River Bank Erosion Prediction Using Multivariable Linear Regression

Hao Duc Do

FPT University, Ho Chi Minh city, Vietnam

Corresponding author. E-mail: haodd3@fe.edu.vn

Received: Aug. 29, 2023; Accepted: Nov. 28, 2023

This research proposes a new approach using multivariable linear regression to predict the riverbank erosion speed. As a simple and interpretable model, the proposed approach gains two main achievements. First, it can specify the main factors causing riverbank erosion. Notably, the method identifies the river's depth and the water flow's hydraulic gradient, contributing primarily to the erosion speed. Second, multivariable linear regression can be learned from such a small dataset. This aspect makes the range of applications for the method much broader. The Experimental results show that the multivariable linear regression can predict erosion speed well. With a dataset with only 27 records, the method can predict the erosion speed with an error of around 2 meters per year. In the future, a more extensive training dataset or a more complicated regression model is requested to gain a better result.

Keywords: riverbank erosion prediction, multivariable linear regression, machine learning

http://dx.doi.org/10.6180/jase.202412_27(12).0006

1. Introduction

Riverbank erosion, one of the traditional topics in environmental research, is the wearing away of the banks of a stream or river [1, 2]. Although several methods are proposed to predict erosion and balance ecosystems, the problems still need to be solved. Riverbank erosion causes many problems in social, economic, and environmental for the surrounding areas. Riverbank occurs suddenly and cannot be easily measured or predicted effectively by any model.

There are many approaches for estimating the erosion speed for riverbanks. One of the first and widely used methods is proposed by Nanson and Hickin [3]. This method and its contemporary methods [4, 5] follow the same process. First, they collect the data related to the riverbank and analyze it manually. Then, they propose or design a mathematical system to model the impact of natural and artificial factors on erosion. Finally, they apply the designed system to predict erosion in the future.

Although achieving good accuracy, the traditional method takes a long time to design a solution for a new place. Forming a mathematical system for a new river could take years or decades. This paper proposed a new approach to deal with this problem. We designed a new solution using multivariable linear regression, a machine learning technique, to estimate the erosion speed quickly. This method can establish a solution quickly using a small dataset. It can yield a well-designed solution within some minutes without the contribution of any experts. In this research, our main contributions are as follows:

- Demonstrating the potential of machine learning and data science techniques in a new field. We apply a multivariable linear regression to the riverbank erosion prediction and gain a good solution.

- Specifying the key factors contributing most to the erosion speed. After the learning process, the regression model yields that the river's depth and the water flow's hydraulic gradient are the two main contributors to erosion.

- Building a predictor for riverbank erosion prediction. The model can make a good prediction from a small dataset with only 27 samples with an error under 2 meters per year.

The remaining of this paper is structured with 4 main sections. Section 2 mentions some representatives in traditional river bank erosion prediction methods. In this section, we focus on the method presented by Nanson and Hickin [3], which is a landmark in this field. Next, in section 3, the paper presents our proposed method using multivariable linear regression in detail. We present the mathematical formula system, model design, and metrics to evaluate the prediction models. After that, section 4 shows our experimental results for the proposed method compared with the traditional method. Finally, the paper ends with section 5, including a summary and conclusion for our research.

2. Related works

2.1. Some research related to erosion prediction

There are many pieces of research to predict the erosion of the riverbank. Some of them are listed in the Table 1. Most were invented before the 2000s or the era of modern machine learning and deep learning.

Table 1. Some traditional methods for riverbank erosion prediction.

Author(s)	Year of conduct
Keady P.D. [6]	1977
J.M. [4]	1980
J.C. [5]	1982
Nanson and Hickin [3]	1983
Odgaard and Spoljaric [7]	1986
Briaud et al. [8]	2001

The research in Table 1 is similar to the handcraft feature designs in traditional machine learning. This means that the researchers carefully observe the data collected in the real world and then design a complicated mathematical formula system to model the data. Then, they apply that mathematics system to predict the phenomenon for the future. Among these models, the solution suggested by Nanson and Hickin is one of the best models. The following subsection introduces this method briefly.

2.2. Nanson and Hickin method

The method proposed by Nanson and Hickin is used in many real systems, such as Digital Shoreline Analysis System (DSAS) Version 4.0 [9]. This method proposes a multivariable function to model the balance status between erosion speed and the activities of the water flow as follows:

$$M = f(\Omega, S, H, R, B) \tag{1}$$

with:

- M: The estimated speed of erosion using this method.
- Ω : Total energy of the water stream in a 1 -meter band.
- S: Average coefficient of resistance to erosion of the riverbank soil.
- H: Average depth of the water stream R : Bend radius of the riverbank.
- B: Width of the water stream.

While S, H, R, B are identified by sensors or measuring devices, the total energy Ω is computed by:

$$\Omega = \Delta * g * J * Q \tag{2}$$

with Δ, g, J, Q denote the density of water, the gravity acceleration, the hydraulic gradient along the water stream, and the bandwidth of the water stream. Particularly, Nanson and Hickin use two equations below to define the $f(.)$:

$$f(\Omega, S, H, R, B) = 2.5 * K * (R/B)^{-1} \tag{3}$$

for the cases with (R/B)>2.5, and:

$$f(\Omega, S, H, R, B) = (2/3) * K * (R/B - 1) \tag{4}$$

for (R/B)>2.5, with K defined by:

$$K = \Omega/(H * S) \tag{5}$$

2.3. Advantages and disadvantages of the traditional methods

The traditional methods, especially the method of Nanson and Hickin, work well in reality. These approaches mainly depend on a long time observation of researchers to reach the optimal equation. Because this function comes from expert knowledge, created by careful distillation, it keeps and reflects most of the properties of nature.

The limit of these methods is the duration to deploy the solution for a new river. It could take years or decades to propose a good solution for a new river to gather, analyze, and establish a new mathematical model.

3. Multivariable linear regression for prediction problem

3.1. The strength of machine learning technique for multivariable problems

Unlike the traditional approach, which is based on the observation skills of the experts, machine learning techniques use a data set to build the processing model. It could take many years or decades to train a scientist or expert, but the time to train a machine learning model is much faster. On the other hand, training an expert takes much effort, while collecting data can be deployed widely in a short time.

Machine learning models use collected data under a mathematical system, or learning model, to yield a solution for many problems, including prediction, in this research. There are three main strengths of machine learning compared with the traditional methods as follows:

- Easy to do without any expert knowledge. The learning models only need data to learn the rules in nature.
- Fast to train and gain a good solution. After the 2010s, with the development of GPUs, the time to train a machine learning model reduced significantly. It could take weeks or months to train a very deep network and only a few minutes to train a regression model or SVM model.
- Able to be deployed by everyone. Because machine learning does not depend on expert knowledge, everyone who can pass the data into machine learning models can use these techniques for prediction problems.

3.2. Multivariable linear regression

When the input data x distributes in the k-dimension space, the regression problem becomes multivariable linear regression [10–12]. Let $x_i = \left[{x_1}^i, {x_2}^i, {x_3}^i, \ldots, {x_k}^i\right]$ denote a data point in the input space. The regression problem becomes finding the optimal β_i with:

$$f(x_i) \approx \beta_0 + \beta_1 * {x_1}^i + \beta_2 * {x_2}^i + \beta_3 * {x_3}^i + \ldots + \beta_k * x_k^i \tag{6}$$

The β_i is the coefficient of the i^{th} dimension in the data space. Let x_i, β denote the more general forms in a multidimension space as follows: $x_i = \left[1, {x_1}^i, {x_2}^i, {x_3}^i, \ldots, {x_k}^i\right]; \beta = [\beta_0, \beta_1, \beta_2, \quad \beta_3, \ldots, \beta_k]$. The form of the regression model becomes:

$$\beta = \text{argMin}_\beta \int \left(f(x_i) - \langle x_i, \beta\rangle\right)^2 dx \tag{7}$$

with $\langle x_i, \beta\rangle = \beta_0 + \beta_1 * {x_1}^i + \beta_2 * {x_2}^i + \beta_3 * {x_3}^i + \ldots + \beta_k * {x_k}^i$. Let $y_i = f(x_i)$, after stacking x_i, y_i into X, Y, the regression problem becomes an optimization aiming to minimize the objective function defined by:

$$L(\beta) = \|X^*\beta - Y\|^2 \tag{8}$$

Then the objective function $L(\beta)$ is rewritten as follows:

$$L(\beta) = \left(X^*\beta - Y\right)^T \left(X^*\beta - Y\right) \tag{9}$$

with $(.)^T$ is the transposed matrix. After doing the multiplying operator in the equation above, $L(\beta)$ becomes:

$$L(\beta) = \beta^T X^T X\beta - \beta^T X^T Y - Y^T X\beta + Y^T Y \tag{10}$$

Because the loss function in the equation above is convex, the minimum value is identified at gradient zero. The 1-order gradient of $L(\beta)$ is defined by:

$$\begin{aligned} &\delta(L(\beta))/\delta(\beta) = \\ &\delta\left(\beta^T X^T X\beta - \beta^T X^T Y - Y^T X\beta + Y^T Y\right))/\delta(\beta) \end{aligned} \tag{11}$$

or

$$\delta(L(\beta))/\delta(\beta) = -2X^T Y + 2X^T X\beta \tag{12}$$

The gradient zero, or $\delta(L(\beta))/\delta(\beta) = 0$, is equal with:

$$-2X^T Y + 2X^T X\beta = 0 \text{ or } 2X^T Y = 2X^T X\beta \tag{13}$$

Then, we have:

$$\beta = \left(X^T X\right)^{-1} X^T Y \tag{14}$$

The achieved value β is the solution for the optimization problem in this research.

3.3. The proposed framework for riverbank erosion prediction using multivariable linear regression

We separate the dataset $S = (X, Y)$ into training set $\left(X_{\text{training}}, Y_{\text{training}}\right)$, accounting 90% of S and testing set $\left(X_{\text{testing}}, Y_{\text{testing}}\right)$, containing 10% of the dataset. For each data point (x_i, y_i), the internal information is organized as follows:

$$(x_i, y_i) = \left((B_i, H_i, R_i, J_i, Q_i, \Omega_i, S_i), E_i\right) \tag{15}$$

with all notions are described in the subsection 2.2, except E denotes the real erosion speed. The process of learning the regression model is presented in the algorithm 1.

In the prediction problem, three main metrics are used to evaluate the model's performance. Let us denote:

$$(Y, \hat{Y}) = (\text{ output, prediction }) \tag{16}$$

Algorithm 1. Multivariable linear regression for prediction - Training phase

1: procedure TRAINING $(X_{training}, Y_{training})$
2: $X \leftarrow X_{training}$
3: $Y \leftarrow Y_{training}$
4: Compute $\beta = (X^T X)^{-1} X^T Y$
5: Return β

Algorithm 2. Multivariable linear regression for prediction - Testing phase

1: procedure TESTING $(X_{testing}, Y_{testing}, \beta)$
2: $\hat{Y} \leftarrow$ An empty list
3: for $x_i \in X_{testing}$ do
4: $\hat{Y}_i \leftarrow \langle \beta, x_i \rangle$
5: MAE, MSE, RMSE $\leftarrow$ Compute from $(Y_{testing}, \hat{Y})$
6: Return MAE, MSE, RMSE

They are the actual output and the predicted output of the input data X. When testing in a dataset with n samples (x_i, y_i), the three metrics Mean absolute error (MAE), Mean square error (MSE), Root mean square error (RMSE) are defined as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{17}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{18}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{19}$$

Both of the three methods above are similar in the aspect of measuring the difference between the predicted values and the actual values. Each metric differs from the others by the contribution of one data sample to the whole dataset. Because all three measures are error rates, the low result corresponds with a good solution. The testing process is shown in the algorithm 2. This process is a combination of evaluation and testing phases.

4. Experiments

4.1. Dataset and data visualization

We use the dataset from "Do Quang Thien et al. [13] to evaluate our proposed method for riverbank erosion prediction. The erosion information in this dataset was collected at the Gianh-Nhat Le riverbanks in the province of Quang Binh, Vietnam, from 2013 to 2015. This dataset includes 27 records, corresponding with 27 data points. The details are recorded by manual methods combined with a remote sensing approach. Each point has recorded all information related to the notions in the subsection 2.2. This means that each record contains the average coefficient of resistance to erosion of the riverbank soil, the average depth of the water stream, the bend radius of the riverbank, the width of the water stream, the density of water, the hydraulic gradient along the water stream, and the bandwidth of the water stream; and the average erosion for the corresponding position. The distribution of the actual erosion speed is shown in Fig. 1, and the cross-correlation of the aspects is shown in Fig. 2.

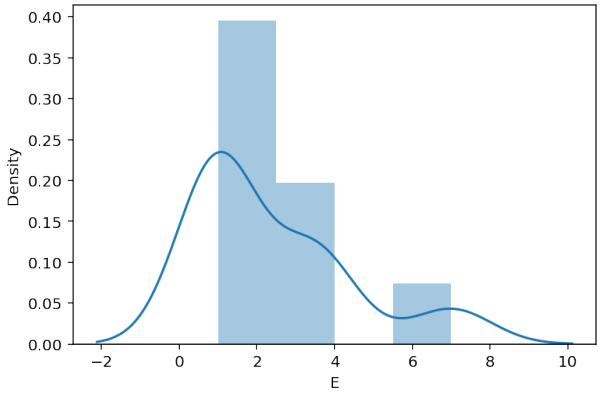

Fig. 1. Distribution of the actual erosion speed. The columns show the dataset's real distribution of the erosion. The curve represents the 1-order derivation of the data to make the distribution more smooth.

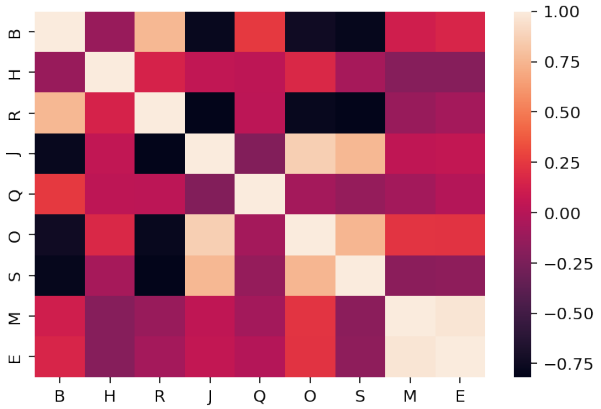

Fig. 2. Cross correlation of many aspects related to the erosion speed.

In Fig. 1, the blue columns reflect the numerical information in the dataset, while the curve is the smooth form (first-order derivative) of the values. The label E in x-axis represents the real speed of erosion (meters/year), and the

y-axis show the corresponding density (%) in the dataset. The actual erosion speed is too difficult to record in the data collection process. Hence, the collectors use numerical rounding techniques to simplify the recorded values and approximate the actual values. This causes some empty positions in the actual value band.

As can be seen, Fig. 2. shows the cross-correlation of the different parameters. The dataset contains nine pieces of information, including seven aspects B, H, R, J, Q, O(Ω), S, actual erosion speed E, and M, resulting from the the Hickin and Johnson method. As can be seen in this figure, there are two crucial things. First, E has close correlations with all seven parameters above. It implies that the proposed method can work and even perform so well. On the other hand, because M has close correlations with these parameters, Hickin and Johnson's method also becomes a prominent opponent of our method.

Experimental results

Table 2 presents the results of the training phase, which are yielded by the algorithm 1 . In this experiment, we split the training set and testing set randomly five times. We use each training set to pass through the training algorithm each time to gain the different coefficient β.

In Table 2, the contributions of different parameters are too different. The output E mainly depends on the contribution of H and J, which are the average depth of the water stream and the hydraulic gradient along the water stream. This is a great result because it specifies precisely the factor contributing most to the riverbank erosion phenomenon.

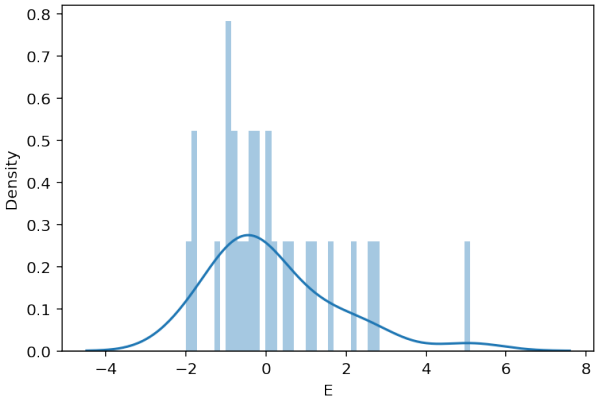

Fig. 3. The difference between actual erosion and predicted erosion speeds models.

Fig. 3 shows the MSE of the proposed method. As can be observed in the diagram, most values concentrate around 0, mainly in [-2, 2]. It means that the predicted erosion speeds are too close to the actual speeds. In the smooth curve, the bell shape, with the center located near 0, also shows the same results.

Table 3 compares our five proposed regression models with the traditional Hickin and Johnson's method. The Hickin and Johnson method results are computed from E and M, while our results are computed via algorithm 2. In this table, the best results in both three metrics come from the Hickin and Johnson method, while our five models gain the following positions.

4.2. Discussion

Via the experiment, we can reach some important observations for applying the machine learning techniques to the erosion prediction problem:

- The multivariable linear regression is an effective and interpretable machine learning technique because it can specify precisely the factor contributing most to the riverbank erosion phenomenon. Notably, via our proposed method, the average depth of the water stream and the hydraulic gradient along the water stream are the key factors causing riverbank erosion.
- After showing the main factors related to erosion, the multivariable linear regression also predicts the erosion speed with deficient error. In terms of MAE, MSE, RMSE, the difference between predicted and actual values is too small. This means that the proposed method can be applied to real applications.
- The results yielded by our method are not as good as the traditional Hickin and Johnson method. This can be easily understood due to the limit of training data. In this research, we train the regression model with a tiny dataset to verify the idea. A bigger dataset or a more complicated regression model is needed for better results.

5. Conclusion

This paper presents a multivariable linear regression approach for riverbank erosion prediction. We have demonstrated the applicability of the machine learning and data science approach to environmental research. After learning from a dataset, our regression method can specify the main factors mainly contributing to the erosion of the riverbank, which are the river's depth and the water flow's hydraulic gradient. Although only using a small dataset with 27 records, experimental results show that the multivariable linear regression can predict the erosion speed so well in terms of MAE, MSE, RMSE, with a low error compared with the ground truth.

In the future, two approaches can be used to gain better performance. First, training the learning models with a

Table 2. Coefficients of five multivariable linear regression models.

Model	B	H	R	J	Q	O	S
LR#1	0.007350	-0.577201	-0.000426	0.369176	-0.001157	0.000710	-0.007316
LR#2	0.003554	-0.480612	-0.000498	0.358429	-0.000662	0.000672	-0.009815
LR#3	0.002054	-0.510923	-0.000665	-0.411171	0.000838	0.000592	-0.010656
LR#4	0.002230	-0.535833	-0.000350	-0.377883	-0.000607	0.000603	-0.008258
LR#5	0.005528	-0.523391	-0.000363	-0.675655	-0.002234	0.000912	-0.008356

Table 3. Comparison of different models in riverbank erosion prediction.

Method	MAE	MSE	RMSE
N&H	0.49	0.32	0.56
LR#1	1.09	1.92	1.39
LR#2	1.06	1.69	1.30
LR#3	1.12	1.78	1.33
LR#4	1.05	1.73	1.32
LR#5	1.16	2.49	1.58

bigger dataset is necessary. In this work, the models' performances could be better than the traditional method because of the small training dataset. The more extensive dataset could improve the predicted results significantly. On the other hand, we can design a more powerful regression model to present more relationships among the parameters. This approach can significantly increase the model's capacity to predict the erosion speed more accurately.

Acknowledgement

Hao D. Do was funded by Vingroup JSC and supported by the Ph.D Scholarship Programme of Vingroup Innovation Foundation (VINIF), Institute of Big Data, code VINIF.2022.TS.037.

References

[1] A. Ghosh, M. B. Roy, and P. K. Roy, (2022) *"Evaluating lateral riverbank erosion with sediment yield through integrated model in lower Gangetic floodplain, India"* **Acta Geophysica *70***: 1769–1795. DOI: 10.1007/s11600-022-00822-7.

[2] B. K. Ghosh, (2022) *"An empirical study of riverbank erosion in Charbhadrasan Upazila of Faridpur District, Bangladesh"* **Urban, Planning and Transport Research *10***: 502–513. DOI: 10.1080/21650020.2022.2123034.

[3] E. J. Hickin and G. C. Nanson, (1984) *"Lateral Migration Rates of River Bends"* **Journal of Hydraulic Engineering *110***: 1557–1567. DOI: 10.1061/(ASCE)0733-9429(1984)110:11(1557).

[4] H. J.M., (1980) *"Magnitude and Distribution of Rates of River Bank Erosion"* **Earth Surface Processes**: 143–157. DOI: 10.1002/esp.3760050205.

[5] B. J.C., (1982) *"Stream Channel Stability Assessment"* **Report No. FHWA/RD-82/021**: 41.

[6] P. M. Keady P.D., (1977) *"The Downstream Migration Rate of River Meandering Patterns"* **12th Mississippi Water Resources Conference**: 29–34.

[7] A. J. Odgaard and A. Spoljaric, (1986) *"Sediment Control by Submerged Vanes"* **Journal of Hydraulic Engineering *112***: 1164–1180. DOI: 10.1061/(ASCE)0733-9429(1986)112:12(1164).

[8] J.-L. Briaud, F. C. K. Ting, H.-C. Chen, Y. Cao, S.-W. Han, and K. Kwak, (2001) *"Erosion function apparatus for scour rate predictions"* **Journal of Geotechnical and Geoenvironmental Engineering *127***: 105–113. DOI: 10.1061/(ASCE)1090-0241(2001)127:2(105).

[9] E. R. Thieler, E. A. Himmelstoss, J. L. Zichichi, and A. Ergul. "The Digital Shoreline Analysis System (DSAS) Version 4.0 - An ArcGIS extension for calculating shoreline change". In: 2009. DOI: 10.3133/ofr20081278.

[10] A. Q, (2020) *"Multiple Linear Regression"* **Principles of Managerial Statistics and Data Science**:

[11] S. K. Gupta and A. P. Agarwal, (2021) *"Predicting Total Sugar Production Using Multivariable Linear Regression"* **2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)**: 465–469. DOI: 10.1109/ICCCIS51004.2021.9397078..

[12] H. G. Perros, (2021) *"Multivariable Linear Regression"* **An Introduction to IoT Analytics**:

[13] D. Q. Thien, T. T. Nhan, N. Q. Tuan, and H. T. Thanh, (2017) *"Experiment with semi-empirical method of Hickin E.J - Nanson G.C to predict the erosion of Gianh - Nhat Le riverbanks in Quang Binh province"* **Journal of Science of Lac Hong University *Special issue***: 60–67.