Yellow River Water and Sand Monitoring Data Model based on SARIMAX Rediction

Xingyu Li*

Changsha Vocational and Technical College, Changsha, China

*Corresponding author: 927895075@qq.com

Abstract. With the evolution of history, the problems faced by the Yellow River Basin have been an important research topic in hydrology. In this paper, for the problems in the monitoring of water and sand in the Yellow River, a seasonal time series model of the monitoring data is established based on the time series prediction model, and the CUSUM algorithm is used to test the model. First, the data are processed to calculate the spearman correlation for each pair of categories. Then, for analyzing the description of single dimensional data, data descriptive analysis was used. Next, the data were forecasted using a seasonal time series forecasting model. Finally, the seasonal time series prediction model was used to predict water fluxes and solved using double interpolation to obtain the predicted effect plots for the next ten years. The proposed model is close to the reality, can reasonably solve the proposed problem, and is characterized by high practicality and high efficiency of the algorithm.

Keywords: Time Series; Yellow River Water and Sand Monitoring; SARIMAX.

1. Introduction

The Yellow River is one of the many great rivers in China and an important birthplace of Chinese history and civilization. However, for a long time, the Yellow River has brought about a lot of harm to the coast by flooding. In order to solve the problem of Yellow River management, China has invested a lot of strength to carry out sand transfer work, and Xiaolangdi Water Conservancy Hub Project is one of the very important part, the role played by the water transfer and sand transfer is also very significant. However, the management of the Yellow River has also brought some negative impacts [1].

In order to better monitor the water and sand of the Yellow River, this paper studies the relationship between the sand content of the Yellow River water at this hydrological station and the time, water level and water flux, and estimates the total annual water flux and total annual sand discharge at this hydrological station in the recent 6 years. Analyze the characteristics of sudden change, seasonality and periodicity of water and sand flux in this hydrological station in the recent 6 years, and study the change rule of water and sand flux. According to the change rule of water and sand flux in this hydrological station, predict and analyze the change trend of water and sand flux in this hydrological station in the next two years, and formulate the optimal sampling and monitoring plan for this hydrological station in the next two years (the number of times of sampling and monitoring and the specific time, etc.), so as to make it not only grasp the dynamic change of water and sand flux in a timely manner, but also minimize the cost of resources for monitoring. Based on the changes in water and sand fluxes and river bottom elevation at this hydrological station, we analyze the actual effect of "water and sand transfer" from Xiaolangdi Reservoir in June and July each year. Predict the river bottom elevation at the station 10 years later if no "water and sand transfer" is carried out.

In this paper, a seasonal time series model of the monitoring data is developed based on a time series forecasting model and the CUSUM algorithm is used to test the model. First, the data were processed to calculate the spearman correlation for each pair of categories. Then, for analyzing the description of single dimensional data, data descriptive analysis was used. Next, the data were forecasted using a seasonal time series forecasting model. Finally, the seasonal time series prediction model was used to predict water fluxes and solved using double interpolation to obtain the predicted effect plots for the next ten years. The proposed model is close to the reality, can reasonably solve the proposed problem, and is characterized by high practicality and high efficiency of the algorithm. This model can also be used for new cases of monkeypox worldwide [2].

2. Model Establishment and Solving

2.1. Study of trends in water and sand fluxes

Firstly, the relationship between the sand content of the Yellow River water at this hydrological station and the time, water level and water flow rate is analyzed, and then data analysis, visualization and correlation analysis are carried out to determine the linear correlation between them by calculating the correlation coefficients between the water level, water flow rate and sand content. The use of line graphs and scatter plots is considered for elaboration. The summation formula is used to estimate the total annual water flow and total annual sand discharge at this hydrological station for the last 6 years.

The total annual streamflow and total annual sand discharge can be expressed as respectively:

$$\begin{aligned}&\text{total year water flux}\ (\mathrm{m}^3) = \sum\left(\text{each time water flux}\ (\mathrm{m}^3/\mathrm{s})\right)\\&\text{total year sand flux}\ (\mathrm{kg}) = \sum\left(\text{each time sand flux}\ (\mathrm{kg/m}^3 \times \text{water flux}\ (\mathrm{m}^3/\mathrm{s})\right)\end{aligned} \tag{1}$$

Spearman's correlation was calculated for each pair of categories, as shown in Table 1.

Table 1 Correlation between categories

	Year	Month	day	water level (m)	Flux (m^3/s)	Sand content (kg/m^3)
Year	1.000000	0.029047	0.016027	0.419851	0.498170	0.494153
month	0.029047	1.000000	0.001066	0.171034	0.169708	0.088361
day	0.016027	0.001066	1.000000	0.026092	0.028952	0.023780
water level (m)	0.419851	0.171934	0.026092	1.000000	0.963965	0.881845
Flux (m^3/s)	0.498170	0.169708	0.028952	0.963965	1.000000	0.911847
Sand content (kg/m^3)	0.494153	0.088361	0.023780	0.881845	0.911847	1.000000

The conclusion is drawn by applying the difference diff as follows. The annual total water flow for year 2016 is 14374641600.000000 m³ and the annual total sand discharge is 18333055159.199997 kg. the annual total water flow for year 2017 is 15338538000.000000 m³ and the annual total sand discharge is 19047051414.000000 kg. the 2018 annual Total annual water flow is 38889187200.000000 m³, and total annual sand discharge is 293539431582.000000 kg. year 2019 total annual water flow is 38723788800.000000 m³, and total annual sand discharge is 304010272620.000000 kg. year 2020 total annual water flow is 43375467600.000000 m³, and the total annual sand discharge is 350911312786.799988 kg. The total annual water flow for year 2021 is 47255094000.000000 m³, and the total annual sand discharge is 225527339538.000000 kg.

Studying the trend of water flux (which can be expressed in terms of total water flow) and sand flux (which can be expressed in terms of total sand), it can be learned that it is a very typical time series analysis, which mainly analyzes the point in time when the water and sand fluxes change significantly or not, and plots the seasonal components of the graphs are first imported into the first question of the data that has been processed, and data visualization and analysis is carried out on it, to see the relationship between the data, and then mutation detection is used methods, seasonal decomposition methods, and consider the use of the SARIMAX model.

The temporal correlation coefficient model was built to solve whether there is correlation or not, and at the same time we imported it into SPSS for correlation test. The autocorrelation coefficient model is

$$ACF(k) = \rho_k = \frac{Cov(y_t, y_{t-k})}{Var(y_t)} \tag{2}$$

The partial autocorrelation coefficient is modeled as

$$\text{PACF}(k) = \frac{\text{cov}[(Z_t - \bar{Z}_t),(Z_{t-k} - \bar{Z}_{t-k})]}{\sqrt{\text{var}(Z_t - \bar{Z}_t)}\sqrt{\text{var}(Z_{t-k} - \bar{Z}_{t-k}}} \tag{3}$$

The "CUSUM" algorithm was used to detect mutations in the data [3]. For mutability, the water-sand flux was calculated based on the data from the first question, and the time series of water-sand flux was plotted as shown in Fig. 1 to observe whether there were obvious mutations.

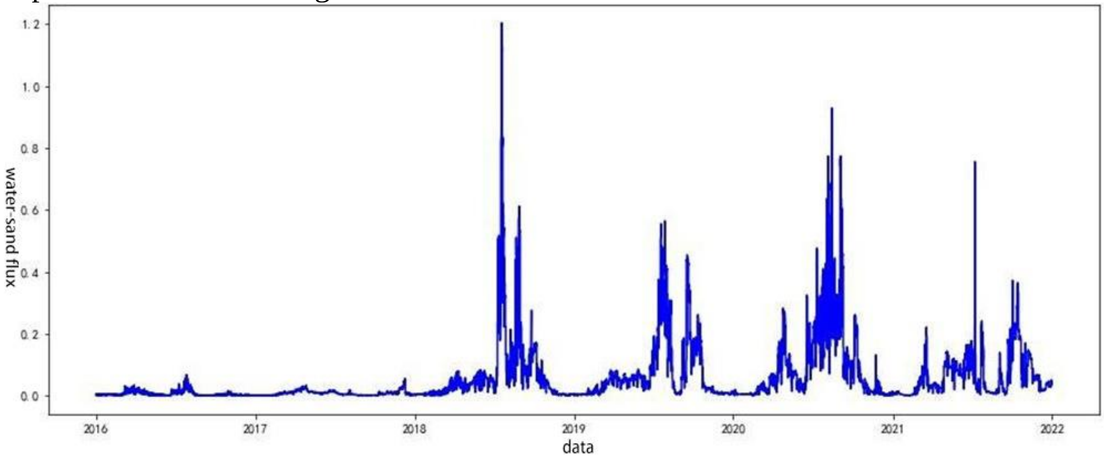

Fig. 1 Time series diagram of water-sand flux

The seasonal decomposition method was used to further analyze the seasonality of water and sand fluxes, and the decomposition was plotted as shown in Fig. 2.

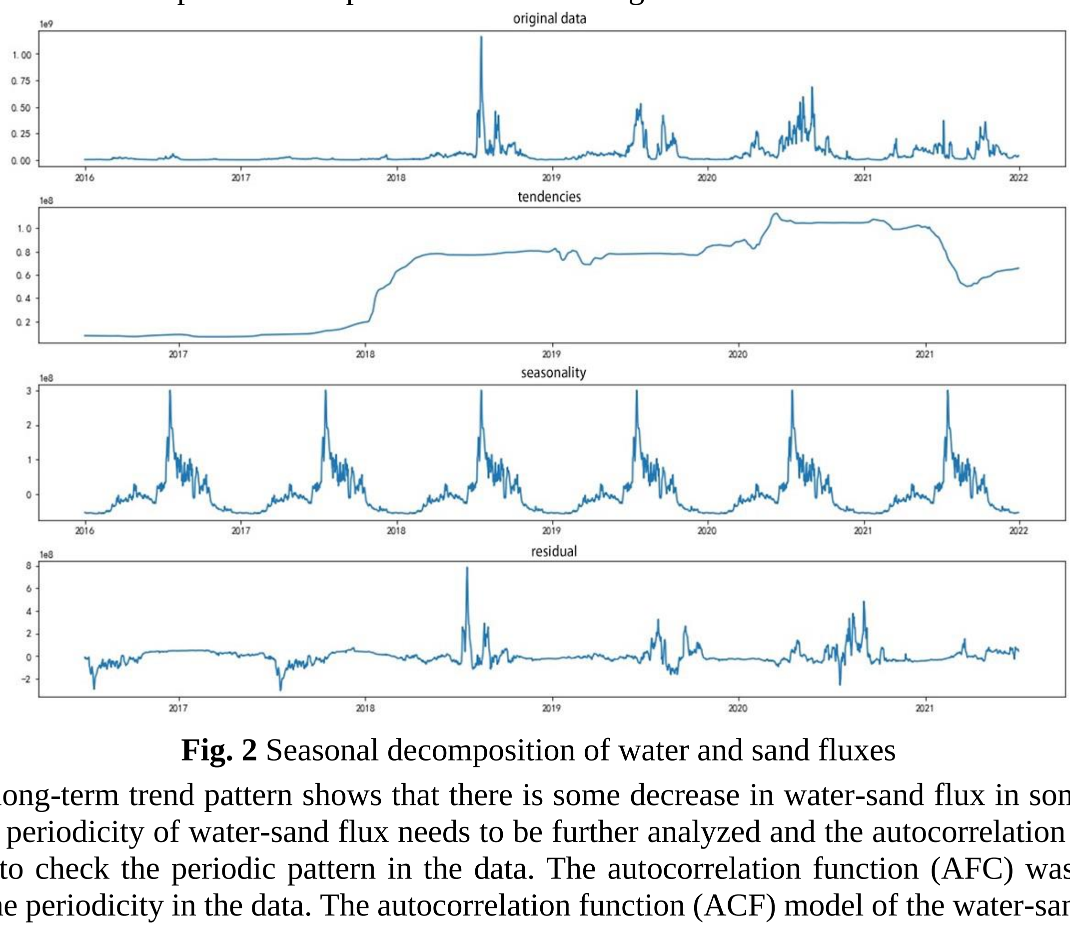

Fig. 2 Seasonal decomposition of water and sand fluxes

The long-term trend pattern shows that there is some decrease in water-sand flux in some years, thus the periodicity of water-sand flux needs to be further analyzed and the autocorrelation function is used to check the periodic pattern in the data. The autocorrelation function (AFC) was used to check the periodicity in the data. The autocorrelation function (ACF) model of the water-sand flux is shown in Fig. 3, and it can be seen that there are obvious periodic peaks, and the surface water-sand flux has annual periodicity. There are some small peaks reflecting that there is some short-term periodicity. The water-sand flux mutation diagram is shown in Fig. 4, the water-sand flux shows a clear seasonal and periodic pattern, there are a total of 37 mutations in a certain period of time, which shows the mutability, the threshold value is between 1e5 and 2e5 by the algorithm, and the change is within this interval.

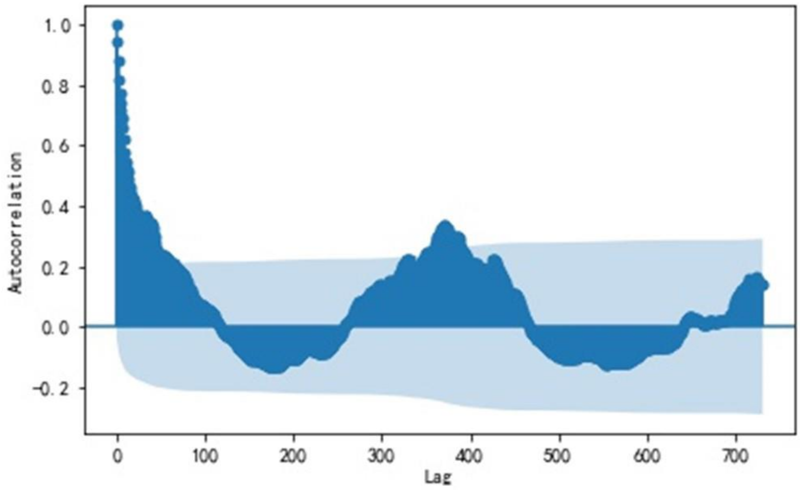

Fig. 3 Autocorrelation function (ACF) of water and sand fluxes

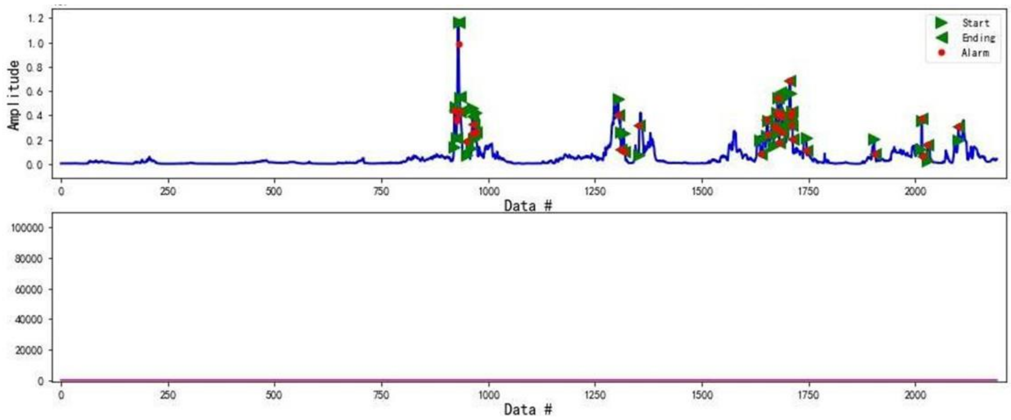

Fig. 4 Water-sand flux mutation map

2.2. Water and sand flux projections for the next two years

A model is built to make it predict the water and sand fluxes of this hydrological station for the next two years, we can predict only the residual values, then invert their predicted values based on seasonal analysis, and finally apply the predicted values to the seasonal analysis method and develop the optimal sampling and monitoring program for the next two years for this water stabilization station.

The formula for the SARIMA model is

$$\text{ARIMA}\left(p,d,q\right)\left(P,D,Q,m\right) \quad (4)$$

Simplified parameter search, the SARIMA model found the best combination of parameters (p,d,q) as (1,1,1). The seasonal parameters (p,d,q,m) are (0,1,1,12).

The projected trends in water and sand fluxes for the next two years are shown in Fig. 5. The red line indicates the predicted value, while the pink area indicates the confidence interval of the prediction. Based on the predictions, we can develop the following sampling and monitoring program for this hydrological station. High-frequency monitoring during critical periods: considering the obvious peaks and valleys in the prediction, we recommend increasing the sampling frequency during these critical periods. Low-frequency monitoring during stable periods: Sampling frequency can be reduced during periods when the predicted water and sand fluxes are relatively stable.

The specific program is as follows. High-frequency sampling: during the predicted peak and trough months of the year, sampling is conducted once a week, totaling approximately 4 times/month. Low-frequency sampling: in other months, sampling is conducted once a month. Such a program provides timely information on the dynamics of water and sand fluxes while minimizing monitoring cost resources.

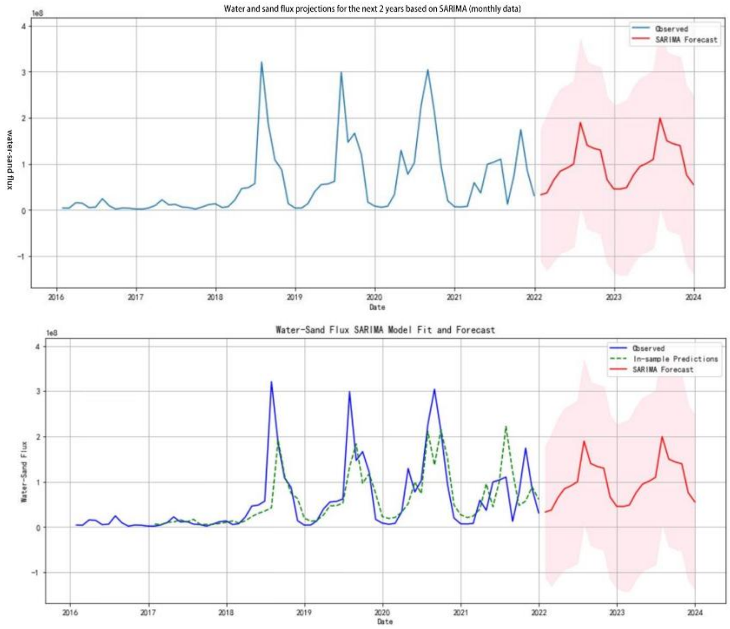

Fig. 5 SARIMA water and sand flux projections for the next two years

Analyze the actual effect of "water and sand transfer" in Xiaolangdi Reservoir in June-July each year, simulate and predict the data in January-June, get the predicted effect in the following June-December, and then calculate the difference between the predicted value and the actual value, so as to assess the actual effect of "water and sand transfer" on water and sand fluxes. Then calculate the difference between the predicted value and the actual value, then we can assess the actual effect of "water and sand transfer" on the water and sand flux. Then take the mean value of its river bottom elevation, establish a prediction model to predict the next ten years without "water and sand transfer", and finally get the data. Double difference DID (double difference method), optimization algorithm, and SARIMAX model are considered [4].

The DID model can be expressed as

$$Y_{it} = \alpha_0 + \alpha_1 du + \alpha_2 dt + \alpha_3 du \cdot dt + \varepsilon_{it} \tag{5}$$

A linear regression equation is used. Given N data pairs (a, b) find the parameters m* and c* so that the model can fit these data better. The loss function is expressed as

$$L(\breve{e}, e) \tag{6}$$

The error between the predicted and true values can be expressed as

$$L(\breve{e}_i, e_i) = (\breve{e}_i - e_i)^2 \tag{7}$$

The error between the predicted and true values at a single sample point can be expressed as follows

$$L(\breve{e}_i, e_i) = \frac{1}{N}\sum_{i=1}^{N}(\breve{e}_i - e_i)^2 \tag{8}$$

$$L(\breve{e}_i, e_i) = \frac{1}{N}\sum_{i=1}^{N}[(ma_i + c) - y_i]^2 \tag{9}$$

$$m^*, c^* = \arg\min L(m,b) = \arg\min \frac{1}{N}\sum_{i=1}^{N}[(ma_i + b) - e_i]^2 \tag{10}$$

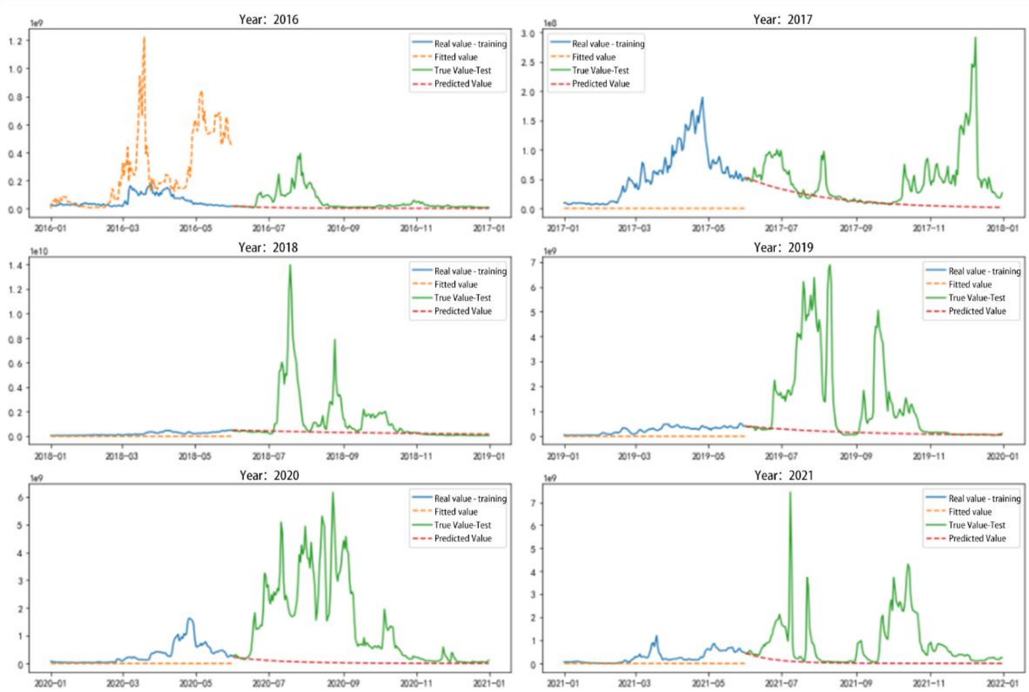

Fig. 6 Dataset for training and testing

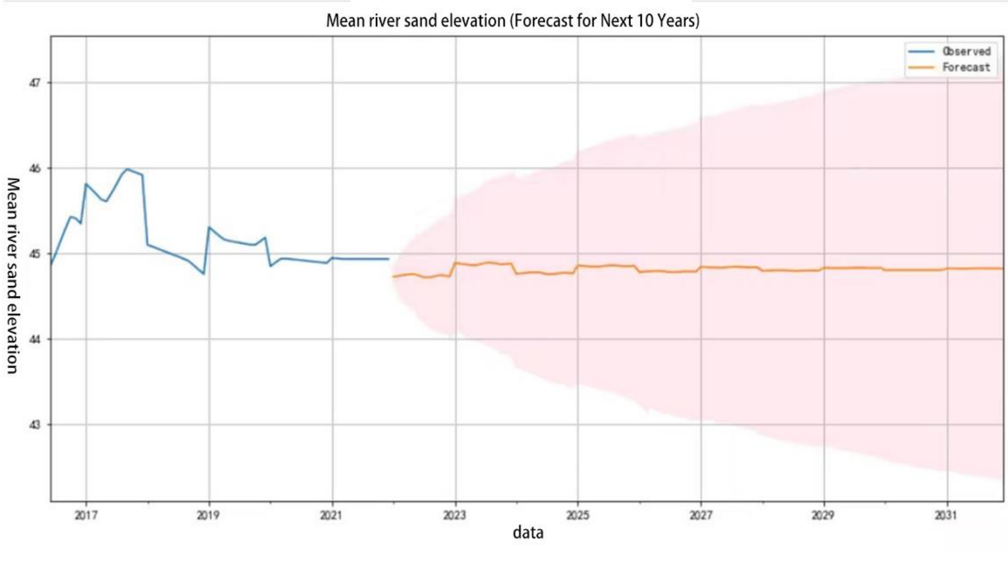

Fig. 7 Fitted plot of mean river sand elevation for the next ten years

DID estimates were calculated for each year from dataset 2016 through 2021 as depiced in Fig. 6. These estimates indicate the actual effect of water and sand fluxes that would increase or decrease as a result of the water and sand transfer after accounting for the effects of the predicted values (i.e., the expected results in the no-intervention scenario).The DID estimates are shown in Fig. 7. Negative values indicate that the actual water and sand fluxes are lower than the predicted values in the absence of the intervention, and positive values indicate that the actual water and sand fluxes are higher than the predicted values indicating that the values of water and sand fluxes are all reduced after the intervention.

The values of the river bottom height change based on the DID model are shown in Table 2.

Table 2 Values of river bottom height changes according to DID modeling

year	Value of change in river bottom height
2017	-0.5299242558685364
2018	0.2626274788472074
2019	-0.33924952352026594
2020	0.04495141189909191
2021	-0.021475463592651067

The water-sand fluxes predicted from the linear regression equation after ten years are shown in Table 3.

Table 3 Ten-year post-water-sand flux scenario

dates	water-sand flux
2022-01-31	44.857046
2022-02-28	44.961070
2022-03-31	45.050889
2022-04-30	45.128562
2022-05-31	45.189609
2031-08-31	47.250909
2031-09-30	47.260689
2031-10-31	47.269726
2031-11-30	47.280754
2031-12-31	47.292275

3. Model Evaluation and Extension

3.1. Model Benefits

(1) The model is fully integrated with reality and can capture the correlation of time, which helps to understand more accurately the changing pattern of data and predict the future. (2) The model can make predictions about the future based on historical data. (3) The model can provide short- and medium- to long-term forecasting guidance to help relevant departments and decision makers in water and sand scheduling and management.

3.2. Model Drawbacks

(1) Geographic location and weather factors may also be important in practical applications, but this paper fails to take into account the effects of these factors, which affects the accuracy of the model to some extent. (2) The model proposed in this paper is more effective for the use of existing conditions, and due to time issues did not test for other situations. For other situations geographic location and human factors, etc., may not be able to achieve better results. (3) In fact, the effects of time, water level, flow volume and sand content are not necessarily linear, and this paper treats them as linear factors, ignoring the influence of marginal effects.

3.3. Model promotion

Agricultural production and decision support: the Yellow River water and sand information data are of great significance for agricultural decision making such as crop irrigation and agricultural products cultivation [5]. The extension of time series based prediction model can provide accurate water resources prediction and decision support for agricultural production and management.

4. Conclusion

In this paper, the Yellow River water and sand monitoring problem is studied, and based on the time series prediction model, a seasonal time series model of the monitoring data is established, and the CUSUM algorithm is used to test the model. Firstly, the data are processed, the spearman correlation of each pair of categories is calculated, and the correlation graph between each variable is plotted. Total annual streamflow was obtained by summing the water flow at each time point, and the product of sand content and water flow at each time point was summed to learn the total annual sand discharge. For analyzing the description of single dimensional data, data descriptive analysis was used. Water-sand fluxes were calculated, time-series plots of water-sand fluxes were plotted to observe abrupt and seasonal changes, cyclic variation data discovery was performed, seasonal decomposition was performed and seasonal decomposition was plotted. Seasonal time series forecasting model was utilized to forecast the data. The merging is first done by time period, and grid search optimization is performed by setting the period parameters as a way to find the optimal parameters. In addition, time series and seasonal pdq values were obtained from the forecasting model and used in the fitted plots. The seasonal time series prediction model was used to predict water fluxes and solved using the double interpolation method. Simulations constructed six datasets to predict the end-decade effects using the seasonal time series model and used an optimization-seeking algorithm to find the pda, seasonal pda, and periodicity to obtain the predicted effect plots for the next decade.

Reference

[1] Gao Xing, Zhu Cheng Hao, Chen Cui Xia et al. Optimization study of water and sand transfer from Xiaolangdi Reservoir considering the change of Yellow River water and sand abundance and depletion [J/OL]. People's Yellow River: 1-6 [2023-09-14]. http://kns.cnki.net/kcms/detail/41.1128.TV.20230911.1709.002.html

[2] Ouyang Zelong, Li Jinqiong, Wang Guanju et al. Forecasting new cases of monkeypox weekly globally based on time series modeling. Chinese Journal of Animal Infectious Diseases, Published online. https://doi.org/10.19958/j.cnki.cn31-2031/s.20230906.004.

[3] Li Cheng, Du Kun, Zhou Ming et al. Analysis of pipe burst detection in water supply network based on CUSUM algorithm. Water Supply and Drainage, 2016, 52(11): 118-121.

[4] Zhou X, Li Y, Zeng Y, et al. Power prediction of photovoltaic power generation based on SARIMAX-SVR. Journal of Power Systems and Automation, Published online. https://doi.org/10.19635/j.cnki.csu-epsa.001322.

[5] WANG Haisong, ZHANG Xiaowei, TAN Linyuan et al. Analysis of the fluctuation characteristics and trend prediction of export price of fresh apples in China. China Fruit Tree, 2023, 2023(9): 122-127.