Automatic recognition of loess landforms using Random Forest method

ZHAO Wu-fan [1, 2, 3] http://orcid.org/0000-0002-0265-3465; e-mail: wufan_zhao@163.com

XIONG Li-yang [1, 2, 3, 4*] http://orcid.org/0000-0001-7930-3319; e-mail: xiongliyang@njnu.edu.cn

DING Hu [1, 2, 3] http://orcid.org/0000-0002-5695-5485; e-mail: nivsop@163.com

TANG Guo-an [1, 2, 3] http://orcid.org/0000-0002-1443-6134; e-mail: tangguoan@njnu.edu.cn

* Corresponding author

1 Key Laboratory of Virtual Geographic Environment (Nanjing Normal University), Ministry of Education, Nanjing 210023, China

2 Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China

3 State Key Laboratory Cultivation Base of Geographical Environment Evolution (Jiangsu Province), Nanjing 210023, China

4 Department of Geography, University of Wisconsin-Madison, Madison 53706, USA

Abstract: The automatic recognition of landforms is regarded as one of the most important procedures to classify landforms and deepen the understanding on the morphology of the earth. However, landform types are rather complex and gradual changes often occur in these landforms, thus increasing the difficulty in automatically recognizing and classifying landforms. In this study, small-scale watersheds, which are regarded as natural geomorphological elements, were extracted and selected as basic analysis and recognition units based on the data of SRTM DEM. In addition, datasets integrated with terrain derivatives (e.g., average slope gradient, and elevation range) and texture derivatives (e.g., slope gradient contrast and elevation variance) were constructed to quantify the topographical characteristics of watersheds. Finally, Random Forest (RF) method was employed to automatically select features and classify landforms based on their topographical characteristics. The proposed method was applied and validated in seven case areas in the Northern Shaanxi Loess Plateau for its complex and gradual changed landforms. Experimental results show that the highest recognition accuracy based on the selected derivations is 92.06%. During the recognition procedure, the contributions of terrain derivations were higher than that of texture derivations within selected derivative datasets. Loess terrace and loess mid-mountain obtained the highest accuracy among the seven typical loess landforms. However, the recognition precision of loess hill, loess hill–ridge, and loess sloping ridge is relatively low. The experiment also shows that watershed-based strategy could achieve better results than object-based strategy, and the method of RF could effectively extract and recognize the feature of landforms.

Keywords: Landform recognition; Random Forest; Feature fusion; DEM; Loess landform

Introduction

Landforms are regarded as "one type of geomorphological objects" that divide the surface of the Earth into fundamental spatial entities. These entities define the boundary conditions for operative processes in geomorphology, hydrology,

ecology, and pedology (Evans 2012; Drăguţ et al. 2012; Tang 2014). Landform quantification and recognition gained popularity worldwide because of their vital significance in elucidating the formation mechanism and spatial heterogeneity of landform evolution (Cheng et al. 2011; Jasiewicz et al. 2013; Martins et al. 2016).

Previous approaches (Hammond 1964; Dikau 1991; Hervás et al. 2009) of landform recognition fairly rely on field investigations and manual interpretations from topographic maps and aerial photographs. However, these recognition processes are time consuming, labor intensive, and subjective (Galli et al. 2008). Significant benefits of digital classifications were achieved since the development of remote-sensing data interpretation and the acquisition of high-quality digital elevation models (DEMs).

Landform recognition and classification resemble human cognition to some extent; thus, scholars were inspired to transfer existing knowledge into machine-executable rule sets (Stumpf et al. 2011). Two main approaches of geomorphic feature extraction could be identified and summarized from previous studies, e.g. (1) pixel-based and (2) object-based techniques. Pixel-based approaches consider terrain derivatives as main characteristics to distinguish landform types. Several terrain derivatives were calculated in these approaches based on neighborhood analysis to build their multidimensional feature space (Wang et al. 2009; Cao et al. 2011). And then, segmentation was implemented by using unsupervised classification method, such as ISODATA (Irvin et al. 1997; Tang et al. 2003; Stepinski et al. 2006). These methods are considered simple and produce clear geographical meanings. However, pixel-based method does not sufficiently consider the topological relationships of neighborhood, embeddedness, or shape information of different objects.

Object-based image analysis (OBIA) became popular among scholars in the field of remote sensing in the past decade. Unlike traditional pixel-based methods, OBIA satisfies the conceptual model of landform objects (Blaschke 2010), which is widely used in landform classification from DEMs, such as classification of landform elements and topography (Drăguţ et al. 2006; Drăguţ et al. 2012), archaeological predictive mapping (Verhagen et al. 2012), landslide detection (Danneels et al. 2007), and land-unit delineation (Van Niekerk 2010). These applications demonstrate that OBIA is sensitive to morphological discontinuities for different types of data and test areas. However, significant shortages exist due to the geographical meaning of segmentation boundary appears to be consistently unclear. Scholars recently investigated the classification of landforms based on watershed units because of their clear geographical meanings in surface morphology and landform evolution (Huang et al. 1990; Caratti et al. 2004; Gooding 2014; Liu et al. 2015). These studies suggested that this type of unit should be used for object selection in landform recognition and classification.

Feature selection in multi-dimensional datasets is another important process in various fields, such as bioinformatics (Saeys et al. 2007) and hyperspectral remote sensing (Guo et al. 2008). Various methods, such as neural network classifier (Benediktsson et al. 1989), boosting (Freund 1996), and bagging (Breiman 1996), were proposed to improve the performance in classifying data and investigating the causal relationships of these data. These classifiers achieved good accuracy. However, large computational resources are needed for the aforementioned classifiers, and their results seem to overfit if training samples are insufficient and sensitive to any outlier in the training samples (Xu et al. 2014). Random Forest (RF) (Breiman 2001) is a well-known and an excellent ensemble method that analyzes feature selection and classification from multisource remote sensing and geographic data (Gislason et al. 2003; Pal 2005; Ham et al. 2005). The results of landform recognition largely depend on the selection of various factors and a suitable classification procedure. The RF classifier uses predictions derived from an ensemble of decision trees. Thus, the RF classifier can be considered reliable because of the high dimensionality of landform data with nonlinear and complex interactions among variables. This classifier can be also successfully used to select and rank these variables with the greatest ability to discriminate between target classes (Belgiu et al. 2016).

The current study aims to investigate the applicability and performance of the RF classifier in feature selection, and to test the feasibility of

watershed-based strategy in Loess landform recognition. Small watersheds, which have highly homogeneous topographic characteristics, were derived as analysis units based on the data of Shuttle Radar Topography Mission (SRTM) DEMs in the Loess Plateau of China. Then, terrain derivatives and texture derivatives, which provide information on the physical attributes and image properties of landform objects, respectively, were extracted to quantify the watershed characteristics. Multi-dimensional derivation selection and landform recognition were accomplished automatically by taking advantage of RF.

2 Study Site and Data

2.1 Study site

The Northern Shaanxi area, which is the core region of the Loess Plateau of China, was selected as the case area in the current study. The geocoordinates of the area are between E 107°28′ to E 111°15′ and N 35°21′ to N 39°34′ based on the World Geodetic System 1984 (WGS84), with a total area of 80,606 km². The elevation of the area ranges from 325–2375 m. The thickness of loess deposits in this area changes gradually from north to south with a range of approximately 50 m to 200 m. The average temperature per year ranges from 7°C to 12°C, and the average annual precipitation ranges from 200 mm to 500 mm. In addition, rainstorms are concentrated in the summer season, and the main vegetation cover consists of shrubs, grass, and wood forests. Dry land, accelerated soil erosion, and high sediment yield are serious problems in this area (Xiong et al. 2014; Zhu et al. 2014).

The landform of this area exhibits significant topographic variability, which corresponds to the different development stages and patterns of the Loess Plateau. Seven typical loess landform types were selected as sample areas based on the evolution and morphological characteristics. The loess landform types include: (1) loess terrace, (2) loess mid-mountain, (3) loess hill, (4) loess hill–ridge, (5) loess ridge, (6) loess broken tableland, and (7) loess tableland. The distribution of the test areas is shown in Figure 1.
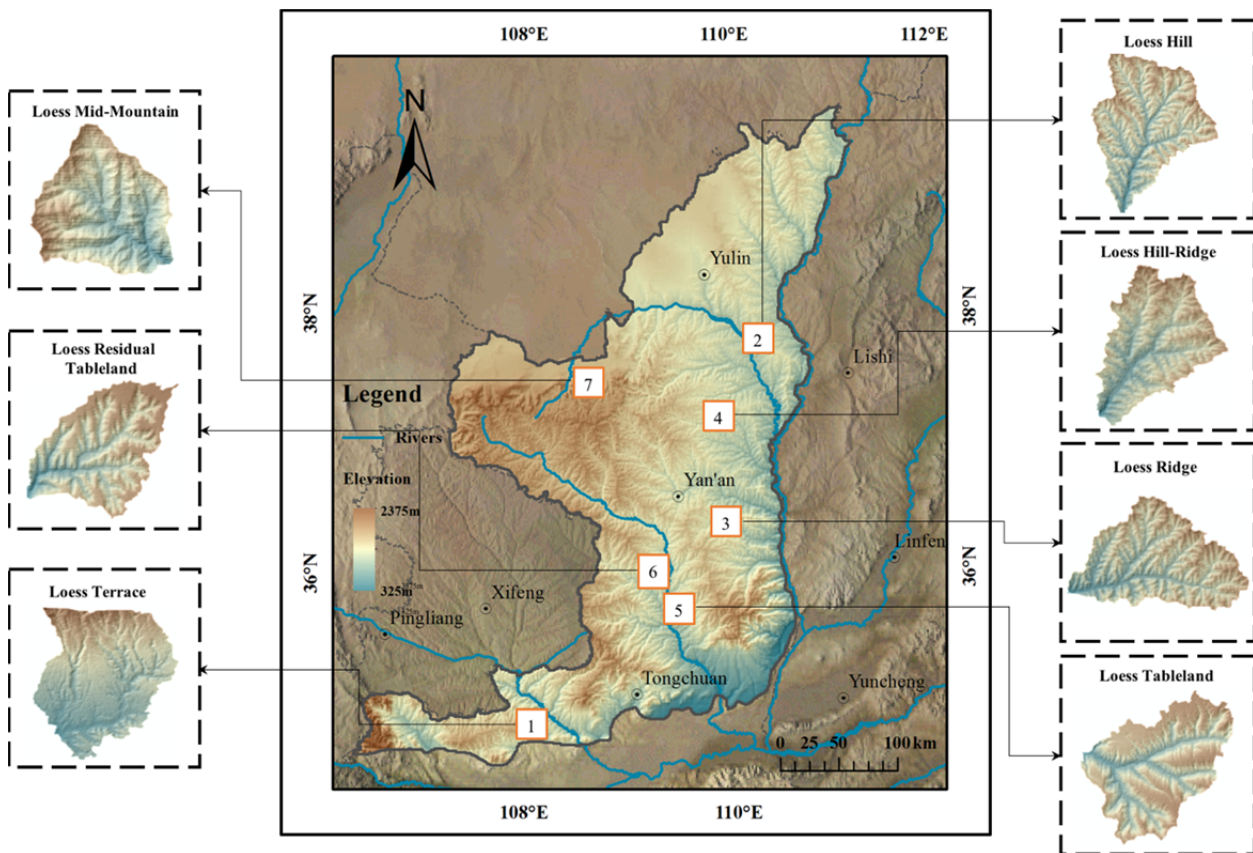
Figure 1 Distribution of test areas in the Northern Shaanxi Loess Plateau, China.

2.2 Data

The DEM used in the study was provided by the recently released SRTM DEM, which is an international project that obtains a DEM for a near-global scale from 56°S to 60°N (Nikolakopoulos 2006). In order to unify the resolution, all the DEMs were resampled using the nearest neighbor method. The final cell size of DEM is 30 m × 30 m. In addition, landform types were digitized from the 1: 1000000 Digital Geomorphology Mapping of China (Zhou et al. 2009).

3 Methods

3.1 Recognition objects

Recognition objects were determined to represent the different landform types, and these selected objects play a significant role in ensuring the quality and accuracy of the subsequent recognition process. Traditional methods obtain objects by dividing an entire test area into different rectangular blocks to guarantee that each sample has the same size (Tang et al. 2015). Numerous image segmentation algorithms based on object-oriented strategy were proposed in past decades to obtain spatial entities with a maximized internal homogeneity and a minimized external homogeneity (Blaschke et al. 2001; Blaschke 2010). However, the geographical meanings of boundaries of different spatial entities that were obtained by these methods are always fuzzy. To avoid these results, small watersheds, which are natural geographical elements with high homogeneity, were selected as the recognition units in the current study.

The maximum gradient single flow algorithm was adopted to derive watersheds (O'Callaghan et al. 1983). A previous study (Tang et al. 2015) determined that the area of a watershed significantly influences the stability of terrain statistical analysis and image texture analysis. After conducting preliminary experiments in combination with local hydrological information, threshold of 10000 was selected to delineate watersheds, and areas of the delineated watersheds that range from 30 km² to 50 km² were selected as

samples for further analysis. These selected samples were segmented based on their morphologic and attributive characteristics. Thirty-six watersheds of the seven landform areas (36×7 in total) were selected to build the training and test datasets to ensure that the recognition results are unbiased.

3.2 Feature extraction

3.2.1 Terrain derivative extraction

The terrain derivatives extracted from DEM are the most effective indexes to quantitatively describe landforms. Since Evans (1972) introduced an integrated system of geomorphometry, significant development was achieved in digital terrain analysis, particularly in developing new automatic algorithms to calculate terrain derivatives. Combined terrain derivatives can comprehensively indicate the morphological characteristics of landscape entities. However, multi-dimensional derivatives could result in redundant information because of the high correlation between each pair of derivatives. Thus, after conducting a pre-experiment on several sample areas, weak related terrain derivatives, e.g. elevation, slope, curvature, and slope of slope (SOS), are selected as basic derivatives for further landform recognition. The terrain derivatives' confusion matrix is shown in Table 1.

The statistical characteristics of terrain derivatives are significantly related with geomorphologic shapes and landscape development stages (Li et al. 2015). By conducting a statistical analysis of terrain derivative datasets, the central tendency, dispersion degree, and distribution shape of the derivatives were calculated, which can help identify internal homogeneity and external heterogeneity. Therefore, eight statistical indicators (e.g. maximum, minimum, average, standard deviation, skewness, kurtosis, amplitude, and entropy) of each terrain derivative of watersheds were calculated to construct their feature space.

3.2.2 Texture derivative extraction

Texture refers to the spatial arrangement of grayscale intensities in a selected region of a grayscale image. Terrain textures derived from

DEM can indicate the specific spatial arrangement of elevation values of an image on a macro scale. Since the 1970s, various scholars have adopted texture analysis method in landscape-pattern definition, image classification, and segmentation (Haralick et al. 1973; Ilea et al. 2011).

Grey level co-occurrence matrix (GLCM), which was presented by Haralick (1973), has been widely used in image-texture analysis. Co-occurrence matrix is the probability matrix of the transition from a pixel of *i* value to a pixel of *j* value by a certain θ orientation and *d* displacement. The principle is expressed as Eq. (1).

$$P(i,j,d,\theta)=\# \ \{(x_1,y_1)(x_2,y_2)|f(x_1,y_1)=i,f(x_2,y_2)$$
$$=j,|(x_1,y_1)\text{-}f(x_2,y_2)|=d,\angle \ x\left((x_1,y_1),(x_2,y_2)\right)=\theta\} \tag{1}$$

where # expresses the number of occurrences inside the window sizes where the intensity of pixel ranges from *i* to *j*, the first pixel is (x_1, y_1), and the second pixel is (x_2, y_2).

Four important parameters (i.e. quantization levels-*g*, displacement value-*d*, orientation value-θ, and moving window size) must be considered when the co-occurrence matrix is determined for texture analysis. Based on previous study (Liu et al. 2012),

g is set to 8 and *d* is 5. The mean value among all four directions were calculated with a 3×3 moving window.

Various derivatives of GLCM are adopted to quantify surface textures. Considering the large computational burden and strong correlations among several GLCM derivatives, a subset of eight texture measures was selected for current studies (Ulaby et al. 1986; Liu et al. 2012). These measures are angular second moment (ASM), contrast, correlation, variance, inverse difference moment (IDM), sum of mean, entropy, and variance of subtraction. The detailed formulation of these measures can be found in previous research (Haralick et al. 1973). Aside from the original DEM image, texture derivatives of hill-shading and slope were also calculated to build the feature space. Hill-shading provides a visual enhancement effect, and the slope presents the topographic relief. The features used to identify landforms in the current study are shown in Table 2.

3.3 Feature selection by random forest

Terrain derivatives and texture derivatives depict the information of surface morphology from

Table 1 The confusion matrix of terrain derivatives

Terrain derivatives	Slope	SC_depth	Terrain relief	Surface roughness	Elevation	ECV	Hillshading	SOS	Curvature
Slope	1.00								
SC_depth	0.94	1.00							
Terrain relief	0.96	0.97	1.00						
Surface roughness	0.40	0.49	0.49	1.00					
Elevation	0.11	0.07	0.11	0.05	1.00				
ECV	0.37	0.40	0.39	0.23	-0.33	1.00			
Hillshading	0.72	0.72	0.75	0.44	0.34	0.10	1.00		
SOS	-0.16	-0.15	-0.16	-0.06	-0.05	-0.02	-0.10	1.00	
Curvature	0.05	0.08	0.03	-0.04	0.02	-0.02	0.02	-0.01	1.00

Notes: SC_depth, Surface cutting depth; ECV, Elevation coefficient of variation; SOS, Slope of slope.

Table 2 Overview of factures used in landform recognition

Dataset	Factors	Count
Terrain derivatives	Elevation (Max., Ran., Ave., Var., Std., Ske., Kur. and Ent.)* Slope (Max., Ran., Ave., Var., Std., Ske., Kur. and Ent.) Curvature (Max., Ran., Ave., Var., Std., Ske., Kur. and Ent.) Slope of slope (Max., Ran., Ave., Var., Std., Ske., Kur. and Ent.)	32
Texture derivatives	Elevation (ASM, Con., Cor., Var., IDM, SOM, Ent., VOS) * Slope (ASM, Con., Cor., Var., IDM, SOM, Ent., VOS) Hillshading (ASM, Con., Cor., Var., IDM, SOM, Ent., VOS)	24

Notes: Max., Maximum; Ran., Range; Ave., Average; Var., Variance; Std., Standard deviation; Ske., Skewness; Kur., Kurtosis; Ent., Entropy; ASM, Angular Second Moment; Con., Contrast; Cor., Correlation; IDM, Inverse Difference Moment; SOM, Sum of mean; VOS, Variance of subtraction.

different aspects. Terrain derivatives reflect the formation mechanism and the physical properties of landscapes, whereas texture derivatives, which are auxiliary features, are used to enhance the description of the visual texture characteristics of landforms. In addition, the vector dimension of a training set can be increased by integrating derivatives under the fusion strategy and so as to improve the accuracy of automatic recognition. However, a simple superposition may lead to an excessive consumption of computation and storage space, thus resulting in additional data noise. The precision of recognition may decrease because of the limited samples. Consequently, adopting a reasonable derivative selection method is necessary. The major contribution derivatives of automatic landform recognition are determined in the current study by using wrapper variable selection procedure of RF based on the out-of-bag data error-estimate characteristics.

The reliability of RF, which primarily works on ensemble decision trees, is verified in ecology, remote sensing, and morphometry, (Cutler et al. 2007; Stumpf et al. 2011; Du et al. 2015). During the training process, the RF algorithm creates multiple CART-like trees. A bootstrapped sample of the original training data was trained by individual trees with a high variance. Each tree searches across the selected subset of the input variables to randomly determine a split. Each tree casts a unit vote for the most popular class during classification and assigns the respective class according to the majority of the votes. Variable assessment and selection were accomplished during the procedures. The RF algorithm diagram is shown in Figure 2.

Based on the wrapper variable selection method, the characteristics were classified according to the measured contributions of variables. By adopting reversed-sequence searching strategy, the lowest scoring derivative was deleted successively and classification accuracy was calculated iteratively. The feature set with the least number of variables and the highest classification accuracy was chosen as the selection result. The overview of the processing steps is shown in Figure 3. During sample selection, DEM and topography data were used to extract the watershed units. The terrain and texture derivatives were calculated and selected based on the variable importance measurement function of RF. Finally, landform recognition was carried out based on the optimal derivative dataset.

4   Results and Analysis

4.1 Parameters calibration

Ensemble methods are generally considered as black-box type classifiers. Hence, the calibration of the parameters of the RF model is needed to achieve an optimal recognition effect. The number of available reference samples of each landform is
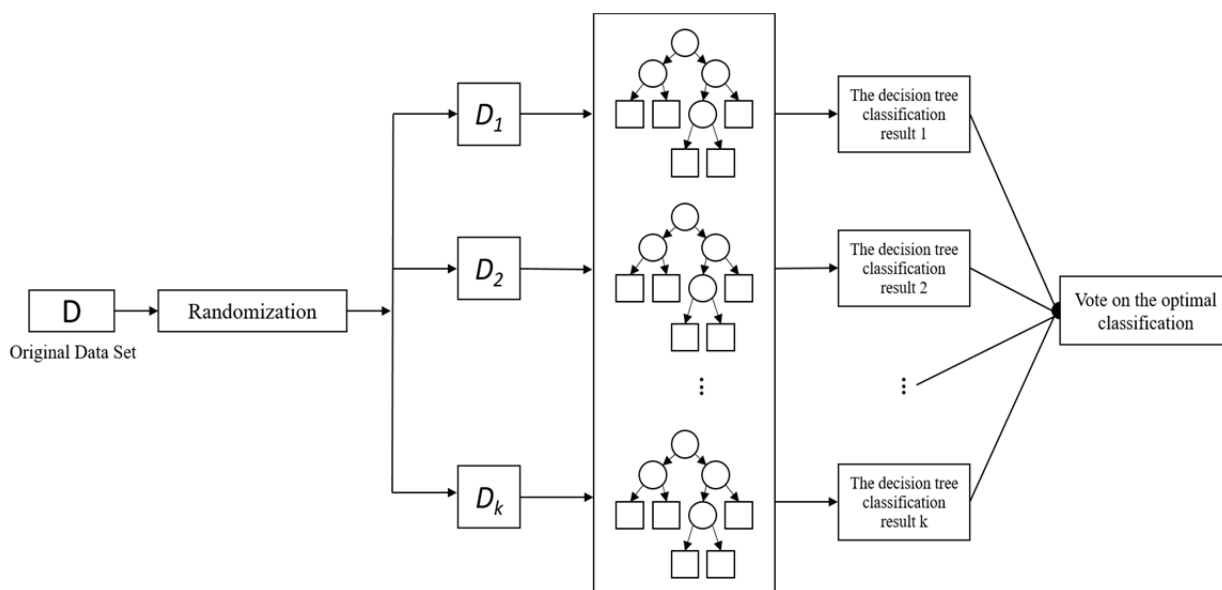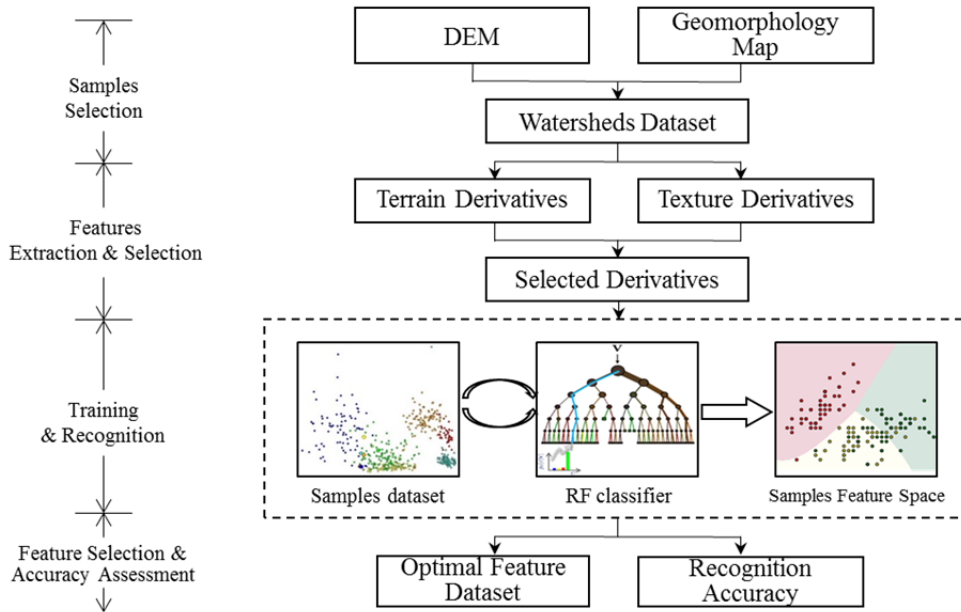

Figure 2 Random Forest algorithm diagram.

Figure 3 Overview of the processing steps.

36, and all samples are split almost evenly between the training and test samples by selecting each sample for test and the rest for training. Often, the 'N predictors' value is automatically set to the square root of the number of inputs. In addition, the number of trees and the minimum value of the parent node (split variables) are two parameters that must be defined by users. Preliminary experiments of parameter calibration are executed, and the results are shown in Table 3.

Table 3 Parameters calibration for the random forest classifier

Trees	Split variables	Runtime (min: s)	Test accuracy (%)
50	2	00:02	88.80
100	2	00:02	90.08
200	2	00:03	90.48
500	2	00:04	90.48
50	3	00:03	88.80
100	3	00:03	90.48
200	3	00:04	91.20
500	3	00:05	90.08
50	4	00:03	87.20
100	4	00:03	88.80
200	4	00:05	89.29
500	4	00:07	89.60

The result shows that the overall accuracy is insensitive to variable settings. The overall accuracy based on the original setting (200 trees and 2 split variables) was close to the most accurate result (200 trees and 3 split variables).

This property is important because the classifier can be run without human guidance. Furthermore, time complexity analysis of the algorithm is unnecessary because of the small dimension of the feature set. After completing feature extraction and selection, recognition will be conducted by the classifier.

Classification accuracy is used to estimate the proportion of correct recognition. About one third of the instances are left out of the training set and remain as out-of-bag sample (OOB). The classification decision is then made by averaging the class assignment probabilities calculated by all produced trees, and the membership class with the maximum votes will be the one that is finally selected (Rodriguez-Galiano et al. 2012). With the confusion matrix, the classification accuracy can be calculated as Eq.(2).

$$\text{Classification accuracy}=\frac{TP+TN}{N} \qquad (2)$$

Notes: TP is true positive, TN is true negative, and N is total sample.

4.2 Recognition based on dataset of single-class features

First, recognition process was implemented based on the dataset of terrain derivatives and texture derivatives. This process aims to compare the difference of the recognition results between the two types of derivatives. The recognition results

are presented in Table 4.

The table indicates that high-accuracy and terrain-derivative datasets could be identified during the recognition process. The highest accuracy was observed in the slope gradient, with a value of 81.32%. The statistical characteristics of the slope gradient, which are important factors to describe the terrain-feature information, performed well in distinguishing landform types. By contrast, the accuracy, which is based on the dataset of curvature and elevation, is relatively low. This result demonstrates that the statistical characteristics of these two factors do not exhibit evident differences in spatial distribution. In addition, the recognition result, which is based on the full-factor dataset, possesses an accuracy of 86.29%, which is higher than that of single-factor datasets.

Recognition accuracy, which is based on single-class image texture derivatives, is relatively low and has the lowest value of 54.37%. The texture feature of the DEM image itself appears low. However, hill-shade image and slope gradient image, which are two types of derived data, can enhance the information of visual texture and terrain texture of the image from different angles. Hence, these images improve the diversity of quantitative results between different areas. Thus, a high-recognition accuracy was observed in the table. The recognition accuracy of the hillside dataset and slope gradient dataset is 60.32% and 73.41%, respectively.

4.3 Recognition based on dataset of selected derivatives

Random forest offers a number of internal measures to estimate the importance of employed variables for the accuracy of a given classification. The importance of variables was estimated by looking at how much prediction error increases when OOB data for that variable is permuted while all others are left unchanged (Breiman 2002; Lombardo et al. 2015). Two datasets were integrated in the study to validate the recognition results. By adopting reversed-sequence searching strategy, the derivative collection with the least number of variables and the highest classification accuracy was screened. The best recognition accuracy achieved by the RF classifier is 92.06%

Table 4 Recognition accuracy based on dataset of single-class derivatives

Dataset	Factor	Accuracy (%)
Terrain derivatives	Elevation	72.98
	Slope gradient	81.32
	Curvature	71.83
	SOS	76.16
	Full-factor	86.29
Texture derivatives	DEM	54.37
	Hillshade	60.32
	Slope gradient	73.41
	Full-factor	75.00

Table 5 Best derivative sequence of recognition

Sorting	Name	Data set	Variable importance
1	Ave. slope	TD1	100
2	Var. slope	TD1	84.92
3	Var. curvature	TD1	84.81
4	Ent. slope	TD1	63.44
5	Ave. slope of slope	TD1	62.05
6	Ran. elevation	TD1	57.51
7	Ent. hillshading	TD2	51.62
8	VOS hillshading	TD2	49.54
9	Ent. slope	TD2	45.87
10	Var. elevation	TD1	41.2
11	Std. slope of slope	TD1	40.01
12	Std. curvature	TD1	38.09
13	Con. Hillshading	TD2	28.74
14	Max. slope	TD1	27.94
15	Ran. slope	TD1	22.12
16	Kur. slope	TD1	18.92
17	Var. elevation	TD2	17.76
18	Ave. elevation	TD1	15.45

Notes: Ave., Average; Var., Variance; Ent., Entropy; Ran., Range; VOS, Variance of subtraction; Std., Standard deviation; Con., Contrast; Max., Maximum; Kur., Kurtosis. TD1= Terrain Derivatives; TD2= Texture Derivatives.

with the best collection composed of 18 derivatives. The sequence of derivatives is shown in Table 4. Compared with the results conducted in Section 4.2, it indicates that the recognition results using the selected fusion derivatives are better than the results based on the single-class features datasets.

Table 5 suggests that the texture derivatives did not contribute as much as the terrain derivatives to the recognition process. The recognition accuracy of each landform type using
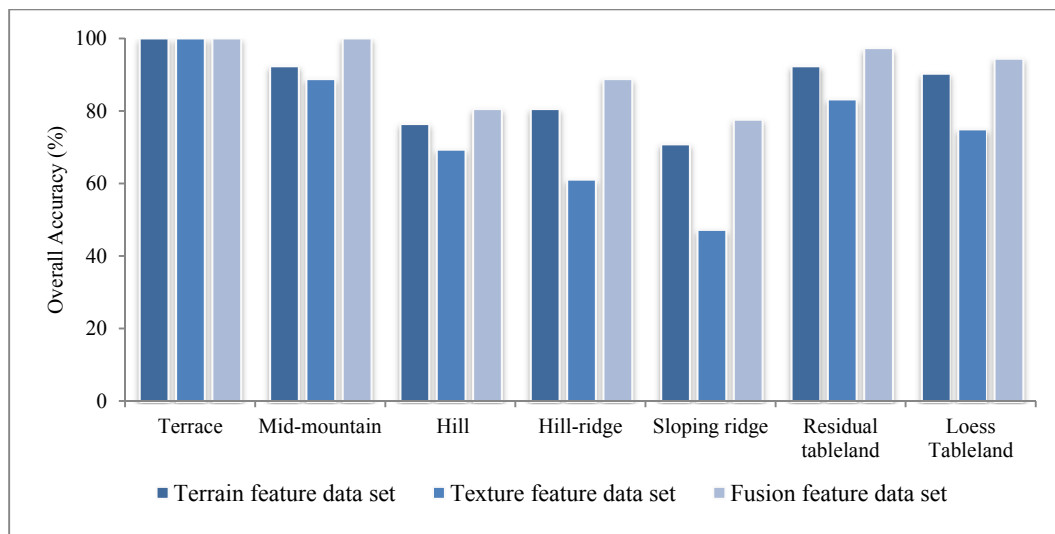
Figure 4 Recognition accuracy of each landform type.

different methods is presented in Figure 4. Based on geomorphic types, high accuracies could be observed in the landforms of loess terrace, loess-broken tableland, and loess mid-mountain. This result indicates that evident and divisible terrain features exist in these landforms unlike other landforms. Loess mid-mountains, which have rough topography and large relief intensity, can be easily distinguished. Strong separability was observed in the flat and wide loess tableland area. However, under evolutionary stages, significant differences of gully incision situation exist on the edges of tableland. Thus, this landform can be easily distinguished from other landforms. The accuracy of the landforms of loess hill, loess hill–ridge, and loess ridge, which are typical landforms in different evolutionary stages of the loess hill–gully area, may be low. These results may be attributed to the presence of similar patterns between the complex landforms, which leads to similar distributions of the statistical characteristics of terrain derivatives and close arrangement rule of image pixels.

5 Discussion

Terrain and texture derivatives were extracted in loess small-scale watershed and RF classifier was adopted in the current study to select features and recognize landforms. The following sections must be elaborated considering the aforementioned analysis.

5.1 Recognition results based on different segmentation strategies

The optimal split variables (classification rules) were obtained by extracting and training the quantitative geomorphic feature set, and test samples were then identified. However, defining the recognition objects is the main step of the process. Considering the characteristics of the consistency the formation process and morphological variation of the loess landforms, watersheds were selected as analysis objects to identify the landform types. These watersheds could be regarded as natural geomorphological units and geomorphic evolution units. The multi-resolution segmentation (MRS) algorithm, which is a part of OBIA, exhibited the highest sensitivity to the morphological discontinuities in DEMs. Further experimentation was conducted to compare the recognition results between the two strategies.

A test area from Loess broken tableland with an area of 8000 km² was selected. The image was segmented by using slope surface runoff simulation method and MRS method. The segmentation results are shown in Figure 5, and the recognition results carried by the RF classifier is presented in Table 5.

Figure 5 indicates that DEM was divided into various spatial objects. Although the multi-resolution segmentation method was designed to capture the spatial entities with maximized internal homogeneity and minimized external homogeneity,
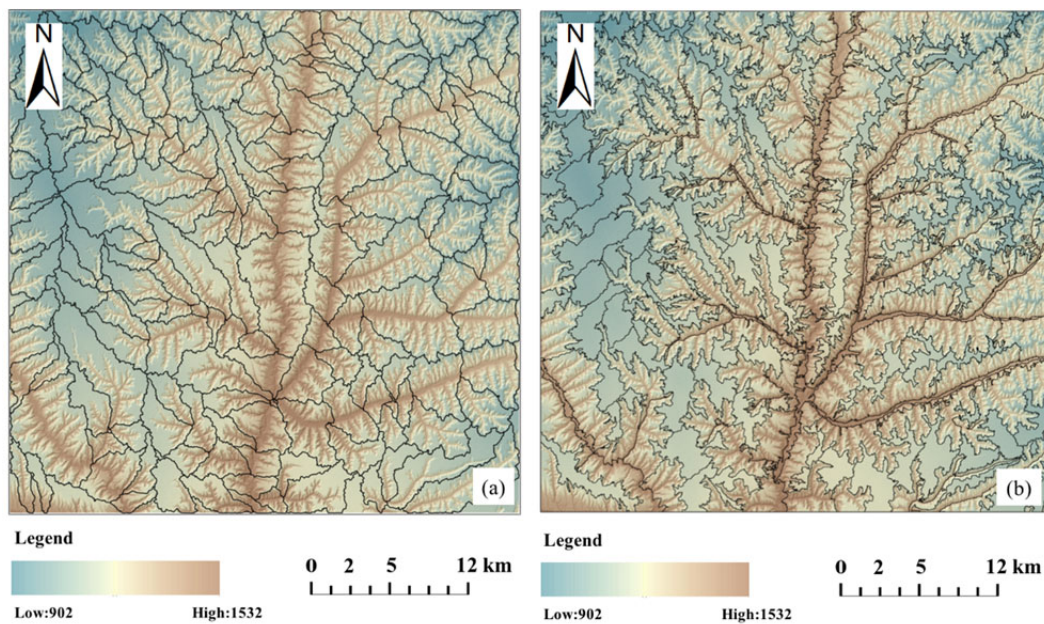
Figure 5 Image segmentation based on different strategies, (a) watershed segmentation, (b) multi-resolution segmentation.

Table 6 Recognition results based on different strategies

Segmentation strategy	Object number	Wrong number	Recognition accuracy (%)
Object-based	216	41	81.02
Watershed-based	199	28	85.93

the boundaries of these objects are fuzzy and irregular unlike segmented watersheds. Table 6 shows that the accuracy of the recognition result with watershed-based strategy is higher than that with object-based recognition strategy. This result suggests that watershed-based segmentation provides more reliable and efficient spatial units for landform classification and landform data mining than object-based segmentation.

5.2 Recognition results based on different classifiers

The recognition accuracy obtained by the RF method was compared with the results of two other popular algorithms of machine learning, e.g. CART trees and TreeNet gradient boosting machine. These classifiers can consistently generate accurate models and estimate the importance of variables during the classification process (Timofeev 2004; Elish et al. 2009). The basic parameter settings and optimal derivatives of the dataset selection strategy were consistent with RF. Figure 6 shows the correlation between derivative quantity and recognition accuracy of these three methods.

Figure 6 shows that recognition accuracy increases with derivative number. The highest recognition accuracy of the RF algorithm was observed at 18 to 22 derivatives, and the recognition accuracy gradually decreased afterwards. Thus, training dimensions improved because the derivative numbers were increased, whereas accuracy reduced when derivative redundancy was enlarged. The accuracy of TreeNet model is relatively high and exhibits a steady improvement. However, the highest accuracy was observed in the succeeding period, which indicates that the characteristics of TreeNet model in selecting effects are weaker than that of RF. Only 20 derivatives contributed to the recognition progress. Thus, the accuracy of CART algorithm is relatively low and does not substantially change when the derivative number increases. The recognition effect of the CART algorithm is relatively poor, and TreeNet algorithm is stable and exhibits comparably high precision. However, the RF algorithm can obtain the highest accuracy, which can also effectively filter high-dimensional derivatives. In this work, Random Forest, TreeNet, CART model and their extensions for variable selection were implemented in Salford Predictive Modeler software suite (SPM Version 8.0).
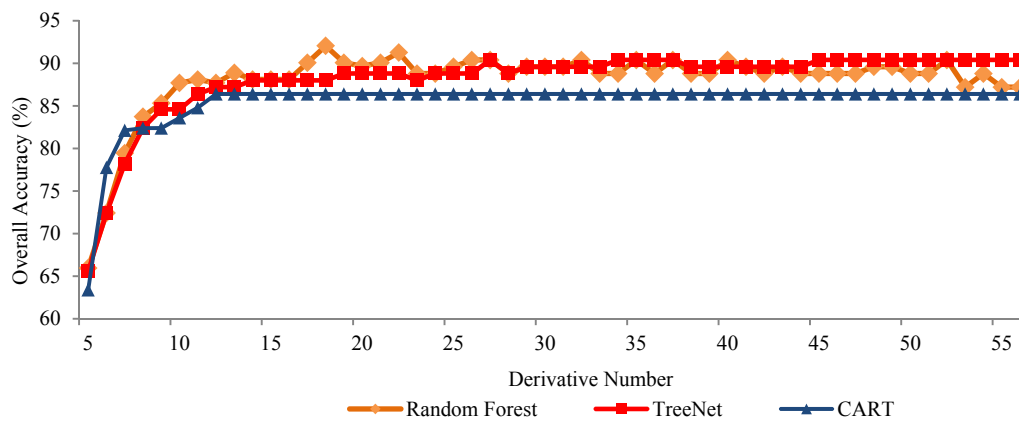
Figure 6 Correlation between recognition accuracy and derivative number.

During the training process, RF can automatically select high-dimensional geomorphologic characteristics and obtain unbiased estimations of prediction error with high precision and efficiency without significant human intervention. Moreover, RF can also detect outliers, which can be very useful when some cases are mislabeled. Thus, RF can be extensively applied in geomorphic information extraction and geomorphic type classification due to its high efficiency and accuracy.

6   Conclusion

The current study discusses the process of automatic recognition of loess landforms using watershed-based strategy and measures the effect of feature selection and recognition accuracy by using Random Forest (RF) classifier. The results show that automatic recognition achieves a high accuracy based on integrated terrain and texture derivatives, which are selected by the RF classifier. These employed methods are also expected to be applied in further landform classification, not only for loess landform area but also for other landform areas.

Terrain derivatives provide complementary and morphological characteristics of landform entities, while texture derivatives enhance above characteristics. In addition, terrain derivatives contribute more than texture derivatives during the process of recognition. Among the tested seven typical loess landforms during the experiment, loess terrace and loess mid-mountain, which have evident separable derivatives, obtained the highest accuracy. Distinguishing loess hill, loess hill–ridge, and loess ridge is difficult because some similarities can be observed between the morphologies of these landforms. Furthermore, the recognition accuracy of watershed-based segmentation strategy was higher than that of object-based recognition. RF obtained the best result in feature selection and landform recognition compared to that of two other machine-learning methods.

However, the scale effect of watershed-based segmentation and landform feature selection must be further explored. Future research can focus on the following aspects: 1. Multi-source data, such as remote sensing data and geological exploration data, should be included to comprehensively describe terrain derivatives. 2. Combined characteristics of topographic elements should be determined to achieve automatic recognition based on landform patterns.

References

Belgiu M, Drăguţ L (2016) Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry & Remote Sensing 114: 24-31. DOI: 10.1016/j.isprsjprs.2016.01.011

Benediktsson JA, Swain PH, Ersoy OK (1989) Neural Network Approaches Versus Statistical Methods in Classification of Multisource Remote Sensing Data. IEEE Transactions on Geoscience & Remote Sensing 28(4): 489-492. DOI: 10.1109/TGRS.1990.572944

Blaschke T (2010) Object based image analysis for remote sensing. ISPRS Journal of Photogrammetry & Remote Sensing 65(1): 2-16. DOI: 10.1016/j.isprsjprs.2009.06.004

Blaschke T, Strobl J (2001) What's wrong with pixels? Some recent developments interfacing remote sensing and GIS. GeoBIT/GIS 6.1: 12-17.

Breiman L (1996) Bagging Predictors. Machine Learning 24(2): 123-140. DOI: 10.1007/BF00058655

Breiman L (2001) Random Forests. Machine Learning 45(1): 5-32. DOI: 10.1023/A:1010933404324

Breiman, L (2002) Manual on setting up, using, and understanding random forests v3.1. Statistics Department University of California Berkeley, CA, USA. p 1.

Cao W, Tao HP, Kong B, et al. (2011) Topographic automatic recognition based on optimal topography feature space — taking southwest china as an example. Geomatics & Information Science of Wuhan University 36(11): 1376-1380. (In Chinese)

Caratti JF, Nesser JA, Maynard C (2004) Watershed Classification Using Canonical Correspondence Analysis and Clustering Techniques: A Cautionary Tale. Journal of the American Water Resources Association 40(5): 1257-1268. DOI: 10.1111/j.1752-1688.2004.tb01584.x

Cheng WM, Zhou CH, Li B, et al. (2011) Structure and contents of layered classification system of digital geomorphology for China. Journal of Geographical Sciences 21(5): 771-790. (In Chinese) DOI: 10.11820/dlkxjz.2014.01.003

Cutler DR, Edwards TC, Beard KH, et al. (2007) Random forests for classification in ecology. Ecology 88(11): 2783-92. DOI: 10.1890/07-0539.1

Danneels G, Pirard E, Havenith HB (2007) Automatic landslide detection from remote sensing images using supervised classification methods, Geoscience and Remote Sensing Symposium, 2007. IGARSS 2007. IEEE International, pp. 3014-3017. DOI: 10.1109/IGARSS.2007.4423479

Dikau R, Brabb EE, Mark RM (1991) Landform classification of New Mexico by computer. Open-File Report.

Drăguţ L, Blaschke T (2006) Automated classification of landform elements using object-based image analysis. Geomorphology 81(3-4): 330-344. DOI: 10.1016/j.geomorph.2006.04.013

Drăguţ L, Eisank C (2012) Automated object-based classification of topography from SRTM data. Geomorphology 141-142: 21-33. DOI: 10.1016/j.geomorph.2011.12.001

Du P, Samat A, Waske B, Liu S, et al. (2015) Random Forest and Rotation Forest for fully polarized SAR image classification using polarimetric and spatial features. ISPRS Journal of Photogrammetry & Remote Sensing 105: 38-53. DOI: 10.1016/j.isprsjprs.2015.03.002

Elish MO, Elish KO (2009) Application of TreeNet in predicting object-oriented software maintainability: a comparative study. European Conference on Software Maintenance and Reengineering. pp 69-78. DOI: 10.1109/CSMR.2009.57

Evans IS (2012) Geomorphometry and landform mapping: What is a landform? Geomorphology 137(1): 94-106. DOI: 10.1016/j.geomorph.2010.09.029

Freund Y (1996) Experiments with a new boosting algorithm. Thirteenth International Conference on Machine Learning. pp 148-156.

Galli M, Ardizzone F, Cardinali M, et al. (2008) Comparing landslide inventory maps. Geomorphology 94(3-4): 268-289. DOI: 10.1016/j.geomorph.2006.09.023

Gislason PO, Benediktsson JA, Sveinsson JR (2003) Random Forests for land cover classification. Pattern Recognition Letters 27(4): 294-300. DOI: 10.1016/j.patrec.2005.08.011

Gooding MP (2014) A watershed classification scheme for lower Michigan. (Doctoral dissertation)

Guo B, Damper RI, Gunn SR, et al. (2008) A fast separability-based feature-selection method for high-dimensional remotely sensed image classification. Pattern Recognition 41(5): 1653-1662. DOI: 10.1016/j.patcog.2007.11.007

Ham J, Chen Y, Crawford MM, et al. (2005) Investigation of the random forest framework for classification of hyperspectral data. IEEE Transactions on Geoscience & Remote Sensing 43(3): 492 - 501. DOI: 10.1109/TGRS.2004.842481

Hammond EH (1964) Analysis of properties in land form geography: an application to broad–scale land form mapping. Annals of the Association of American Geographers 54(54): 11-19. DOI: 10.1111/j.1467-8306.1964.tb00470.x

Haralick RM, Shanmugam K, Dinstein I (1973) Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics 3(6):610-621. DOI: 10.1109/TSMC.1973.4309314

Hervás J, Bobrowsky P (2009) Mapping: inventories, susceptibility, hazard and risk. Springer Berlin Heidelberg. pp 321-349. DOI: 10.1007/978-3-540-69970-5_19

Huang SL, Ferng JJ (1990) Applied land classification for surface water quality management: II. Land process classification. Journal of Environmental Management 31(2): 127-141. DOI: 10.1016/S0301-4797(05)80003-9

Ilea DE and Whelan PF (2011) Image segmentation based on the integration of colour–texture descriptors—a review. Pattern Recognition 44(10):2479-2501. DOI:10.1016/j.patcog.2011.03.005

Irvin BJ, Ventura SJ, Slater BK (1997) Fuzzy and isodata classification of landform elements from digital terrain data in Pleasant Valley, Wisconsin. Geoderma 77(2-4): 137-154. DOI: 10.1016/S0016-7061(97)00019-0

Jasiewicz J, Stepinski TF (2013) Geomorphons - a pattern recognition approach to classification and mapping of landforms. Geomorphology 182: 147-156. DOI: 10.1016/j.geomorph.2012.11.005

Li F, Tang GA, Wang C, et al. (2015) Slope spectrum variation in a simulated loess watershed. Frontiers of Earth Science: 1-12. DOI: 10.1007/s11707-015-0519-2

Liu K, Tang GA, Tao Y, et al. (2012) GLCM Based Quantitative Analysis of Terrain Texture from DEMs. Journal of Geo-Information Science 14(6): 751-760. (In Chinese) DOI: 10.3724/SP.J.1047.2012.00751

Liu K, Tang GA, Jiang S (2013) Research on the classification of terrain texture from DEMs based on BP neural network. Geomorphometry Org.

Liu SL, Li FY, Jiang RQ, et al. (2015) A Method of Loess Landform Automatic Recognition Based on Slope Spectrum. Journal of Geo-Information Science 17(10). (In Chinese) DOI: 10.3724/SP.J.1047.2015.00000

Lombardo L, Cama M, Conoscenti C, et al. (2015) Binary logistic regression versus stochastic gradient boosted decision trees in assessing landslide susceptibility for multiple-occurring landslide events: application to the 2009 storm event in Messina (Sicily, southern Italy). Natural Hazards 79(3): 1621-1648. DOI: 10.1007/s11069-015-1915-3

Martins FMG, Fernandez HM, Isidoro JMGP, et al (2016) Classification of landforms in Southern Portugal (Ria Formosa Basin). Journal of Maps 12(3): 422-430. DOI: 10.1080/17445647.2015.1035346

Van Niekerk A (2010) A comparison of land unit delineation

techniques for land evaluation in the Western Cape, South Africa. Land Use Policy 27(3): 937-945. DOI: 10.1016/j.landusepol.2009.12.007

Nikolakopoulos KG (2006) SRTM vs ASTER elevation products. Comparison for two regions in Crete, Greece. International Journal of Remote Sensing 27(21): 4819-4838. DOI: 10.1080/01431160600835853

O'Callaghan JF, Mark DM (1984) The extraction of drainage networks from digital elevation data. Computer Vision Graphics & Image Processing 28(3):323-344. DOI: 10.1016/S0734-189X(84)80011-0

Pal M (2005) Random forest classifier for remote sensing classification. International Journal of Remote Sensing 26(1): 217-222. DOI: 10.1080/01431160412331269698

Rodriguez-Galiano VF, Ghimire B, Rogan J, et al. (2012) An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing 67: 93-104. DOI: 10.1016/j.isprsjprs. 2011.11.002

Saeys Y, Inza I, Larrañaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23(19): 2507-2517. DOI: 10.1093/bioinformatics/btm344

Stepinski TF, Ghosh S, Vilata R (2006) Automatic Recognition of Landforms on Mars Using Terrain Segmentation and Classification. Discovery Science 255-266. DOI: 10.1007/11893318_26

Stumpf A, Kerle N (2011) Object-oriented mapping of landslides using Random Forests. Remote Sensing of Environment 115(10): 2564-2577. DOI: 10.1016/j.rse.2011.05.013

Tang GA, Yang W, Yang X, et al. (2003) Some Key Points in Terrain Variables Deriving from DEMs. Science of Surveying and Mapping 28(1):28-32. (In Chinese)

Tang GA (2014) Progress of DEM and digital terrain analysis in China. Acta Geographica Sinica 69(9): 1305-1325. (In Chinese)

Tang GA, Song XD, Li FY, et al. (2015) Slope spectrum critical area and its spatial variation in the Loess Plateau of China. Journal of Geographical Sciences 25(12): 1452-1466. DOI: 10.1007/ s11442-015-1245-0

Timofeev R (2004) Classification and Regression Trees (CART) Theory and Applications. Humboldt-Universität zu Berlin, Wirtschaftswissenschaftliche Fakultät.

Ulaby FT, Kouyate F, Brisco B, et al. (1986) Textural infornation in SAR images. IEEE Transactions on Geoscience and Remote Sensing (2): 235-45. DOI: 10.1109/TGRS.1986.289643

Verhagen P, Drăguţ L (2012) Object-based landform delineation and classification from DEMs for archaeological predictive mapping. Journal of Archaeological Science 39(3): 698-703. DOI: 10.1016/j.jas.2011.11.001

Wang C, Hu P, Liu XH, et al. (2009) Automated Classification of Martian Landforms Based on Digital Terrain Analysis(DTA) Technology. Geomatics & Information Science of Wuhan University 34(4): 483-487. (In Chinese) DOI: 10.13203/j.whugis2009.04.021.

Xiong LY, Tang GA, Li FY, et al. (2014) Modeling the evolution of loess-covered landforms in the Loess Plateau of China using a DEM of underground bedrock surface. Geomorphology 209(3): 18-26. DOI: 10.1016/j.geomorph. 2013.12.009

Xu L, Li J, Brenning A (2014) A comparative study of different classification techniques for marine oil spill identification using RADARSAT-1 imagery. Remote Sensing of Environment 141(141): 14-23. DOI: 10.1016/j.rse.2013.10.012

Zhou CH, Cheng WM, Qian JK, et al. (2009) Research on the Classification System of Digital Land Geomorphology of 1:1000000 in China. Journal of Geo-Information Science 11(6):707-724. (In Chinese)

Zhu S, Tang G, Xiong L, Zhang G (2014) Uncertainty of slope length derived from digital elevation models of the Loess Plateau, China. Journal of Mountain Science 11(5): 1169-1181. DOI: 10.1007/s11629-013-2788-0