Environment, Resource and Ecology Journal (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/erej.2023.070708
ISSN 2616-3756 Vol. 7 Num. 7

Analysis model of Yellow River water and sediment monitoring data based on multiple regression

Zhenpeng Shi[*]

School of Information Engineering, Yangzhou Polytechnic College, Yangzhou, 225009, China
[*]*Corresponding author*

Keywords: Multivariate nonlinear regression, Pearson correlation, Yellow river water and sediment

Abstract: The Yellow River is one of the most important rivers in China, and the monitoring of ecological environment in the Yellow River basin is very important to protect the ecosystem and sustainable development of the Yellow River. The research on the analysis of water and sediment detection data in the Yellow River Basin has a certain theoretical guiding significance for water resource allocation and water and sediment coordination in the Yellow River basin. In order to explore the relationship between the sediment content of the Yellow River and the time, discharge and water level, we introduced a multivariate nonlinear regression model based on the general linear regression model. A multivariate nonlinear regression analysis model of Yellow River water and sediment is constructed by using the least square theory and Pearson correlation visualization in processing big data and data relationships.

1. Introduction

"The governance of the Yellow River basin should focus on protection and governance." Clarifying the effect of ecological management on the Loess Plateau and the changes in the water-sediment relationship of the Yellow River under the influence is helpful to provide scientific and technological support for the coordination of water-sediment relationship of the Yellow River. By sorting out the history and changes of soil and water loss control measures in the Loess Plateau, the sediment reduction effects of control measures were analyzed. The results show that: (1) The Loess Plateau goes through five typical treatment stages, the main color changes from yellow to green, and the underlying surface changes irreversibly; (2) Since 2000, the flood season rainfall in the main sand-producing areas has been more abundant and the extreme rainfall has increased, while the amount of sand in the Yellow River has decreased by 85% compared with that in 1919-1959, the frequency of moderate and regular floods in the middle reaches has decreased, and the downstream channels have changed from siltation to erosion; (3) The sediment reduction effect of forest and grass, terraced fields and silting dam is remarkable, and the ratio of silting sediment discharge from silting dam to silting dam is lost. 15%, it is difficult to occur "fractional deposit and withdrawal" phenomenon; (4) Compared with the historical extreme rainfall events, under similar rainfall conditions after 2000, the sub-flood amount and sediment amount of typical watershed decreased by 30%-78% and 53%-88%, respectively, indicating significant soil and water conservation results. (5) Under the new water and sediment situation, the pattern of soil and water loss control on the Loess Plateau should be timely adjusted, the water and sediment control system should be improved, the downstream river channel should be reformed, and the beach area should be liberated. Since the implementation of the unified allocation of water resources in the Yellow River Basin in 1999, the occurrence of downstream flow interruption has been effectively curbed. From 2002 to 2006, continuous water and sediment transfer has been carried out. These two important measures have significantly changed the water and sediment environment in the Yellow River estuary under the action of human resources. The change of the environment will inevitably cause the change of the Yellow River estuary. It is of great significance to analyze the variation of water and sediment discharge in the Yellow River basin and its influence on people's life, climate change and environmental regulation. The fitting analysis of the relationship between the sediment content of the Yellow River and the water level, discharge and time is an important part of the study of the Yellow River sediment. The general linear regression has certain defects, so we consider using the combination of multiple linear regression and multiple nonlinear regression to study the monitoring data of the Yellow River sediment[1-3].

2. Basic principles of linear regression model

2.1 Modeling the relationship between sediment content and time, water level and water flow

2.1.1 Correlation analysis

In view of the first question of the topic, this paper asks to explore the relationship between sediment content, time, water level and water flow. Firstly, Pearson correlation coefficient is used to analyze the above variables[4].

Pearson correlation coefficient is used to detect the correlation between each variable. The value range is [-1,1]. A positive value indicates a positive correlation, and a negative value indicates a negative correlation. The calculation formula is shown in equation (1).

$$\rho_{m,n} = \frac{\operatorname{cov}(m,n)}{\sigma_m \sigma_n} = \frac{E((m-\mu_m)(n-\mu_n))}{\sigma_m \sigma_n} \tag{1}$$

In formula (1), it represents the covariance of m variable and n variable, which is the variance of m variable and n variable respectively. In order to further refine the influence of time on sediment content, the time unit is divided into three variables, namely, year, month and day. In summary, Pearson correlation was conducted on six variables, namely year, month, day, sediment content, water level and water flow, and thermal maps were drawn, as shown in Figure 1.

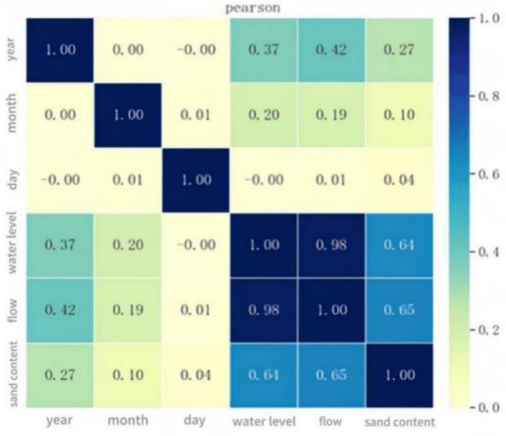

Figure 1: Correlation analysis diagram

From the thermal map in Figure 1, it can be seen that the Pearson coefficients of sediment concentration on discharge and water level are 0.65 and 0.64, respectively, indicating that sediment concentration and discharge are positively correlated with water level. From the relationship between sediment concentration and time, it can be seen that the significance of year to sediment concentration is the highest 0.27, and the other time variables are 0.1 and 0.04, respectively. On the whole, the correlation between sediment concentration and time is low, but there is still a certain relationship.

2.1.2 Multiple linear regression and nonlinear regression models

(1) Establishment of multiple linear regression and nonlinear regression models

According to the correlation analysis in 4.3.1, there is a certain correlation between sediment concentration and discharge, water level and time. Aiming at this correlation, this paper established the functional relationship between sediment content, discharge, water level and time, namely:

$$SC = f(x_1, x_2, x_3) \tag{2}$$

In formula (2), SC stands for sediment content, flow rate, water level and time. This paper selects a common regressionmodel - multiple regression model.

multiple linear regression model: Firstly, the relationship between sediment content and discharge, water level and time is established, as shown in equation (3)

$$SC = k_0 + k_1x_1 + k_2x_2 + k_3x_3 \tag{3}$$

Where is the linear regression coefficient. The least square method is often used to solve the regression coefficient[5-6].

The ordinary least square method is by finding the best function and finding the best function. The coefficient matrix is solved by matrix operation, as shown in equation (4).

$$k^T = (X^T X)^{-1} X^T SC \tag{4}$$

Multiple nonlinear regression models: nonlinear models consider the interactions between variables and can be applied to a wider range of data than linear regression models. According to the Pearson heat map, it can be seen that there is a positive correlation between water level, flow rate and time, and in particular, the Pearson correlation coefficient between flow rate and water level is 0.98. It can be shown that there is a certain interaction between the three independent variables. The method of multivariate polynomial regression is to take into account the cross-action between variables. The regression coefficient is usually solved by the least square method. After testing, cubic multinomial fitting has the best effect, and the formula of ternary cubic polynomial fitting is shown in equation (5) :

$$\text{y=b}_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{11}x_1^2$$

$$\begin{aligned}&+b_{22}x_2^{\;2}+b_{33}x_3^{\;2}+b_{12}x_1x_2\\&+b_{13}x_1x_3+b_{23}x_2x_3+b_{111}x_1^{\;3}\\&+b_{222}x_2^{\;3}+b_{333}x_3^{\;3}\\&+b_{123}x_1x_2x_3+b_{112}x_1^{\;2}x_2\\&+b_{113}x_1^{\;2}x_3+b_{221}x_2^{\;2}x_1\\&+b_{223}x_2^{\;2}x_3+b_{331}x_3^{\;2}x_1\\&+b_{332}x_3^{\;2}x_2\end{aligned} \tag{5}$$

Where, is the flow, is the water level, is the time; bi is a polynomial coefficient [7-8].

Model evaluation criteria

The fit degree evaluation of the established multiple linear regression model and nonlinear regression model is evenly evaluated by the following three indexes:

Goodness of fit (R^2) evaluates the result of a polynomial fit, as in equation (6):

$$R^2=\frac{SSR}{SST}=\frac{\sum\left(\hat{y}_i-\bar{y}\right)^2}{\sum\left(y_i-\bar{y}_i\right)^2} \tag{6}$$

Mean square sum error (RMSE) evaluates the results of polynomial fitting and is used to check whether there is a deviation between the observed value and the true value, as shown in equation (7):

$$RMSE=\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i-\hat{y}_i\right)^2} \tag{7}$$

Mean absolute error (MAE) evaluates the results of polynomial fitting, which can better reflect the actual situation of predicted value error, as shown in equation (8):

$$MAE=\frac{1}{m}\sum_{i=1}^{m}\left|\left(y_i-\hat{y}_i\right)\right| \tag{8}$$

3. Results

3.1 Multiple linear regression and multiple nonlinear regression solution

The result of solving multiple linear regression model.

Linear regression: Firstly, this paper uses linear regression method to solve the relationship between sediment content and flow, water level and time. The placement of the regression model is shown in equation (9).

$$SC=-38.524+0.001x_1+0.913x_2+0.002x_3 \tag{9}$$

Linear regression predicted the variation of sediment content, as shown in Figure 2.

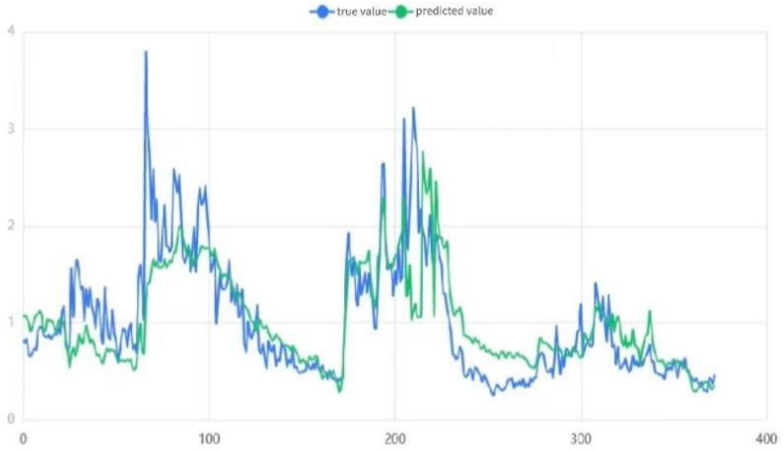

Figure 2: Variation of sediment content obtained by linear regression

A statistical analysis table was obtained for the linear regression results obtained from equation (9) and Figure 2.

3.2 Results of nonlinear regression model

Similar to the solution of multiple regression linear model, the least square method is used to regression the ternary cubic polynomial model, and the coefficients of the model are returned. The results obtained are as follows:

```
[0.00000000e+00          8.18206063e+00            -1.78017462e+04
1.10923423e+01           7.22903815e-03            -3.82780845e-01
3.66987341e-04          4.20653874e+02           -5.29078568e-01
3.15083521e-04          1.47618025e-05            -1.79665742e-04
-4.79034380e-08         4.47370824e-03            -7.93286054e-06
-1.76436750e-08         -3.31333431e +00           6.30922372e-03
-7.51353463e-06         3.04183878e-09]
```

The relationship curves of sediment content, water level, discharge and time are shown in Figure 3.

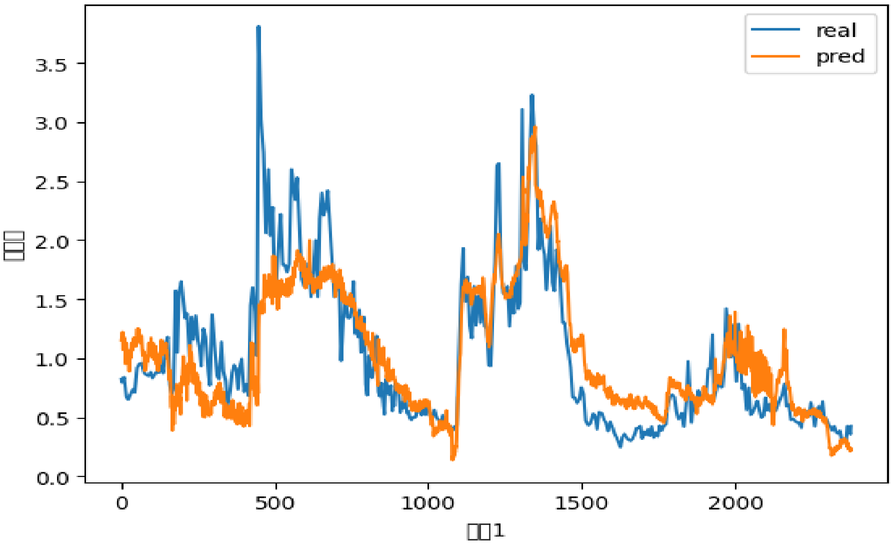

Figure 3: Relation curve of sediment content with water level, discharge and time

In order to check the stability of the ternary nonlinear regression equation established. We carried out random sampling on the data in Attachment 1 to verify the stability of the regression model. The results of the second and second regressions are shown in Figure 4. The results of coefficients are as follows:

[0.00000000e+00 3.26978969e-01 -6.72967603e+01 1.06267786e-01 -3.77439536e-05 -7.71905929e-03 5.32022325e-06 8.14681438e-01 -2.47502937e-03 6.39108888e-07]

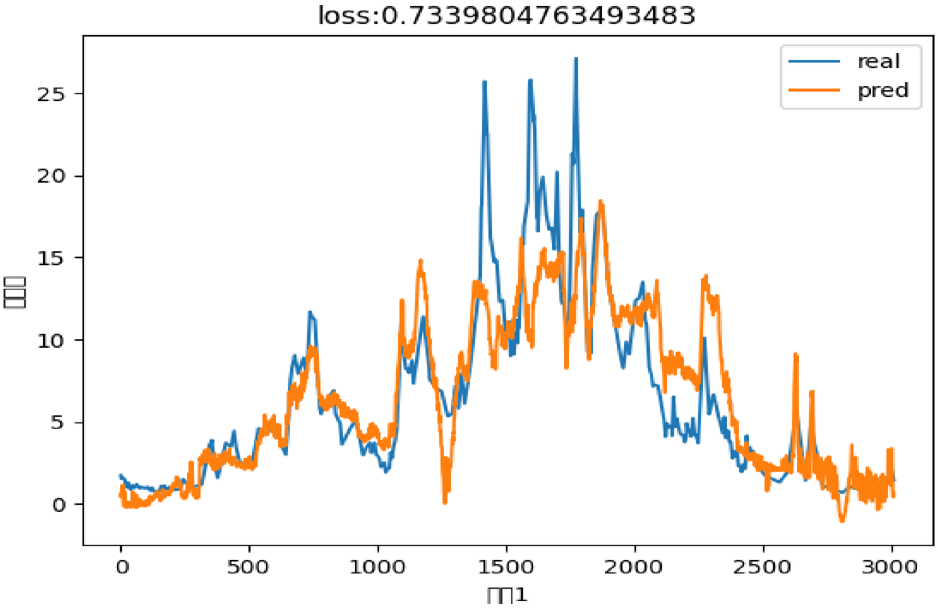

Figure 4: Second multivariate nonlinear regression diagram

According to the analysis and comparison of the two regression renderings, it is found that the model can get similar results on different data sets, so the model can be considered stable. The model has strong stability.

For the ternary cubic polynomial regression equation of sediment quantity, water level, flow rate and time, the bi coefficient solved for the first time is selected, and the results are shown in equation (10).

$$SC = 0 + 8.182x_1 - 1.78\times10^4 x_2 + ... + 3.0418\times10^{-9} x_3^2 x_2 \tag{10}$$

The error evaluation criteria of equations (8) - (10) are used below, and the evaluation results are shown in Table 1.

Table 1: Model fitting comparison

Model	Degree of fitting(R^2)	Root mean square error(MSE)	Mean absolute error(MAE)
Linear regression	0.587	9.443	2.364
Nonlinear regression	0.734	8.099	1.805

To sum up, the evaluation results of the two models were obtained. Formula (10) was selected to describe the relationship between sediment content, discharge water level and time[9-10].

Therefore, this paper uses the nonlinear regression model to complete the incomplete data of sediment content in Annex I, and part of the data list is shown in Table 2.

Table 2: Prediction of sediment content by multivariate nonlinear regression

Year	Month	Day	Water level	Flow rate	Sediment concentration	Date
2016	1	1	42.80142857	363.4285714	0.804285714	2016-1-1
2016	1	2	42.77571429	348.4285714	0.827428571	2016-1-2
2016	1	3	42.74666667	331.3333333	0.728	2016-1-3
2016	1	4	42.71857143	315	0.657	2016-1-4
2016	1	5	42.74142857	328.5714286	0.681	2016-1-5
2016	1	6	42.77285714	358.7142857	0.722285714	2016-1-6
2016	1	7	42.78714286	381.4285714	0.717571429	2016-1-7
2016	1	8	42.80375	389.375	0.841625	2016-1-8
2021	12	22	43.05833	1051.667	2.65	2021-12-22
2021	12	23	43.05667	1050	2.65	2021-12-23
2021	12	24	43.13667	1161.667	2.65	2021-12-24
2021	12	25	43.12667	1156.667	2.09	2021-12-25
2021	12	26	43.14667	1176.667	1.81	2021-12-26
2021	12	27	43.14667	1176.667	1.81	2021-12-27

4. Conclusion

The trend of the Yellow River detection data provides a basis for the categorization and quantitative relationship model, but the traditional linear regression method cannot bear such a huge consumption of time and computing resources. The overfitting problem of large sample sets will affect the model accuracy. In this paper, multiple linear regression and multiple nonlinear regression are used to establish the relationship model between sediment content and time, water level and water flow. The coefficient of multiple linear regression model can be directly used to explain the influence of independent variables on dependent variables, which can provide us with a deeper understanding of data relations. The multivariate nonlinear regression model is more convenient to fit the data among nonlinear relations, and the nonlinear relations existing in the data can be studied by using nonlinear functions or nonlinear terms, so as to improve the accuracy of the model.

References

[1] Hu Chunhong, Zhang Xiaoming. Soil and water loss control and sediment change of Yellow River in Loess Plateau [J]. Water Resources and Hydropower Technology ,2020(1):1-11.

[2] Bainbridge Z T , Lewis S E , Smithers S G ,et al.Fine ‐ suspended sediment and water budgets for a large, seasonally dry tropical catchment: Burdekin River catchment, Queensland, Australia[J].Water Resources Research, 2015, 50(11):9067-9087.DOI:10.1002/2013WR014386.

[3] Gao Zongjun, Feng Guoping. Runoff sediment variation trend and cause analysis [J]. Journal of groundwater, and 2020.147-151.

[4] Zhiyi W,Tingyu W,Yongqiang Y, et al. Differential Confocal Optical Probes with Optimized Detection Efficiency and Pearson Correlation Coefficient Strategy Based on the Peak-Clustering Algorithm.[J]. Micromachines,2023,14(6).148-155

[5] Ziyun Y,Lei C,Gang L, et al. Physics-based Bayesian linear regression model for predicting length of mixed oil[J]. Geoenergy Science and Engineering,2023.223.-235

[6] Jürgen G,Annette M. A Note on Cohen’s d From a Partitioned Linear Regression Model[J]. Journal of Statistical Theory and Practice,2023.72-80

[7] Yangguang R,Ziqi L,Zhiqiang X, et al. Slurry-ability mathematical modeling of microwave-modified lignite: A comparative analysis of multivariate non-linear regression model and XGBoost algorithm model[J]. Energy,2023,281.69-82

[8] Zhao H , Kou H , Xia R .Research on the evaluation system of Yellow River water and sediment mathematical

model[J].Yellow River, 2014.

[9] Pan Bin, Change of water and sediment in the Yellow River and its response to Climate Change and Human Activities, Shandong Normal University, 2021.57-63

[10] Liu Hongwei, Wang Hongtao, Ma Haibing, Yellow River Water Pollution Monitoring based on Remote Sensing and GIS technology, Journal of Henan Vocational College of Water Resources and Hydropower,2010.45-51