

# 不完备样本条件下基于支持向量回归模型的滑坡易发性评价

胡德勇<sup>1,2</sup>, 赵文吉<sup>\* 1,2</sup>, 李小娟<sup>1,2</sup>, 李 京<sup>3</sup>, 李家存<sup>1,2</sup>

(1. 三维信息获取与应用教育部重点实验室, 首都师范大学资源环境与旅游学院, 北京 100037;

2. 资源环境与地理信息系统北京市重点实验室, 北京 100037;

3. 民政部/教育部减灾与应急管理研究院, 北京 100875)

**摘要:** 区域滑坡易发性评价对灾害中长期预测预报具有重要意义, 在基于统计模型进行评价过程中, 样本选取对评价结果有较大影响, 构建较稳健的、受样本数量影响小的分析模型非常重要。本文以马来西亚热带雨林地区为例, 选择坡度、坡向、地表曲率、地貌类型、岩性、构造、土地覆盖、道路和排水系统等 9 大要素作为评价因子, 结合支持向量回归 (SVR) 模型计算研究区滑坡易发性指数, 并探讨不完备样本条件下易发性评价方法, 分析样本数量和评价精度之间的关系。结果显示, 基于 SVR 模型进行该区滑坡易发性分析评价, 其成功率验证法的描述精度约为 95.9%; 同时, 样本数量的增减对分析精度影响较小; SVR 方法是一种适于热带雨林地区高植被覆盖条件下的分析模型, 可为今后同类地区的滑坡灾害管理工作提供支持。

**关键词:** 不完备样本; 支持向量回归模型; 滑坡易发性指数; 精度分析

**文章编号:** 1000-0585(2008)04-0755-08

区域滑坡空间预测预报的系统理论与方法是当前滑坡研究的重要内容<sup>[1,2]</sup>, 滑坡易发性评价 (landslide susceptibility evaluation) 针对孕灾环境的综合性和区域性等特征, 开展区域滑坡的敏感度、易发度分析, 因此它对滑坡灾害中长期预测预报具有重要作用, 在实际中得到了广泛应用<sup>[3~5]</sup>。

概括说来, 区域滑坡易发性评价模型可以分为知识驱动和数据驱动两类。基于专家经验的知识驱动型方法, 它依靠专家先验知识来确定灾害发生的敏感程度, 因此具有一定的主观性和不确定性; 数据驱动型方法根据历史滑坡调查数据、孕灾环境背景资料, 对评价指标进行分级和确定权重, 并对孕灾环境的敏感性进行定量评价。数据驱动型模型应用较多的有指数综合法、多元统计分析等<sup>[6]</sup>。自 20 世纪 80 年代开始, Carrara A<sup>[7]</sup> 开始运用多元统计理论开展滑坡相关研究, 此后该理论在灾害分析领域得到了广泛应用<sup>[8~10]</sup>; 随着计算科学、自动控制论的日益发展, 多元统计分析发展到机器学习阶段, 如人工神经网络 (ANN)<sup>[11]</sup>、支持向量机 (SVM)<sup>[12~14]</sup> 等理论在滑坡分析中得到应用。通常情况下基于统计模型的滑坡易发性评价受到研究区历史滑坡样本的影响, 在寻求最大样本数量、减少

收稿日期: 2007-10-26; 修订日期: 2008-04-07

基金项目: 国际科技合作计划项目 (2007DFA20640), 国家高技术研究发展计划“减灾救灾应用示范”项目 (2007AA120306) 资助。

作者简介: 胡德勇 (1974), 男, 博士, 教师。主要从事遥感与地理信息系统在资源环境、自然灾害等领域的应用研究。E-mail: deyonghu@163.com

\* 通讯作者: 赵文吉 (1967), 男, 教授。E-mail: zhwenji1215@163.com

和降低统计模型理论误差的同时,研究者面临样本选择的工作量大、可信度低等问题,工作效率和分析精度难达平衡。因此,降低样本选择对统计拟合结果精度的影响,探寻适于小样本空间(不完备样本条件)、更加稳健的统计分析模型,对滑坡易发性评价具有重要意义。

本文以马来西亚热带雨林地区滑坡易发性评价为例,采取随机分割样本的方式来获取不完备样本条件,基于支持向量回归机(support vector regression, SVR)计算研究区滑坡易发性指数(landslide susceptibility index, LSI),并讨论不完备样本条件下分析精度的变化特征,通过热带雨林地区的滑坡易发性评价实证研究,为其他滑坡孕灾环境类似地区的灾害管理提供技术支持。

## 1 研究区、数据和方法

### 1.1 研究区

金马伦高原(Cameron Highland)位于马来西亚中部彭亨(Penang)州西北角,东北和西部分别毗邻吉兰丹(Kelantan)州和霹雳(Perak)州,地理坐标为东经 $101^{\circ}20'$ ~ $101^{\circ}36'$ ,北纬 $4^{\circ}19'$ ~ $4^{\circ}37'$ 。整个研究区轮廓呈马蹄形状,东西宽32km,南北长34km,总面积约为690km<sup>2</sup>。(图1)

该区位于马来半岛中央大山脉(Rang of Peninsular Malaysia)中部,距离马来西亚首都吉隆坡约150km,是马来西亚国内重要的农作物生产基地。植被类型在平地区主要为经济林,包括油棕、橡胶和茶叶等,山地丘陵区主要为热带雨林,植被冠层浓密。基岩岩床主要由花岗岩组成。随着经济快速发展,人口数量急剧增加,人们对热带雨林的垦殖加剧,无节制地“围林造地”导致大片的雨林变为农用地,失去“保护外衣”的裸露地表在致灾因子的诱发下,极易形成灾害事件。特别是最近几年,滑坡灾害频繁发生,造成人员伤亡和经济损失,如2000年发生的一次滑坡造成6人死亡,公路损毁;2004年也出现了几次大的滑坡。

### 1.2 数据

滑坡发生是多因子的综合产物,不同区域影响因子之间也存在差异性,它们对分析结果具有较大影响。根据热带雨林地区的具体特点和前人对马来西亚境内滑坡的部分研究结果<sup>[4,15]</sup>,选取坡度、坡向、地表曲率等地形特征参数为重要影响因子,地表曲率是指地面坡度的变化率,可以通过计算地面坡度而求得;地貌类型在区域滑坡评价中为重要影响因子,所以也选入其中;另外还包括岩性、构造、土地覆盖、地貌类型、道路和排水系统等共9个因子。

研究中收集的数据主要包括历史滑坡资料、基础地图两大类(表1)。基础数据中包括1:50000比例尺地形图、地质图、土地利用图和地貌类型图等。从地形图上提取等高

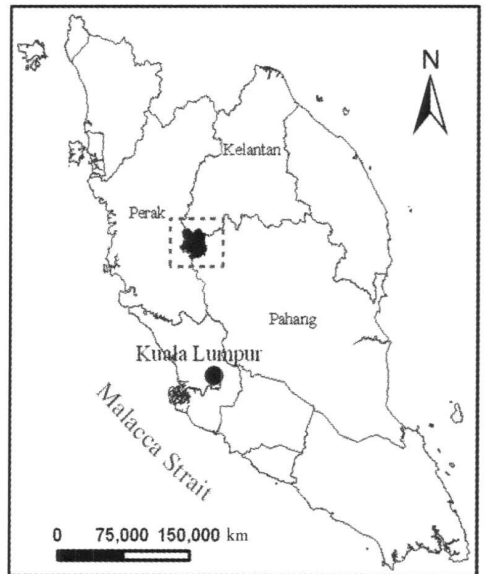


图1 研究区地理位置示意图

Fig 1 The location of the study area

线、河流和道路等信息; 由地图等高线和高程点在 ArcGIS 软件中采用 Kriging 插值法衍生出 DEM 数据, 对 DEM 数据进行进一步处理, 得到坡度、坡向和地表曲率等描述地表形态特征的数据; 地质图主要包括滑坡易发性评价有关的构造和岩性分布数据。为了研究的方便, 应用 ArcGIS 软件将所有数据分别处理成点 (point)、线 (line)、面 (polygon) 和格网 (grid) 等类型, 待后续进一步分析和调用。

1.3 方法

图 2 为基于栅格的滑坡易发性评价过程示意图, 主要包括专题地图选择、地图数据库构建、特征应用和输出 4 个主要步骤, 通过地理编码将专题地图转换为地图数据库, 特征选择从地图数据库中按照一定要求选择出适合滑坡易发性分析的特征因子, 然后逐像素基于专题模型进行易发性评价, 最后得到易发性指数专题图。

表 1 滑坡易发性评价 GIS 专题数据分类及特性

Tab 1 The GIS thematic data for landslide susceptibility evaluation

分类	子类	GIS 数据类型	比例尺
滑坡	滑坡灾害	Point、Polygon	1: 50000
	等高线	Line	1: 50000
	DEM	Grid	1: 50000
	坡度	Grid	1: 50000
	坡向	Grid	1: 50000
基础地图	地形图	Grid	1: 50000
	地表曲率	Grid	1: 50000
	河流分布	Line	1: 50000
	道路分布	Line	1: 50000
	地质图	Polygon	1: 50000
	岩性	Polygon	1: 50000
	构造	Line	1: 50000
	土地利用图	Polygon	1: 50000
	地貌类型图	Polygon	1: 50000

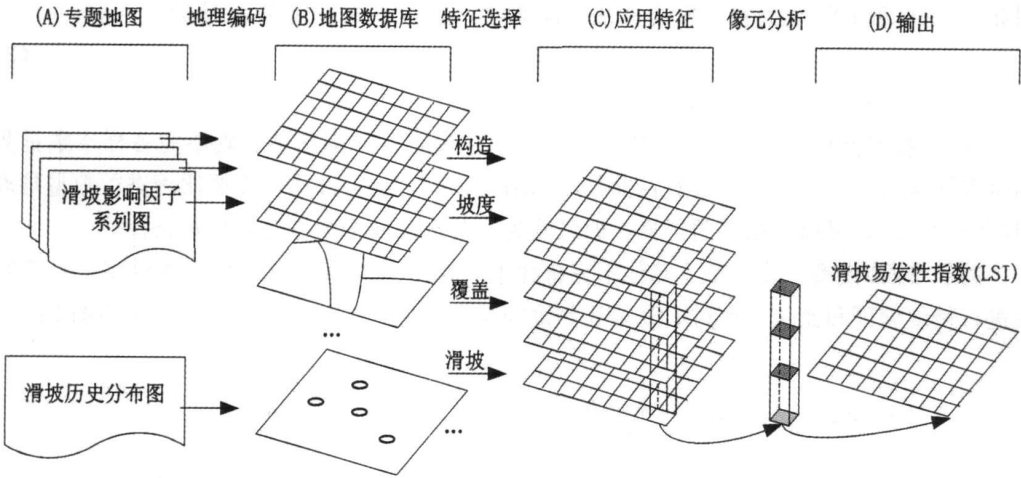


图 2 滑坡灾害易发性评价过程示意图

Fig 2 The procedure of landslide susceptibility evaluation based on raster analysis

在整个灾害易发性分析评价过程中, 评价单元的划分、评价模型构建方法对最终结果将产生一定的影响。

(1) 网格单元划分 李军等曾对不同比例尺图层下栅格单元大小选择进行过研究, 提出了计算栅格单元大小的经验公式<sup>[16]</sup>:

$$G_s = 7.49 + 0.0006S - 2.0 \times 10^{-9} S^2 + 2.9 \times 10^{-15} S^3$$

(1)

式 (1) 中,  $G_s$  为适宜格网大小,  $S$  为地图数字比例尺分母。对于 1: 50000 比例尺地形图, 由式 (1) 计算出最佳网格单元为 32. 8 m, 为了计算的方便, 评价单元采用 30 m 正方形网格。

相关矢量图层可以通过矢-栅转换获得滑坡分析的网格单元。其中, 坡向按照地表 8 个方位分别进行编码、坡度以  $4^\circ$  为间隔划分为 10 种类型、地表曲率按照计算值域等分为 7 种类型、河流和道路等线性地物要素以 50 m 为单位进行 buffer 分析而依次编码, 其他因子按照具体类型分别编码。这些数据按照  $30\text{m} \times 30\text{m}$  正方形格网处理为栅格图像, 为下一步基于 SVR 模型的像素分析做准备。

(2) SVR 与 LSI 支持向量机(Support Vector Machine, SVM)是基于 VC (Vapnik-Chervonenkis) 理论的创造性机器学习方法。SVR 是 SVM 理论的重要组成部分, SVR 理论在解决小样本空间、非线性问题等方面存在较大的优势, 关于 SVR 相关理论这里不作赘述, 见参考文献 [17]。根据相关的理论和算法, 对于二维空间分划问题, 最优线性回归函数可以表示为:

$$f(x) = (w \cdot \phi(x)) + b = \sum_{i=1}^n (a_i - a_i^*) K(x, x_i) + b \tag{2}$$

式 (2) 中,  $n$  为支持向量的个数,  $a_i$ ,  $a_i^*$  和  $b$  为确定最优超平面的参数,  $K(x, x_i)$  为核函数。

在滑坡易发性评价过程中, 通过样本数据集的机器学习, 可以建立滑坡发生事件和影响因素之间的决策函数, 然后利用它对研究区数据进行回归分析, 回归值的大小  $f(x)$  反映滑坡对影响因子  $x_1, x_2, \dots, x_n$  的敏感程度, 它从侧面反映了滑坡发生的可能性高低, 因此  $f(x)$  和  $LSI$  之间存在密切关系, 这里定义基于 SVR 模型的  $LSI$  为:

$$LSI = k \cdot f(x) + c \tag{3}$$

式 (3) 中,  $k$ 、 $c$  为常数, 用来调整 SVR 回归值的大小。

(3) 技术流程 假定业已采集的全部样本为较完备样本条件, 则不完备样本数据集为全部样本的一个子集, 通过对目前样本的随机分割, 将全部样本按照比例进行数量抽取 (10% ~ 90%), 可以获取不同子集, 即可认为这些部分样本为不完备样本条件。

基于 SVR 模型计算完备、其他不完备样本条件下滑坡易发性指数, 并对多个计算结果进行精度验证和比较, 探讨 LSI 描述精度与样本条件的关系, 技术流程表示为图 3:

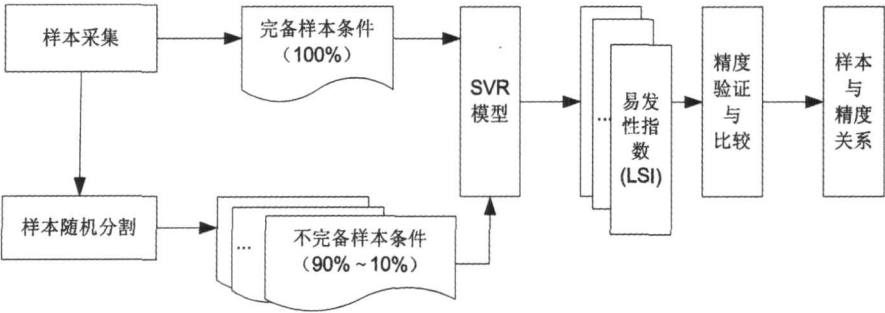


图 3 不完备样本条件下易发性评价精度分析

Fig 3 The accuracy analysis procedure of landslide susceptibility evaluation in incomplete sample conditions

## 2 基于 SVR 模型的易发性评价

### 2.1 样本采集与数据标准化

从栅格化的图层数据中，采集历史上曾经出现过滑坡地段样本，共 2963 个像素（因变量赋值+1），占研究区总面积的 0.40%，然后从研究区中随机采集未发生滑坡地区的样本 3557 个像素（因变量赋值-1），约占研究区总面积的 0.48%，形成训练样本数据集。在进行 SVR 模型分析评价前，为消除多特征空间属性数据对分析评价结果的影响，需进行数据的标准化，研究中将各因子值标准化到-1 到+1 之间，然后生成标准化模型，然后根据这个模型来标准化图像数据。

### 2.2 SVR 模型参数设置

对样本对象进行训练过程中，核函数的选择对分类结果存在一定的影响，在常用的四类核函数中，多项式、S 形核函数都有  $\gamma$ 、 $r$  两个参数，而径向基核函数（RBF）只有  $\gamma$  一个参数，在参数确定过程中，径向基核函数由于比多项式核和 S 形核函数少一个变化参量，因此，变化范围容易控制，成为非线性分划的推荐核函数<sup>[18]</sup>。

为了确定回归模型中 RBF 核函数  $C$ 、 $\gamma$  参数的取值，假定支持向量分类机（support vector classification, SVC）的优化 RBF 参数同样也适合支持向量回归机，因此凭借适于 SVC 交叉验证“网格搜寻”方法，来确定  $C$ 、 $\gamma$  的取值<sup>[18,19]</sup>，即将选择的训练样本分成  $V$  部分，其中  $V-1$  部分作为模型的训练样本，剩下的一部分作为模型参数确定的经验样本，利用检验样本来验证  $V-1$  部分数据分类结果的精度，不断改变  $C$ 、 $\gamma$  来获取更高的样本预测精度。

### 2.3 易发性评价结果

确定下来  $C$ 、 $\gamma$  参数取值后，用生成的模型对图像数据逐像元进行 SVR 回归分析，计算出 LSI，最终得到基于 SVR 的滑坡易发性评价结果专题图（图 4）。

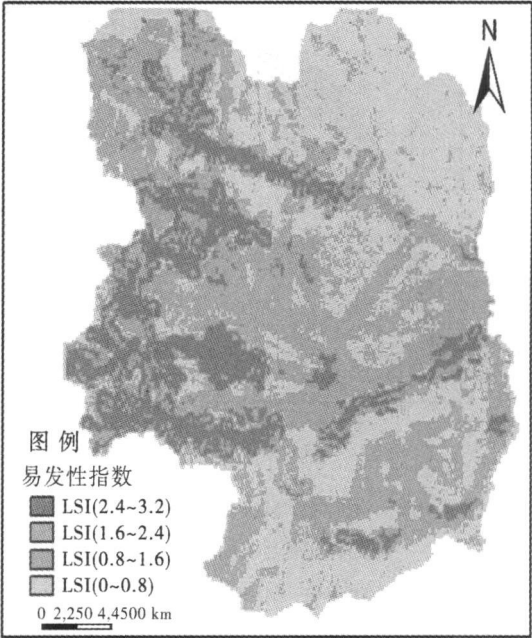


图 4 基于 SVR 的滑坡灾害易发性指数分布图

Fig 4 The thematic map of landslide susceptibility index based on SVR

## 3 不完备样本条件结果分析

### 3.1 成功率验证法

比较滑坡易发性的计算结果和历史滑坡发生的位置，可以评价易发性分析的有效性，成功率验证法是评价相对精度的常用方法，由 Chung C F 等首先提出<sup>[20]</sup>，Lee S 等在研究中也应用了该方法来评价滑坡分析的精度<sup>[4,21]</sup>，它以滑坡易发性指数累计百分比为 X 轴（%），滑坡发生累计百分比为 Y 轴（%），来分析 LSI 对滑坡发生的概括程度。对于 X 轴，将易发性指数（LSI）规一化到 0 ~ 100 范围内，按照降序进行排列，然后等分成 100 个单位，分别计算各个单位内滑坡发生的百分比；Y 轴显示的是滑坡发生样本的累计频率百分比。

很显然,对于滑坡发生累计频率(%) - 滑坡易发性指数(%)曲线,由X和Y轴共同组成的绘图区的面积视为1,可以代表滑坡易发性评价精度为100%,因此,频率累计曲线下部和X轴围成的面积(面积 $\leq 1$ ,即精度 $\leq 100\%$ )可以作为定量参数,描述滑坡易发性分析精度高低;同时,由于滑坡发生和未发生既对立又统一,较高的滑坡发生预测精度对应较低的滑坡未发生预测精度,因此,频率累计曲线上部和Y轴围成的面积可以相对地表示为滑坡未发生分析精度。

3.2 样本-精度影响分析

为了讨论不完备样本条件下滑坡易发性评价精度变化特征,采用随机选点方式将采集到的全部样本进行分割和抽取,分10%~90%共9级,这样加上最初的全样本100%这一级,可供研究的样本数量共为10级。

利用多种样本(10%~90%)分别进行学习,计算不同样本条件下基于SVR模型的滑坡易发性指数(LSI),同样绘制各个样本条件下的LSI专题分布图(专题图不再展示),然后绘制滑坡发生累计频率与易发性指数关系曲线,通过成功率验证法来横向比较不同样本条件下的易发性分析评价的结果,分析样本数量对SVR模型的影响(图5):

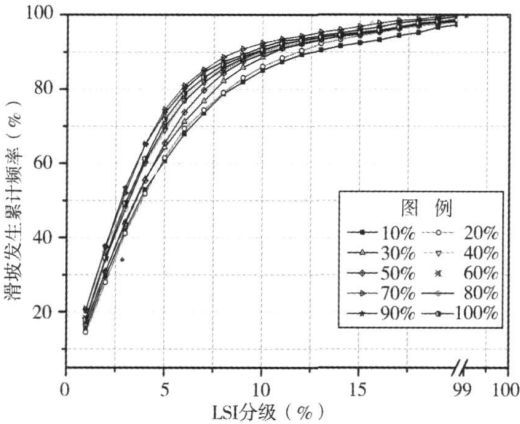


图5 不同样本条件下滑坡发生  
累计频率与LSI的关系

Fig. 5 The relation between LSI and landslide  
cumulative frequency in different sample sizes

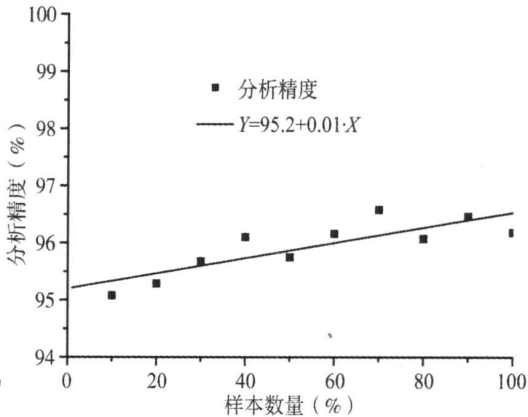


图6 滑坡易发性分析精度与  
样本数量之间的关系

Fig. 6 The relation between landslide susceptibility  
analysis accuracy and sample size

根据图5绘制样本数量和易发性分析精度的关系图6,可以看出,随着样本数量的增加,滑坡易发性分析精度有所提高;同时,SVR法的分析精度受样本数量影响较少,表现为图中离散点波动较小,其分析精度均值为95.9,标准差为0.5。即便在10%样本条件下,滑坡易发性指数的描述精度也可以达到约95.3左右,这对于小样本条件下的易发性描述精度已经非常高。

对结果进行直线拟合,得到关系式:

$$Y = 95.2 + 0.01 \times X \tag{4}$$

式(4)中,X表示样本数量,Y表示分析精度。直线方程的Y轴截距为95.2,斜率为0.01。因此,总体上说,分析精度均在95.2以上,波动较小,即受到样本数量的影响较小,且随样本数量的增加精度有升高趋势。分析其原因,SVR依据支持向量来建立最终的决策函数,它考究的是因子群构造的样本空间与分析对象的关系,适合解决本质上非

线性的问题, 对于滑坡易发性评价可以得到较好的解决。

## 4 结论和展望

在热带雨林地区的滑坡易发性评价与制图工作中, 囿于热带雨林的高植被覆盖, 大范围的历史滑坡调查很难保证全面、彻底; 同时, 滑坡发生是地质环境综合变异的结果, 强行将孕灾背景两分为滑坡和未滑坡来采集样本, 很难保证样本呈正态分布, 因而影响滑坡易发性评价的精度。从样本角度考虑, 不完备样本条件在滑坡易发性分析中普遍存在。

基于 SVR 模型对研究区进行滑坡易发性分析和评价, 其成功率描述精度可达到约 95.9%; 同时, 针对不完备样本条件, SVR 模型可以根据有限样本信息在模型复杂性和学习能力之间寻求最佳折衷, 以期获得最好的推广能力, 表现出样本数量对滑坡易发性分析精度影响较小, 这表明 SVR 模型适合样本采集难以达到理想状况下的滑坡易发性评价工作。

该模型对于其他滑坡孕灾环境类似地区的灾害管理工作具有重要应用价值, 可以利用历史滑坡资料和其他专题数据实现区域群发性滑坡的分析、评价和区划等工作, 从而为中长期灾害预测预报服务。

## 参考文献:

- [1] 殷坤龙, 朱良峰. 滑坡灾害空间区划及 GIS 应用研究. 地学前缘, 2001, 8(2): 279~ 284
- [2] 戴福初, 李军. 暴雨滑坡泥石流的的研究进展与趋向. 地理研究, 1998, 17(增): 117~ 124.
- [3] 张丽君, 江思宏. 区域性滑坡敏感性评价的数据驱动权重模型及应用. 水文地质工程地质, 2004(6): 33~ 36.
- [4] Lee S. Application of logistic regression model and its validation for landslide susceptibility mapping using GIS and remote sensing data. International Journal of Remote Sensing, 2005, 26(7): 1477~ 1491.
- [5] Ercanaglu M, Gokceoglu C. Assessment of landslide susceptibility for a landslide prone area (north of Yenice, NW Turkey) by fuzzy approach. Environmental Geology, 2002, 41: 720~ 730.
- [6] 程凌鹏, 杨冰, 刘传正. 区域地质灾害风险评价研究述评. 水文地质工程地质, 2001(3): 75~ 78
- [7] Carrara A. Multivariate models for landslide hazard evaluation. Mathematical Geology, 1983, 15(3): 403~ 426
- [8] Baeza C, Corominas J. Assessment of shallow landslide susceptibility by means of multivariate statistical techniques. Earth Surface Processes and Landforms, 2001, 26: 1251~ 1263
- [9] Lee S, Min K. Statistical analysis of landslide susceptibility at Yongin, Korea. Environmental Geology, 2001, 40: 1095~ 1113
- [10] Lee S, Chwae U, Min K. Landslide susceptibility mapping by correlation between topography and geological structure: the Janghung area, Korea. Geomorphology, 2002, 46: 149~ 162
- [11] Lee S, Ryu J H, Won J S, *et al*. Determination and application of the weights for landslide susceptibility mapping using an artificial neural network. Engineering Geology, 2004, 71: 289~ 302
- [12] 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. 地质灾害与环境保护, 2005, 16(3): 328~ 330
- [13] 戴福初, 姚鑫, 谭国焕. 滑坡灾害空间预测支持向量机模型及其应用. 地学前缘, 2007, 14(6): 153~ 159
- [14] 马志江, 陈汉林, 杨树锋. 基于支持向量机理论的滑坡灾害预测. 浙江大学学报, 2003, 30(5): 592~ 596
- [15] Jasmi Ab Talib. Slope instability and hazard zonation mapping using remote sensing and GIS techniques in the area of Cameron Highlands, Malaysia [CP/OL], 1997 [2008-03-17]. <http://www.gisdevelopment.net/aars/aars/1997/ts3/ts3001.asp>
- [16] 李军, 周成虎. 基于栅格 GIS 滑坡风险评价方法中格网大小选取分析. 遥感学报, 2003, 7(2): 86~ 92
- [17] 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. 遥感学报, 2007, 11(6): 852~ 859

- [ 18] Chang Chih-chung, Lin Chih-jen. LIBSVM: A library for support vector machines [ C/ OL], 2001[ 2008-03-17]. <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
- [ 19] 张锦水, 何春阳, 潘耀忠, 等. 基于 SVM 的多源信息复合的高空间分辨率遥感数据分类研究. 遥感学报, 2006, 10( 1): 49~ 57
- [ 20] Chung C F, Fabbri A G. Probabilistic prediction models for landslide hazard mapping. Photogrammetric Engineering & Remote Sensing, 1999, 65( 12): 1388~ 1399
- [ 21] Lee S, Chol J, Min K. Probabilistic landslide hazard mapping using GIS and remote sensing data at Boun, Korea International Journal of Remote Sensing, 2004, 25( 11): 2037~ 2052

## Landslide hazard susceptibility evaluation based on small-sample SVR method: A case study in Malaysia tropical rainforest area

HU De-yong<sup>1,2</sup>, ZHAO Wen-ji<sup>1,2</sup>, LI Xiao-juan<sup>1,2</sup>, LI Jing<sup>3</sup>, LI Jia-cun<sup>1,2</sup>

( 1 Key Laboratory of 3D Information Acquisition and Application of Ministry of  
Education, Capital Normal University, Beijing 100037, China)

( 2 Beijing Municipal Key Laboratory of Resources Environment and GIS, Beijing 100037, China)

( 3 Academy of Disaster Reduction and Emergency Management, Ministry of Civil Affairs &  
Ministry of Education, Beijing 100875, China)

**Abstract:** Landslide hazard susceptibility relates to middle- and long- term predicting and forecasting, and it is very important to landslide managements. In the process of evaluation based on statistical model, the result is greatly influenced by landslide sample size, so the more conservative and less influencing model must be applied to the susceptibility evaluation in order to reduce the system error. The study area is located in Malaysia tropical rainforest, where nine factors were selected as topographic slope, aspect, surface curvature, geomorphology, lithology, structure, land cover, road and drainage and so on. The Landslide Hazard Susceptibility Index (LSI) was constructed based on support vector regression (SVR) theory, then the susceptibility evaluation methodology was discussed in incomplete sample conditions, and the relation between the sample size and the result accuracy was analysed too. The result show that the success-rate analysis accuracy based on SVR model was about 95.9%, an obviously high value; the fluctuation of sample size influenced the accuracy slightly; SVR was a better model suited to landslide hazard evaluation in high vegetation cover conditions, which could provide a technique support for landslide management in similar areas.

**Key words:** incomplete sample; SVR model; landslide susceptibility index; accuracy analysis