Journal of Engineering Science and Technology
Vol. 13, No. 7 (2018) 1916 - 1925

PREDICTION OF CONCRETE MIX COMPRESSIVE STRENGTH USING STATISTICAL LEARNING MODELS

A. A. SHAQADAN[1,*], M. AL-RAWASHDEH[2]

[1]Civil Engineering Department, Zarqa University, Zarqa, 13132, Jordan
[2]Department of Civil Engineering, Faculty of Technical Engineering, Balqa Applied University, Salt, 19117, Jordan
*Corresponding Author: ashrafshaq10@gmail.com

Abstract

Laboratory analysis is expensive. In addition, recent developments in data mining techniques present a great opportunity for researchers to utilize available experimental data in a more productive way by building forecasting models. Today, there are several data-driven learning models that can predict outputs using input variables. Relevance vector machine is emerging statistical learning algorithm used in prediction of complex systems. Random Forest model is a novel algorithm used in a regression where it can assess the importance of input variables in a robust manner for a large number of variables. The two algorithms are developed and compared to concrete mix properties. Adding Silica to concrete mix produce high early strength material, which is a desirable feature. We analyse the results of 90 experimental samples that tested the effect of adding Silica and other physical properties on the compressive strength of concrete. Several silica/cement mixing ratios are evaluated and compressive strength was measured after 3 days and 28 days. Concrete strength increase is proportional with the increase of silica/cement ratio. In this study, the Random Forest model was developed to predict compressive strength using input variables and compared to Relevance Vector Machines model. The R^2 for Random Forest model is high at 0.97 but less than Relevance Vector Machines model at 0.989. However, the Random Forest model evaluated variables importance and indicated the curing time and milling time has a higher impact than increasing silicate percent.

Keywords: Compressive strength, Concrete, Milling time, Prediction model, Random forest, Rattle, RVM, Silica percent.

1. Introduction

Today, there is a high demand for concrete with enhanced properties to serve specific infrastructure needs. Demand for High-Performance Concrete (HPC) with high durability is on the rise. Virgin Silica (SiO_2) is an additive proven to enhance concrete durability, especially compressibility as shown in the experimental results in Fig. 1.

Virgin Silica (SiO_2) is a natural resource available in abundance in many regions in the world and in Jordan. Therefore, it is considered an attractive option for its environmental safety and economic efficiency [1, 2]. Experimental investigations of virgin silica are somewhat restrained by the high time and labour costs. Therefore, well-developed prediction models based on extensive experimental data are needed to help in forecasting properties of a new concrete mix.

Recently, statistical learning methods have seen significant development and growth in its application in various applications. Statistical learning models utilize data mining techniques, which make them versatile and valuable tools in various research fields [3]. Statistical learning methods use data or observations to build a structure of the relations among variables that can be used to predict system response in a given set of conditions [4].

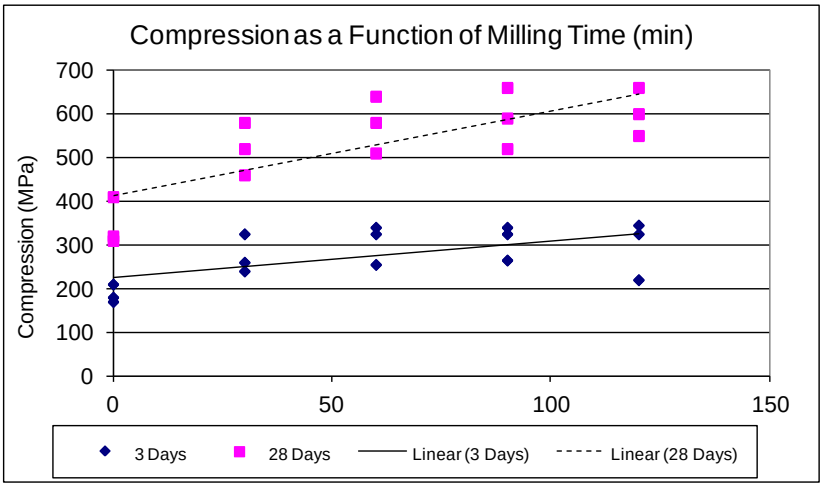

Fig. 1. Impact of silica % on compressive strength of concrete samples at two sampling periods.

1.1. Random Forest prediction model

Recently, Random Forest (RF) has received a lot of attention in various fields because they can handle large numbers of variables with the relatively small number of events [5], which makes it ideal for costly and time-consuming experiments. RF can be used for classification and regression [6]. Also, RF provides an assessment of variable importance [6, 7]. RF is a structure grown by a procedure called bagging, which is short for bootstrap aggregating, where trees are independently built by using a bootstrap sample (with replacement) of the entire dataset. Each node of the trees is split using a subset of the explanatory variables chosen randomly for each tree. Nodes are homogeneous groups of data.

Constructed trees are random, and overall prediction of the forest is the average of predictions from the individual trees. The parameters that must be tuned for growing a Random Forests are the number of attributes chosen at each split and the number of trees to be grown.

Once a tree has been built, the response for any observation can be forecasted by tracing the path from the root node down to the appropriate terminal node of the tree, based on the observed values for the splitting variables, and the predicted response value simply is the average response in that terminal node. For regression analysis, data are partitioned into more homogeneous groups called nodes, each split is based on data of splitting variables. After a tree is split, the response variable can be predicted by following the path from the root node.

Random Forest is grown from many regression trees forming an ensemble. This ensemble can be described as an r-dimensional random vector, *X =(X1, ..., Xr) T* where *X* represent real-valued predictor variable and *Y* represent real-valued response variable. The joint distribution of *XY* is assumed unknown. So, for RF we do not need to assume any distribution for variables. RF goal is to find a function *F(X1,...Xr*) = *Y*. The function F is a collection of " base learners" *h1(x), ..., hJ(x).*

1.2. Relevance vector machine prediction model

Relevance Vector Machine (RVM) is a sparse learning method for training linear models. RVM is used in a variety of forecasting applications. In this research, RVM is used to predict the compressive strength of concrete to compare with Random Forests. RVM view data as a chaotic system in which data are assumed to contain information about the response variable [8]. RVM simplifies complex systems by building "structured" models; therefore parameterization process that fits the information content. The key advantage of RVM is the generalization ability and the sparse formulation of the resulting model that utilizes few kernel functions. RVM fits yield full probability distributions of the output. However, it does not identify variables importance separately. RVM uses the following formula for the prediction of output (*y*).

$$y = a^{'} \emptyset(x) = \sum_{i=1}^{n} a_i K(x, x_i) + a_0 \quad (1)$$

where *x*, x_i are input variables, *K*(*x*, *xi*) is kernel function, *n* is a number of data and a_0 is weight.

The likelihood of the complete data set can be written as:

$$p(y|a, \sigma^2) = (2\pi\sigma^2)^{-n/2} exp\left\{-\frac{1}{2\sigma^2}\|y - a\varphi\|^2\right\} \quad (2)$$

$$p(y|a, \sigma^2) = (2\pi\sigma^2)^{-n/2} exp\left\{-\frac{1}{2\sigma^2}\|y - a\varphi\|^2\right\} \quad (3)$$

To prevent over-fitting, automatic relevance detection (ARD) prior is set over the weights as follows:

$$p(a/\propto) = \prod_{i=0}^{n} N(\propto_i/0, \propto_i^{-1}) = \prod_{i=0}^{n} \frac{\propto_i}{\sqrt{2\pi}} exp\left(-\frac{(\alpha_i \alpha_j)2}{2}\right) \quad (4)$$

where α is a hyperparameter vector that determines how far from zero each weight is allowed to deviate [9]. The posterior distribution over the weights is thus given by:

$$p(a/y,\alpha,\sigma^2) = \frac{p(y/a,\sigma^2)p(a/\alpha)}{p(y/\alpha,\sigma^2)} = (2\pi)^{-(n+1)/2}|\sum \quad|^{-1}/_2 exp\left[-\frac{1}{2}(a-\mu)'\right.$$
$$\left.\sum^{-1}(a-\mu)\right] \tag{5}$$

where the posterior covariance $\sum\left(\sigma^2\phi'\phi + A\right)^{-1}$ and the mean $\mu = \sigma^2\sum\phi' y$.

For uniform hyperpriors over α and σ^2, one needs only to maximize the term. $p(y/\alpha,\sigma^2)$.

$$p(y/\alpha,\sigma^2) = \int p(y/a,\sigma^2)p(a/\alpha)$$
$$= (2\pi)^{-n/2}|\sigma^2 I + \phi A^{-1}\phi'|^{-1/2} \times exp\left[\frac{-1}{2}y'(\sigma^2 I + \phi A^{-1}\phi')^{-1}y\right] \tag{6}$$

Maximization of this quantity is known as the type II maximum likelihood method [9, 10] or "evidence for hyperparameter" [11]. Hyperparameter estimation is carried out in iterative formulae, e.g., gradient descent on the objective function [4]. The outcome of this optimization is that many elements of this go to infinity such that we will have only a few nonzero weights that will be considered as relevant vectors. R programming language is a powerful tool for data mining. The R language and tools for statistical computing are gaining popularity among researchers [12]. *R* platform was utilized and the Rattle package [13, 14] was used.

2. Experimental Data

The data set contains experimental data about Cement Ratio (CR), Silica Ratio (SR), Milling Time (MT), Sample Age (SA), and Compressive Strength (CS). The dataset has compressive strength measurements after 3 days and 28 days. Three silica percentages are considered with five milling time periods and compressibility was measured at 3 and 28 days as explained in Table 1.

Table 1. Description of experimental data used in the study.

Virgin silica %	Age of sample	Milling time
0 **15** **30**	3 and 28 days	0,30, 60, 90, 120 minutes

3. Modelling Approach

In this study, inputs are cement percent, milling time, the age of the sample, Silicate percent, and compression strength. The Random Forests and RVM models are trained on 70% of the data, tested, and validated on remaining 30% of the dataset.

3.1. Random Forest

Setting-up RF requires choosing an *m* subset of predictor variables used to determine decision at a node of the tree. Also, a number of trees to be fitted is also selected iteratively. RF provides metrics for assessing its performance, the

assessment is conducted on Out-Of-Bag (OOB) cases that were left out of the bootstrapped training set. Also, residual error and pseudo R2 can be computed for the OOB cases. RF also provides variable importance score for each of the predictor variables. The data is imported to R, which was stored in CSV format. The target data was continuous data and it was partitioned to 70 % training, 15% validation, and 15% testing the partition chosen was 70/15/15.

Random Forest model requires setting *m* which is a number of variables used to determine the decision at a node of the tree. For this study, Cement and Silica percentages, milling time, and age of sample are used as regressors (*m*=4) and compression is used as a target. The number of trees is selected to be 1000. The RF model performance is compared with linear regression and neural networks models as a way to assess its prediction accuracy.

3.2. RVM

In RVM, a similar process was followed to import, train and test model. The developed RVM gives the best performance at $\sigma = 0.059$. The calibration process is iterative to maximize regression measures R^2 by changing σ.

4. Results and Discussion

4.1. Random Forest

Random Forest model fitting performance is evaluated using the OOB cases in the training set that have been left out of the bootstrapped training set. Figure 2 shows the error in the prediction of data outside of the training set is reduced with increasing number of trees after 600 trees error becomes minimal.

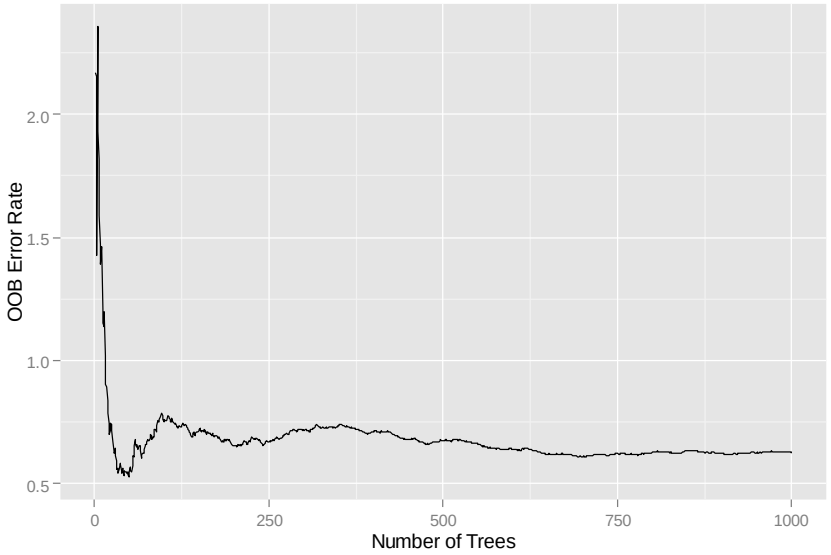

Fig. 2. The decrease in error as a function of number of trees.

RF model prediction can be evaluated by running the model on a new subset of data. In Fig. 3, the model was validated on a subset composed of 15% of the data with R^2 of 0.89 achieved.

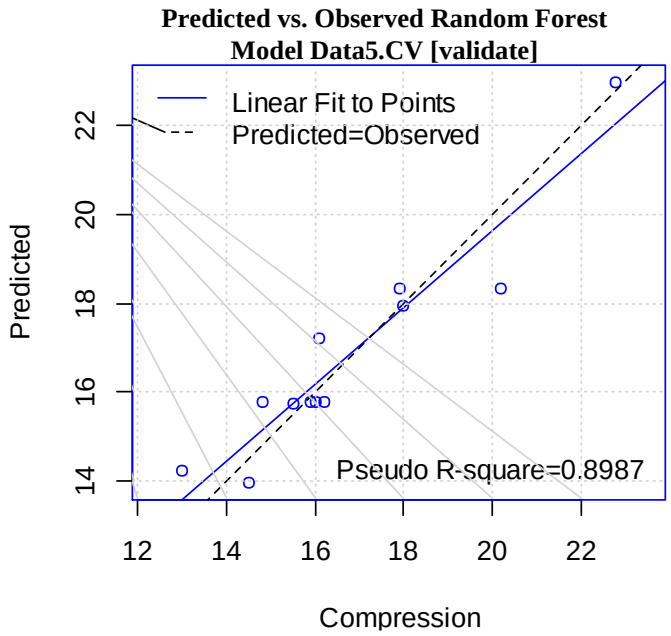

Fig. 3. RF model correlation coefficient for validation subset data.

Testing of the calibrated model on new 15% subset shows high R^2 of 0.97 as shown in Fig. 4.

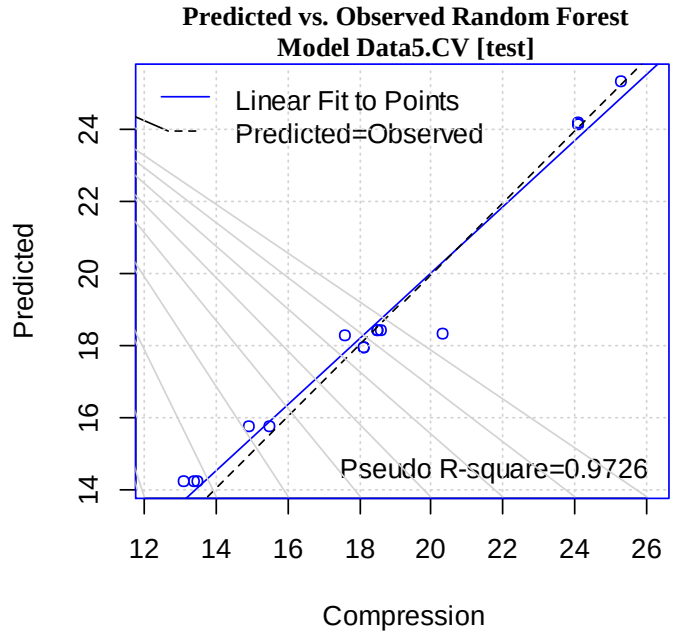

Fig. 4. RF model correlation coefficient for testing subset data.

Random Forests can evaluate variables importance by quantifying how the prediction error increases when data for a variable is permuted while keeping other variables constant. A randomly selected subset of explanatory variables is used for building single trees. Variable importance in RF can be evaluated by looking at how much the prediction error increases when the OOB data are permuted for a certain variable while keeping all others constant. Only a randomly selected subset of explanatory variables are used for the induction of the single trees, which means the relative importance of every variable can be determined and displayed, even if explanatory variables are correlated. The most important variables are the age of samples, followed by milling time, then cement percent, and Silica as shown in Fig. 5.

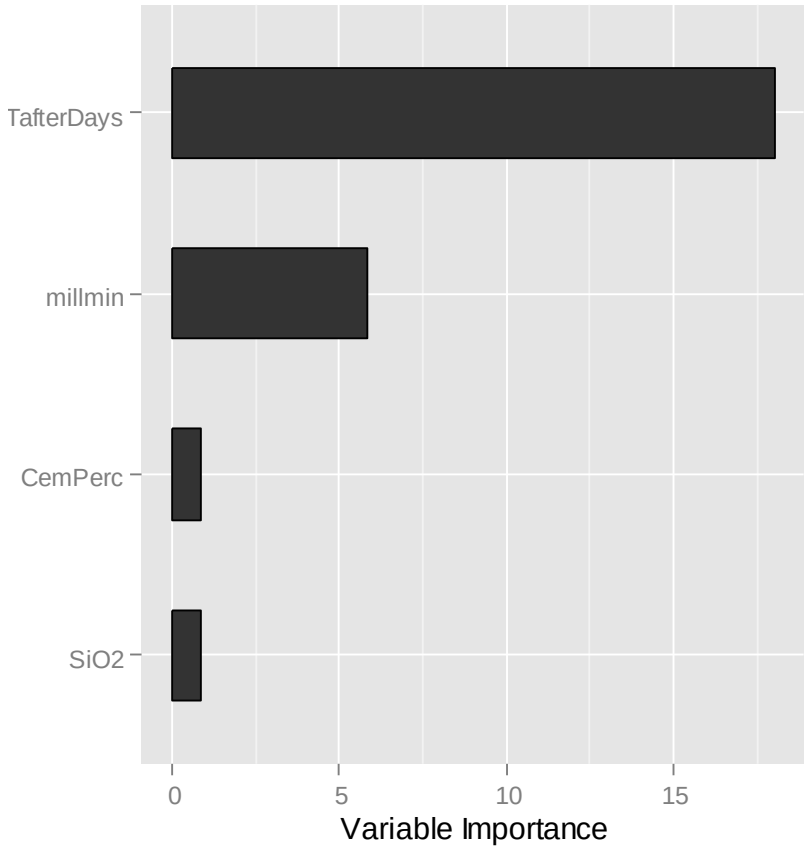

Fig. 5. Importance of the explanatory variables using Random Forest.

Variable importance is shown by the depth of trees where a variable with shallow depth explains more variability than variables with larger depth as shown in Fig. 6. So, sample age explains more variability than SiO_2.

For validation, the RF model is compared with linear regression and neural nets models. The RF model R^2 is comparable to neural nets and it is slightly higher than the linear regression model as shown in Table 2.

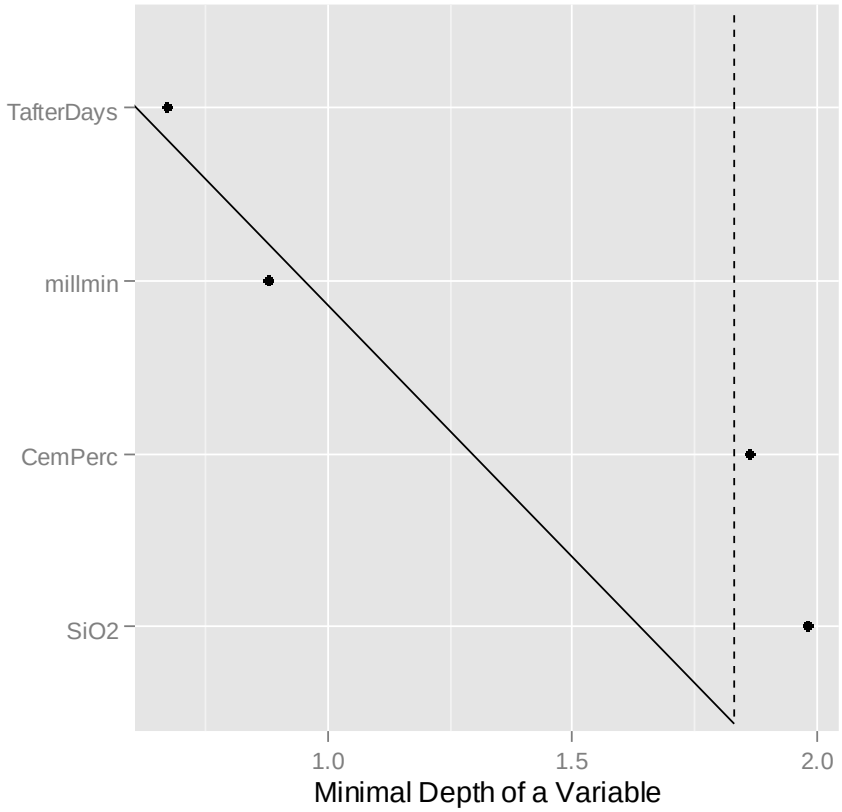

Fig. 6. A plot of depth for input variables as obtained by Random Forest model.

Table 2. Comparison of correlation coefficients of RF model with linear regression and neural nets models.

	Random Forest (RF)	Relevance vector machine
R^2	0.9726	0.99
Error		0.07

4.2. RVM

After training was conducted on 70% of the data, testing was conducted on 30% of the data. The trained model proved to be capable of predicting compressive strength over the entire experiment period with insignificant error as shown in Fig. 7. The model predictions are shown to match testing data very closely with correlation (R^2) of 0.99 and residual standard error of 0.07.

Therefore, the RVM model is capable of predicting compressive strength (CS) adequately with no unexplained trends in data. However, input variables importance cannot be assessed in a robust way.

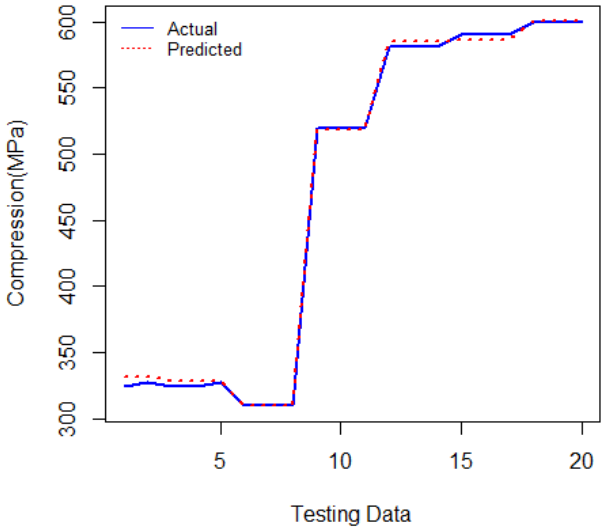

Fig. 7. Actual and predicted training data set RVM model.

5. Conclusion

Random Forest modelling is an intuitive and effective statistical approach for predicting response variables in regression problems.

- Random Forest model is capable of predicting compression using experimental inputs with high correlation (0.97) which is slightly inferior to RVM (0.99).
- A unique feature of the RF model compared to RVM is the advantage of assessing variables importance in a robust way.
- RF shows the high importance of curing time, which is the age of sample followed by milling time.
- Silicate percent impact is observed but it is overshadowed by sample age. Sample Age effect on compressibility is expected to be prominent. Silicate percent show an increasing trend incompressibility with higher silicate percent.

Nomenclatures

$F(X)$	Base learners function
$K(x.x_i)$	Kernel function
X	Predictor variable
Y	Response variable

Greek Symbols

α	Hyperparameter vector
Φ	Kernel function
μ	Mean function
a_i , a_0	Weight parameter

Abbreviations

ARD	Automatic Relevance Detection

CR	Cement Ratio
CS	Compressive Strength
HPC	High-Performance Concrete
MT	Milling Time
OOB	Out of Bag Data
RF	Random Forest Algorithm
RVM	Relevance Vector Machine Algorithm
SA	Sample Age
SiO_2	Virgin Silica Additive

References

1. Suliman, K.M.R.; and Awwad, M.T. (2005). Virgin silica improves the durability of Portland cement concrete. *Journal of Engineering Science*, 33(1), 1-9.
2. Suliman M.R.; and Awwad, M.T. (2000). Utilizing of silica in early-high strength concrete. Cement and concrete technology in the 2000s. *Proceedings of the Second International Symposium*. Istanbul, Turkey.
3. Vapnik V.N. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.
4. Tipping M.E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1, 211-244.
5. Denil, M.; Matheson, D.; and De Freitas, N. (2014). Narrowing the Gap: Random forests in theory and in practice. *Proceedings of the 31st International Conference on Machine Learning*. Beijing, China, 9 pages.
6. Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.
7. Ishwaran, H. (2007). Variable importance in binary regression trees and forests. *Electronic Journal of Statistic*, 1, 519-537.
8. Khalil, A.; McKee, M.; Kemblowski, M.; and Asefa, T. (2005). Sparse Bayesian learning machine for real-time management of reservoir releases. *Water Resources*, 41, W11401.
9. Berger, J.O. (1985). *Statistical decision theory and Bayesian analysis* (2nd ed.). New York: Springer.
10. Wahba, G. (1985). Comparison of GCV and GML for choosing the smoothing parameters in the generalized spline-smoothing problem. *The Annals of Statistics*, 13(4), 1378-1402.
11. MacKay, D.J. (1992). *Bayesian methods for adaptive models*. Ph.D. Thesis, Computation and Neural Systems, California Institute of Technology, Pasadena, California.
12. R Development Core Team (2008). R: A language and environment for statistical computing. R foundation for statistical computing, Vienna. Retrieved, June 22, 2017 from http://www.R-project.org.
13. Williams, G. (2011). *Data mining with rattle and R: the art of excavating data for knowledge discovery*. New York: Springer.
14. Liaw, A.; and Wiener, M. (2002). Classification and regression by Random Forest. *R News*, 2(3), 18-22.