Journal of Hydroinformatics

Journal of Hydroinformatics Vol 00 No 0, 1 doi: 10.2166/hydro.2022.003

Evaluation of effective parameters of Manning roughness coefficients in HDPE culverts via kernel-based approaches

Ghazaleh Nassaji Matin

Department of Water Resource Engineering, Faculty of Civil Engineering, University of Tabriz, Tabriz, Iran
Corresponding author. E-mail: ghazaleh.matin@yahoo.com

GNM, 0000-0002-8336-0831

ABSTRACT

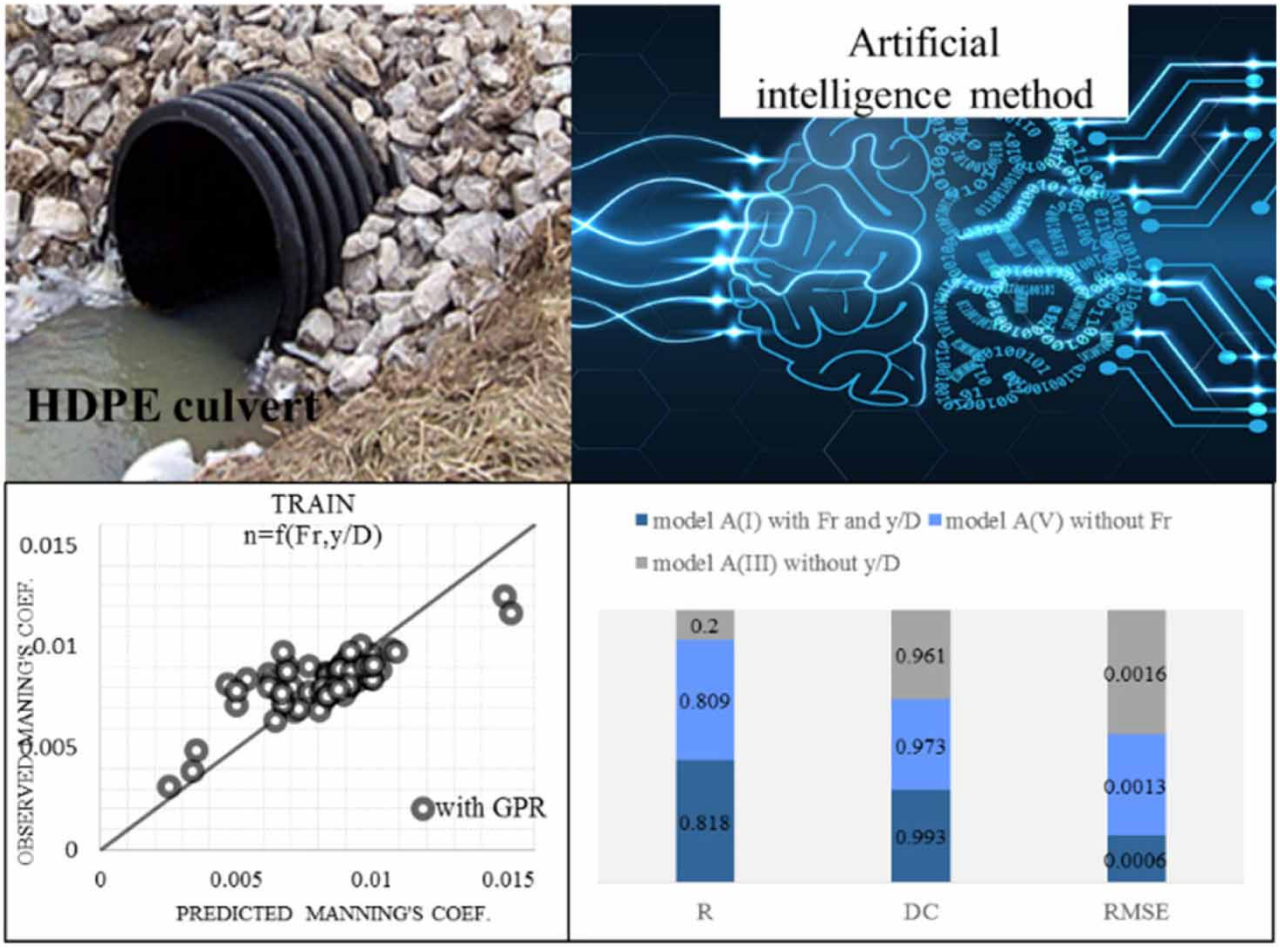
The prediction of Manning coefficients plays a prominent role in the estimation of head losses along culvert systems. Although the Manning coefficient is treated as a constant, previous studies showed the dependency of this coefficient on several parameters. This study aims to evaluate the effective parameters of the Manning roughness coefficient using intelligence approaches such as Gaussian process regression (GPR) and support vector machines (SVM), in which the input variables were considered as dimensionless and dimensional. In addition to the enhanced efficiency of the SVM approach compared to the GPR approach in model development with dimensionless input variables, the accuracy of model A(I) with input parameters of Fr (Froude) and y/D (the ratio of water depth to culvert diameter) and performance criteria of correlation coefficient (R) = 0.738, determination coefficient (DC) = 0.0962, root mean square errors (RMSE) = 0.0015 and R = 0.818, DC = 0.993 and RMSE = 0.0006 for GPR and SVM approaches were the highest. Thus, for the second category, a model with an input parameter of discharge (Q), hydraulic radius (R_H), and culvert's slope (S_0) showed good efficiency in predicting the Manning coefficient, in which the performance criteria of GPR and SVM approaches were (R = 0.719, DC = 0.949, RMSE = 0.0013) and (R = 0.742, DC = 0.991, RMSE = 0.007), respectively. Furthermore, developed OAT (one-at-a-time) sensitivity analysis revealed that relative depth y/D and Q are the most important parameters in the prediction of the Manning coefficient for models with dimensionless and dimensional input variables, respectively.

Key words: culvert systems, Gaussian process regression, intelligence approaches, Manning roughness coefficient, support vector machine

HIGHLIGHTS

- Effective parameters in predicting the Manning coefficient was evaluated, and the efficiency of GPR and SVM was evaluated in predicting the Manning roughness coefficient of culverts.
- Although the Manning roughness coefficient is treated as a constant, it was observed that the Manning coefficient depends on several parameters.
- Results of developed models revealed the uncertainty of friction loss in culvert systems.

GRAPHICAL ABSTRACT



INTRODUCTION

Among various channel hydraulic parameters, the channel's Manning roughness (n) plays a crucial role in the study of channel flow, particularly in the hydraulic modeling of culvert systems. Manning's roughness of culvert systems greatly influences the velocity and depth of water inside the culvert (Lang *et al.* 2004) and is known to be a key parameter for a realistic simulation of flows but remains especially complicated to determine (Niazkar *et al.* 2018). Hence, the reliable estimation of discharge capacity is essential for the design, and the hydraulic criteria influencing discharge capacity in culverts are flow rates, material roughness, diameter, and slope of the culvert. Therefore, the determination of (n) of culvert systems has become a challenge in practice.

Manning's equation has been widely used to calculate the flow rate, and researchers have simulated floods using the calibrated Manning's n slope area method. Boyer (1954) stated that n in open channels plays an important role in the determination of discharge value (Pradhan & Khatua 2018b). Li & Zhang (2001) refer to n as one of the most important parameters for analyzing water flow over the ground and provide a technique for the calculation of field n. Hosseini *et al.* (2000) discussed the key role of n and Manning's equation in evaluating the accuracy of energy and momentum principles for the analysis of one-dimensional open channel flow. According to the literature, estimating resistance coefficients may be conducted using different approaches (Kitsikoudis *et al.* 2015). Ardıçlıoğlu & Kuriqi (2019) presented an eight-step scheme that was developed to predict n when grain and form roughness are the major sources of friction. Boulomytis *et al.* (2017) estimated n of the main channel and floodplain of the Juqueriquere River basin using the Cowan method based on field observations. The prediction of the friction factor in pipes using the model tree was performed by Najafzadeh (2017). Bardestani

et al. (2016) predicted turbulent flow friction coefficients using an adaptive neural fuzzy inference system (ANFIS) technique, and the friction factor in pipes was estimated by using the ANFIS and a grid partition method. It was found that the ANFIS model was more accurate than other empirical equations in modeling friction factors. Furthermore, the soft computing technique was used as a robust method to predict n in grassed channels, high gradient streams, and other environmental problems (Roushangar *et al.* 2017).

Water distribution systems are considered an important public infrastructure (Pandey et al. 2020), and the design of the systems includes a large number of parameters (Bhave & Gupta 2006). Previous studies have shown that n, as one of the key parameters, depends on various factors including the change of the stream conditions, water depth, rainfall intensity, discharge, tailwater level, pipe material, pipe diameter, corrugations, the slope of energy grade line, and hydraulic radius (Bloodgood & Bell 1961; American Concrete Pipe Association 2007; Devkota 2012; Ardıçlıoğlu & Kuriqi 2019).

Numerous models have been developed for the prediction of the n using experimental and numerical studies. In the past decades, artificial intelligence approaches such as artificial neural networks (ANNs), neuro-fuzzy models (NF), genetic programming (GP), gene expression programming (GEP), support vector machine (SVM), and Gaussian process regression (GPR) have become popular in water resources engineering, leading to numerous publications in this field (Amaranto *et al.* 2018; Carvalho *et al.* 2018; Owen & Liuzzo 2019; Roushangar *et al.* 2019; Tinelli & Juran 2019; Zhu *et al.* 2019; Roushangar & Shahnazi 2020). Therefore, utilizing artificial intelligence was considered as a tool for predicting the roughness coefficient (Mangin 2010; Saghebian *et al.* 2020). However, in addition to artificial intelligence methods, other numerical and experimental methods have been proposed by many researchers (Mohammadpour *et al.* 2019; Lavoie & Mahdi 2020) and, from another point of view, some researchers presented an innovative method for calibrating Manning roughness (Boulomytis *et al.* 2017; Attari & Hosseini 2019). Since the defections of the water flow in culvert systems could cause problems such as the overflow of water on the road, the prediction of n plays a prominent role in the determination of losses along water distribution and drainage infrastructures. The optimum design of culvert depends on many factors including but not limited to the n which is the concentration point in this paper. Since many previous studies have pointed out the dependency of Manning coefficient on geometric and hydraulic parameters, it seems that having sufficient knowledge of parameters affecting the n could be useful in the accurate calculation of the upcoming optimum design of culverts. Despite the fact that the n is treated as a constant, many studies and available methods indicate that the n is a dependent parameter of hydraulic and geometric parameters (McKay & Fischenich 2011; Ferguson 2013). Due to the uncertainties and considering the reviewed literature, a lack of comprehensive studies on the prediction of n in culvert systems using artificial intelligence was observed. Therefore, in order to survey the subject comprehensively, utilizing artificial intelligence was considered as a tool for predicting the roughness coefficient (Mangin 2010; Saghebian *et al.* 2020). In the present study, the effective parameters in predicting the n were evaluated, and the efficiency of kernel-based approaches such as GPR and SVM was assessed. Hence, the OAT (one-at-a-time) sensitivity analysis was implemented to recognize the most effective input variables in predicting Manning roughness coefficient.

MATERIALS AND METHODS

Kernel-based approaches

Kernel-based approaches (such as GPR and SVM) are one of the common methods for solving nonlinear problems which are based on a statistical learning theory. They are also fairly robust against overfitting, especially in a high-dimensional space (Roushangar *et al.* 2019). The availability of sufficient input data will enable these models to predict any variable. However, the model covers only the relationships found within the given dataset (Bobivic 2009).

Gaussian process regression (GPR)

GPR is a newly developed learning approach that works based on the concept of kernel functions. GPR presents probabilistic models, which means that the Gaussian process provides the reliability of responses to the given input data. In addition, the GPR method is flexible as it can handle nonlinear problems and also non-parametric as it does not need parameter selection (Roushangar & Shahnazi 2019). Such models are capable of adapting themselves to predict any variable of interest using sufficient inputs. The training of these methods is fast and has high accuracy. GPRs can model nonlinear decision boundaries, and there are many kernels to choose from. They are also fairly robust against overfitting, especially in a high-dimensional space. However, the appropriate selection of kernel type is the most important step in the GPR due to its direct impact on the training and classification precision (Saghebian *et al.* 2020).

GPR models are based on the assumption that adjacent observations should convey information about one other. Due to prior knowledge regarding data and functional dependencies, no validation process is required for generalization, and GP regression models are able to understand the predictive distribution corresponding to the test input (Rasmussen & Williams 2006). Considered input space $x = R^n$ of n-dimensional vectors to an output space $\gamma = R$ of real-valued targets, in which n pair (x_i, y_i) is drawn independently and identically distributed. For regression, assuming that $y \subset R$, then, a GP on γ is defined by a mean function $\mu{:}x \to R$ and a covariance function $k{:}X \times X \to R$.

In GP regression, the main assumption is that the y value can be calculated from $y = f(x) + \xi$, where $\xi \sim N\ (0,\ \sigma^2)$. In GP regression, for every input x, there is an associated random variable $f(x)$, which is the value of the stochastic function f at that location. In this study, it is assumed that observational error ξ is normal dependent and identically distributed, with a mean value of zero ($\mu(x) = 0$), a variance of σ^2 and $f(x)$ drawn from the Gaussian process on x specified by k. That is, $Y = (y_1,\ldots,y_n) \sim (0,\ K+\ \sigma^2 I)$, where $K_{ij} = k(x_i,\ x_j)$, and I is the identity matrix. Because $Y/X \sim N(0,K+\ \sigma^2 I)$ is normal, so is the conditional distribution of test labels given the training and test data of $p(Y^*0Y,\ X,\ X^*)$. Then, one has $Y^*/Y,\ X,\ X^*N$ $(\mu,\ \Sigma)$ where

$$\mu = K(X^*, X^*)(K(X, X) + \sigma^2 I)^{-1} Y \tag{2}$$

$$\Sigma = K(X^*, X^*) - \sigma^2 I - K(X^*,\ X)(K(X, X) + \sigma^2 I)^{-1} K(X, X^*) \tag{3}$$

If N training data and N^* test data were available, then we have $K(X^*, X^*)$; here, X and Y are the vector of training data and training data labels y_i, whereas X^* is the vector of test data. A specified covariance function is required to generate a positive semi-definite covariance matrix K, where $K_{ij} = k(x_i,x_j)$. The term of kernel function used in the SVM is equivalent to the covariance function used in GP regression. With the known values of kernel k and the degree of noise σ^2, Equations (2) and (3) would be enough for inference. During the training process of GP regression models, one needs to choose a suitable covariance function as well as its parameters. In the case of GP regression with a fixed value of Gaussian noise, a GP model can be trained by applying Bayesian inference, i.e., maximizing the marginal likelihood. This leads to the minimization of the negative log-posterior:

$$P(\sigma^2, k) = \frac{1}{2} y^T (K + \sigma^2 I)^{-1} y + \frac{1}{2} \log|K + \sigma^2 I| - \log p(\sigma^2) - \log p(k) \tag{4}$$

To assess the hyperparameters, the partial derivation of Equation (3) can be obtained with respect to σ^2 and k. For a detailed discussion on GP regression, Kuss (2006) is suggested. The optimal value of capacity constant (c) and the size of error-intensive zone (ε) in GPR are required due to their high impact on the accuracy of the mentioned regression approaches. To achieve the optimum values of these parameters, a trial-and-error process was executed.

Support vector machine (SVM)

SVM as an intelligence approach is implemented in information categorization and dataset classification. This approach, which was developed by Vapnik (1995), is known as structural risk minimization (SRM), which minimizes an upper bound on the expected risk, as opposed to the traditional empirical risk (ERM), which minimizes the error on the training data. The primary concept of SVM is the optimal hyperplane that separates samples of two classes by considering the widest gap between these two classes. Many researches have been carried out in various fields of engineering by using SVM. Therefore, only a summary of the employed SVM model is presented here. It is assumed that for dataset $(x_i,\ y_i)$, SVM equations founded on Vapnik theory approximate the function as:

$$f(x) = w\varphi(x) + b \tag{5}$$

where $\varphi(x)$ represents a nonlinear function in feature of input x, w-vector known as the weight factor, b is known as bias. These coefficients are predicted by minimizing regularized risk function as shown below:

$$R_{\mathrm{SVM}}(c) = \frac{1}{2}||w||^2 + c\frac{1}{2}\sum_{i=1}^{n} L_\varepsilon(t_i, y_i) \tag{6}$$

where

$$L_\varepsilon(t_i, y_i) = \begin{cases} 0 & |t_i - y_i| \le \varepsilon \\ |t_i - y_i| - \varepsilon & \text{otherwise} \end{cases} \tag{7}$$

The constant c is the cost factor, $\frac{1}{2}||w||^2$ stands for the regularization term, ε is the radius of the tube within which the regression function must lie, n is the number of elements and $L_\varepsilon(t_i, y_i)$ denotes the loss function, in which y_i is the forecasted value and t_i stands for the desired value in the period i . The parameters w and b are estimated by the minimization process of the regularized risk function after introducing positive slack variables ξ_i and ξ_i^* that express upper and lower excess deviation.

$$\text{Minimize} \quad R_{\text{SVM}}(w,\ \xi^*, \xi) = \frac{1}{2}||w||^2 + c\sum_{i=1}^{n}(\xi_i, \xi_i^*)$$

$$t_i - W_{i\varphi}(X_i) - b \le \varepsilon + \xi_i, \quad W_{i\varphi} + b - t_i \le \varepsilon + \xi_i^*, \quad \xi_i + \xi_i^* \ge 0 \tag{8}$$

Equation (5) can be solved by introducing Lagrange multipliers and optimality constraints, therefore obtaining a general form of function given by:

$$f(x) = \sum_{i=1}^{n}(\beta_i - \beta_i^*)K(x_i,\ x_j) + b \tag{9}$$

where β_i and β_i^* are Lagrange multipliers, $K(x_i,\ x_j)$ is $\varphi(x_i)\varphi(x_j)$, and the term $K(x_i, x_j)$ refers to the kernel function, which is an inner product of two vectors x_i and x_j in the feature space $\varphi(x_i)$ and $\varphi(x_j)$, respectively.

Among the existed approaches, SVM is one of the best known techniques to optimize the expected solution. The extraordinary generalization capability of SVM, along with its optimal solution and its discriminative power, has attracted the attention of data mining, pattern recognition, and machine learning communities in the last years. SVM has been used as a powerful tool for solving practical binary classification problems. It has been shown that SVMs are superior to other supervised learning methods (Bhowmik *et al.* 2009; Cervantes *et al.* 2019). Due to its good theoretical foundations and good generalization capacity, in recent years, SVMs have become one of the most used classification methods. The advantage of SVM is to neatly solve the inner product operation in the high-dimensional space by introducing kernel function, so the prediction accuracy would be enhanced with appropriate kernel function (Roushangar & Shahnazi 2020). Other advantages includes unique solution due to the convex nature of the optimal problem, the use of high-dimensional spaced set of kernel functions which discreetly comprise nonlinear transformation, and no assumption in functional transformation which makes data linearly separable indispensable (Kisi *et al.* 2015).

The extraordinary generalization capability of SVM, along with its optimal solution and its discriminative power, has attracted the attention of data mining, pattern recognition, and machine learning communities in the past years.

Performance criteria

In the current study, the model's performance was evaluated using three statistics: correlation coefficient (R), determination coefficient (DC), and root mean square errors (RSME). The expressions for performance criteria are presented in Equation (10).

$$R = \frac{\sum_{i=1}^{N}(I_0 - \overline{I_0}) \times (I_P - \overline{I_P})}{\sqrt{\sum_{i=1}^{N}(I_0 - \overline{I_0})^2 \times (I_P - \overline{I_P})^2}}, \quad \text{DC} = 1 - \frac{\sum_{i=1}^{N}(I_0 - I_P)^2}{\sum_{i=1}^{N}(I_0 - \overline{I_P})^2}, \quad \text{RMSE} = \sqrt{\sum_{i=1}^{N}\frac{(I_0 - I_P)^2}{N}} \tag{10}$$

where I_0,I_P, $\overline{I_0}$, $\overline{I_P}$, and N, respectively, are the measured values, predicted values, mean measured values, mean predicted values, and the number of data samples.

Model and simulation and development

Data characterization

To evaluate the effective parameters of the n, the laboratory data of Devkota (2012) were utilized. The laboratory testing was performed in the Fluid Mechanics Laboratory of the Civil and Environmental Engineering Department at the Youngstown State University. The HDPE culverts used in this research were externally corrugated but smooth inside with diameters of 1, 2, and 3.5 ft. The profile and cross-section of the 2 ft.

In the study, Devkota (2012) varied discharge, slope and culvert diameter, and measured water depth. The slope of the energy grade line could be computed from the measured water depth at four different flume slopes. In order to determine the discharge, a pump rating curve was created. The data collection, therefore, consisted of measurements of discharge for the creation of the discharge rating curve and water depth in the culvert under various diameter, slope, and discharge conditions in HDPE culverts. The water from the sump area is discharged into the headbox of the flume, and an outlet was provided at the downstream end of the flume which returned the water to the sump, thus recirculating the water. Upstream from the pump, a control valve has been installed to limit the discharge of water into the headbox. As the number of opened turns of the valve is increased, the discharge increases. The data collection process consisted of two steps:

1. Developing a discharge rating curve (discharge measurement)
2. Collecting water depths within the culverts (depth measurement)

Input variables

The accuracy of the predictions mainly depends on the appropriate allocation of input parameters. Based on Saghebian *et al.* (2020) and Devkota (2012), the most important parameters in predicting the n of culvert include the culvert diameter (D), culvert's bed slope (S_0), discharge (Q), water depth in the culvert (y), hydraulic radius (R_H), Froude number (Fr), and Reynolds number (Re). Accordingly, the input parameters consist of culvert diameter (D), bed slope (S_0), discharge (Q), water depth in culvert (y), hydraulic radius (R_H), and Reynolds number (Re). Also, the data range of various parameters used in this study is presented in Table 1.

$\mathrm{Fr} = V/(g \times H_w)^{0.5}$ is the Froude number a dimensionless value that describes flow regimes, V is the flow velocity in the culvert, H_w is flow depth in the culvert, $\mathrm{Re} = vL/v$ is the Reynolds number which is a dimensionless number used to categorize the effect of viscosity in controlling velocities or the flow pattern of a fluid, L is fluid length, v is the static viscosity, and y/D is the ratio of water depth in the culvert to culvert diameter. Two categories were considered for model development: (1) models with dimensionless input variables and (2) models with dimensional input variables. Therefore, in order to develop the appropriate models, input variables of the first category with five models were considered to be dimensionless, whereas the input variables of the second category with seven models had dimensions (Table 2). Providing appropriate dataset is a critical step in the prediction of n via artificial intelligence methods. The examination of models showed that considering 75% of the dataset for training goals and the remaining 25% for testing goals lead to more accurate results. Because of more accurate estimation and more efficient use of data and to avoid model bias, the training and testing datasets were divided using v-fold cross-validation which was developed in STATISTICA software (v. 8). The separated dataset is then utilized in further model developments. A v-fold cross-validation is a standard approach for model selection, and the main idea of cross-validation is data splitting (Arlot & Lerasle 2016). There are a total of 156 data, of which 75% (116 number of data) are dedicated to training data and 25% (39 number of data) are related to the testing data.

Table 1 | Details of various parameters from laboratory experiments used in this study

Parameters	Data range	No. of data
Slope (S_0)	0.0012–0.0262	112
Discharge (Q)	0.217–10.281	
Hydraulic radius (R_H)	0.057–1.856	
Reynolds number (Re)	15,602.53–370,560.1	
Froude number (Fr)	0.705–23.9	
Relative depth (y/D)	0.044–0.751	

Table 2 | Developed models

Inlet loss in slope-tapered circular culvert

Models using dimensionless parameters assigned to GPR and SVM approaches		Models using parameters with dimensions assigned to GPR and SVM approaches	
Model	**Input variables**	**Model**	**Input variables**
A(I)	Fr, y/D	B(I)	Q, S_0, R_H
A(II)	Re, y/D	B(II)	Q, S_0
A(III)	Fr	B(III)	Q, R_H
A(IV)	Re	B(IV)	S_0, R_H
A(V)	y/D	B(V)	Q
		B(VI)	S_0
		B(VII)	R_H

It is worst to note that the understanding of the correlations between data could help the model to achieve higher accuracy. At the same time, training time will be reduced because of the reduced dimensions of data. It is correct that correlation is often used for feature selection. But, there is no strong reason either to keep or remove features that have a low correlation with the target response, other than reducing the number of features if necessary. However, to the author's knowledge, if the number of features is not a problem, the correlation seems to be not important for modeling.

Model justification

To show the benefits or values of GPR and SVM methods, the results of these models were compared to the GEP method which can serve as a benchmark for comparison purposes. The better performance of considered methods justified the use of GPR and SVM methods.

GEP was developed by Ferreria (2001) using fundamental principles of genetic algorithms (GA) and genetic programming (GP). GAs are the heuristic search and optimization techniques that mimic the process of natural evolution. Thus, GAs implement the optimization strategies by simulating the evolution of species through natural selection (Bhattacharjya 2012). The problems are encoded in linear chromosomes of fixed-length as a computer program (Ferreria 2001), which are then expressed or translated into expression trees (ETs). GEP algorithm begins by selecting the five elements such as the function set, terminal set, fitness function, control parameters, and stop condition. There is a comparison between predicted values and actual values in a subsequent step. When desired results in accord with the error criteria initially selected are found, the GEP process is terminated. If desired error criteria could not be found, some chromosomes are chosen by a method called roulette wheel sampling and they are mutated to obtain new chromosomes. After the desired fitness score is found, this process terminates and then the chromosomes are decoded for the best solution to the problem (Teodorescu & Sherwood 2008). In the present study, GEP has been trained for model A(I). Four basic arithmetic operators ($+$, $-$, $\times$, $/$) and some basic mathematical functions ($\sqrt{}$, X2, X3, X1/3) were utilized as a GEP function set. GEP models were evolved till the fitness function remains unchanged for 10,000 runs for each pre-defined number of a gene, then the program was stopped. The model parameters and the size of the developed GEP models were then tuned (optimized) throughout refining (optimizing) the trained and fixed model as a starter.

The statistical parameters of (R, DC, and RMSE) of model A(I) were used to compare the results of GEP with GPR and SVM methods. The method which led to the highest R and DC and the lowest RMSE was selected as the best method. According to the performance criteria, in the GEP model, using the methods of SVM and GEP led to more accurate predictions. The performance criteria of mentioned approaches are presented in Table 3.

SVM and GPR model development

The design of GPR- and SVM-based regression approaches involves the use of kernel functions. The appropriate selection of kernel type leads to accurate results. In general, there are several types of kernel functions, including linear, polynomial, radial basis function (RBF), and sigmoid functions. The kernel-based models are capable of adapting themselves to predict

Table 3 | Comparison of SVM, GPR, and GEP models

		Evaluation criteria					
		Training			Testing		
Models		*R*	DC	RMSE	*R*	DC	RMSE
A(I)	GPR	0.786	0.989	0.0012	0.738	0.962	0.0015
	SVM	0.934	0.99	0.0007	0.818	0.993	0.0006
	GEP	0.645	0.63	0.0016	0.603	0.582	0.0017

any variable of interest in the presence of sufficient data. The training of these methods is fast and accurate, with a low probability of data overtraining.

In order to select the best kernel function, model A(I) was predicted via various kernels. The statistical parameters of (*R*, DC, and RMSE) were used to find optimum kernel functions. The kernel function, which led to the highest *R* and DC and the lowest RMSE, was selected as optimum kernel function. According to the performance criteria of *R*, DC, and RMSE, in GPR model development, using the kernel function of squared-exponential led to more accurate predictions. Whereas in SVM model development, the RBF kernel function led to more accurate results. Figure 1 shows the results of the statistical parameters for different kernels for model A(I). Kernel functions of presented kernel types for SVM and GPR approaches are indicated in Table 4 (Rasmussen & Williams 2006; Neal 2012; Roushangar & Shahnazi 2019).

RESULTS AND DISCUSSION

In order to investigate the influence of various variables on *n* of culverts, several models were developed based on flow characteristics (Froude number, Reynolds number, and discharge) and geometric parameters (relative depth, culvert slope, and hydraulic radius). In order to achieve a prediction of *n*, all GPR and SVM models were trained and tested. Results are presented in Tables 2 and 3 and Figures 2–5.

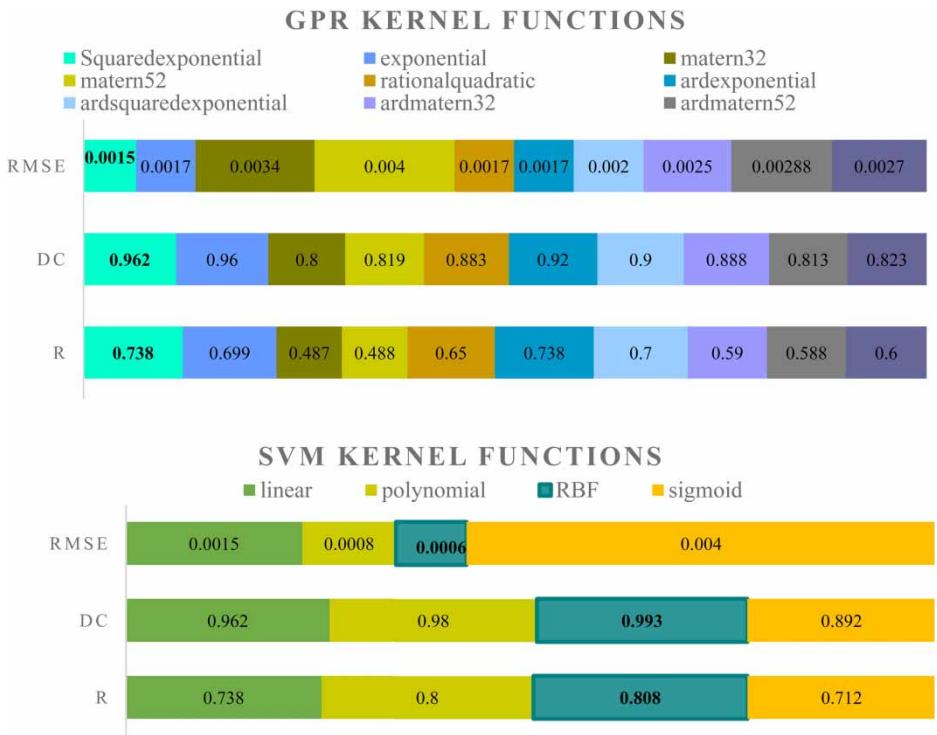


Figure 1 | Statistical parameters for GPR and SVM kernel function types for a testing set of model A(I).

Table 4 | Kernel functions for SVM and GPR approaches

SVM approach		GPR approach	
Kernel type	**Function**	**Kernel type**	**Function**
Linear	$k(x_i,\ x_j) = (x_i, x_j)$	Squared-exponential	$k(x_i, x_j\|\theta) = \sigma_f^2 \exp\left[-\frac{1}{2}\frac{(x_i - x_j)^T(x_i - x_j)}{\sigma_l^2}\right]$
Polynomial	$k(x_i,\ x_j) = ((x_i, x_j) + 1)^d$	Exponential	$k(x_i, x_j\|\theta) = \sigma_f^2 \exp\left(-\frac{r}{\sigma_l}\right)$
RBF	$k(x_i,\ x_j) = \exp(-\|\|x_i, x_j\|\|)^2/2\gamma^2$	Matern 3/2	$k(x_i, x_j\|\theta) = \sigma_f^2\left(1 + \frac{\sqrt{3}r}{\sigma_l}\right)\exp\left(-\frac{\sqrt{3}r}{\sigma_l}\right)$
Sigmoid	$k(x_i,\ x_j) = \tanh(-\alpha(x_i, x_j) + c)$	Matern 5/2	$k(x_i, x_j\|\theta) = \sigma_f^2\left(1 + \frac{\sqrt{5}r}{\sigma_l} + \frac{5r^2}{3\sigma_l^2}\right)\exp\left(-\frac{\sqrt{5}r}{\sigma_l}\right)$
	–	Rational quadratic	$k(x_i, x_j\|\theta) = \sigma_f^2\left(1 + \frac{r^2}{2\alpha\sigma_l^2}\right)^{-\alpha}$
	–	ARD squared-exponential	$k(x_i, x_j\|\theta) = \sigma_f^2 \exp\left[-\frac{1}{2}\sum_{m=1}^{d}\frac{(x_{im} - x_{jm})^2}{\sigma_m^2}\right]$
	–	ARD exponential	$k(x_i, x_j\|\theta) = \sigma_f^2 \exp(-R)$
	–	ARD Matern 3/2	$k(x_i, x_j\|\theta) = \sigma_f^2(1 + \sqrt{3}R)\exp(-\sqrt{3}R)$
	–	ARD Matern 5/2	$k(x_i, x_j\|\theta) = \sigma_f^2\left(1 + \sqrt{5}R + \frac{5}{3}R^2\right)\exp\left(-\sqrt{5}R\right)$
	–	ARD rational quadratic	$k(x_i, x_j\|\theta) = \sigma_f^2\left(1 + \frac{1}{2\alpha}\sum_{m=1}^{d}\frac{(x_{im} - x_{jm})^2}{\sigma_m^2}\right)^{-\alpha}$

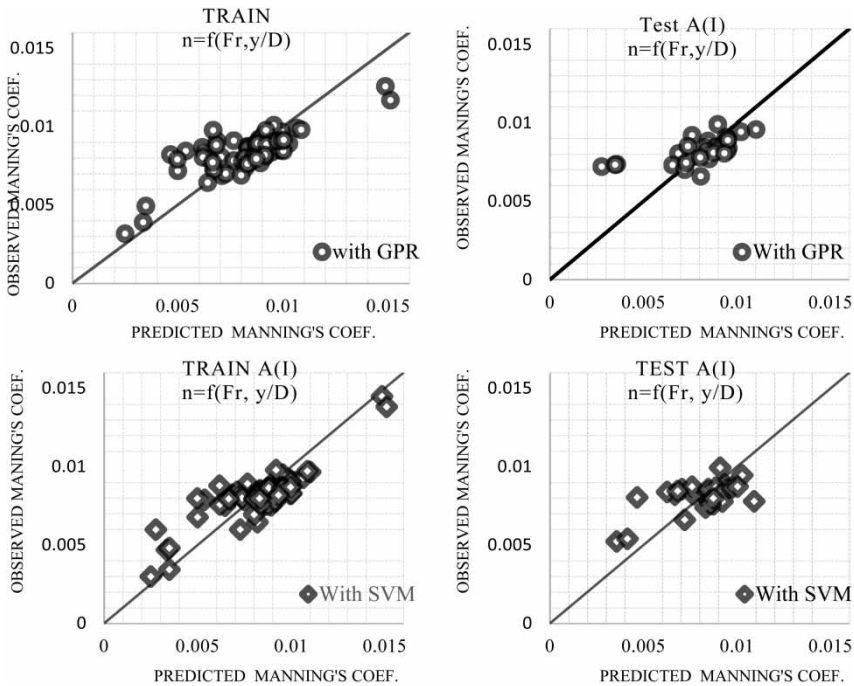


Figure 2 | Comparison of observed and predicted Manning coefficient for the best dimensionless model (A(I)).

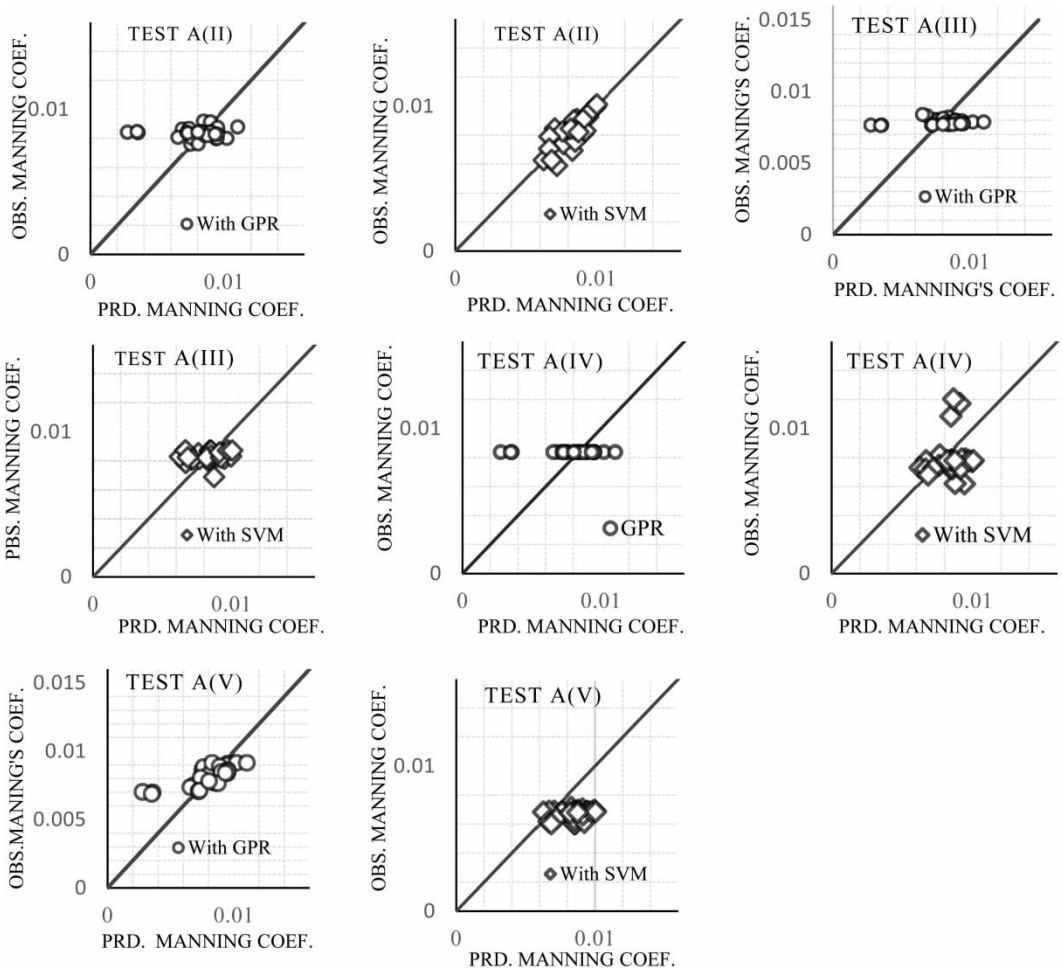


Figure 3 | Comparison of observed and predicted Manning coefficient for dimensionless models.

Models with dimensionless input variables

GPR-based models

From Table 5, it can be concluded that considering the Froude number and relative depth as input variables led to more accurate results, performance criteria were $R = 0.786$, DC = 0.989, and RMSE = 0.0012 for training and $R = 0.738$, DC = 0.962, and RMSE = 0.0015 for testing. Therefore, model A(I) is identified as superior among the five models. Model A(V) showed acceptable results, and the analysis revealed that the geometric parameter y/D had a significant impact on obtained results, and adding the variable input of Froude number caused an increase in model efficiency. Furthermore, performance criteria (R, DC, RMSE) of models A(II), A(III), and A(IV) exhibited an undesired efficiency in the prediction of n. The n is more affected by the Froude number than by the Reynolds number. This outcome revealed that, in the presence of insufficient information, the n could be estimated using the input variables Fr and y/D. Figures 2 and 3 demonstrate the comparison of predicted and observed n of GPR models.

SVM-based models

Table 5 features the efficiency of the SVM method, in which the accuracy of results is increased compared to the GPR method. Table 4 and Figures 1 and 2 demonstrate the enhanced efficiency of the SVM method in predicting the n. Model A(I) with input parameters of Fr and y/D leads to better results and high accuracy with R, DC, and RMSE equal to ($R = 0.934$, DC = 0.99, RMSE = 0.0007) and ($R = 0.808$, DC = 0.993, RMSE = 0.0006) for training and testing data, respectively. From model A(V), it is concluded that parameter y/D seems to be the most influential parameter in predicting the n. However, adding the Froude number as an input variable increases the model efficiency. Since the results obtained

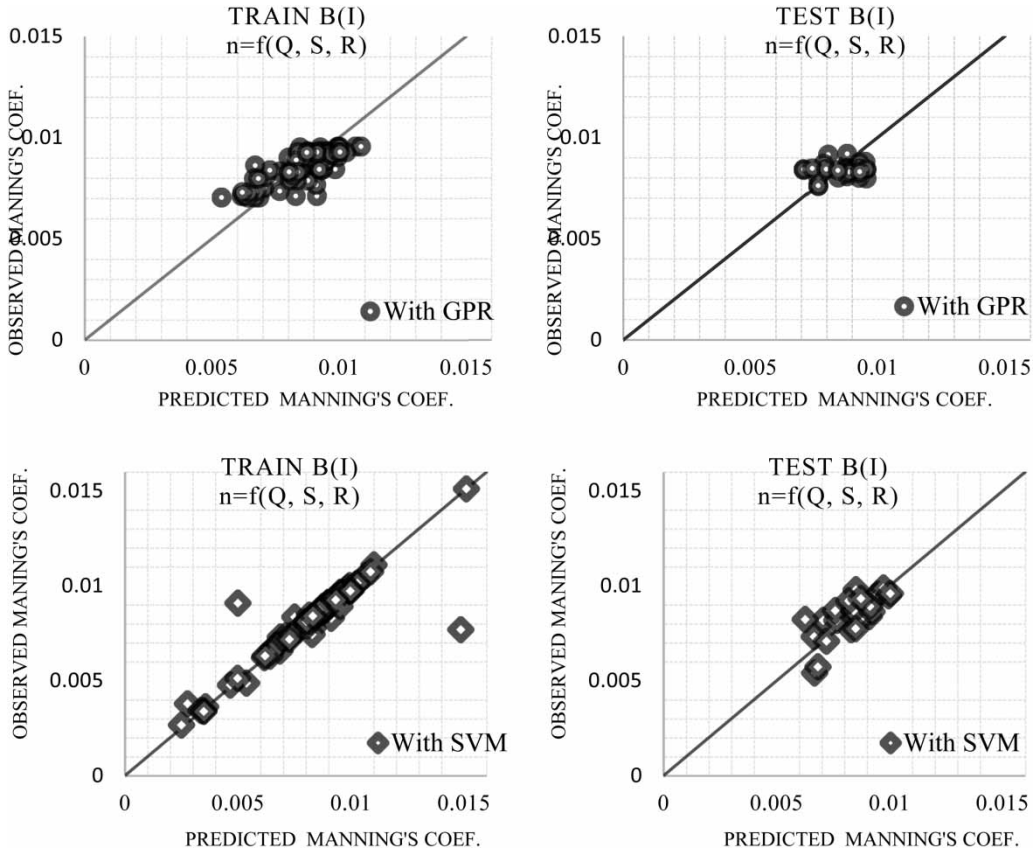


Figure 4 | Comparison of observed and predicted Manning coefficient for the best model with dimensional variables (B(I)).

from the GPR method are consistent with the results obtained from the SVM method, incapability of A(II), A(III), and A(IV) models in acquiring reasonable results were expected.

From Figure 2, it is observed that the proximity of the observed and predicted values to the $x = y$ line and the appropriate clutter of data confirms the high accuracy of model A(I), in which the predicted test data range values for SVM and GPR approaches were 0.00356–0.01 and 0.0027–0.011, respectively. Comparison of observed and predicted data for GPR and SVM approaches in Figure 2 confirms the high efficiency of the SVM approach compared to the GPR approach.

According to Figure 3, the distribution of data along the 1:1 line in models A(II) and A(IV) showed the inability of model in predicting the n. Furthermore, it is observed that model A(III) with the input parameter of Fr had an inappropriate distribution along the line 1:1 compared to model A(V) with the input parameter of y/D, which confirms that y/D is the most effective parameter in modeling.

Models with dimensional input variables

GPR-based models

The comprehensive evaluation of models was performed considering the second category of input variables, in which the variables have dimensions (Table 6 and Figures 4 and 5). Although models B(II), B(III), and B(IV) did lead to acceptable results, from R, DC, and RMSE values, it was concluded that model B(I) with input variables of Q, S_0, and R_H was the best model in predicting n (training = 0.760, 0.992, 0.00075 and testing = 0.742, 0.991, 0.0007), respectively. Also, comparing models B(II), B(III), B(IV) with model B(I) showed that the input variable Q was the most important parameter in the prediction of n. Identifying the process of the most effective input variable will further be discussed.

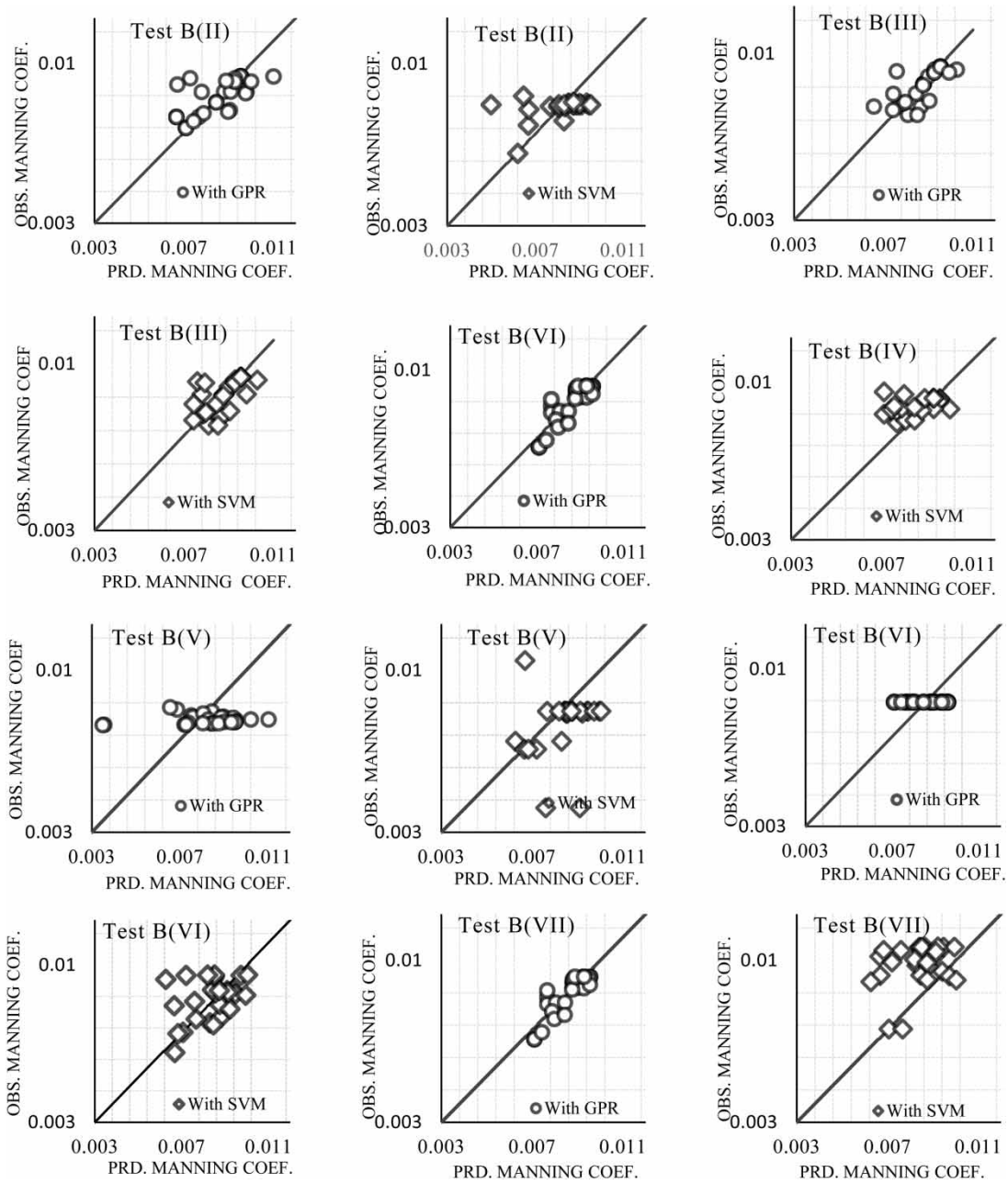


Figure 5 | Comparison of observed and predicted Manning coefficient for models with dimensional variables.

SVM-based models

Evaluation criteria of the SVM method are presented in Table 6, in which the increase in the accuracy of SVM method was evident. Exploring performance criteria showed that performance criterion of model B(I), which is equal to (training: $R = 0.902$, DC = 0.987, RMSE = 0.0009) and (testing: $R = 0.742$, DC = 0.991, RMSE = 0.0007), provides better results and therefore is selected as the best model. The results of the SVM method are in good accordance with the GPR method; therefore, models B(II), B(III), B(IV), B(V), B(VII), and B(VII) are unable in predicting the n. Comparison of observed and predicted Manning coefficients via the SVM method is represented in Figures 4 and 5.

Although the data destitution of data around line 1:1 for models B(II), B(III), and B(IV) in Figure 5 was observed to be acceptable, more proportional dispersion and proximity of data to line 1:1 in Figure 4 confirms that model B(I) is the best

Table 5 | Statistical parameters of the GPR and SVM models with dimensionless input variables

		Evaluation criteria					
		Training			Testing		
Models		DC	*R*	RMSE	*R*	DC	RMSE
A(I)	GPR	0.786	0.989	0.0012	0.738	0.962	0.0015
	SVM	0.934	0.99	0.0007	0.818	0.993	0.0006
A(II)	GPR	0.543	0.950	0.0012	0.340	0.854	0.0017
	SVM	0.550	0.953	0.0018	0.27	0.984	0.001
A(III)	GPR	0.510	0.962	0.0017	0.260	0.904	0.0020
	SVM	0.305	0.930	0.020	0.200	0.961	0.0016
A(IV)	GPR	0.490	0.940	0.0017	0.190	0.831	0.0020
	SVM	0.262	0.900	0.0025	0.228	0.940	0.0019
A(V)	GPR	0.770	0.970	0.0013	0.520	0.968	0.0013
	SVM	0.849	0.984	0.0012	0.809	0.973	0.0013

Table 6 | Statistical parameters of the GPR and SVM models with dimensional input variables

		Evaluation criteria					
		Training			Testing		
Models		DC	*R*	RMSE	*R*	DC	RMSE
B(I)	GPR	0.760	0.992	0.00075	0.719	0.949	0.0013
	SVM	0.902	0.987	0.0009	0.742	0.991	0.0007
B(II)	GPR	0.720	0.915	0.0008	0.700	0.909	0.0014
	SVM	0.812	0.925	0.0015	0.592	0.901	0.0015
B(III)	GPR	0.746	0.971	0.00079	0.733	0.930	0.0013
	SVM	0.96	0.992	0.0007	0.610	0.983	0.001
B(IV)	GPR	0.85	0.992	0.00076	0.652	0.897	0.0010
	SVM	0.625	0.914	0.001	0.435	0.920	0.0018
B(V)	GPR	0.51	0.976	0.017	0.261	0.940	0.0020
	SVM	0.356	0.923	0.0023	0.311	0.953	0.0017
B(VI)	GPR	0.354	0.405	0.0020	0.287	0.935	0.0021
	SVM	0.317	0.900	0.002	0.282	0.949	0.0018
B(VII)	GPR	0.827	0.990	0.0079	0.630	0.887	0.0011
	SVM	0.501	0.950	0.0018	0.413	0.983	0.0010

model in predicting the n value. The predicted test data range values for model B(I) in SVM and GPR approaches are 0.0066–0.0097 and 0.0076–0.0095, respectively. Scattered arrangement of data around the line 1:1 for models B(V), B(VI), and B(VII) showed the inability of models in predicting the n value.

Sensitivity analysis

The determination of the contribution of each parameter on the n of a culvert is evaluated with a sensitivity analysis. There are different methods for doing sensitivity analysis such as local or global, quantitive or qualitative, or OAT. One of the simplest and most common approaches is that of changing one-factor-at-a-time, to see what effect this produces on the output. AOT sensitivity analysis essentially consists of selecting a base parameter setting (nominal set) and varying one parameter at a time while keeping all other parameters fixed (hence, it is referred to as a local method). An important use of OAT is to reveal the form of the relationship between the varied parameter and the output, given that all other parameters have their nominal values. (Holvoet *et al.* 2005; Roushangar *et al.* 2019).

This study applied the OAT sensitivity analysis for two groups of models with different input variables for the SVM method. To develop an OAT analysis, the superior models were run with all variables, then one of the parameters was omitted and the process was repeated. According to Figure 6, for models with dimensionless variables, it was deduced that the elimination of y/D from the best model would decrease the model efficiency significantly from ($R = 0.808$, DC $= 0.993$, RMSE $= 0.006$) to ($R = 0.2$, DC $= 0.961$, RMSE $= 0.0016$). Consequently, it could be stated that y/D is the most effective parameter in predicting n using dimensionless input variables. Because of the unfavorable results of model A(II), OAT analysis was not performed with this model.

The OAT sensitivity analysis was performed for the second category, in which the input variables have dimensions. Figure 7 indicates that eliminating parameters of discharge (Q) and hydraulic radius (R_H) would decrease the model efficiency from ($R = 0.742$, DC $= 0.991$, RMSE $= 0.0007$) to ($R = 0.435$, DC $= 0.92$, RMSE $= 0.0018$) and ($R = 0.592$, DC $= 0.90$, RMSE $=$ 0.00015), respectively. Therefore, it was concluded that the most effective parameters in predicting n are discharge (Q) and hydraulic radius (R_H).

CONCLUSION

The calculation of n is inherently a challenging matter since multiple factors influence it. Many previous studies have pointed out the dependency of the n on geometric and hydraulic parameters. Therefore, the comprehensive evaluation of parameters affecting the Manning coefficient could be useful for accurate calculation of head loss in culvert systems for the optimal design of culverts. In the present study, the capability of GPR and SVM models as kernel-based approaches was verified for predicting n of a culvert system. In this regard, the laboratory data of Devkota (2012) for an HDPE culvert were used. In order to perform a detailed investigation, two categories of models were developed, in which the input data of the first category were dimensionless and the input data of the second category have dimensions. The obtained results are as follows:

- Among various kernel functions, using kernel functions of squared-exponential in GPR and RBF in the SVM approach led to more accurate predictions.
- Considering the Froude number and relative depth as dimensionless input variables in model A(I) leads to more accurate results. The geometric parameter y/D had a significant impact on the results and adding the Froude number caused an increase in model efficiency.
- The efficiency of the SVM method is better than the GPR method. Similar to the GPR method, model A(I) with input parameters of Fr and y/D leads to the best results. From model A(V), it is concluded that parameter y/D seems to be the most influential parameter. However, adding the Froude number as an input variable increases model efficiency. Moreover, models A(II), A(III), and A(IV) do not lead to accurate results.

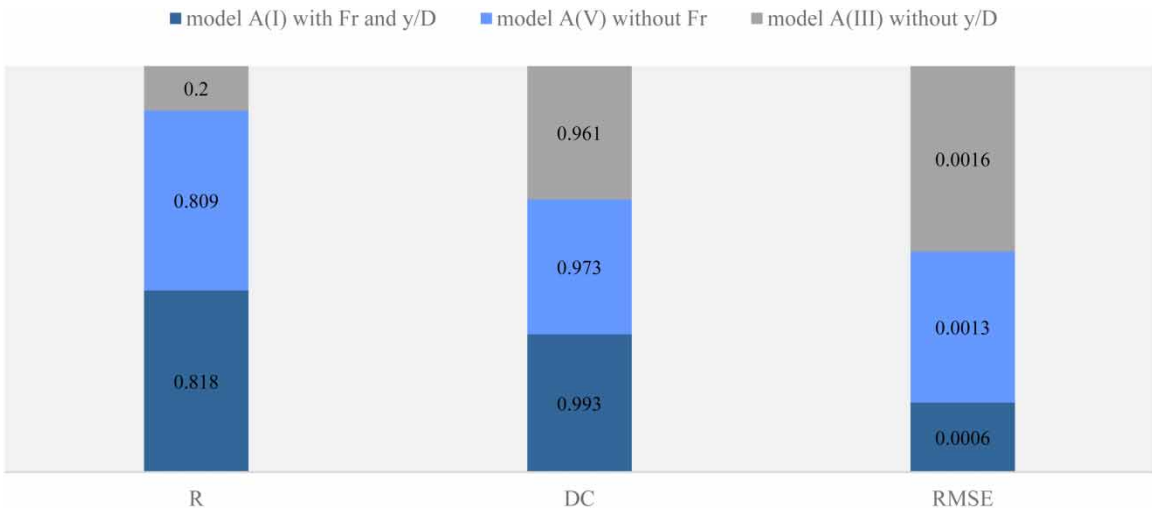


Figure 6 | Relative significance of each input parameters of the best models with dimensionless variables.

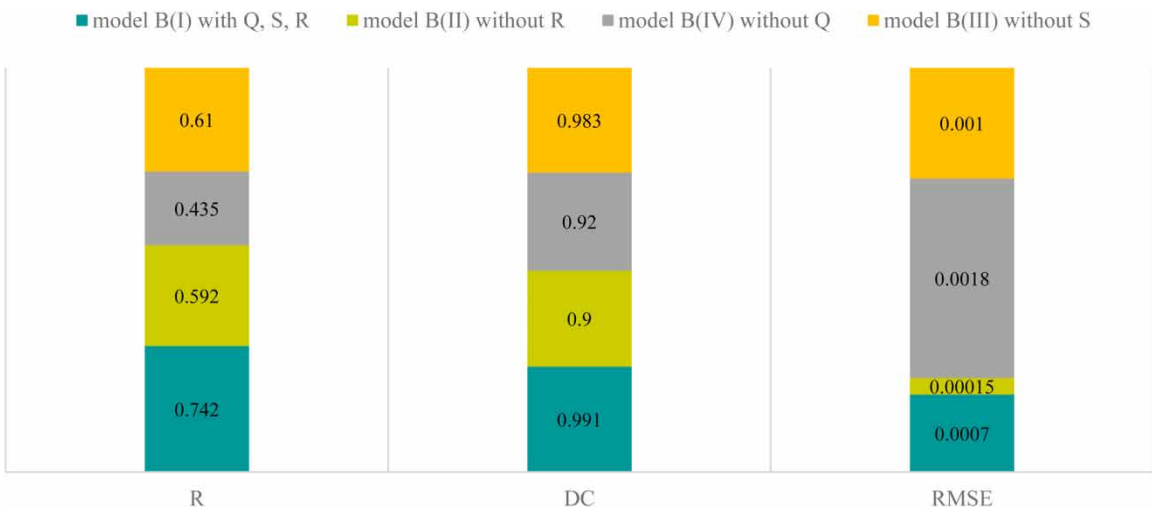


Figure 7 | Relative significance of each input parameters of the best models with dimensional variables.

- GPR models for dimensional input variables showed that model B(I) with input variables of Q, S_0, and R_H is the best model (testing = 0.742, 0.991, 0.0007). Also, comparing models B(II), B(III), and B(IV) with model B(I) showed that the input variable Q was the most important parameter in estimating n.
- The obtained evaluation criteria of the SVM method for dimensional input variables indicated that model B(I) with (R = 0.902, DC = 0.987, RMSE = 0.0009) for training data and (R = 0.742, DC = 0.991, RMSE = 0.0007) for testing data led to better results and is selected as the best model. The results of the SVM method are in good accordance with the GPR method, and the results of B(II), B(III), B(IV), B(V), B(VII), and B(VII) models do not provide good accuracy.
- Applying the OAT sensitivity analysis for the SVM method in models with dimensionless variables, the elimination of parameter y/D from the best model decreased the model efficiency significantly from (R = 0.808, DC = 0.993, RMSE = 0.006) to (R = 0.2, DC = 0.961, RMSE = 0.0016) and is the most effective parameter.
- Furthermore, OAT sensitivity analysis on the second category indicated that eliminating parameters of discharge (Q) and hydraulic radius (R_H) would decrease the model efficiency. Therefore, it was concluded that the most effective parameters in predicting n are discharge (Q) and hydraulic radius (R_H), respectively.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Amaranto, A., Munoz-Arriola, F., Corzo, G., Solomatine, D. P. & Meyer, G. 2018 Semi-seasonal groundwater forecast using multiple data-driven models in an irrigated cropland. *Journal of Hydroinformatics* **20** (6), 1227–1246.

American Concrete Pipe Association 2007 *Manning's n Value A History of Research*.

Ardıçlıoğlu, M. & Kuriqi, A. 2019 Calibration of channel roughness in intermittent rivers using HEC-RAS model: case of Sarimsakli creek, Turkey. *SN Applied Sciences* **1** (9), 1080.

Arlot, S. & Lerasle, M. 2016 Choice of V for V-fold cross-validation in least-squares density estimation. *The Journal of Machine Learning Research* **17** (1), 7256–7305.

Attari, M. & Hosseini, S. M. 2019 A simple innovative method for calibration of Manning's roughness coefficient in rivers using a similarity concept. *Journal of Hydrology* **575**, 810–823.

Ballesteros, J. A., Bodoque, J. M., Díez-Herrero, A., Sanchez-Silva, M. & Stoffel, M. 2011 Calibration of floodplain roughness and estimation of flood discharge based on tree-ring evidence and hydraulic modelling. *Journal of Hydrology* **403** (1–2), 103–115.

Bhave, P. R. & Gupta, R. 2006 *Analysis of Water Distribution Networks*. Alpha Science Int'l Ltd.

Bhowmik, T. K., Ghanty, P., Roy, A. & Parui, S. K. 2009 SVM-based hierarchical architectures for handwritten Bangla character recognition. *International Journal on Document Analysis and Recognition (IJDAR)* **12** (2), 97–108.

Bloodgood, D. E. & Bell, J. M. 1961 Manning's coefficient calculated from test data. *Journal (Water Pollution Control Federation)* 176–183.

Boulomytis, V. T. G., Zuffo, A. C., Dalfré Filho, J. G. & Imteaz, M. A. 2017 Estimation and calibration of Manning's roughness coefficients for ungauged watersheds on coastal floodplains. *International Journal of River Basin Management* **15** (2), 199–206.

Boyer, M. C. 1954 Estimating the Manning coefficient from an average bed roughness in open channels. *Eos, Transactions American Geophysical Union* **35** (6), 957–961.

Carvalho, J., Santos, J. P. V., Torres, R. T., Santarém, F. & Fonseca, C. 2018 Tree-based methods: concepts, uses and limitations under the framework of resource selection models. *Journal of Environmental Informatics* **32** (2).

Devkota, J. 2012 *Variation of Manning's Roughness Coefficient with Diameter, Discharge, Slope and Depth in Partially Filled HDPE Culverts*. *Doctoral Dissertation*.

Ferreria, C. 2001 Gene expression programming: a new adaptive algorithm for solving problems. *Complex Systems* **13** (2), 87–129.

Hosseini, S. M., Bousmar, D., Zeck, Y. & De Almeida, B. 2000 Energy and momentum in one dimensional open channel flow. *Journal of Hydraulic Research* **38**.

Kisi, O., Shiri, J., Karimi, S., Shamshirband, S., Motamedi, S., Petković, D. & Hashim, R. 2015 A survey of water level fluctuation predicting in Urmia Lake using support vector machine with firefly algorithm. *Applied Mathematics and Computation* **270**, 731–743.

Kitsikoudis, V., Sidiropoulos, E., Iliadis, L. & Hrissanthou, V. 2015 A machine learning approach for the mean flow velocity prediction in alluvial channels. *Water Resources Management* **29** (12), 4379–4395.

Lavoie, B. & Mahdi, T. F. 2020 Manning's roughness coefficient determination in laboratory experiments using 2D modeling and automatic calibration. *La Houille Blanche* (1), 22–33.

Li, Z. & Zhang, J. 2001 Calculation of field Manning's roughness coefficient. *Agricultural Water Management* **49** (2), 153–161.

McKay, S. K. & Fischenich, J. C. 2011 Robust prediction of hydraulic roughness (No. ERDC/CHL-CHETN-VII-11). In *Engineer Research and Development Center Vicksburg MS Coastal and Hydraulics Lab*.

Mohammadpour, R., Zainalfikry, M. K., Zakaria, N. A., Ghani, A. A. & Weng Chan, N. 2019 Manning's roughness coefficient for ecological subsurface channel with modules. *International Journal of River Basin Management* 1–13.

Neal, R. M. 2012 *Bayesian Learning for Neural Networks*, Vol. 118. Springer Science & Business Media.

Owen, N. E. & Liuzzo, L. 2019 Impact of land use on water resources via a Gaussian process emulator with dimension reduction. *Journal of Hydroinformatics* **21** (3), 411–426.

Pandey, P., Dongre, S. & Gupta, R. 2020 Probabilistic and fuzzy approaches for uncertainty consideration in water distribution networks – a review. *Water Supply* **20** (1), 13–27.

Quinonero-Candela, J., Rasmussen, C. E. & Williams, C. K. 2007 Approximation methods for Gaussian process regression. In: *Large-scale Kernel Machines*. MIT Press, pp. 203–223.

Rasmussen, C. E. & Williams, C. K. I. 2006 *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, MA.

Roushangar, K. & Shahnazi, S. 2020 Prediction of sediment transport rates in gravel-bed rivers using Gaussian process regression. *Journal of Hydroinformatics* **22** (2), 249–262.

Roushangar, K., Saghebian, S. M. & Mouaze, D. 2017 Predicting characteristics of dune bedforms using PSO-LSSVM. *International Journal of Sediment Research* **32** (4), 515–526.

Roushangar, K., Matin, G. N., Ghasempour, R. & Saghebian, S. M. 2019 Evaluation of the effective parameters on energy losses of rectangular and circular culverts via kernel-based approaches. *Journal of Hydroinformatics* **21** (6), 1014–1029.

Saghebian, S. M., Roushangar, K., Ozgur Kirca, V. S. & Ghasempour, R. 2020 Modeling total resistance and form resistance of movable bed channels via experimental data and a kernel-based approach. *Journal of Hydroinformatics* **22** (3), 528–540.

Teodorescu, L. & Sherwood, D. 2008 High energy physics event selection with gene expression programming. *Computer Physics Communications* **178** (6), 409–419.

Tinelli, S. & Juran, I. 2019 Artificial intelligence-based monitoring system of water quality parameters for early detection of non-specific bio-contamination in water distribution systems. *Water Supply* **19** (6), 1785–1792.

Zhu, S., Luo, X., Xu, Z. & Ye, L. 2019 Seasonal stream-flow forecasts using mixture-kernel GPR and advanced methods of input variable selection. *Hydrology Research* **50** (1), 200–214.

First received 1 January 2022; accepted in revised form 12 July 2022. Available online 23 July 2022