

文章编号: 1004-4353(2015)02-0179-06

基于 Logistic 回归模型的延边地区 渤海国遗址预测研究

董振, 金石柱*

(延边大学理学院 地理系, 吉林 延吉 133002)

摘要: 以延边地区的渤海国遗址为研究对象, 借助 GIS 软件获取同遗址分布相关的高程、坡度、坡向、与河流之间的距离、与村屯之间的距离等因素值, 利用 Logistic 回归模型建立延边地区渤海国遗址预测模型, 并分析预测结果. 结果表明: 预测模型的预测准确率达 89.7%, 采用自然分裂法(Jenks)分级的高概率区面积占研究区域的 14.7%. 高概率区具有靠近河流分布的特点, 主要分布在海兰江流域、布尔哈通河流域、嘎呀河流域、牡丹江流域、图们江干流域等. 高概率区相对比重高的地区有龙井、图们、延吉、珲春等县市, 而高概率区绝对比重高的地区有敦化、汪清、珲春、龙井、安图等县市. 该研究结果有助于提高渤海国考古工作效率.

关键词: 遗址; 渤海国; Logistic 回归模型; 预测模型; 延边地区

中图分类号: K878

文献标识码: A

Prediction research on Bohai Kingdom ruins in Yanbian area based on the logic regression model

DONG Zhen, JIN Shizhu*

(Department of Geography, College of Science, Yanbian University, Yanji 133002, China)

Abstract: We aim to establish a forecasting model to analyze the predicting results on the Bohai Kingdom ruins in Yanbian area by using logic regression model, which gets the value of height, gradient, slope aspect, the distance from the river, and the distance from village by using the GIS. The results show that the accuracy of the prediction model reaches to 89.7%, and the highly probability region of the whole study area is 14.7% by using Jenks classification method. The results show that the distribution of the highly probability region is always near the river, which are Hailan River Basin, Buerhatong River Basin, Gaya River Basin, Mudan River Basin, Tumen River Basin etc. The highly relatively proportion of high probability region includes Longjing, Tumen, Yanji, Hunchun etc. And the highly absolutely proportion of highly probability region includes Dunhua, Wangqing, Hunchun, Longjing, Antu etc. The research result has significant influencing on improving the efficiency in archeology work in Bohai Kingdom.

Key words: ruins; Bohai Kingdom Site; logic regression model; predictive model; Yanbian area

考古遗址预测模型是基于对特定区域内已知遗址进行的环境因素分析, 如高程、坡度、与水系的距离、土壤类型等, 找出遗址分布的统计性规律和特征, 然后在这个区域的其他地方用多变量判

别函数对遗址存在的可能性进行评价, 给出潜在遗址的概率分布^[1]. 在国外, 遗址预测模型可追溯到 20 世纪 50 年代的聚落考古研究, 其开拓者为美国考古学家 Willey^[2]; Michael Märker 等利用

空间数据挖掘技术对伊朗扎格罗斯山脉上的旧石器时代聚落遗址位置进行了预测^[3]。在国内,倪金生^[4]、乔文文等^[5]利用 Logistic 回归模型先后对山东莒县沭河上游流域的大汶口、龙山和岳石文化时期遗址、郑龙地区龙山文化遗址分别进行了预测研究;彭淑贞等^[6]利用环境因素设定权重的方法针对山东省汶泗流域的大汶口文化时期的遗址进行了预测研究。本文以延边朝鲜族自治州(以下简称延边地区)的渤海国遗址为研究对象,利用 Logistic 回归方法建立模型并生成遗址分布概率图,分析遗址分布特点。

1 研究区和数据源

1.1 研究区概况

延边朝鲜族自治州位于吉林省东部,地理位置为北纬 $41^{\circ}59'47''\sim 44^{\circ}30'42''$,东经 $127^{\circ}27'43''\sim 131^{\circ}18'33''$ 之间^[7]。延边地区水资源丰富,主要河流有图们江、牡丹江、绥芬河、第二松花江四大水系 8 条主要江河和 487 条大小河流^[8]。延边地区属于中温带大陆性季风气候区,春季干燥多风,夏季炎热多雨,秋季凉爽少雨,冬季寒冷漫长;年均气温为 $2\sim 6^{\circ}\text{C}$,年降水量为 $450\sim 700\text{ mm}$ 。

延边地区总面积约为 4.27 万 km^2 ,下辖延吉、图们、珲春、龙井、和龙、敦化 6 个市和汪清、安图 2 个县。延边地区是吉林省内人类繁衍历史最长的地区之一,1963 年在安图县明月镇的洞穴内发现距今 2.6 万年前的“安图人”牙齿化石,而且在龙井、和龙、汪清、延吉、珲春等地也发现过新旧石器时期的遗址。资料显示,延边地区较早的居民主要有沃沮人、肃慎人、女真人等,高句丽、渤海、辽金、明朝、清朝等历代王朝均把延边地区作为领土的一部分进行治理,其中渤海国同延边地区的关系最为密切,它最初定都于现今的敦化市,之后在延边地区设置中京和东京,使延边地区成为渤海国的中心地之一,从而在延边地区留存了大量的渤海国时期遗址和遗物,这使延边地区成为渤海国史研究的重要地区之一。

1.2 数据来源

本研究所涉及的延边地区渤海国遗址信息来自《中国文物地图集—吉林分册》^[9]、《高句丽渤海

古城址研究汇编》^[10]、《延吉市文物志》^[11]、《图们市文物志》^[12]、《敦化市文物志》^[13]、《珲春县文物志》^[14]、《龙井县文物志》^[15]、《和龙县文物志》^[16]、《安图县文物志》^[17]、《汪清县文物志》^[18]等文献资料。经过资料整理后共得到 226 处遗址,其分布情况参见图 1。本文中所用的 DEM 的空间分辨率为 30 m ,并利用该 DEM 获取高程、坡度、坡向、山脊线、山谷线等地形地貌数据。从矢量化的“延边朝鲜族自治州行政区划图”^[19]中获取研究区的道路图、村屯分布图、水系图等专题图。

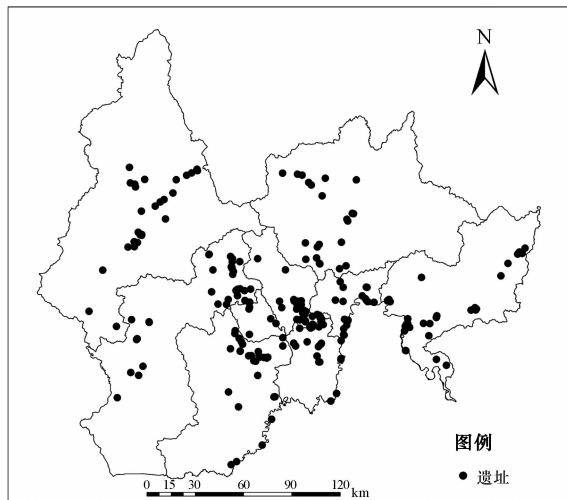


图 1 延边地区渤海国遗址分布图

2 遗址预测模型

设 P 为某事件发生的概率,取值范围为 $[0, 1]$, $1 - P$ 为该事件不发生的概率,将比数 $\frac{P}{1 - P}$ 取自然对数得 $\ln \frac{P}{1 - P}$,即对 P 作 logit 转换,记为 $\text{logit}P$, $\text{logit}P$ 的取值范围为 $(-\infty, +\infty)$ 。以 P 为因变量,建立如下线性回归方程^[20]:

$$\text{logit}P = \alpha + \beta_1\chi_1 + \beta_2\chi_2 + \cdots + \beta_m\chi_m,$$

由上式可得 $P = 1 / (1 + e^{-L})$,该模型即为 Logistic 回归模型。该模型实际上是普通多元线性回归模型的推广,但它的误差项服从二项分布而非正态分布,模型中 α 为常数项, β_i 为 Logistic 回归系数。

2.1 样本选取

本文中使用的样本是建立模型和验证模型时所需的数据。为了准确性,样本中要包含遗址和非遗址数据。遗址数据是由前述文献资料汇总而得,

非遗址数据是利用 ArcGIS 的随机点生成工具在遗址点以外的区域中生成的随机点,然后假设这些随机点为非遗址点。

通过上述样本选取方法,在研究区域中随机选取的建模样本数量共为 238 个,其中遗址和非遗址数量各为 119 个。遗址中聚落址、山城址、平原城址、墓葬墓群、寺庙址、古建筑址、其他等遗址数量依次为 49、11、21、22、4、9、3 个,占整个遗址数量的 52.7%。

通过同样方法获取的验证样本数共为 214 个,其中遗址和非遗址数量各为 107 个。遗址中聚落址、山城址、平原城址、墓葬墓群、寺庙址、古建筑址、其他等遗址数量依次为 48、6、23、17、2、9、2 个,占整个遗址数量的 47.3%。遗址类型的具体分布状况如图 2 所示。

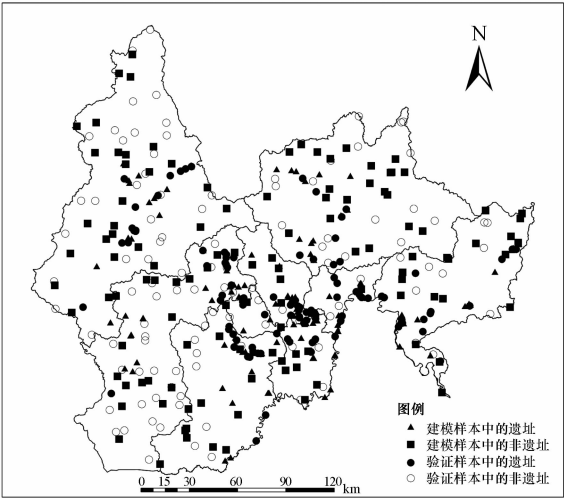


图 2 建模样本和验证样本中遗址和非遗址点分布图

2.2 模型变量选取

Logistic 回归模型的因变量只有两个值,即遗址点为 1,非遗址点(随机点)为 0。

模型的自变量是影响遗址分布的自然因素和人文因素,在充分考虑数据的可获取性和模型建立必要性的基础上,本文选取海拔高度、坡度、坡向(方位)、与河流之间的水平距离、地形起伏度、与山脊线和山谷线之间的距离、土壤类型、植被类型等自然因素之外,还选取了与道路之间的距离、与村屯之间的距离和土地利用类型等人文因素。具体自变量及取值范围如表 1 所示,其中土地利用类型、植被类型、土壤类型属于定性变量,因此

未给出取值范围。

表 1 自变量及其取值范围

自变量	自变量取值范围		
	最小值	最大值	平均值
高度/m	14.24	1375.9	481.3
坡度/(°)	0.0	61.8	6.4
坡向/(°)	平地(-1)	359.7	127.2
地形起伏度/m	0.0	170.3	9.3
与山脊线之间的距离/m	0.4	2124.2	358.3
与山谷线之间的距离/m	4.4	2346.9	396.2
与河流之间的水平距离/m	0.0	3212.3	512.5
与道路之间的距离/m	0.0	43714.1	1 244.3
与村屯之间的距离/m	0.0	32 420.7	2 701.2
土地利用类型	—	—	—
植被类型	—	—	—
土壤类型	—	—	—

2.3 Logistic 回归模型

建立模型时,自变量进入模型的方法有“输入”、“向前:条件”、“向前:LR”、“向前:Wald”、“向后:条件”、“向后:LR”、“向后:Wald”等方法,为了获取预测效果最好的模型,本文对各方法下的建模样本和验证样本的预测准确率进行了对比,其结果见表 2。表 2 表明,用“输入”方法将自变量选入模型时,其对样本的预测准确率最高。对建模样本的 119 个非遗址中,正确预测 107 个,准确率为 89.92%;对 119 个遗址中正确预测 112 个,准确率为 94.12%;总准确率为 92.02%。对验证样本的 107 个非遗址中实际参与 104 个,正确预测 92 个,准确率为 88.46%;对验证样本的 107 个遗址中实际参与 101 个,正确预测 87 个,预测准确率为 86.14%;总准确率为 87.32%。因此,本文最终选取“输入”方法下所建立的模型。最终回归模型为

$$P = 1/(1 + e^{-L}),$$

其中 $L = -0.003 \times \text{高度} - 0.017 \times \text{坡度} - 0.001 \times \text{坡向} + 0.025 \times \text{地形起伏度} + 0.001 \times \text{与山脊线之间的距离} + \dots + 23.654 \times \text{常量}$ (因模型参数过多,在此部分省略,具体参数见表 3)。

3 预测结果分析

3.1 不同概率区的遗址分布

基于已建立的 Logistic 回归模型,利用 Arc-

GIS 的栅格计算器计算研究区域的遗址分布概率图,为了分析不同概率值的分布情况,采用自然分等级,等级越高表示遗址存在的概率越高.各概率图,为了分析不同概率值的分布情况,采用自然分区遗址分布和面积比如图 3 和表 4 所示.

裂法(Jenks)将概率图重新分类为低、中、高 3 个

表 2 各方法对遗址样本分类结果

方法	观测	预测					
	是否遗址	建模样本			验证样本		
		是否遗址		%	是否遗址		%
		否	是		否	是	
输入	否	107	12	89.92	92	12	88.46
	是	7	112	94.12	14	87	86.14
	%			92.02			87.32
向前:条件	否	104	15	87.39	88	16	84.62
	是	9	110	92.44	13	88	87.13
	%			89.92			85.85
向前:LR	否	104	15	87.39	88	16	84.62
	是	9	110	92.44	13	88	87.13
	%			89.92			85.85
向前:Wald	否	93	26	78.15	82	22	78.85
	是	11	108	90.76	10	91	90.10
	%			84.45			84.39
向后:条件	否	104	15	87.39	88	16	84.62
	是	9	110	92.44	13	88	87.13
	%			89.92			85.85
向后:LR	否	104	15	87.39	88	16	84.62
	是	9	110	92.44	13	88	87.13
	%			89.92			85.85
向后:Wald	否	101	18	84.87	85	19	81.73
	是	11	108	90.76	15	86	85.15
	%			87.82			83.41

表 3 遗址预测模型的参数

自变量	B	自变量	B	自变量	B
高度	−0.003	植被类型(农地)	−42.287	土壤类型(未分类)	−61.904
坡度	−0.017	植被类型(杨桦林)	−19.585	土壤类型(白浆土)	−22.721
坡向	−0.001	植被类型(针阔混交林)	−38.935	土壤类型(草甸土)	−22.045
地形起伏度	0.025	植被类型(柞树林)	−19.525	土壤类型(灰棕壤)	−22.951
与山脊线之间的距离	0.001	植被类型(灌木林)	0.811	土壤类型(冲积土)	−21.176
与山谷线之间的距离	0.001	植被类型(荒山荒地)	−19.301	土壤类型(暗棕壤)	−21.083
与河流之间的水平距离	−0.002	植被类型(牧草地)	−37.089	土壤类型(灰化土)	−38.489
与道路之间的距离	0.000	植被类型(岩石裸露地)	−18.334	土壤类型(水稻土)	−20.970
与村屯之间的距离	−0.001	土地利用类型(未利用土地)	21.335	土壤类型(新积土)	−21.438
植被类型(火烧迹地)	−41.089	土地利用类型(耕地)	21.513	土壤类型(石灰岩土)	−40.398
植被类型(疏林地)	−19.327	土地利用类型(林地)	21.368	土壤类型(风沙土)	−45.672
植被类型(沼泽地)	−39.155	土地利用类型(牧草地)	41.520	土壤类型(黑土)	−2.864
植被类型(灌丛地)	−19.002	土地利用类型(居民点及工矿用地)	38.152	常量	23.654
植被类型(阔叶林)	−20.178	土壤类型(沼泽地)	−22.177		

注: B 为偏回归系数.

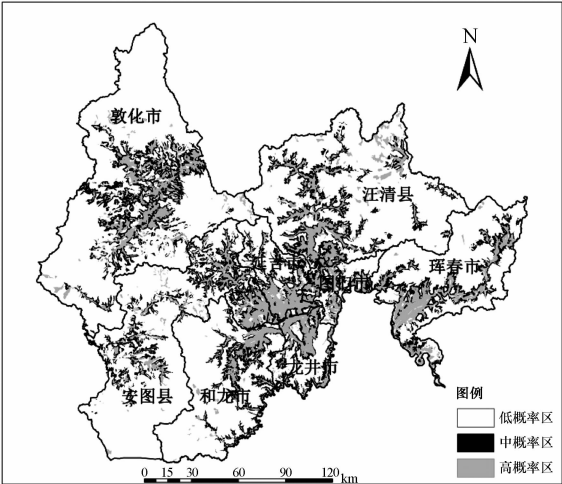


图 3 延边地区渤海国遗址分布概率示意图

表 4 各概率区遗址分布数和面积比

遗址存在 概率等级	遗址 数量(%)	面积 (栅格数量)(%)
低概率区	17(7.5)	35 127 114(73.2)
中概率区	20(8.8)	5 787 519(12.1)
高概率区	189(83.6)	7 041 938(14.7)

3.2 高概率区在各流域的分布

为了更为详细地观察遗址在各河流流域的分布状况,将研究区分为图们江干流流域、松花江流域、牡丹江流域、布尔哈通河流域、海兰江流域、嘎呀河流域、绥芬河流域、珲春河流域等.将河流专题图和遗址分布高概率区重叠后发现,高概率区主要分布在布尔哈通河流域(干流、长兴河、福兴河、细鳞河、依兰河)、海兰江流域(干流、福洞河、长仁河)、嘎呀河流域(干流、汪清河、新兴河)、珲春河流域(干流、松林河)、牡丹江流域(干流、大石河、沙河、官地河)、图们江干流流域等(图 4).此外,松花江流域的古洞河和五道白河流域也比较集中.

3.3 各类概率区在各县市的分布

各类概率区在各县市的分布情况如表 5 所示.从表 5 中可知:各县市面积中高概率区面积所占比重(相对比重)较高的有龙井、图们、延吉、珲春等县市,其所占比重依次为 36.9%、35.8%、30.3%、17.7%,这说明这些地区渤海遗址分布密度高的可能性大;各县市高概率区面积占延边地区高概率区总面积的比重(绝对比重)较高的是敦化、汪清、珲春、龙井、安图等县市,其所占比重依

次为 20.9%、18.4%、14.2%、12.9%、10.5%,这说明这些地区渤海遗址分布绝对量多的可能性大.

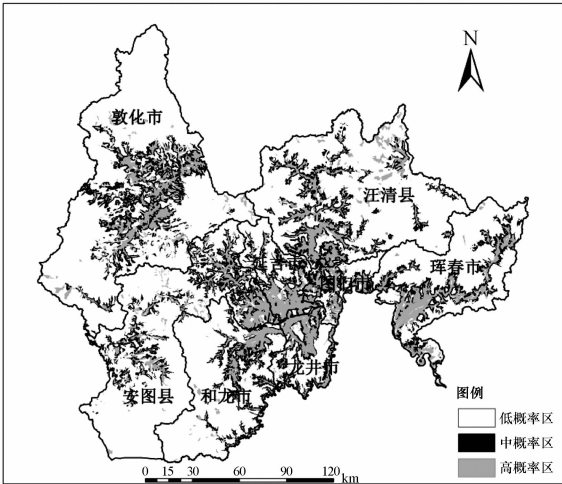


图 4 延边地区渤海国遗址高概率区水系分布图

表 5 各县市的各等级概率区分布情况 %

县市	各县市各概率区相对比重			各县市高概率区绝对比重
	低概率区	中概率区	高概率区	
延吉市	52.2	17.5	30.3	8.2
图们市	43.3	20.8	35.8	6.4
敦化市	77.7	11.0	11.3	20.9
珲春市	69.4	12.8	17.7	14.2
龙井市	44.3	18.8	36.9	12.9
和龙市	78.4	10.8	10.8	8.6
汪清县	75.6	11.1	13.3	18.4
安图县	80.4	10.7	9.0	10.5
合计				100.0

4 结束语

本文选取遗址和非遗址的海拔高度、坡度、坡向(方位)、与河流之间的水平距离、地形起伏度、与山脊线和山谷线之间的距离、土壤类型、植被类型、土地利用类型等自然因素和与道路之间的距离、与村屯之间的距离、土地利用类型等人文因素,利用 Logistic 回归模型建立了延边地区渤海国遗址的预测模型,并给出了遗址分布概率图.研究表明:

1) 模型建立过程中,采用“输入”方法的预测效果最好,对建模样本的预测准确率为 92.02%,对验证样本的预测准确率为 87.32%,总预测准确率为 89.7%.

2) 用自然分裂法将遗址分布概率图分为高、中、低 3 个等级概率区,其中高概率区占研究区的 14.7%,遗址数量为 189 个(83.6%).

3) 高概率区具有向河流聚集分布的特征,主要分布在海兰江流域、布尔哈通河流域、嘎呀河流域、牡丹江流域、图们江干流流域等.

4) 各县市面积中高概率区面积所占比重较高的地区是龙井、图们、延吉、珲春等县市,其所占比重依次为 36.9%、35.8%、30.3%、17.7%;各县市高概率区面积占延边地区高概率区总面积的比重较高的是敦化、汪清、珲春、龙井、安图等县市,其所占比重依次为 20.9%、18.4%、14.2%、12.9%、10.5%.

5) 目前为止,渤海国还有很多遗址没有被发现,本文所得结论可为制定渤海国考古计划和选定考古范围提供有效的依据,从而有助于提高考古工作效率,节省人力和资金等.

参考文献:

- [1] 高立兵. 时空解释新手段:欧美考古 GIS 研究的历史现状和未来[J]. 考古,1997(7):89-95.
- [2] Willey G R. Prehistoric Settlement in the Virúvalley, Peru[M]. Washington: Bureau of American Ethnology Bulletin 155, 1953.
- [3] Michael Märker, Saman Heydari-Guran. Application of datamining technologies to predict Paleolithic site locations in the Zagros Mountains of Iran[J]. Computer Applications to Archaeology 2009 Williamsburg, Virginia, USA, 2009:1-7.

- [4] 倪金生. 山东沭河上游流域考古遗址预测模型[J]. 地理科学进展,2009,28(4):489-492.
- [5] 乔文文,毕硕本,王启富,等. 郑洛地区龙山文化遗址预测模型[J]. 测绘科学,2013,38(6):172-181.
- [6] 彭淑贞,张伟,陈栋栋. 汶泗流域大汶口文化考古遗址模型预测[J]. 泰山学院学报,2010,32(6):34-39.
- [7] 延边朝鲜族自治州编撰委员会. 延边朝鲜族自治州土地志[M]. 延吉:延边人民出版社,2002.
- [8] 吉林省延吉市地方志编撰委员会. 延吉市志[M]. 北京:新华出版社,1994.
- [9] 国家文物局. 中国文物地图集:吉林分册[M]. 北京:中国地图出版社,1992.
- [10] 王禹浪,王宏北. 高句丽渤海古城址研究汇编[M]. 哈尔滨:哈尔滨出版社,1994.
- [11] 《吉林省文物志》编委会. 延吉市文物志[M]. 长春:吉林省文物志编修委员会,1983.
- [12] 《吉林省文物志》编委会. 图们市文物志[M]. 长春:吉林省文物志编修委员会,1985.
- [13] 《吉林省文物志》编委会. 敦化市文物志[M]. 长春:吉林省文物志编修委员会,1985.
- [14] 《吉林省文物志》编委会. 珲春县文物志[M]. 长春:吉林省文物志编修委员会,1984.
- [15] 《吉林省文物志》编委会. 龙井县文物志[M]. 长春:吉林省文物志编修委员会,1984.
- [16] 《吉林省文物志》编委会. 和龙县文物志[M]. 长春:吉林省文物志编修委员会,1984.
- [17] 《吉林省文物志》编委会. 安图县文物志[M]. 长春:吉林省文物志编修委员会,1985.
- [18] 《吉林省文物志》编委会. 汪清县文物志[M]. 长春:吉林省文物志编修委员会,1983.
- [19] 延边朝鲜族自治州民政局. 延边朝鲜族自治州行政区划图[M]. 长沙:湖南地图出版社,2009.
- [20] 王济川,郭志刚. Logistic 回归模型:方法与应用[M]. 北京:高等教育出版社,2001.