引用格式：曹泽涛,方子东,姚瑾,等.基于随机森林的黄土地貌分类研究[J].地球信息科学学报,2020,22(3):452-463. [Cao Z T, Fang Z D, Yao J, et al. Loess landform classification based on random forest[J]. Journal of Geo- information Science, 2020,22(3):452- 463.] **DOI:** 10.12082/dqxxkx.2020.190247

基于随机森林的黄土地貌分类研究

曹泽涛[1,2,3],方子东[1,2,3],姚　瑾[4],熊礼阳[1,2,3*]

1. 南京师范大学地理科学学院,南京 210023；2. 虚拟地理环境教育部重点实验室(南京师范大学),南京 210023；
3. 江苏省地理信息资源开发与利用协同创新中心,南京 210023；4. 自然资源部第一地理信息制图院,西安 710054

Loess Landform Classification based on Random Forest

CAO Zetao[1,2,3], FANG Zidong[1,2,3], YAO Jin[4], XIONG Liyang[1,2,3*]

1 .School of Geography, Nanjing Normal University, Nanjing 210023, China; 2. Key Laboratory of Virtual Geographic Environment (Nanjing Normal University), Nanjing 210023, China; 3. Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China; 4. The First Institute of Geoinformation Mapping, Ministry of Natural Resources of the People's Republic of China, Xi'an 710054, China

Abstract: Landform classification is one of the most important steps tor eveal the mechanisms of surface matter flows and energy conversion, which could inform the scale and layout of human construction activities. However, traditional landform classification methods based on Digital Elevation Model (DEM) often use a small number of topographical derivatives or landform characteristics, resulting in insufficiently precise classification results. However, object- oriented landform classification performs better in that reliable classification can be achieved by maximizing the homogeneity within and between objects. But how to set conditions in object segmentation remains a challenge. In this paper, a geomorphological classification method based on watershed unitwas proposed, by accounting for many characteristics of watershed unit including statistics of basic topographic factors, feature point and feature line, basin and texture characteristics. Firstly, hydrological analysis based on DEM divided the study area into different small basins as the experimental units. Then, 29 features were extracted within each unit to represent watershed morphology using digital terrain analysis; feature selection and parameter calibration were carried out based on Random Forest (RF) algorithm. RF is a supervised integrated learning model aggregating different outputs of a single decision tree to reduce variances that may lead to classification errors in the decision tree. Finally, the watersheds in training set were selected to train the RF classifier, and the trained classifier was used to classify the landform of the whole study area, based on which we studied the landform spatial differentiation pattern. This experiment achieved good results in the landform classification of the Loess Plateau in northern Shaanxi Province. It is one of the areas with the most serious soil erosion and the most fragile eco- environment in the world. Most of them are covered by thick loess, and the topography is fluctuant. Result shows: (1) Compared with manual interpretation, excellent classification results

收稿日期：2019-05-23；**修回日期**：2019-12-09.

基金项目：国家自然科学基金项目(41601411、41671389)；江苏高校优势学科建设工程资助项目。[**Foundation items:** National Natural Science Foundation of China, No.41601411, 41671389; Priority Academic Program Development of Jiangsu Higher Education Institutions.]

作者简介：曹泽涛(1997—)，男，江苏扬州人，硕士生，主要从事研究DEM数字地形分析研究。E-mail: zetao_cao_1997@163.com

***通讯作者**：熊礼阳(1989—)，男，江西南昌人，副教授，主要从事黄土继承性DEM数字地形分析研究。
E-mail: xiongliyang@163.com

based on small watershed in the study area were obtained, with the classification accuracy reaching 85% and the Kappa coefficient 0.83. (2)All small watersheds were divided into eight types of landforms. The same type of landforms showed obvious spatial aggregation. There were boundaries and transitional zones between different types of landforms. (3) Different geomorphological regions explained different situations of loess deposition and runoff erosion in different regions. Our findings suggest that the combination of RF algorithm and DEM data can achieve better classification results.

Key words: Landform; random forest; loess plateau; terrain feature; feature selection; geomorphological classification; DEM; watershed

***Corresponding author:** XIONG Liyang, E-mail：xiongliyang@163.com

摘要：地貌分类在指导人类建设活动的规模与布局中有着重要的意义。然而，传统的基于数字高程模型（DEM）的地貌分类方法使用的地形因子和考虑到的地貌特征往往比较单一。本文提出了一种基于流域单元的地貌分类方法，该方法考虑了流域单元的多方面特征，包括基本地形因子统计量、地形特征点线统计量、小流域特征和纹理特征。本研究首先基于DEM进行水文分析将研究区域划分成不同的小流域。然后利用数字地形分析提取29个不同方面的特征来表征流域的形态，并基于随机森林（RF）算法进行了特征选择和参数标定。RF是一种基于决策树算法的集成分类器，能有效地处理高维数据，分类精度高。最后选择训练集小流域对RF分类器进行训练，使用训练完成的分类器对整个研究区域的地貌进行分类，研究地貌分异的规律。该实验在我国陕北黄土高原典型黄土地貌区域的地貌分类中取得了较好的结果，结果表明不同的地貌之间存在明显的区域界线，特定的地貌类型在空间上表现出明显的聚集性。通过人工判读进行验证的分类精度达到了85%，Kappa系数为0.83。

关键词：地貌；随机森林；黄土高原；地形特征；特征选择；地貌分类；DEM；小流域

1 引言

地貌作为地球表层系统中的一个最重要的基本自然地理要素，影响甚至决定着其他要素的特征，并直接地影响人类活动，是地理学研究的核心与基础内容之一[1-4]。地貌分类是划分和识别地表形态的最重要的步骤之一[5]，能够反映地表沉积物和表层营养物质的水平流动，进而揭示地表物质流动和能量转换的内在机理；能够反映初级生产力、水资源质量[6-7]；能够帮助环境质量与生物多样性的监测与维护[8-9]；同时，地貌分类反映出的景观格局在指导人类建设活动的规模与布局中具有重要现实意义。

传统的地貌分类方法主要是依据地貌的形成机制进行野外观测，根据航摄影像、地形图进行判别来实现的[10-11]，其分类结果可以将山地、平原等地貌区分，但由于不足的地貌学知识和有限的数据，往往不能够得到更加精细的地貌分类结果。并且，地形图中包含的高程信息非常有限，形态信息不够准确，这些都制约了对复杂地貌的判别。数字高程模型（DEM）的出现解决了数据上的问题，是目前进行自动的地形特征分析和地貌规律研究的主要数据源。近年来，在DEM数字地形分析基本理论与技术、基于DEM的地貌分类与制图等方面有了大量的研究成果[12-18]。

基于DEM数据的地貌类型的分类主要集中在基于像素单元的划分和面向对象宏观地貌形态分类2方面。前者的主要思路为，将地形因子作为主要的区分特征，对研究区的栅格单元进行地形因子计算并构建其特征空间，采用非监督分类或决策树的分类方法，实现地貌类型的识别[19-22]。然而，基于像素的方法并没有充分考虑像素的邻域以及地貌作为一个区域尺度单元的整体性。另一个方面是面向对象的方法，这个方法广泛应用于基于遥感图像或DEM数据的分类识别当中，满足地貌对象的概念模型[23]。使用基于对象的图像分析（OBIA）或数字地形分析技术（DTA）从DEM数据中获取信息对地貌进行分类，可以取得较好的分类结果[24-25]，但由于分割边界的地理意义一直不明确，存在着明显的不足。相比其他面向对象的方法划分的单元，流域单元在地表形态和地貌演变方面具有明确的地理意义，地貌的特征和变化规律较易被掌握，并且能够契合地貌和水文之间紧密的关系，所以基于流域单元的地貌识别和分类有着很好的效果[26-28]。

然而，无论是基于像素单元还是面向对象的地貌分类方法，所使用的指标往往比较单一或者数量较少[19-22,25-31]，不能全面地体现出地貌的特征，而且使用高维特征的地貌分类相比使用单一特征的地貌

分类精度更高[26]。但原先的分类算法不再适合高维数据的处理,集成分类器的出现使得基于高维特征的分类成为可能,随机森林便是其中之一[32],随机森林方法已经被广泛地运用在基于遥感数据的地表分类上,并取得了很好的分类效果[33-35]。已有研究将随机森林与DEM数据结合用于地貌的识别上[26],但结合两者进行地貌的分类还很少有研究涉及。结合集成分类器与更加全面的地貌特征进行地貌分类,可以从更加全面的角度对地貌进行描述,从而提高分类精度。

本研究选取我国陕西北部的黄土高原作为研究区域,选用ASTER GDEM 30 m分辨率的DEM作为实验数据,首先使用水文分析,设立特定的汇流累积量阈值对研究区域进行小流域划分,使得划分得到的小流域地貌特征达到稳定。并初步选取了小流域单元不同角度的29个特征,从而较为全面地描述每个小流域。进一步地,在所有的小流域中选择出部分作为地貌样区小流域,基于样区小流域的特征,使用随机森林算法对这些特征进行重要性评价,进行特征选择。最后基于选择的特征,以样区小流域作为训练集对随机森林分类器进行训练,使用训练后的分类器对所有小流域进行分类。

2 研究区概况与研究方法

2.1 研究区概况

本文以陕西境内关中平原以北的黄土高原为研究对象。黄土高原作为中国四大高原之一,以其独特的地理条件、地貌研究价值、黄土地貌特征、独特的自然景观和人文景观而备受世界关注[36-37]。研究区地貌形态以黄土地貌为主,黄土塬、梁、峁及沟谷等地貌发育十分典型。此外,陕北黄土高原各种地貌类型的空间分异规律也比较明显:北部沿线为风沙地貌区,向南逐步过渡到黄土低丘地貌,再向南到延安地区渐变为峁状丘陵沟壑地貌。因此,选择陕北黄土高原作为研究区域,能够更全面的反映黄土地貌形态的总体特征,既有地貌学上的理论意义,也具有实际意义。

本研究在研究区域中选择了均匀分布的典型地貌单元作为研究样区。这些地貌样区分别选自于沙丘草滩盆地、黄土低丘、黄土峁状丘陵沟壑、黄土梁状丘陵沟壑、黄土残塬丘陵沟壑、石质山地、黄土塬和黄土台塬这些典型的黄土地貌类型区域[31,36,38],基本包含了黄土高原典型的地貌组合及景观形态[36],这些地貌样区的分布及样区流域的遥感影像如图1所示。

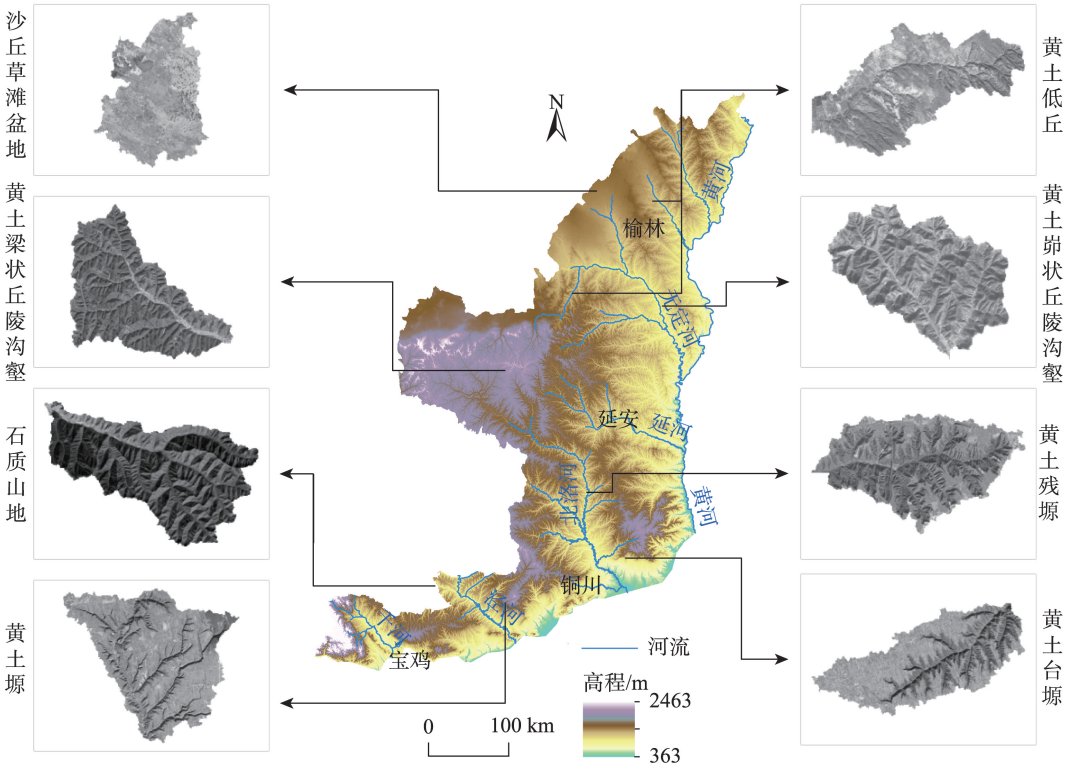

图1 陕西境内关中平原以北的黄土高原高程及地貌样区分布

Fig. 1 Elevation of the study area and distribution of sampled geomorphic types

2.2 数据源

本实验的实验数据为研究区域内的ASTER GDEM数据[39]。ASTER GDEM数据由日本METI和美国NASA联合研制，公众可免费获取。ASTER GDEM数据产品基于“先进星载热发射和反辐射计（ASTER）”数据计算生成，是目前唯一覆盖全球陆地表面的高分辨率高程影像数据，数据的分辨率为30 m，数据来源于http://www.gscloud.cn/。

2.3 小流域划分与特征提取

通过水文学的方法，设立不同的汇流累积量阈值，可以将一个区域划分成尺度不同的小流域[40]。所以在利用DEM对小流域进行分割时，首先要确定小流域分割的大小。若小流域分割的太大，可能会出现一个小流域内包含不止一种地貌类型的情况；若小流域分割的太小，则会因为数据分辨率不足产生错误。

相关实验表明，当小流域面积过小时，小流域的地形指标不稳定，对小流域的地貌类型不具有代表性[41-43]。大量研究发现，在黄土高原地区，当小流域面积增大到10 km^{2}时，小流域划分的结果基本正确，相关地貌指标趋于稳定[30-31,42-43]。因此，本研究确定的小流域分割的最小面积阈值为10 km^2，设定的汇流累积量阈值为16 000。

对整个研究区域内30 m分辨率的DEM数据进行小流域的分割。分割过程可以分为4个步骤：① 对DEM数据进行填洼；② 使用填洼后的DEM数据进行流向和汇流累积量的计算[40,44]；③ 设定汇流累积量的阈值为16 000，提取栅格河网数据并将其矢量化；④ 使用流量数据和矢量河网数据进行小流域的划分[45]。划分结果如图2所示。

从DEM中提取的地形因子是定量描述地貌最有效的指标[26]。地形因子一方面与地貌形态具有对应性，另一方面又是对地貌特征的抽象，反映地貌在空间上连续变化的特征。选择合适的地形因子作为小流域单元的特征是实现地貌分类的关键。

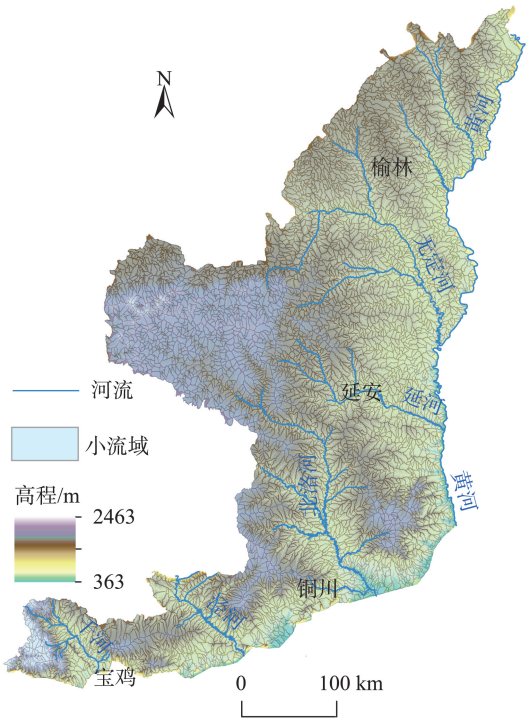

图2　小流域划分结果

Fig. 2　Result of small watershed division

作为小流域单元的抽象，小流域特征应该尽可能全面地、从多个尺度反映地貌形态，并且在此基础上，特征之间应该相对独立。其次，考虑到数据精度及现有条件，选取的特征的计算算法应尽可能简洁。

依照上述的条件，综合考虑先前学者对各种地形因子的研究[31,39,46-53]，本研究初步选取并计算了29个小流域特征，这些小流域特征彼此之间较为独立且能够代表小流域单元的多个方面，这些特征如表1所示。

2.4 随机森林算法

随机森林算法（Random Forest，RF）由Breiman等[32]在2001年提出，RF具有很高的预测准确率，对异常值和噪声有很强的容忍度，能够处理高维数

表1　初步选取的小流域特征

Tab. 1　Preliminary selection of basin features

类型	小流域特征	数量/个
基本地形因子统计量	平均高程、高程标准差、平均坡度、坡度标准差、平均起伏度、起伏度标准差、平均切割深度、切割深度标准差、平均平面曲率、平面曲率标准差、平均剖面曲率、剖面曲率标准差、平均坡向、坡向标准差	14
地形特征点线统计量	山顶点密度、山顶点高程标准差、沟沿线密度、沟沿线平均高程、割裂度、沟谷线密度、沟谷线平均高程	7
小流域特征	相对高程差、沟谷深度、面积高程积分、坡谱信息熵	4
纹理特征	纹理对比度、纹理角二阶矩、纹理信息熵、纹理逆差矩	4

据，有效地分析非线性、具有共线性和交互作用的数据，并能够在分析数据的同时给出变量重要性评分（Variable Importance Measures，VIM）。这些特点使得RF特别适用于高维数据的研究。

2.4.1 随机森林的基本原理

RF是一种基于决策树算法的集成分类器，它通过自助法（Bootstrap）重采样技术，从原始训练集中有放回地重复随机抽取n样本数据集生成新的训练自助样本集合，以降低分类模型间的相关性，从而提高分类器整体的识别精度。然后根据自助样本集生成n个决策树组成RF，目标的分类结果按决策树投票多少形成的分数而定。

RF中的每一棵决策树都为二叉树，根节点包含全部训练自助样本，按照一定的原则，在每个节点从一组随机选取的变量中选择使分枝后节点“不纯度”最小的变量作为分枝变量，分裂得到左节点和右节点，它们分别包含训练数据的一个子集，分裂后的节点按照同样规则继续分裂，直到满足分枝停止规则而停止生长，具体过程见图3。“不纯度”的衡量标准包括Gini指数和熵。

2.4.2 随机森林的变量重要性评分

RF可以用来评估一组变量对于分类的重要程度，可以作为特征选择的依据。RF常规的变量重要性评分（VIM）计算方法可以根据Gini指数计算得到[54-55]。假设有变量$X_1, X_2, \cdots, X_M$，其中变量X_j的VIM可由式（1）到式（4）计算得到。

一棵决策树中，节点m的Gini指数的计算公式为：

$$G_m = \sum_{k=1}^{K} P_{mk}(1 - P_{mk}) \tag{1}$$

式中：G_m为节点m的Gini指数；K为样本集中类的个数；P_{mk}为样本在节点m属于第k类的概率估计值。

变量X_j在节点m的重要性，即节点m分裂前后Gini指数变化量，计算公式为：

$$V_{jm}^{Gini} = G_m - G_{ml} - G_{mr} \tag{2}$$

式中：V_{jm}^{Gini}为变量X_j在节点m的重要性；G_{ml}和G_{mr}分别表示由节点m分裂的左右节点的Gini指数。

如果第i棵树中有M个节点包含变量X_j，则变量X_j在第i棵树的重要性计算公式为：

$$V_{ij}^{Gini} = \sum_{m=1}^{M} V_{jm}^{Gini} \tag{3}$$

式中：V_{ij}^{Gini}为变量X_j在第i棵树的重要性；M为第i棵树中包含变量X_j的节点数量；V_{jm}^{Gini}为变量X_j在节点m的重要性。

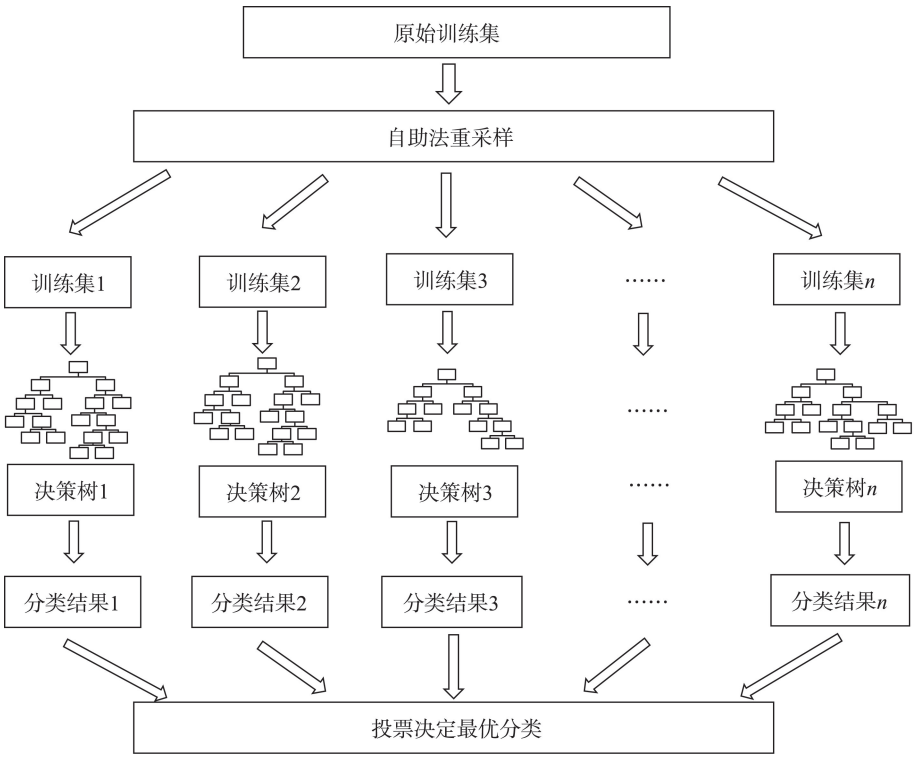

图3 随机森林原理[56]

Fig 3 Principle of Random Forest[56]

若RF中有n棵树，则变量X_j在RF中的Gini重要性定义为变量X_j在RF所有树的重要性的平均值，计算公式为：

$$V_j^{Gini}=\frac{1}{n}\sum_{i=1}^{n}V_{ij}^{Gini} \tag{4}$$

式中：V_j^{Gini}为变量X_j在RF中的Gini重要性；n为RF中决策树的数量；V_{ij}^{Gini}为变量X_j在第i棵树的重要性。

2.4.3 随机森林精度评估

在RF的自助法重采样过程中，原始训练集（N个样本）中每个样本未被抽取的概率为$(1-1/N)^N$，这表明训练集中有大约1/3的数据是没有被选取的。这些没有被选取的数据被称为袋外样本（Out of Bag，OOB），通过袋外样本，可以得到每棵决策树的OOB精度估计。将森林中所有决策树的OOB精度估计取平均，即可得到RF的泛化精度估计[56]。

3　结果及分析

3.1　特征重要性评估

从多个尺度反映地貌形态的小流域特征可以很好地表达出小流域单元，较高维度的特征也可以提高自动识别的精度。但仅仅将所有的特征简单地叠加在一起可能会导致计算和存储空间的过度消耗，不适合的特征可能也会带来额外的噪声从而影响分类精度。所以在分类前需要对所有的特征进行特征选择，筛选出对于分类较为重要的特征。本研究采用RF算法，基于Gini指数，对特征的重要性进行评估，选取重要性较高的特征，确定了进行小流域单元分类的主要特征。

研究从沙丘草滩盆地、黄土低丘、黄土峁状丘陵沟壑、黄土梁状丘陵沟壑、黄土残塬丘陵沟壑、石质山地、黄土塬和黄土台塬8种地貌类型中选取了210个小流域作为样区小流域，地貌类型通过观察遥感图像确定。实验使用样区小流域的地貌类型和特征作为训练集，基于Gini指数，使用RF算法对初步选取的29个特征进行重要性计算，对计算得到的重要性进行最大值归一化处理后结果如图4所示。在进一步的特征选择中可以依次选取重要性得分较高的特征。

3.2　特征选择与参数标定

集成的分类方法一般被认为是黑箱的分类器。作为集成分类器的一种，RF分类器需要进行参数的标定，才能获得最佳的分类效果。RF分类器中，需要标定的参数有RF中决策树的数量（*nTrees*），分裂节点时需要考虑的特征子集的大小（*nFeatures*），树中一个节点所需要用来分裂的最少样本数（*nSamples*）和树的最大深度（*max_depth*）。已有结论表明，一般在分类时：① *nFeatures*的最佳大小为数据集中特征数量（N）的算术平方根；② 当*max_depth*没有限制且*nSamples*=2时分类效果最好[32]。所以在使用RF分类器进行地貌分类时，该结论仍然适用。

因此，本研究中需要进行调整的参数为输入的特征数量N和RF中树的数量*nTrees*。构建RF分类器时，将样区小流域作为训练集，首先改变N的值，通过RF自身的OOB精度估计，计算*nTrees*为50、100、200和500时N对应的OOB精度的均值，进行参数N的标定实验，结果如图5所示。

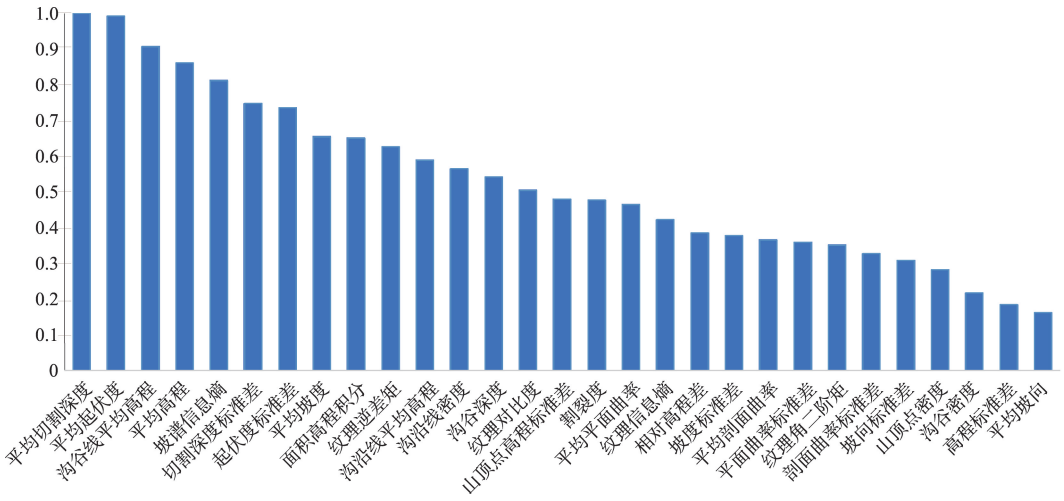

图4　训练样区小流域地形特征的重要性评分

Fig. 4　Importance ranking of each feature of watershed in training area

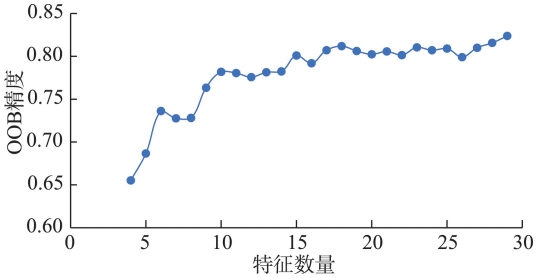

图5 预测精度随特征数量变化情况

Fig. 5 Relationship between prediction accuracy and features number

特征数量的参数标定结果表明RF分类精度总体上随着特征的数量增多而上升，但当N上升到18之后，精度变化不明显，部分情况下精度有所降低，说明部分重要性较低的特征对分类结果的影响很小，甚至带来额外的噪声使精度下降。虽然特征数量达到最高时预测精度很高，但相比N=18精度上升不明显，且不能排除是RF分类器产生了过拟合的现象。因此，最终选取重要性排名前18的特征：平均切割深度、平均起伏度、沟谷线平均高程、平均高程、坡谱信息熵、切割深度标准差、起伏度标准差、平均坡度、面积高程积分、纹理逆差矩、沟沿线平均高程、沟沿线密度、沟谷深度、纹理对比度、山顶点高程标准差、割裂度、平均平面曲率、纹理信息熵。

接下来对RF中树的数量*nTrees*进行标定，将样区小流域作为训练集，输入选取的18个特征，改变*nTrees*的值，计算不同*nTrees*值对应的OOB精度，结果如图6所示。

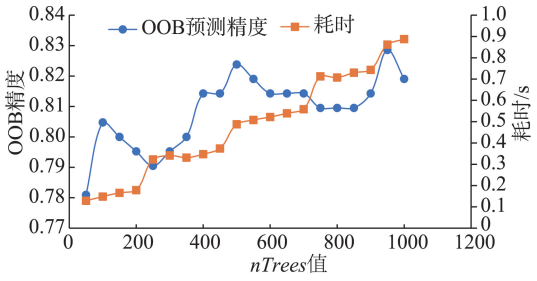

图6 *nTrees*参数标定结果

Fig. 6 Calibration results of the *nTrees* parameter

对*nTrees*的参数标定结果表明，RF分类精度总体上随着RF中树的数量增多而上升。当*nTrees*大于500时，精度随着*nTrees*增多而上升的趋势不明显，且树的数量增多会导致计算时间的增加，所以最终*nTrees*的值选择为500。

特征选择与参数标定完成后，使用标定的参数和所选的特征集合进行RF分类器的构建和训练。

3.3 基于随机森林的小流域单元分类

本研究选取8个地貌类型中的210个样区小流域，每个地貌类型中选取的样区小流域的数量如表2所示。结合3.2节确定的参数构建RF分类器并进行训练。训练完成后，使用RF分类器对面积大于10 km^{2}的小流域单元进行分类；其他面积较小的小流域的地貌类型由其邻域中面积最大的地貌类型决定。

表2 各地貌类型中样区数量

Tab. 2 Number of sample areas per geomorphological type

地貌类型	样区小流域数量/个
沙丘草滩盆地	30
黄土低丘	30
黄土峁状丘陵沟壑	30
黄土梁状丘陵沟壑	30
黄土残塬丘陵沟壑	30
石质山地	30
黄土塬	16
黄土台塬	14
汇总	210

分类结果如图7所示。为了清晰表达地貌类型分布的整体特征，使用邻域小流域对分类结果中零散的小流域进行众数处理和融合，处理结果如图8所示。分类结果表明不同的地貌之间存在明显的区域界线，特定的地貌类型在空间上表现出明显的聚集性。相比基于单一指标或决策树方法进行地貌分类的结果[19-22,25,28-31]，基于随机森林算法的分类显示出不同地貌类型的边界存在着过渡区，出现了不同地貌类型的小流域交错分布的结果，这一结果符合地貌类型在空间上的渐变的特征。

研究区内的地貌类型，包括沙丘草滩盆地、黄土低丘、黄土峁状丘陵沟壑、黄土梁状丘陵沟壑、黄土残塬丘陵沟壑、石质山地、黄土塬和黄土台塬，地貌类型在空间上存在连续变化。

在研究区西北部存在从沙漠区到黄土丘陵沟壑区的分界线(T1与T2之间)，此分界线大致与长城重合。T1属于毛乌素沙漠地区，是一个平坦的、黄沙覆盖的草滩盆地，这一地貌类型与其他黄土丘陵沟壑地貌具有明显的地表性质差异和地貌形成机理差异。该分界线对毛乌素沙漠的治理与管理和地貌发育的解释具有重要意义。T2是干旱风沙地貌

区和黄土丘陵沟壑地貌之间的过渡区，兼具二者特征，是一个分布着黄土低丘，受到风沙影响的区域。

子午岭以东、吕梁山以西的陕北黄土高原，在地质构造上属鄂尔多斯台向斜的主体部分，黄土覆盖面广，是黄土高原的核心部分，也是典型黄土地貌集中分布区。其中绥德、米脂等地分布着黄土峁状丘陵沟壑(T3)，是黄土高原中地形最为破碎的区域；白于山及吴起、志丹、甘泉等地分布着黄土梁状丘陵沟壑(T4)；黄龙山周围的宜君、宜川以及淳化等地分布着不同类型的黄土塬(T5，T7，T8)，这些不同类型的塬是由于受到不同程度的流水侵蚀而形成。这3种类型的地貌(塬、梁、峁)所在区域在第四纪以来以间歇性上升为主，同时临近黄河，受黄河基准面下降影响较大，因而地面遭受强烈侵蚀的历史较长。这3个典型黄土地貌区反应了它们在小流域尺度上的地表形态的特殊地貌过程，是最容易发生水土流失的地区，因此分类得到范围对水土流失研究和治理具有重要意义。另外，本研究表明六盘山内的千阳县千河两侧也分布有黄土残塬，这是与以往基于少数地形因子的分类的不同之处。

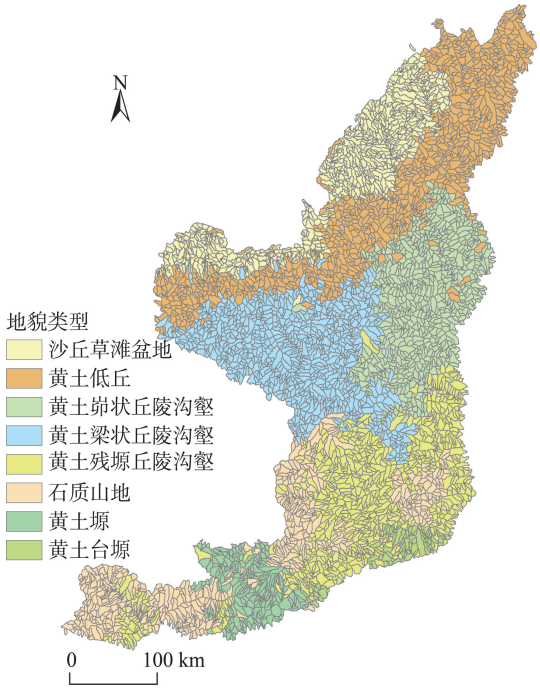

图7　基于随机森林的小流域单元分类结果

Fig. 7　Classification result of watershed based on random forest

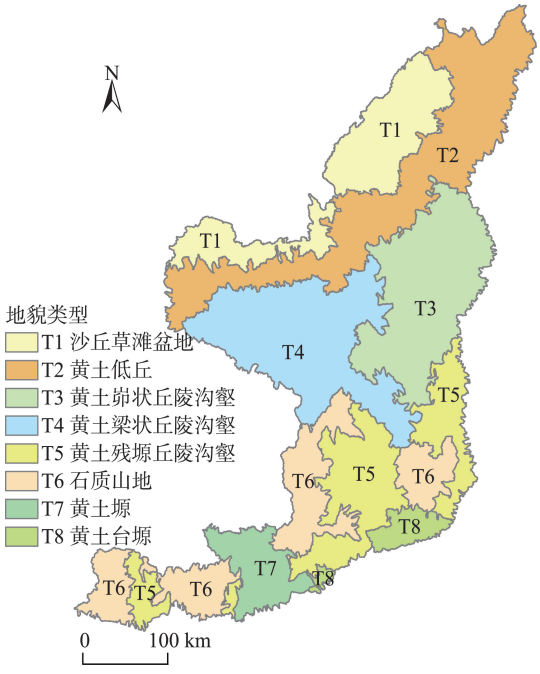

图8　分类后小流域融合处理结果

Fig. 8　Result of fusion processing of classified watershed

分类结果还表明，研究区内分布着黄龙山、子午山和六盘山3个石质山地(T6)，这些地区古地层基准较高，黄土土层薄，受到侵蚀后露出基岩，显示出石质山地的特征。

这些不同地貌区域及其分界线说明了区域地貌分类的重要性，强调了本研究的意义。

3.4　分类精度验证

精度评价是分类中的一项非常重要的工作，它是分类结果适用性评价的标准。本研究在划分的小流域中随机选取100个小流域，通过人工判读得到这些小流域的地貌类型，将人工判读的地貌类型于RF分类器自动分类的结果进行比较。

图9显示了研究区域内随机采样的用于精度验证的小流域的空间分布。RF自动分类结果与人工判读结果的混淆矩阵如表3所示。对比结果表明，100个小流域中有85个小流域的人工判读结果与RF自动分类结果一致。15个不一致的小流域中有7个小流域的内部地形十分复杂，是2种地貌类型的混合，其地理位置也处于分类结果中不同地貌类型的分界线处，但RF自动分类只能将其归类成一种地貌类型。这些小流域在人工判读时被认为时复杂的混合地形，标识为T2/T1、T4/T3。复杂的混合地形出现在沙丘草滩盆地和黄土低丘的交界处，黄土梁状丘陵沟壑和黄土峁状丘陵沟壑的交界处。

在本次实验中，将复杂的混合地形作为误差进行分类精度评价，分类结果可达到85%精度，Kappa系数为0.83；若将复杂地形作为误差的一半进行精度评价，达到88.5%的精度。

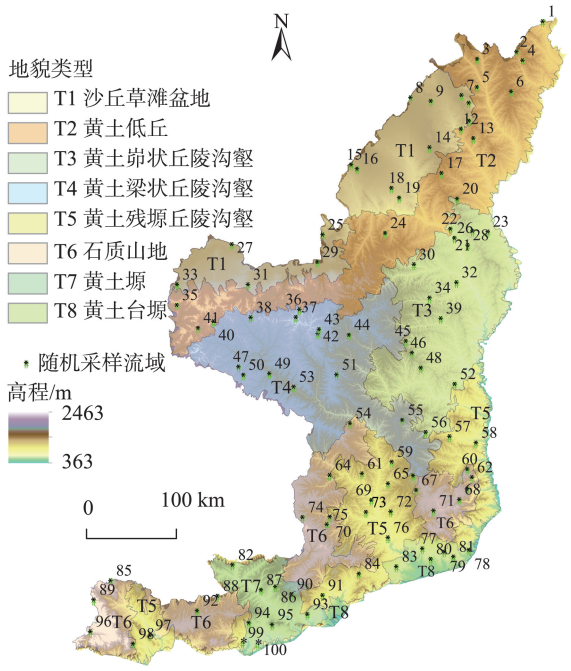

图9 随机采样小流域(用于精度验证)的空间分布

Fig. 9 Spatial distribution of randomly sampled watersheds (used for accuracy verification)

表3 RF分类结果与人工判读结果的混淆矩阵

Tab. 3 Confusion Matrix between RF classification result and manual interpretation result

		RF分类结果									
		T1	T2	T2/T1	T3	T4	T4/T3	T5	T6	T7	T8
人工判读结果	T1	13	1	0	0	0	0	0	0	0	0
	T2	0	10	0	0	0	0	0	0	0	0
	T2/T1	1	2	0	0	0	0	0	0	0	0
	T3	0	1	0	12	0	0	0	0	0	0
	T4	0	1	0	0	10	0	0	0	0	0
	T4/T3	0	0	0	3	1	0	0	0	0	0
	T5	0	0	0	0	0	0	13	0	1	0
	T6	0	0	0	0	1	0	2	13	0	0
	T7	0	0	0	0	0	0	1	0	9	0
	T8	0	0	0	0	0	0	0	0	0	5

表4为随机选取的100个小流域中不同地貌类型的小流域分类精度评价结果。8个地貌类型中T1(沙丘草滩盆地)、T6(石质山地)、T7(黄土塬)和T8(黄土台塬)的分类精度较高,达到90%以上,说明分类方法对破碎程度较低(T1、T7、T8)和具有明显异质特征(T6)的地貌类型有着较好的结果。其次,T3(黄土峁装丘陵沟壑)、T4(黄土梁状丘陵沟壑)、T5(黄土残塬丘陵沟壑)的分类精度低一些,这

表4 不同地貌分类精度

Tab. 4 Classification accuarcies of different landforms

地貌类型	正确分类数量/个	错误分类数量/个	分类精度/%
T1	13	1	92.9
T2	10	5	66.7
T3	12	3	80.0
T4	10	2	83.3
T5	13	3	81.3
T6	13	0	100.0
T7	9	1	90.0
T8	5	0	100.0

与这些地区地表破碎程度高、形态复杂有关,部分小流域内存在着混合的地貌。T2(黄土低丘)处于风沙区和黄土丘陵区的过渡区域,特征稳定性差;并且T2与丘陵沟壑区相同,区域内的部分小流域内部也存在着混合的地貌,这些特点都导致T2的分类精度较低。总体而言,基于RF的小流域尺度地貌分类的精度达到85%,取得了较好的分类结果。

4 结论与讨论

4.1 结论

本文以ASTER GDEM 30 m分辨率DEM为数据源,将小流域作为地貌分类的单元;通过数字地形分析计算了小流域多方面的特征;为了减少特征之间不必要的重复和部分噪声较大的特征的影响,本研究使用随机森林算法,基于Gini指数进行了特征的选择,通过随机森林分类器的OOB精度评估对随机森林分类器输入的参数进行了确定;进而构建随机森林分类器对所有的小流域进行了分类。相比基于单一地形因子或少量特征的地貌分类,使用随机森林基于高维特征的地貌分类具有更高的精度。主要结论如下:

(1)陕西省北部黄土高原上划分出的小流域得到了较好的分类结果,与人工判读相比,分类精度达到了85%,Kappa系数为0.83,表明RF算法与DEM数据结合进行地貌分类可以取得较好的效果。

(2)所有的小流域被分为8种地貌类型,同一地貌类型在空间上表现出明显的聚集性,不同地貌类型之间存在着边界和过渡区。具体来说,从北部的沙丘草滩盆地、黄土低丘到南部的黄土丘陵沟壑区、黄土塬区,存在着地貌类型的连续变化,其中部分地区由于高程较高和黄土层较薄显

现出石质山地的特征。

（3）不同的地貌区解释了不同区域的黄土沉积和流水侵蚀的不同情况，而不同的边界是黄土沉积、黄河侵蚀和构造分异组合过程的重要变化边界。

4.2 讨论

此研究将随机森林方法运用在了地貌分类上，取得了较好的分类结果，对地貌形态监督分类及自动分类的方法学研究具有较为重要的意义。但本研究也存在一些问题：

（1）首先实验使用的DEM数据的分辨率为30 m，可能会导致计算出的一些小流域特征存在较大的噪声；部分噪声进而会在训练随机森林分类器时使分类器出现过拟合的现象，还需要进一步的研究来选取适用于特定分辨率DEM的地形因子。

（2）随机森林算法需要进行训练，因此训练集的选择将会对分类结果产生很大的影响；对于本研究来说，地貌样区的数量和位置是分类结果的决定性因素；采用不同的地貌样区会导致分类结果的变化，这是监督分类难以避免的问题。

（3）研究发现，为了使小流域的属性稳定，划分出的各小流域面积较大，从而导致部分小流域内部含有不同类型地貌的现象，处于过渡区域（如黄土低丘）的小流域由于这个原因特征往往不够稳定，这些小流域难以进行分类；若使用分辨率更高的DEM数据能够使得小流域属性在较小的面积上稳定，从而解决混合地貌的问题。

（4）研究中基于小流域对研究区域进行地貌单元的划分，但地貌类型单元的还有其他的划分标准，如山麓线、沟谷线、坡折线，基于不同的划分标准进行地貌的划分和识别，是进一步研究的思路。

参考文献(References)：

[1] 李炳元,潘保田,韩嘉福.中国陆地基本地貌类型及其划分指标探讨[J].第四纪研究,2008,28(4):535-543. [Li B Y, Pan B T, Han J F. Basic terrestrial geomorphological types in China and their circumscriptions[J]. Quaternary Science, 2008,28(4):535-543.]

[2] 张寿根.现代地貌学[M].北京:科学出版社,2005. [Zhang S G. Mordern geomorphology[M]. Beijing: Science Press, 2005.]

[3] 杨景春,李有利等.地貌学原理[M].北京:北京大学出版社,2001. [Yang J C, Li Y L. Principles of geomorphology [M]. Beijing: Peking University Press, 2001.]

[4] 周成虎,程维明,钱金凯,等.中国陆地1∶100万数字地貌分类体系研究[J].地球信息科学学报,2009,11(6):707-724. [Zhou C H, Cheng W M, Qian J K, et al. Research on the classification system of digital land geomorphology of 1:1 000 000 in China[J]. Journal of Geo-information Science, 2009,11(6):707-724.]

[5] 裘善文,李风华.试论地貌分类问题[J].地理科学,1982,2(4):327-335. [Qiu S W, Li F H. On the problem of geomorphological classification in China[J]. Scientia Geographica Sinica, 1982,2(4):327-335.]

[6] Lee S W, Hwang S J, Lee S B, et al. Landscape ecological approach to the relationships of land use patterns in watersheds to water quality characteristics[J]. Landscape and Urban Planning, 2009,92(2):80-89.

[7] Wondzell S M, Cunningham G L, Bachelet D. Relationships between landforms, geomorphic processes, and plant communities on a watershed in the northern Chihuahuan Desert[J]. Landscape Ecology, 1996,11(6):351-362.

[8] Gordon J E, Brazier V, Lees G. Geomorphological systems: Developing fundamental principles for sustainable landscape management[J]//o'Halloran D, Green C, Harley M, Stanley M, Knill J(Eds.). Geological landscape conservation[M]. London: Geological Society, 1994:185-189.

[9] O'Neill R V, Hunsaker C T, Jones K B, et al. Monitoring environmental quality at the landscape scale: using landscape indicators to assess biotic diversity, watershed integrity, and landscape stability[J]. BioScience, 1997,47(8):513-519.

[10] Mark D M. Computer analysis of topography: A comparison of terrain storage methods[J]. Geografiska Annaler: Series A, Physical Geography, 1975,57(3-4):179-188.

[11] Speight J G. Landform pattern description from aerial photographs[J]. Photogrammetria, 1977,32(5):161-182.

[12] 周廷儒,施雅风,陈述彭.中国地形区划草案[J]//中国自然区划草案[M].北京:科学出版社,1956,56:21-56. [Zhou T R, Shi Y F, Chen S P. Draft of terrain regionalization of China[J]//Draft of natural regionalization of China[M]. Beijing: The Science Publishing Company, 1956,56:21-56.]

[13] Wood J. The geomorphological characterisation of digital elevation models[D]. Leicester: University of Leicester, 1996.

[14] Peucker T K, Douglas D H. Detection of surface-specific points by local parallel processing of discrete terrain elevation data[J]. Computer graphics and Image processing, 1975,4(4):375-387.

[15] O'Callaghan J F, Mark D M. The extraction of drainage networks from digital elevation data[J]. Computer vision, graphics, and image processing, 1984,28(3):323-344.

[16] 周启鸣,刘学军.数字地形分析[M].北京:科学出版社,

2006. [Zhou Q M, Liu X J. Digital Terrain Analysis[M]. Beijing: Science Press, 2006.]

[17] 汤国安,刘学军,闾国年.数字高程模型及地学分析的原理与方法[M].北京:科学出版社,2006. [Tang G A, Liu X J, Lv G N. Principles and Methods of Digital Elevation Model and Geological Analysis[M]. Beijing: Science Press, 2006.]

[18] Gallant J P W J C. Terrain analysis: Principles and applications[M]. London: John Wiley & Sons, 2000.

[19] Guth P L. Quantifying and visualizing terrain fabric from digital elevation models[C]. International Conference on GeoComputation, 4th, Fredericksburg VA, Mary Washington College. 1999:25-28.

[20] MacMillan R A, Pettapiece W W, Nolan S C, et al. A generic procedure for automatically segmenting landforms into landform elements using DEMs, heuristic rules and fuzzy logic[J]. Fuzzy sets and Systems, 2000,113(1):81-109.

[21] Yokoyama R, Shirasawa M, Pike R J. Visualizing topography by openness: A new application of image processing to digital elevation models[J]. Photogrammetric engineering and remote sensing, 2002,68(3):257-266.

[22] Irvin B J, Ventura S J, Slater B K. Fuzzy and isodata classification of landform elements from digital terrain data in Pleasant Valley, Wisconsin[J]. Geoderma, 1997,77(2-4):137-154.

[23] Blaschke T. Object based image analysis for remote sensing[J]. ISPRS journal of photogrammetry and remote sensing, 2010,65(1):2-16.

[24] Drăguţ L, Eisank C. Object representations at multiple scales from digital elevation models[J]. Geomorphology, 2011,129(3-4):183-189.

[25] Verhagen P, Drăguţ L. Object-based landform delineation and classification from DEMs for archaeological predictive mapping[J]. Journal of Archaeological Science, 2012, 39(3):698-703.

[26] Zhao W, Xiong L, Ding H, et al. Automatic recognition of loess landforms using Random Forest method[J]. Journal of Mountain Science, 2017,14(5):885-897.

[27] Huang S L, Ferng J J. Applied land classification for surface water quality management: II. Land process classification[J]. Journal of Environmental Management, 1990,31 (2):127-141.

[28] 刘双琳,李发源,蒋如乔,等.黄土地貌类型的坡谱自动识别分析[J].地球信息科学学报,2015,17(10):1234-1242.[Liu S L, Li F Y, Jiang R Q, et al. A Method of Loess Landform Automatic Recognition Based on Slope Spectrum[J]. Journal of Geo- information Science, 2015,17 (10):1234-1242.]

[29] 周毅.基于DEM的黄土高原正负地形及空间分异研究[D].南京:南京师范大学,2011. [Zhou Y. DEM based research on positive-negative terrains and their spatial variation on loess plateau[D]. Nanjing: Nanjing Normal University, 2011.]

[30] Xiong L Y, Zhu A X, Zhang L, et al. Drainage basin object-based method for regional-scale landform classification: a case study of loess area in China[J]. Physical Geography, 2018,39(6):523-541.

[31] 张磊.基于核心地形因子分析的黄土地貌形态空间格局研究[D].南京:南京师范大学,2013. [Zhang L. Core factor analysis based research on spatial characteristics on loess plateau[D]. Nanjing: Nanjing Normal University, 2013.]

[32] Breiman L. Random forests[J]. Machine learning, 2001,45 (1):5-32.

[33] Gislason P O, Benediktsson J A, Sveinsson J R. Random forests for land cover classification[J]. Pattern Recognition Letters, 2006,27(4):294-300.

[34] Pal M. Random forest classifier for remote sensing classification[J]. International Journal of Remote Sensing, 2005,26(1):217-222.

[35] Ham J, Chen Y, Crawford M M, et al. Investigation of the random forest framework for classification of hyperspectral data[J]. IEEE Transactions on Geoscience and Remote Sensing, 2005,43(3):492-501.

[36] Tang G A, Li F Y, Liu X J, et al. Research on the slope spectrum of the Loess Plateau[J]. Science in China Series E: Technological Sciences, 2008,51(1):175-185.

[37] 蔡凌雁,汤国安,熊礼阳,等.基于DEM的陕北黄土高原典型地貌分形特征研究[J].水土保持通报,2014,34(3):141-144. [Cai L Y, Tang G A, Xiong L Y. An analysis on fractal characteristics of typical landform patterns in northern Shaanxi loess plateau based on DEM[J]. Bulletin of Soil and Water Conservation, 2014,34(3):141-144.]

[38] 汤国安,李发源,杨昕,等.黄土高原数字地形分析探索与实践[M].北京:科学出版社,2015. [Tang G A, Li F Y, Yang X,et al. Exploration and practice of digital terrain analysis on Loess Plateau[M]. Beijing: Science Press, 2015.]

[39] Nikolakopoulos K G, Kamaratakis E K, Chrysoulakis N. SRTM vs ASTER elevation products. Comparison for two regions in Crete, Greece[J]. International Journal of remote sensing. 2006,27(21):4819-38.

[40] Tarboton D G, Bras R L, Rodriguez-Iturbe I. On the extraction of channel networks from digital elevation data [J]. Hydrological processes, 1991,5(1):81-100.

[41] 王春.基于DEM的陕北黄土高原地面坡谱不确定性研

究[D].西安:西北大学,2005. [Wang C. The uncertainty of slope spectrum derived from DEM in the loess plateau of northern Shaanxi Province[D]. Xi'an: Northwest University, 2005.]

[42] 贾旖旎.基于DEM的黄土高原流域边界剖面谱研究[D].南京:南京师范大学,2010. [Jia Y N. Research on catchment boundary profile spectrum based on digital elevation models[D]. Nanjing: Nanjing Normal University, 2010.]

[43] 祝士杰.基于DEM的黄土高原流域面积高程积分谱系研究[D].南京:南京师范大学,2013. [Zhu S J. Research on watershed hypsometric integral in the loess plateau based on digital elevation models[D]. Nanjing: Nanjing Normal University, 2013.]

[44] Jenson S K, Domingue J O. Extracting topographic structure from digital elevation data for geographic information system analysis[J]. Photogrammetric engineering and remote sensing, 1988,54(11):1593-1600.

[45] Strahler A N. Quantitative analysis of watershed geomorphology[J]. Eos, Transactions American Geophysical Union, 1957,38(6):913-920.

[46] 陈浩.陕北黄土高原沟道小流域形态特征分析[J].地理研究,1986,5(1):82-92. [Chen H. A preliminary study on geomorphic features of small drainage basins on the loess plateau in northern Shaanxi[J]. Geographical Research, 1986,5(1):82-92.]

[47] 崔灵周,李占斌,朱永清,等.流域地貌分形特征与侵蚀产沙定量耦合关系试验研究[J].水土保持学报,2006,20(2):1-4,9. [Cui L Z, Li Z B, Zhu Y Q, et al. Experimental Study on quantitative coupling relationship between topographic fractal feature and sediment yield in small watershed[J]. Journal of Soil and Water Conservation, 2006,20(2):1-4,9.]

[48] 张婷,汤国安,王春,等.黄土丘陵沟壑区地形定量因子的关联性分析[J].地理科学,2005,25(4):85-90. [Zhang T, Tang G A, Wang C, Long Y, Wu L C, Wang Z. Correlation of quantitative terrain factors in Gully Hill areas of China Loess Plateau[J]. Scientia Geographica Sinica, 2005,25(4):85-90.]

[49] 朱红春,刘海英,张继贤,等.基于DEM的流域地形因子提取与量化关系研究——以陕北黄土高原的实验为例[J].测绘科学,2007(2):138-140,182. [Zhu H C, Liu H Y, Zhang J X, et al. Research on the topographic factors and its mathematical simulation based on DEMs - a case study in the loess plateau of northern Shaanxi Province [J]. Science of Surveying and Mapping, 2007(2):138-140, 182.]

[50] 汤国安,李发源,刘学军,等.数字高程模型教程[M].北京:科学出版社,2016. [Tang G A, Li F Y, Liu X J. Digital Elevation Model Course[M]. Beijing: Science Press, 2016.]

[51] 陶旸.基于纹理分析方法的DEM地形特征研究[D].南京:南京师范大学,2011. [Tao Y. Texture analysis based on research on terrain morphology characteristics[D]. Nanjing: Nanjing Normal University, 2011.]

[52] 谢轶群,朱红春,汤国安,等.基于DEM的沟谷特征点提取与分析[J].地球信息科学学报,2013,15(1):61-67. [Xie Y Q, Zhu H C, Tang G A, et al. Extraction and analysis of gully feature points based on DEMs[J]. Journal of Geo-information Science, 2013,15(1):61-67.]

[53] 薛凯凯,熊礼阳,祝士杰,等.基于DEM的黄土崾岘提取及其地形特征分析[J].地球信息科学学报,2018,20(12):1710-1720. [Xue K K, Xiong L Y, Zhu S J, et al. Extraction of loess dissected saddle and its terrain analysis by using digital elevation models[J]. Journal of Geo-information Science, 2018,20(12):1710-1720.]

[54] Goldstein B A, Polley E C, Briggs F B S. Random forests for genetic association studies[J]. Statistical applications in genetics and molecular biology, 2011,10(1):32.

[55] Friedman J H. Greedy function approximation: A gradient boosting machine[J]. Annals of statistics, 2001,29(5):1189-1232.

[56] 方匡南,吴见彬,朱建平,等.随机森林方法研究综述[J].统计与信息论坛,2011,26(3):32-38. [Fang K N, Wu J B, Zhu J P, et al. A Review of Technologies on Random Forests[J]. Statistics & Information Forum, 2011,26(3):32-38.]