DOI:10.13203/j.whugis20150260　　文章编号:1671-8860(2017)03-0390-05

利用逻辑回归的南极常年考察站选址定量研究

刘海燕[1,2,3]　庞小平[1,2,3]　王　跃[1,3]　李忠香[1,3]

1　武汉大学中国南极测绘研究中心,湖北 武汉,430079
2　武汉大学资源与环境科学学院,湖北 武汉,430079
3　武汉大学极地测绘科学国家测绘地理信息局重点实验室,湖北 武汉,430079

摘　要:南极常年考察站是人类在南极地区科考、探索以及观光等活动的据点,其选址对南极自然环境产生一定影响并制约考察站本身功能。本文突破传统考察站选址方法,基于地理信息系统,采用逻辑回归模型模拟已建站情况,使用定量化方法建立指标体系,评价常年站建站适宜性程度。实验过程摒除了人工指标适宜性分类以及专家打分等过程,降低了人为因素带来的不确定性。实验证明,其对已建站情况的拟合精度大于全指标模型,可以更好地模拟常年站的选址情况。

关键词:逻辑回归;定量化模型;选址;南极考察站

中图法分类号:P208　　**文献标志码**:A

随着对全球气候变化研究的日益深入,各国纷纷建立南极考察站以支持人类在南极地区的科考、探索及观光等活动。考察站位置的选择将直接影响科考活动的类别和范围、南极自然生态环境[1]以及考察站本身的运行寿命等。区别于先专家选择后对若干备选地址一一实地验证的传统选址方法[2-4],定量与定性结合的多目标选址已被用于南极考察站建站选址研究中[5-6]。模糊层次分析法等定量方法的引入虽然让选址过程更加科学合理,为决策提供定量化依据,但因将所有指标都纳入评价并进行人为处理和专家打分,受人为因素干扰较多,存在过度拟合的可能性。

针对上述问题,本文将定量与定性的结合分析完全转化为定量化模型。将从各个数据源收集的南极相关数据不作人工分类,利用逻辑回归的方法求出所有指标对于建站的相关性情况及适宜性程度,进一步得到适宜性评价图。该模型完全以已建立考察站环境数据为依托,使用定量化方法建立指标体系,以数学统计方法剔除与其无明显相关性的指标以及互相高度相关的指标[7-9],使用最小数据包模拟建站适宜性程度,缩减模型的复杂性,脱离人为因素的干扰,使评价更加客观具体,为决策提供科学判断依据。

1　实验区域及数据

本文研究区域覆盖了南纬 60°以南的所有南极区域,包括了陆地和冰架。陆地冰或与大陆架相连的冰体延伸到海洋的部分称为冰架,因其有较高的冰流速在传统认知上不利于考察站的长期活动。但是在已有的常年考察站中,如英国哈雷站、德国纽梅因站等都设立在冰架之上,故本次研究本着存在即有合理性的原则,将冰架也纳入分析区域。

不同类型考察站对选址要求各不相同,常年站要满足后勤保障人员和考察队员的生活工作需要,建设和维护费用高,故对站址选择要求高;夏季站每年夏天开放,一般规模相对较小,设备和建筑都较简易,大多数建立在条件恶劣而又特别具有科学研究意义的地区;无人站安装各种自动化检测设备,对建站环境要求最低。本文以常年站选址为例进行分析,研究方法可推至其他类型考察站选址评价。逻辑回归作为一种经典的近似方法,其指标体系的选择是由数据驱动而不是人为选择,大大降低了人工决策的主观性,但是分析指标的输入仍需要人为采集[7]。

本文在分析大量文献和各国建站的综合环境

收稿日期:2016-03-07
项目资助:中国极地科学战略研究基金(20150407);中国博士后科学基金(2016M592373)。
第一作者:刘海燕,讲师,主要从事地图学与地理信息系统、数据模型分析和不确定性等方面研究。liuhaiyan@whu.edu.cn
通讯作者:庞小平,博士,教授。pxp@whu.edu.cn

评价报告的基础上[2-4],从权威机构获取最新公开发布的数据,以保证分析数据的正确性和完整性。因变量指标 y"是否在已建成考察站范围之内"为二值型,自变量指标(X_1,…,X_{21})皆为连续型,分别对应高程、坡度、冰厚度、冰流速、积雪率、到冰架的距离、到海冰边缘的距离、到海岸线的距离、到露岩的距离、到湖泊的距离、到机场的距离、到植物存在区域的距离、到海鸟和企鹅活动范围的距离、到南极特别保护区的距离、到南极特别管理区的距离、到冰碛(冰川沉积物)的距离、到雪丘的距离、一月风速、七月风速、一月温度以及七月温度等,距离统一为欧几里得直线距离。

2　考察站选址适宜性评价模型的构建

模型构建可分为以下几个步骤:① 全指标适宜性指数计算;② 根据指标相互之间的相关性和指标与分析结果的显著程度选取部分指标,构建最优模型;③ 根据最优模型中的指标及其对应的适宜性指数(权重)输出整个研究地区的适宜性程度图。

2.1　指标适宜度指数(权重)计算

逻辑回归(logistic regression)是对二分类因变量(即 $y=0$ 或 $y=1$)进行回归分析时最为普遍应用的多元量化分析方法,因其因变量与自变量之间的为非线性关系且不需要假设自变量之间存在多元正态分布,在社会学、经济学、自然地理学及公共卫生学中得到广泛应用[8-10]。本文分析南极考察站的建站适宜性程度,其因变量为适宜建站和不适宜建站两类,其自变量都为连续型,在逻辑回归中当多元共线性不太严重时,其系数估计基本还是无偏且有效的,几乎可以忽略共线性带来的影响。因此,采用二元逻辑回归进行分析建模,其计算公式为:

$$y = a + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n \quad (1)$$

$$y = \ln\left(\frac{P}{1-P}\right) \quad (2)$$

$$P = \frac{e^y}{1+e^y} \quad (3)$$

式中,y 为因变量,本文中 1 代表此地已建站或完全适宜建站,0 代表此地未建站或完全不适宜建站;(X_1,X_2,…,X_n)为自变量,本次研究中代表各个数据指标,用于拟合最优的建站适宜程度;(b_1,b_2,…,b_n)为待求的回归系数,通过加权最小二乘法进行求解;P 表示事件发生的概率,即适应建站程度。

2.2　指标选择及模型的建立

为了保证数据的完整性,我们收集了与建站相关的所有指标,但是在这些指标中有的相对建站适宜性显著水平较低,有的指标之间相关度较高,有时对于特定的数据可能有几个候选的模型。在建立模型时,要检验所有可能的模型,以可解释性、检阅性、变量的易得性等作为依据,从中选择一个模型作为"最优"模型。

本文实验采用自动逐步向前 wald(先对无约束模型得到参数的估计值,再代入约束条件检查是否成立)方法,对模型进行调整,以确定最优指标体系。这样自动逐步避免了多次运行回归程序取得不同模型最大似然函数的对数之差(似然比 LR 检验)比较,从而一次得到统计结果,简化了运算过程[9],提高计算效率。得到模型需要通过拟合优度、准确性和模型 χ^2 检验等多个参数综合评价,确定最优模型。

2.3　建站适宜性图及验证

将自变量数值代入拟合模型中,利用回归方法确定指标的权重(逻辑回归系数),根据式(3)得到最后的建站适宜性评价图。将逻辑回归得到的适宜性地图与模糊层次分析法得到的结果图进行对比,以此来验证模型的准确性。

3　实验结果和分析

3.1　数据预处理及样本点的选择

由于数据来源以及类型的多样化,为分析的简便性将所有数据进行统一处理,在被转换为极方位投影后,统一将矢量数据转换为与地形数据相同的 1 km 分辨率的栅格数据。将已建成的考察站范围的因变量 y 设为 1,用于模拟适宜建站的条件,对不在已建成考察站范围随机选择相同数量的样本点,设其因变量 $y=0$,视为不适合建站。将选择样本点的所有指标对应值进行提取,作为自变量 X 的值,用于拟合逻辑回归模型。所有数据都在研究区域范围内进行处理,数据的准备和分析工作在 ArcGIS 和 SPSS 19.0 中进行。

对整个南极地区而言,人类活动所能到达的地区很少,考察站建立的位置就更小,为了取得足够的数据来支撑逻辑回归分析,需对已建站区域进行加密采样,依次间隔 1 000 m、800 m、500 m 和 300 m 取 $y=1$ 的因变量样点,然后在整个研究范围内平均随机抽取同数量相同 $y=0$ 的样本点。对不同的样本分别进行建模分析,对比不同

样本数量下指标的显著性情况，以选取最优采样比例，稳定最优模型。经过比对，随着采样点的加密，所有指标都对结果有影响，但所得模型不能很好地拟合已有数据(300 m间隔样本)。而800 m样本最优模型与500 m样本最优模型中自变量及其对应系数正负相同，只在数值上有细小差异，故采用500 m样本的13 766个随机样本点(占总样本的6.7%)分析建站适宜性趋势。

3.2 模型及对应指标权重的确定

采用向前wald的自动逐次方法，依次添加重要预测变量，经过9次迭代得到最优模型(表1)。最优模型中所有指标的都在0.05层次上与结果显著相关(sig<0.05)，显示建站适宜性程度可以由坡度、冰厚度、冰流速、到冰架距离、到海冰边缘距离、到湖泊距离、到机场距离、到雪丘距离以及七月风速等9个指标模拟而成。因不同数据保留其原有的单位和量纲，所以最优模型显示指标本身单位尺度对建站适宜性的影响程度。

表1 逻辑回归的全指标模型及最优模型

Tab.1 Model Including all Criteria and the Best Model Based on Logistic Regression

	全指标模型			最优模型		
	b	Sig	Exp(b)	b	Sig	Exp(b)
X_1	−0.001	0.01	0.999	/	/	/
X_2	−0.583	0.00	0.558	−0.666	0.00	0.514
X_3	−0.002	0.00	0.998	−0.002	0.00	0.998
X_4	−0.007	0.01	0.993	−0.006	0.02	0.994
X_5	0	0.49	1	/	/	/
X_6	0.003	0.03	1.003	0.004	0.00	1.004
X_7	0	0.93	1	−0.002	0.00	0.998
X_8	−0.006	0.02	0.994	/	/	/
X_9	−0.006	0.03	0.994	/	/	/
X_{10}	0.002	0.06	1.002	0.001	0.02	1.001
X_{11}	−0.008	0.00	0.992	−0.008	0.00	0.992
X_{12}	0.007	0.00	1.007	/	/	/
X_{13}	−0.003	0.04	0.997	/	/	/
X_{14}	−0.002	0.06	0.998	/	/	/
X_{15}	0	0.85	1	/	/	/
X_{16}	0	0.81	1	/	/	/
X_{17}	0	0.82	1	−0.001	0.00	0.999
X_{18}	0.439	0.03	1.551	/	/	/
X_{19}	0.576	0.00	1.779	0.712	0.00	2.038
X_{20}	−0.076	0.18	0.927	/	/	/
X_{21}	−0.045	0.27	0.956	/	/	/
a	−10.309	0.00	0	−7.102	0.00	0.001

从最优模型中可以看出，已建成站对坡度、到机场距离以及七月风速较为敏感。其中坡度每增加1%，建站的可能性变为原来的51.4%(Exp(b)=0.514)，即坡度增加更不利于建站。随着人类对南极考察的逐渐深入，对南极内陆地区的考察日益增多，固定翼飞机和直升机的使用大大便利了内陆站的考察和后勤补给，也使得机场的作用日益突出。

从全指标模型和最优模型的对比可以发现，在全模型中无明显统计意义(sig>0.05)的数据，如到海冰边缘、雪丘距离、七月风速等，在最优模型中较好拟合已建站环境。这是由于在全指标模型中，部分数据相关性较高(如高程与冰厚度、到海冰边缘与海岸线距离、七月风速与七月温度等、到雪丘与冰碛距离等)，故在全模型中的弱相关，在迭代分析中随着相关变量的移除可能变成重要的自变量。这也解释了为何建站适宜性对七月风速较为敏感，七月风速与一月风速、一月温度以及七月温度都有较高的相关性，后三者都没有出现在最优模型中，七月风速在最优模型中成为天气情况的唯一依托指标。

3.3 选址适宜性评价模型与对比验证

经过Hosmer和Lemeshow检验(表2)，最优模型的概率值0.310大于全指标模型的0.068，且大于迭代过程中产生的其他模型概率值，显示此模型可以更好地拟合已建站数据。

表2 Hosmer和Lemeshow检验结果

Tab.2 Result of Hosmer and Lemeshow Test

	全指标模型	最优模型
卡方检验	14.568	9.394
df(自由度)	8	8
Sig(P)	0.068	0.310

将得到回归系数代入式(1)得到建站适宜性图(图1)，适宜建站地区保留南极特别管理区，排除了南极特别保护区及雪丘区域。虽然南极特别保护区除非符合管理计划的要求并取得进入许可证，否则禁止进入[11]，所以在建站层面上不予考虑。而南极特别管理区部分是对人类活动较多需要协调和规划的区域进行管理，原则上可以进入[12]，并且已有建成站(如美国阿蒙森斯科特站)在特别管理区之内，所以在本次研究中，特别管理区并不做要求。雪丘地区由于特殊的地形条件，也并不适合建站，故都在图1中予以标明。

与同区域的模糊层次分析法实验结果相比较(图2)[5]，本次实验具有以下优点：①添加了冰流速等数据，涵盖数据范围更广，为整个南极环境的建模提供了更多的细节，降低了玛丽伯德地(Marie Byrd Land)大部分区域建站适宜性程度，提高了数据模拟的精确程度；②避免指标的人为分层和专家打分，降低了人工决策的主观性，提高了数据的容积，得出整个南极的建站适宜性程度；③选取与结果拟合程度最高的指标体系，简化运算体

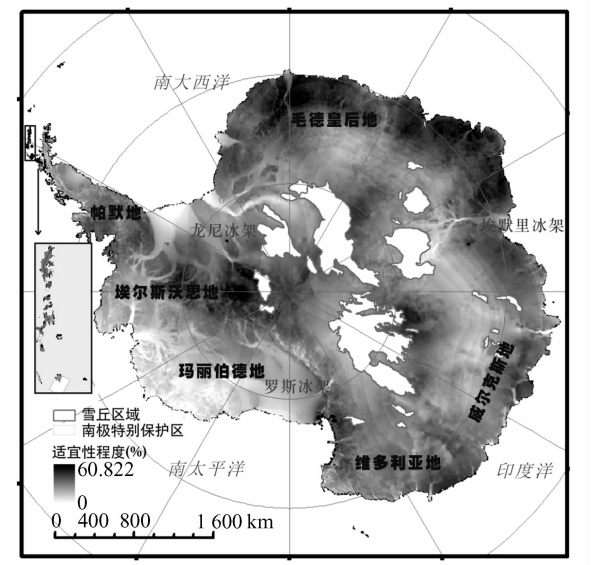

图 1　建站适宜性图,放大图为南设得兰群岛

Fig. 1　Map of Site Suitability, the Area Covering South Shetland Islands is Enlarged

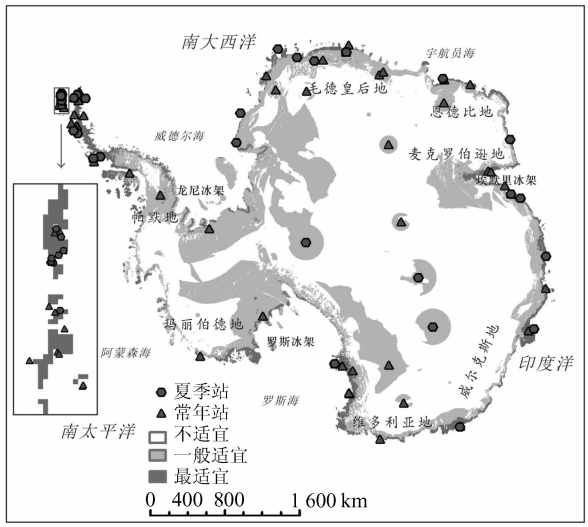

图 2　模糊层次分析法获得建站适宜区域

Fig. 2　Suitable Sites Derived from Fuzzy Analytical Hierarchy Process

系,避免数据的过度拟合,提高了运算效率。

由图 1 可见,虽然研究将冰架区域涵盖在内,可是较高的冰流速仍使该区域建站适宜性程度较低。具体适宜性分析为:① 南设得兰群岛(South Shetland Islands)和海岸线周围由于气候状况良好,有较好的可达性,为生态环境研究,海洋物理化学研究和南极观光的热点地区,故为考察站建设集中地区;② 随着全球气候变化,冰雪成为南极研究的新焦点。帕默地(Palmer Land)(靠近龙尼冰架(Ronne Ice Shelf)),埃默里冰架(Amery Ice Shelf)周边,以及维多利亚地(Victoria Land)周边地区都为适宜研究区域;③ 毛德皇后地(Queen Dronning Maud Land)航空网络项目(DROMLAN)的协议达成,使得毛德皇后地成为东南极的交通枢纽,也增加了毛德皇后地的建站适宜性程度;④ 空中运输的广泛应用,使得建站范围不局限于海洋周边,埃尔斯沃思地(Ellsworth Land)为冰原地形,海拔较高(有南极大陆最高点),适宜高空物理等项目的研究,且大部分区域未被开发,可以视为新的建站拓展区域。

4　结　语

南极地区因为其环境的特殊性,选址考虑的内容也相对复杂,对已建成考察站进行模拟和对全南极适宜建站程度分析往往受制于模型。寻求一种简单可行易于构建的模型,探求建站过程中各种因素对选址决策的影响机理,对人类的南极活动具有重要的实际意义。将逻辑回归与南极考察站选址相结合,将传统定性分析或者定量与定性结合的方法完全转化为定量分析,减少了人为数据分类和专家打分带来的不确定性,增加了环境模拟的科学性和系统性。因变量的可变性和自变量的灵活添加修改使得此模型在南极环境模拟方面具有普遍适用性。但是由于本实验使用已建成考察站作为模拟因变量,相对于整个南极环境的样本较少,使得模拟精度有待进一步验证。

参　考　文　献

[1] Chen Jie, Blume H P. Anthropic Impacts on the Antarctic Terrestrial Ecosystem[J]. *Chinese Journal of Polar Research*, 2000, 12(1): 62-75(陈杰, Blume H P. 人类活动对南极陆地生态系统的影响[J]. 极地研究, 2000, 12(1): 62-75)

[2] Belgian Federal Science Policy Office. The International Polar Foundation. Final Comprehensive Environmental Evaluation: Construction and Operation of the New Belgian Research Station, Dronning Maud Land, Antarctica[C]. The 28th Antarctic Treaty Consultative Meeting, Stockholm, Sweden, 2005

[3] British Antarctic Survey. Final Comprehensive Environmental Evaluation: Proposed Construction and Operation of Halley VI Research Station, and Demolition and Removal of Halley V Research Station, Brunt ice Shelf, Antarctica[C]. The 28th Antarctic Treaty Consultative Meeting, Stockholm, Sweden, 2005

[4] Korea Polar Research Institute, Korea Environment Institute. Final Comprehensive Environmental Eval-

uation: Construction and Operation of the Jang Bogo Antarctic Research Station, Terra Nova Bay, Antarctica[C]. The 35th Antarctic Treaty Consultative Meeting, Hobart, Australia, 2012

[5] Pang Xiaoping, Liu Haiyan, Zhao Xi. Selecting Suitable Sites for an Antarctic Research Station: A Case for a New Chinese Research Station[J]. *Antarctic Science*, 2014,26(5): 479-490

[6] Liu Haiyan, Pang Xiaoping. Antarctic Research Station Site Selection Based on GIS and Fuzzy AHP [J]. *Geomatics and Information Science of Wuhan University*, 2015,40(2):249-252(刘海燕,庞小平. 基于GIS和模糊层次分析法的南极考察站选址研究[J]. 武汉大学学报·信息科学版,2015,40(2):249-252)

[7] Ho Zhiyong, Lo C P. Modeling Urban Growth in Atlanta Using Logistic Regression[J]. *Computer Environment and Urban Systems*, 2007, 31: 667-688

[8] Menard S. Applied Logistic Regression Analysis [M]. Shanghai: True & Wisdom Press, 2009(Menard S. 应用Logistic回归分析[M]. 上海:格致出版社,2009)

[9] Wang Jichuan, Guo Zhigang. Logistic Regression Model-Method and Application[M]. Beijing: Higher Education Press, 2001(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)

[10] Jiang Wenliang, Zhang Xiaotong, Li Lin, et al. Urban Spatial Expansion Forecast Based on GIS & Spatial Logistic Regression Model[J]. *Science of Surveying and Mapping*, 2008,33(4):172-174(姜文亮,张晓通,李霖,等. 基于GIS和空间逻辑回归模型的城市空间拓展预测[J]. 测绘科学,2008,33(4):172-174)

[11] Ling Xiaoliang, Chen Danhong, Zhang Xia, et al. Review of the Status, Feature and Prospect of Antarctic Specially Protected Areas[J]. *Chinese Journal of Polar Research*, 2008,20(1):48-63(凌晓良,陈丹红,张侠,等. 南极特别保护区的现状与展望[J]. 极地研究,2008,20(1):48-63)

[12] Gu Yueting, Sun Bo, Chen Danhong, et al. Review of the Current Status, Feature and Future Development of Antarctic Specially Managed Area[J]. *Chinese Journal of Polar Research*, 2010,22(4):431-440(顾悦婷,孙波,陈丹红,等. 南极特别管理区现状分析与未来展望[J]. 极地研究,2010,22(4):431-440)

Quantitative Research for Site-selection of Antarctic Year-round Research Stations Based on Logistic Regression

LIU Haiyan[1,2,3] *PANG Xiaoping*[1,2,3] *WANG Yue*[1,3] *LI Zhongxiang*[1,3]

1 Chinese Antarctic Center of Surveying and Mapping, Wuhan University, Wuhan 430079, China

2 School of Resources and Environment Science, Wuhan University, Wuhan 430079, China

3 Key Laboratory of Polar Surveying and Mapping, National Adminstration of Surveying, Mapping and Geoinformation, Wuhan University, Wuhan 430079, China

Abstract: Antarctic year-round research stations are the foothold for Antarctic expedition, exploration, and sightseeing activities, whose selection affects the natural environment of Antarctica as well as the function and operational efficiency of the stations. Different from traditional site selection methods, this study utilizes the logistic regression, a quantitative method, to build the criteria system and identify their weights and model the construction environment of Antarctica with the aids from geographical information systems. Due to the elimination of artificial data classification and expert scoring process, the logistic model generated in this study largely reduced the uncertainties brought by human decisions. It turned out that the logistic model was more suitable to the construction environment than the model with all criteria, and more precise and detailed, showing the feasibility of the results.

Key words: logistic regression; quantitative model; site selection; Antarctic research station

First author: LIU Haiyan, lecturer, specializes in cartography and GIS application, data model analysis and uncertainty. E-mail: liuhaiyan@whu.edu.cn

Corresponding author: PANG Xiaoping, PhD, professor. E-mail:pxp@whu.edu.cn

Foundation support: The Chinese Polar Strategy Fund, No. 20150407;the China Postdoctoral Science Foundation, No. 2016M592373.