River Runoff Influence Factors Recognition Using Stepwise Regression Analysis: The Case of a Northern Chinese Coal Mining Area

Xi-jun Wu¹,²*, Ying Dong¹,², Ya-ning Zhang¹, Jing Liu¹

¹School of Civil Engineering, Yulin University, Yulin 719000, Shaanxi, China
²Shaanxi Key Laboratory of Ecological Restoration in Shanbei Mining Area, Yulin University, Yulin 719000, Shaanxi, China

Received: 18 December 2018
Accepted: 24 January 2019

Abstract

Northern Shaanxi Province of China has been affected by severe water shortages, especially in coal mining areas. A systematic method of identifying the influencing factors and their significance on river runoff is essential for paving the way towards environmental protection. Based on stepwise regression analysis, this paper analyzed the runoff influence factors of the Kuye River for 1961-1979, 1980-1998, 1999-2016 and 1961-2016, and calculated the runoff reduction caused by various factors in different periods. We found that the main influence factors on Kuye runoff are disparate in different periods. In 1961-2016, the main factors were rainfall, temperature, soil and water conservation measures, and coal mining. Compared with the base period (1961-1979), the reduction in Kuye runoff stood at 21,569×10⁴ m³ per year in 1980-1998. In 1999-2016, runoff reduction of 52,992×10⁴ m³ was attributable to water conservation measures (57%), temperature (21%), and coal mining (25%) – but was also partially offset by the rainfall increase. These findings could serve as important references for the ecological restoration of river ecosystems in other coal-mining areas.

Keywords: northern China, river runoff, stepwise regression analysis, influence factors, coal mining, recognition

Introduction

In general, the multiple regression method has been adopted for establishing an optimal regression equation in order to predict or estimate a dependent variable as a function of selected independent variable(s). The optimal regression equation would include the independent variables that dominate the dependent variables [1, 2]. Based on this principle, stepwise regression analysis has been widely used in environmental prediction, monitoring and assessment [3]. In terms of river runoff studies, researchers typically adapt basic data to choose a desired statistical, conceptual or distributed hydrological model to study related issues [4, 5]. Stepwise regression analysis is often used to identify the main influence factors and calculate the influence degree of each factor on river runoff [6].
Growing coal mining activities in many regions across China have been causing the fragile natural environment to deteriorate in recent years [7, 8]. Although many studies have investigated river water quality in the mining area, only a few studies have focused on the river runoff because the influence mechanism of mining activities on water resources is relatively complicated [9, 10]. Therefore, statistical methods are often utilized to analyze the influence factors and response relationship of river runoff. Zhang et al. established the correlation between the coal mining activities and river runoff in the Daliuta mining area, and concluded that river runoff change is predominantly a function of coal-mining activities [11]. Jiang et al. calculated the runoff reduction caused by mining activities through statistical methods and the water balance model [12]. More recently, Shu et al. proposed a new statistical method to evaluate the significance of coal mining activities on river runoff [13]. The Kuye River basin is the main component of Northern Shaanxi Energy and Chemical Base, where water resources are generally scarce [14]. Nonetheless, little is known regarding the significance of various factors on river runoff; and the response relationship between each factor and runoff. To this end, this study adopts stepwise regression analysis to investigate them systematically. The study outcomes might provide useful insights for the ecological restoration of river ecosystems in other coal-mining areas.

**Materials and Methods**

**Study Area**

The Kuye is a tributary located in the middle reaches of the Yellow River and originating in Inner Mongolia Province and flowing through Shaanxi Province (Fig. 1). The river is 242 km long and the area of the basin is 8,706 km², covering 4,499 km² of Shenmu County. In terms of hydrology, the average annual pan evaporation is 1,788 mm per year; the average annual precipitation is 413 mm per year; and the average annual runoff is 510 million m³ per year. The Kuye basin is near the Muus Desert, an Aeolian sand area in the upstream which stores a large amount of groundwater. However, its downstream loess gully area generates surface runoff easily, and it is among the major sources of sediment transports of the Yellow River [15]. The Kuye basin contains rich coal resources where nearly 312 x 10⁴ t were mined per year in the 1980s. Large-scale mining activities from 2000 onwards saw that figure expanded to 21,500 x 10⁴ t per year in 2016 [16]. However, over-mining is adversely affecting the water circulation system (e.g., pit wastewater, land surface fracture/collapse) and destroying the original groundwater storage conditions [17].

**Materials**

Kuye River runoff data are collected from the “Yellow River Basin Hydrology Yearbook”, which was published by the Yellow River Committee [18]. The rainfall and temperature data are obtained from the China Meteorological Science Data Sharing Service Network. The remaining data are gathered from the Shaanxi Water Resources Statistical Data and Shenmu County Statistical Yearbook.

**Methods**

A previous study demonstrated that there is a strong correlation between rainfall, temperature, water consumption, Soil and Water Conservation Measures (SWCM) area, coal mining and river runoff in the Kuyebasin [19, 20]. Although the correlations between these factors and runoff are different, it is possible to use a linear regression approach to express the relationship. The influence mechanism of coal mining on water resources is complex [21]. On top of this, the mechanism analysis requires groundwater storage and migration data and relevant geological data, which are difficult to obtain [22]. Therefore, based on the available data, this paper adopts the stepwise regression method to analyze the influence factors and response relationships of river runoff in mining areas, and calculates the influence degree of each factor on river runoff. This study uses IBM SPSS software to perform all analyses.
Stepwise Regression

Conceptually, stepwise regression enables establishing the relationship between the independent variables and the dependent variable in a systematic manner. The equation of such multiple linear regression models could be expressed as:

\[ Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_i X_i \]

\[ i = 1, 2, 3, \ldots, k; \]

(1)

...where \( Y \) is the dependent variable; \( \alpha_i \) is the partial regression coefficient of the \( i \)th independent variable; and \( X_i \) is the \( i \)th independent variable.

All the independent variables \( X_i \) are entered into the regression equation one by one in accordance with their influence on the dependent variable \( Y \). When each independent variable is entered, a series of t-tests are conducted for selected variables one by one in order to retain the statistically significant independent variables [23]. The process of entering or removing an independent variable from the regression equation (e.g., Equation (1)) is a step of the stepwise regression. It also performs the significance test using the F-test at each step. If it is passed, the t-test results of each variable in the equation are checked until the P-value is less than 0.10. This process is repeated until no significant independent variables are entered into or removed from the regression equation. From there, an optimized regression equation could be obtained [24].

Standardization Analysis of the Data

Since the units among the independent variables are different, it is difficult to draw a correct conclusion by analyzing the weights of the partial regression coefficients. However, the data could be standardized, i.e., the original data could be first subtracted from the mean of corresponding variables and then divided by the standard deviation of variable, and hence we get the standardized regression coefficient [25]. The data standardization process does not affect the testing of independent variables and equations. The standardized regression coefficient reflects the significance of the independent variables and is related to the dispersion degree of the independent variables. It is more important if it fluctuates more, otherwise it is less important. According to the standardized coefficient of each variable, the weight can be calculated. The influence degree of each independent variable on the dependent variable is reflected in the weight, given by:

\[ q_i = \frac{L_i}{\sum |L_i|} \]

(2)

...where \( L_i \) is the standardized coefficient of the \( i \)th independent variable and \( q_i \) is the weight of the \( i \)th independent variable.

Results

According to Wu et al. [26], the main influence factors for the Kuye River runoff are rainfall, temperature, SWCM area, water consumption and coal mining. In this study, the study duration is divided into four periods: 1961-1979, 1980-1998, 1999-2016 and 1961-2016. The Kuye River runoff and basin’s coal mining in 1961 to 2016 is shown in Fig. 2. To develop the runoff regression equation of the Kuye River for each period, it is possible to first conceptualize an initial equation:

\[ y = \alpha_0 + \alpha_1 R + \alpha_2 T + \alpha_3 U + \alpha_4 W + \alpha_5 M \]

(3)

...where \( \alpha_i \) is the partial regression coefficient of each variable; \( R \) is the rainfall of the Kuye River basin in each year, mm; \( T \) is the temperature of the Kuye River Basin in each year, ºC; \( U \) is the production and living water consumption of each year in Shenmu County, 10⁴ m³; \( W \) is the soil and water conservation measures (SWCM) area in each year in Shenmu County, km²; and \( M \) is the amount of coal mined in each year in Shenmu County, 10⁴ t.

Fig. 2. The Kuye River runoff and coal mining in 1961-2016(1).
Large-scale coal mining activities started in 1998 resulted in decreases of Kuyerunoff, but the increase of runoff occurring since 2010 was driven by a small increase in rainfall. Within this period, temperatures remained above 9ºC and the water conservation measures area expanded and water consumption increased.

Table 1 displays the outcomes of regression analysis. The stepwise regression is performed in two steps, where two independent variables (rainfall R and coal mining M) are tested. Through this process, a representative regression equation is obtained:

\[ y = 1016.20 + 48.36R - 0.33M, \quad r^2 = 0.73 \quad (4) \]


In 1961-1979, there was no coal mining and few SWCM in the Kuye River basin, hence the impact of human activities on its runoff were insignificant. Water consumption (for production and living) was basically constant, and the runoff was mainly affected by natural factors such as rainfall and temperature. Hence it can be perceived as the base period. Therefore, the runoff equation can be fitted as a function of rainfall R, given by,

\[ Y_1 = 5536.45 + 157.80R, \quad r^2 = 0.67 \]

Apart from rainfall and temperature in 1980-1998, the impacts of SWCM, water consumption and coal mining on Kuye River runoff are considerable. The effect of SWCM on river runoff may vary in different areas. The general perception is that SWCM plays an important role on the Yellow River runoff reduction. The rationale is that when there is no vegetation in the loess steep slope area, rainfall would quickly form runoff that rushes into the river. With the water conservation measures, a fraction of the rainfall would be trapped and absorbed by vegetation cover. Based on these three independent variables (rainfall R, temperature T and SWCM area W), the optimized regression equation is formulated as:

**Table 1. Stepwise regression computation sheet in 1999-2016.**

<table>
<thead>
<tr>
<th>Model</th>
<th>Partial correlation coefficient</th>
<th>Standard coefficient</th>
<th>t</th>
<th>P_1</th>
<th>P_2</th>
<th>r^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>α_0 1,440.60</td>
<td>0.43</td>
<td>0.67</td>
<td>0.00</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>T 39.20</td>
<td>0.81</td>
<td>5.43</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>α_0 1,016.20</td>
<td>0.33</td>
<td>0.75</td>
<td>0.00</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R 48.36</td>
<td>0.99</td>
<td>5.97</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>M -0.33</td>
<td>-0.33</td>
<td>-1.98</td>
<td>0.07</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 2. Stepwise regression computation sheet in 1961-2016.**

<table>
<thead>
<tr>
<th>Model</th>
<th>Partial correlation coefficient</th>
<th>Standard coefficient</th>
<th>t</th>
<th>P_1</th>
<th>P_2</th>
<th>r^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>α_0 75,323.53</td>
<td>18.71</td>
<td>0.00</td>
<td>0.00</td>
<td>0.57</td>
<td></td>
</tr>
<tr>
<td></td>
<td>W -16.31</td>
<td>-0.76</td>
<td>-8.49</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>α_0 28,400.10</td>
<td>3.93</td>
<td>0.00</td>
<td>0.00</td>
<td>0.78</td>
<td></td>
</tr>
<tr>
<td></td>
<td>W -17.21</td>
<td>-0.80</td>
<td>-12.34</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>R 112.04</td>
<td>0.46</td>
<td>7.09</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>α_0 17,621.28</td>
<td>2.50</td>
<td>0.02</td>
<td>0.00</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td></td>
<td>W -12.73</td>
<td>-0.59</td>
<td>-7.45</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>R 131.76</td>
<td>0.54</td>
<td>8.78</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>M -1.34</td>
<td>-0.32</td>
<td>-3.83</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>α_0 77,405.32</td>
<td>2.88</td>
<td>0.01</td>
<td>0.00</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td></td>
<td>W -9.87</td>
<td>-0.46</td>
<td>-4.79</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>R 125.36</td>
<td>0.51</td>
<td>8.53</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>M -1.14</td>
<td>-0.27</td>
<td>-3.30</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T -7,270.14</td>
<td>-0.21</td>
<td>-2.30</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Period from 1961 to 2016

From the stepwise regression analysis, it can be deduced that the main influence factors of Kuye River runoff are different in each period. However, the runoff is primarily a function of rainfall \( R \), followed by the SWCM area \( W \) and temperature \( T \) from 1980 to 1998; and coal mining \( M \) in the recent period. To understand the extent of these independent variables on the Kuye River runoff for 1961-2016, we need to conduct another stepwise regression analysis.

The outputs of regression analysis are presented in Table 2. The quantitative contribution of four independent variables (precipitation \( R \), temperature \( T \), the SWCM area \( W \) and coal mining \( M \)) are evaluated. The fitted regression equation is:

\[
Y_2 = 80920.52 + 106.87R - 6647.26T - 9.99W, \quad r^2 = 0.76
\]

A comparison between the measured runoff and estimated runoff (using Equation (5)) of the Kuye River in 1961 to 2016 is shown in Fig. 3. Overall, the simulated runoff is close to the measured values over the period 1961-1998. However, the simulated runoff fluctuates more in 1999-2016, mainly because coal mining in the Kuye River basin has rapidly risen during this period. Since selecting a suitable model to simulate the influence degree of coal mining on river runoff is difficult, more runoff data would be needed to verify it.

Discussion

In many locations, climate change may significantly reduce river runoff and urban water supplies [27-29]. Many recent studies have also shown that human activities, including land use change [30], agricultural irrigation, reservoir storage, coal mining, industrial and domestic water consumption have become the main reasons for effects on surface water resources

\[
Y_3 = 77408.32 + 125.36R - 7270.14T - 9.87W - 1.14M, \quad r^2 = 0.85
\]

The Weights of Influence Factors

For the period 1961-1979, a t-test is performed on each independent variable. The \( P_1 \) values for the variables \( R \) and \( M \) are all less than 0.10; and \( P_1 \) for \( U \), \( T \) and \( W \) are all greater than 0.10. Thus the equation only keeps the variables \( R \) and \( M \). From F-test, the \( P_2 \) value is far below 0.01, so the equation is significant. The weight of the variable \( M \) is -25%, where the absolute value is only one-third of the rainfall \( R \). However, the former weight is negative while the latter weight is positive. While runoff Equation (4) can better simulate the Kuye River runoff from 1999 to 2016, the data series is quite short to yield satisfactory validation. Hence long periods of relevant data are needed to test the significance of various factors (especially coal mining) on the river runoff.

From 1961 to 2016, the t-test and F-test were performed on each variable and equation. The variables \( R \), \( T \), \( W \) and \( M \) are kept in the equation, while water consumption \( U \) is discarded. The outcomes of these statistical tests demonstrated that the equation is statistically significant. The weights of the variables \( R \), \( T \), \( W \) and \( M \) are 35%, -15%, -31%, and -19%, respectively. The natural factors \( R \) and \( T \) contribute half of the weights, and the weight of SWCM area \( W \) is greater than that of coal mining \( M \). Among these variables, only the rainfall \( R \) has a positive impact on river runoff, while the rest indicate negative impact. From the long-term study we found that the weight of human activity factor has increased significantly, surpassing those of natural factors in recent years. Although other factors also have some impact on Kuyerunoff, these four factors (\( R \), \( T \), \( W \), \( M \)) are dominant.

Runoff Reduction in Different Periods

Relative to the base period 1961-1979, the changes in average values of each of the influence factors are calculated for the two recent periods (1980-1998, 1999-2016). The changes could be multiplied by the runoff coefficient from equation (5), and hence we get

Fig. 3 The stimulated values and measured values of Kuye River runoff in 1961-2016 (2).
the runoff reduction caused by each influence factor. By aggregating up the simulated runoff reduction caused by various factors in the same period, it should match the actual runoff reduction in the Kuye River. The calculation process is shown in Table 3. Although the weight of each variable has been calculated, the variation value of each variable is different. Hence the actual runoff reduction proportion is inconsistent with the weight.

According to the results in Table 3, the average runoff of Kuye River in 1980-1998 dropped by 21,569 × 10^4 m^3 from the base period (1961-1979). It was slightly less than the actual decrease. The SWCM accounted for nearly two-thirds of all runoff reduction in this period, followed by temperature and coal mining (to a lesser extent). During the coal mining period (1999-2016), the average runoff of the Kuye River decreased by 52,992 × 10^4 m^3 relative to the base period, which was also slightly less than the actual runoff reduction. It should be noted that during 1999-2016, the river runoff was 1,700 × 10^4 m^3 above that of the base period. It was the outcome of an increase in rainfall, partially offset by other factors. Among them, the SWCM reduced the runoff by 30,468 × 10^4 m^3 (the absolute value increased, but the proportion reduced). Secondly, the temperature led to the runoff reduction up to 10,905 × 10^4 m^3, as about the same proportion of the previous period. The coal mining is more than 50 times that of the previous period, so it caused runoff reduction by as much as 13,319 × 10^4 m^3, accounting for 25%. In this period, the runoff reduction due to natural factors such as rainfall and temperature was 9,205 × 10^4 m^3 (the absolute value was slightly higher, but the proportion was only half of the previous period). However, the runoff reduction due to coal mining was 40 times that of the previous period.

**Conclusions**

This study investigated the significance of different factors on river runoff and the response relationship between each factor and runoff. The overall conclusions follow.

The key influence factors on the Kuye River runoff in 1999-2016 are rainfall and coal mining, while those in 1961-2016 are rainfall, temperature, SWCM, and coal mining. From 1999 to 2016, the average runoff of the Kuye decreased by 52,992 × 10^4 m^3 relative to the base period (1961 to 1979). The rainfall caused the runoff to increase by 1,700 × 10^4 m^3, which is in contrast to the previous period (from 1980 to 1998). The SWCM caused the runoff to be reduced by 30,468 × 10^4 m^3, but the proportion is less than the previous period. The runoff reduction caused by temperature rise is also as high as 10,905 × 10^4 m^3, and the proportion does not change much. The runoff reduction caused by coal mining is as high as 13,319 × 10^4 m^3, with the proportion increased significantly to 25%.
Based on the stepwise regression method, rainfall is the main influence factor of the Kuye River runoff for all periods. The SWCM and temperature rise caused the runoff to decrease in the medium-term (from 1980 to 1998), and the rapid increase in coal mining activities has a larger impact on the runoff more recently (from 1999 to 2010). These results are generally consistent with the actual situation. This study demonstrated that the required data are easy to gather; the analysis process is simple and easy to operate; and it can be accomplished using the IBM SPSS software. Therefore, a further extension of the current methodology to other coal mining areas might provide helpful reference for the ecological restoration of river ecosystems in those areas.

Due to data scarcity, this paper only considered the influence of rainfall, temperature, SWCM, water consumption and coal mining on the Kuye River runoff. The rainfall increase in 1999-2016 led to the runoff increase, so the influence weight of coal mining accounted for only one-third of the rainfall. In a future study we plan to consider longer time series data and more influence factors.

Acknowledgements

This research was sponsored by the National Natural Science Foundation of China (No. 51969031), Shaanxi Education Department Special Scientific Research Funds (No.18JK0905), Key Laboratory of Water Resources Protection and Efficient Utilization in Yulin (No. 50502), Shaanxi Key Laboratory of Ecological Restoration in Shanbei Mining Area opening funds (No.KFXM201810), and Yulin University PhD scientific research funds (No.16GK07).

Conflict of Interest

The authors declare no conflict of interest.

References

River in northern Shaanxi, Arid Land Geography, 39, 246, 2016 [In Chinese].