Partial least-squares regression for linking land-cover patterns to soil erosion and sediment yield in watersheds

Z.H. Shi\textsuperscript{a,b,*}, L. Ai\textsuperscript{b}, X. Li\textsuperscript{b}, X.D. Huang\textsuperscript{b}, G.L. Wu\textsuperscript{a}, W. Liao\textsuperscript{c,d}

\textsuperscript{a} State Key Laboratory of Soil Erosion and Dryland Farming on the Loess Plateau, Institute of Soil and Water Conservation, Chinese Academy of Sciences, Yangling, Shaanxi 712100, China
\textsuperscript{b} College of Resources and Environment, Huazhong Agricultural University, Wuhan 430070, China
\textsuperscript{c} Huazhong Normal University, Wuhan 430079, China
\textsuperscript{d} The Center of Soil and Water Conservation Monitoring, Hubei Province, Wuhan 430071, China

\textbf{Article info}

Received 6 May 2013
Received in revised form 13 June 2013
Accepted 15 June 2013
Available online 27 June 2013
This manuscript was handled by Geoff Syme, Editor-in-Chief

\textbf{Keywords:}
Landscape metrics
Co-dependent data
Soil erosion
Sediment yield
Sediment delivery ratio

\textbf{Summary}

There are strong ties between land cover patterns and soil erosion and sediment yield in watersheds. The spatial configuration of land cover has recently become an important aspect of the study of geomorphological processes related to erosion within watersheds. Many studies have used multivariate regression techniques to explore the response of soil erosion and sediment yield to land cover patterns in watersheds. However, many landscape metrics are highly correlated and may result in redundancy, which violates the assumptions of a traditional least-squares approach, thus leading to singular solutions or otherwise biased parameter estimates and confidence intervals. Here, we investigated the landscape patterns within watersheds in the Upper Du River watershed (8973 km\textsuperscript{2}) in China and examined how the spatial patterns of land cover are related to the soil erosion and sediment yield of watersheds using hydrological modeling and partial least-squares regression (PLSR). The results indicate that the watershed soil erosion and sediment yield are closely associated with the land cover patterns. At the landscape level, landscape characteristics, such as Shannon’s diversity index (SHDI), aggregation index (AI), largest patch index (LPI), contagion (CONTAG), and patch cohesion index (COHESION), were identified as the primary metrics controlling the watershed soil erosion and sediment yield. The landscape characteristics in watersheds could account for as much as 65% and 74% of the variation in soil erosion and sediment yield, respectively. Greater interspersion and an increased number of patch land cover types may significantly accelerate soil erosion and increase sediment export. PLSR can be used to simply determine the relationships between land-cover patterns and watershed soil erosion and sediment yield, providing quantitative information to allow decision makers to make better choices regarding landscape planning. With readily available remote sensing data and rapid developments in geographic information system (GIS) technology, this practical and simple PLSR approach could be applied to a variety of other watersheds.

\textcopyright 2013 Elsevier B.V. All rights reserved.

1. Introduction

Soil erosion depends on the interaction of different physical and anthropogenic factors, including soil properties, topography, climatic characteristics, land use and its management. Soil properties and topography are relatively constant in the short term, and changes in land use and climate features are the dominant variables (Wei et al., 2007). Soil erosion is largely determined by the absence of protective land cover, whereas sediment export to rivers is determined by on-site sediment production and the connections between sediment sources and rivers (Bakker et al., 2008). The latter factor is also a function of land use because hydrological processes and sediment transport capacity vary for different types of land cover (Van Oost et al., 2000). The types of land cover are closely related to the characteristics of human activities, which in turn determine the anthropogenic substances carried into erosion systems through soil detachment, runoff process, sediment transport, and deposition. Previous studies have often focused on the composition of land cover within the watershed to explain variations in soil erosion and sediment yield (Phippen and Wohl, 2003; Pelacani et al., 2008; Casali et al., 2010; Feng et al., 2010; Nie et al., 2011; Yan et al., 2013). From a landscape ecology perspective, land cover patterns within the watershed may play a critical role in determining hydrological connectivity processes, the temporal storage of runoff and run-on,
and sediment delivery. Therefore, understanding the relationships between land cover patterns and erosion processes is of practical importance for watershed planning and management.

Within the last few decades, the spatial configuration of land cover has become an important aspect of studies of geomorphological processes related to erosion (Valentini et al., 1999; Imeson and Prinsen, 2004; Ludwig et al., 2005; Puigdefábregas, 2005; Fu et al., 2009). The spatial configuration, including the extent, distribution, and intensity of land uses, is an important factor in understanding the erosion processes linking land use and sediment export. Bartley et al. (2006) reported large differences in the runoff and sediment yield measurements from three hill-slopes with similar total plant cover but different plant cover arrangements. The location of medium-to-high cover patches close to the bottom of hill-slopes, together with other co-occurring attributes such as topography and soil characteristics, were thought to determine the hydrologic response. Jordan et al. (2005) investigated the impact of historical land use changes on soil erosion and sediment transport using a modeling approach. Their results demonstrated that land use changes introduced by property ownership and agricultural changes have decreased sediment production in the catchment while increasing the relative sediment export to downstream areas within the basin. This increase in sediment export is due to changes in the land cover pattern that allow more sediment to be transported to the river system. Ziegler et al. (2007) quantified the effects of patchiness and the optimized patch arrangement of different land cover types to reduce runoff in two catchments in Vietnam. Independent of the modeled event size, increasing patchiness and an optimized patch arrangement, which maximizes the number of transfers between patches with different hydrologic behaviors, substantially reduced catchment runoff without changing the proportions of different land uses.

Land cover patterns can be described by using a variety of landscape metrics, and many different landscape metrics have been developed (Chen et al., 2008). As one of most important processes, soil erosion within a heterogeneous landscape has received significant attention. Not surprisingly, many studies have used multivariate regression techniques (a landscape metric approach) to examine the soil erosion and sediment yield response to different land cover patterns (Xiao and Ji, 2007; Lee et al., 2009; Ouyang et al., 2010; Huang et al., 2011; Memarian et al., 2012). Despite the great potential of these landscape metric approaches, they also present particular analytical challenges. Many landscape metrics are highly correlated, which can result in redundancy (Hargis et al., 1998). The use of correlated parameters violates the assumptions of the traditional least-squares approach, thereby leading to singular solutions or otherwise biased parameter estimates and confidence intervals. Therefore, significant caution must be exercised when using landscape metrics, particularly when establishing relationships between landscape patterns and ecological processes using these metrics (Corry and Nassauer, 2005; Chen et al., 2008; Yang et al., 2012). The inherent limitations of traditional multivariate regression approaches in handling multi-collinear and noisy data can be overcome by applying techniques based on multivariate statistical projection, such as principal component regression (PCR) and partial least-squares regression (PLSR). PLSR is a recent technique that combines features from the principal component analysis (PCA) technique and the multiple linear regression technique and generalizes them (Abdi, 2010). The PLSR technique handles highly correlated noise-corrupted data sets by explicitly assuming the dependency between variables and estimating the underlying structures, which are essentially linear combinations of the original variables (Carrascal et al., 2009; Singh et al., 2013).

We previously developed quantitative relationships between sediment yield and the compositions of land cover types within the Upper Du River watershed in China (Yan et al., 2013). However, spatial land cover patterns can exert a significant influence on runoff and sediment transport at different scales (Fu et al., 2009). The quantification of the effects of spatial land cover patterns on sediment yield is important to develop effective soil erosion control through spatial planning for land use. Therefore, for this study, the Upper Du River watershed was chosen as the case study area. The SWAT model and partial least-squares regression were used to explore the relationship between the watershed landscape characteristics and the soil erosion and sediment yield. The objectives of this study were to (i) investigate the relationship between the spatial configurations (landscape metrics) of land cover and the watershed soil erosion and sediment yield using partial least-squares regression at the sub-basin scale and (ii) model the sediment delivery ratio (SDR) using landscape metrics at the watershed level.

2. Materials and methods

2.1. Study area

The case study area is the Upper Du River watershed (31°30'N–32°27'N, 109°11'E–110°25'E), which is located in the Danjiangkou Reservoir Area and has a total drainage area of 8973 km² (Fig. 1). The Danjiangkou Reservoir on the Han River, the largest tributary of the Yangtze River, is the water source for the Middle Route Project under the South-to-North Water Transfer Scheme, and it supplies 13.8 billion m³ of water annually to the North China Plain. This area has a typical subtropical monsoon climate. The average annual precipitation is approximately 973 mm, which mainly falls during the monsoon season (June–October), and the average yearly temperature is approximately 14.3 °C. The topography of the watershed is characterized by mountain ranges, steep slopes, and deep valleys, with altitudes of 220–2833 m. Forest is the principal land cover type in this watershed. The villages, small towns, and agricultural land are concentrated along the river. The major crops are corn (Zea mays L.) and wheat (Triticum aestivum L.).

2.2. Land-cover data and pattern analysis

Land cover maps for the years 1978, 1987, 1999, and 2007 were obtained from the Changjiang River Water Resources Commission (Fig. 2). The land use maps were generated from Landsat images, which were obtained from the Landsat archive (http://glovis.usgs.gov/). These images included Multi Special Scanner imaging for 1978, Thematic Mapper imaging for 1987 and 2007, and Enhanced Thematic Mapper imaging for 1999. The accuracy of the land cover types in the area was assessed before the data were released. The four-year land cover maps were prepared for use in the SWAT model and to calculate landscape metrics. Table 1 lists the information on the different land-cover types.

Although many landscape metrics have been proposed and utilized to quantify landscape patterns or characteristics, not all of these landscape metrics are informative for delineating patterns because of the multi-collinearity among the metrics and the erratic behavior of some of the metrics across different scales, such as the minimum mapping unit or the extent of the map (McGarigal et al., 2012). In this paper, 15 metrics were analyzed to describe the landscape features (Table 2). These selected metrics have been commonly used in previous studies of the role of land cover patterns in soil erosion or sediment yield (e.g., Ouyang et al., 2010; Huang et al., 2011; Wang et al., 2011; Memarian et al., 2012). These metrics reflect the major components of land use planning (i.e., shape, distance, connectivity, and diversity), which enable the results of this study to be readily applied to landscape and land use planning. Furthermore, these metrics are important in understanding the
ecological processes and the effect of human activities in a landscape. To calculate the metrics, we used the program FRAGSTATS 4.0 (McGarigal et al., 2012), which is a widely accepted tool for landscape metrics quantification.

2.3. SWAT databases and validation

The data, including topography, land cover, soil type, and hydro-meteorological characteristics, were prepared for analysis. The current version, SWAT2009, was used to develop the necessary databases. The topographical information used in this study was derived from a Digital Elevation Model (DEM) with a resolution of 25 m × 25 m, which was purchased from the National Geomatics Center of China. The soil data, including a soil type map (1:100,000) and information on related soil properties, were obtained from the Soil Survey Office of Hubei Province. Watershed management information was added to improve the modeling accuracy. The watershed climatic features were simulated based on daily historical monitoring data from nine weather stations from the period of 1965 through 2010 (Fig. 1). The daily averages of stream flow and sediment yield at the Zhushan gauging station (the watershed outlet) were used to calibrate and validate the SWAT model.

The Upper Du River watershed was divided into 107 sub-basins using the SWAT model. By using historical monitoring data for stream flow and sediment yield, the dominant factors...
were calibrated and the SWAT model was validated. The calibration and validation simulation periods ran from January 1971 to December 1980 and from January 1981 to December 1990, respectively. After performing a sensitivity analysis, 10 principal factors associated with stream flow were determined. Next, the five indices related to sediment yield were calibrated and validated. The coefficient of determination ($R^2$) value was calculated to evaluate the relationship between the measured and the modeled data for the stream flow and sediment yield, and the Nash–Sutcliffe efficiency ($E_{NS}$) was used to compare the modeled values to the measured values (Nash and Sutcliffe, 1970). As a general rule, $R^2$ and $E_{NS}$ values greater than 0.5 are considered acceptable in watershed simulations (Moriasi et al., 2007). On a monthly time step, the calibrated model performed very well, yielding $R^2$ and $E_{NS}$ values of 0.94 and 0.88, respectively, for the stream flow and $R^2$ and $E_{NS}$ values of 0.84 and 0.67, respectively, for the sediment yield. The statistical analysis of the data demonstrated a reasonable agreement between the observed and simulated values during the validation period. The observed $R^2$ and $E_{NS}$ values were 0.92 and 0.87, respectively, for the stream flow, and the observed $R^2$ and $E_{NS}$ values were 0.81 and 0.64, respectively, for the sediment yield. The detailed procedure for the calibration and validation was reported previously (Yan et al., 2013).

### 2.4. Partial least-squares regression

The PLSR technique is used to determine the relationship between two sets of variables, the matrix $X_{m×n}$, which consists of $m$ variables (columns) and $n$ objects (rows), and a response vector $y_{n×1}$. Similar to principal component analysis (PCA), PLSR strives to identify a few linear combinations (components or factors) of the original $x$-values that describe most of the inherent variable information in $y$. In contrast to PCA, only the most important linear combinations are used in the regression equation in PLSR, which is achieved mathematically by maximizing the covariance between $y$ and all possible linear functions of $x$. Details on the theory,
Table 2
Landscape metrics used in the present study (McGarigal et al., 2012).

<table>
<thead>
<tr>
<th>No.</th>
<th>Metrics</th>
<th>Abbreviation</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Patch density</td>
<td>PD</td>
<td>Number of patches per unit area (number per 100 ha)</td>
</tr>
<tr>
<td>2</td>
<td>Largest patch index</td>
<td>LPI</td>
<td>Percentage of the landscape in the largest patch (unit: %)</td>
</tr>
<tr>
<td>3</td>
<td>Edge density</td>
<td>ED</td>
<td>Total length of all edge segments per hectare for the considered landscape (unit: m/ha)</td>
</tr>
<tr>
<td>4</td>
<td>Landscape shape index</td>
<td>LSI</td>
<td>A standardized measure of the total edge or edge density that adjusts for the size of the landscape (unit: none)</td>
</tr>
<tr>
<td>5</td>
<td>Mean patch size</td>
<td>AREA_MN</td>
<td>Average area of a patch for a particular class of land cover (unit: ha)</td>
</tr>
<tr>
<td>6</td>
<td>Mean shape index</td>
<td>SHAPE_MN</td>
<td>Patch perimeter divided by the minimum perimeter possible for a maximally compact patch (unit: none)</td>
</tr>
<tr>
<td>7</td>
<td>Mean perimeter-area ratio</td>
<td>PARA_MN</td>
<td>Mean perimeter-to-area ratio of a particular class of land cover (unit: none)</td>
</tr>
<tr>
<td>8</td>
<td>Mean Euclidian nearest-neighbor distance</td>
<td>ENN_MN</td>
<td>Distance to the nearest neighboring patch of the same land use type based on the edge-to-edge distance (unit: m)</td>
</tr>
<tr>
<td>9</td>
<td>Perimeter-area fractal dimension</td>
<td>PAFRAC</td>
<td>An index of the patch shape complexity across a wide range of spatial scales (unit: none)</td>
</tr>
<tr>
<td>10</td>
<td>Interspersion and juxtaposition index</td>
<td>IJI</td>
<td>Based on patch adjacencies, directly measures the patch type interspersion or intermixing (unit: %)</td>
</tr>
<tr>
<td>11</td>
<td>Aggregation index</td>
<td>AI</td>
<td>Number of like adjacencies involving the corresponding land use type, divided by the maximum possible number of like adjacencies involving the corresponding land use type (unit: %)</td>
</tr>
<tr>
<td>12</td>
<td>Patch cohesion index</td>
<td>COHESION</td>
<td>Indicates the physical connectedness of the corresponding patch type (unit: none)</td>
</tr>
<tr>
<td>13</td>
<td>Contagion</td>
<td>CONTAG</td>
<td>Tendency of the patch types to be aggregated (unit: %)</td>
</tr>
<tr>
<td>14</td>
<td>Shannon’s diversity index</td>
<td>SHDI</td>
<td>Based on information theory, indicates the patch diversity in a landscape (unit: none)</td>
</tr>
<tr>
<td>15</td>
<td>Simpson’s diversity index</td>
<td>SIDI</td>
<td>Equal to zero when the landscape contains only 1 patch and approaches 1 as the number of different patch types increases and the proportional distribution of the area among patch types becomes more equitable (unit: none)</td>
</tr>
</tbody>
</table>

Fig. 3. Map showing the Digital Elevation Model (DEM), sub-basins, and soil types in the Upper Du River watershed.
principles, and application of PLSR can be found in the literature (Abdi, 2010), and thus only a brief description of this technique is presented here. Soil erosion, sediment yield, and the sediment delivery ratio (SDR) at the sub-basin scale were revealed to be related to land-cover patterns using the PLSR approach: the dependent variables were soil erosion, sediment yield, and SDR, and the independent variables were the landscape metrics. All of the analyses were performed using the PLSR procedure implemented in SIMCA-P (Umetrics AB, Sweden). The criterion used to determine the number of significant PLSR components was cross-validation. Within SIMCA, \( Q^2 \) (the fraction of the total variation of the dependent variables that can be predicted by a component) and \( Q^2_{cum} \) (the cumulative \( Q^2 \) over all the selected PLSR components) were computed using the following equations:

\[
Q^2 = 1.0 - \frac{PRESS}{SS}
\]

\[
Q^2_{cum} = 1.0 - \prod \left( \frac{PRESS}{SS} \right)_a \quad (a = 1, 2 \ldots m)
\]

where \( PRESS \) is the prediction error sum-of-squares, \( SS \) is the residual sum-of-squares, and \( m \) is the number of PLSR components. When \( Q^2_{cum} \) is greater than 0.5, the model is considered to exhibit good predictive ability. In addition, the root-mean-squared error-of-prediction (RMSEP) provides useful information for calibrating and developing the regression model. The RMSEP is calculated by the following:

\[
RMSEP = \sqrt{\frac{\sum_{i=1}^{n} (Y_{i,\text{predicted}} - Y_{i,\text{measured}})^2}{n}}
\]

In the PLSR modeling, the importance of a predictor for both the independent and the dependent variables is given by the variable importance for the projection (VIP). Terms with large VIP values are the most relevant for explaining the dependent variable. The regression coefficients reveal the direction and strength of the impact of each variable in the PLSR model.

Not all landscape metric variables must be included in a PLSR model. The redundant variables can lead to PLSR models with low statistical significance. Therefore, the following PLSR analysis process was followed to obtain an optimal model. First, a simulation using the PLSR model with all predictor variables was conducted. Next, a series of simulations of new PLSR models were performed in which each new PLSR analysis was performed with a variable eliminated. The new PLSR model exhibiting the largest \( Q^2_{cum} \) was selected. When there were several models with the same VIP, the model that eliminated the variable for which the VIP was the lowest in the previous model was selected. This procedure was repeated until two predictor variables remained. Finally, the model with the largest \( Q^2_{cum} \) was selected as the optimal PLSR model.

3. Results

3.1. Soil erosion and sediment yield within the watershed

The annual soil erosion and sediment yield distributions for the years 1978, 1987, 1999, and 2007 were simulated with the SWAT model (Fig. 4). The soil erosion and sediment yield of the 107 sub-basins in each of the four years were statistically analyzed (Table 3). The spatial distributions of soil erosion were similar in each of the four years, and the most intensively eroded sub-basins were found to be situated in the northern part of the study area. The southern part of the study area, however, experienced relatively slight soil erosion, perhaps because its dominant land cover type was forest in steep mountain areas. The simulations demonstrated that the watershed-averaged soil erosion rates for the years 1978, 1987, 1999, and 2007 were 9.47, 10.40, 14.14, and 7.64 t/ha/yr, respectively. The maximum sub-basin loads of soil erosion in 1978, 1987, 1999, and 2007 occurred in sub-basins 8, 8, 34, and 34, respectively, and the erosion rates in 1978, 1987, 1999, and 2007 were 45.95, 38.90, 54.75, and 43.57 t/ha/yr, respectively. Each of these sub-basins is in a mountainous area and is covered mainly by farmland (the percentage of farmland in each sub-basin is 44.6, 39.8, 48.5, and 39.6, respectively). The sediment yield of each sub-basin is the amount of sediment removed from the sub-basin. In the study area, not all of the eroded soil was transported out. The watershed-averaged sediment yields for the years 1978, 1987, 1999, and 2007 were 3.72, 5.36, 7.30, 3.69 t/ha/yr, respectively. The sediment yield of the individual sub-basins varied significantly. The maximum sediment yields in 1978, 1987, 1999, and 2007 were located in sub-basins 8, 6, 39, and 21, respectively.

The dynamics of the annual soil erosion and sediment yield were a result of the coupling of land use change and climate variability. Compared to the baseline in 1978, the mean annual soil erosion rate over the watershed was 0.93 t/ha/yr higher in 1987 and 4.67 t/ha/yr higher in 1999 (increases of 9.8% and 49.3%, respectively), whereas it was 1.83 t/ha/yr lower in the 2000s. The mean annual sediment yield also increased 96.2% from 1978 to 1999 and then remarkably decreased in 2007. The changes in soil erosion and sediment yield matched the land use dynamics, particularly for forest and farmland (Fig. 2 and Table 1). A very strong positive correlation was observed between soil erosion and sediment yield and the proportion of farmland (with \( R^2 \) values of 0.88 and 0.83, respectively). By contrast, a negative relationship between the proportion of forest and the soil erosion and sediment yield was observed (\( R^2 \) is 0.79 and 0.85, respectively).

3.2. Landscape pattern characteristics at the sub-basin scale

Table 4 shows the mean value and dispersion of the 15 selected landscape metrics. The landscape characteristics of the 107 sub-basins used in the analysis varied widely. Interestingly, the coefficient of variation (CV) of mean patch size (AREA_MN), patch density (PD), Shannon’s diversity index (SHDI), edge density (ED), Mean Euclidian nearest-neighbor distance (ENN_MN), and largest patch index (LPI) were 129.0, 74.9, 46.8, 46.5, 41.1, and 38.9, respectively, showing greater variances than other measures. PD varied greatly from 0.61 to 24.59/100 ha, and AREA_MN also varied greatly from 4.94 to 241.46 ha. The CV of aggregation index (AI), patch cohesion index (COHESION), mean perimeter-area ratio (PARA_MN) and contagion (CONTAG) were relatively small, with values of 10–20%. The correlation coefficients between individual landscape metrics are listed in Table 5. The preliminary analysis had already indicated that many of the landscape metrics were highly correlated. The preliminary results indicated that the area and edge metrics (e.g., PD, LPI, ED, and AREA_MN) are the most basic indices within the landscape metrics. These area and edge metrics correlated with many landscape metrics, such as aggregation and diversity in the landscape. The shape metrics were more independent than the other groups of metrics. Many of the diversity metrics were correlated among themselves, particularly the Shannon and Simpson metrics. However, the Shannon indices were more sensitive to the minor cover type in the landscape, while the Simpson indices were more sensitive to the common cover types. Furthermore, the Shannon indices were more sensitive to patch density, edge density, and average patch area than the Simpson indices. The diversity metrics use the proportions of the cover types and the number of classes to describe the landscape diversity, and the contagion metrics use the adjacency matrix to describe the landscape edge diversity. Therefore, the diversity indices are often correlated with the contagion indices. The diversity and contagion indices exhibit different aspects of the heterogeneity of the landscape in the study area.
3.3. Soil erosion and sediment yield for different landscape patterns

A summary of the two PLSR models constructed separately for soil erosion and sediment yield is presented in Table 6. The soil erosion model extracts three PLSR components that are relevant to 11 predictor variables (landscape metrics, Table 7). The first component accounted for 53.9% of the variance in the dataset with soil erosion (Table 6). The addition of two more components cumulatively accounted for 64.5% of the total variance in the soil erosion rates. Adding more components to the PLSR models did not substantially improve the description of the contributions to the variance (Table 6). The first component of the soil erosion model is dominated by SHDI and PD with positive PLSR weights and LPI, CONTAG, ED, and COHESION with negative PLSR weights, whereas the second component is dominated by ENN_MN, AREA_MN, SHDI, and PARA_MN on the positive side (Table 7). Although the weight values indicate how important the individual landscape metrics are for the soil erosion and sediment yield, a more convenient and comprehensive expression of the relative importance of the landscape metrics can be obtained by exploring their VIP values and their regression coefficients (RCs). In the case of soil erosion, the highest VIP value was obtained for the SHDI (VIP = 1.26; RCs = 0.438), followed by the LPI (VIP = 1.24; RCs = −0.407), CONTAG (VIP = 1.18; RCs = −0.253), AI (VIP = 1.17; RCs = −0.288), and ED (VIP = 1.05; RCs = −0.039). Soil erosion appeared to decrease with higher LPI, CONTAG, AI, and ED indices (due to the negative regression coefficients), whereas the SHDI contributed to higher soil erosion. All of the considered landscape metrics are related to soil erosion to some extent, yet only some of them have VIP > 1. Landscape metrics with VIP values below 1 are of minor importance for soil erosion. Hence, further discussion is restricted only to the predictors with VIP values >1. For the sediment yield model, the RMSEP minimum was achieved with three components, of which the first component accounted for 56.2% of the sediment yield (Table 6). The second and third components accounted for 13.2% and 4.1% of the sediment yield, respectively. The addition
of further components did not significantly improve the accounting of the factors contributing to the variance. The SHDI appears to dominate the first and second components of the sediment yield PLSR model, while LPI and CONTAG dominate on the negative side of the first component components (Table 7). COHESION and AI are positive PLSR weights with the second and third components, respectively. For this model, the highest VIP value was observed with SHDI (VIP = 1.16; RCs = 0.536), followed by LPI (VIP = 1.12; RCs = −0.549), COHESION (VIP = 1.06; RCs = −0.493), CONTAG (VIP = 1.05; RCs = −0.202), and AI (VIP = 1.02; RCs = −0.373). The $Q_{cum}^2$ values for soil erosion and sediment yield models are 0.62 and 0.72, respectively, indicating the good predictive ability and robustness of the two models. The best PLSR prediction models are also illustrated by the lowest RMSEP, e.g., soil erosion with 5.92 t/ha/yr and sediment yield with 2.62 t/ha/yr.

### 3.4. Modeling the SDR incorporated with the landscape metrics

The sediment delivery ratio (SDR) is defined as the ratio of sediment delivered at the catchment outlet to the gross soil erosion within the catchment. The quantitative estimations of SDR have important implications for the study of off-site environmental impact due to exported sediment as well as on-site erosion control. Information on the spatially distributed sediment delivery is useful in determining the relative importance of the sediment sources

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Soil erosion</td>
<td>Minimum</td>
<td>0.97</td>
<td>1.19</td>
<td>1.84</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>45.95</td>
<td>38.90</td>
<td>54.75</td>
<td>43.57</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>9.47</td>
<td>10.40</td>
<td>14.14</td>
<td>7.64</td>
</tr>
<tr>
<td></td>
<td>Standard Deviation</td>
<td>11.98</td>
<td>8.57</td>
<td>11.69</td>
<td>10.27</td>
</tr>
<tr>
<td>Sediment yield</td>
<td>Minimum</td>
<td>0.16</td>
<td>0.34</td>
<td>0.72</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>16.78</td>
<td>22.32</td>
<td>21.91</td>
<td>20.43</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>3.72</td>
<td>5.36</td>
<td>7.30</td>
<td>3.69</td>
</tr>
<tr>
<td></td>
<td>Standard Deviation</td>
<td>4.31</td>
<td>4.55</td>
<td>5.34</td>
<td>5.18</td>
</tr>
</tbody>
</table>
value of model 4 is 0.55, indicating the good predictive ability and robustness of the model. Validation of the best PLSR prediction model is illustrated by comparison of the plots of the actual and predicted SDR values (Fig. 5). The predictive relationship was reasonable for the SDR values (Fig. 5) with a data range between 0.209 and 0.557. Model 4 extracted two PLSR components that define the scores, they can be used to describe the quantitative relationship between the predictors and the response. SDR is a scaling factor that relates the sediment availability and deposition at different spatial scales (De Vente et al., 2007). Therefore, the sub-basin area (AREA) is added as a predictor in addition to the selected landscape metrics. A PLSR regression process for the 107 sub-basins in the four simulated years resulted in the following model 4 as the optimal equation:

\[
SDD = 2.185 - 0.369 \times X_{\text{AREA}} + 0.048 \times X_{\text{PD}} - 0.142 \times X_{\text{CONTAG}} - 0.219 \times X_{\text{COHESION}} + 0.141 \times X_{\text{SHDI}}
\]

\[(Q_{\text{cum}}^2 = 0.55, \ R^2 = 0.58, \ \text{RMSEP} = 0.084)\]  

The sur-face value of model 4 is 0.55, indicating the good predictive ability and robustness of the model. Validation of the best PLSR prediction model is illustrated by comparison of the plots of the actual and predicted SDR values (Fig. 5). The predictive relationship was reasonable for the SDR values (Fig. 5) with a data range between 0.209 and 0.557. Model 4 extracted two PLSR components that define the scores, they can be used to describe the quantitative relationship between the predictors and the response. SDR is a scaling factor that relates the sediment availability and deposition at different spatial scales (De Vente et al., 2007). Therefore, the sub-basin area (AREA) is added as a predictor in addition to the selected landscape metrics. A PLSR regression process for the 107 sub-basins in the four simulated years resulted in the following model 4 as the optimal equation:
are relevant to six predictor variables (Table 8). The factors governing SDR can be interpreted by the VIP and PLSR weights of the variables included in model 4. The sub-basin area appears to dominate the first and second components of the PLSR model for SDR, and it has the highest VIP (1.426). As expected, in the sub-basin area, COHESION and CONTAG are associated with lower SDR (negative regression coefficient in model 4), while the SHDI and PD contribute to a higher SDR. To validate model 4, approximately 70% of the 107 sub-basins were randomly selected and used to develop a new PLSR model using the same descriptors. Then, the new model was used to predict the SDR values of the remaining 30% of the sub-basins. The procedure was repeated 10 times, and the final results were as follows: average $Q^2_{cum} = 0.52$, average correlation coefficient for the SDRs excluded from the models, $R^2 = 0.54$. These results indicate the good predictive ability and robustness of model 4.

### 4. Discussion

From a landscape ecology perspective, the landscape structure of the watershed determines the material and energy flow, such as solar radiation, temperature, evapotranspiration, surface runoff, nutrients, soil erosion, and sediments (Forman, 1995). It is conceivable that land cover patterns affect the runoff processes that carry eroded soils into streams and rivers, while the proportions of certain land cover types determine the erosion rates within watersheds. Thus, given the same proportions of certain land cover types, the spatial patterns of land cover define, at least partially, the characteristics of erosion and sediment transport processes by accelerating or reducing the runoff speed. He et al. (2000) suggested that nonpoint source pollution loading, including erosion, sediment, and nutrient runoff, in rural watersheds is closely tied to the landscape structure.

Forman (1995) recognized three landscape components: patches, corridors, and the matrix. The landscape matrix is defined as the most extensive and connected habitat of a landscape, and as such, it can potentially have great influence on the dynamics of the species in the landscape. Within the study watershed, forest was the dominant landscape and constituted 70.9%, 70.4%, 69.3%, and 76.2% of the total area in 1978, 1987, 1999, and 2007, respectively (Table 1). At the landscape level, the presence of many small patches of various land cover types is more likely to accelerate soil erosion and to increase sediment export. In this study, soil erosion and sediment yield were positively associated with PD and ED and were negatively related to LPI (Table 7). By definition, PD and ED will increase while LPI will decrease when many small patches of land cover types occupy the watershed (McGarigal et al., 2012). Both PD and ED reflect the degree of forest fragmentation. Thus, highly fragmented forests may not function effectively to increase infiltration and decrease surface runoff and erosion from agricultural areas. However, Ouyang et al. (2010) reported that properly implementing fragile landscape status can prevent soil erosion by infiltrating and decrease surface runoff and erosion from agricultural areas. Ouyang et al. (2010) reported that properly implementing fragile landscape status can prevent soil erosion by infiltrating and decrease surface runoff and erosion from agricultural areas.
dynamics of the relationships between land-cover patterns and the watershed soil erosion and sediment yield. The COHESION and AI metrics reflect the physical connectedness and aggregation of land covers within watersheds, respectively, and have higher values when land cover types are more clumped and aggregated (McGarigal et al., 2012). Here, both the COHESION and AI metrics consistently exhibited negative relationships with the watershed soil erosion and sediment yield. Thus, our results suggest that soil erosion and sediment yield are more likely to occur when the land cover types are scattered over the watershed area with many land cover patches. The CONTAG metric is associated with both the dispersion and interspersed of land cover types, and it is high when there are low levels of dispersion and interspersed of land cover types. The CONTAG metric value approaches 0 when land cover types are maximally disaggregated and interspersed and approaches 100 when all land use types are maximally aggregated. SHDI increases as the number of different land cover types increases (McGarigal et al., 2012). In our study, the CONTAG metric was consistently and negatively related to soil erosion and sediment yield. These results are in agreement with former reported negative relationships of the CONTAG with non-point sources of pollution in watersheds (Xiao and Ji, 2007; Lee et al., 2009). By contrast, the SHDI value of land cover types was positively correlated with soil erosion and sediment yield. This positive correlation indicates accelerating soil erosion and increasing sediment export when watersheds include many different land cover types that are small and interspersed. Thus, from the perspective of watershed management, watersheds with highly interspersed land cover types may be the worst scenario. Our previous studies indicate that erosion rates and sediment loads are related to the proportion of farmland, while forests are associated with better soil and water conservation (Table 3), suggesting that special consideration must be given to the spatial distribution and number of farmland areas (Yan et al., 2013). The negative impacts of the interspersion and diversity of land cover types within watersheds are primarily associated with human dominated land uses, such as urban area and farmland.

This study revealed a strong influence of landscape metrics on watershed erosion and sediment yield at the sub-basin scale. Although our goal was not to develop a prediction model but to identify the main controls affecting soil erosion and sediment yield at the watershed scale, we suggest that incorporating landscape metrics such as SHDI, CONTAG, and COHESION, which can be easily computed from a digital land cover map, can substantially improve the assessment of soil erosion and sediment yield in watersheds without regular monitoring at the gauging station. To accommodate the highly co-dependent data of landscape metrics, we utilized the PLSR approach in conjunction with the variable influence of the projection approach as a suitable technique. The PLSR methodology is beneficial because it enables the elimination of co-dependency among the variables and encourages a more unbiased view of the contribution of landscape metrics to soil erosion and sediment yield. Of course, there are also limitations to this approach. For example, the PLSR model only provides general insights into the relationships between the different landscape metrics and the watershed soil erosion and sediment yield, whereas more detailed information about sediment delivery pathways, sources, and sinks cannot be obtained by this methodology. The land use type had significant influence on soil erosion and sediment yield (Wei et al., 2007; Yan et al., 2013). There is need to identify the relationship between landscape metrics at class-scale and hydrological processes. Obtaining such detailed information will be the focus of future research efforts in this area.

5. Conclusions

In this study, we investigated how landscape patterns affect the watershed soil erosion and sediment yield in the Upper Du River watershed by using hydrological modeling and PLSR. The results indicate that soil erosion and subsequent sediment delivery are closely associated with the spatial configuration of the land cover types within the watershed. At the landscape level, Shannon’s diversity index (SHDI), the aggregation index (AI), the largest patch index (LPI), contagion (CONTAG) and the patch cohesion index (COHESION) were the primary landscape metrics controlling watershed soil erosion and sediment yield. Greater interspersion and higher patch numbers of land cover types may significantly accelerate soil erosion and increase sediment export. Land planners and managers may therefore need to control land cover patterns through zoning regulations and planning practices to minimize the adverse effects of land cover. PLSR analysis produced useful information for modeling the sediment delivery ratio. Our results also suggest the need for care when generalizing relationships between land cover and soil erosion and subsequent sediment delivery. That is, we must consider distinct regional or local characteristics of the landscape matrix that might affect these relationships.

In large-scale basins, measurements of sediment yield are sparse, and the majority of sub-basins are not monitored by gauges. Therefore, at this scale, modeling method appears to be the most practical approach to providing spatial information for watershed management. With the availability of remote sensing data and rapid developments in GIS technology, decision makers are able to quantify the spatial patterns of land cover over broad study areas using landscape pattern metrics. The PLSR approach provides a simple means to determine the relationships between land-cover patterns and watershed soil erosion and sediment yield, providing quantitative information that enables decision makers to make better choices regarding landscape planning.

Acknowledgments

Financial support for this research was provided by the National Natural Science Foundation of China (41071190), the National Science and Technology Supporting Programs (2012BAC06B03), the Program for New Century Excellent Talents in University (NCTE-10-0423), and the Key Project of Chinese Ministry of Education (No. 108165).

References


