Hindawi
Discrete Dynamics in Nature and Society
Volume 2021, Article ID 3594271, 12 pages
https://doi.org/10.1155/2021/3594271

Research Article

A New Dual-Mode GEP Prediction Algorithm Based on Irregularity and Similar Period

Lei Yang,[1,2] **Zexin Xu**,[1] **Rui Xu**,[1] **Jianfan Lu**,[1] **Zhenlin Xu**,[2] **and Kangshun Li**[1]

[1]*College of Mathematics and Informatics, South China Agricultural University, Guangzhou 510642, China*
[2]*Guangdong Provincial Key Laboratory of Food Quality and Safety, South China Agricultural University, Guangzhou 510642, China*

Correspondence should be addressed to Lei Yang; yanglei_s@126.com

Received 26 July 2021; Accepted 25 August 2021; Published 16 September 2021

Academic Editor: Shi Cheng

Gene expression programming (GEP) uses simple linear coding to solve complex modeling problems. However, the performance is limited by the effectiveness of the selected method of evaluating population individuals, the breadth and depth of the search domain for the solution, and the ability of accuracy of correcting the solution based on historical data. Therefore, a new dual-mode GEP prediction algorithm based on irregularity and similar period is proposed. It takes measures to specialize origin data to reserve the elite individuals, reevaluate the target individuals, and process data and solutions via the similar period mode, which avoids the tendency to get stuck in local optimum and the complexity of the precisions of correcting complex modeling problems due to insufficiency scope of the search domain, and subsequently, better convergence results are obtained. If we take the leek price and the sunspot observation data as the sample to compare the new algorithm with the GEP simulation test, the results indicate that the new algorithm possesses more powerful exploration ability and higher precision. Under the same accuracy requirements, the new algorithm can find the individual faster. Additionally, the conclusion can be drawn that the performance of new algorithm is better on the condition that we take another set of sunspot observations as samples, combining the ARIMA algorithm and BP neural network prediction algorithm for simulation and comparison with the new algorithm.

1. Introduction

In the field of predictive modeling, there are many models. Many scholars conduct in-depth research in this area. Wang et al. [1–3] have made great efforts in the optimization of prediction algorithms and achieved certain results. This paper is mainly to study and improve the GEP model and use ARIMA and BP-ANN models for experimental comparison.

Time series prediction is a typical method in data mining, which is widely used in the fields of financial economy, meteorology, hydrology, signal processing, and disaster warning. The autoregressive integral moving average model (ARIMA) is the most common model used for time series forecasting. Atanu et al. [4] used the ARIMA model to predict a country's GDP, and Thiruchelvam et al. [5] used the ARIMA model to determine the spatial effect of dengue fever cases on neighboring areas. The error back-propagation (BP) algorithm is also a current method used for time series forecasting. Li et al. [6] predicted the passive torque of the human shoulder joint based on BP-ANN (Artificial Neural Network), and Kianpour et al. [7] used BP-ANN to predict the acute oral toxicity of organophosphates. ARIMA is essentially a linear prediction model and requires data to be stable. And, there exist many anthropogenic factors in the training of the BP-ANN prediction model. GEP inherits the advantages of the simple linear coding of genetic algorithms (GAs) and the expression tree of genetic programming (GP) and expresses the population from both the genotype and phenotype [8]. Because of the simplicity, comprehensibility, and high efficiency of GEP, it has been widely utilized in data prediction. Oulapour et al. [9] used GEP to find the best equation for the relationship between the width and depth of the possible crack area and the

geometric parameters of the valley cross section. Khan et al. [10] used GEP to predict the compressive strength of geopolymer concrete. Yang et al. [11] proposed a new spectral model for leaf area index estimation based on GEP. Mallick et al. [12] used GEP to evaluate the surface average pressure coefficient of the building surface. Ali et al. [13] used GEP to predict the air ratio parameters of mineral tailings. Deng [14] et al. used hybrid GEP to recognize numerical sensitive data in active distribution networks. Majidifard et al. [15] used GEP to develop a prediction model of asphalt mixture derusting depth. Murad [16] et al. used GEP to predict the shear strength of internal reinforced concrete beam-to-column joints subjected to cyclic loading.

GEP is based on the principle of survival of the fittest in biological evolution and the unique mechanism which can decode the results into functions making it extraordinary in the family of prediction algorithms. But there exist many shortcomings, such as tendencies to fall into the local optimum [17], the slowness of later convergence, and the complicacy of accuracy correction of nonlinear complex problem solutions [18, 19]. In this regard, Zhang et al. [20] used regularization methods to enhance the generalization ability of GEP, increase gene diversity, and jump out of local optimality. Jiang et al. [21] accelerated GEP through measures such as adaptive parameters, population age stratification, and transplantation of the Spark framework. Wang et al. [22] introduced a multipreference-driven coevolutionary algorithm in GEP to improve the quality of the target solution, while reducing the complexity of the algorithm. However, because the above improvements require the additional constraint information or the integration of other algorithms, they do not have high generalization. This paper proposes a new dual-mode GEP (DM_GEP) prediction algorithm based on irregularity and similar period. The data processing objects of the first mode are those that are irregular. The data processing objects of the second mode are those with similar periodic fluctuations. In general, if the data object has similar periodic fluctuation law, we use the second mode. Otherwise, use another mode. We compared the results with the basic GEP algorithm. The experimental results prove that DM_GEP has a wider and deeper search area and convergence efficiency, and thus, it can achieve higher prediction accuracy. In the experiment, DM_GEP is compared with the ARIMA model and BP neural network prediction model. The experimental results further verify that DM_GEP has better prediction performance.

2. Methods

2.1. Gene Expression Programming. GEP is a new type of adaptive evolution algorithm proposed by Portuguese scientist Candida Ferreira in 2001. GEP inherits the rapidity and usability of GAs and the variability and versatility of GP. It can utilize simple coding to solve complex problems [23]. Meanwhile, the separation of genotype and phenotype makes the evolutionary efficiency of GEP to solve practical problems' 2–4 orders of magnitude higher than GA and GP [24].

2.2. Chromosome. In the process of gene expression programming modeling, a random initial population is first generated, and the population is composed of chromosomes. The processing object of GEP is a chromosome composed of a single gene or multiple genes. Genes consist of linear and fixed-length strings of symbols, which can be divided into heads and tails. The chromosomes generated according to a certain rule can be decoded according to the rule to generate an expression tree. The expression tree can be further transformed into mathematical expressions. So, the essence of chromosomes is a series of mathematical expressions. If F is the set of function symbols and T is the set of terminal symbols, the heads of the genes can be randomly composed of any symbols in F and T, and the tails of genes can only be composed of any symbols in T. If we let the length of a gene, the length of its head and the length of its tail be L, H, and T, respectively, and the maximum number of operations of the function in the function symbols contained in the gene is N. The following formulas are established:

$$T = H \times (N - 1) + 1, \tag{1}$$

$$L = H + T = N \times H + 1. \tag{2}$$

Figure 1 shows a double-gene chromosome with head length of 4 and tail length of 5.

The chromosome in Figure 1 has two open reading frames (ORF), which correspond to the subtree (sub-ET) of Figure 2. In the multilevel structure tree, each subexpression tree is not only an independent evolutionary individual but also a part of the hierarchical evolution system.

Figure 2 is the expression tree generated by decoding corresponding to Figure 1, which is connected by "+."

2.3. Fitness. Fitness is an index to evaluate the ability of an individual to adapt to the environment. The solution step of fitness needs to decode the chromosome to get the corresponding expression tree and then generate the corresponding mathematical expression. Finally, the value of the objective function is obtained by substituting the value of the variable into the mathematical expression. The smaller the gap between the objective function value and the actual value, the higher fitness that the individual has. According to the individual's fitness value, the quality of the individual in evolution can be evaluated. There are two classic evaluation models in GEP, absolute error (equation (3)), and relative error (equation (4)):

$$\text{Fitness} = \sum_{i=1}^{n} \left(M - \left| Y_i - \widehat{Y}_i \right| \right), \tag{3}$$

$$\text{Fitness} = \sum_{i=1}^{n} \left(M - \left| \frac{Y_i - \widehat{Y}_i}{Y_i} \right| \right), \tag{4}$$

where Y is the training dataset which contains n data needed for modeling, Y_i represents the input of the ith group of training data, $\widehat{Y}_i$ represents the predicted value of the corresponding ith group of data, and M is a constant representing the selection range.

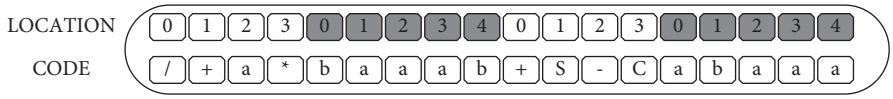

Figure 1: Randomly generated dual-gene chromosomes.

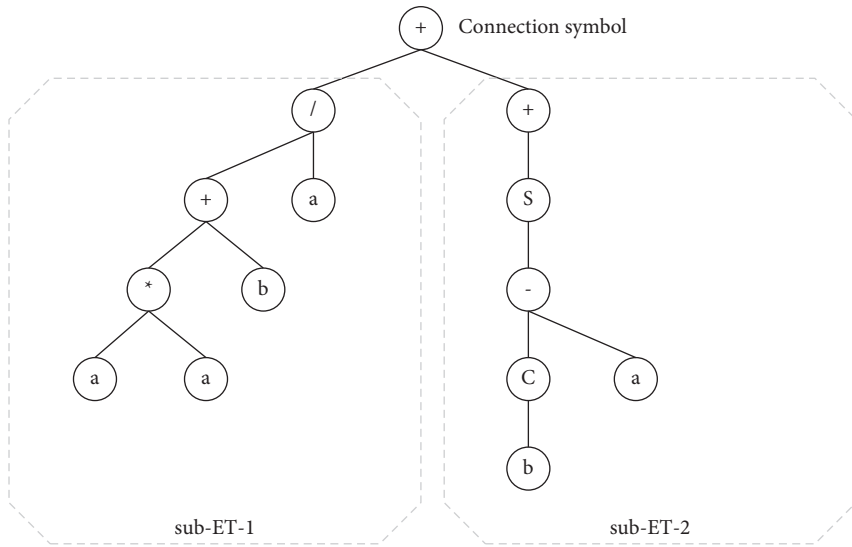

Figure 2: Corresponding expression tree generated by chromosome decoding in Figure 1.

3. A Dual-Mode GEP Prediction Algorithm

If the observation data used for modeling has obvious similar waveform trends that do not strictly limit the height of peaks and troughs in intervals with similar spans, we define that it has a similar period (SP), and this type of data is nonlinear and often contains high ambient white noise. The basic GEP adopts different individual evaluation standards for different types of input data, which does not have generalization. However, if a unified evaluation system is used to evaluate individuals from different data types, the situation that the individual fitness value is high but the regression fitness keeps low is easy to occur. In case of the search area of the algorithm is small, it would make the algorithm fall into the local optimum. When the individual fitness value is approaching the theoretical value, the time consuming is not proportional to the improved accuracy, which leads to the slow convergence rate of the algorithm in the later stage. When processing data with SP characteristics, the accuracy correction is difficult and the process is complicated. The DM_GEP proposed in this paper is based on the basic GEP algorithm and consists of the irregular prediction mode (IPM) and similar period mode (SPM). Among them, IPM can be used for regression prediction problems of various types of data. SPM only processes SP data without complicated and difficult processes, so it has an improved accuracy compared with IPM.

3.1. General Mode-IPM. On the basis of GEP evolution, a unified evaluation system is applied to deal with different types of modeling data, while expanding the search domain and accelerating the evolution efficiency. IPM is described as follows, and the pseudocode of the algorithm is shown in Algorithm 1.

① Appropriately lower the preset solution fitness value to reduce the huge amount of time it takes to approach the value in the later stage of model evolution. Reduce the number of individuals who obtain high fitness but the regression fitting effect does not meet the needs, and accelerate the acquisition of the target solution.

② When evaluating the chromosome, the strict parameter values are not used as the denominator for a few parameter values approaching the zero point, and these parameter values are specially treated to reduce the impact on the individual in calculating the error value of prediction effect.

③ If the individual has reached the preset fitness value, the target individual obtained by the model is re-evaluated. Use modeling data to calculate the average error of the individual for regression. If the average error value is less than the preset limit value, it will be given as the model result. Otherwise, restart the entire model to re-evolve and combine with ① and ② to accelerate the evolution efficiency under the premise of ensuring the accuracy of the target individual's prediction.

The calculation of BEST-FIT complies with the special calculation principle in ② above. K is the optimal number of individuals retained to the next generation, and the value depends on the population size requirements. MAX_FIT is the ideal fitness.

```
Begin:
(1)   CHROMOSME [ ] ⟵ produce stochastic original population;
(2)   While (True)
(3)     While (True)
(4)       BEST ⟵ find the best one from CHROMOSME [ ];
(5)       BEST_FIT ⟵ calculate fitness of BEST;
(6)       If (BEST_FIT ≥ MAX_FIT) then
(7)           AVE_ERROR ⟵ calculate error of BEST;
(8)         If (AVE_ERROR ≤ LIMIT) then
(9)             Return BSET;
(10)        End if;
(11)        Else Break; //restart evolutionary model
(12)      End if;
(13)      SONS [ ] ⟵ produce empty population same as CHROMOSME [ ];
(14)      SONS [ ] ⟵ the top K best individuals; //retain dominant individuals
(15)      SONS [ ] ⟵ Roulette and Genetic manipulation of CHROMOSME [ ];
(16)      CHROMOSME [ ] ⟵ SONS [ ]; //new population
(17)    End While;
(18)  End While;
End.
```

Algorithm 1: IPM.

3.2. Dedicated Mode-SPM. IPM can process SP data, but it is more difficult. The SPM proposed in this paper is based on the larger search domain of IPM and effective convergence in the later stage. Because it is aimed at the SP data processing model and using of compound individuals as solutions, there is no complicated and difficult accuracy correction process, and the convergence efficiency and accuracy are further improved on the basis of IPM.

3.2.1. Original Data Processing Model

① The number of SP data in a group is P, and the average value L of its "period" is obtained, and G is N times L (P is rounded down), that is, there are N "periods" in this group of data. Among them, G, L, and N are a positive integer.

② The sliding size of the SPM window is W, and set $W = L$, that is, the coding parameter of a chromosome is W dimension.

③ A set of modeling data consists of two parts: continuous L points form W-dimensional calculation parameters, and the $(L+1)$th data is used as the correction value. For example, the first set of modeling data $Y_1 = \{\{D_1, D_2, D_3 \dots D_L\}, \{D_{(L+1)}\}\}$. The first $(G-L)$ pieces of data constitute the modeling data, and the remaining L pieces of data are reserved for observing the effect of simulation prediction.

④ From the modeling data group, a group of $(N-2)$ continuous modeling data with each data interval of L is selected as the target child chromosome SP modeling data group. For example, the first target child chromosome modeling dataset Datas $[1] = \{Y_1, Y_{(L+1)}, Y_{(2L+1)} \dots Y_{[(N-3)*L+1]}\}$, and so on, and the modeling data set in ③ is assigned to each target child chromosome for evolutionary modeling.

3.2.2. Compound Individual Solutions. SPM uses compound individuals as the modeling result, that is, an array containing multiple target child chromosomes is used as a solution to solve related problems. Suppose the composite target Com_Chromosme $[W]$ is obtained, that is, there are W target child chromosomes. According to the Datas $[1 \dots W]$ constructed in 3.2.1, Com_Chromosme $[1 \dots W]$ is obtained as the modeling data input into the IPM submodel in sequence. For the Nth cycle data $\{D_{(N-1)*L}, D_{(N-1)*L+1} \dots D_{(N*L)}\}$ reserved for simulation prediction, it is, respectively, obtained by Com_Chromosme $[1 \dots W]$. In the same way, predict future data, and calculate the remainder of L through the historical data coordinate value of the point to obtain the remainder i. Select Com_Chromosme $[i]$, and then, predict the point to get theoretical data.

3.2.3. Model Schematic. The ideal periodic function trend chart is convenient to describe the basic principle of SPM, as shown in Figure 3, and the pseudocode of the algorithm is shown in Algorithm 2 below.

where L is the length of the SP period, W is the sliding size of the SPM window, and P is the number of SP data in a group.

4. Results

4.1. Evaluation Standard. In this paper, the following criteria are used to evaluate the algorithm model.

4.1.1. MSE and MAPE. The MSE is the average value of the sum of squares of errors during the fitting process of the linear regression model, and the value range is $[0, +\infty)$. The closer the value approaches to 0, the better the data obtained from the model fits the original data:

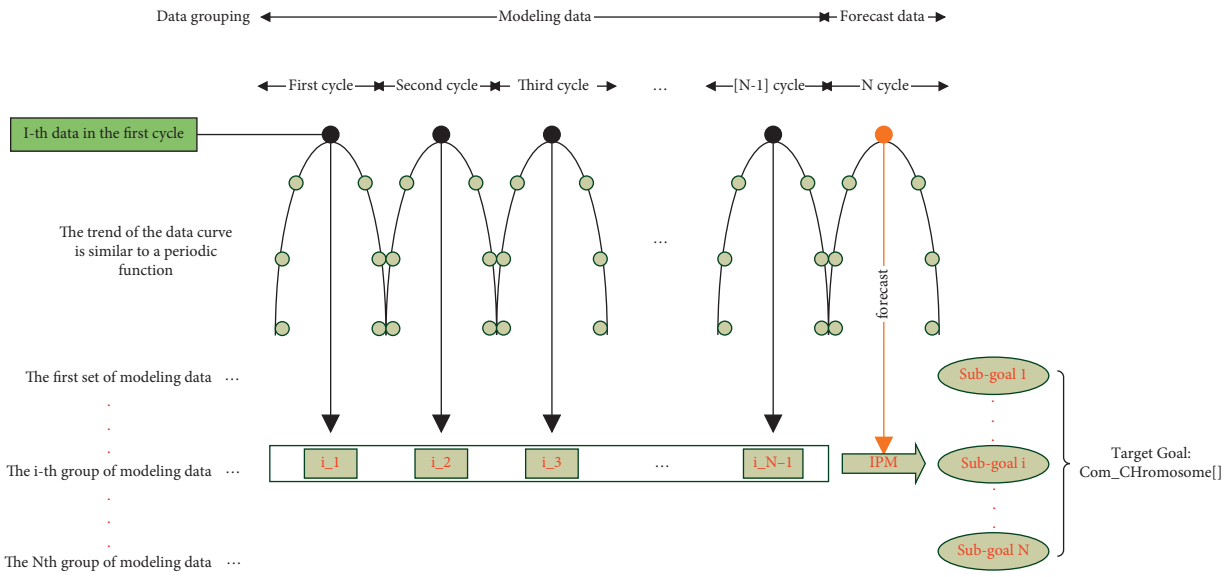

Figure 3: Principle of SPM.

Begin:
Step1:Data preprocessing
(1) **For** $i = 1$ to $G\text{-}L$ **do**//base array
(2) $Y\_datas[i] \leftarrow \{\{D_i \ldots D_{i+L-1}\}, \{D_{L+i+1}\}\}$;
(3) **End for**;
(4) **For** $j = 1$ to W **do**//subgoal modeling and forecast sets
(5) Datas $[j] \leftarrow Y\_datas\ [j] \ldots Y\_datas\ [j + (N\text{-}3) * L]$;
(6) **End for**;
(7) Com_Chromosme $[W] \leftarrow$ structure container for complex subgoals;
Step2:genetic evolution process
(8) **For** $t = 1$ to W do
(9) CHROMOSME [ ] $\leftarrow$ stochastic original population using Datas $[t]$;
(10) BEST $\leftarrow$ Using IPM algorithm get the best individual;
(11) Com_Chromosme $[t] \leftarrow$ BEST
(12) **End for**;
(13) **Return** Com_Chromosme $[W]$ as result;
Step3:use compound target
(14) **For** $x = 1$ **to** P **do**
(15) FORECAST gets using Com_Chromosme $[x \bmod L]$;
(16) SUM_ERROR += gets calculate the error;
(17) **End for**;
(18) AVE_ERROR gets SUM_ERROR/P;
End;

Algorithm 2: SPM.

$$\mathrm{MSE} = \frac{1}{m}\sum_{i=1}^{m}\left(\widehat{Y}_i - Y_i\right)^2, \tag{5}$$

$$\mathrm{MAPE} = \frac{100\%}{m}\sum_{i=1}^{m}\left(\left|\frac{\widehat{Y}_i - Y_i}{\widehat{Y}_i}\right|\right). \tag{6}$$

where m refers to m samples, i refers to the I dimension of quantity Y, $\widehat{Y}_i$ refers to the original value, and Y_i refers to the predictive value.

The MAPE is the mean absolute percentage error, and the value range of MAPE is $[0, +\infty)$. The closer the value approaches to 0, the better the data obtained from the model fits the original data:

4.1.2. Coefficient of Determination R^2. In some cases, MSE is not comprehensive and unable to describe the goodness of fit of the model accurately. If R^2 is used in combination, the performance of the model can be better explained, and the value range of R^2 is $[-\infty, 1]$. The value of R^2 is divided into the following situations:

① $R^2 = 1$: ideal model, and the predicted value is equal to the true value.

② $R^2 = 0$: one possibility is that all predicted values are equal to the average value of the sample. Of course, there are other possibilities.

③ $R^2 < 0$: predictive ability of the model keeps weak, which means that the wrong model may be used or the model assumptions are unreasonable.

$$R^2 = 1 - \frac{\sum_{i=1}^{m} \left(Y_i - \widehat{Y}_i\right)^2}{\sum_{i=1}^{m} \left(Y_i - \overline{Y}_i\right)^2}, \tag{7}$$

where Y is the sample average.

4.1.3. Residual Diagram. The residual diagram is to visually evaluate the performance of the model and obtain outliers by drawing the difference or vertical distance between the true value and the predicted value. For a good regression model, the expected error is randomly distributed, and the residuals are also randomly distributed near the center line.

4.2. Experimental Data. Data group 1 is the daily average price of leeks in the Jiangnan agricultural and sideline product market in Guangzhou City, Guangdong Province from January 1, 2020 to April 20, 2020, obtained from the national agricultural product price database, with a total of 108 price data. This piece of data can represent a normal time series and is used to test the performance of the model.

Data group 2 is taken from NCEI Sun-Geophysics in Space Weather [25]. The sunspot detection data values from 1770–1869, a total of 100. And, its trend is shown in Figure 4. The observational value sequence of sunspots has the characteristics of nonlinearity and multiple time scales [26]. At the same time, due to the large environmental interference and large noise when observing and collecting values, it has become a classic use case for testing the effectiveness and prediction accuracy of predictive model analysis to solve complex real-world problems.

Data group 3 selects the sunspot data from 1919 to 2018 [25]. Because the sunspot time series observation is a typical example of detection and prediction model, it is used to test the performance of several typical algorithms.

For details of the data, please refer to the two documents in the supplementary materials (available here).

4.3. Experimental Setup. In order to highlight the advantages of DM_GEP over GEP, the simulation and comparison experiment parameter settings are simplified to the greatest extent: the function set only selects the most basic four arithmetic operations, and the chromosome structure and length are simplified. At the same time, the average value of multiple experiments is obtained as the conclusion. The same parameter settings of GEP and DM_GEP are shown in Table 1.

BP neural network is a multilayer feedforward neural network trained according to the error backpropagation algorithm [27]. The gradient descent method is used to

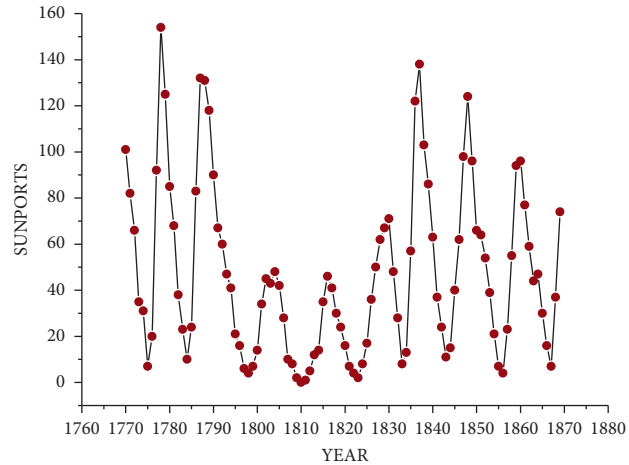

Figure 4: Trend of sunspot observations from 1770 to 1869.

minimize the mean square deviation between the actual output value and the expected output value of the target network. The trained target network can intelligently process the input information of similar input samples and then output the information obtained from the linear transformation with the current minimum error. The ARIMA model has three heaviest parameters. The parameters p, q, and d represent the autoregressive parameter, moving average, and order of transforming the original sequence into a stationary sequence [28]. This paper compares BP, ARIMA, and DM_GEP through a set of time series simulation experiments to verify the advantages of DM_GEP algorithm performance. Table 2 shows the parameter settings of ARIMA and BP-ANN as the comparison algorithm. Among them, the selection of the number of hidden nodes in the BP neural network refers to the empirical formula that scholars have obtained for a long time:

$$h = \sqrt{m + n} + a, \tag{8}$$

where h refers to the number of hidden layer nodes, m refers to the number of input layer nodes, n refers to the number of output layer nodes, a is a constant and its range is [1, 10]. In Table 2, logis and purelin refer activation functions.

4.4. GEP and IPM. For the comparison between GEP and IPM, the data selected in this paper is data group 1, in which the first 103 data constituted 98 sets of modeling data, and the last 5 data will be retained for simulation prediction.

The two target chromosomal individuals representing the average performance of GEP and IPM are as follows:

① GEP-/d+/+deecdbad [+] bac/e * ebdecbcda [+] b-cbc/dedcbccdd

② IPM a/e-c * /eeeddbdd [+]-e + e + cabbaeebba [+] cda * ec-cdabadbe

The regression prediction effect of leek price data is shown in Figure 5. The trend of the IPM target individual regression prediction curve basically completely fits the sample data, and each inflection point changes sensitively

Table 1: GEP and DM_GEP comparison test parameter settings.

Parameter setting	GEP/IPM (leek price)	GEP/IPM (sunspot)	SPM (sunspot)
Head length	7		
Population capacity	50	100	
Select range	1000		
Number of genes	3		
Connection function	{+}		
Symbol set	{+, −, *, /}		
Terminal symbols	{a, b, c, d, e}	{a, b, c, d, e, f, g, h, i, j}	
Mutation rate	0.044		
Single-point mutation rate	0.3		
Double-point mutation rate	0.3		
Gene recombination rate	0.1		
Insertion transposition rate	0.1		
Root insert transposition rate	0.1		
(ROOT) Length of that insert seat	{1, 2, 3}		
Best fitness value	97900/97800	79100/79000	7900
Number of experiments	1000	50	

Table 2: BP neural network and ARIMA experimental parameter settings.

	BP-ANN	ARIMA
Input layer nodes	10	—
Output layer nodes	1	—
Hidden nodes	15	—
Maximum number of training	5000	—
E-learning efficiency	0.025	—
Target network expected error	0.2	—
Excitation function	Hidden layer: logis Output layer: purelin	—
p	—	10
q	—	9
d	—	2

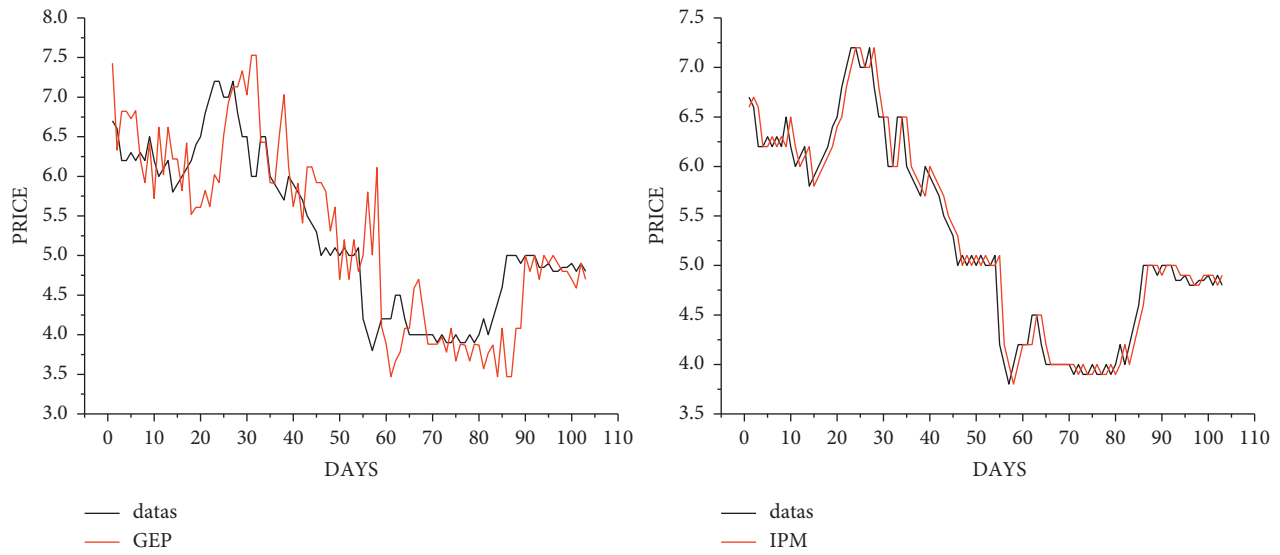

Figure 5: GEP-IPM regression prediction fitting effect of leek price.

with the sample data, and there is no significant mutation point that deviates from the sample data curve. Although the GEP target individual can better reflect the fluctuation of the sample data in the curve trend, it obviously has a certain delay, and many data points deviate greatly from the sample data curve. The residual mean value of the sample data of the GEP target and IPM target is plotted as a residual diagram, as shown in Figure 6. And, the conclusion mentioned above

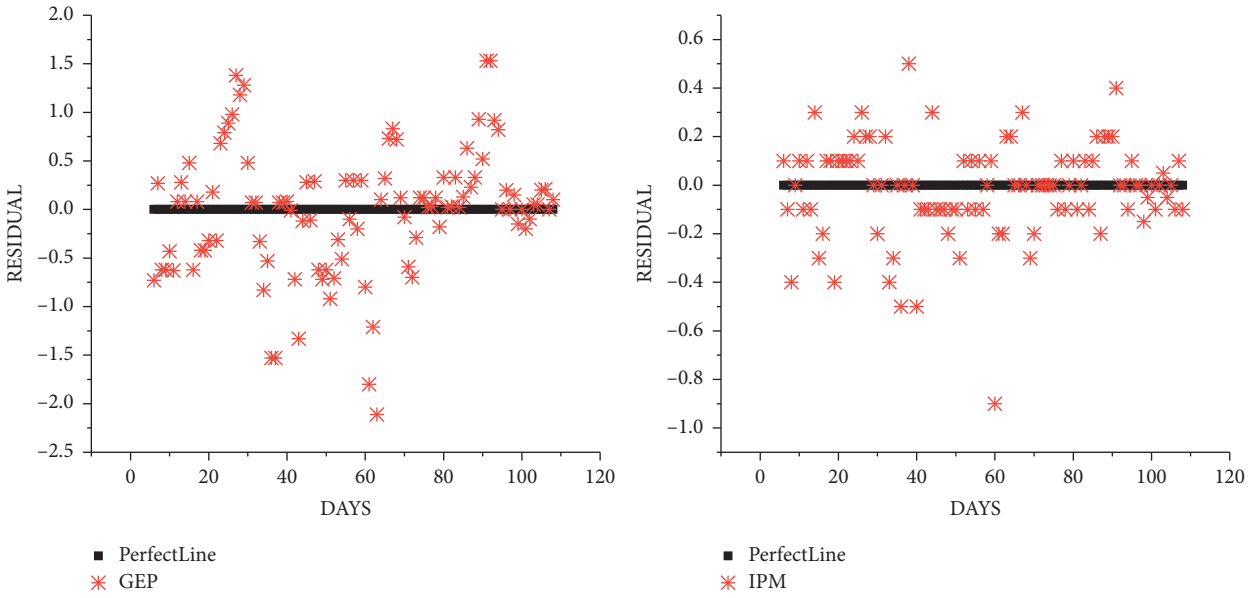

Figure 6: GEP-IPM prediction result residual of leek price.

can be demonstrated. The IPM target individual residual data points are concentrated and uniformly and randomly distributed on the upper and lower sides of the $Y = 0$ line. The number of perfect fit points with an absolute value of 0 for residuals reached 24, accounting for 23.3%. Among them, data points between 0 and 0.2 accounted for 84.5%. The data points between 0 0.4 accounted for 96.2%, and they have accounted for most of the data points. And, the proportion between 0 and 0.6 is 99.1%. The prediction data of the GEP target is scattered and unevenly distributed on the upper and lower sides of $Y = 0$. The mutation points far away from the center line account for a relatively large number. There are only 5 perfect fitting points, accounting for 3.9%, which is between 0 and 0.2. The proportion of data points between 0 and 0.4 is 38.9%, 56.4% is between 0 and 0.4, which is barely more than half, and the proportion between 0 and 0.6 is 65.1%.

Table 3 shows the average conclusion of the experimental data. It can be seen that, under the same or even better evolution time, the MSE and calculation error of the IPM target is still better than the GEP target. After the IPM model has increased the calculation accuracy by 10%, the time-consuming is still close to that of GEP, achieving an MSE close to 0 under a large R^2. It is proved that IPM can break out of the defects of local optimality and precocity better than GEP and find better target individuals. Setting aside time constraints, the IPM target prediction error is only 2.71%, and the MSE and R^2 are 0.2 and 0.96, respectively. The prediction model has basically fitted the actual data. This indicates that IPM has a better ability to explore a wider and deeper search field and to approach the limits with high accuracy than GEP.

4.5. GEP, IPM, and SPM. In this paper, data group 2 is selected as the dataset to compare GEP, IPM, and SPM. The data trend chart is shown in Figure 3. It can be seen that it is composed of several waveforms with large peak fluctuations, similar waveforms, and an average time span of 10, which has obvious SP characteristics and is suitable for detecting SPM performance. Among the 100 years of data, the first 90 years of data are used to train the model, and the remaining 10 years of data are used to simulate predictions.

The target chromosomal individuals of GEP, IPM, and SPM are as follows:

① GEP: /e-ejhibhijeegj [+]-jijbb + ebieejge [+] jhihhbhbhijeebd

② IPM: j-/fe * fiicjeifj [+]-/+a+e-biigdhij [+] eb *-/bfbiicgeje

Table 3: Comparison experiment data of leek price GEP-IPM.

	Expected accuracy	Average time (ms)	MAPE	MSE mean	R^2 mean
GEP	—	5.58	15.23	8.325	0.7273
	0.15	3.39	9.64	0.4948	0.7251
IPM	0.05	7.89	4.1	0.2959	0.9142
	0.03	48.09	2.71	0.2002	0.96

③ SPM:

Com_Chromosme[1]	-i+i-i-ahaicjij	[+]	/e+fbj*jacdjgab	[+]	+j/fcacjacdjgab
Com_Chromosme[2]	g+g/i-hecbgbhia	[+]	/j/*+gejagbaiea	[+]	-/faigdjedeiiea
Com_Chromosme[3]	/dfdfdadjeagbdj	[+]	/cjgfa/cdedbiab	[+]	-jce/g-gbbdffia
Com_Chromosme[4]	-/ifeccaagdcabd	[+]	jjiej*jfecdddcd	[+]	jabcb*ieedchjad
Com_Chromosme[5]	jhfhcb-fgjiicig	[+]	/gja-chabjieahf	[+]	/d-bc/bdfeggfhe
Com_Chromosme[6]	–j+fc/ihfagaig	[+]	/cbe+gbaeegfabb	[+]	/jcji/fiabccbac
Com_Chromosme[7]	-i-i-ijgfddajhj	[+]	jh+fifibfghajfc	[+]	–i-fcadiijgigg
Com_Chromosme[8]	-jij-/cajjcchih	[+]	/h-a+-+aejachih	[+]	-j/cccjabiegjcf
Com_Chromosme[9]	-/ajb/gbgjaedfg	[+]	-bj///jadfjijee	[+]	/cj/ec/idfgajhi
Com_Chromosme[10]	jhi-g-jehjbcaah	[+]	/g/e/h-ggbfhjji	[+]	/b-ggedagchhcgh

Figure 7 shows the forecast data of the set of sunspot data with the above targets, and it can be seen that all three can closely follow the trend of the sample data. Although the GEP target reflects the fluctuation trend of the sample data, the troughs are not well-fitted and the prediction curves are abrupt and jagged in many places. On the contrary, the SPM target has the best effect. The entire prediction curve only appears once with a large abrupt change. Figure 8 shows the residual mean value of these three models, which shows that the distribution of GEP residual points is scattered. The mean error values of each data point of the prediction results of the three models are made into Figure 9 to support the above discussion.

The average value of the experimental data conclusions is made in Table 4, and the performance gap between GEP and DM_GEP can be analyzed. Under the same experimental conditions, the performance of SPM and IPM is close. With only 31% and 25% of the time consumption of GEP, an accuracy of more than 30% higher than GEP was achieved. At the same time, MSE improved by more than 700 and R^2 improved by more than 15% of the target. With the knowledge that the DM_GEP dual mode is better than GEP, analyze the advantages of SPM compared to IPM in SP data; when the expected accuracy is 0.5, SPM is 56.28 higher than IPM on MSE, and the rest is close. When the expected accuracy is 0.35, the experiment shows that this is the performance bottleneck of the IPM experiment. At the same time, SPM can achieve a target of 5.68% increase in accuracy of IPM, 80.41 increase in MSE, and slightly better R^2 than IPM with only 17% of the time consumed by IPM. It shows that SPM has higher evolution efficiency and better target exploration ability than IPM when processing SP data to obtain higher precision targets and has a higher performance threshold. Exploring the SPM performance threshold, the experiment shows that the threshold is 0.2, and the average achieved accuracy is 17.73%.

In summary, SPM is more efficient than IPM and GEP in processing SP data to obtain individuals with uniform residual distribution, larger R^2 values, lower MSE values, and higher accuracy thresholds.

4.6. ARIMA, BP-ANN, and DM_GEP. In this paper, data group 3 is selected as the dataset to compare ARIMA, BP-ANN, and DM_GEP. Among them, the DM_GEP experiment parameter settings are the same as Table 1, the budget accuracy of the change is set to 0.2, and the expected fitness is set to 7850.

The sliding window of BP neural network and DM_GEP is 10, and the last 10 data of three models are reserved as simulation prediction data. Figure 10 shows the experimental fitting effects of the three algorithms. Table 5 shows the statistical analysis of the experimental results.

It can be seen from Table 5 that the mean absolute percentage error (MAPE), MSE, and R of the DM_GEP model are better than the BP neural network and the ARIMA model in the prediction. It can be seen from Figure 10 that the regression prediction curve of the DM_GEP model is basically fitted to the modeling sample data curve, and the MAPE is 16.27. The prediction trend of the last ten sets of prediction data conforms to the future development trend, and the effect is the best.

Some scholars have also made up the shortcomings of ARIMA and BP neural networks: Min et al. [29] used SVM

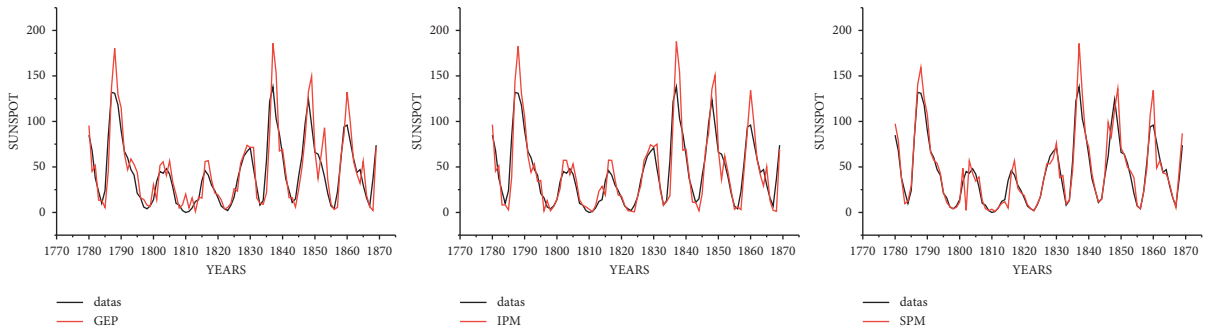

Figure 7: GEP-IPM-SPM regression prediction fitting effect.

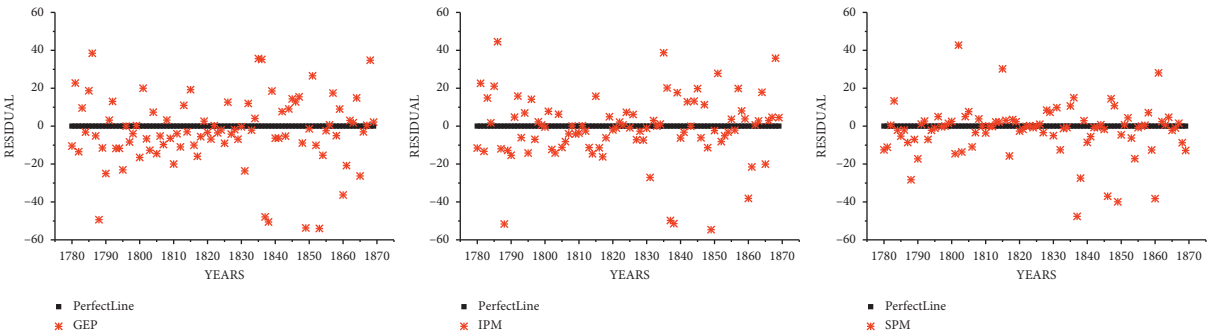

Figure 8: GEP-IPM-SPM regression prediction residual.

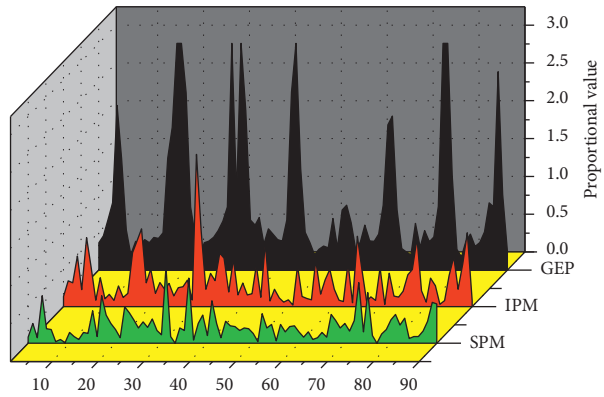

Figure 9: GEP-IPM-SPM mean error.

Table 4: Sunspot GEP-IPM-SPM comparative test data conclusion.

	Expected accuracy	Average time (ms)	MAPE	MSE mean	R^2 mean
GEP	—	8278	75.66	1003.21	0.6515
IPM	0.5	2595	38.16	316.91	0.8294
	0.35	63933	34.37	298.85	0.8436
SPM	0.5	2030	39.39	260.63	0.8032
	0.35	10442	28.69	218.44	0.8512
	0.25	232161	22.91	185.96	0.862
	0.2	1870369	17.73	157.72	0.8971

to map data to a high-dimensional space to try to weaken the interference caused by nonlinearity. He et al. [30] used BP neural networks to pass the PSO algorithm that optimizes the weights of each connection layer accordingly. But in contrast, DM_GEP has excellent linear and nonlinear analysis and modeling capabilities, which not only requires no special requirements for the amount of historical data but also enables excellent accuracy of regression prediction for

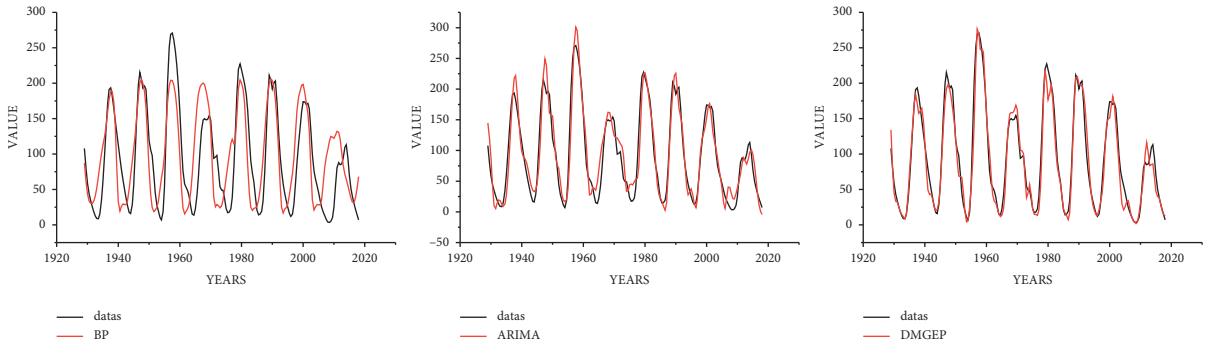

Figure 10: DM_GEP, ARIMA, and BP neural network simulation comparison experiment fitting effect.

Table 5: Sunspot GEP-IPM-SPM comparative test data conclusion.

	Expected accuracy	MAPE	MSE mean	R^2 mean
DM_GEP	0.2	16.27	270.459	0.9443
BP-ANN	0.2	160	2436.504	0.39
ARIMA	—	46.69	503.012	0.898

modeling data that causes large environmental noise without noise reduction. These characteristics make DM_GEP have more prospects and better forecast applicability in the field of forecasting.

5. Conclusion

In the experiment of leek price prediction, the experimental results show that the IPM mode can find better individuals in a shorter time than ordinary GEP. In the experiment of predicting the observed value of sunspots, the SPM mode has higher accuracy and shorter time than ordinary GEP and IPM mode. In addition, the results of experiments with ARIMA and BP-ANN in the prediction of sunspot observations also show that the accuracy of SPM is higher.

GEP has unique advantages in the family of prediction algorithms. However, there are shortcomings such as tendencies to fall into local optimum and difficulties to regress complex nonlinear data. In this regard, a new DM_GEP prediction algorithm was proposed in this paper, which is compatible with the high efficiency of GEP's regression of simple linear problems and excellent nonlinear data analysis and construction capabilities. For the sake of avoiding overly premature models, the algorithm expands the algorithm search space by reducing the rigor of error judgments for those true values close to 0 and the precomputation of individuals. By changing the single mode of modeling data and using specific methods, the complicated and difficult correction process for SP data was avoided. At the same time, it improved the deficiencies of GEP, simplified the threshold of GEP prediction application, and enhanced practicality and generalization. In the experiment of leek price prediction, the experimental results show that IPM can find better individuals in a shorter time than ordinary GEP.

In this thesis, DM_GEP only used simple four arithmetic operations as a function set and simple structure chromosome to solve the problem in order to compare the experimental effect. The next research direction is to add a rich set of functions and diverse connection functions and to study the more general data preprocessing method to DM_GEP so that the new GEP algorithm will be more convenient and general, and the accuracy will be better.

Data Availability

The data used to support the findings of this study are included within the Supplementary Information file.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported in part by the Natural Science Foundation of Guangdong Province of China (Grant no. 2020A1515010691), Science and Technology Project of Guangdong Province of China (Grant no. 2018A0124), Guangdong Provincial Key Laboratory of Food Quality and Safety (Grant no. 2020B1212060059), and National Natural Science Foundation of China (Grant nos. 61573157 and 61703170).

Supplementary Materials

The price of leek in Jiangnan agricultural and sideline products market in Guangzhou City, Guangdong Province from January 1, 2020 to April 20, 2020 are recorded in 1.txt. The sunspot detection data values from 1770 to 2018 are recorded in 2.txt. (*Supplementary Materials*)

References

[1] F. Wang, Y. Li, A. Zhou, and K. Tang, "An estimation of distribution algorithm for mixed-variable Newsvendor problems," *IEEE Transactions on Evolutionary Computation*, vol. 24, no. 3, pp. 479–493, 2020.

[2] F. Wang, H. Zhang, and A. Zhou, "A particle swarm optimization algorithm for mixed-variable optimiza-tion problems," *Swarm & Evolutionary Computation*, vol. 60, Article ID 100808, 2021.

[3] F. Wang, Y. Li, F. Liao, and H. Yan, "An ensemble learning based prediction strategy for dynamic multi-objective optimization," *Applied Soft Computing*, vol. 96, Article ID 106592, 2020.

[4] E. Y. Atanu, H. E. Ette, K. Nwuju, and W. Chimee Nwaoha, "ARIMA Model for gross domestic product (GDP): evidence from Nigeria," *Archives of Current Research International*, vol. 20, no. 7, pp. 49–61, 2020.

[5] L. Thiruchelvam, S. C. Dass, V. S. Asirvadam, H. Daud, and B. S. Gill, "Determine neighboring region spatial effect on dengue cases using ensemble ARIMA models," *Scientific Reports*, vol. 11, no. 1, p. 5873, 2021.

[6] S. Li, P. Dario, and Z. Song, "Prediction of passive torque on human shoulder joint based on BPANN," *Applied Bionics and Biomechanics*, vol. 202010 pages, 2020.

[7] M. Kianpour, E. Mohammadinasab, and T. M. Isfahani, "Prediction of oral acute toxicity of organophosphates using QSAR methods," *Current Computer-Aided Drug Design*, vol. 17, no. 1, pp. 38–56, 2021.

[8] Y. Peng, C. Yuan, X. Qin, J. Huang, and Y. Shi, "An improved gene expression programming approach for symbolic regression problems," *Neurocomputing*, vol. 137, no. 1, pp. 293–301, 2014.

[9] M. Oulapour, A. Adib, and S. Gholamzadeh, "GEP prediction of the cracking zones in earthfill dams," *Arabian Journal of Geosciences*, vol. 14, no. 7, pp. 1–11, 2021.

[10] M. A. Ali Khan, A. Zafar, A. Akbar, M. F. Javed, and A. Mosavi, "Application of gene expression programming (GEP) for the prediction of compressive strength of geopolymer concrete," *Materials*, vol. 14, no. 5, p. 1106, 2021.

[11] L. Yang, S. Deng, and Z. Zhang, "New spectral model for estimating leaf area index based on gene expression programming," *Computers & Electrical Engineering*, vol. 83, p. 106604, 2020.

[12] M. Mallick, A. Mohanta, A. Kumar, and K. C. Patra, "Gene-expression programming for the assessment of surface mean pressure coefficient on building surfacefficient on building surfaces," *Building Simulation*, vol. 13, no. 2, pp. 401–418, 2020.

[13] X. Li, Z. He, and M. Shahidi, "Prediction model for connected voids ratio of the porous asphalt mixture," *Advances in Civil Engineering*, vol. 2020, no. 2, 11 pages, 2020.

[14] S. Deng, X. Xie, C. Yuan, L. Yang, and X. Wu, "Numerical sensitive data recognition based on hybrid gene expression programming for active distribution networks," *Applied Soft Computing*, vol. 91, p. 106213, 2020.

[15] H. Majidifard, B. Jahangiri, P. Rath, L. U. Contreras, W. G. Buttlar, and A. H. Alavi, "Developing a prediction model for rutting depth of asphalt mixtures using gene expression programming," *Construction and Building Materials*, vol. 267, p. 120543, 2020.

[16] Y. Murad, R. Hunifat, and W. Al-Bodour, "Interior reinforced concrete beam-to-column joints Subjected to cyclic loading: shear strength Prediction using gene expression programming," *Case Studies in Construction Materials*, vol. 13, no. 432, pp. 1–9, 2020.

[17] Y. Peng, C. Yuan, X. Mai, and X. Qin, "A summary of theoretical research on gene expression programming," *Ji Suan Ji Ying Yong Yan Jiu*, vol. 28, no. 2, pp. 413–419+43, 2011, in Chinese.

[18] W. Li, H. Li, X. Guo, W. Yang, and S. Ma, "Analysis and trend forecast of sunspot activity period," *Shui Li Shui Dian Ji Shu*, vol. 50, no. 5, pp. 53–62, 2019, in Chinese.

[19] W. Cui, W. Wang, Z. Huang, and Q. Tan, "Predictive model of higher order ordinary differential equation based on GEP algorithm," *Ji Suang Ji Gong Cheng Yu Ying Yiong*, vol. 54, no. 18, pp. 256–262, 2018, in Chinese.

[20] X. Zhang, L. He, Z. Huang, and Q. Tan, "Predictive model of higher order ordinary differential equation based on MFR-GEP," *Ji Suang Ji Gong Cheng Yu Ying Yiong*, vol. 55, no. 21, pp. 247–253, 2019, in Chinese.

[21] Z. Jiang and G. Wang, "Improvement of gene expression programming algorithm," *Ji Suang Ji Gong Cheng Yu She Ji*, vol. 38, no. 12, pp. 3298–3305, 2017, in Chinese.

[22] L. Wang, J. Peng, F. Qiu, and L. Mo, "High-dimensional target evolution algorithm based on multi-preference adaptive cooperation," *Xiao Xing Wei Xing Ji Suang Ji Xi Tong*, vol. 37, no. 6, pp. 1308–1312, 2016, in Chinese.

[23] C. Ferreira, "Gene expression programming: a new adaptive algorithm for solving problems," *Complex Systems*, vol. 13, no. 2, pp. 87–129, 2001.

[24] C. Ferreira, *Gene Expression Programming*, First edition, 2002.

[25] https://www.ngdc.noaa.gov/stp/solar/ssndata.html NOAA > NESDIS > NCEI (formerly NGDC) > STP > Space Weather.

[26] J. Tang and X. Liu, "Analysis of the multiple time scales and chaotic characteristics of the relative number of sunspots," *Zhongguo Ke Xue Wu Li Xue Li Xue Tian Wen Xue*, vol. 48, no. 2, pp. 103–110, 2018, in Chinese.

[27] H. Ding, W. Wang, L. Wan, and J. Luo, "Research on power grid material demand forecast based on BP neural network," *Ji Suang Ji Ji Shu Yu Fa Zhan*, vol. 29, no. 6, pp. 138–142, 2019, in Chinese.

[28] F. Dong, L. Shi, and N. Wu, "Dynamic evaluation of wind power performance Based on DEA-TOPSIS-time series," *Dian Li Ke Xue Yu Gong Cheng*, vol. 34, no. 11, pp. 20–29, 2018, in Chinese.

[29] Y. Min, D. Jian, and W. Wei, "Hybrid dwell time prediction method for bus rapid transit based on ARIMASVM model," *Journal of Southeast University*, vol. 46, no. 3, pp. 651–656, 2016.

[30] Y. He and Q. Xu, "Short-term power load forecasting based onself-adapting PSO-BP neural network model," in *Proceedings of the Fourth International Conference on Computational and Information Sciences*, pp. 1096–1099, Chongqing, China, August 2012.