Prediction of river suspended sediment load using machine learning models and geo-morphometric parameters

Maryam Asadi¹ · Ali Fathzadeh² · Ruth Kerry³ · Zohre Ebrahimi-Khusfi⁴ · Ruhollah Taghizadeh-Mehrjardi^{5,6}

Received: 25 June 2020 / Accepted: 26 June 2021 / Published online: 31 August 2021
© The Author(s) 2021

Abstract

Estimating sediment load of rivers is one of the major problems in river engineering that has been using various data mining algorithms and variables. It is desirable to obtain accurate estimates of sediment load while using techniques that limit computational intensity when datasets are large. This study investigates the usefulness of geo-morphometric factors and machine learning (ML) models for predicting suspended sediment load (SSL) in several river basins in Lorestan and Gilan, Iran. Six ML models, namely, multiple linear regression (MLR), artificial neural networks (ANN), K-nearest neighbor (KNN), Gaussian processes (GP), support vector machines (SVM), and evolutionary support vector machines (ESVM), were evaluated for estimating minimum and average SSL for the study regions. Geo-morphometric parameters and river discharge data were utilized as the main predictors in modeling process. In addition, an attribute reduction technique was applied to decrease the algorithm complexity and computational resources used. The results showed that all models estimated both target variables well. However, the optimal models for predicting average sediment load and minimum sediment load were the GP and ESVM models, respectively.

Keywords Evolutionary support vector machine · Sediment load · Geo-morphometric parameters · Gaussian processes · River basin

Introduction

Estimating suspended sediment load (SSL) of rivers is a major objective in water resource planning because it has a key role in the design and construction of water-related structures and has major implications for erosion management and sediment redistribution within watersheds (Halbe et al. 2013; Rajaei 2011). Sediments removed by upstream components of

turbulent flow that remain suspended for a long time are defined as suspended loads (Kaveh et al. 2017). Erosion of soil throughout a drainage basin and sediment input to river flow cause a reduction in water resource quality and the useful lifetime of hydraulic structures. Prediction of SSL could help to control water quality and increase the effectiveness of hydro-electric power installations. Measuring SSL requires the installation of hydrometric stations, which is costly

Responsible Editor: Stefan Grab

✉ Ali Fathzadeh
fatzade@ardakan.ac.ir

Maryam Asadi
m.asadi.96@ut.ac.ir

Ruth Kerry
ruth_kerry@byu.edu

Zohre Ebrahimi-Khusfi
zohreebrahimi2018@ujiroft.ac.ir

Ruhollah Taghizadeh-Mehrjardi
ruhollah.taghizadeh-mehrjardi@mnf.uni-tuebingen.de;
rtaghizadeh@ardakan.ac.ir

¹ College of Natural Resources, Tehran University, Tehran, Iran

² Faculty of Agricultural and Natural Resources, Ardakan University, Ardakan, Iran

³ Department of Geography, Brigham Young University, Provo, USA

⁴ Department of Ecological Engineering, Faculty of Natural Resources, University of Jiroft, Jiroft, Iran

⁵ Department of Geosciences, Soil Science and Geomorphology, University of Tübingen, Rümelinstr. 19-23 Tübingen, Germany

⁶ Faculty of Agricultural and Natural Resources, Ardakan University, Ardakan, Iran

financially and in terms of human labor, and it is impossible in many remote regions. In such regions, using methods that can estimate SSL, indirectly and less expensively, would be desirable for improved sediment and water resource management.

Asian countries are very sensitive to flood flows, which can carry large SSL and lead to severe socio-economic damage (Das et al. 2018). Climate change and human interference exacerbate these conditions (Das et al. 2018; Malik and Pal 2020b). Therefore, predicting natural disasters and the SSL can be effective in reducing the negative consequences of such events (Malik and Pal 2020a; Melesse et al. 2011). Several authors have already used different mathematical modeling techniques to predict SSL of rivers (Verstraeten and Poesen 2001; Ward et al. 2009; Bezak et al. 2014; Choi and Lee 2015; Wang et al. 2015; Si et al. 2017). Such mathematical modeling requires a large amount of input data that is costly and time-consuming to collect. One of the most common methods for predicting SSL is through using sediment rating curves that show the relationship between sediment load and flow. Unfortunately, there are many errors associated with using this method (Rajaei et al. 2009; Asselman 2000).

There are many different factors in watersheds that could influence the amount of SSL in watersheds. By gathering data on these factors and using them with experimental models, SSL can be estimated. So far data such as drainage characteristics, topography, river discharge, rainfall levels, and vegetation cover have been used to estimate SSL with a range of data mining techniques (Kisi 2012; Lafdani et al. 2013; Agarwal et al. 2006; Haddadchi et al. 2013; Vafakhah 2013; Nourani and Andalib 2015).

Data mining science incorporates aspects of several disciplines such as statistics, artificial intelligence, and machine learning. The methods used involve making predictions, using carefully collected data, and for each model suitably processing databases. Zounemat-Kermani et al. (2016) modeled SSL based on data-oriented models such as ANNs and SVR to allow comparison with predictions from sediment rating curves. Results indicated that SVR models produced by radial basis functions (RBFs) were the most suitable for predicting SSL. Lafdani et al. (2013) also used ANNs and SVR to forecast SSL, but they considered streamflow and precipitation as explanatory variables and showed that the SVR model produced the most reliable predictions. Kisi (2012) modeled the relationship between SSL and discharge using SVR, ANNs, and sediment rating curves at the downstream and upstream stations in California. For the upstream stations, the SVR model was the best, but for the downstream stations, sediment rating curves produced the best results. In flood studies, the boosted regression tree (BRT) model was found to be the best predicting flash flood susceptibility using climatic and geo-environmental variables (Band et al. 2020a). In another study, flood susceptibility mapping was successfully performed using logistic regression compared to multi-criteria

decision analysis approaches (Malik et al. 2020a). The ensemble of MLR with an evidence belief function (EBF) has also shown good performance in predicting flood-sensitive areas (Chowdhuri et al. 2020).

Using additional climatic and hydrological data can lead to increased model accuracy in prediction of SSL. Zhu et al. (2007) modeled SSL by using ANNs and mean precipitation, temperature, precipitation intensity, and discharge data related to the Yangtse River, China. Their results showed that ANNs and these data could provide better predictions of monthly SSL compared with MLR models.

The physiographic characteristics of drainage basins have a critical role in determining the amount of water and sediment in channels as well as channel morphology downstream. Some of these characteristics are slope and elevation (Malik and Pal 2020a, b; Malik et al. 2020b). With regard to the specific role of different physiographic characteristics in erosion and sedimentation processes, only very limited research has taken place such as the work of Talebi et al. (2016) who reviewed the role of curvature and its types on the rate of erosion. With current geographic information system (GIS) software, extraction of topographic and physiographic features from digital elevation models (DEMs) is relatively simple, and they have the potential to improve prediction of suspended sediment loads compared to the studies mentioned earlier that did not use such a key group of variables as inputs to modelling.

While physiographic attributes can be easily extracted from DEMs and data from many sources can be included in modeling suspended sediment loads, there can be computational issues in terms of sufficient memory and computer storage space as well as computational time when making predictions over large geographical areas like multiple drainage basins. For many algorithms as the space used for input data is extended, the complexity of algorithm implementation is increased. Given that some input data have a small role in prediction yet they increase the space needed for input data and the computational complexity, attribute reduction techniques could decrease input data space without a significant loss in prediction accuracy (Taghizadeh-Mehrjardi et al. 2016).

Gilan and Lorestan provinces are rainy and important to the water supply of Iran. The majority of dams in Iran are on rivers that originate in these two provinces. Therefore, being aware of the minimum and average SSL in a year can assist in the proper management of the dams. In this regard, discharge data are often used to predict minimum and average sediment load (Kakaei Lafdani et al. 2013; Zounemat-Kermani et al. 2016). Due to the non-linear relationship between discharge and deposition and the complexity of sediment transport processes, and factors affecting them (Kisi 2012), this study will use geomorphometric parameters that can be easily extracted from DEMs in addition to discharge data. This study not only investigates what are the best ML techniques for doing this, but also uses the SVM method to determine which are the most

important geo-morphometric parameters for predicting SSL in these regions of Iran.

The major objective of this work was to evaluate six ML models for predicting SSL for large geographic areas covering many drainage basins. Using such methods could greatly reduce the cost of predicting SSL compared to traditional methods. The methods compared were MLR, ANNs, KNN, GP, SVM, and ESVM. The novelty of this work lies in the range of methods compared as well as the use of topographic/geo-morphometric attributes for improving prediction as they have largely been ignored in previous research. Data reduction methods should also enable this key wealth of topographic/geo-morphometric data to be utilized without a proportionate increase in computational complexity and input data storage capacity given the large size of the datasets used.

Materials and methods

Study sites

The study sites consist of 68 watersheds located in the Gilan and Lorestan provinces, Iran. These watersheds have a suitable data bank compared with other watersheds in Iran. The altitude varies between −89 and 3702 m above sea level in the study sites. The two regions have different vegetation types. The geographic location of the study regions and hydrometric stations are shown in Fig. 1.

Data used

ML models need appropriate input parameters for predicting and modeling processes. In this study, discharge and geo-morphometric parameters were used as input parameters. DEM with a resolution of 30 m was downloaded from the United States Geological Survey (USGS). In addition, 26 geo-morphometric parameters were extracted as auxiliary input parameters using SAGA (System for Automated Geoscientific Analyses) GIS.

In this study, SSL recorded at the hydrometric stations for 1983–2014 were used as the target or dependent variable. Minimum and average SSL were calculated to validate the output of the models used in the study.

The methodology used for the current study is represented in Figure 2, and more details are described below.

Principal component analysis (PCA)

Reducing the number of variables is a useful tool for decreasing the volume of input data when modeling large datasets (Ebrahimi-Khusfi et al. 2021). PCA was used to decrease the complexity of computation and determine which were the most important variables (Taghizadeh-Mehrjardi et al.

2016). PCA was done prior to modeling of sediment loads by various methods. As all variables are used in the formation of the principal components, there is a low loss of information from the primary variables (Johnson and Wichern 2007).

ML methods used in predicting the SSL

In this work, minimum and average SSL were predicted using the MLR, ANNs, KNN, GP, SVM, and ESVM models. To investigate the contribution of auxiliary variables in the predicting of SSL, first minimum and average SSL were modeled with the corresponding discharge.

Multiple linear regression (MLR)

MLR aims to specify association strength between one dependent variable and several independent variables (Aiken et al. 2003). Relations among a dependent variable and one or more independent variables are modeled through fitting a linear equation to actual data, as follows (Eq. 1).

$$y = a_0 + a_1x_1 + a_2x_2 + \dots + a_px_p \quad (1)$$

where y is dependent variable; $x_1, x_2, \dots, x_p$ are independent variables; a_0 is the y intercept; and $a_1, a_2, \dots, a_p$ are the coefficients of variables under investigation. In this study, the SSL was considered as the dependent variable, and the geo-morphometric parameters were used as the independent variables.

Artificial neural networks (ANNs)

ANNs are characterized with one or more hidden layers (Sahana et al. 2020). ANNs are a large parallel system which consists of enormous processing components connected by weight links. The most widely used ANN model is the feed-forward back-propagation network (FFBP) (Haykin and Lippmann 1994). The ANN includes some layers of parallel processing components called neurons. Each layer is completely connected to the previous one via inter-connecting weights. At each iteration, determined weight amounts are initially altered as the calibration or training process proceeds. This compares predicted outputs with real outputs and back-propagates any errors to determine appropriate weight adjustments so that errors can be minimized (Kişi 2010).

Initially, the number of hidden layers was changed from 0 to 20 to show that 16 hidden layers gave the best results. Training cycle tools were used to determine the number of training cycles that was best, and 500 training cycles were shown to give the best results. The model also had a learning rate of 3/0 and a momentum of 2/0 as analysis was conducted to optimize these parameters.

Figure 1 Location of hydrometric stations in the study area

Figure 2 Flowchart of steps involved in the study

In this method, a function of sigmoid $\phi(x)$ was expressed by Eq 2:

$$d^i = \phi \left(\sum_i w_i \cdot \phi \left(\sum_k w'_{ik} \left(\sum_j f_{kj} \cdot w_{jk} + k_n \right) + k'_m \right) + k'' \right) \quad (2)$$

where f_p is a feature of input kn and k'_m and k'' are neurons of different layers. Also, w_i , w'_{ij} , and w'_{jk} are weights of the different input layers.

K-nearest neighbor (KNN)

The KNN model was suggested by Nemes et al. (2006). It does not need particular assumptions about the distribution of predictors, which is regarded as one of the advantages of this algorithm. The KNN samples are categorized based on the mean values of k neighbor responses in a predictor space (Mahmoudzadeh et al. 2020). The training examples are characterized by n attributes. Every example denotes a point in an n -dimensional space. Hence, all training examples will be maintained in an n -dimensional pattern space (Jayawardena et al. 2002). In this model, model optimization results showed that $k=3$ gave the best results.

Gaussian process (GPs)

As a useful non-parametric ML technique, GP can be applied to build comprehensive probabilistic models of real-world issues. A GP is a stochastic procedure whose substantiation is constituted of randomized amounts which are related with every point in spatial and temporal scales such that every random variable is normally distributed (Roushangar and Shahnazi 2020). In addition, every finite set of those random variables is characterized with a multivariate normal distribution. GPs play a substantial role in statistical modeling thanks to properties inherited from the normal distribution. In GPs, multivariate Gaussian distributions are extended to unlimited dimensionality. As a whole, a GP produces data nested over some domain such that every finite subset will follow a multivariate Gaussian distribution. In GP models, different kernel functions give different model accuracies. In this study, the functions investigated included radial basis functions (RBF), Cauchy, Laplace, poly, sigmoid, Epanechnikov, Gaussian combination, and multi-quadratic functions.

Support vector machine (SVM)

The SVM (Cortes and Vapnik 1995) has been widely used to predict different target variables, such as gully erosion and SSL (Rashidi et al. 2016; Band et al. 2020b). This algorithm measures functional dependency $f(x)$ between X and Y while assuming that these samples do not follow a specific probability distribution function $P(\vec{x}, y)$:

$$F = \left\{ f | f(\vec{x}) = (\vec{K}, \vec{x}) + C : \vec{K} \in R^n, R^n \rightarrow R \right\} \quad (3)$$

where K and C are coefficients obtained from sampling data (X, Y) . Considering that the probability distribution function of sampling data is unknown, it cannot reduce the risk function. Therefore, the empirical risk function is calculated using Eq. 4 (Cimen 2008).

$$R_{emp} [f(\vec{x})] = \frac{1}{T} \sum_{i=1}^N l(y_1 - f(\vec{x}_i)) \quad (4)$$

where T denotes the number of sampled data. However, the empirical risk minimization is not appropriate to reduce the risk function. To reduce it, the regularized risk function is used as expressed by Eq. 5:

$$R_{reg} [f(\vec{x})] = d \sum_{x_i \in X} l_\varepsilon(y_1 - f(\vec{x}_i)) + \frac{1}{2} \|\vec{K}\|^2 \quad (5)$$

In Eq. 5, $\|\vec{K}\|$ and d refer to a regression vector and a coefficient constant, respectively. l_ε is the loss function and is defined by Eq. 6:

$$l_\varepsilon \left(y_1 - f \left(\vec{x}_i \right) \right) = \begin{cases} 0 & \text{for } |y_1 - f(\vec{x}_i)| < \varepsilon \\ y_1 - f(\vec{x}_i), & \text{otherwise} \end{cases} \quad (6)$$

The non-linear regression function is computed as (Cimen 2008; Gunn 1998):

$$f(x) = \sum_{i=1}^N (l_i^* - l_i) K(x, x_i) + bc \quad (7)$$

where bc is a bias component, l_i^* and l_i are the Lagrange multipliers, and $k(x, x_i)$ is the kernel function (Kisi and Cimen, 2011).

In the SVM and ESVM models, different kernel functions have differing accuracies. The functions investigated in this study for SVM and ESVM were Cornell dot, radial, polynomial, neural, ANOVA, Epachnenikov, Gaussian combination, and multi-quadratic.

Support vector machine evolutionary (ESVM)

The ESVM is a SVM implementation using evolutionary algorithms to deal with the dual optimization problems of a SVM. This model can be applied for several aims learning that makes parameter C selection before learning unnecessary (Huang and Chang 2007). Through designing an efficient GA¹-chromosome representation as well as an intelligent crossover operation, ESVM makes the best use of intelligent

¹ Genetic algorithm

Table. 1 Summary statistics of input parameters

Parameters	Minimum	Maximum	Mean	Standard deviation	Coefficient of variation
Aspect	1.5	3.8	2.98	0.34	1.95
Catchment area	17604.3	767530.2	322149	200851.1	0.11
Channel length	1.210036	2777.37	316.357	635.774	2
Channel network base level	-19.83	2553.03	1167.2	798.25	1.6
Convergence index	-0.06	3.06	0.04	0.37	8.75
Cross-sectional curvature	-0.098	0.098	0.00021	0.00071	3.28
Dem	-15.92	2785.85	1287.74	818.24	1.57
Discharge	0.586556	153.5434	27.2948	28.2489	1.03
Drainage density	0.04026	1.734043	0.78928	0.33425	0.42
Flow accumulation	3450.8	35170.25	15905.1	7439.57	2.14
Flow directions	1.68	4.4	3.18	0.49	6.51
General curvature	-0.020	0.020	0.000009	0.0011	-120.56
Longitudinal curvature	-0.12	0.13	-0.00023	0.00040	-1.77
LS factor	0.08	36069	5.72	3.36	1.7
MRRTF ^a	0.01	5.16	0.73	1.49	0.57
MRVBF ^b	0.01	6.6	1.03	1.81	3.89
Plan curvature	-0.01	4.7	0.07	0.57	0.12
Profile curvature	-0.029	0.033	-0.00014	0.00013	-0.98
Relative slope position	0.11	0.72	0.45	0.11	1.93
Slope	0.01	0.51	0.29	0.15	6.23
Strahler order	1.08	1.61	1.47	0.08	19.1
Stream power index	214.64	65887.08	25481.9	18141.8	1.4
Suspended sediment	27.52	295.97	116.6	65.02	1.79
Tangential curvature	-0.033	0.038	0.00012	0.00033	2.72
Topographic wetness index	5.4	11.28	7.05	1.22	5.78
Vertical distance to channel network	1.18	358.9	121.2	62.96	1.46
Watershed area	2.6325	7733.127	638.714	1555.35	2.43
Watershed basins	1.05	5752.8	1851.13	1407.84	1.31

^a Multi-resolution ridge top flatness index^b Multi-resolution index off valley bottom flatness

Bold items indicate the most important geo-morphometric input parameters

GA in optimizing system parameters. The intelligent cross-over operation is contingent upon an orthogonal experimental design applying a divide-and-conquer strategy to deal with intractable optimization problems which consist of many system parameters. Common intelligent GAs are concisely presented while having the strengths of orthogonal experimental design and intelligent crossover and have shown superiority in several studies (Ho et al. 2004). Rapid Miner software (Ver. 5) was used to run all the models mentioned.

Model validation

To evaluate the efficiency of the study models, a leave-one-out (LOO) cross-validation approach was used where one data point is excluded and the included data are utilized to estimate the value at the missing point. The advantage of this method is

that all data are used for evaluation, so extra data collection to provide a separate validation set is not needed and there is good coverage of the study area. However, a disadvantage of this method is that each data point is used several times for estimating other points that are spatially close to it.

To compare the ML techniques used here, the root mean squared error (RMSE) and Pearson's correlation coefficient (r) were computed using Eqs. 7 and 8, respectively.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (o_i - p)^2}{n}} \quad (7)$$

$$r = \frac{\sum_{i=1}^n (p_i - \bar{p})(o_i - \bar{o})}{\sqrt{\sum_{i=1}^n (p_i - \bar{p})^2 \sum_{i=1}^n (o_i - \bar{o})^2}} \quad (8)$$

where o is observed data, p is predicted data, $\bar{p}$ is the mean value of predicted data, $\bar{o}$ is the mean value of observed data, and n is the total number of observations.

Weighting parameters

Not all variables received equal weight in the prediction methods mentioned hereafter. Some variables showed greater correlation with the model output and therefore have a greater impact on predictions. In this study, the SVM algorithm was used for weighting variables (Sani Abade et al., 2014).

Results

Summary statistics

Summary statistics for each geo-morphometric input parameter/variable are given in Table 1. The results show that the range of most variables is wide for the 68 watersheds as they represent very different conditions that accompany a wide range in elevation from a minimum of -15.92 m (Gilan Province) to a maximum of 2785.85 m (Lorestan province). Table 2 also indicates wide variation in the area of the different watersheds studied with the smallest being just 2.63 km² and the largest being 7733 km². The wide range in values for these two key variables, elevation and watershed area, in turn causes a wide range in values of other variables.

To reduce the amount of input data, PC analysis was used. PCA results showed that the dataset was summarized by three components. The first PC explains by far the greatest proportion of variation in the data (99 percent). The second and the third PCs determined the remaining 1% of changes in the data. The PCA results are provided in Table 2.

Modeling

Modeling with discharge parameters as inputs

The results of applying models for the collection of average SSL showed that if the discharge parameter is used to predict

Table. 3 Validation statistics of different prediction methods for estimating average suspended sediment load using discharge

Model	RMSE	r
ANN	64.23	0.11
KNN*	50.45	0.63
LR	64.54	0.01
GPs	52.57	0.58
SVM	63.98	0.14
ESVM	181.4	0.33

*Best model

average SSL, all models show less accuracy than the models that used geo-morphometric parameters. The results for models applied to the dataset with discharge as the input are provided in Table 3.

The results for the SVM method showed that this model performed best with a polynomial kernel function (minimized RMSE=63.98). Also for the ESVM method, the Epachnenikov kernel function had the lowest RMSE=181.40, and for the GP method, the RBF kernel showed the best result. Overall, when discharge was used as the input data for KNN, model h was most accurate at predicting average sediment load with RMSE=50.45 and $r=0.63$.

The outcomes for predicting minimum SSL also indicated that the RBF had the lowest RMSE for the SVM method. For the ESVM method, the polynomial kernel function was most accurate. The results for all methods for predicting minimum SSL using discharge as the inputs are presented in Table 4. Overall, the KNN method was the most accurate at predicting minimum SSL.

Modeling with geo-morphometric parameters as inputs

For the six ML approaches used in this study, the best results for each method are summarized in Table 5, and the correlations between actual and predicted data are shown in Figure 4. The best-performing ANN model had 16 hidden layers, moment of 0.2, and learning rate 0.2 and had a RMSE =44.56

Table. 2 Principal component analysis results

Component	Standard deviation	Proportion of variance	Cumulative variance
PC 1	206885928	0.991	0.991
PC 2	18392165.5	0.008	0.999
PC 3	6739975.35	0.001	1
PC 4	1343642.14	0	1
PC 5	82046.085	0	1

Table. 4 Validation statistics of different prediction methods for estimating minimum suspended sediment load using discharge

Model	RMSE	r
ANN	1.483	0.25
KNN*	0.882	0.64
LR	1.122	0.18
GPs	1.102	0.26
SVM	1.1	0.28
ESVM	16.541	0.5

*Best model

Table. 5 Validation statistics of different prediction methods for estimating average suspended sediment load using the geo-morphometric parameters

Model	RMSE	<i>r</i>
ANN	44.56	0.83
KNN	49.68	0.63
LR	53.52	0.55
GPs*	11.00	0.98
SVM	65.61	0.74
ESVM	24.95	0.95

*Best model

Table. 6 Validation statistics of different prediction methods for estimating minimum suspended sediment load using the geo-morphometric parameters

Model	RMSE	<i>r</i>
ANN	0.87	0.82
KNN	1.20	0.59
LR	0.53	0.93
GPs	0.19	0.99
SVM	0.99	0.82
ESVM*	0.17	0.99

*Best model

(Table 4) for predicting average SSL. For the KNN model, prediction was best when $K=3$ was used.

The results showed that the GP model with a RBF kernel function had better results than other kernel functions. For average SSL with SVM, the Epachnenikov kernel function had the lowest RMSE=65.61 and highest correlation coefficient, $r = 0.74$. In contrast, for predicting average SSL with ESVM, the ANOVA kernel function had the lowest RMSE=24.95 and highest $r=0.95$.

When comparing the performance of the most appropriate models for each method, Table 3 shows that the GP model had the best validation statistics (RMSE=11 and $r=0.98$) for estimating average SSL with geo-morphometric parameters (Table 6).

The relationship between actual and predicted SSL is shown when discharge (Figure 3a) and geo-morphometric parameters (Figure 3b) were used for modeling average SSL using the best models (GP). These graphs show that the correlation between actual and estimated values is far stronger when geo-morphometric parameters are used rather than discharge.

The ESVM model had acceptable validation statistics close to those of the GP model (RMSE=24.95 and $r=0.95$) for predicting average SSL. The same range of model parameters and kernel functions was investigated for each method for predicting minimum SSL as were used for average SSL. The best parameters/kernel functions were generally the same for predicting minimum SSL as for predicting average SSL apart from for SVM and ESVM. The validation statistics for the best models for each method are shown in Table 5, and the correlations between actual and predicted SSL are shown in Figure 5. For forecasting minimum SSL with the SVM model, the Epanechnikov kernel was best with a RMSE=0.99 and $r=0.82$. However, for the ESVM model, the radial and Epachnenikov functions were the best performing with the same RMSEs=0.17 and $r=0.99$. When comparing all methods for predicting minimum SSL, the ESVM model was the best performing with RMSE=0.17 and $r=0.99$ (Table 6). The difference in validation statistics between ESVM and GPs was very low. Figure 4a and b show the relationship between actual and predicted SSL when minimum SSL was predicted using discharge and geo-

Figure 3 Scatter plots showing the relationship between measured and estimated amounts of average SSL produced by the GP: **a** using discharge, **b** using the geo-morphometric parameters

Figure 4 Scatter plots showing the relationship between measured and estimated amounts of minimum SSL produced by the ESVM: **a** using discharge, **b** using the geo-morphometric parameters

morphometric parameters, respectively, and based on modeling using the best model (ESVM).

For predicting both average and minimum SSL, the difference between GPs and ESVM model was very low so it can be concluded that among the methods implemented here, GPs and ESVM would be the best to use.

Weighting parameters

In this study, geo-morphometric parameters were used as auxiliary data for predicting SSL, but it was recognized that not all of these parameters have equal importance in making such predictions. Therefore, each parameter was weighted to give effective parameters. The results of weighting input parameters showed that discharge and drainage density had the highest weight (1) and among the geo-morphometric parameters the highest weightings were for profile curvature (1), LS factor (0.819), longitudinal curvature (0.810), and flow accumulation (0.77) for predicting SSL. Maps of each of these parameters for one studied sub-basin are presented in Figure 5.

The results of applying weighting to minimum SSL showed that stream power index (1) and Strahler order (0.98) and aspect (0.95) were the parameters with the highest weight and maximum impact for predicting minimum SSL. The maps for these parameters are presented in Figure 6 for one sub-basin.

Discussion

In this study, the use of the geo-morphometric parameters for optimizing the estimation of river SSL using different ML models was investigated.

Cross-validation results from using six ML models for estimation of SSL indicated that the behavior of average and minimum SSL was not the same in terms of the accuracy metrics. According to our outcomes, GP was found to be the best method to forecast the average SSL (Table 5). However, the ESVM model was identified as the optimal model for predicting the minimum SSL in the study regions (Table 6). It was also found that integrating these techniques along with the geo-morphometric parameters can markedly increase the accuracy of SSL prediction.

In predicting average values of SSL using the selected models, the correlation coefficient increased from 0.63 using discharge data to about 0.98 using the geo-morphometric properties. However, the prediction error value decreased from 54 to 11 mg L⁻¹.

In studies where only climatic data have been used to predict SSL, less accurate predictions have been reported (Cobaner et al. 2009; Liu et al. 2013; Zounemat-Kermani et al. 2016; Kisi 2012; Rajaei et al. 2011).

Comparing our findings with other researchers' findings can highlight the effective role of geo-morphometric features in increasing the accuracy of SSL predictions.

As a whole, compared to discharge data, the use of the geo-morphometric parameters improved the accuracy of SSL predictions in our study regions by 30%.

Determining the contribution of factors affecting SSL variations is required to better understand the conditions where large SSL occur and to try and prevent adverse consequences of such large SSL (Shi et al. 2017). Our results showed that in addition to discharge and drainage density, some geo-morphometric parameters, especially profile curvature, slope factor, longitudinal curvature, and cumulative flow, had the most influence on average SSL in the study areas. However,

Figure 5 Map of profile curvature, LS factor, longitudinal curvature, and flow accumulation parameters in one of the studied sub-basins

according to the best model for predicting minimum SSL, stream power index and Strahler order and aspect had the highest share in the variations of the minimum SSL across the study regions. According to Sabzevari and Talebi (2019), the curvature shape has more effect on erosion compared to plan shape. Malik and Pal (2020a) reported that slope,

HSI, TSI, and SSI were the main factors controlling channel capacity that can be a function of changes in SSL in the rivers. Also, stream power was shown as the best predictor for soil detachment rate in research done by Wu et al. (2019). These reports are quite consistent with outcomes of the current research.

Figure 6 Maps of stream power index, Strahler order, aspect, and vertical distance to channel network in one of the sub-basins studied

Changes in longitudinal and profile curvature alter potential gradients and influence flow rates and hence transport processes in rivers. The high weighting for LS factor is understandable given that according to Moore and Wilson (1992), it was equal to the sediment transport index and represents erosion and sediment transport processes showing the impact of

erosion on the slope. The flow accumulation map (Figure 3c) shows a numerical value for each cell which indicates the total number of cells that ultimately drain into that cell. The longitudinal curvature parameter indicates whether there is an acceleration or deceleration in stream flow and hence whether erosion or deposition is the dominant processes in a given cell.

Stream power index is a demonstration of the power of erosion surface currents and is calculated using specific surface area and shelves. Aspect represents a line in which the greatest reduction in height takes place (see Table 1 for further definitions of geo-morphometric parameters).

The Strahler order parameter is used to define stream size based on a hierarchy of tributaries. It is clear from the definition of each of the important geo-morphometric parameters that these factors are of key importance in determining patterns of SSL and that the weightings produced are sensible theoretically and likely to produce acceptable results. In the future, such parameters could also include other variables that influence erosion processes and basin sediment load such as vegetation coverage and type as well as land use which could be extracted from satellite imagery. In addition, it would be useful to examine the efficiency of other ML models and algorithms based on artificial intelligence. This study investigates annual SSL, but temporal changes in SSL on a daily basis are very important in river engineering studies. Given the success of using geo-morphometric parameters in predicting average and minimum annual SSL, it is likely that they could be incorporated with weather data into daily or monthly predictions, but this would require a separate study.

Conclusion

In this study, six ML models, namely MLR, ANN, KNN, GPs, SVM, and ESVM, were evaluated to predict the minimum and average SSL in 68 river basins of Gilan and Lorestan provinces, in Iran. For this purpose, the optimal independent features were selected among the 26 geo-morphometric parameters using the PCA method. Lastly, model yield was evaluated based on r and RMSEs. The best models for predicting the average SSL were recognized as GPs, ESVM, ANN, SVM, KNN, and MLR, in order of accuracy. However, ESVM, GPS, MLR, SVM, ANN, and KNN models were, respectively, identified as the best predictive models to forecast the minimum SSL, using geo-morphometric characteristics. Furthermore, using discharge data, KNN was determined as the optimal predictive model for predicting minimum and average SSL. It was also found that discharge, drainage density, and profile curvature were the most important variables for predicting the average SSL. However, stream power index and Strahler order were dominant parameters for predicting minimum SSL in our study regions. These results encourage the future use of GPs and ESVM for predicting SSL at the basin scale. Indeed, this work provides a good reference for evaluating the effects of geo-morphometric parameters on SSL variation using GPs and

ESVM methods. These findings are useful for mitigating the undesirable impacts of SSL in different basins, decreasing soil erosion and increasing sustainable development.

Funding Open Access funding enabled and organized by Projekt DEAL. This study was financially supported by the Ardakan University. The work of Ruhollah Taghizadeh-Mehrjardi was supported by the Alexander von Humboldt Foundation under the grant number: Ref 3.4-1164573-IRN-GFHERMES-P.

Availability of data and material The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy restrictions.

Code availability Not applicable.

Declarations

Conflict of interest The authors declare that they have no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Agarwal A, Mishra S, Ram S, Singh J (2006) Simulation of runoff and sediment yield using artificial neural networks. *Biosystems Eng* 94: 597–613
- Aiken LS, West SG, Pitts SC (2003) Multiple linear regression. In: Schinka JA, Velicer WF (eds) *Handbook of psychology: Research methods in psychology*, Vol, vol 2. John Wiley & Sons Inc., pp 483–507
- Asselman N (2000) Fitting and interpretation of sediment rating curves. *J Hydrol* 234:228–248
- Band SS, Janizadeh S, Chandra Pal S, Saha A, Chakraborty R, Melesse AM, Mosavi A (2020a) Flash flood susceptibility modeling using new approaches of hybrid and ensemble tree-based machine learning algorithms. *Remote Sens* 12:3568
- Band SS, Janizadeh S, Chandra Pal S, Saha A, Chakraborty R, Shokri M, Mosavi A (2020b) Novel ensemble approach of deep learning neural network (DLNN) model and particle swarm optimization (PSO) algorithm for prediction of gully erosion susceptibility. *Sensors* 20: 5609

- Bezak N, Mikoš M, Šraj M (2014) Trivariate frequency analyses of peak discharge, hydrograph volume and suspended sediment concentration data using copulas. *Water Resour Manage* 28:2195–2212
- Choi S-U, Lee J (2015) Assessment of total sediment load in rivers using lateral distribution method. *J Hydro-Environ Res* 9:381–387
- Chowdhuri I, Pal SC, Chakraborty R (2020) Flood susceptibility mapping by ensemble evidential belief function and binomial logistic regression model on river basin of eastern India. *Adv Space Res* 65:1466–1489
- Cimen M (2008) Estimation of daily suspended sediments using support vector machines. *Hydrol Sci J* 53(3):656–666
- Cobaner M, Unal B, Kisi O (2009) Suspended sediment concentration estimation by an adaptive neuro-fuzzy and neural network approaches using hydro-meteorological data. *J Hydrol* 367:52–61
- Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn* 20:273–297
- Das B, Pal SC, Malik S (2018) Assessment of flood hazard in a riverine tract between Damodar and Dwarkeswar River, Hugli District, West Bengal. *Spat Info Res* 26:91–101
- Ebrahimi-Khusfi Z, Taghizadeh-Mehrjardi R, Mirakbari M (2021) Evaluation of machine learning models for predicting the temporal variations of dust storm index in arid regions of Iran. *Atmos Pollut Res* 12:134–147
- Gunn S (1998) Support vector machines for classification and regression. ISIS Technical Report, U of Southampton
- Haddadchi A, Movahedi N, Vahidi E, Omid MH, Dehghani AA (2013) Evaluation of suspended load transport rate using transport formulas and artificial neural network models (case study: Chelchay catchment). *J Hydrodyn Ser B* 25:459–470
- Halbe J, Pahl-Wostl C, Sendzimir J, Adamowski J (2013) Towards adaptive and integrated management paradigms to meet the challenges of water governance. *Water Sci Technol* 67:2651–2660
- Haykin S, Lippmann R (1994) Neural networks, a comprehensive foundation International Int. *J Neural Syst* 5:363–364
- Ho S-Y, Shu L-S, Chen J-H (2004) Intelligent evolutionary algorithms for large parameter optimization problems. *IEEE Trans Evol Comput* 8:522–541
- Huang H-L, Chang F-L (2007) ESVM: Evolutionary support vector machine for automatic feature selection and classification of microarray data. *Biosystems* 90:516–528
- Jayawardena A, Li WK, Xu P (2002) Neighbourhood selection for local modelling and prediction of hydrological time series. *J Hydrol* 258:40–57
- Johnson RA, Wichern DW (2007) Applied multivariate statistical analysis, 6th edn. Prentice Hall, Upper Saddle River
- Kaveh K, Bui MD, Rutschmann P (2017) A comparative study of three different learning algorithms applied to ANFIS for predicting daily suspended sediment concentration. *Int J Sediment Res* 32:340–350
- Kişçi Ö (2010) River suspended sediment concentration modeling using a neural differential evolution approach. *J Hydrol* 389:227–235
- Kisi O (2012) Modeling discharge-suspended sediment relationship using least square support vector machine. *J Hydrol* 456:110–120
- Lafdani EK, Nia AM, Ahmadi A (2013) Daily suspended sediment load prediction using artificial neural networks and support vector machines. *J Hydrol* 478:50–62
- Liu Q-J, Shi Z-H, Fang N-F, Zhu H-D, Ai L (2013) Modeling the daily suspended sediment concentration in a hyperconcentrated river on the Loess Plateau, China, using the wavelet-ANN approach. *Geomorphology* 186:181–190
- Mahmoudzadeh H, Matinfar HR, Taghizadeh-Mehrjardi R, Kerry R (2020) Spatial prediction of soil organic carbon using machine learning techniques in western Iran. *Geoderma Regional* 21:e00260
- Malik S, Pal SC (2020a) Characterization of downstream channel morphology of a monsoon dominated Dwarkeswar River in West Bengal. *J Geol Soc India* 96:539–556
- Malik S, Pal SC (2020b) Downstream decreasing channel capacity of a monsoon-dominated Bengal basin river: a case study of Dwarkeswar River, Eastern India Chin. *Geogr Sci* 1–21
- Malik S, Pal SC, Chowdhuri I, Chakraborty R, Roy P, Das B (2020a) Prediction of highly flood prone areas by GIS based heuristic and statistical model in a monsoon dominated region of Bengal Basin. *Remote Sens Appl Soc Environ* 19:100343
- Malik S, Pal SC, Sattar A, Singh SK, Das B, Chakraborty R, Mohammad P (2020b) Trend of extreme rainfall events using suitable global circulation model to combat the water logging condition in Kolkata Metropolitan Area. *Urban Clim* 32:100599
- Melesse A, Ahmad S, McClain M, Wang X, Lim Y (2011) Suspended sediment load prediction of river systems: an artificial neural network approach Agric. *Water Manag* 98:855–866
- Moore ID, Wilson JP (1992) Length-slope factors for the revised universal soil loss equation: simplified method of estimation. *J Soil Water Conserv* 47:423–428
- Nemes A, Rawls WJ, Pachepsky YA (2006) Use of the nonparametric nearest neighbor approach to estimate soil hydraulic properties. *Soil Sci Soc Am J* 70:327–336
- Nourani V, Andalib G (2015) Daily and monthly suspended sediment load predictions using wavelet based artificial intelligence approaches. *J Mt Sci* 12:85–100
- Rajaei T (2011) Wavelet and ANN combination model for prediction of daily suspended sediment load in rivers. *Sci Total Environ* 409:2917–2928
- Rajaei T, Mirbagheri SA, Zounemat-Kermani M, Nourani V (2009) Daily suspended sediment concentration simulation using ANN and neuro-fuzzy models. *Sci Total Environ* 407:4916–4927
- Rajaei T, Nourani V, Zounemat-Kermani M, Kisi O (2011) River suspended sediment load prediction: application of ANN and wavelet conjunction model. *J Hydrol Eng* 16:613–627
- Rashidi S, Vafakhah M, Lafdani EK, Javadi MR (2016) Evaluating the support vector machine for suspended sediment load forecasting based on gamma test. *Arab J Geosci* 9:1–15
- Roushangar K, Shahnazi S (2020) Determination of influential parameters for prediction of total sediment loads in mountain rivers using kernel-based approaches. *J Mt Sci* 17:480–491
- Sabzevari T, Talebi A (2019) Effect of hillslope topography on soil erosion and sediment yield using USLE model. *Acta Geophys* 67:1587–1597
- Sahana M et al. (2020) Rainfall induced landslide susceptibility mapping using novel hybrid soft computing methods based on multi-layer perceptron neural network classifier. *Geocarto Int* :1–25
- Shi H, Hu C, Wang Y, Liu C, Li H (2017) Analyses of trends and causes for variations in runoff and sediment load of the Yellow River. *Int J Sediment Res* 32:171–179
- Si W, Bao W, Jiang P, Zhao L, Qu S (2017) A semi-physical sediment yield model for estimation of suspended sediment in loess region. *Int J Sediment Res* 32:12–19
- Taghizadeh-Mehrjardi R, Toomanian N, Khavaninzadeh AR, Jafari A, Triantafyllis J (2016) Predicting and mapping of soil particle-size fractions with adaptive neuro-fuzzy inference and ant colony optimization in central Iran. *E J Soil Sci* 67:707–725
- Talebi A, Hajiabolfhasemi R, Hadian MR, Amanian N (2016) Physically based modelling of sheet erosion (detachment and deposition processes) in complex hillslopes. *Hydrol Process* 30:1968–1977

- Vafakhah M (2013) Comparison of cokriging and adaptive neuro-fuzzy inference system models for suspended sediment load forecasting. *Arab J Geosci* 6:3003–3018
- Verstraeten G, Poesen J (2001) Factors controlling sediment yield from small intensively cultivated catchments in a temperate humid climate. *Geomorphology* 40:123–144
- Wang Y-G, Wang SS, Dunlop J (2015) Statistical modelling and power analysis for detecting trends in total suspended sediment loads. *J Hydrol* 520:439–447
- Ward PJ, van Balen RT, Verstraeten G, Renssen H, Vandenberghe J (2009) The impact of land use and climate change on late Holocene and future suspended sediment yield of the Meuse catchment. *Geomorphology* 103:389–400
- Wu B, Wang Z-l, Zhang Q-W, Shen N, Liu J (2019) Response of soil detachment rate by raindrop-affected sediment-laden sheet flow to sediment load and hydraulic parameters within a detachment-limited sheet erosion system on steep slopes on Loess Plateau, China. *Soil Till Res* 185:9–16
- Zhu Y-M, Lu X, Zhou Y (2007) Suspended sediment flux modeling with artificial neural network: an example of the Longchuanjiang River in the Upper Yangtze Catchment, China. *Geomorphology* 84:111–125
- Zounemat-Kermani M, Kişi Ö, Adamowski J, Ramezani-Charmahineh A (2016) Evaluation of data driven models for river suspended sediment concentration modeling. *J Hydrol* 535:457–472