文章类型：论文|刊号（ISSN）：2630-4740 /（中图刊号）：715GL012

基于 Logistic 回归模型的成都遥望雪山景观预报研究

文雯[1,2]　张新龙[3]　刘自牧[1,2*]　黄瑶[1,2]
1 四川省气象服务中心　2 高原与盆地暴雨旱涝灾害四川省重点实验室　3 成都信息工程大学
DOI:10.12238/eep.v8i6.2738

[摘　要] 本文利用成都地区气象站各气象要素资料、空气质量自动监测站和观山记录分析了遥望雪山景观的预报关键性因素。对各气象因素与景观出现事件做Spearman秩相关检验,对景观事件出现与否这样的二分类事件进行了分区域的Logistic回归,并对预报模型进行了回代检验。结果表明：(1)在对各气象因素和景观是否出现的相关性分析当中,可以发现景观是否出现与最高气温、平均湿度、降水量、平均风速和能见度显著相关。(2)将成都市分为四个地区,分别根据Logistic回归对各地区建立预报方程,再利用ROC曲线和约登指数,筛选各地区预报模型的最优临界值,最后将最优临界值回代进预报方程中进行测试检验,结果表明成都中部地区预报方程的预报效果最好。
[关键词] 成都；遥望雪山；预报；二分类事件；Logistic回归
中图分类号：TV149.2　文献标识码：A

Research on the Forecast of Chengdu Viewing Snow Mountain Landscape Based on Logistic Regression Model

Wen Wen[1,2]　Xinlong Zhang[3]　Zimu Liu[1,2*]　Yao Huang[1,2]
1 Sichuan Meteorological Service Center
2 Sichuan Provincial Key Laboratory of Heavy Rain, Drought and Flood Disasters in Plateaus and Basins
3 Chengdu University of Information Technology

[Abstract] The key factors for predicting the distant snow mountain landscape were analyzed using various meteorological element data from meteorological stations in Chengdu, automatic air quality monitoring stations, and mountain observation records. Spearman rank correlation test was performed on each meteorological factor and the occurrence of landscape events. Logistic regression was conducted on binary events such as the occurrence of landscape events in different regions, and a regression test was conducted on the prediction model. The results indicate that, firstly, in the analysis of the correlation between meteorological factors and the appearance of landscape, we can find that the appearance of landscape is significantly correlated with the maximum temperature, average humidity, precipitation, average wind speed and visibility. Secondly Chengdu is divided into four regions, and prediction equations are established for each region according to Logistic regression. Then, Receiver operating characteristic and Youden's J statistic are used to screen the optimal critical value of each region's prediction model. Finally, the optimal critical value is replaced into the prediction equation for testing and testing. The results show that the prediction equation in central Chengdu has the best prediction effect.
[Key words] Chengdu; Viewing the snowy mountains; Forecast; Category two events; Logistic regression

引言

成都是全球目前唯一能看到数座海拔超5000m雪山的千万级人口城市,其中以幺妹峰和贡嘎山最为出名。成都处于四川盆地西侧,与青藏高原接壤,当天气情况良好时,在市内就能够遥望到西侧海拔超过4000m的雪山。随着成都市生态环境的改善以及空气质量的提升,成都中心城区出现遥望雪山景观的次数从2016年的35次增加到了2021年的63次,总体上呈现逐年增加的趋势。

空气污染对于城市能见度也是很重要的影响因素,故空气污染指数以及能见度也应该纳入景观出现原因的考虑之中,有学者对于中国城市污染时空分布做了研究,发现成都所在的中国西南地区空气质量好转[1]。目前国内气象工作者对于雾凇以及云海等景观天气现象的预报进行了分析和探索,对于遥望雪山的形成原因以及预报研究较少。有学者利用逻辑回归和机

器学习等多种方法对云海的出现与否进行了研究和分析[2]，还有学者基于Logistic回归对城市雾天气出现与否做出了预报模型[3]。开展对遥望雪山景观的预报关键因素进行研究，对于景观形成的原因以及后续的预报方式的探索具有十分重要的科学意义和实际运用价值。

1 资料和方法

1.1研究区概况

成都市区与幺妹峰的直线距离约为120公里，与贡嘎山的直线距离约为240公里。且成都位于四川盆地西侧的成都平原，视野开阔，与雪山之间没有阻挡，与雪山之间的高度差十分巨大，市区与贡嘎山的海拔高度差约为7000m，与幺妹峰的海拔高度差约为5700m，使得成都能够在天气状况良好的情况下看到远方越过云层的雪山山顶。

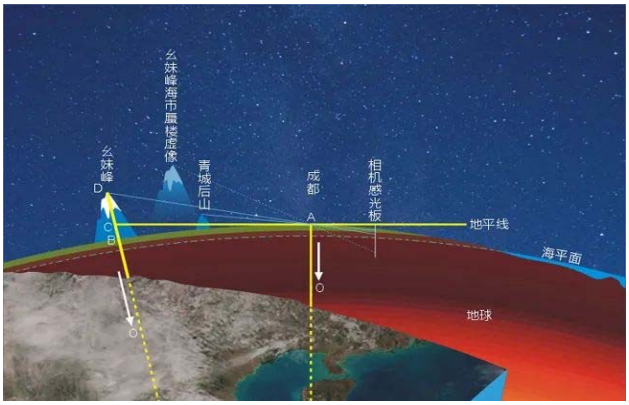


图1 地球弧度对遥望幺妹峰的影响示意

1.2资料概况

本文选取了包括“遥望雪山”景观出现的人工观测和记录数据。气象资料使用气象局历史数据，主要选取了成都市崇州、温江、都江堰、彭州等14个站点的气象资料数据，以及成都市区空气质量自动观测站的AQI和IAQI数据，资料时间年限为2017年-2021年。使用python编程和SPSS对数据进行分析和计算，用ArcGIS、Origin对数据进行可视化处理。

1.3研究方法

利用成都市气象环境观测站的湿度、降水量等气象资料，采用Spearman秩相关系数对因变量和自变量之间进行相关性检验。Logistic回归是概率型非线性回归模型，是研究分类因变量(Y)与自变量(X)之间关系的一种多变量分析方法。由于雪山景观出现的结果为二分类事件，因此使用Logistic回归[4]。设自变量在样本因变量的影响下发生概率记作，则不发生的概率为，Logistic函数如下：

$$F(p)=\ln\left(\frac{p}{1-p}\right) \tag{1}$$

因此，Logistic回归模型为：

$$F(p)=\ln\left(\frac{p}{1-p}\right)=bX \tag{2}$$

其中b为回归系数矩阵，X为因子矩阵。

从式(2)中可以看出，p的变化在(0,1)之间，$F(p)$在$(-\infty,+\infty)$之间，自变量$x_1,x_2\cdots\cdots x_m$ 取值可在任意范围。根据式(2)可以得到预报方程：

$$P=\frac{e^a}{1+e^a} \tag{3}$$

其中

$$a=bX$$

将实际个例和预测结果转化为二分类事件，并对预测结果进行TS评分[8]。

TS评分：

$$TS=\frac{N_A}{N_A+N_B+N_C}\times 100\% \tag{4}$$

漏报率：

$$PO=\frac{N_C}{N_A+N_C}\times 100\% \tag{5}$$

空报率：

$$FAR=\frac{N_B}{N_A+N_B}\times 100\% \tag{6}$$

总体准确率：

$$AR=\frac{N_A+N_D}{N_A+N_B+N_C+N_D}\times 100\% \tag{7}$$

N_A为遥望雪山预报正确次数；N_B为空报次数；N_C为漏报次数；N_D为无遥望雪山预报正确次数(表1)。

表1 遥望雪山景观检验分类表

类别	预报有	预报无
实况有	N_A	N_C
实况无	N_B	N_D

2 遥望雪山景观预报模型

2.1预报因子分区选取

由于成都地区能见度、云量、降水等气象要素的地区差异较大，在分析当中将整个成都地区划分为四个地区，即中部地区(双流、温江、郫县、新都和龙泉驿)，西南地区(崇州、大邑、邛崃、蒲江以及新津)，东部地区(简阳、龙泉驿和金堂)以及西北区域(都江堰、彭州和郫县)，并分地区对数据进行标准化处理，且由于成都所在的盆地西部云量一直处于较高水平，平均值达到了8到9成，故在各气象因子之外再加上一个云量是否大于等于平均值的二分类变量，最后利用Spearman秩相关系数对因变量和自变量进行相关性检验。

其中X_1为平均气压，X_2为平均气温，X_3为最高气温，X_4为最低气温，X_5为平均湿度，X_6为总云量，X_7为该地区总降水量，X_8为前一天该地区总降水量，X_9为平均风速，X_{10}为能见度，X_{11}为云量是否大于等于平均值(二分类变量，1代表大于等于平均值，0代

表小于平均值)，X_{12}为AQI。

将上述12个预报因子与观山是否出现(因变量,1代表景观出现,0代表景观未出现)进行相关性分析(表2),可以看出在中部地区平均气压、最低气温、总云量与云量是否大于等于平均值未通过显著性检验,平均气温与前一天该地区总降水量通过了显著性水平为0.05的检验,其他因子通过了显著性水平为0.01的检验。西南地区平均气压、最低气温、总云量与云量是否大于等于平均值未通过显著性检验,平均气温、前一天该地区总降水量与平均风速通过了显著性水平为0.05的检验,其他因子通过了显著性水平为0.01的检验。东部地区平均气压、最低气温、总云量与云量是否大于等于平均值未通过显著性检验,平均气温与平均风速通过了显著性水平为0.05的检验,其他因子通过了显著性水平为0.01的检验。北部地区平均气压、最低气温、前一天该地区总降水量与云量是否大于等于平均值未通过显著性检验,平均气温、总云量与平均风速通过了显著性水平为0.05的检验,其他因子通过了显著性水平为0.01的检验。

表2 2020-2021年成都遥望雪山预报因子与预报对象的相关性分析

因子	中部地区		西南地区		东部地区		西北地区	
	R	sig	R	sig	R	sig	R	sig
X_1	-0.042	0.37	-0.046	0.324	-0.052	0.268	-0.051	0.272
X_2	0.108*	0.02	0.104*	0.025	0.107*	0.022	0.106*	0.022
X_3	0.159**	0.001	0.143**	0.002	0.150**	0.001	0.160**	0.001
X_4	0.023	0.625	0.024	0.606	0.035	0.454	0.021	0.657
X_5	-0.177**	0	-0.206**	0	-0.139**	0.003	-0.152**	0.001
X_6	-0.056	0.265	-0.022	0.688	0.011	0.844	-0.109*	0.031
X_7	-0.210**	0	-0.224**	0	-0.210**	0	-0.202**	0
X_8	0.108*	0.02	0.100*	0.031	0.123**	0.008	0.089	0.056
X_9	0.131**	0.005	0.094*	0.042	0.102*	0.028	0.098*	0.035
X_{10}	0.183**	0	0.202**	0	0.156**	0.001	0.125**	0.007
X_{11}	-0.038	0.456	0.051	0.339	0.053	0.318	-0.076	0.132
X_{12}	-0.024	0.599						

注：*表示在0.05水平(双侧)上显著相关；**表示在0.01水平(双侧)上显著相关

2.2 Logistic回归模型的建立及检验

将13个因子以及自变量导入SPSS中进行计算,选择Enter法进入Logistic模型,表3和表4中的列是回归方程当中的自变量系数；S.E.为Standard Error标准误差[5]；Wals是检验自变量对因变量的影响程度,数值越大,自变量对因变量的影响就越显著。将自变量系数代入到式(5)便能得到不同地区的回归方程,可以发现回归方程的值的取值范围在0到1之间。

表3 中部和西南地区Logistic回归方程中的变量

统计量	中部地区			西南地区		
	b	S.E.	Wals	b	S.E.	Wals
X_1	0.071	0.039	3.336	0.075	0.041	3.276
X_2	0.35	0.288	1.477	0.688	0.335	4.208
X_3	0.248	0.135	3.345	-0.047	0.147	0.1
X_4	-0.415	0.182	5.208	-0.513	0.209	6.051
X_5	-0.025	0.029	0.789	0.037	0.032	1.333
X_6	0.008	0.008	1.009	0.005	0.008	0.33
X_7	-0.009	0.004	5.537	-0.009	0.003	7.68
X_8	0.01	0.003	9.558	0.006	0.003	5.477
X_9	0.717	0.451	2.533	1.27	0.42	9.149
X_{10}	0	0	7.237	0	0	3.863
X_{11}	-0.031	0.008	17.082	-0.356	0.407	0.764
X_{12}	0.049	0.41	0.014			
常量	-73.376	38.363	3.658	-81.412	40.475	4.046

表4 东部和西北地区Logistic回归方程中的变量

统计量	东部地区			西北地区		
	b	S.E.	Wals	b	S.E.	Wals
X_1	0.097	0.04	6.03	0.097	0.04	6.03
X_2	-0.078	0.258	0.091	-0.078	0.258	0.091
X_3	0.338	0.126	7.201	0.338	0.126	7.201
X_4	-0.151	0.158	0.912	-0.151	0.158	0.912
X_5	0.037	0.027	1.935	0.037	0.027	1.935
X_6	0.011	0.007	2.231	0.011	0.007	2.231
X_7	-0.011	0.005	4.877	-0.011	0.005	4.877
X_8	0.015	0.004	11.416	0.015	0.004	11.416
X_9	0.49	0.503	0.949	0.49	0.503	0.949
X_{10}	0	0	8.668	0	0	8.668
X_{11}	0.4	0.39	1.05	0.4	0.39	1.05
常量	-102.113	38.901	6.89	-102.113	38.901	6.89

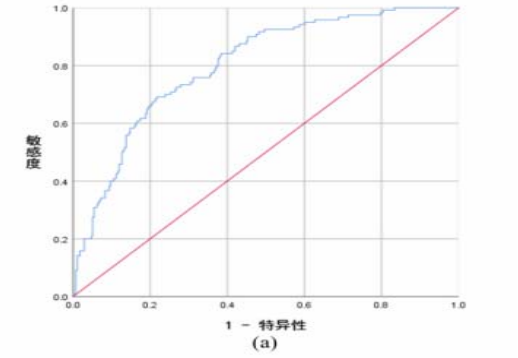
Logistic回归模型的拟合优度检验主要有以下三种指标,-2对数释然值越大,以及Cox-Snell R方与Nagelkerke R方理论值越接近1,拟合效果越好[6],但通常在0.3至0.5之间。根据表5可以看出各个区域的模型拟合效果较为一般。

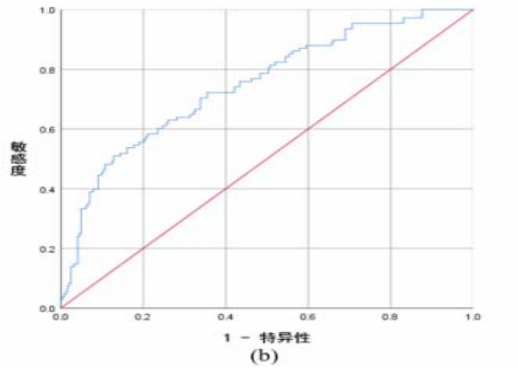
表5 各地区模型的综合检验

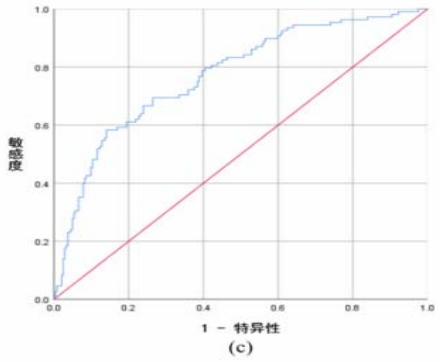
模型	-2对数似然	Cox-Snell R 方	Nagelkerke R 方
中部地区	387.177	0.22	0.312
西南地区	370.348	0.163	0.23
东部地区	418.701	0.152	0.215
西北地区	365.655	0.174	0.245

2.3 Logistic回归模型临界值的选定

根据Logistic回归预报方程,计算得到的预测值是一个在0和1之间的值,因此需要找到适当的临界值用作判断观山景观是否发生的标准,故可以利用到ROC曲线以及约登指数[7],如果ROC曲线下的所占的面积越大,则诊断准确性越高。可以看出中部区域ROC曲线下面积为0.801,西南地区面积为0.748,东部区域面积为0.771,西北区域面积为0.736,四个区域Logistic回归预报方程的诊断实验准确度都较为理想。当约登指数最大时对应的临界值距离对角参考线,此时的灵敏度以及特异度都处于比较高的值,因此可以利用ROC曲线来寻找最优的临界值[8]。根据计算,可以看出中部地区Logistic回归的约登指数取临界值为0.35；西南地区取临界值为0.29；东部地区取临界值为0.32；西北地区取临界值为0.35。







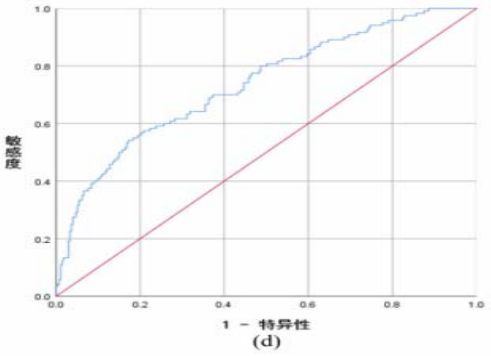


图2 不同区域Logistic模型的ROC曲线

2.4模型的应用检验与结果对比

将上面得到的临界值作为切割值重新引入各个地区的Logistic回归,得出训练集检验结论。中部地区：当观测为景观不出现,预测正确率为78.3%；当观测为景观出现,预测正确率为69.2%；总体正确率为75.5%。西南地区：当观测为景观不出现,预测正确率为64.0%；当观测为景观出现,预测正确率为72.2%；总体正确率为66.6%。东部地区：当观测为景观不出现,预测正确率为72.7%；当观测为景观出现,预测正确率为69.4%；总体正确率为71.7%。西北地区：当观测为景观不出现,预测正确率为76.2%；当观测为景观出现,预测正确率为58.3%；总体正确率为70.7%。

将各地区的Logistic回归所得结果评分汇总,可以得到成都不同地区回归不同的评分。其中PO为漏报率,FAR为空报率,AR为总体准确率,可以发现TS评分在中部地区和东部地区较好,中部地区的总体准确率最高。

表6 不同地区Logistic回归模型预报评分对比

地区	TS评分/%	PO/%	FAR/%	AR/%
中部地区	46.1	30.8	42.0	75.5
西南地区	40.0	27.8	52.7	66.6
东部地区	43.1	30.6	46.8	71.7
西北地区	37.8	41.7	48.1	70.7

文章类型：论文|刊号（ISSN）：2630-4740 /（中图刊号）：715GL012

3 结论

本文利用了成都地区站点的气象信息对遥望雪山景观的预报关键性因素做了分析，对景观出现与否这样的二分类事件进行了分区域的Logistic回归，并对预报模型进行了研究，得出以下结论：

(1)在对各气象因素和景观是否出现的相关性分析当中，可以发现最高气温、平均湿度、降水量、平均风速和能见度通过了显著性水平为0.01的双侧显著性检验，平均气温以及前一天降水量通过了显著性水平为0.05的双侧显著性检验，平均气压、最低气温、总云量、云量是否大于等于平均云量以及AQI与景观是否出现相关性不显著。

(2)将成都市分为四个地区，分别对各地区根据Logistic回归建立预报方程，再利用ROC曲线和约登指数，筛选各地区预报模型的最优临界值，最后将最优临界值回代入预报方程中进行测试检验，结果表明成都中部地区和东部地区的拟合度较好，其中以中部地区最佳。

[参考文献]

[1]危诗敏,冯鑫媛,王式功.四川盆地多层逆温特征及其对大气污染的影响[J].中国环境科学,2021,41(3):1005-1013.

[2]丁国香,刘安平,杨彬.黄山冬半年云海预报研究[J].气象与环境学报,2019,35(2):97-101.

[3]陈荣泉,彭端,赖燕冰.基于Logistic回归的肇庆市区雾天气的预报模型[J].广东气象,2019,41(2):19-23.

[4]吴安坤,郭军成,黄天福.Logistic回归联合ROC曲线模型在雷电潜势预报中的应用[J].中低纬山地气象,2022(002):046.

[5]庞古乾,伍志方,郭春迓.广东省前汛期分区强对流潜势预报方法研究[J].热带气象学报,2016,32(2):265-272.

[6]邓飞,张荣稳,余靖,等.基于ROC曲线的逻辑回归切割值寻优方法研究及应用[J].昆明理工大学学报:自然科学版,2023,48(1):5.

[7]张锴,黄京平,吴文心.三清山雾凇景观资源分析与雾凇旅游指数预报[J].决策探索,2020(2):93-94.

[8]王彦兵,王聪,赵亚丽,等.基于ROC曲线的永久散射体识别最佳阈值定量筛选[J].遥感学报,2021,025(010):2083-2094.

作者简介：

文雯(1994--),女,汉族,四川省成都市人,硕士,工程师,主要从事综合气象服务、污染气象及生态气象。

***通讯作者：**

刘自牧(1991--),男,汉族,四川省成都市人,硕士,工程师,主要从事专业气象服务、高原气象研究。