using the mixture method is higher than the observed upper bound distribution. However, the mixture result is closer to the observed upper bound than the individual prediction of ANN. The integrated prediction of Cu using the average method well enveloped the *P*-box for the observed concentrations. However, the majority part of the mixture upper bound is higher than the observed upper bound.

The comparison between the integrated (ANN–SVM) prediction of Cd and observed *P*-boxes is shown in Fig. 5. This figure shows that any of the integration methods could not improve the prediction uncertainty of individual models. Also, it shows that the integrated predictions are exactly the predictions of individual ANN and SVM. As discussed previously, these poor predictions are attributed to lack of heterogeneity in the Cd data.

The comparison between the integrated (ANN–SVM) prediction of Zn and observed *P*-boxes is shown in Fig. 6. This figure shows that the integrated prediction using the envelop method well bounded the observed *P*-box except the slight portion of the upper bound. It is interesting to note that this method improved the poor predictions of ANN at the lower part of the upper bound distribution. The integrated prediction using the intersection method has not enveloped the upper bound of the observed distribution. The integrated prediction of Zn using the mixture and average methods has not enveloped the upper portion of the observed upper bound.

The correlation coefficients between variables reported in this study are quite different from values reported in previous studies. For instance, the correlations in this study between Cu (i.e., dependent) and pH, conductivity, sulfate, and alkalinity (i.e., independent) are equal to 0.8, 0.3, 0.4, and 0.3, respectively as seen in Table 1. Aryafar et al. (2012) reported that the relationship between Cu (i.e., dependent) and pH, conductivity, sulfate, and alkalinity (i.e., independent) is equal to −0.697, 0.757, 0.663, and −0.199, respectively. Betrie et al. (2012) reported that the relationship between Cu (i.e., dependent) and pH and conductivity is equal to −0.74, and 0.52, respectively. The absolute values of these and previous studies for Cu and pH show that there is an agreement, whereas the absolute values of this study and previous works disagree for correlation between Cu and the other variables. This disagreement is likely attributed to the effect of outliers. All previous studies used linear correlation methods (Spearman and Pearson) that are highly sensitive to outliers (Gideon et al., 1987). Therefore, the correlation results of previous studies are relatively inflated compared to this study that does not have outliers.

The evaluation among machine learning techniques based on predictive accuracy by Aryafar et al. (2012) and Betrie et al. (2012) reported that SVM performed best compared to other techniques. Furthermore, Betrie et al. (2012) reported that SVM performed best compared to other methods based on predictive uncertainty. However, this study shows that ANN is better than SVM since it envelops very well the observed data as depicted in Fig. 3. This finding indicates that the higher predictive uncertainty of ANN, which is often considered as a limitation, has enabled ANN to envelop the observed *P*-box. It is interesting to note that this finding indicates that selection of an optimal model based on model accuracy (i.e., evaluation techniques) presented in Section 2.3.3 could be misleading. This agrees with that of Cherkassky and Mulier (2006) who stated that a selection of model based on evaluation techniques cannot guarantee an optimal model in critical situations.

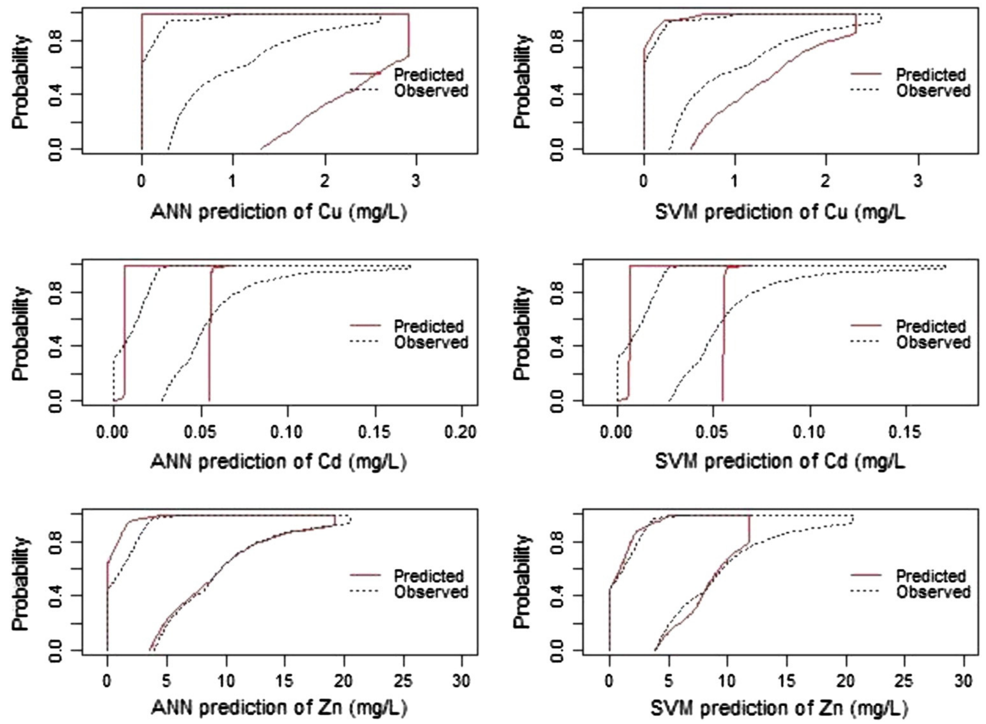


Fig. 3. Predicted and observed *P*-boxes for Cu, Cd, and Zn.

Therefore, it could be reliable to select an optimal model based on predictive uncertainty rather than predictive accuracy of models.

It is worth noting that while the gap between the lower and upper bounds informs the degree of epistemic uncertainty, the shapes of lower and upper bounds inform the degree of aleatory uncertainty (i.e., variability). The predicted and observed lower bounds have no variability as they are almost straight, whereas the upper bounds have some variability as shown in the figures. The lack of variability in the lower bound could be attributed to analytical detection limits. Both the observed and predicted data have epistemic uncertainty as shown in the figures. Thus, this indicates that there is a need for further study and data collection to reduce the epistemic uncertainty.

Although the machine learning techniques predicted well the observed distribution of Cu and Zn, they have not performed well for the Cd prediction as seen in Fig. 3. Most of the observed data (over 90%) are below 0.05 mg/L (see Fig. 7) and those data above this value do not have enough information to identify an empirical model. Subsequently, the maximum concentration value predicted by ANN and SVM is slightly over 0.05 mg/L. This result indicates that machine learning techniques work best if the data have heterogeneity. Of course, one of

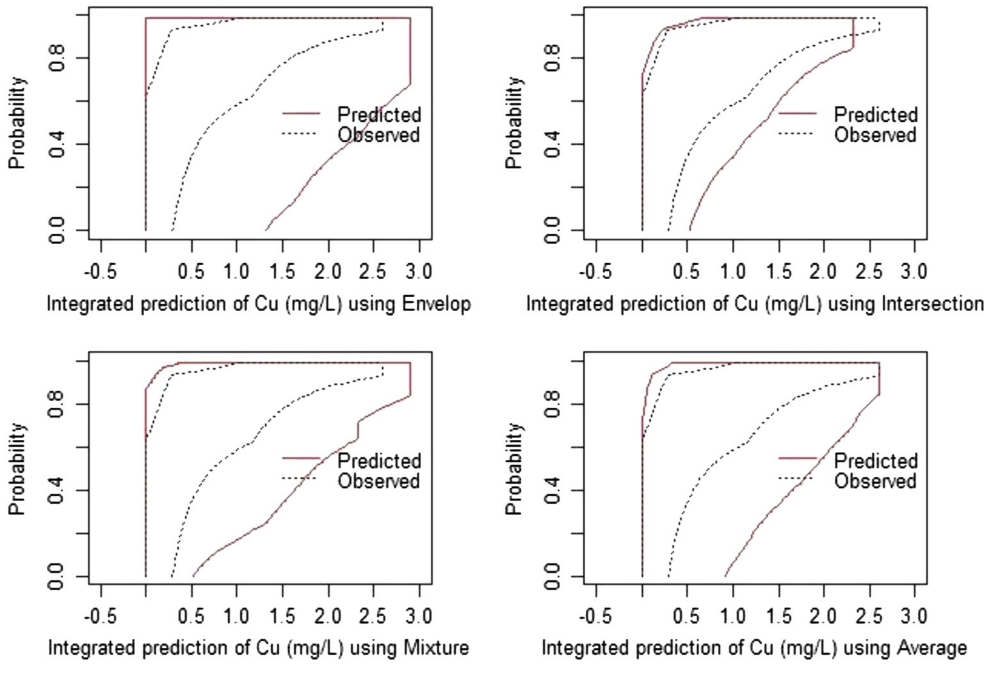


Fig. 4. Integrating ANN and SVM predictions for Cu using four methods and observed *P*-boxes.

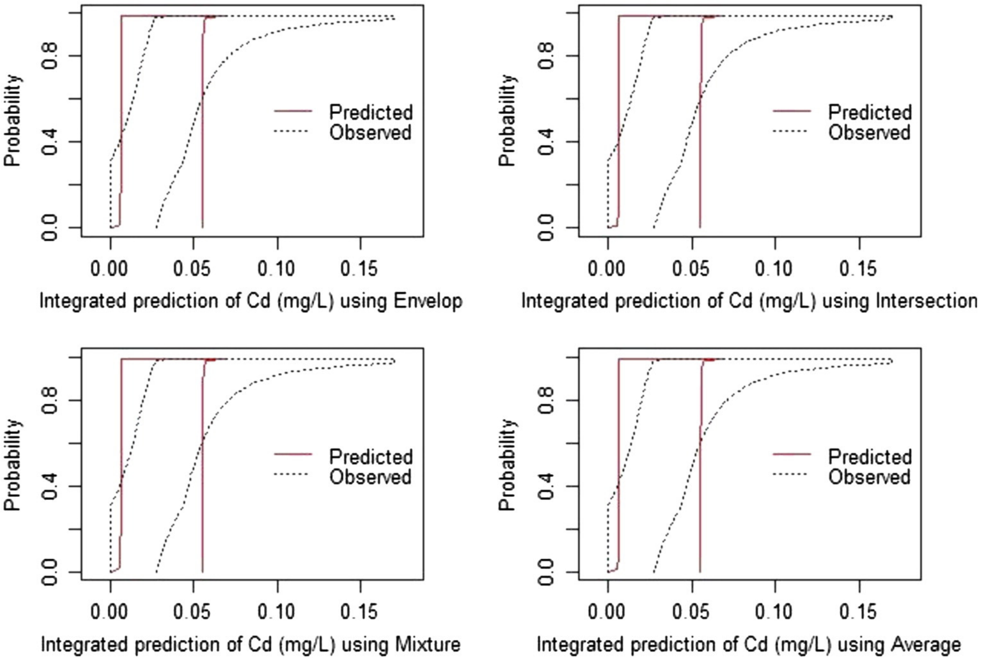


Fig. 5. Integrating ANN and SVM predictions for of Cd using four methods and observed *P*-boxes.

the limitations of data-driven paradigms including machine learning techniques is the requirement of a large database (Cherkassky and Mulier, 2006; Solomatine & Ostfeld, 2008). In case the data lack heterogeneity, a classical modeling approach (e.g., multiple linear regression) should be used instead of machine learning techniques.

Integrating the prediction of ANN and SVM using the aggregation methods improved the prediction results except for the prediction of Cd as shown in Figs. 4-6. For the integration of Cu prediction, the average method performed best followed by mixture, envelope and intersection. On the other hand, the envelope method performed best followed by mixture, average and intersection in the integration of Zn prediction. It is interesting to note that the intersection method has the worst performance and no single method performed best for the prediction of Cu and Zn. This result indicates that all aggregation methods should be investigated by modelers if the intention is to improve prediction results.

4. Summary and conclusions

This study quantifies the predictive uncertainty of two machine learning techniques for predicting acid rock drainage (ARD) chemistry using the probability bounds (*P*-boxes) approach. Also, it integrates the prediction of machine learning techniques using four aggregation methods to improve the individual prediction. The two machine

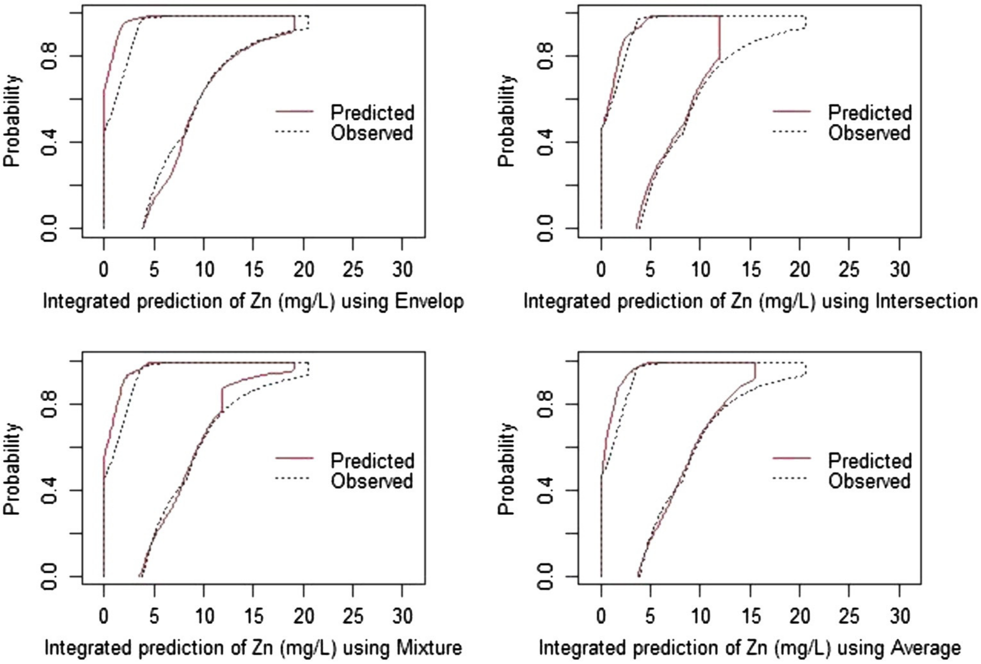


Fig. 6. Integrating ANN and SVM predictions for Zn using four methods and observed *P*-boxes.

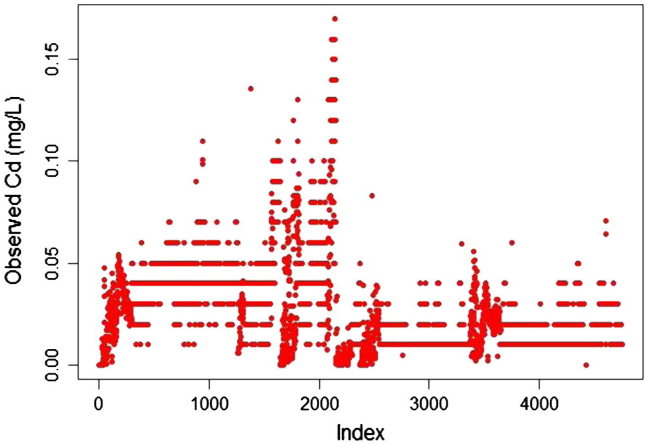


Fig. 7. Observed concentrations of Cd.

learning techniques used are artificial neural networks (ANN) with multilayer perceptrons and support vector machine with polynomial kernel.

Missing values in the data were estimated using the iterative robust model-based imputation algorithm. Multivariate outlier analysis was conducted using the adaptive outlier detection algorithm to remove outlier values from the data. Note that a selection of percentile value for outlier analysis should be done in consultation with an expert opinion since it has a serious implication for risk analysis. A nonlinear correlation analysis was conducted using the maximum information criterion algorithm to identify the relationship between independent and dependent variables. Based on the correlation analysis, heavy metals (Cu, Cd, and Zn) were predicted as a function of pH, alkalinity, sulfate, acidity, and flow. The predictive accuracy of ANN and SVM algorithms was evaluated using the Root Mean Squared Error, Mean Absolute Error, Root Relative Squared Error, and Relative Absolute Error. The epistemic and aleatory uncertainties in the prediction results were quantified using *P*-boxes and compared with the observed *P*-boxes graphically. A visual comparison of observed and predicted *P*-boxes was used as a measure of prediction uncertainty. The predictions of ANN and SVM were integrated using envelope, intersection, mixture, and average methods.

The results of this study show that the predictions of ANN enveloped well the observed Cu and Zn concentrations than the SVM prediction. On the other hand, both algorithms did not envelope the observed Cd concentrations. The prediction of Cd was not good because there was little heterogeneity in the Cd dataset since 90% of the dataset is below analytical detection limit. These results indicate that the success of machine learning techniques depends not only on the amount of data but also on heterogeneity within the data.

Integrating the prediction of ANN and SVM using the aggregation methods improved the prediction results except for Cd. While the envelope, mixture and average methods showed good performances, the intersection method showed poor performances. These results indicated that there is no best aggregation method, but rather analysts should investigate to determine which one improves predictions.

This study not only quantified prediction uncertainty, but also identified the sources of uncertainties, and whether there is a possibility to reduce them. In addition, the study showed the danger of selecting an optimal technique model using predictive accuracy and thus highlights the use of predictive uncertainty using *P*-boxes for selecting an optimal model.

In general, this study presented a novel methodology to apply machine learning techniques for the prediction of ARD chemistry in minesites and quantify prediction uncertainty. This methodology could be used as an integral part of ARD risk assessment and management framework.

References

Acuna E, Rodriguez C. The treatment of missing values and its effect on classifier accuracy. In: Banks D, et al, editors. Classification, clustering and data mining applications. Chicago: Springer Berlin Heidelberg; 2004. p. 639–47.

Aryafar A, Gholami R, Rooki R, Doulati Ardejani F. Heavy metal pollution assessment using support vector machine in the Shur River. Sarcheshmeh copper mine, Iran, Environ Earth Sci 2012;67(4):1191–9.

Azapagic A. Developing a framework for sustainable development indicators for the mining and minerals industry. J Clean Prod 2004;12(6):639–62. [Available at: http://linkinghub.elsevier.com/retrieve/pii/S0959652603000751 [Accessed May 24, 2013]].

Bello AL. Imputation techniques in regression analysis: looking closely at their implementation. Comput Stat Data Anal 1995;20(1):45–57.

Betrie GD, Tesfamariam S, Morin KA, Sadiq R. Predicting copper concentrations in acid mine drainage: a comparative analysis of five machine learning techniques. Environ Monit Assess 2012;185(5):4171–82.

Betrie GD. On the issue of incomplete and missing water-quality data in minesites databases: comparing three imputation methods. Mine Water Environ 2014. [Submitted for publication].

Bouckaert RR, Frank E, Hall M, Kirkby R, Reutemann P, Seewald A, Scuse D, et al. WEKA manual (version 3.6.4). University of Waikato, New Zealand: Hamilton; 2010.

Cherkassky VS, Mulier F. Learning from data: concepts, theory, and methods. New Jersey: John Wiley & Sons; 2007.

Cherkassky VS, Mulier F. Computational intelligence in earth sciences and environmental applications: issues and challenges. Neural Netw 2006;19(2):113–21.

Ferson S. RAMAS risk calc 4.0 software: risk assessment with uncertain numbers. Lewis Publishers; 2000 [Available at: http://www.ramas.com/riskcalc.htm].

Ferson S, Kreinovich V, Ginzburg L, Myers DS, Sentz K. Constructing probability boxes and Dempster–Shafer structures. SAND2002-4015. Albuquerque, NM: Sandia National Laboratories; 2003.

Filzmoser P, Garrett RG, Reimann C. Multivariate outlier detection in exploration geochemistry. Comput Geosci 2005;31(5):579–87.

Gideon RA, Hollister RA, Hollister A. A rank correlation coefficient resistant to outliers. J Am Stat Assoc 1987;82(398):656–66.

Gray NF. Field assessment of acid mine drainage contamination in surface and ground water. Environ Geol 1996;27(4):358–61. [Available at:http://link.springer.com/10.1007/BF00766705].

Gray NF. Acid mine drainage composition and the implications for its impact on lotic systems. Water Res 1998;32(7):2122–34. [Available at:http://linkinghub.elsevier.com/retrieve/pii/S0043135497004491].

Khandelwal M, Singh TN. Prediction of mine water quality by physical parameters. J Sci Ind Res 2005;64(August):564–70.

Lottermoser BG. Mine wastes characterization, treatment and environmental impacts third. Heidelberg Dordrecht London New York: Springer; 2010.

Maest AS, Kuipers JR, Travers CL, Atkins DA. Predicting water quality at hardrock mines: methods and models, uncertainty and state-of-the-Art. Kuipers & Associate and Buka Environmental; 2005.

Miller GC, Kempton H, Figueroa L, Pantano J. Engineering issue: management and treatment of water from hard rock mines. EPA/625/R-06/014. U.S. Environmental Protection Agency; 2006.

Mitchell T. Machine learning. McGraw Hill; 1997.

Morin KA, Hutt NM. Environmental geochemistry of minesite drainage: practical theory and case studies. Vancouver: MDAG Publishing; 2001.

Morin KA, Hutt NM, Aziz M. Case studies of thousands of water analyses through decades of monitoring: selected observations from three minesites in British Columbia, Canada. Proceedings of the 2012 International Conference on Acid Rock Drainage, Ottawa, Canada, May 22–24. Ottawa; 2012. [Available at:http://www.mdag.com/MDAGPaperDatabase/M0070-Morinetal2012-ThousandsofWaterAnalysesOverDecades.pdf.].

Price WA. Prediction manual for drainage chemistry from sulphidic geologic materials. Canadian Mine Environment Neutral Drainage (MEND), report 1.20.1; 2009.

Reimann C, Filzmoser P, Garrett RG. Background and threshold: critical comparison of methods of determination. Sci Total Environ 2005;346(1–3):1–16.

Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, et al. Detecting novel associations in large data sets. Science (New York, NY) 2011;334(6062):1518–24.

Rooki R, Doulati Ardejani F, Aryafar A, Bani Asadi A. Prediction of heavy metals in acid mine drainage using artificial neural network from the Shur River of the Sarcheshmeh porphyry copper mine. Southeast Iran, Environ Earth Sci 2011;64(5):1303–16.

RStudio. RStudio software. Available at http://www.rstudio.org, 2012.

Sentz K, Ferson S. Combination of evidence in Dempster–Shafer theory. SAND 2002–0835. Albuquerque, NM: Sandia National Laboratories; 2002.

Smola AJ, Scholkopf B. A tutorial on support vector regression. Report 1998-030. London: Royal Holloway College; 2003.

Solomatine DP, Ostfeld A. Data-driven modelling: some past experiences and new approaches. J Hydroinf 2008;10(1):3–22.

Templ M, Kowarik A, Filzmoser P. Iterative stepwise regression imputation using standard and robust methods. Comput Stat Data Anal 2011;55(10):2793–806.

Tucker WT, Ferson S. Probability bounds analysis in environmental risk assessments. Setauket, NY: Applied Biomathematics; 2003.

Vapnik V. Statistical learning theory. New York: Willey; 1998.

Vapnik V. The nature of statistical learning theory. Springer; 2000.

Walley P. Statistical reasoning with imprecise probabilities. New York: Chapman and Hall; 1991.