基于加权Logistic回归模型的森林火灾预测*

文 斌[1] 谢献强[2] 孙 萌[3] 杜治国[1]
李 溯[2] 黄 平[2] 朱宇浩[2] 谢柏联[2]
（1. 华南农业大学 数学与信息学院，广东 广州 510642；2. 广东省航空护林站 / 广东省林火卫星监测中心，广东 广州 510173；3. 中国南方电网有限公司，广东 广州 510663）

摘要 林火预警是保障超高压输电网安全的重要工作，是森林防火部门和电网公司深度关注的领域。研究以 2007—2017 年广东省超高压输电网线路途经地区的气象数据和林火监测的数据为基础，通过加权 Logistic 回归分析方法构建了广东省超高压输电线路区域森林火险预警模型，并用 2017 年实际林火发生数据对该模型进行检验。模型预测准确率达到 92.6%，证明该模型具有良好的预测效果，反映了广东省区域森林火险等级与相关气象因子的密切关系。

关键词 森林火灾；动态加权；预测
中图分类号 :S762 **文献标志码**：A **文章编号**：2096-2053（2019）04-0079-05

Forest Fire Prediction Based on Weighted Logistic Regression Model

WEN Bin[1] XIE Xianqiang[2] SUN Meng[3] DU Zhiguo[1]
LI Su[2] HUANG Ping[2] ZHU Yuhao[2] XIE Bailian[2]
(1. School of Mathematics and Information, South China Agricultural University, Guangzhou, Guangdong 510642, China; 2. Guangdong Aerial Forest Fire Protection Station/Guangdong Forest Fire Satellite Monitoring Center, Guangzhou, Guangdong 510173, China; 3. CSG EHV Maintenance，Guangzhou，Guangdong 510663，China）

Abstract Based on the meteorological data and forest fire records from 2007 to 2017, the dynamic weight of forest fire was obtained by cluster analysis, and the weighted Logistic regression model was established. Finally, the nonlinear regression equation between forest fire risk, meteorological factors and forest fire historical distribution factors in Guangdong province was obtained. The data of forest fire in 2017 were predicted by the equation. The accuracy rate is 92.6%. The availability of the model was verified. Reflecting the close relationship between the regional forest fire risk rating and related meteorological factors in Guangdong province. It can provide reference for the forest fire warning work of the ultra-high voltage transmission network in Guangdong province.

Key words forest fire; dynamic weighted; forecast

穿越林区的超高压输电线路，是林区的重要基础设施之一。近年来，山火引起的林区架空输电线路闪络跳闸或停运故障频频发生，给人民群众的生产、生活带来的严重影响。如果可以及时地预测超高压输电线路周边森林火灾，对森林防火和降低林火灾害损失都有重要意义，因此林火

* **基金项目**：广东省林业科技创新项目 (2017KJCX046) ; 华南农业大学校级教学改革项目（JG17088）。
第一作者：文斌（1981 — ），男，实验师，主要从事数据分析教学研究，E-mail:wenzip@scau.edu.cn。
通信作者：谢献强（1980 — ），男，教授级高级工程师，主要从事森林防火工作，E-mail: xqxie@126.com。

预警是森林防火部门和超高压电网公司共同关注的领域。

林火预测是通过对气象数据、时空数据、人员活动等数据进行综合分析，获得可测量数据与林火发生风险之间的相关性的过程[1-3]。通过分析林火发生的历史数据，建立准确有效的森林火灾预警模型，对提高林火管理水平至关重要[4-5]。由于气象数据测量的准确性与林火的密切相关性，众多学者基于气象数据建立数学模型用于森林火灾预测[6-7]，其中Logistic回归模型应用得比较广泛并取得一些成果，如曾钦文等[8]通过Logistic回归分析方法建立了龙川县森林火灾预警系统，对模型检验表明林火高风险预警准确率有62.9%，但是检验的数据量比较少，存在稳定性问题。王亚琴等[9]提出一种改进的直连神经网络模型用于林火预测，提高了两个不同地区（火灾多和火灾少）的林火预测精度，该模型也是先通过回归模型来分析气象因子与火灾结果的相关性，不过该方法采用神经元的选定会直接影响预测结果。

针对上述问题，本次研究通过分析其他行业对回归模型应用的成果[10-11]，发现预测模型需要具有灵活的非线性拟合能力，比如影响林火发生的因素是多维的，包括不同地区地形、植被、林区人员活动情况等都会影响预测的准确度。本研究通过对广东省南方电网超高压输电线路途经位置的2007-2017林火数据和气象数据进行数据分析，构造出反映林火高发的季节因素和人为活动因素的权重数据，并以此建立了基于林火活动性的Logistic回归模型（Fire-Dynamic Weighted Logistic Regression，以下简称FDWLR模型）。通过对已有数据的验证，该模型可以很好地预测每日林火风险，对南方电网超高压输电公司在粤输电线路的林火预警工作有很好的参考作用。

1 材料与方法

1.1 数据来源

气象数据来源于中国气象大数据共享平台，包括2007-2017年广东省内24个国家级气象观测站的日气象数据。气象站的选取是依照超高压输电网线路的地理坐标就近选取，原始数据有18个气象因子。参考相关研究结论[7-9]，我们选取与森林火险密切相关的8个气象因子进行分析，包括日平均气温AvgTEM（℃），日最高气温MaxTEM（℃），日平均风速AvgWin（m/s），日最大风速MaxWind（m/s），24 h降水量PRE（mm），日相对湿度RHU（%），日蒸发量EVP（mm），连续无降水天数（连续降水量低于3 mm的天数）。2007-2017年广东省24个县（市、区）超高压输电线路沿线位置发生的林火数据来源于广东省航空护林站（广东省林火卫星监测中心）。

1.2 数据筛选

广东省2007-2017年的24个国家级气象监测

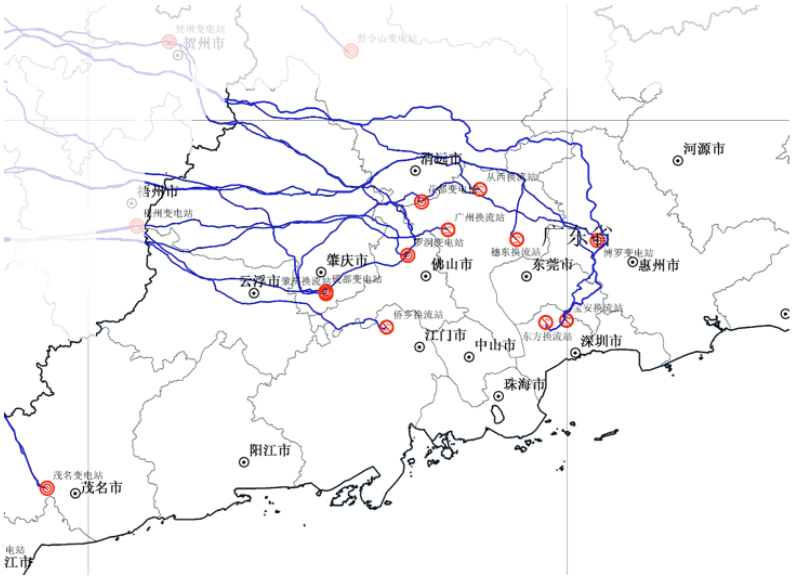

图1 广东省超高压输电线路地理位置

Fig. 1 Geographical location of UHV transmission lines in Guangdong Province

表 1 筛选的气象数据统计

Tab.1 Statistical table of filtered regional meteorological data

气象因子 Meteorological factor	极小值 Min	极大值 Max	均值 Mean	标准差 Std	偏度 Skewness	峰度 Kurtosis
最大风力 /（m · s^{-1}）MaxWind	0.60	24.50	4.89	1.96	1.76	6.88
24 h 降水量 /mmPRE	0.00	2 812.00	23.26	88.93	7.50	86.33
湿度 /%RHU	8.10	100.00	73.17	13.60	–0.54	0.22
最高气温 /℃ MaxTEM	8.60	36.45	22.78	5.61	–0.51	–0.17
连续无降水天数 Continuous precipitation-free day	0	58	6.51	8.19	1.93	4.32

站的日气象数据有 56 702 条，数据分析前剔除部分缺失值。表 1 列出了预处理后主要气象数据的基本统计描述信息，确认各项气象数值都在正常范围内。从中可以看出广东省超高压输电线路途径位置降雨量的波动范围很大，而且分布不平均。

林火记录数据项有：（1）发生林火的地理坐标和精确到村级的地址；（2）林火发生时间；（3）林火过火面积等详细信息。经过数据筛选和数据标准化处理，最后得到 11 928 条林火记录，时间跨度为每年 1-4 月、10-12 月共 7 个月。本文模型中使用了林火发生时间和林火规模两项数据。从图 1 可以看出 12，1，2 月合计林火数目约占全年 80%。

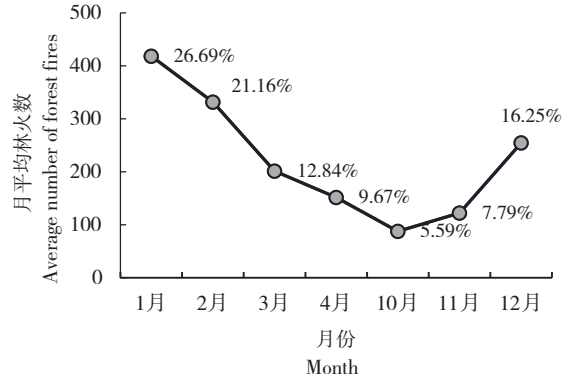

图 2 2007–2016 年各月份林火平均数及占全年总数的百分比

Fig.2 The average number of forest fires from 2007 to 2016

1.3 研究方法

1.3.1 Logistic 回归模型　Logistic 回归（LR）模型常用来解决因变量为分类变量的回归分析问题，如火灾的发生、不发生，风险等级的高、低等，可以很好地预测非线性回归事件发生的可能性 [10-11]。本文研究目的是通过可以被测量和记录的数据，预测未来一定时期内发生林火的风险。如果用 1 表示有林火发生，0 表示没有林火发生，则问题适合用因变量二元 logistic 回归分析，解释变量可以从多个影响林火发生的因素中选出，最后得到回归方程可以用于预测。

设用 y 表示林火风险预测值，则建立 Logistic 回归模型公式 [13]：

$$y = \ln\frac{P}{1-P} = a + \beta x + \varepsilon$$

公式是 n 个解释变量 x 与因变量 y 之间的关系表达式，P 为二元因变量发生的概率，a 为模型截距常数项，ε 表示误差。应用 LR 模型分析本次研究的数据时，以选取的 5 个气象因子作为解释变量，预测准确度最低为 22.6%。从林火数据与气象数据散点图分布情况可以看出，许多日期具有近似的气象因子数据，但是林火发生存在较大差异。

1.3.2 基于林火活动性加权 LR 模型　为了识别林火发生受气象因素之外的其他因素（如林火季节因素，历史同期林火频率等）影响的水平，采用基于信息熵理论设置林火动态权重，当某个日期最邻近区域的林火发生频次和气象因子相关度都高时，代表该日期林火活动性强，对应权重越高。第 n 天权重 W_n 的计算公式如下 [14]

$$W_n = 1 - \frac{\sum_{i=1}^{k} \frac{D_i}{T} \ln_2\left(\frac{D_i}{T}\right)}{\ln_2 k}$$

公式中，k 表示样本数据日期分类区间内天数，D_i 表示区间内第 i 日林火数，T 表示样本数据中第 n 天林火累计数，区间是通过对历史林火记录数据和 5 个气象因子进行聚类分析得到。W_n 值越大，则表示同等气候条件下该日期区间内林火活动性越高。由此构造 FDWLR 模型回归方程为：

$$y_n = \alpha + \beta_{w_n} X_n + \varepsilon$$

其中第 n 天的回归模型参数系数由加权最小二乘法得到。该模型首先考察5个气象因子和林火是否发生之间的相关性，选取相关性最显著的因子作为FDWLR方程的解释变量。

2 结果分析与验证

使用SPSS软件和Python数据分析软件包计算模型得到以下结果，从表2可知，连续无降水日、相对湿度、最大风力、最高气温、蒸发量这5个气象因子与林火发生在 P=0.01水平上有显著相关性。然后把权重加入方程求出各项回归系数。通过霍斯默 - 莱梅肖检验可以判断回归方程拟合度良好（显著性0.328>0.05）。

从表3可以得到FDWLR模型回归公式为：y =（1.888 +0.356× 连续无降水天数 –0.106 × 相对湿度 +0.128× 最高气温 – 0.057× 蒸发量 – 0.06× 最大风力)× FD 权重 W。其中 y 值越大，表示林火风险越高。

表4列出了FDWLR模型预测准确度，其中对有林火发生情况的预测准确度提升了50.1个百分点（对比上面列举的LR回归分析模型），而这也是反映林火预警模型实用性的最重要指标。

用2017年林火数据作为检验样本，把2017年对应观测点的气象因子和已经取得的2007-2016年林火动态权重值代入上面的FDWLR模型进行检验，对预测为林火发生的情况进行标记，得到林火预测散点图。图3显示了预测的每日林火数（灰色虚线）与2017年实际林火数据（黑色实线）有很好的拟合度。

3 结论

利用加权Logistic回归分析方法，对南方电网超高压输电公司在粤线路沿线2007年至2016年10年间的林火数据进行分析，结合每日林火动态权重和5个气象因子建立了林火预警FDWLR模型，用FDWLR模型预测2017年超高压输电线路沿线防火期内每日林火是否发生，对有林火发生的情况预测准确度有72.7%。另外通过对模型结果进一步分析也可以得到以下结论：连续无降水天数和日最高气温对火险发生影响较大；有火险发生的日期里，日最大风速对火险规模（火险持续时间）影响较大；火险等级随日降水量增加而迅速降低。因此高压输电线路林火防控预警系统需考虑增加线路沿途气象信息监测点并与之动态关联以提高性能。

表2 气象因子相关性

Tab. 2 Correlation of meteorological factors

相关事件 Related Events	连续无降水天数 Number of days without precipitation	最大风力 MaxWin	最高气温 MaxTEM	蒸发量 EVP	湿度 RHU
林火发生 Forest fire	0.273*	–0.102*	0.088*	0.047*	–0.356*

注：* 在0.01水平上显著相关 Note: Significant correlation at 0.01 level

表3 FDWLR回归方程的回归系数

Tab.3 Regression coefficient in FDWLR regression equation

气象因子 Meteorological factors	回归系数 B	标准误差 S.E.	瓦尔德 Wald	自由度 df	显著性 Sig.	OR值 Exp (B)
连续无降水天数 Number of days without precipitation	0.356	0.002	411.958	1	显著	1.039
相对湿度 RHU	–0.106	0.002	3 097.586	1	显著	0.900
最高气温 /℃ MaxTEM	0.128	0.004	941.011	1	显著	1.137
蒸发量 /mm EVP	–0.057	0.002	1 071.135	1	显著	0.944
最大风力 /（$m \cdot s^{-1}$）MaxWind	–0.060	0.012	26.560	1	显著	0.942
常量 Constant	1.888	0.149	160.133	1	显著	6.604

表 4 FDWLR 模型预测林火发生的正确率

Tab. 4 The correct rate of FDWLR model for forest fire prediction

预测类别 Forecast category	实际数量 Actual quantity	预测数量 Forecast quantity	正确百分比 /% Correct percentage
无火点 No fire spots	44 774	43 849	97.9
有火点 Fire spots exist	11 928	8 669	72.7
总体百分比 Overall percentage			92.6

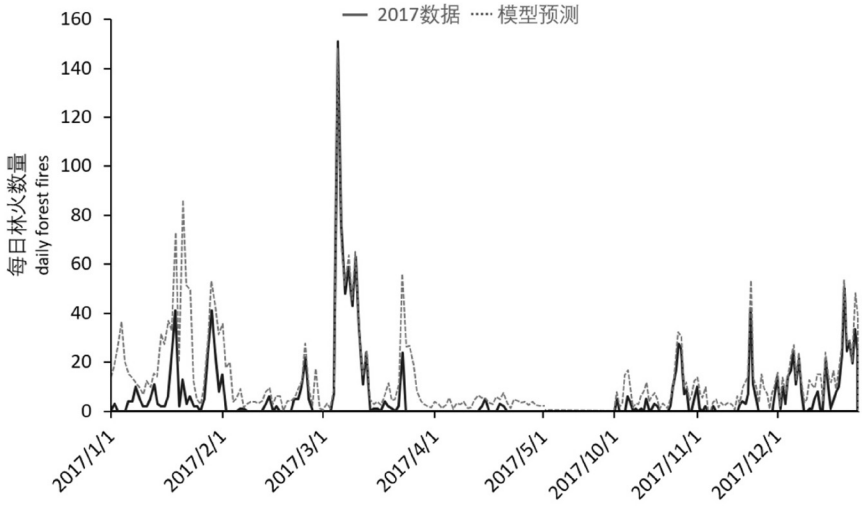

图 3 2017 年防火期内每日林火数量预测数据与实际发生数对比

Fig. 3 Comparison of forecast data and actual occurrence number of daily forest fires during the 2017 fire prevention period

参考文献

[1] 吴恒, 朱丽艳, 刘智军, 等. 中国森林火灾发生规律及预测模型研究[J].世界林业研究, 2018, 31(5): 64-70.

[2] 王振师, 周宇飞, 李小川, 等. 无人机在森林防火中的应用分析[J].林业与环境科学, 2016, 32 (1) : 31-35.

[3] 王振师, 魏书精, 苏润鸿, 等. 林火燃烧环境对灭火效果的影响研究[J].林业与环境科学, 2019, 35(2): 84-88.

[4] 吴善材, 王成．回归方程在开平市高温天气预报中的应用[J].广东气象, 2009, 21(2) : 38-39.

[5] 袁春明, 文定元．林火行为研究概况[J].世界林业研究, 2000, 13(6): 27-28.

[6] 于文颖, 周广胜, 赵先丽, 等. 大兴安岭林区火灾特征及影响因子[J].气象与环境学报, 2009, 25(4) : 1-5.

[7] 田晓瑞, 赵凤君, 舒立福, 等．西南林区卫星监测热点及森林火险天气指数分析[J].林业科学研究, 2010, 23(4) : 523 -529.

[8] 曾钦文, 曾思亮, 王辉, 等. 龙川县森林火险等级预报方法的建立及应用[J].广东气象, 2017, 39(4): 52-55.

[9] 王亚琴, 王耀力, 郭学斌, 等.基于直连BP神经网络模型的森林火险预测[J].森林防火, 2018, 137(2): 46-50.

[10] 段永春．茶树干燥型大冻害气象因素的LOGISTIC回归分析[J]．中国农学通报, 2016, 32(7): 152-156.

[11] 潘泽清.企业债务违约风险LOGISTIC回归预警模型[J].上海经济研究, 2018(8): 73-83.

[12] 李明华, 崔少萍, 罗凤明, 等. 统计软件 SPSS 在气象中的应用[J].广东气象, 2007, 29(1) : 50 -52.

[13] 陈胜可.SPSS统计分析从入门到精通[M]. 2版. 北京: 清华大学出版社, 2014: 254-262.

[14] 杨文涛, 邓敏, 王玉朝, 等. 一种基于信息熵的时空点模式分析方法[J].地理与地理信息科学, 2016, 32(5): 71-75.