CHAPTER 5

GENE EXPRESSION PROGRAMMING FOR SCOUR

5.1 Introduction

The scour phenomenon is a natural aspect of the structural changes that occur in rivers. Changes in river structure can be caused by natural decay processes,human activity, or any disturbance inside the river. Any man made or natural disturbance to the hydrodynamic forces and sediment transport capacity contributed to the natural state of the riverine system instability. The emergence of diverse hydraulic structures has significantly altered river regimes in recent years, having a considerable impact on sediment transport capacity and deposition. The dynamic equilibrium of normal free-flowing reverse is sought by striking a compromise between flow conditions and sediment transport capacity. The sediment transport capacity of natural flowing water is at the rate where it is provided during the dynamic equilibrium phase (Vanoni, 1975; Chang, 1988; Alekseevskiy et al., 2008). Understanding river form and natural qualities is, in general, the most difficult phenomena to grasp. Many scientists, researchers, and engineers from all around the world are working to find the best answers to the river ecological and engineering problems that are part of everyday life (Garcia, 2008).

In earlier research investigations, several challenges were met in terms of addressing river bed material and scour variation through time, as well as the equilibrium scour development stage. The study's focus has been on scour phenomena, particularly the physics of the scour process, which could lead to a better understanding of the phenomenon. It's important to look into the shear stress that's exerted at the riverbed-structure interface.

Simple linear regression models or non-linear models based on artificial neural networks are examples of inductive models (ANNs). Inductive models based on (GP) and (GA) have recently been utilised to forecast scour geometry in a variety of disciplines, including maximum scour depth, bridge pier, culvert, and sediment transport. Jothiprakash et al., (2006) applied the GA to Pechiparai reservoir, Tamil Nadu, India, and derived rule curves on the basis of reliability of the GA model. The limits of the more simple regression methods can be linked to the goal of employing these highly complicated, accurate, and non-linear procedures.

As a result, these methods are limited to a single function, such as linear, polynomial, or exponential functions. A simple linear regression model to a complicated non-linear model based on ANNs, GP, and GA are all examples of inductive models. Because of its consistency when applied to non-linear issues, artificial intelligence-based models garner greater attention than classic regression techniques.

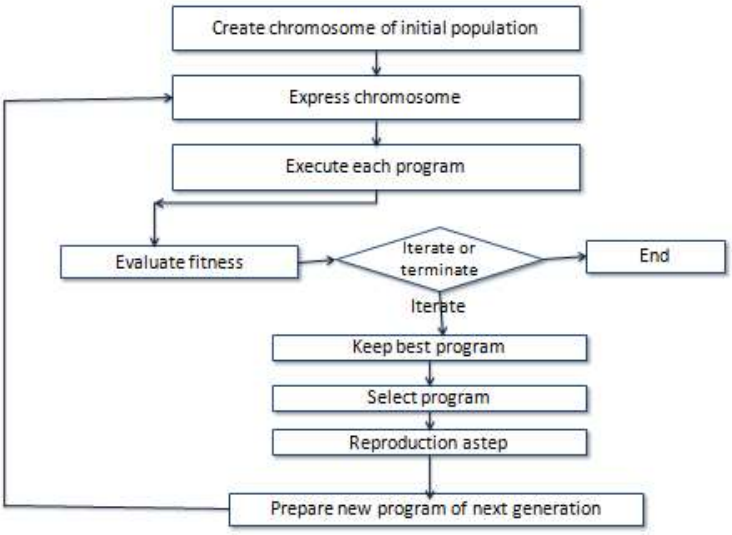

Figure 5.1 Gene expression algorithm flow diagram

The present study is mainly deal with the prediction of maximum scour around the spur dyke using GEP (Gene expression programming) algorithm and the accuracy of this

method validate with pre-existing results Froehlich, (1989); Lim, (1997); Husain, (1998); Dey, and Barbhuiya, (2005) and field data available in the literature. The used mathematical model is shown in Figure 5.1.

5.2 An overview of Gene Expression Programming (GEP)

GEP is a modified type of Genetic Programming (GP) created by Koza (1999), and Ferreira (2001). Inductive models are developed using GEP, which is an evolutionary technique. To produce a closed-form functional approximation for the modeled process by fitting an approximation to the provided data set. GEP can combine the benefits of its predecessors, GA and GP, while also resolving their drawbacks (Duan et al., 2006). Individuals in the initial populations in GA are simple strings (chromosomes) of fixed lengths, whereas sparse trees in GP are non-linear entities of various shapes and sizes. Individuals in GEP are non-linear entities of various shapes and sizes, similar to expression trees in GP, which can be encoded into simple strings of fixed lengths, similar to GA. Any changes in the chromosome during reproduction will always result in a structurally valid programmed, according to the gene expression programming algorithm. This good quality can make GEP preferable to GP in terms of dealing with random numerical constants (Ferreira, 2001b). GEP is also capable of solving Mathematical expressions, neural networks, decision trees, polynomial structures, and logical expressions are all examples of GEP output (Piotr, and Ewa, 2009). GEP is made up of two parts: chromosomes and expression trees. In GEP, there are two languages: one for the gene (chromosomes for initial populations) and another for the expression tree (chromosomes encoded with genetic information). A simple gene in GEP is made up of a set of functions made up of a large set of functions from a library of functions and terminals. The user can choose the amount of functions to be used based on the needs of the task at hand, or even all of the library's functions. The accuracy of the resulting model

is diminished when all of the functions in the problem are selected, making the model more difficult. The following is an example from the gene:

$$Qb[ba(Qa + b)] \tag{5.1}$$

Where Q, *, and + are functions while a, b are the terminals. The ETs for Equation(5.1) is shown in Figure 5.2.

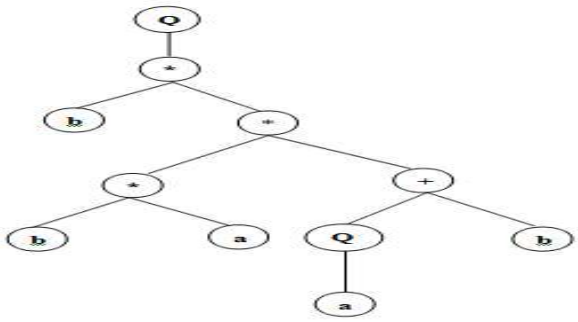

Figure 5.2 Expression tree for the gene shown in Equation (5.1)

The figure shows that ET began with Q, the root of ET, and has one argument, followed by one node in the next line marked with a "*." Because the ET has two arguments, there will be two nodes denoted by "*" and "+" in the next line. This will continue until you have a horizontal line with only terminals. Each gene on a chromosome is divided into two sections: the head contains symbols that are mostly utilised for encoding functions, and the tail contains only terminals that ensure valid expression trees or programmes. The terminals function as intermediary structures and give the ends of chromosomes (Zhang, and Xiao, 2010). The user chooses the head length while the tail length is calculated in Equation (5.3) given by (Ferreira, 2001b).

$$t = h.(n_{max} - 1) + 1 \tag{5.2}$$

Where h represents the head length and n_{max} shows the maximum number of function arrangement.

GEP can solve recent difficult real-world problems that necessitate multi-gene chromosomes. The length of each gene in such chromosomes will be equal to the evolutionary start, and each will result in a sub-expression tree, which will be linked together by some linking mechanism (Ferreira, 2001b). Arithmetic operators such addition, subtraction, multiplication, and division, or BOLENS, are different linking functions in GEP (NOT, OR, AND). These "n" arguments can likewise be used to create links.

They have proven to be troublesome because they might lead to long and complicated expressions. Post-translational interaction is the term for this mechanism. Mutation, cross-over, transposition, and inversion are some of the genetic operators employed in GEP systems for genetic manipulation (Ferreira, 2006). The following are the genetic operators:

Mutation is the most powerful and crucial operator in GEP modelling, and it can occur anywhere in the genome. The chromosomes' structural organisation however, is the same. As a result, all structural individuals produced through mutation are identical.

Inversion: The operator is used for choosing the sequence from the genes and chromosomes.

Transposition of the insertion sequence (IS): This operator involves sorting a piece of the genome that has function or terminals in the first position.

Root insertion sequence (RIS) transposition: Operator is similar to IS transposition with a few differences. It will be moved to the beginning of the gene, replacing the root.

Gene transposition: The complete gene functions as an element that transposes to the chromosome's beginning. Gene transposition differs from other forms of transposition in that the transposed elements are erased at the point of origin.

Single or double cross-over/recombination: After the parent chromosomes have been paired, a single spot is chosen as the cross-over point. The genes downstream of the cross-over location are transferred between the two chromosomes (Guven, and Aytek, 2010).

Gene cross-over: When chromosomes are multi-genetic, this operator is useful. The entire gene is transferred between the two parental chromosomes, resulting in the formation of two new offspring chromosomes that include both parent genes. The genes that are swapped are chosen at random and are located in the same place on the original chromosomes. Because random numerical constants are an important part of a mathematical model and can have a significant impact on the accuracy of the resulting model. The reproduction, in which the individuals were chosen based on their fitness. Following that, several genetic operators such as mutation, inversion, transposition, and recombination are used to create genetic changes in the initial population. Following that, a new population of people is created. This method was repeated until the problems optimal expression was discovered. According to Zhang and Xiao (2010), problem-solving with GEP consists of seven components;

GEP= <Np,Ng,h,Fs,T_s,M,F>,

where N_p = number of individuals at initial population, N_g = number of genes, h is head length, F_s = function set, T_s = set of the terminal, M = range of selection, and F = linking function.

The GEP model begins by selecting one individual chromosome from the original population at random. The chromosome could be a single gene or numerous genes, depending on the complexity of the problem. In GEP, the root means square fitness function is the most popular fitness function.

The fitness values of these people are then recreated using a roulette wheel selection procedure. After spinning the wheel a few times, the populations individuals for the following generation are chosen. In GEP, the likelihood of picking individuals is mostly determined by their fitness value, with the maximum fitness value determining the likelihood of selection. Individuals are exposed to genetic modification and a new generation is created by using various genetic operators. In contrast to other GEPs, mutation, inversion, transposition, and one, two, and three-point cross-over are all genetic operators. The process will be repeated until the best option has been found (Ferreira, 2001b).

Individual chromosomes in GA are simple, and the length of the chromosomes is fixed, whereas chromosomes in GP are non-linear, and the length of the chromosomes is variable in shape and size. The individuals in the GEP's initial populations are a mix of these two.

GA yields the best decision variable values. GP produces lengthy expressions with varying sizes, whereas GEP produces a relatively basic, easy-to-use, and compact representation of the modelled process.

In comparison to GP and GA, GEP is better at dealing with random numerical constants. In GP, genetic operators are applied directly to expression trees, which are the most delicate. Otherwise, an incorrect computer programme may occur. Only two genetic operators, mutation and cross-over, are utilised in GP to account for genetic variation. Point mutation is inefficient and produces faulty programmes multiple times. As a result, the cross-over operator is commonly utilised in GP for problem-solving methodologies. In GA, the same genetic operators are directly applied to individual chromosomes,

whereas in GEP, the number of genetic operators stated above is used to account for genetic variation.

5.3 Non-linear Regression for Maximum scour depth

In the case of maximum scour depth, Pandey et al., (2015) proposed the equilibrium condition at the maximum scour depth.

$$\left(d_{se}/_{l}\right) = 5.686 F_Z^{0.276}\, y^{0.248}\, l^{0.163} \qquad (5.3)$$

Where d_{se}= equilibrium scours depth, Fr = Froude number, and it depends on approach velocity, Fe = abutment Froude number, y = depth of approaching flow, l = transverse length of spur Dyke.

Pandey et al., (2015) show the relationship given by Dey, and Barbhuiya, (2005) the experiments with eighty percent data within twenty-five percent error when having data used from other publications.

Currently, soft computing techniques are successfully applied to hydraulics engineering (Samadi et al., 2014, 2015; Samadi et al., 2020a,b). This paper includes GEP for obtaining a new relationship in the field of spur Dyke.

5.4 Prediction of scour depth using GEP model

For accuracy in the measurement of depth in the paragraph, the scour depth d_{se} is modeling's using the GEP programming technique. The "training set" can choose from the data, and the rest is used as the "test set" except this. When we chose the training set of systems, it was defined as the system learning environment. The next part of the model (test set) has 5 key processes in preparation for use in GEP. In starting, fitness function can be selected first. In the current experiment f_i, consider as a fitness program for an individual; hence f_i is represented as

$$f_i = \sum_{j=1}^{C_t} \left(M - \left| C_{(i,j)} - T_j \right| \right) \tag{5.4}$$

For particular above equation: M, represents the range of selection, and $C_{(i,j)}$ gives the value returned by the chromosome, the subscript s_i for its fitness case j, T_j= target values (for fitness case j). If $|C_{(i,j)} - T_j|$ (precision) is ≥ 0.01, at this condition, precision = 0 , and f_i = f_{max} = C_t, M. for particular case used as, M = 100, therefore, f_{max} = 1000. Here the advantage is that the optimal solution of the problem finds itself because of this kind of fitness function. Next, a set of terminal state functions T& F is selected to form chromosomes. The only independent variable i.e., $T = \{h\}$. Choosing a suitable function is not very easy, except that a good approximation that includes all the functions can be helpful. This paper includes 4 common arithmetic operators (+, -, *, /) and the widespread mathematical function (√).

The third important step is choosing the chromosomal architecture, and it involves the length & number of genes. Using single-gene and two lengths heads accordingly rose in their numbered genes and heads, one by one in each round. It observes the training and test performed by every model in the experiment. It seems that if the gene counts increase greater than two and the length of the head greater than eight, there is not much more rise in training & testing performance of the gene expression programming model in the experiment. Hence for better performance, used two genes in each chromosome and the head length is 8 for every GEP model in the study. Linking of the functions is the fourth major step. Next, we are trying for the addition & multiplication of the link function. This is obtained that gives a good fitness value of linking sub-trees Equation (5.4). Finally, choosing the genetic operators set causes the difference in its rate was the fifth major step of this problem-solving approach. For this programming, all common genetic operators were used in Table 5.1.

Table 5.1 Parameters for optimizing the GEP model.

Parameters	Description of parameters	Setting of parameters
P_1	Function set	+, -, *, /, √
P_2	Mutation rate %	30
P_3	Inversion rate %	30
P_4	One point and two point recombination rate respectively %	30,30
P_5	Gene recombination rate	95
P_6	Gene transportation rate	0.1

In the calibration process for GEP, we used 120 input-target pairs of data collected in Table 5.2

Table 5.2 Data ranges (Pandey et al. 2015)

Parameters	**Present study**	**Lim (1997)**	**Husain et al. (1998)**	**Dey and Barbhuiya (2005)**	**Nasrollahi et al. (2008)**	**Coleman et al (2003)**
l(m)	0.06-0.2	0.05-0.15	0.4-0.8	0.06-0.12	0.25	0.05-4.75
V(m/s)	0.159-0.21	0.24-0.325	0.1991	0.219-0.67	0.3125	0.21-0.443
y (m)	0.1	0.1-0.15	0.22	0.58-0.2	0.128	0.05-0.53
d_{50}(mm)	0.27	0.94	0.775	0.26-1.866	1.3	0.8-1.02
B(m)	1	0.6	1.37	0.9	2	-

Among 220, only 55 (25%) data sets are reserved for validation, and 165 data sets are used for calibration purposes and then for the validation of genetic programming validation. The best generation individual, chromosomes 30, has fitness 680.5 for d_{se}. Clear formulations of GEP for d_{se} the value is:

$$d_{se} = \frac{y}{-57.177Fr + 6.717} + \frac{y}{FrL^{-1} + \sigma + 2Fr + 6.547} + 5.666L\sigma^2 yFr + L^2\sigma^2 yFr + 5.668Fr^2 L + L^2\sigma^2 Fr^2 + 1 \quad (5.5)$$

5.4.1 The Error Measures

In present study, the following error criteria have been used to assess the computational derivations from observed data.

1. Correlation coefficient (R);

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} \tag{5.6}$$

$$AE = \frac{\sum \frac{X - Y}{X} * 100}{n} \tag{5.7}$$

3. Mean Square Error (MSE):

$$RMSE = \left[\frac{\sum (X - Y)^2}{n} \right] \tag{5.8}$$

4. Average absolute deviation (δ):

$$\delta = \frac{\sum |(Y - X)|}{\sum X} * 100 \tag{5.9}$$

5.5 Development of ANN model

Artificial neural networks (ANNs algorithms) can give an irregular mapping between input and an output vector, usually consisting of three layers of neurons, viz., hidden input and output (Samadi et al., 2021). Every neuron acts non-dependent computational element. The neural network can find its strength with a high degree of freedom related to its architecture. Before the field experiment, the network is trained for observing the data sets. In general, the feed-forward kind of network has been trained using trainlm. Out of 215 input-output pairs, approximately 75% of 161 sets, choose irregularly are used for training, while the remaining 25% of 54 sets are employed for the testing procedure.

The equation also was tested for sensitivity analysis using testing (validation) data sets and shown in Table 5.3. The table shows that the coefficient of determination, R^2 of the proposed equation, was moderately strong with R^2 = 0.934. Table5.3 compared to the ANN model; one independent parameter can be removed in each case. A larger RMSE and lower R^2 value can be obtained by removing an independent parameter from the input set. These parameters can affect the $d_{se,}$ and the functional relationship is given in Eq. 5.1

can be used for GEP modeling. The resulting GEP approach has a high non-linear relationship with d_{se}. GEP model provides highly reliable data with minimum error. So the proposed GEP model has a good testing performance and the highest generalization capacity with R^2=0.88, MAE=0.015, and MSE=0.000424.

Table 5.3 Sensitivity Analysis (GEP) of Independent Parameters for the Testing

Model	MSE	RMSE	MAE	R^2
$d_{se} = f(d_{50}, F_r, l, y, \sigma)$	0.006	0.077	0.56	0.934
$d_{se} = f(d_{50}, F_r, l, y)$	0.0979	0.313	0.878	0.856
$d_{se} = f(d_{50}, F_r, l)$	0.0954	0.308	0.945	0.789
$d_{se} = f(d_{50}, F_r)$	0.119	0.345	0.825	0.756
$d_{se} = f(d_{50})$	0.38		0.91	0.78

5.6 Result and discussion of GEP

Comparisons of present GEP model with sediment transport Equation (5.3) by (Pandey et al., 2015). We concluded that particularly in laboratory measurement, GEP models prediction gives more accurate prediction than pre-existing models. It is seen that the result of the model gives standard data and shows very little error for tested data $R^2 = 0.94$, $MAE = 0.06566$, $MSE = 0.0056$ shown in Figure 5.3.

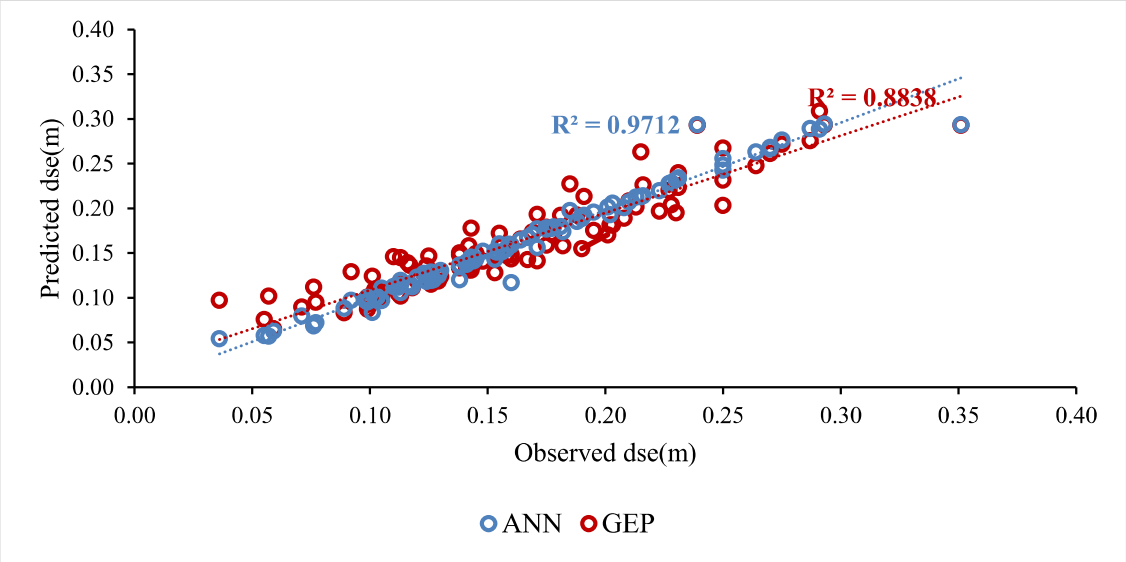

Figure 5.3 Observed versus predicted relative scour depth

The most important advantage of the GEP model in comparison with previously reported regular reverse analysis conventional equations is that GEP can map data to a higher dimensional feature space. In this study, various methods were used to predict the relationships in data. So the relation found from the mapping of the data is quite simple. In this study ANN with (R^2=0.97, MAE=0.005, RMSE=0.010) was more accurate than GEP with (R^2=0.88, MAE=0.016, RMSE=0.021). GEP provided an explicit equation for prediction d_{se}. GEP found a direct relationship between input and out parameters. In contrast, ANN had higher accuracy; however, it was considered as a black box model for prediction d_{es}.

5.7 Conclusions

The detection of local scours depth of a dyke we used relatively very general to understand the new soft computing technique approach, GEP. GEP & ANNs models help for predicting the relative scour depth value from laboratory measurement. A new technical approach to estimating the equilibrium depth scours of spur dyke with optimum data sets and genetic programming, artificial neural network logarithm, or model soft

technique. These programming models also help in predicting depth estimation methods for pipes. We concluded that the new soft computing technique gives a more precise and accurate depth scour in rivers than previous depth elevations. The overall result shows that the ANNs model is superior to the GEP model. The high rise of coefficient of determination ($R^2 = 0.88$) proves that the GEP model has shown a perfect fit for the data set measured. The result was shown that the GEP model was useful for practitioners. However, compared to ANN, the significant advantage of GEP is to provide an explicit equation for the prediction of scour depth.