文章编号: 1008 - 2786 - (2016) 4 - 432 - 10
DOI: 10.16089/j.cnki.1008 - 2786.000148

滑坡危险度评价对 BCS 负样本采样的敏感性

缪亚敏[1], 朱阿兴[1,2,3], 杨 琳[2], 白世彪[1], 刘军志[1], 邓永翠[1]

(1. 虚拟地理环境教育部重点实验室(南京师范大学), 江苏省地理环境演化国家重点实验室培育建设点, 江苏省地理信息资源开发与利用协同创新中心, 江苏 南京 210023;
2. 中国科学院地理科学与资源研究所 资源与环境信息系统国家重点实验室, 北京 100101;
3. Department of Geography, University of Wisconsin - Madison, Madison, WI 53706, USA)

摘 要: 滑坡负样本在基于统计方法的滑坡危险度制图中具有重要作用, 能够抑制统计方法对滑坡危险度的高估。缓冲区控制采样(Buffer controlled sampling, BCS)是一种广泛使用的负样本采样方法, 其原理是认为滑坡点附近一定范围内的地理环境与滑坡点所在的地理环境相似, 易发生滑坡, 因而应当在灾害点一定缓冲区以外的区域采集负样本。目前缓冲区大小主要是根据专家对研究区的经验知识确定, 具有主观性。缓冲区大小对基于统计方法的滑坡危险度制图的影响研究较少。因此, 有必要分析缓冲区大小与滑坡危险度制图精度之间的关系, 探究适宜的缓冲区大小。以陇南山区的油坊沟流域为研究区, 基于 BCS 负样本采样方法, 探究不同缓冲区大小对基于支持向量机(Support vector machine, SVM)的滑坡危险度制图结果的影响。结果表明: 缓冲区过小会导致与滑坡点地理环境相似的假的负样本的存在, 从而导致滑坡危险度的低估; 缓冲区过大会导致负样本在环境特征空间中太局限, 负样本集的全局代表性差, 从而导致滑坡危险度的高估。在本研究区基于 SVM 的滑坡危险度制图中, 200 ~ 500 m 是使用 BCS 采集负样本的较理想的缓冲区大小。

关键词: SVM; 负样本; 缓冲区控制采样; 缓冲区大小; 滑坡危险度制图
中图分类号: P642.22 **文献标志码:** A

滑坡灾害的频繁发生, 造成经济损失与人员伤亡, 不利于社会的安定与发展[1-3]。鉴别未来可能发生滑坡的区域, 是应对滑坡灾害的基础, 也是当前滑坡研究中的重点内容。滑坡危险度是指在当地地形等条件下区域发生滑坡的可能性, 即回答“什么地方容易发生滑坡”[4-6]。通过对滑坡危险度进行推测制图, 识别滑坡高危险区域, 有针对性地实施相关措施, 可以有效减少滑坡带来的损失, 对地区规划和建设有重要参考意义[7-8]。

统计方法, 又称为数据驱动模型, 包括常规统计方法和机器学习算法, 是滑坡危险度制图中最常使用的方法[9-12]。其原理是“过去决定未来”, 即与过去发生过滑坡的区域具有相似地理环境的地区也极易发生滑坡[13-14]。统计方法通常是从滑坡点(正样本)和非滑坡点(负样本), 以及这些点所在的影响因素中获取滑坡发生可能性与影响因素之间的关系, 然后将这种关系应用到整个研究区, 实现区域滑坡危险度的推测与制图。

滑坡负样本在基于统计方法的区域滑坡危险度制图中具有重要作用, 能够抑制统计方法对滑坡危险度的高估[15], 以合理区划滑坡高危险区与低危险区, 因此滑坡负样本的研究越来越引起研究者们的关注[15]。考虑到样本采集的随机性和空间均匀性, 滑坡负样本一般是在研究区中没有发生过滑坡的区

收稿日期(Received date): 2015 - 11 - 02; 修回日期(Accepted): 2015 - 11 - 16。
基金项目(Foundation item): 国家自然科学基金项目(41431177, 41471178); 江苏省高校自然科学研究重大项目(14KJA170001); 国家重点基础研究发展计划 973 项目(2015CB954102)。[This study is supported by the National Natural Science Foundation of China (41431177), the Natural Science Research Program of Jiangsu(14KJA170001)], the National Basic Research Program of China (2015CB954102).]
作者简介(Biography): 缪亚敏(1991 -), 女, 江苏泰州, 硕士研究生, 从事滑坡危险度评价研究。[Miao Yamin, Female, Master Candidate, Born in Taizhou, Jiangsu, Major in landslide susceptibility mapping.] E - mail: miaopaper@163.com

域随机采样获得。缓冲区控制采样法(Buffer controlled sampling，BCS) 是一种使用较广泛的滑坡负样本采样方法[16-18]，其原理是：认为滑坡发生点附近一定范围内的地理环境与滑坡点所在的地理环境相似，因而其孕育滑坡的可能性非常大。因此，在采集滑坡负样本时，需要避开滑坡点，在滑坡点一定距离(缓冲区) 之外的区域随机采样。目前 BCS 方法中缓冲区大小主要是根据专家对研究区的经验知识确定。方苗等[18]根据对研究区的分析和认知，主观确定缓冲区大小。Xiao 等[16]将缓冲区大小定为所有滑坡面大小的平均值，缓冲区大小的确定没有统一标准。此外，不同缓冲区大小对基于统计模型的滑坡危险度制图的影响研究甚少。因此，有必要分析缓冲区大小与滑坡危险度制图精度之间的关系，探究适宜的缓冲区大小。

本文以在滑坡危险度评价方面广泛使用的支持向量机(Support vector machine，SVM) 为推测模型，选择滑坡频发的陇南山区油坊沟流域为研究区，根据 BCS 负样本采样方法，设置不同的缓冲区大小，在缓冲区以外且避开水系的区域随机采集滑坡负样本作为负样本集；组合负样本集与滑坡点以构成训练样本集，对油坊沟流域构建基于 SVM 的滑坡危险度推测模型，实现油坊沟流域滑坡危险度制图；分析缓冲区大小与基于 SVM 的滑坡危险度制图精度之间的关系，并探究油坊沟流域适宜的负样本采样缓冲区大小。

1　研究区与数据

1.1　研究区与历史滑坡编目

油坊沟流域位于甘肃省武都县安化镇(图 1)，流域面积 49.4 km^2，周边分布着青藏高原、黄土高原和四川盆地，地质构造复杂，地震频发。油坊沟上游分水岭为石灰岩溶蚀夷平面，是天然草场，下游堆积着冲积物；地势北高南低，平均海拔 2 000 m 以上，地势起伏较大，河谷深切，平均坡度在 20°以上。流域内泥盆系、志留系分布广泛，岩性多为千枚岩、板岩、粉砂岩、泥岩等。该研究区亚热带季风气候显著，气候温暖湿润，受山地地形效应的影响，全年降水量达到 400～900 mm[19]，降水集中在 5—9 月，且多以暴雨的形式出现。受特殊地质环境和气候环境的影响，流域内滑坡灾害频繁发生，造成极大的人员伤亡，严重制约着该区工农业生产和社会经济的发展[3]。

根据室内遥感解译与野外检核，记录滑坡名称、地理位置、前缘高程、后缘高程、面积和体积等，构建油坊沟流域历史滑坡编目数据库。本研究区共识别 65 个滑坡，分别为 12 个基岩滑坡、17 个崩塌和 36 个黄土滑坡。所有滑坡的总面积达 7.24 km^2，总体积达 107.8 km^3[20-21]，其中最小的滑坡面积为 5.23 $\times 10^{-2}$ km^2，最大的滑坡面积达到 0.46 km^2。

滑坡正样本以滑坡编目为数据源，其采样方法

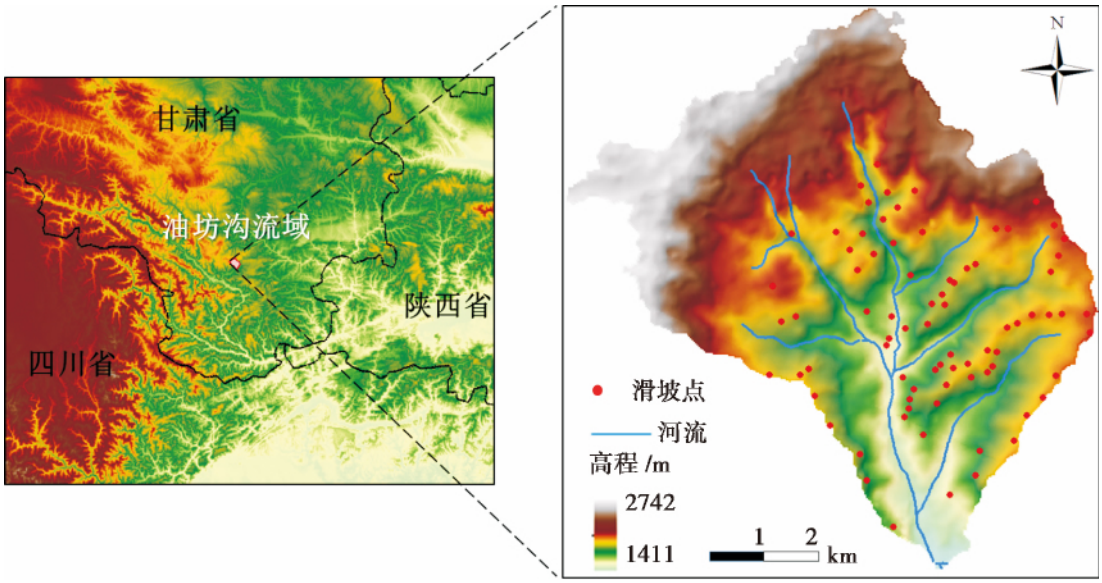

图 1　油坊沟流域地理位置和滑坡灾害点分布图

Fig. 1　Location of the Youfang catchment and the spatial distribution of landslides

较多,如 Seed cell、滑坡发生内斜坡区、滑坡发生内斜坡区的顶部中心、滑坡发生区模糊 C 均值聚类等[22]。本文用滑坡发生内斜坡区的顶部中心点代表滑坡点,构成滑坡正样本,不仅可以避免由于面状滑坡边界难以确定而带来的误差,也可以兼顾不同大小的滑坡[22]。基于这一正样本采样方法,油坊沟流域共采集 79 个滑坡正样本(图 1)。

1.2 影响因素

滑坡是内外影响因素综合作用的产物。内部因素又称为孕灾环境,是指斜坡本身具备的易于产生崩滑的内在条件,一般包括地貌、地质、土地利用、排水状况等。外部因素又称为诱发因素,是使斜坡内在条件发生变化,导致斜坡失去平衡形成滑坡的关键因素,如:地震、强降水等[23]。滑坡危险性是对斜坡所处的内在因素的综合评估,不考虑滑坡动态诱发因素,不涉及滑坡发生时间和滑坡量级问题[7-8]。

根据研究区的地质环境特征和前人已有研究成果,选取高程、坡度、坡向、平面曲率、剖面曲率、距河流的距离、距道路的距离、岩性、距断层线的距离、土地利用类型等 10 个影响因素(表 1),用以对研究区内的滑坡孕灾环境进行定量描述。

表 1 影响因素与数据源

Tab. 1 Predisposing factors and data sources

类别	影响因素	数据源	比例尺
地貌	高程	地形图	1:5 万
	坡度		
	坡向		
	平面曲率		
	剖面曲率		
排水状况	距河流的距离		
人类活动	距道路的距离		
地质条件	岩性	地质图	1:2.5 万
	距断层线的距离		
土地利用	土地利用类型	TM 影像	1:10 万

地貌是滑坡发育的空间因素[24]。高程影响水系发育程度、植被覆盖、土地利用等[25];坡度制约着重力和流水侵蚀作用的强度,一定程度上为滑坡的形成与位移提供了临空面[18];坡向影响斜坡的光热分配和水分分配,间接影响斜坡地下水孔隙压力分布和岩土体物理学特征,对斜坡的稳定性有重要作用[23];曲率是重要的地表几何形态变量,直接影响地表径流的汇集,进而影响岩土体的强度变化和坚硬程度[26-27]。对研究区 1:5 万地形图数字化,生成 30 m 分辨率的数字高程模型(DEM),在 ArcGIS10.1 软件下基于 DEM 派生出一系列地形因子,包括:高程、坡度、坡向、平面曲率、剖面曲率,以描述研究区的地貌条件。

水系的分布与滑坡发生有重要关联,河流侵蚀作用对坡面具有极大的影响,河流不断的掏蚀坡脚,为滑坡的发生提供了许多临空面[25];道路体现了人类工程活动对滑坡发生的影响,公路的修建会改变斜坡坡体的自然结构,修建后留下的大量松散固体物质,是滑坡发生的物质基础[18]。本文对研究区 1:5万地形图数字化,获得矢量格式的河流和道路数据,河流和道路对滑坡的影响是根据距它们的距离表达。

对研究区 1:2.5 万地质图数字化,获得地层分布和断层构造数据。同时代的地层可由不同成因的岩石类型组成,根据相似的物质组成和物理力学特征,把研究区的地层组合并划分为厚层石灰岩、板岩岩组,粉砂岩、泥岩、薄层砂砾岩岩组,以及千枚岩、板岩、薄层灰岩岩组 3 类岩性单元。断层构造对滑坡的影响是根据距断层线的距离表达。

土地利用类型决定着植被覆盖,影响地表涵养水源的能力,影响斜坡坡体的稳定性[28]。对研究区的 TM 影像进行遥感解译,获得土地利用类型,并将其分为四类:农田、林地、居民地与工业用地、以及未利用土地。

本研究中,为避免因离散型影响因素(如坡向、岩性和土地利用类型)而导致哑变量过多的问题,使用滑坡发生频率表达离散型影响因素[28]。例如岩性数据包括 3 类岩性单元,每一类岩性单元可以表达为该单元中出现滑坡数量与所有岩性单元中出现滑坡数量的比值。此外,为方便空间分析与计算,需要保证影响因素数据尺度的一致性。根据研究区的尺度和数据源的尺度,选择 30 m 作为所有影响因素数据的分辨率。

2 研究方法

2.1 基于不同缓冲区大小的负样本采样设计

基于油坊沟流域历史滑坡编目获得 79 个滑坡点,将滑坡点分为滑坡正训练样本 63 个(80%)和正检验样本 16 个(20%)。以所有滑坡点(79 个滑坡点)为中心,分别设置从 50 m 到 1 500 m 不等的、

间距为 50 m 的缓冲区大小。在缓冲区以外且避开水系的区域,随机采集与滑坡正训练样本同等数量(63 个)的滑坡负样本数据,与滑坡正训练样本一起,构成不同缓冲区大小下的训练样本集。在每个缓冲区大小下,重复采样 10 次,得到同一个缓冲区大小下 10 套不同的训练样本集,以避免偶然现象,真实反映同一个缓冲区大小下基于统计方法的滑坡危险度制图的一般规律。

2.2　基于 SVM 的滑坡危险度制图

支持向量机(SVM)是 Vapnik 等提出的一种新的机器学习方法[29-30]。其基本思想是: 将训练样本通过某一核函数映射到一个高维特征空间,在高维特征空间中建立线性回归函数,寻找最优分类超平面,使得该超平面能够尽可能多的将两类训练样本正确分开,同时使得分开的两类样本距离分类超平面最远(图 2)[31-32]。

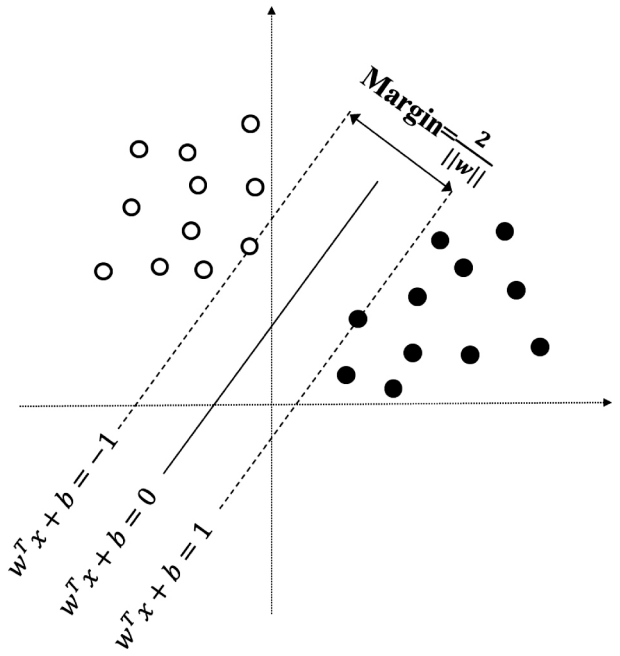

图 2　SVM 的基本思想

Fig. 2　Basic idea of SVM

给定训练样本集$\{x_i, y_i\}$, $i=1,2,\cdots,n$; $x_i \in R^m$, $y_i \in \{-1,+1\}$; n 为训练样本数, m 为输入向量的维数(本研究中是指影响因素的个数),假设有一个线性分类超平面可以将这两类训练样本完全分开,定义该超平面如下:

$$w \cdot x + b = 0 \tag{1}$$

SVM 要求构建分类面将所有样本正确分类,同时使得分开的两类训练样本距离分类超平面最远,这是一个最优化问题,构造如下目标函数:

$$\max \quad \frac{2}{|w|^2} \tag{2}$$

$$\text{s.t.} \quad y_i((w \cdot x_i) + b) \geqslant 1 \quad i=1,2,\cdots,n \tag{3}$$

此外,考虑到有些训练样本会被错误分类,Vpanik 和 Cortes 等引入非负的松弛变量 ξ_i,则上述带约束的最优化问题修改成如下目标函数:

$$\min \quad \frac{|w|^2}{2} + + C\sum_i \xi_i \tag{4}$$

$$\text{s.t.} \quad y_i((w \cdot x_i) + b) \geqslant 1 - \xi_i \qquad \xi_i \geqslant 0, i=1\cdots n, i=1,2,\cdots,n \tag{5}$$

当样本分类错误的时候 ξ_i 就会大于 0。C 是对误判样本的惩罚程度, $C>0$,是一可调常数, C 越大代表对分类错误样本惩罚越重。因此,在求分类超平面的时, $C\sum_i \xi_i$ 的值越小越好[29]。

通常训练样本集是线性不可分的,SVM 通过引入核函数将线性不可分的训练样本集映射到高维空间中,使其线性可分。SVM 中常使用的核函数包括: 线性核函数、多项式核函数和高斯核函数。

基于训练样本集构建 SVM 时,需要选择适宜的核函数并设置相应的参数。高斯核函数由于其具有较强的非线性映射能力,在滑坡危险度制图中广泛使用[33-35]。本文选择高斯核函数作为 SVM 的核函数。Y 参数是高斯核函数中的重要参数,其大小关系着高斯核函数的形状。惩罚因子 C 是对分类错误样本的惩罚程度。这两个参数的组合直接影响支持向量机的泛化能力。本文采用交叉验证寻求最小 MSE 的方法,基于台湾大学林智仁教授开发的 LIBSVM 工具[36],通过网格搜索方法对训练样本进行分组交叉验证,找到 SVM 的最优参数 C 和 Y。

根据训练样本和最优参数,构建基于 SVM 的滑坡危险度推测模型。将整个研究区的影响因素数据输入到构建好的基于 SVM 的滑坡危险度预测模型中,即可推测整个研究区滑坡危险度的空间分布。根据上述负样本采样方法中获得的不同训练样本集分别构建 SVM 滑坡危险度预测模型,推测不同训练样本集下的区域滑坡危险度空间分布。

2.3　滑坡危险度制图的精度和有效性评价

本文通过以下四个指标对滑坡危险度制图的精度和有效性进行评价,指标一、二是对制图的精度进行评价,指标三、四是对制图的有效性进行评价。

评价指标一为建模精度。建模精度是统计模型将训练样本正确分类的比率,用以衡量模型对训练样本的拟合程度。本文对参与建模的 126 个训练样

本进行建模精度评价。

评价指标二为验证精度。验证精度是验证样本的分类正确率,用以衡量统计模型的的预测能力。本文旨在探讨不同缓冲区下的负样本对制图精度的影响,由于负样本数据的质量无法直接判断,因此只使用 16 个正检验样本来衡量模型的验证精度。通过对正检验样本设置一个滑坡危险度阈值(本文设置阈值为 0.5),认为大于该阈值的样本分类正确,否则分类错误,如此即可度量检验样本分类正确的比率。

评价指标三为滑坡危险度空间分布。与大部分地理现象一样,滑坡危险度本质上是一种空间连续型地理变量,其在空间域中应具有一定的连续性和渐变性。本文对滑坡危险度推测结果图的空间分布进行定性分析,评价模型的有效性。

评价指标四为滑坡危险区面积所占比重与落在该危险区域内的滑坡点比重之间的关系(下文简称"危险面积 - 滑坡点关系")。一个有效的推测模型不仅应该具有高验证精度,还应该使得推测出的滑坡高危险区面积尽可能小,即应具有"使更多的滑坡灾害点落在有限面积的高危险区域内"这一特点[16],这样的模型可以高效地识别出真正的滑坡危险区域,有利于土地资源的充分利用和价值实现。本文使用滑坡发生内斜坡面作为滑坡灾害点,将研究区内的所有栅格点按滑坡危险度由大到小排序,统计不同危险度值域下的累积面积比重,并统计落在该危险度值域范围内的滑坡灾害点比重,以获得危险面积 - 滑坡点关系,基于这一关系评价模型的有效性。

3 结果

3.1 不同缓冲区大小下的建模精度和验证精度

统计不同缓冲区大小下 10 次重复负样本采样的建模精度(图 3),可以发现,在缓冲区较小(50 ~ 150 m)时,SVM 的建模精度低于 70.0%;随着缓冲区的增大,建模精度逐渐增大,模型对训练样本的拟合程度逐渐增高;当缓冲区大于 800 m 以后,建模精度趋于平稳达到 100%,模型对所有训练样本都拟合。

统计不同缓冲区大小下 10 次重复负样本采样的验证精度(图 4),可以发现,随着缓冲区的增大,滑坡正检验样本的分类正确率不断上升;当缓冲区超过 700 m 时,滑坡检验样本的正确率趋于平稳达到 100%。以上说明模型对滑坡灾害点的识别能力随着缓冲区的增大而增强。

3.2 不同缓冲区大小下的模型有效性

通过以上分析以及观察不同缓冲区大小下推测的滑坡危险度图和危险面积 - 滑坡点关系曲线图,将缓冲区大小分成 4 个区间进行分析:50 ~ 200 m,200 ~ 500 m,500 ~ 800 m,800 ~ 1500 m,每一区间内滑坡危险度空间分布图类似,危险面积 - 滑坡点关系曲线图也类似。在每一个缓冲区区间内各选择一个缓冲区大小下的推测危险度图进行分析,以增加

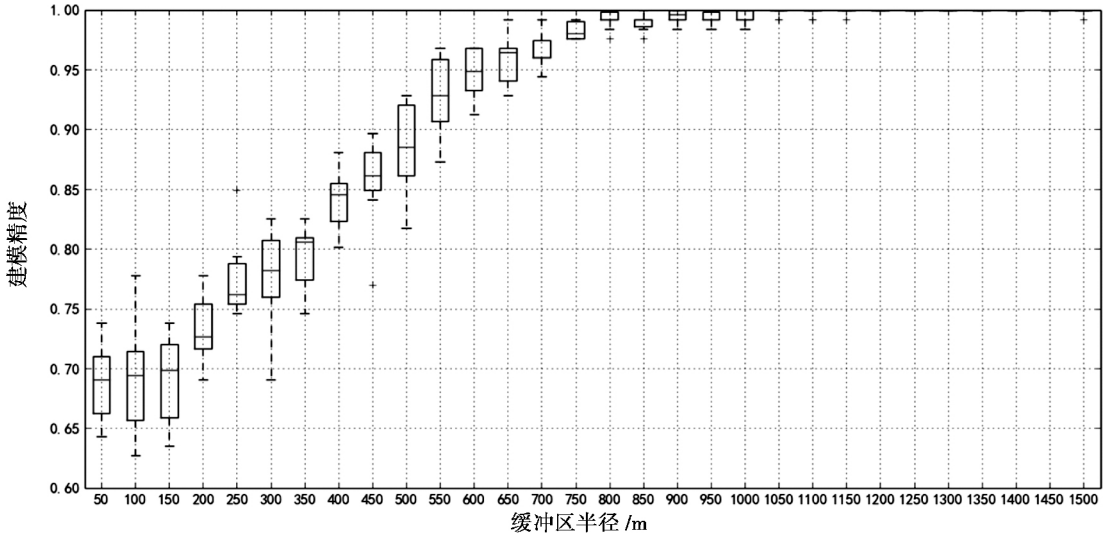

图 3 建模精度盒图

Fig. 3 Boxplot of modelling accuracy

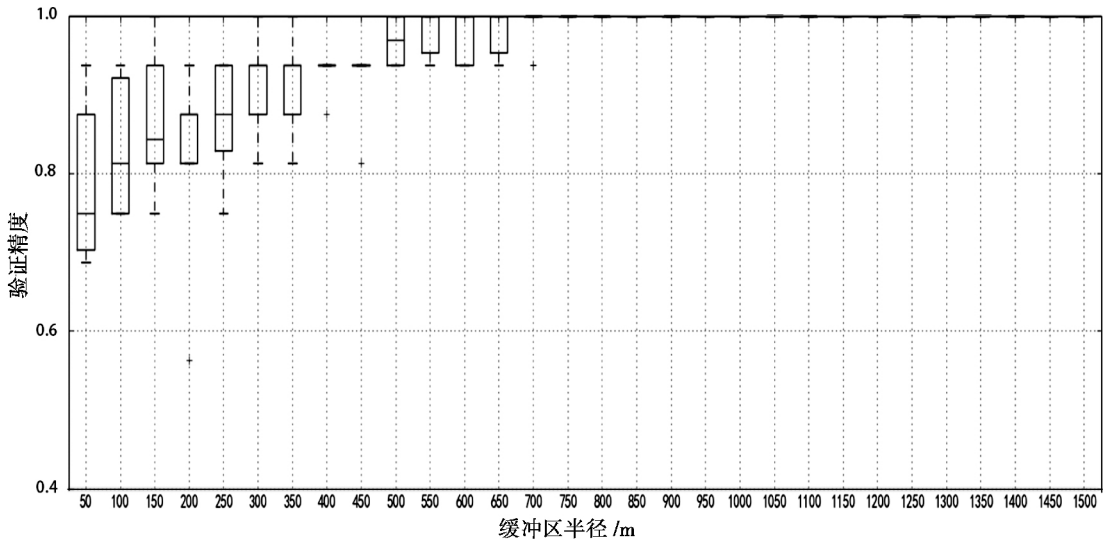

图 4　验证精度盒图

Fig. 4　Boxplot of validation accuracy

成图的可读性,便于对结果进行归纳。每个缓冲区区间内选取区间的中值作为这一区间的代表,如 50 ~200 m 取 100 m,200 ~500 m 取 350 m,500 ~800 m 取 650 m,800 ~1 500 m 取 1 150 m。对遴选出的 4 个缓冲区大小下的滑坡危险度分布图(图 5) 和危险面积 - 滑坡点关系曲线图(图 6) 进行分析。

3.2.1　滑坡危险度空间分布

缓冲区大小在 50 ~200 m 时,以缓冲区大小等于 100 m 为例,滑坡危险区围绕在滑坡发生点周围,较少向外扩展,斜坡面的危险度较大,沟谷处危险度较小,滑坡危险度在空间分布上具有一定的连续性和渐变性 [图 5(a)]。当缓冲区大小增大到一定的阈值(200 ~500 m) 时,以缓冲区大小等于 350 m 为例,滑坡危险区不仅围绕在滑坡发生点周围,向周围空间也有一定程度的扩张。沿着斜坡面向下到达沟谷,滑坡危险度逐渐发生变化,空间变异性较好 [图 5(b)]。当缓冲区大小增大到一定的阈值(500 ~ 800 m) 时,以缓冲区大小等于 650 m 为例,滑坡危险区向外继续扩张,整个斜坡面和部分沟谷的滑坡危险度很高,空间变异性较差 [图 5(c)]。当缓冲区大小在 800 ~1 500 m 时,以缓冲区大小等于1 150 m 为例,研究区中大部分区域的滑坡危险度很高,无论是斜坡面还是沟谷都被推测为高危险区,空间变异性极差 [图 5(d)]。

3.2.2　危险面积 - 滑坡点关系

缓冲区大小在 50 ~200 m 时,以缓冲区大小等于 100 m 为例,根据危险面积 - 滑坡点关系曲线图

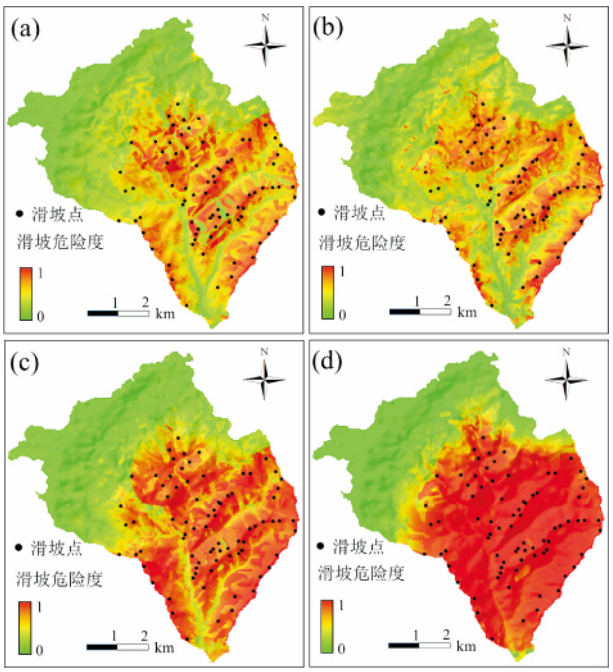

图 5　不同缓冲区大小下的滑坡危险度空间分布图

[a) —100 m,(b) —350 m,(c) —650 m,(d) —1 150 m]

Fig. 5　Derived landslide susceptibility maps from different buffer sizes

[a) —100 m,(b) —350 m,(c) —650 m,(d) —1 150 m]

(图 6 红线) 可以发现,SVM 滑坡危险度推测模型可以在较小的滑坡高危险面积内高效地识别出部分滑坡点,但是其需要在很大的面积内(75%) 才能识别出所有滑坡点,这说明模型对滑坡点的预测能力较低,滑坡危险度图中存在不少落在滑坡低危险区的滑坡灾害点,这也印证了在该缓冲区阈值下模型较

低的建模精度和验证精度。

当缓冲区大小增大到一定的阈值(200 ~ 500 m)时,以缓冲区大小等于 350 m 为例,根据危险面积 - 滑坡点关系曲线图(图 6 蓝线)可以发现,SVM 滑坡危险度推测模型可以在较小的滑坡高危险面积内高效地识别出部分滑坡点,在较小的面积内(60%)识别出所有滑坡点,滑坡点多落在推测滑坡危险度高的区域,这证明该缓冲区阈值下 SVM 推测模型的有效性,也印证了在该缓冲区阈值下模型较高的建模精度与验证精度。

当缓冲区大小增大到一定的阈值(500 ~ 800 m)时,以缓冲区大小等于 650 m 为例,根据危险面积 - 滑坡点关系曲线图(图 6 绿线)可以发现,SVM 滑坡危险度推测模型可以在较小的面积内(63%)高效地识别出所有滑坡点,这也印证了在该缓冲区阈值下模型较高的建模精度与验证精度,但是模型在较小的滑坡高危险面积内识别出的滑坡点较少,这说明该缓冲区阈值下 SVM 模型推测能力有限。根据图 5(c)发现,模型虽然能识别出所有滑坡点,但是其却以很大面积的滑坡高危险区为代价,这是没有意义的,这样的危险度推测结果会高估滑坡发生的可能性,造成土地资源的浪费和不合理使用。

当缓冲区大小在 800 ~ 1 500 m 时,以缓冲区大小等于 1 150 m 为例,根据危险面积 - 滑坡点关系曲线图(图 6 黑线)可以发现,与缓冲区大小在 550 ~ 800 阈值时类似,SVM 滑坡危险度推测模型可以在较小的面积内(63%)高效地识别出所有滑坡点,这也印证了在该缓冲区阈值下模型极高的建模精度与验证精度,但是相同面积的滑坡高危险区域中所包含的滑坡点却更少,这说明该缓冲区阈值下 SVM 模型推测能力更差。根据图 5(d)发现,模型虽然能识别出所有滑坡点,但是其却以极大面积的滑坡高危险区为代价,比缓冲区大小在 550 ~ 800 阈值时模型的预测能力更为恶化,模型的过度高估造成推测结果的不合理与不可信。

4 讨论

对以上不同缓冲区大小阈值分区结果进行原因分析,当缓冲区大小在 50 ~ 200 m 内时,缓冲区较小,负样本可以采样的区域较大,负样本的空间均衡性较好。但是由于缓冲区过小,有可能采集到不少与滑坡发生点地理环境相似的样本,这些样本本该

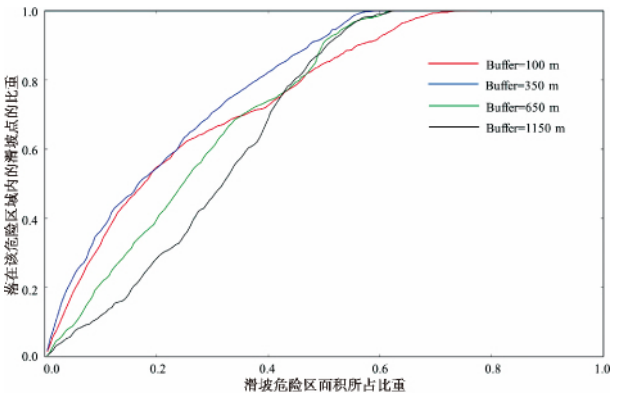

图 6 不同缓冲区大小下的危险面积——滑坡点关系曲线图

[a) —100 m,(b) —350 m,(c) 650 m,(d) —1 150 m]

Fig. 6 Relationships between the proportion of landslide prone zones and the percentage of landslides from different buffer sizes

[a) —100 m,(b) —350 m,(c) —650 m,(d) —1 150 m]

是具有滑坡高危险性,却被选作为负样本,会导致负样本集中出现假的负样本。SVM 是在由影响因素构成的环境特征空间中区分正负样本,当负样本中掺杂一些假的负样本时,SVM 很难在环境特征空间中正确区分正负样本,SVM 的建模精度会较低。此外,SVM 最优分类超平面更关心两类样本中更靠近分类面的样本,当负样本中掺杂一定量的假的负样本时,原本应该合理分割正负样本的 SVM 分类超平面会向正样本处偏移,造成 SVM 预测的滑坡危险度出现低估,研究区中划分为低危险区的面积较大,滑坡检验样本更有可能落在非危险区,因此验证精度较低。甚至,当训练样本集中假的负样本达到一定数量之后,SVM 会无法区分正负样本,导致推测的滑坡高危险区和低危险区颠倒,实际滑坡点都落在预测的滑坡低危险区中。

当缓冲区大小在 200 ~ 500 m 内时,随着缓冲区大小的增大,负样本的空间采样范围会逐渐减小,随着与滑坡点的逐渐远离,采集的负样本会逐渐避开滑坡发生点,避开与滑坡点相似的地理环境,假的负样本被采到的机率会降低,负样本集的质量会有所好转,SVM 模型能够在影响因素构成的环境特征空间中更为正确地区分这样的正负样本,具有更高的建模精度,对滑坡高危险区的预测能力更强。

当缓冲区大小在 500 ~ 800 m 内时,滑坡负样本的可采样空间范围不断缩小和局限,负样本的空间分布均衡性变差,无法全面包含研究区内典型的滑坡低危险情况(全局代表性差)[37],空间范围上的局限性会造成由影响因素构成的环境特征空间的局

限性。SVM 更关心两类样本中更靠近分类面的样本，当负样本在环境特征空间所占据的范围被局限，原本应该合理分割正负样本的 SVM 分类超平面会向负样本处偏移，导致 SVM 的过拟合和推测滑坡危险度的高估。由于滑坡危险度的高估，较大面积的研究区被划分为滑坡高危险区，滑坡检验样本更有可能落在推测滑坡危险区，因此验证精度较高。此外，由于正负样本在环境特征空间中的间距较大，导致 SVM 的分类超平面极容易将正负样本集区分开，建模精度极高。

当缓冲区大小在 800 ~ 1500 m 内时，负样本的空间采样范围进一步缩小，缓冲区以外采得负样本的全局代表性极差，无法全面反映流域内低危险区的典型地理环境。采集的负样本在环境特征空间中所占据的范围很小，且趋于稳定，导致 SVM 分类面的波动不大，几乎不会继续向负样本处偏移，研究区滑坡危险度仍然会被高估，极大面积的研究区被划分为滑坡高危险区，滑坡检验样本更有可能落在推测滑坡危险区，因此验证精度非常高，达到 100%。此外，正负训练样本集在环境特征空间中极大的间距也使得分类超平面能够非常容易将二者区分，建模精度达到 100%。

5　结论

本文旨在探究 BCS 负样本采样方法中不同的缓冲区大小与基于 SVM 的滑坡危险度制图精度之间的关系。研究发现，缓冲区大小对基于 SVM 的滑坡危险度制图影响很大，缓冲区过小会导致与滑坡点地理环境相似的假的负样本的存在，从而导致滑坡危险度的低估，当假的负样本达到一定量时，甚至会使得推测的滑坡高危险区与低危险区颠倒; 缓冲区过大会导致负样本在环境特征空间中太局限，负样本集的全局代表性差，从而导致滑坡危险度的高估。

本研究中缓冲区大小在 200 ~ 500 m 内时使用 BCS 采集的负样本较为符合油坊沟流域基于 SVM 的滑坡危险度制图。这一阈值内采集的负样本集中假的负样本较少，且具有较好的全局代表性。该阈值的确定依赖于研究区的尺度(流域尺度) 、数据源的尺度和质量、正样本的数量和质量、滑坡面的面积等，因此这一值域是否具有普适性仍然需要探讨。

参考文献(References)

[1] 胡新丽，唐辉明. 斜坡工程 GIS 系统研究与应用 [M]. 武汉: 中国地质大学出版社，2005: 1 - 136 [Hu Xinli, Tang Huiming. Research on the GIS system of slope engineering GIS and its application [M]. Wuhan: China University of Geosciences Press, 2005: 1 - 136]

[2] Iswar D, Sashikant Sahoob, Cees van Westena, et al. Landslide susceptibility assessment using logistic regression and its comparison with a rock mass classification system, along a road section in the northern Himalayas (India) [J]. Gepmorphology, 2010, 114(4) : 627 - 637

[3] Bai Shibiao, Wang Jian, Zhang Zhigang, et al. Combined landslide susceptibility mapping after Wenchuan earthquake at the Zhouqu segment in the Bailongjiang Basin, China [J]. Catena, 2012, 99: 18 - 25

[4] Guzzetti F, Carrara A, Cardinali M, et al. Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy [J]. Geomorphology, 1999, 31: 181 - 216

[5] Guzzetti F, Reichenbach P, Cardinali M, et al. Probabilistic landslide hazard assessment at the basin scale [J]. Gepmorphology, 2005, 72(1 - 4) : 272 - 299

[6] Guzzetti F, Reichenbach P, Ardizzone F, et al. Estimating the quality of landslide susceptibility models [J]. Gepmorphology, 2006, 20(1 - 2) : 166 - 184

[7] Dai Fuchu, Lee Chaofen, Zhang Xiaohui. GIS - based geo-environmental evaluation for urban land-use planning: a case study [J]. Engineering Geology, 2001, 61(4) : 257 - 271

[8] 尹志华. 基于 RS 和 GIS 技术对区域滑坡进行高效快速敏感性评价的模型研究——以北川县为例 [D]. 成都理工大学，2011. [Yin zhihua. Rapid and efficient regional landslide susceptibility assessment model based GIS and RS technology——a case study in Beichuan County [D]. Chengdu University of Technology, 2011.]

[9] Carrara A, Cardinali M, Detti R, et al. GIS techniques and statistical models in evaluating landslide hazard [J]. Earth surface processes and landforms, 1991, 16: 427 - 445

[10] Süzen M L, Doyuran V. Data driven bivariate landslide susceptibility assessment using geographical information systems: a method and application to Asarsuyu catchment, Turkey [J]. Engineering Geology, 2004, 71(3 - 4) : 303 - 321

[11] Das I, Sahoo S, van Westen C, et al. Landslide susceptibility assessment using logistic regression and its comparison with a rock mass classification system, along a road section in the northern Himalayas (India) [J]. Geomorphology, 2010, 114: 627 - 637

[12] Yilmaz I. The effect of the sampling strategies on the landslide susceptibility mapping by conditional probability and artificial neural networks [J]. Environmental Earth Sciences, 2010, 60: 505 - 519

[13] Carrara A, Cardinali M, Guzzetti F, et al. GIS - based techniques for mapping landslide hazard [G] // Carrara A, Guzzetti F. (Eds.) . Geographical Information Systems in Assessing Natural Hazards. Kluwer Academic Publishers, Dordrecht, The Nether-

lands, 1995: 135 – 175

[14] 祁元,刘勇,杨正华,等. 基于 GIS 的兰州滑坡与泥石流灾害危险性分析[J]. 冰川冻土,2012,34(1):96 – 104 [Qi Yuan, Liu Yong, Yang Zhenghua, et al. GIS – based analysis of landslide and debris flow hazard in Lanzho [J]. Journal of Glaciology and Geocryology,2012,34(1): 96 – 104]

[15] Guo Qinghua, Maggi K, Catherine H G. Support vector machines for predicting distribution of Sudden Oak Death in California [J]. Ecological Modelling, 2005, 182: 75 – 90

[16] Xiao Chenchao, Tian Yuan, Shi Wenzhong, et al. A new method of pseudo absence data generation in landslide susceptibility mapping with a case study of Shenzhen [J]. Science China Technological Sciences, 2010, 53(1): 75 – 84

[17] Yao Xin, Tham L G, Dai Fuchu. Landslide susceptibility mapping based on support vector machine: a case study on natural slopes of Hong Kong, China [J]. Geomorphology, 2008, 101: 572 – 582

[18] 方苗,张金龙,徐填. 基于 GIS 和 Logistic 回归模型的兰州市滑坡灾害敏感性区划研究[J]. 遥感技术与应用,2011,24(6):845 – 852 [Fang Miao, Zhang Jinlong, Xu Zhen. Landslide susceptibility zoning study in Lanzhou city based on GIS and logistic regression model [J]. Remote Sensing Technology and Application,2011,24(6): 845 – 852]

[19] 谌文武,赵志福,刘高,等. 兰州 – 海口高速公路甘肃段工程地质问题研究[M]. 兰州: 兰州大学出版社,2006: 19 – 22 [Chen Wenwu, Zhao Zhifu, Liu Gao, et al. The engineering geological problems study of Gansu section of Lanzhou – Haikou highway [M]. Lanzhou: Lanzhou University Press,2006: 19 – 22]

[20] 陈耀乾. 甘肃省武都县地质灾害调查与区划报告[R]. 兰州: 甘肃省地质环境监测总站,2001. [Chen Yaoqian. Geo-hazard survey and zone report in Wudu country of Gansu province [R]. Lanzhou: General Monitoring Station of Geological Environment of Gansu Province, China,2001.]

[21] 董抗甲. 甘肃省武都县地质灾害调查与区划报告[R]. 甘肃省地质环境监测总站,2003. [Dong Kangjia. Geo-hazard survey and zone report in Zhouqu country of Gansu province [R]. Edited by General Monitoring Station of Geological Environment of Gansu Province, China,2003.]

[22] Atkinson P M, Massari R. Autologistic modelling of susceptibility to landsliding in the Central Apennines, Italy [J]. Geomorphology, 2011, 130(1 – 2): 55 – 64

[23] 陈晓利,叶洪,程菊红. GIS 技术在区域地震滑坡危险性预测中的应用——以龙陵地震滑坡为例[J]. 工程地质学报,2006,14(03):333 – 338 [Chen Xiaoli, Ye Hong, Cheng Juhong. Use of GIS in regional risk assessment of earthquake induced landslides—— a case study of earthquake induced landslides in Longling in 1976 [J]. Journal of Engineering Geology,2006,14(03): 333 – 338]

[24] 谭龙,陈冠,王思源,等. 逻辑回归与支持向量机模型在滑坡敏感性评价中的应[J]. 工程地质学报,2014,22(1):56 – 63 [Tan Long, Chen Guan, Wang Siyuan, et al. Landslide susceptibility mapping based on logistic regression and support vector machine [J]. Journal of Engineering Geology,2014,22(1): 56 – 63]

[25] 齐识,张雅莉,张鹏,等. 白龙江流域滑坡危险性评价指标体系的构建[J]. 长江科学院院报,2014,31(01):23 – 38 [Qi Shi, Zhang Yali, Zhang Peng, et al. An assessment index system for landslide risk in Bailong river basin [J]. Journal of Yangtze river scientific research institute,2014,31(01): 23 – 38]

[26] 许冲,戴福初,姚鑫,等. 基于 GIS 的汶川地震滑坡灾害影响因素确定性系数分析[J]. 岩石力学与工程学报,2010,29(01): 2372 – 2381 [Xu Chong, Dai Fuchu, Yao Xin, et al. GIS based certainty factor analysis of landslide triggering factors in Wenchuan earthquake [J]. Chinese Journal of rock mechanics and engineering,2010,29(01): 2372 – 2381]

[27] 白世彪,阎国年,盛业华. GIS 技术在三峡库区滑坡影响因素分析中的应用[C]//中国地理信息系统协会第八届年会论文集,2004. [Bai Shibiao, Lü Guonian, Sheng Yehua. The application of GIS technology in the analysis of the landslide triggering factors in three Gorges reservoir area [C]//The eighth annual meeting proceedings of China association of geographic information system, 2004.]

[28] Bai Shibiao, Wang Jian, Lü Guonian, et al. GIS – Based and Data – Driven Bivariate Landslide – Susceptibility Mapping in the Three Gorges Area, China [J]. Pedosphere, 2009, 19: 14 – 20

[29] 傅文杰. GIS 支持下基于支持向量机的滑坡危险性评价[J]. 地理科学,2008,28(6):838 – 841 [Fu Wenjie. Landslide hazard evaluation based on GIS and SVM [J]. Scientia geographica sinica,2008,28(6): 838 – 841]

[30] 李秀珍,孔纪名,王成华. 多分类支持向量机在滑坡稳定性判识中的应用[J]. 吉林大学学报: 地球科学版,2010,40(3): 631 – 637 [Li Xiuzhen, Kong Jiming, Wang Chenghua. Application of multi-classification support vector machine in the identifying of landslide stability [J]. Journal of Jilin University: Earth Science Edition,2010,40(3): 631 – 637]

[31] 姜琪文,许强,何政伟. 基于 SVM 多类分类的滑坡区域危险性评价方法研究[J]. 地质灾害与环境保护,2005,16(3):328 – 330 [Jiang Qiwen, Xu Qiang, He Zhengwei. Study on landslide hazard zonation based on multi-classification support vector machine [J]. Journal of Geological Hazards and Environment Preservation,2005,16(3): 329 – 330]

[32] 胡德勇,李京,陈云浩,等. GIS 支持下滑坡灾害空间预测方法研究[J]. 遥感学报,2007,11(6):852 – 859 [Hu Deyong, Li Jing, Chen Yunhao, et al. GIS – based landslide spatial prediction methods, a case study in Cameron highland, Malaysia [J]. Journal of Remote Sensing,2007,11(6): 852 – 859]

[33] Xu Chong, Xu Xiwei, Dai Fuchu. Comparison of different models for susceptibility mapping of earthquake triggered landslides related with the 2008 Wenchuan earthquake in China [J]. Computers & Geosciences, 2012, 46: 317 – 329

[34] Xu Chong, Dai Fuchu, Xu Xiwei, et al. GIS – based support vector machine modeling of earthquake-triggered landslide susceptibility in the Jianjiang River watershed, China [J]. Geomorphology, 2012, 145: 70 – 80

[35] Pradhan B. A comparative study on the predictive ability of the de-

cision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS [J]. Computers & Geosciences, 2013, 51: 350 - 365

[36] Chang Chihchuang, Lin Chihjen. LIBSVM: a library for Support Vector Machines [J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(3): 27. Software available at http://www.csie.ntu.edu.tw/ ~ cjlin/libsvm

[37] 刘京,朱阿兴,张淑杰,等. 基于样点个体代表性的大尺度土壤属性制图方法 [J]. 土壤学报,2013,50(1):12 - 20 [Liu Jing, Zhu Axing, Zhang Shujie, et al. Large scaled soil attribute mapping method based on individual representativeness of sample sites [J]. Acta Pedologica Sinica,2013,50(1): 12 - 20]

Sensitivity of BCS for Sampling Landslide Absence Data in Landslide Susceptibility Assessment

MIAO Yamin[1], ZHU Axing[1,2,3], YANG Lin[2], BAI Shibiao[1], LIU Junzhi[1], DENG Yongcui[1]

(1. *Key Laboratory of Virtual Geographic Environment (Nanjing Normal University), Ministry of Education, State Key Laboratory Cultivation Base of Geographical Environment Evolution (Jiangsu Province), Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing* 210023, *China*;
2. *State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing* 100101, *China*;
3. *Department of Geography, University of Wisconsin - Madison, Madison, WI* 53706, *USA*)

Abstract: Landslide absence data plays an important role in data-driven models for landslide susceptibility mapping. It can constrain the overestimation of predicted landslide susceptibility values. Buffer controlled sampling (BCS) is widely used in sampling landslide absence data. It is based on the general principle that the area near the landslide occurrences has similar geo-environment with landslides, resulting it prone to landslides. Thus landslide absence data should be sampled from the areas beyond the buffer zones of the landslide sites. Currently the buffer size is decided subjectively based on the experts' knowledge of the study area. The study of the effect of buffer size on data-driven models for landslide susceptibility mapping is rare. It is important to study the general relationships between buffer size and mapping accuracy and find an appropriate buffer size for an given area. In this study, BCS sampling strategy was used in the Youfang ravine in the south Gansu of China for sampling landslide absence data and Support Vector Machine (SVM) was used to deliniute landslide susceptibility across the study area. Results show that if the buffer size be small, false absence data would be included in the generated absence datasets and result in the underestimation of the predicted landslide susceptibility values. If the buffer size be large, the representativeness of generated landslide absence data for the whole study area is low, resulting in the overestimation of the predicted landslide susceptibility values. In the Youfang ravine, the appropriate range of buffer size in BCS for sampling landslide absence data is from 200 m to 500 m in SVM for landslide susceptibility mapping.

Key words: SVM; landslide absence data; BCS; buffer; landslide susceptibility mapping