Environmental Science and Pollution Research (2019) 26:1821–1833
https://doi.org/10.1007/s11356-018-3751-y

RESEARCH ARTICLE

Evaluation of the bias and precision of regression techniques and machine learning approaches in total dissolved solids modeling of an urban aquifer

Conglian Pan[1] · Kelvin Tsun Wai Ng[1] · Bahareh Fallah[1] · Amy Richter[1]

Received: 20 August 2018 / Accepted: 12 November 2018 / Published online: 19 November 2018

Abstract

TDS is modeled for an aquifer near an unlined landfill in Canada. Canadian Drinking Water Guidelines and other indices are used to evaluate TDS concentrations in 27 monitoring wells surrounding the landfill. This study aims to predict TDS concentrations using three different modeling approaches: dual-step multiple linear regression (MLR), hybrid principal component regression (PCR), and backpropagation neural networks (BPNN). An analysis of the bias and precision of each models follows, using performance evaluation metrics and statistical indices. TDS is one of the most important parameters in assessing suitability of water for irrigation, and for overall groundwater quality assessment. Good agreement was observed between the MLR1 model and field data, although multicollinearity issues exist. Percentage errors of hybrid PCR were comparable to the dual-step MLR method. Percentage error for hybrid PCR was found to be inversely proportional to TDS concentrations, which was not observed for dual-step MLR. Larger errors were obtained from the BPNN models, and higher percentage errors were observed in monitoring wells with lower TDS concentrations. All models in this study adequately describe the data in testing stage ($R^2 > 0.86$). Generally, the dual-step MLR and hybrid PCR models fared better ($R^2_{\text{avg}} = 0.981$ and 0.974, respectively), while BPNN models performed worse ($R^2_{\text{avg}} = 0.904$). For this dataset, both regression and machine learning models are more suited to predict mid-range data compared to extreme values. Advanced regression methods (hybrid PCR and dual-step MLR) are more advantageous compared to BPNN.

Keywords Total dissolved solids · Artificial neural network · Principal component regression · Multivariate statistical analysis · Machine learning methods · Bias and precision

Introduction

Groundwater quality assessments based on water quality indices, drinking water standards, and irrigation guidelines are regularly conducted due to their practical importance. Hassen et al. (2016) analyzed groundwater quality in Tunisia using several water quality indices based on World Health Organization drinking water guidelines. Pan et al. (2017) and Pan and Ng (2018) adopted the Canadian Drinking Water Guideline and other indices to assess groundwater quality near an unlined landfill in Canada. Statistical approaches are often adopted and integrated in water quality studies due to complexity of the hydrogeological environment and interactions between the water constituents in subsurface environments. Specifically, multivariate statistical approaches such as cluster analysis, hierarchical cluster analysis, and principal component analysis are commonly applied to classify groundwater constituents and examine their correlations. Viswanath et al. (2015) proposed a prediction model for total dissolved solids (TDS) concentrations in watersheds by combining principal component analysis (PCA) with multiple linear regression (MLR). This approach is known as principal component regression (PCR) and has been successfully applied in solid waste generation rate prediction (Azadi and Karimi-Jashni 2016), oil refinery forecasts (Rashid

Responsible editor: Marcus Schulz

✉ Kelvin Tsun Wai Ng
kelvin.ng@uregina.ca

[1] Environmental Systems Engineering, University of Regina, Regina, SK S4S 0A2, Canada

et al. 2017), and energy system reliability assessment (Solanki et al. 2018). It appears that there are very limited studies on the use of PCR on groundwater quality assessment, despite the versatility and robustness of the method. In this study, an improved principal component regression model (hybrid PCR) which integrates PCR with machine learning ideology is proposed to simulate groundwater TDS concentration in an urban aquifer near an unlined municipal solid waste landfill in Saskatchewan, Canada. Saskatchewan has the lowest diversion rate among the Western Canadian provinces and relies heavliy on landfill technology (Pan et al. 2018). Unlike other PCR prediction models utilizing an entire dataset for model development with no validation, the proposed hybrid PCR model is developed by a stand-alone training dataset, and validated with an independent testing dataset. The proposed hybrid PCR is believed to be more advantageous in groundwater quality studies, where interactions between the constituents and the potential collinearity between the parameters are expected. The use of a non-overlapping testing dataset allows the proposed model to be validated independently (Tan et al. 2016) and is commonly employed in machine learning techniques.

In addition to the regression methods such as MLR and PCR, machine learning approaches such as artificial neural networks have been increasingly popular in environmental studies in the past decades. A number of studies attempted to compare results between the machine learning models with the conventional multivariate regression models. Sahoo and Jha (2013) developed 17 site-specific MLRs for water-elevation prediction in a basin located in Shikoku Island, Japan, and compared the results with artificial neural network techniques. They found that their backpropagation neural network (BPNN) models provided better results than MLR for most sites; however, MLR was also recommended as an alternate cost-effectiveness tool. Azadi and Karimi-Jashni (2016) compared MLR and artificial neural network prediction models for seasonal solid waste generation rates in Fars Province, Iran, and concluded that the non-linear BPNN model provided more accurate results. It appears that artificial neural network methods are comparable or better than the conventional MLR methods in several chemistry and environmental studies (Civelekoglu et al. 2007; Xu et al. 2011; Ebrahimi and Rajaee 2017). In this paper, prediction results from advanced regression and machine learning models are examined.

The objectives of this study are to (i) develop a robust hybrid PCR approach utilizing a non-overlapping testing dataset; (ii) predict TDS concentrations of an aquifer using dual-step MLR, hybrid PCR, and BPNN models; and (iii) examine the bias and precision of the methods, and systematically compare the results using a set of performance evaluation metrics and statistical indices.

Materials and methods

Regina landfill and Condie aquifer

Regina, the capital city of Saskatchewan, is located in a semi-arid prairie region at 50°26′ N and 104°37′ W (Fig. 1). Regina has a land area of 180 km^2 with a total population of about 215,000 residents in 2016, representing 20% of the province's population (Statistics Canada 2018). The city is located in the western sedimentary basin, and bedrock in the area consists of marine shales, evaporates, and mudstones. Soil around the landfill area is dominated by clayey glaciolacustrine parent material, as well as expanding clay minerals which result in high fracturing near the surface (City of Regina 2009, 2011, 2014, 2015, 2016). Groundwater plays an important role in water supply in Saskatchewan, especially in areas where the availability of surface water is limited, such as south of the boreal shield region (City of Regina 2002, 2013; Pomeroy et al. 2005). The Condie aquifer, located near the city, provides water to the city for anthropogenic and industrial uses. The thickness of the Condie aquifer is quite shallow, and the groundwater table is about 8–10 m below ground surface (City of Regina 2016) protected by the lacustrine clay and silt above (City of Regina 2009, 2011, 2014, 2015, 2016).

One of the potential contaminant sources of the Condie Aquifer is the City of Regina Landfill, located at the North-East corner of the city (Fig. 1). Built in the 1960s, the old cells occupy over 60 ha without an engineered liner (Vu et al. 2017; Bruce et al. 2018). The regional groundwater flow is towards the west and southwest, with a flow rate of 400 m/year (City of Regina 2002). The saturated hydraulic conductivity of the aquifer varies from 1.4×10^{-6} m/s to as high as 2.3×10^{-3} m/s, depending on the saturated thickness and effective grain size of the soil (City of Regina 2016).

Data sampling and grouping

The City of Regina started the groundwater monitoring program at the study area in the 1970s. Over the years, monitoring wells in the vicinity of the landfill were installed and decommissioned, and a total of 27 monitoring wells are currently in operation (Fig. 1). Water sampling data used in this study is collected from City of Regina groundwater monitoring program reports between 2008 and 2015 (City of Regina 2009, 2011, 2012, 2013, 2014, 2015, 2016). In 2013 and 2014, however, only 44% and 33% of well data were respectively reported due to maintenance schedules, well-drying, and decommission. As a result, 2013 and 2014 data sets were excluded from the study to minimize data skewing.

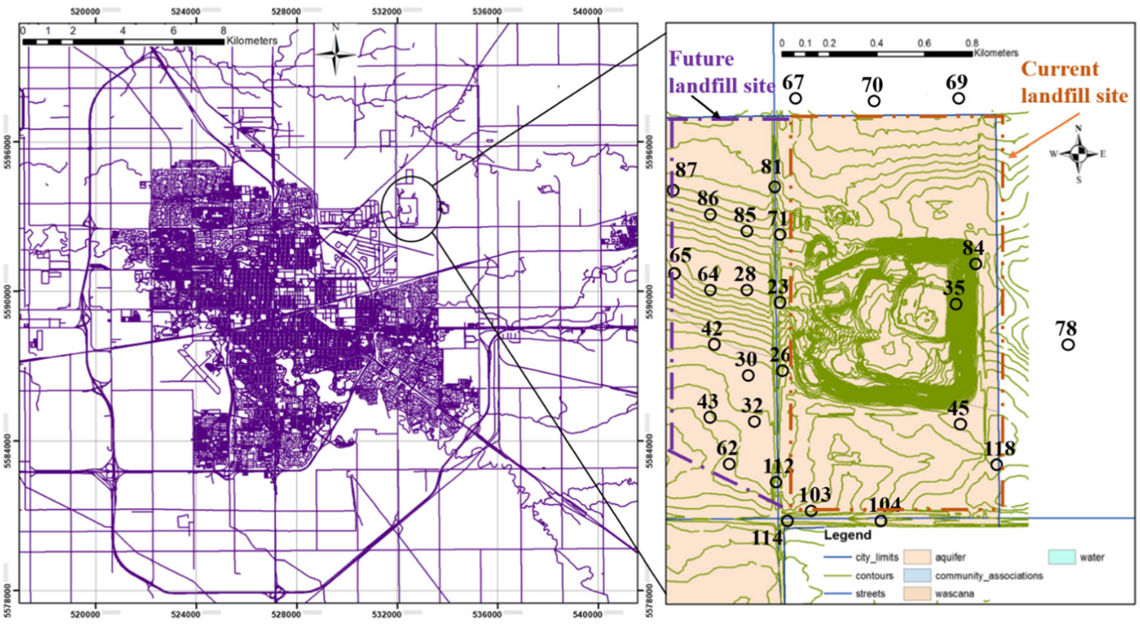

Fig. 1 Location map of Regina and Regina landfill using ArcMap (ver. 10.4.1)

During the study period, some data would occasionally be missing. For example, chloride is not reported for a monitoring well (ID 35) in 2011. Moreover, three monitoring wells south of the disposal area (IDs 112, 114, and 118) were not in operation until 2011. A total of 151 datasets from 27 monitoring wells (monitoring well IDs 23, 26, 28, 30, 32, 35, 42, 43, 45, 62, 64, 65, 67, 69, 70, 71, 78, 81, 84, 85, 86, 87, 103, 104, 112, 114, and 118) were considered, and the samples indexed continuously from no. 1 to no. 151. Figure 1 shows the location of the wells. The monitoring wells were categorized into five groups according to their respective locations of the unlined landfill:

1. Background Group (dataset index no. 1 to 24) is located upstream, outside of the landfill site boundary. It includes four monitoring wells: 67, 69, 70, and 78.
2. East Group (dataset index no. 25 to 44) is located within the landfill area along the East boundary. It includes four monitoring wells: 35, 45, 84, and 118.
3. South Group (dataset index 45 to 63) is located immediately downstream, south of the landfill site. This group includes four monitoring wells: 103, 104, 112, and 114.
4. West Group (dataset index 64 to 109) is located downstream, to the west side of landfill. It includes eight monitoring wells: 23, 26, 28, 30, 32, 71, 81, and 85.
5. Far West Group (dataset index 110 to 151) is located in downstream area in the west, furthest away from the site. It includes seven monitoring wells: 42, 43, 62, 64, 65, 86, and 87.

In this study, a 70:30 ratio (training/testing) is used for the prediction models. To reduce modeling errors and avoid possible bias from the inputs, five different datasets were derived from the original dataset randomly, each with their own training and testing subsets. As such, each of the three modeling approaches (dual step-MLR, hybrid PCR, and BPNN) is evaluated five times with the derived datasets. This approach allows a fair and systematic assessment of the methods and modeling precision by applying consistent training and testing inputs in multiple trials. This also helps to identify and to reveal possible bias from the input data.

Groundwater parameters and indicators

Groundwater parameters are carefully chosen according to their significance, data availability, and concentrations with respect to the guideline values (Saskatchewan Ministry of Environment 2016; Health Canada 2017). In this study, a total 14 parameters are selected, including heavy metals such as arsenic (As), calcium (Ca), magnesium (Mg), manganese (Mn), potassium (K), sodium (Na), and uranium (U), as well as ionic species and general parameters: bicarbonate (HCO_3), chloride (Cl), sulfate (SO_4), total dissolved solids (TDS), total hardness (TH), pH, and electric conductivity (EC). For all modeling and analyses, TDS is selected as the target parameter, as its concentration is affected by many of the studied parameters (Sherrard et al. 1987; Xun et al. 2007) and the remaining 13 parameters are the input parameters used to build the prediction model with respect to TDS. Given the physical-chemical interactions of the selected species at the Condie aquifer (Pan et al. 2017), correlations between variables are expected and advanced numerical techniques are warranted.

Dual-step multiple linear regression

MLR is a statistical technique which is used to establish a linear relationship between one or more independent (or explanatory) variables and a dependent (or response) variable. The general expression form of MLR can be written as below (Bingham and Fry 2010):

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots\cdots + \beta_k X_{k,i} + \varepsilon_i \tag{1}$$

where

Y_i	dependent variable (response variable, TDS in this study)
$X_{1,\mathrm{i}}, X_{2,\mathrm{i}}, \ldots\ldots, X_{\mathrm{k},\mathrm{i}}$	regressors (the *i*th observations of each of the independent variables)
$\beta_0, \beta_1, \ldots\ldots, \beta_\mathrm{k}$	the coefficients of each regressors which are unknown but fixed values
ε_i	the noise (measurement error)

In this study, SPSS (v. 25) is used to identify significant parameters, and to obtain the coefficients $\beta_0, \beta_1, \ldots\ldots, \beta_k$ for regressors. Ideally, the target parameter should be normally distributed and all of the explanatory variables should be independent of each other. The Kolmogorov-Smirnov test (K-S test) has been commonly used in waste studies (Chickering et al. 2018; Han et al. 2018) and is adopted to evaluate the normality of the TDS data. Multicollinearity of the explanatory variables is identified using the critical correlation coefficient R_crit, defined as follows (Sousa et al. 2007; Azadi and Karimi-Jashni 2016):

$$R_\mathrm{crit} = \frac{t_\mathrm{crit}}{\sqrt{df + t_\mathrm{crit}^2}} \tag{2}$$

and

$$df = n{-}k \tag{3}$$

where df is the degree of freedom of the analysis, n represents the number of datasets, and k is the number of comparing variables. In this study, a two-tailed test with significance of 0.05 is adopted, corresponding to an R_crit of 1.6.

MLR does not only generate the linear relationships among parameters but also identifies parameters which contribute to the target parameter (Kicsiny 2016). In this study, a dual-step MLR approach is attempted to better model TDS. Dual-step MLR involves two steps: to conduct MLR on TDS using the 13 variables as independent variables and TDS as dependent variable. During this step, variables that has a statistical significance on impacting TDS will be identified: (i) to conduct MLR on TDS using the 13 selected variables and (ii) to run the regression again using only the significant variables obtained from the first step. In this study, 95% significance is adopted, and only parameters with significance equal to or greater than 95% (p value < 0.05) are used. Five trials are conducted based on different values.

Principal component analysis and regression

PCA is commonly used in environmental studies to reduce the number of variables, extract useful information, and eliminate the noise from data. PCA extracts eigenvalues from the original dataset and forms new principal components (PC) that are linear combinations of the parameters. The resulting PCs are orthogonal to each other after Varimax Rotation (Ravikumar and Somashekar 2017; Abou Zakhem et al. 2017), which helps to avoid multicollinearity between model parameters. PCs with eigenvalues greater than unity are considered significant (Cattell and Jaspers 1967; Abou Zakhem et al. 2017; Selvakumar et al. 2017), and each significant PC explains a portion of the total variance of the dataset. A combination of all significant PCs should explain no less than 80% of the total variance for sufficient coverage (Zhao et al. 2012; Hu et al. 2013; Viswanath et al. 2015; Selvakumar et al. 2017). In this study, only PCs with eigenvalues greater than unity are adopted to build models explaining at least 85% of total variances.

To conduct PCR, the PCs identified by PCA are used as independent variables in MLR. PCR is more advantageous than conventional MLR modeling since it retains more original predictor variables and minimizes multicollinearity between variables. Unlike other PCR studies, independent training and testing datasets are separately utilized in the proposed hybrid PCR. For a given trial, PCs on TDS are first identified from the training data set and MLR is carried out using the significant PCs (total variance > 85%) to obtain a TDS prediction model. Likewise, a different set of PCs are obtained using another testing data set following the same PCA loading vectors for the training set, and they are substituted to the original MLR equation derived from the training dataset to obtain another TDS value for validation purposes. PCA is conducted using R (ver. 3.5.1), from which PCs and loading vectors are obtained, and the MLR equation is obtained using SPSS (ver. 25).

Backpropagation neural network models

A basic BPNN structure contains three layers: the input layer, the hidden layer, and the output layer. The hidden layer represents the transferring function and relationships between the inputs and outputs (Chen et al. 2010). In every layer, processing units that contain values are known as nodes. Weight and bias are assigned to each iteration according to the membership functions to approximate the output. The output therefore can be obtained as the sum of the weighted inputs as shown below (Sahoo and Jha 2013):

$$Y_k = f\left(\sum_i W_{ij} X_i + \theta_j\right) \quad (4)$$

where Y_k is the output at node k, $f(\cdot)$ is the transferring function, W_{ij} is the weight applied between node i and j, X_i is the input at node i, and θ_j is the bias at node j.

While a BPNN model can contain multiple hidden layers, it is found that in most cases, one hidden layer is sufficient to provide the required accuracy (Azadi and Karimi-Jashni 2016). In this study, the same input and target variables are used as other methods, and a BPNN structure of 13-10-1 is adopted for all 5 trials, representing 13 input parameters, 10 nodes in hidden layer, and 1 output variable (TDS). Comparisons between single- and double-hidden layer are conducted, and it is found that single-hidden layer BPNN models provide better TDS estimates than a BPNN trial using double-hidden layer. As such, BPNNs with one hidden layer are adopted to avoid overfitting issues. A major concern of conducting neural network analysis is overfitting, meaning that the network “memorizes” certain combinations during training stage instead of “learning” and building a proper algorithm. A well-fitted model during training and poorly fitted during testing are possible indicators of overfitting. To minimize model overfitting, an early stopping technique is adopted by properly distributing the inputs for all trials: training (70%), testing (15%), and validating (15%). A total of 5 trials of BPNN are conducted, using the same structure of 13-10-1, representing 13 input variables, 10 nodes in hidden layer, and 1 output variable. The Levenberg-Marquardt Backpropagation training method is adopted. BPNN analysis is performed using MATLAB (ver. R2016a).

Model performance and error quantification

A number of statistical indicators are used to examine the characteristics of the models and to quantify the accuracy and precision of results. R^2 describes the portion of variance explained by the linear model. P value in ANOVA model indicates the significance of the models (Sahoo and Jha 2013; Hanley 2016). In the present study, a confidence interval of 95% is used ($p < 0.05$). When evaluating MLR, R^2 and adjusted R^2 are both used:

$$R^2 = \frac{\Sigma\left(\widehat{Y}_i - \overline{Y}\right)^2}{\Sigma\left(Y_i - \overline{Y}\right)^2} = \frac{\mathrm{SSR}}{\mathrm{SST}} \quad (5)$$

and

$$\text{Adjusted } R^2 = 1 - \left(\frac{n-1}{n-p-1}\right)\left(1-R^2\right) \quad (6)$$

where SSR is the sum of squares regression, SST is the total sum of squares, Y_i is the observed values of the target (dependent) variable, $\widehat{Y}_i$ is the estimated (predicted) values, $\overline{Y}$ is the average value of a dependent variable, n is the number of observations, and p is the number of independent variables.

R^2 provides an intuitive measurement between the observed and modeled values, but may be less applicable for non-linear problems (Azadi and Karimi-Jashni 2016). Mean absolute error (MAE) measures how close the predicted values are to the observed values, and provides the mean value of the model errors. A MAE closer to zero represents better model performance. Root mean squared error (RMSE) describes the global discrepancy between the predicted and observed values (Zhao et al. 2011) and is more sensitive to erroneous data. Similar to MAE, a smaller RMSE denotes a better model performance.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right| \quad (7)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2} \quad (8)$$

In addition, a simple error percentage is also used:

$$\text{Error percentage} = \frac{\left|\hat{Y}_i - Y_i\right|}{Yi} \times 100\% \quad (9)$$

In the above equations, Y_i is the observed values of the target (dependent) variable, $\widehat{Y}_i$ is the estimated (predicted) values, and n is the number of observations.

Results and discussion

A summary of parameter concentrations from all wells during the study period is listed in Table 1. The mean concentrations of manganese, sulfate, TDS, and TH are considerably higher than guideline values. It is believed that the operation of the unlined landfill has impacted groundwater quality (Pan et al. 2017; Pan and Ng 2018). TDS measures the total organic and inorganic dissolved substances. It is affected by many parameters in the study and is the target parameter in the present study. TDS is also one of the most important parameters in assessing suitability of water for irrigation (Atta et al. 2018) and for overall groundwater quality assessment (Li et al. 2018).

Table 1 Parameter concentrations and guideline values

	Canadian Guidelines (mg/L)	All 27 monitoring wells			
		Max	Mean	Min	STD[c]
Trace metals (mg/L)					
Arsenic	0.01	0.028	0.006	0.0002	0.006
Calcium	No value[a]	500	302.013	95	106.651
Magnesium	No value[a]	220	112.391	29	47.268
Manganese	0.05 (0.1[b])	190	*2.496*	0.04	15.320
Potassium	No value[a]	86	12.988	4.5	10.344
Sodium	200	380	62.169	16	63.076
Uranium	0.02	0.056	0.016	0.0001	0.009
General parameters (mg/L)					
Bicarbonate	No value[a]	850	418.026	220	111.373
Chloride	250	580	67.526	1.3	112.774
Sulfate	250	1800	*927.232*	170	414.860
Total dissolved solids	1000	3400	*1699.795*	460	700.970
Total hardness $CaCO_3$	500	2000	*1216.556*	360	456.957
Lab pH (–)	6.5–8.5	8.18	7.758	7.27	0.179
Lab conductivity (µs/cm)	No value[a]	4800	2173.576	740	844.081

Concentrations exceeding guideline values are italicized

[a] "No value" indicates no health-related maximum allowable concentrations provided

[b] Indicating WHO guidelines

[c] Standard deviation

Dual-step multiple linear regression

Results obtained from five dual-step MLR models indicate that only several parameters are identified by dual-step MLR to be statistically significant to affect the concentration of TDS. Among the four to six variables identified from all five trials, sodium (Na) and sulfate (SO_4) are two variables identified to be significant in most of the dual-step MLR models. The prediction models are derived by the five trials, and the models are shown in Table 3. It is found that all five MLR models adequately describe the observed TDS concentrations, especially MLR1, MLR4, and MLR5, each with satisfactory R^2, MAE, and RMSE in both training and testing stages. The performance indices of all dual-step MLR models are tabulated in Table 7 and are further discussed in "Evaluation of model performances by statistical indices".

To evaluate the validity of the regression results, correlation analysis is conducted to ensure the target parameter is statistically correlated with the eight independent variables identified by MLR. As shown in Table 2, all statistically significant correlation coefficients are positive, except for pH (− 0.64). This is probably due to a higher amount of dissolved organic and inorganic pollutants present in acidic environment. TH, EC, and Ca have the highest coefficients (≥ 0.95) with respect to TDS concentrations, suggesting stronger correlations with TDS. Moreover, the absolute values of the coefficient are all greater than $R_{crit} = 0.16$, indicating that the relationships between each explanatory variable and TDS are statistically significant (meeting or exceeding the 95% significance).

Figures 2 and 3 compare the modeled results (dual-step MLR, hybrid PCR, and BPNN) with the observed TDS data. Figure 2a graphically compares the predicted values with the observed data for MLR1. There is good agreement between the MLR1 results and field data. The scatter plot (Fig. 2a) shows a very linear relationship ($R^2 = 0.988$) for the entire TDS concentration range (500 mg/L to 3500 mg/L). In Fig. 3a, both the predicted and observed concentrations are plotted using the left-hand-side vertical axis, whereas the percentage error is plotted using the right-hand-side axis. The dual-step regression model MLR1 adequately describes the data, with an average percentage error of 3.7%. A peak is observed at index 127, giving a maximum percentage error of 38.5%. It appears that the Far West Group (index 110–151)

Table 2 Correlation coefficients of statistically significant explanatory variables on TDS

	Ca	Na	HCO_3	Cl	SO_4	TH	pH	EC
TDS	0.95	0.67	0.72	0.64	0.93	0.97	− 0.64	0.97

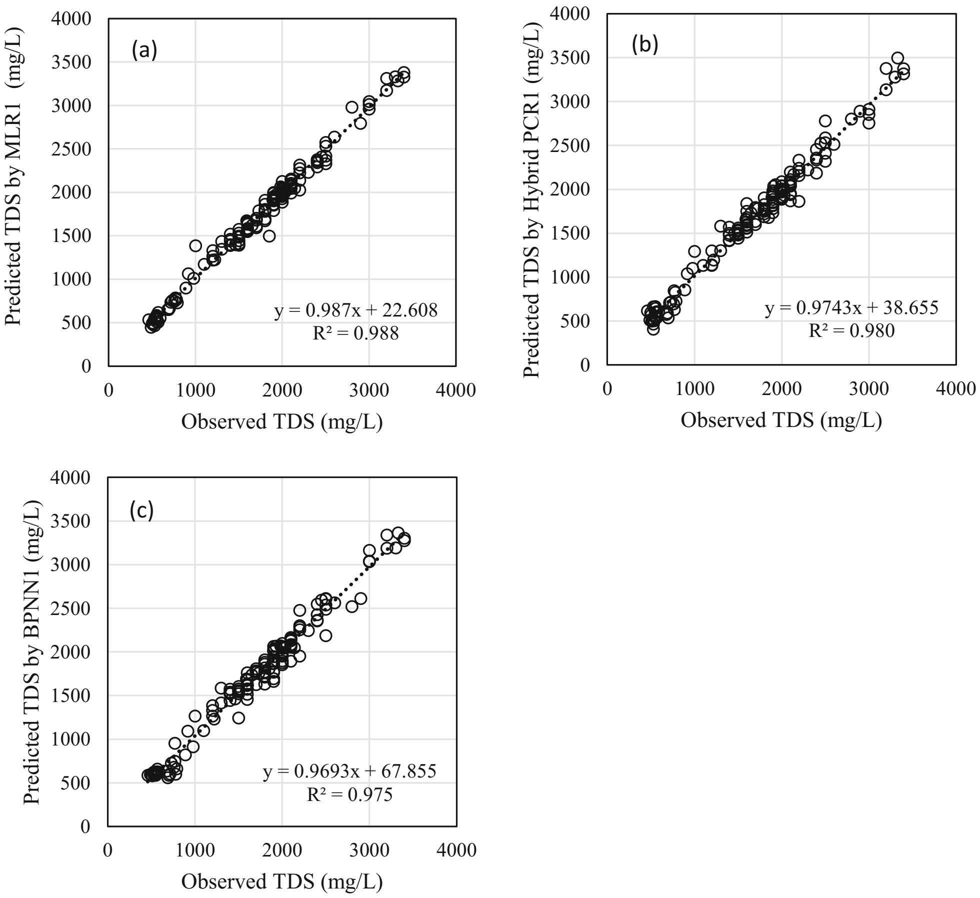

Fig. 2 Performance of the three approaches (MLR1, PCR1, and BPNN1)

has slightly higher percentage errors. No apparent trend is observed between the magnitudes of percentage error and the observed TDS concentrations.

All MLR models assume non-collinear relationship among independent variables; however, the justification of nonlinearity is difficult given the frequent interactions of organic and inorganic constituents in subsurface environments. A check regarding the collinearity among explanatory variables was conducted, and the results from MLR1 are shown in Table 4. Only the absolute values of correlation coefficient larger than the $R_{\mathrm{crit}} = 0.16$ (from Eq. 2) are bolded. It is found that multicollinearity exists among some explanatory variables (Ca-Cl, and Na-SO_4) in MLR1, and may affect the accuracy of the final results. Multicollinearity issue was also found in other MLR trials in this study.

K-S test results suggested that TDS set from the five trials are likely not distributed normally, as the significance of the trials are all less than 0.05 (Table 5). This is probably due to the variability of TDS data in this study (Table 1). Extreme TDS values are consistent with the groundwater flow regime, with the lowest TDS in the Background Group (upstream of the unlined landfill) and the highest TDS in the West Group (downstream). In addition, the mean is smaller than the median in the original TDS data set, and the set is skewed to the left. Although the MLR models adequately describe the field data, multicollinearity exists in the model parameters, and the target variable is likely not distributed normally.

Hybrid principal component regression

PCR models are constructed to minimize multicollinearity of the explanatory variables. In any given set, the largest four PCs provide sufficient coverage of variance (> 85%). The accumulated covariance for hybrid PCR model 1 to 5 are 85.8%, 85.8%, 86.2%, 85.9%, and 86.6%, respectively. MLR is then conducted using the four PCs as independent variables. ANOVA test reveals p value < 0.001 for all models, indicating that the models are statistically valid. All four PCs are effectively contributing to the target parameter, with p values for all factors < 0.001. The hybrid PCR models are shown in Table 6. It is worth noting that the absolute values of the PC1 coefficients are at least three times larger than the PC2 coefficients, and the sign of coefficients varied among trials. The modeling results, however, are quite consistent.

All five hybrid PCR models are validated by both of the testing and the entire dataset (Table 7), as discussed in

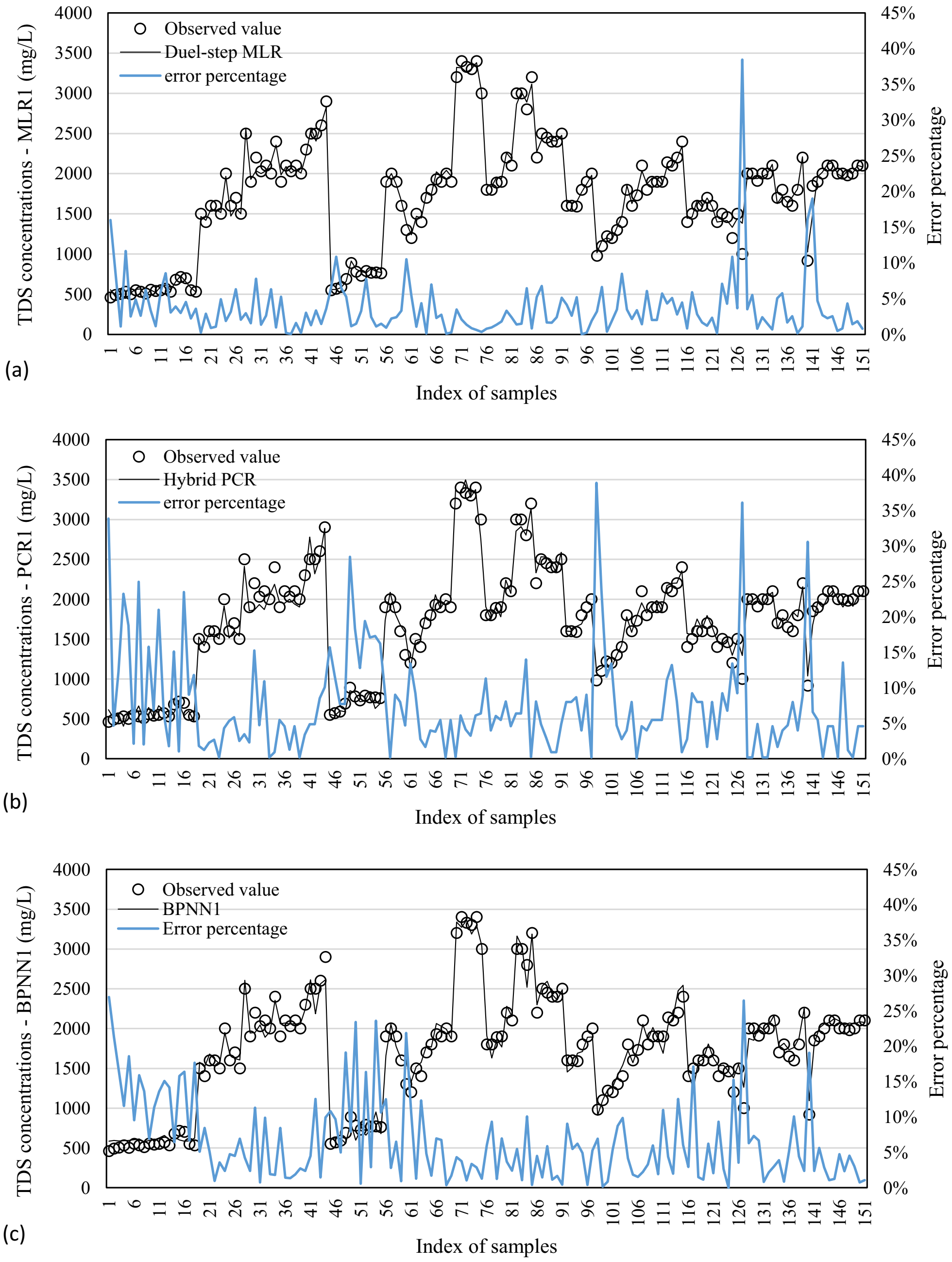

Fig. 3 Predicted TDS and error percentages regarding the geospatial locations of wells

“Evaluation of model performances by statistical indices”. Figure 2b graphically compares the hybrid PCR1 results. The scatter plot shows a good linear relationship (R^2 = 0.980). Slightly more data points are located below the 1:1 line (not shown), suggesting the tendency of the hybrid PCR1 model to underestimate the target parameter.

Table 3 Dual-step MLR prediction models

Trials	MLR equations
MLR1	$TDS = -34.886 + 2.61 \times Ca + 1.315 \times Na + 0.598 \times Cl + 0.484 \times SO4 + 0.173 \times EC$
MLR2	$TDS = 1804.031 + 2.525 \times Na - 0.020 \times HCO3 + 0.305 \times Cl + 1.208 \times TH - 223.878 \times pH$
MLR3	$TDS = 2216.516 + 0.831 \times Na + 1.583 \times Cl + 1.273 \times SO4 - 238.356 \times pH$
MLR4	$TDS = -57.579 + 2.624 \times Na + 0.335 \times HCO3 + 0.554 \times SO4 + 0.776 \times TH$
MLR5	$TDS = 615.925 + 2.318 \times Ca + 1.591 \times Na + 0.497 \times HCO3 + 0.508 \times SO4 - 97.832 \times pH + 0.170 \times EC$

Table 4 Correlation coefficient of variables and multicollinearity check for MLR1

	EC	Cl	Ca	Na	SO_4
EC	1	*− 0.437*	*− 0.541*	*− 0.408*	*− 0.454*
Cl		1	0.017	*− 0.589*	*0.431*
Ca			1	*0.403*	*− 0.455*
Na				1	− 0.053
SO_4					1

Coefficients larger than R_{crit} are italicized

Percentage errors are generally larger than dual-step MLR models, with an average of 7.6% (Fig. 3b). Unlike the MLR1 model (Fig. 3a), it appears that the percentage errors of the PCR1 model are inversely related to the TDS concentrations (Fig. 3b). PCR1 model shows larger percentage errors in the Background group (index 1–14), probably due to the smaller observed TDS values. More peaks in percentage errors are observed than the MLR1. The maximum value is observed at index 98 (West Group), corresponding to a percentage error of 38.9% (Fig. 3b).

Prediction and performance of BPNN

Compared to dual-step MLR and hybrid PCR models, larger errors are obtained from the BPNN models. Results from BPNN1 are used to demonstrate the model performance. The scatter plot (Fig. 2c) indicates a general linear relationship ($R^2 = 0.975$) between the observed and the predicted values. Although BPNN1 is capable of describing the data in general, more data scattering is observed between 2000 and 3000 mg/L. By comparing the slopes of the best fit line to the 1:1 line (not shown), BPNN1 tends to overestimate TDS in the lower concentration range and underestimate in the higher concentration range.

Great variations in error percentage are observed in Fig. 3c, with an average percentage error of 6.5%. A less obvious inverse relationship between the magnitudes of percentage error and TDS is again observed. Higher percentage errors are generally observed in the monitoring wells with lower TDS concentrations, such as the Background group (index 1–14) and the South group (index 45–63). Similar to both regression models, a peak is observed at index no. 127, corresponding to a percentage error of 26.5%. A maximum percentage error of 26.9% occurs at index no. 1 (Fig. 3c).

Table 5 Normality check on TDS for dual-step MLR models

	Kolmogorov-Smirnov		
	Statistic	*df*	Sig.
MLR1	0.090	151	0.004
MLR2	0.091	151	0.004
MLR3	0.079	151	0.023
MLR4	0.082	151	0.015
MLR5	0.084	151	0.011

df degree of freedom, *sig.* significance

Evaluation of model performances by statistical indices

Models are evaluated using performance indices including R^2, MAE, and RMSE for training, testing, and model fitting stages, as shown in Table 7. The best performance trials (i.e., highest R^2 value, or lowest MAE and RMSE values) for each model are bolded. In general, indices during the training stage are not good references for model evaluation accuracy, as inputs from train dataset make the model "aware" of the input values (Bagheri et al. 2017). Model performance in the testing and validation stages provides more insight on the model accuracy and their capability of fitting complex system. R^2 of the models between the training and testing stages are comparable, and "overfitting" issues are not observed (Table 7). Compared to the training stage, slightly larger errors (MAE and RMSE) are observed in the testing stage, with exceptions in some MLR and PCR models, probably due to the characteristics of datasets.

Performance indices from the testing stage are used in the present work to assess the model performance and bias. Performance indices at the model fitting stage (i.e., the complete data set) are included for comparison purposes. It is found that all models adequately describe the observed data in the testing stage, with the R^2 consistently greater than 0.86. R^2 of BPNN models are, however, lower than the statistical regression approaches. The average R^2 values of dual-step MLR models ($R^2{}_{\mathrm{avg}} = 0.981$) and hybrid PCR models ($R^2{}_{\mathrm{avg}} = 0.974$) are higher than the BPNN models ($R^2{}_{\mathrm{avg}} = 0.904$). For a given trial, the model inputs are identical irrespective of the modeling approach; however, no obvious trend is observed among the trials.

The level of disagreement between the predicted and observed TDS of the five BPNN models ($\mathrm{MAE}_{\mathrm{avg}} = 145.875$) is considerably higher than the PCR and MLR models. For example, the highest MAEs of the regression models in this study are 96.210 (MLR2) and 97.553 (PCR2), whereas the highest MAE of the machine learning neural network model is 211.882 (BPNN5). Moreover, it is found that the precision of the hybrid PCR results is generally better than MLR and BPNN models. The variability of the hybrid PCR's performance indicators is more consistent among the five trials. MAE and RMSE results are similar to R^2 results in all stages. It is found that the regression approaches used in this study adequately describe the TDS data and outperform a machine learning method.

Table 6 Hybrid PCR prediction models

Trials	Hybrid PCR equations
Hybrid PCR1	$\mathrm{TDS} = 1754.858 - 257.971 \times \mathrm{PC1} - 78.145 \times \mathrm{PC2} - 33.133 \times \mathrm{PC3} + 29.469 \times \mathrm{PC4}$
Hybrid PCR2	$\mathrm{TDS} = 1689.274 - 244.061 \times \mathrm{PC1} + 59.241 \times \mathrm{PC2} - 16.552 \times \mathrm{PC3} - 20.784 \times \mathrm{PC4}$
Hybrid PCR3	$\mathrm{TDS} = 1783.132 - 248.295 \times \mathrm{PC1} + 71.535 \times \mathrm{PC2} + 24.981 \times \mathrm{PC3} + 27.244 \times \mathrm{PC4}$
Hybrid PCR4	$\mathrm{TDS} = 1634.047 + 263.709 \times \mathrm{PC1} - 70.708 \times \mathrm{PC2} - 21.321 \times \mathrm{PC3} - 17.648 \times \mathrm{PC4}$
Hybrid PCR5	$\mathrm{TDS} = 1739.132 - 252.759 \times \mathrm{PC1} - 75.037 \times \mathrm{PC2} - 31.487 \times \mathrm{PC3} + 2.027 \times \mathrm{PC4}$

The overall performance of the models can be assessed using percentage error. The first row in Table 8 shows the observed TDS data in the field. The skewness of the original TDS set is captured in the models, where medians are larger than the means in all cases. Better modeling results are obtained at mid-range TDS concentrations. With the exception of BPNN3, the percentage errors are generally low (about or less than 2%) for the mean and median values. Lower percentage errors are observed in the dual-step MLR and hybrid PCR models.

Larger percentage errors are observed in the maximum and minimum TDS values in all three modeling approaches. For instance, the minimum predicted TDS value by BPNN3 is 201.81 mg/L, less than half compared to the observed data. It appears that regression and machine learning models are more suited to predict mid-range data than extreme values. The precision of the BPNN models among trials is not as good as the regression models, and larger error ranges are observed. For example, the BPNN percentage error ranges regarding the maximum and minimum TDS are 1.05–20.68%, and 1.0–56.1%, respectively.

Unlike some studies (Sahoo and Jha 2013; Azadi et al. 2016) which reported better performance of machine learning approaches than regression models, this study found that BPNN, a machine learning method, is not superior to the regression models considered in this study. Advanced regression methods such as the dual-step MLR and hybrid PCR are better in terms of model accuracy and precision, at least using TDS concentrations of an urban aquifer considered in this study.

The dual-step MLR models perform well and are less sensitive to the variability of the inputs. However, collinearity of the explanatory variables is difficult to eliminate given the nature of the groundwater parameters. Hybrid PCR models eliminate collinearity issues and provide reasonable estimates of the target parameter. The proposed hybrid PCR approach is more appropriate for complex systems with multiple and interconnected variables. The results are promising, and the

Table 7 Performance indices in training, testing, and model fitting stages

	Training			Testing			Model fit		
	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE
MLR1	0.988	53.932	78.697	*0.990*	*52.269*	*66.872*	0.988	53.481	75.370
MLR2	0.976	77.918	103.476	0.972	96.210	128.463	0.975	83.372	111.510
MLR3	0.965	73.516	127.256	0.986	67.014	83.281	0.973	71.556	115.910
MLR4	0.991	*43.259*	67.128	0.973	68.047	101.189	0.987	50.656	78.833
MLR5	*0.992*	44.481	*62.829*	0.986	56.263	83.610	*0.990*	*47.946*	*69.677*
Hybrid PCR1	*0.982*	*73.701*	*96.050*	0.973	84.377	107.966	*0.980*	*76.882*	*99.750*
Hybrid PCR2	0.975	79.247	104.972	0.974	97.553	123.548	0.975	84.703	110.834
Hybrid PCR3	0.978	78.561	102.095	*0.980*	75.440	99.115	0.979	77.631	101.216
Hybrid PCR4	0.978	79.941	106.970	0.965	91.705	114.441	0.976	83.446	109.249
Hybrid PCR5	0.976	84.342	108.451	*0.980*	*72.645*	*97.949*	0.977	80.856	105.430
BPNN1	*0.982*	76.047	*94.952*	*0.955*	114.666	*139.095*	0.975	87.556	109.976
BPNN2	0.898	*51.678*	213.416	0.951	*102.212*	170.902	0.917	*66.738*	201.686
BPNN3	0.886	168.616	230.499	0.888	167.049	234.883	*0.891*	168.149	231.814
BPNN4	0.981	70.819	99.172	0.865	133.568	226.267	0.955	89.519	*148.867*
BPNN5	0.919	147.356	198.700	0.862	211.882	258.538	0.903	166.586	218.255

Italicized values indicate the best among the trial

R^2 coefficient of determination, *MAE* mean absolute error, *RMSE* root mean squared error

Table 8 Percentage error of the models in different TDS concentration ranges

	Minimum		Mean		Median		Maximum	
	TDS	Percentage of error	TDS	Percentage of error	TDS	Percentage of error	TDS	Percentage of error
Observed	460	–	1699.79	–	1800	–	3400	–
MLR1	447.51	2.71%	1700.31	*0.03%*	1791.20	*0.49%*	3379.25	0.61%
MLR2	481.77	4.73%	1705.93	0.36%	1781.52	1.03%	3639.60	7.05%
MLR3	503.63	9.49%	1706.21	0.38%	1812.90	0.72%	3566.47	4.90%
MLR4	471.12	*2.42%*	1703.33	0.21%	1828.89	1.61%	3461.70	1.81%
MLR5	447.54	2.71%	1704.20	0.26%	1837.53	2.08%	3381.12	*0.56%*
Hybrid PCR1	406.76	11.57%	1694.80	0.29%	1781.45	1.03%	3497.44	*2.87%*
Hybrid PCR2	438.00	*4.78%*	1714.15	0.84%	1807.05	0.39%	3647.04	7.27%
Hybrid PCR3	422.16	8.23%	1710.09	0.61%	1798.57	*0.08%*	3501.75	2.99%
Hybrid PCR4	409.53	10.97%	1697.94	*0.11%*	1790.80	0.51%	3584.00	5.41%
Hybrid PCR5	405.58	11.83%	1705.95	0.36%	1795.27	0.26%	3538.68	4.08%
BPNN1	557.98	21.30%	1715.45	0.92%	1776.09	1.33%	3364.15	*1.05%*
BPNN2	464.51	*0.98%*	1712.56	*0.75%*	1804.74	*0.26%*	4088.32	20.24%
BPNN3	201.81	56.13%	1547.48	8.96%	1687.29	6.26%	2965.57	12.78%
BPNN4	498.02	8.27%	1727.21	1.61%	1844.71	2.48%	4103.15	20.68%
BPNN5	673.05	46.32%	1733.31	1.97%	1790.70	0.52%	3306.01	2.76%

Italicized values indicate the best among the trials

proposed approach is found applicable to groundwater TDS modeling.

Conclusion

Groundwater parameters from 27 strategically located monitoring wells near an unlined landfill cell were studied. Different groundwater parameters are implemented to develop TDS models using regression techniques and machine learning approaches. Bias and precision of the TDS models are evaluated by various statistical indices.

Unlike other studies, dual-step MLR, hybrid PCR, and BPNN models are developed using distinct training and testing sets. The dual-step MLR estimates are close to the observed values, with good linear relationship ($R^2 = 0.988$). Eight out of 13 parameters are identified by the dual-step MLR as significant parameters on TDS. However, correlation coefficients among the variables revealed multicollinearity among the parameters. The proposed hybrid PCR is developed to minimize multicollinearity and retain information from more parameters. It is found that the proposed method performs well compared to other methods ($R^2 = 0.965$ to 0.980, MAE = 75.550 to 97.553, RMSE = 97.949 to 123.548, for testing stage in all models). BPNN, a machine learning method, did not perform as well as others. For instance, predictions of the target value were 56% higher in model BPNN5.

This study demonstrates the potential of integrating multivariate statistical approaches as a predictive modeling approach. Hybrid PCR in this study not only minimizes multicollinearity of the input parameters but also yields accurate and precise results. Results suggest that advanced regression techniques are appropriate for groundwater studies and can outperform machine learning approaches.

Funding information The research reported in this paper was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (RGPIN-385815). The authors are grateful for their support. The views expressed herein are those of the writers and not necessarily those of our research and funding partners.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Abou Zakhem B, Al-Charideh A, Kattaa B (2017) Using principal component analysis in the investigation of groundwater hydrochemistry of Upper Jezireh Basin, Syria. Hydrol Sci J 62(14):2266–2279. https://doi.org/10.1080/02626667.2017.1364845

Atta HSAF, Amer AWM, Atta SAF (2018) Hydro-chemical study of groundwater and its suitability for different purposes at Manfalut District, Assuit Governate. Water Science 32(1):1–15

Azadi S, Karimi-Jashni A (2016) Verifying the performance of artificial neural network and multiple linear regression in predicting the mean seasonal municipal solid waste generation rate: a case study of Fars province, Iran. Waste Manag 48:14–23. https://doi.org/10.1016/j.wasman.2015.09.034

Azadi S, Amiri H, Rakhshandehroo GR (2016) Evaluating the ability of artificial neural network and PCA-M5P models in predicting leachate COD load in landfills. Waste Manag 55:220–230. https://doi.org/10.1016/j.wasman.2016.05.025

Bagheri M, Bazvand A, Ehteshami M (2017) Application of artificial intelligence for the management of landfill leachate penetration into groundwater, and assessment of its environmental impacts. J Clean Prod 149:784–796. https://doi.org/10.1016/j.jclepro.2017.02.157

Bingham NH, Fry JM (2010) Regression: linear models in statistics. Springer Science & Business Media. https://doi.org/10.1007/978-1-84882-969-5

Bruce N, Ng KTW, Vu HL (2018) Use of seasonal parameters and their effects on FOD landfill gas modeling. Environ Monit Assess 190:291. https://doi.org/10.1007/s10661-018-6663-x

Cattell RB, Jaspers J (1967) A general plasmode (No. 30-10-5-2) for factor analytic exercises and research. Multivar Behav Res Monogr 67-3:211

Chen CS, Chen BPT, Chou FNF, Yang CC (2010) Development and application of a decision group back-propagation neural network for flood forecasting. J Hydrol 385(1–4):173–182. https://doi.org/10.1016/j.jhydrol.2010.02.019

Chickering GW, Krause MJ, Townsend TG (2018) Determination of as-discarded methane potential in residential and commercial municipal solid waste. Waste Manag 76:82–89. https://doi.org/10.1016/j.wasman.2018.03.017

City of Regina (2002) State of the environment report 2000, Regina Urban Environment Advisory, May 2002

City of Regina (2009) City of Regina landfill groundwater monitoring report. City of Regina environmental services. Regina, SK

City of Regina (2011) City of Regina landfill groundwater monitoring report. City of Regina environmental services. Regina, SK

City of Regina (2012) City of Regina landfill groundwater monitoring report. City of Regina environmental services. Regina, SK

City of Regina (2013) City of Regina landfill groundwater monitoring report. City of Regina environmental services. Regina, SK

City of Regina (2014) City of Regina landfill groundwater monitoring report. City of Regina environmental services. Regina, SK

City of Regina (2015) City of Regina landfill groundwater monitoring report. City of Regina environmental services. Regina, SK

City of Regina (2016) City of Regina landfill groundwater monitoring report. City of Regina environmental services. Regina, SK

Civelekoglu G, Yigit NO, Diamadopoulos E, Kitis M (2007) Prediction of bromate formation using multi-linear regression and artificial neural networks. Ozone Sci Eng 29(5):353–362. https://doi.org/10.1080/01919510701549327

Ebrahimi H, Rajaee T (2017) Simulation of groundwater level variations using wavelet combined with neural network, linear regression and support vector machine. Glob Planet Chang 148:181–191. https://doi.org/10.1016/j.gloplacha.2016.11.014

Han Z, Liu Y, Zhong M, Shi G, Li Q, Zeng D, Zhang Y, Fei Y, Xie Y (2018) Influencing factors of domestic waste characteristics in rural areas of developing countries. Waste Manag 72:45–54. https://doi.org/10.1016/j.wasman.2017.11.039

Hanley JA (2016) Simple and multiple linear regression: sample size considerations. J Clin Epidemiol 79:112–119. https://doi.org/10.1016/j.jclinepi.2016.05.014

Hassen I, Hamzaoui-Azaza F, Bouhlila R (2016) Application of multivariate statistical analysis and hydrochemical and isotopic investigations for evaluation of groundwater quality and its suitability for drinking and agriculture purposes: case of Oum Ali-Thelepte aquifer, Central Tunisia. Environ Monit Assess 188(3):135. https://doi.org/10.1007/S10661-016-5124-7

Health Canada (2017) Guidelines for Canadian drinking water quality—summary table. Water and air quality bureau, healthy environments and consumer safety branch, Health Canada, Ottawa, Ontario

Hu S, Luo T, Jing C (2013) Principal component analysis of fluoride geochemistry of groundwater in Shanxi and Inner Mongolia, China. J Geochem Explor 135:124–129. https://doi.org/10.1016/j.gexplo.2012.08.013

Kicsiny R (2016) Improved multiple linear regression based models for solar collectors. Renew Energy 91:224–232. https://doi.org/10.1016/j.renene.2016.01.056

Li Z, Wang G, Wang X, Wan L, Shi Z, Wanke H, Uugulu S, Uahengo C (2018) Groundwater quality and associated hydrogeochemical processes in Northwest Namibia. J Geochem Explor 186:202–214

Pan C, Ng KTW (2018) Multivariate analysis and Hydrochemical assessment of groundwater at the Regina landfill site. 33rd international conference on solid waste technology and management, Annapolis, Washington, MD, U.S.A.

Pan C, Ng KTW, Richter A (2017) Hydrochemical assessment of groundwater quality near Regina municipal landfill". Sardinia '17, 16th International Waste Management and Landfill Symposium, Santa Margherita di Pula, Cagliari

Pan C, Bolingbroke D, Ng KTW, Richter A, Vu HL (2018) "The Use of Waste Diversion Indices on the Analysis of Canadian Waste Management Models". Journal of Material Cycles and Waste Management. https://doi.org/10.1007/s10163-018-0809-3

Pomeroy JW, de Boer D, Martz LW (2005) Hydrology and water resources of Saskatchewan (p 25). Saskatoon: Centre for Hydrology Report #1, University of Saskatchewan

Rashid NA, Rosely NAM, Noor MAM, Shamsuddin A, Hamid MKA, Ibrahim KA (2017) Forecasting of refined palm oil quality using principal component regression. Energy Procedia 142:2977–2982. https://doi.org/10.1016/j.egypro.2017.12.364

Ravikumar P, Somashekar RK (2017) Principal component analysis and hydrochemical facies characterization to evaluate groundwater quality in Varahi river basin, Karnataka state, India. Appl Water Sci 7(2):745–755. https://doi.org/10.1007/s13201-015-0287-x

Sahoo S, Jha MK (2013) Groundwater-level prediction using multiple linear regression and artificial neural network techniques: a comparative assessment. Hydrogeol J 21(8):1865–1887. https://doi.org/10.1007/s10040-013-1029-5

Saskatchewan Ministry of Environment (2016) Municipal Drinking Water Quality Monitoring Guidelines. Edition 4. Environmental and Municipal Management Services Division, Water Security Agency. Regina, Saskatchewan.

Selvakumar S, Chandrasekar N, Kumar G (2017) Hydrogeochemical characteristics and groundwater contamination in the rapid urban development areas of Coimbatore, India. Water Resources and Industry 17:26–33. https://doi.org/10.1016/j.wri.2017.02.002

Sherrard JH, Moore DR, Dillaha TA (1987) Total dissolved solids: determination, sources, effects, and removal. J Environ Educ 18(2):19–24. https://doi.org/10.1080/00958964.1987.9943484

Solanki RB, Kulkarni HD, Singh S, Verma AK, Varde PV (2018) Optimization of regression model using principal component regression method in passive system reliability assessment. Prog Nucl Energy 103:126–134. https://doi.org/10.1016/j.pnucene.2017.11.012

Sousa SIV, Martins FG, Alvim-Ferraz MCM, Pereira MC (2007) Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. Environ Model Softw 22(1):97–103. https://doi.org/10.1016/j.envsoft.2005.12.002

Statistics Canada (2018) Census Profile, 2016 Census—Regina, City, Saskatchewan and Canada. Retrieved from https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/page.cfm?Lang=E&Geo1=CSD&Code1=4706027&Geo2=PR&Code2=01&Data=Count&SearchText=Regina&SearchType=Begins&SearchPR=

01&B1=All&GeoLevel=PR&GeoCode=4706027&TABID=1 on June 30, 2018

Tan KC, San Lim H, Jafri MZM (2016) Prediction of column ozone concentrations using multiple regression analysis and principal component analysis techniques: a case study in peninsular Malaysia. Atmos Pollut Res 7(3):533–546. https://doi.org/10.1016/j.apr.2016.01.002

Viswanath NC, Kumar PD, Ammad KK (2015) Statistical analysis of quality of water in various water shed for Kozhikode City, Kerala, India. Aquatic Procedia 4:1078–1085. https://doi.org/10.1016/j.aqpro.2015.02.136

Vu HL, Ng KTW, Richter A (2017) Optimization of first order decay gas generation model parameters for landfills located in cold semi-arid climates. Waste Manag 69:315–324. https://doi.org/10.1016/j.wasman.2017.08.028

Xu J, Wang L, Wang L, Shen X, Xu W (2011) QSPR study of Setschenow constants of organic compounds using MLR, ANN, and SVM analyses. J Comput Chem 32(15):3241–3252. https://doi.org/10.1002/jcc.21907

Xun Z, Hua Z, Liang Z, Ye S, Xia Y, Rui L, Li Z (2007) Some factors affecting TDS and pH values in groundwater of the Beihai coastal area in southern Guangxi, China. Environ Geol 53(2):317–323. https://doi.org/10.1007/s00254-007-0647-4

Zhao X, Wang S, Li T (2011) Review of evaluation criteria and main methods of wind power forecasting. Energy Procedia 12:761–769. https://doi.org/10.1016/j.egypro.2011.10.102

Zhao Y, Xia XH, Yang ZF, Wang F (2012) Assessment of water quality in Baiyangdian Lake using multivariate statistical techniques. Procedia Environ Sci 13:1213–1226. https://doi.org/10.1016/j.proenv.2012.01.115