

刘扬,王维国,2020. 基于随机森林的暴雨灾害人口损失预估模型及应用[J]. 气象,46(3):393-402. Liu Y, Wang W G, 2020. Assessing model of casualty loss in rainstorms based on random forest and its application[J]. Meteor Mon, 46(3):393-402 (in Chinese).

# 基于随机森林的暴雨灾害人口损失预估模型及应用<sup>\*</sup>

刘 扬 王维国

国家气象中心,北京 100081

**提 要:** 基于 2009—2017 年的广西县级暴雨灾情记录,综合考虑致灾因子、孕灾环境和承灾体因素选取 7 个解释变量,运用随机森林算法,构建暴雨灾害人口损失预估模型;并以精细化网格降水实况分析和预报产品驱动模型,预估是否发生人口损失。研究表明:模型训练样本及测试样本的分类准确率均在 90% 以上,致灾因子(降水情况)是最主要的解释变量,重要性从大到小依次是前 10 d 降水距平百分率、过程最大日雨量、最大小时雨量和短时强降水频次。应用智能网格降水产品对广西地区近两年的暴雨灾害过程进行回报试验,准确率超过 70%。

**关键词:** 暴雨,人口损失,随机森林,智能网格

**中图分类号:** P49

**文献标志码:** A

**DOI:** 10.7519/j.issn.1000-0526.2020.03.010

## Assessing Model of Casualty Loss in Rainstorms Based on Random Forest and Its Application

LIU Yang WANG Weiguo

National Meteorological Centre, Beijing 100081

**Abstract:** Based on historic casualty loss records of rainstorm that occurred at county level in Guangxi from 2009 to 2017, seven factors were selected as explanatory variables by comprehensively considering the trigger factors, disaster formative environment and exposure units, and the prediction model of casualty loss caused by rainstorms was built up by using random forest algorithms. The refined grid precipitation analysis and forecast products were used to drive the model to predict loss of life. The results showed that the classification accuracies are both above 90% in training and testing samples. Disaster-triggering factors (precipitation) are the most significant explanatory variables. The importances of these precipitation variables in turn are the anomaly percentage of accumulated precipitation over the previous 10 days, the maximum daily precipitation, the maximum hourly precipitation and the frequency of short-time severe rainfall. By applying the intelligent grid precipitation products, several rainstorm processes in Guangxi in recent two years were used to verify the model, showing that prediction accuracies are above 70%.

**Key words:** rainstorm, loss of life, random forest, intelligent grid

<sup>\*</sup> 国家气象中心青年基金(Q201814)资助

2019年4月29日收稿; 2019年10月4日收修定稿

第一作者:刘扬,从事气象灾害风险评估和决策服务技术研究. Email:liuyang@lasg.iap.ac.cn

## 引言

在暴雨防灾减灾工作中,保障人民生命安全至关重要。随着城市化进程加快,人口趋于集中、人口密度增加,人口受暴雨灾害影响的风险加剧(Hanson et al, 2011; Peduzzi et al, 2012)。

基于灾害系统理论,人口损失(包括人口死亡和失踪)是致灾因子危险性、承灾体脆弱性、孕灾环境不稳定性综合作用的结果(尹卫霞等, 2016),暴雨灾害人口损失评估研究需综合考虑“致灾因子—孕灾环境—承灾体”与人口损失的关系。研究方法主要有情景查找、经验统计(Penning-Rowsell et al, 2005)、多元回归(王志军等, 2008)和概率网络(Peng and Zhang, 2012)等。扈海波和张艳莉(2014)参照北京地区暴雨洪涝的历史灾情资料,运用模糊隶属函数建立等效日雨量与人员损失之间的损失系数估算方程,并用因子乘积方式建立人员损失风险预评估模型。王秀荣等(2016)将各类环境脆弱性影响要素结合暴雨致灾因子要素,由层次分析法得出各指标权重系数,通过加权求和建立了暴雨灾害综合风险评估模型。

“致灾因子—孕灾环境—承灾体”与人口损失之间具有复杂的非线性关系,而机器学习算法对此具有优势。机器学习算法能直接通过训练样本确定模型结构、参数,对变量间依赖关系进行拟合,适合于多变量、非线性问题(Breiman, 2001a)。利用人工智能技术(包括机器学习),发掘历史灾情数据中隐藏的规律对灾害损失进行评估,是洪涝灾害损失评估的主要方法之一(杨帅和苏筠, 2014)。随着数据挖掘技术的发展和历史暴雨灾害人口损失数据的积累,运用机器学习算法建立暴雨灾害人口损失预估模型,是本研究的主要目的。

气象灾害损失预估离不开气象要素预报。以降水预报为前端的下游气象服务,如农业气象、水文气象等对降水预报的精细化程度需求也越来越高(曹勇等, 2016)。我国无缝隙精细化网格预报业务技术的发展起步于 2014 年,现已初步建立了滚动更新、实时共享的全国 5 km 分辨率 0~30 d 无缝隙精细化网格天气预报业务(金荣花等, 2019)。但智能网格对于社会气象服务以及气象防灾减灾的实践尚偏少。进一步将网格气象预报与不同行业的数据和服务需求相融合,提高基于网格气象预报的服务能力,

也是我国无缝隙精细化网格预报发展有待加强的方向(金荣花等, 2019)。由于智能网格产品具有较高的时空分辨率,与暴雨灾害人口损失预估模型结合,可得到精细化的预估结果,更好地为防灾减灾提供依据。

本文基于历史暴雨灾害人口损失记录,计算相应的致灾因子、承灾体、孕灾环境等变量,运用机器学习算法,挖掘人口损失与解释变量之间的关系,构建暴雨灾害人口损失预估模型;并以智能网格降水实况分析和预报产品驱动模型,预估人口损失分布。

## 1 算法和数据处理

### 1.1 方法

#### 1.1.1 随机森林算法

随机森林(random forest, RF)是由 Breiman (2001b)提出的一种基于统计学习理论的机器学习算法,是在分类树算法(Breiman, 1996)基础上发展而来。使用有放回的重重复采样方法从原始数据中随机抽取训练样本,并对每个样本随机选取解释变量进行分类树构建,然后对所有分类树的预测进行组合并通过投票方式得出最终结果。RF 的基本思想是把多个弱分类器集合起来组成一个强分类器,而这些弱分类器间起到互补的作用,可以把单个分类器错误的影响缩小从而提高分类准确率和稳定性。大量理论和实例表明 RF 具有极强的数据挖掘能力和很高的预测准确率,且对异常值和噪声具有很好的容忍度,不容易出现过拟合(Iverson et al, 2008)。

RF 作为一种自然的非线性建模工具,对解决多变量的预测具有很好的效果,因而被应用到水利(赵钢铁等, 2012; 王盼等, 2014; 赖成光等, 2015)、地质(Dong et al, 2013; 李亭等, 2014; 吴小君等, 2018)、生态(Iverson et al, 2008)等众多领域,近年来在气象领域的应用也逐渐增多,例如对强对流天气进行诊断和预报(McGovern et al, 2011; Williams, 2014; 李文娟等, 2018)。

对于二分类问题,运用 RF 算法建立模型后,常使用混淆矩阵(表 1)来评估分类效果(张雪蕾等, 2018),其中:TP 是被模型预测为正的正样本数量, TN 是被模型预测为负的负样本数量, FP 是被模型预测为正的负样本数, FN 是被模型预测为负的正样本数。用分类准确率表示模型总体的分类精度,

计算公式为  $(TP+TN)/(TP+TN+FP+FN)$ ;漏报率表示实际是正类而被分类成负类的比例,计算公式为  $FN/(TP+TN+FP+FN)$ ;空报率表示实际是负类而被分类成正类的比例,计算公式为  $FP/(TP+TN+FP+FN)$ 。在本研究中,正类表示出现人口损失,负类表示没有出现人口损失。

表 1 混淆矩阵

Table 1 Confusion matrix		
	被分类为正类	被分类为负类
实际为正类	TP	FN
实际为负类	FP	TN

### 1.1.2 技术路线

本文基于历史灾情数据,运用 RF 算法构建暴雨灾害人口损失预估模型;并以精细化网格降水产品驱动模型,进行预报试验。技术流程如下(图 1): (1)清洗整理历史暴雨灾情记录,并分析暴雨灾害的时空分布特点;(2)根据灾情记录,计算解释变量,包括过程最大日雨量、最大小时雨量、短时强降水频次、前 10 d 降水距平百分率,以及灾害发生地的人口密度、地质条件、经济水平(GDP),将上述 7 个解释变量和因变量(有、无人口损失)构成样本集;(3)把样本分为训练和测试样本,基于训练样本运用 RF 算法构建模型,并计算各解释变量的重要性,用测试样本对模型进行检验;(4)基于精细化网格降水

实况和预报产品,计算与降水有关的 4 个解释变量,并将人口密度、地质条件、GDP 这 3 个解释变量统一到网格分辨率,共同驱动模型,得到预估结果。其中,根据历史灾情信息计算相应的解释变量并生成可靠的训练样本是构建模型的关键,基于精细化网格降水产品计算有关解释变量是应用模型的关键。本文以广西为例进行研究。

## 1.2 数据处理

### 1.2.1 历史暴雨灾情

原始灾情数据为 2009 年 1 月 1 日至 2017 年 12 月 31 日的广西县级暴雨灾情记录,源自中国气象局灾情直报系统,共计 1471 条,灾情记录包含暴雨过程开始和结束时间、发生地点具体到县(市、区)、死亡和失踪人口数、受灾人口数及报送时间等信息。对原始灾情数据进行清理:(1)去除明显奇异值(死亡人口超过 100 人)、重复报送及信息不完整的灾情记录;(2)去除有人口损失但受灾人口数  $\leq 3$  人的事件,认为是局地偶然事件,不具有可预测性;(3)本文通过 RF 分类算法建立模型,以有、无人口损失为预估对象,因此将历史暴雨灾情中的人口损失情况记为 1(有人口损失)或 0(无人口损失),参与建模。由于有人口损失的灾情记录远少于无人口损失的灾情记录,是一个非平衡数据集,通过在数据层进行欠抽样处理(吴琮等,2013),减少多数类样本数量,降低数据集的不平衡度。最终得到用于建模的暴雨灾害事件 530 件,其中有人口损失的 105 件,有其他灾害而没有人口损失的 425 件。

### 1.2.2 致灾因子数据

本文考虑的致灾因子为降水。构建模型的原始降水数据为 2009 年 1 月 1 日至 2017 年 12 月 31 日广西地区 2411 个区域站逐小时降水数据和 92 个国家站逐日降水数据。根据每条灾情记录的起止时间和地点信息,使用区域站逐小时降水数据,计算县域范围内的过程最大日雨量(08—08 时)、最大小时雨量、短时强降水频次(1 h 降水量超过 30 mm 的小时数),使用国家站逐日降水数据计算过程开始前 10 d 的累计降水量距平百分率。验证和应用模型的降水数据为全国智能网格实况分析产品和预报产品,空间分辨率为  $0.05^\circ \times 0.05^\circ$ ,时间分辨率为 1 h。

### 1.2.3 承灾体数据

基于中国公里网格 2010 年人口分布数据集(付晶莹等,2014),利用地理信息系统的空间分析功能,

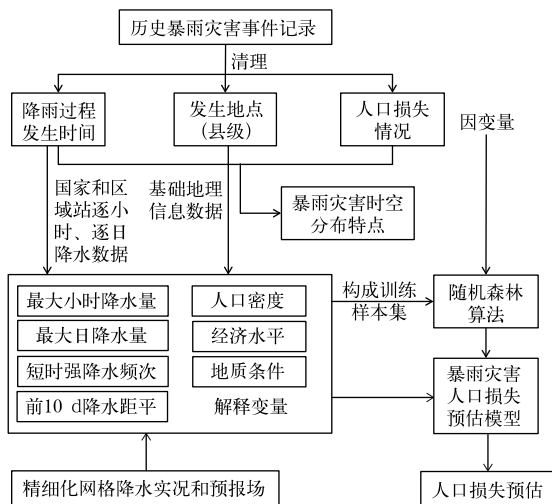


图 1 暴雨灾害人口损失预估模型  
建立和应用技术流程

Fig. 1 Technical process of construction and  
application of assessing model of  
casualty loss in rainstorms

叠加广西县级行政边界,得到各县(市、区)平均人口密度数据(图2a)。

#### 1.2.4 孕灾环境数据

基于中国公里网格2010年GDP分布数据集(黄耀欢等,2014)和1:400万全国地质灾害易发程度区划(自中国地质调查局发展研究中心网站\*),利用地理信息系统,得到各县(市、区)总GDP(图2b)和平均地质灾害易发程度(图2c)。

## 2 模型建立与验证

### 2.1 广西暴雨灾害的时空分布特点

首先,根据2009—2017年暴雨灾情记录,统计分析广西暴雨灾情概况,验证灾情信息可靠性。从空间分布来看(图3),暴雨灾害发生次数较多的地区位于南宁市马山、上林、横县及河池市巴马、天峨县,累计超过30次,平均每年发生3~6次;暴雨造成的人口损失大值区主要位于广西东南部,其中2010年6月1—2日广西玉林市容县发生暴雨洪涝灾害,造成41人死亡,是2009—2017年广西单县人口损失最多的暴雨灾害事件。

从年际变化来看(图4a),2009—2017年广西地区暴雨灾害事件数略有增加趋势,人口损失数基本保持平稳;平均每年发生暴雨灾害133次,造成人口损失24人。2017年广西共发生暴雨灾害210次,造成人口损失48人。

从多年平均的逐月暴雨灾害和人口损失数来看(图4b),大值集中在5—7月,与当地汛期一致,其中发生暴雨灾害和人口损失的峰值均在6月。平均而言,每年6月广西发生暴雨灾害50件,造成10人死亡或失踪。

### 2.2 RF模型建立

本研究基于Python语言机器学习程序包进行建模。以选取的7个预报因子作为解释变量(自变量),2个分类作为目标变量(因变量),建立RF分类模型。具体实现步骤如下:

(1)训练数据。采用五折交叉验证方法(杨柳和王钰,2015;赖成光等,2015),把所有样本 $D$ 随机分成相等容量的5份子集,取其中1份 $D_i$ 作为测试数

据集 $Test_i$ ,剩下的4份作为训练数据集 $Train_i$ ,从而构成第 $i$ 组训练—测试集( $Train_i, Test_i$ ), $i=1, 2, 3, 4, 5$ 。在每组训练—测试集中,用训练集生成模型,并在测试集中进行预测检验。

(2)参数设置。RF算法中有两个重要参数影响预测能力,即树的个数 $m$ 和预选变量个数 $k$ 。 $m$ 决定了随机森林的整体大小, $k$ 决定了单个决策树的情况。 $m$ 越大,RF算法过拟合效应越小; $k$ 越大,子预报模型间差异性越小。根据测试集的分类准确率,对RF算法的参数不断调优,找出准确率最高的对应参数。

(3)精度评判。用5组训练—测试集进行训练和测试,联合这5组结果的平均进行精度评判。若满足训练与测试精度要求即可进行下一步计算,若不满足则需要修改参数或重新取样。

经测算得到 $m$ 取值为50, $k$ 取值为 $\sqrt{7}$ 时模型精度最高,建立的RF模型精度如表2所示,该模型对训练集拟合很好(准确率为98.8%),且测试集的准确率也达到92%。

(4)指标重要性计算。RF算法中用精度下降率来表征各解释变量的重要性,即去除某个解释变量后分类准确率的降低程度。精度下降率越大,说明该变量越重要。图5为7个解释变量的重要性排序。致灾因子(降水情况)是最主要的解释变量,产生的精度下降率均超过10%,重要性从大到小依次为前10d降水距平、最大日雨量、最大小时雨量、短时强降水频次。其次是人口密度、GDP和地质灾害易发程度。致灾因子的重要度累加达85.6%,对人口损失起最关键作用。

### 2.3 基于智能网格实况的模型验证

使用全国智能网格实况分析产品,对2018年以来广西地区多个县(市、区)出现人口损失的较大范围强降雨过程,进行回报试验。

2018年6月20—25日的一次强降水过程造成都安、武鸣、钦州、灵山、凌云5个县(市、区)出现人口死亡(图6)。通过计算这次过程的4个致灾因子变量,即前10d(6月10—19日)累计降水距平百分率(图6a)、6月20日08时至26日08时过程期间最大日雨量、最大小时雨量、短时强降水频次(图6b~6d),结合相应网格化人口密度、GDP、地质灾害

\* [http://www.cigem.cgs.gov.cn/sghdzc/dzzh\\_4869/dzzhdc\\_4871/201510/t20151021\\_406406.html](http://www.cigem.cgs.gov.cn/sghdzc/dzzh_4869/dzzhdc_4871/201510/t20151021_406406.html)

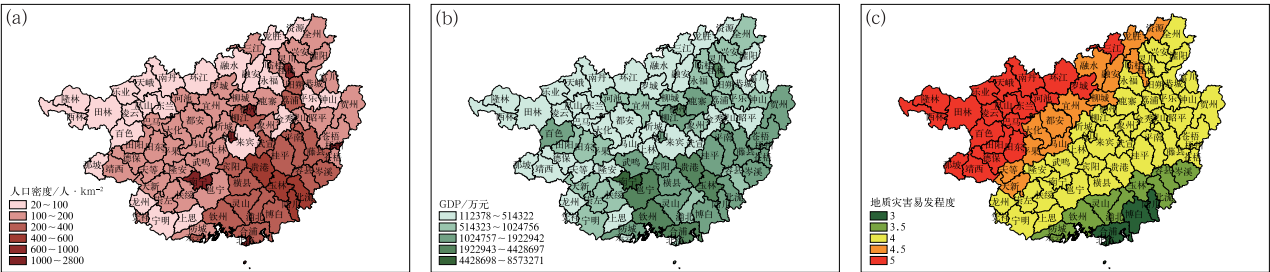


图 2 广西县级(a)平均人口密度,(b)总 GDP,(c)平均地质灾害易发程度  
Fig. 2 Averaged population density (a), total GDP (b), and averaged vulnerability of geological disaster (c) in each county of Guangxi

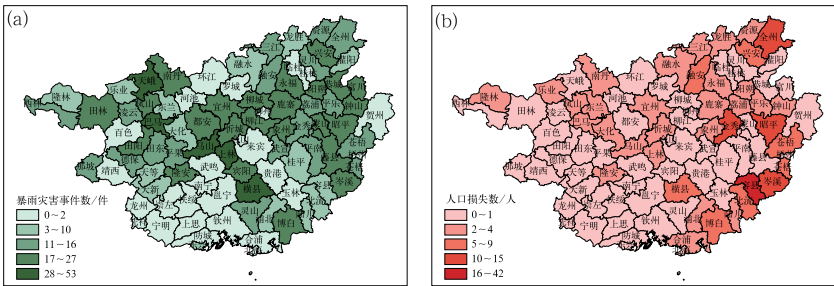


图 3 2009—2017 年广西累计(a)暴雨灾害事件数,(b)人口损失数  
Fig. 3 Distribution of accumulated rainstorm disasters in Guangxi during 2009—2017  
(a) number of rainstorm disasters, (b) number of life losses

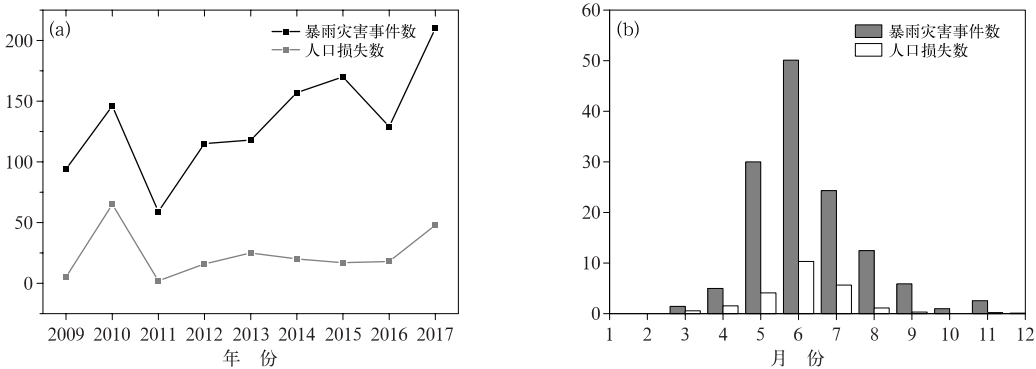


图 4 2009—2017 年广西地区(a)逐年和(b)多年平均逐月暴雨灾害事件数和人口损失数  
Fig. 4 Number of (a) yearly and (b) multi-annual mean monthly rainstorm disasters and life losses during 2009—2017 in Guangxi

表 2 RF 模型精度(单位:%)			
Table 2 Precision of RF model (unit: %)			
	准确率	漏报率	空报率
训练集	98.8	1.2	0
测试集	92	7	1

易发程度,使用之前训练得到的 RF 模型预测是否会发生人口损失,结果如图 6e 所示,红色打点区表示预测有人口损失发生。可以看出,由于前期广西地区大部分地区降水为负距平,人口损失分布主要受强降雨落区(图 6b)影响,在县域精细化结构上则

与最大小时雨量、短时强降水频次(图 6c,6d)有关。模型预测结果(图 6e)与实际发生人口损失的地区(图 6f)有较好的对应关系。基于县域统计,模型分类准确率为 89.6%,空报率为 8.3%,漏报率为 2.1%。

同样对 2018 年 5 月 6—10 日的降水过程(图 7)进行回报试验。这次过程造成博白、浦北两县出现人口死亡(图 7f)。将前 10 d(4 月 26 日至 5 月 5 日)累计降水距平百分率及过程最大日雨量、最大小时雨量、短时强降水频次(图 7a~7d),输入 RF 模型进行预测,结果如图 7e 所示。可以看出,预估有人口损失

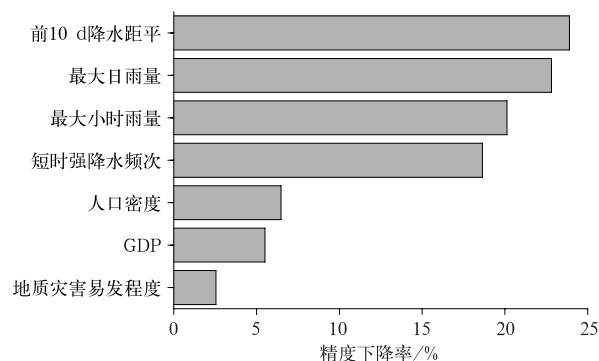


图 5 解释变量重要性排序

Fig. 5 The importance ranking of explanatory variables

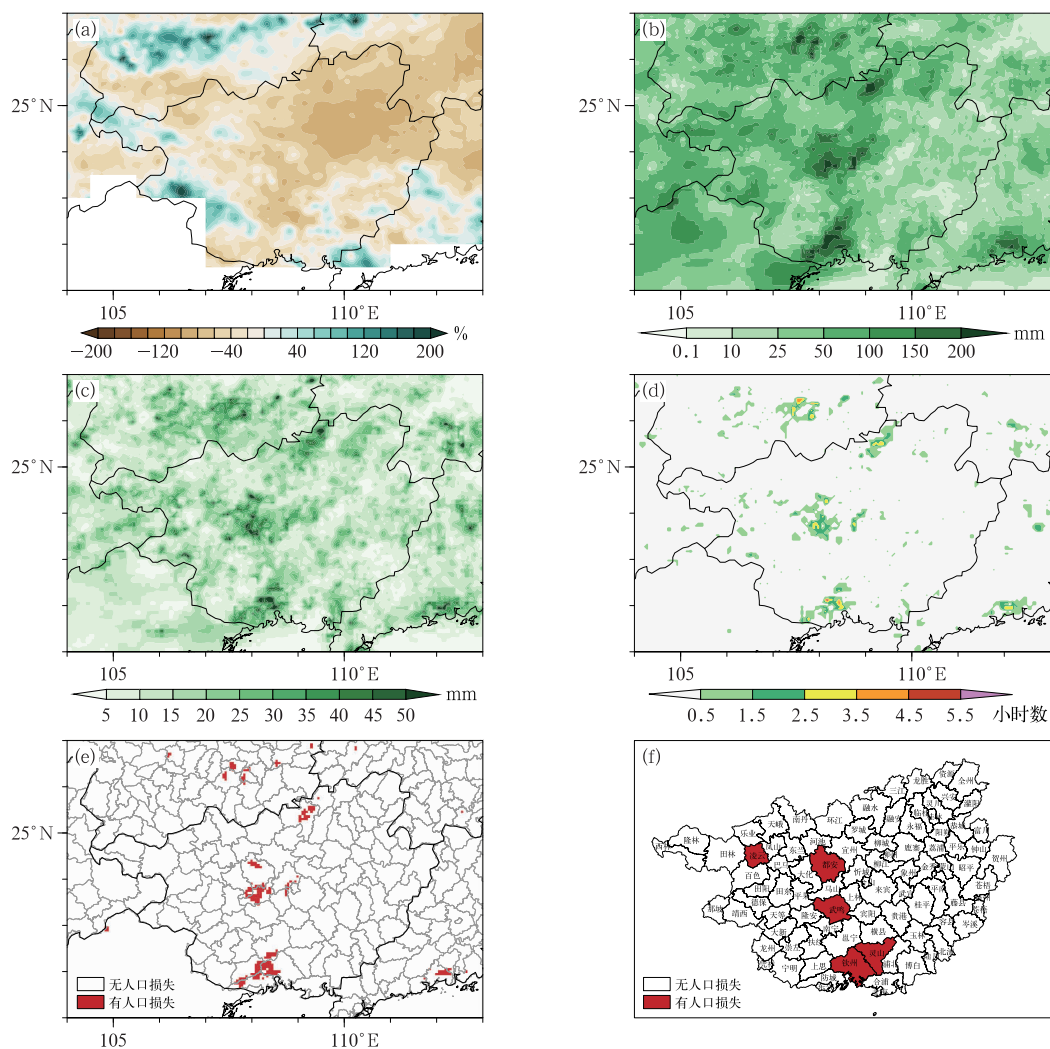


图 6 2018 年 6 月 20—25 日的致灾因子变量

(a)前 10 d 累计降水距平百分率,(b)最大日雨量,(c)最大小时雨量,(d)短时强降水频次;  
以及(e)RF 模型预测的人口损失分布和(f)实际的人口损失分布

Fig. 6 Disaster-triggering factors during 20—25 June 2018

(a) anomaly percentage of accumulated precipitation over the previous 10 days,  
(b) maximum daily rainfall, (c) maximum hourly rainfall, (d) frequency of short-time severe rainfall,  
and (e) distribution of life losses predicted by RF model and (f) distribution of life losses found actually



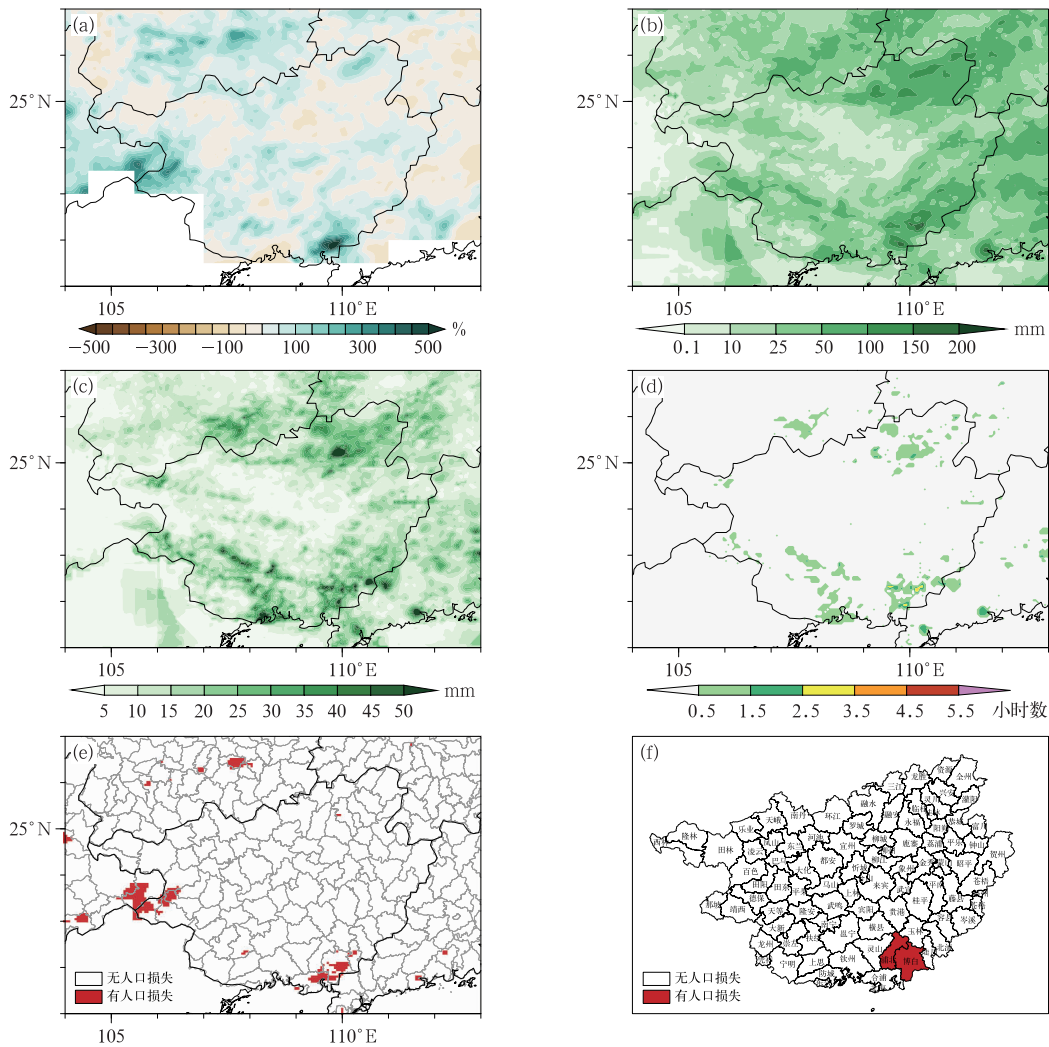


图7 同图6,但为2018年5月6—10日的降雨过程

Fig. 7 Same as Fig. 6, but for the precipitation process during 6—10 May 2018

的地区位于广西东南部 and 西部,与前期累计降水正距平大值区(图7a)有较好的对应关系。尤其是广西东南部博白、浦北县附近,前10 d累计降水量较常年同期偏多5倍,导致人口损失发生的可能性高。基于县域统计,RF模型对这次过程的人口损失分类准确率为91.7%,空报率为8.3%,漏报率为0%。

2019年6月5—13日广西地区发生暴雨灾害,造成全州、资源等13个县(市、区)出现人口死亡。基于全国智能网格实况分析产品,运用RF模型进行回报,分类准确率为81.3%,空报率为17.7%,漏报率为1%(图略)。

### 3 基于智能网格降水预报的模型应用

基于全国智能网格降水预报产品,计算过程最大日雨量、最大小时雨量、短时强降水频次;利用全国智

能网格降水实况分析产品,计算过程开始前10 d的累计降水距平百分率。同时,将承灾体、孕灾环境数据插值到智能网格相同分辨率。将上述解释变量输入之前建立的暴雨灾害人口损失预估模型,便可进行人口损失预报。由于参与建模的训练样本是历史暴雨灾害事件,该模型适用于汛期(5—7月)强降雨时段的人口损失预报。

以2018年5月6—10日的降水过程为例,说明基于智能网格的模型应用效果。使用智能网格降水实况和预报产品计算致灾因子变量,即前10 d降水距平百分率(图8a)和过程最大日雨量(图8b)、最大小时雨量(图8c)、短时强降水频次(由于预报的逐小时降水较弱,没有出现小时降水量超过30 mm的短时强降水,图略),结合网格化的平均人口密度、总GDP、平均地质灾害易发程度,运用之前建立的RF模型,预估人口损失分布(图8d)。预报结果与实际

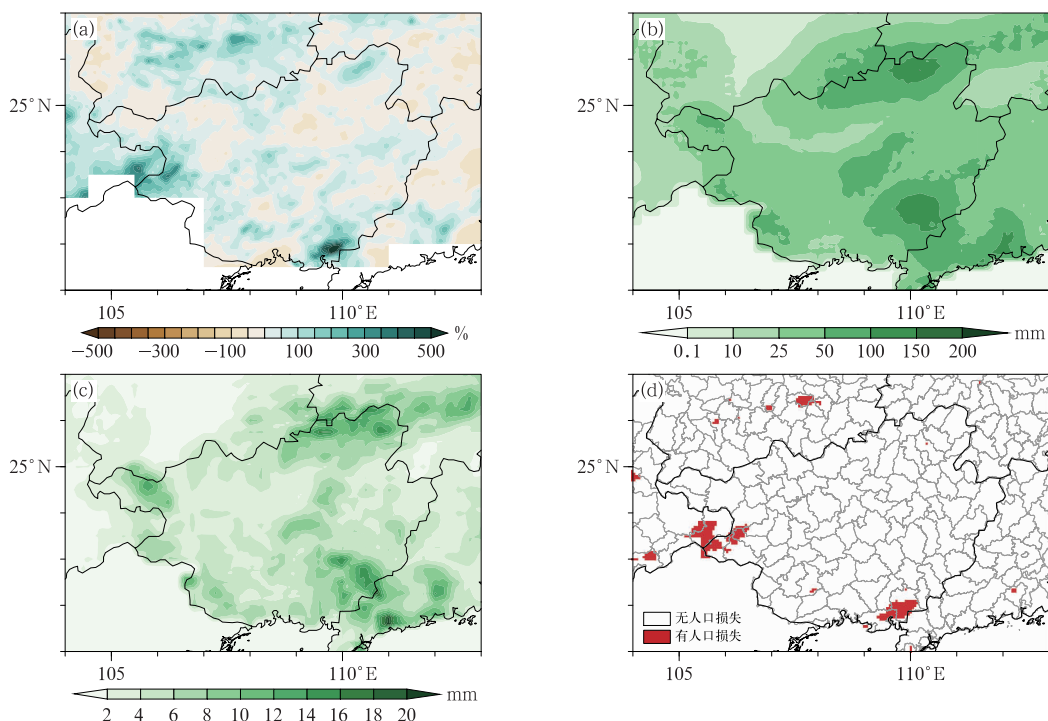


图 8 2018 年 5 月 6—10 日基于智能网格降水预报产品:

(a)前 10 d 累计降水距平百分率,(b)过程最大日雨量,  
(c)最大小时雨量以及(d)RF 模型预估人口损失分布

Fig. 8 (a) Anomaly percentage of accumulated precipitation over the previous 10 days,  
(b) maximum daily rainfall, (c) maximum hourly rainfall,  
(d) distribution of life losses predicted by RF model during 6–9 May 2018  
based on intelligent grid precipitation forecast products

发生人口损失的县(市、区)(图 7f)有较好的对应关系。

对 2019 年 6 月 5—13 日的强降雨过程,由于过程持续时间长,且从 2019 年开始智能网格逐小时降水预报业务产品的最长时效为 72 h,因此运用 RF 模型进行逐日滚动预报,将逐日预报结果合成,得到过程人口损失预报。与实际人口损失相比,预报准确率为 71.9%,空报率为 28.1%,漏报率为 0%(图略)。

## 4 结论和讨论

本研究对 2009—2017 年广西的县级暴雨灾情数据进行了整理分析,综合考虑致灾因子、孕灾环境和承灾体因素选取了 7 个解释变量,基于 RF 分类算法,构建了暴雨灾害人口损失预估模型。结果表明,RF 模型的拟合效果较好,训练样本和测试样本的准确率均在 90%以上。致灾因子(降水情况)是最主要的解释变量,重要性从大到小依次是前 10 d

降水距平百分率、最大日雨量、最大小时雨量和短时强降水频次,其次是人口密度、经济水平和地质灾害易发程度。应用智能网格降水产品驱动模型对近两年广西地区发生的主要暴雨灾害过程进行回报,准确率在 70%以上。

本研究将机器学习算法运用到暴雨灾害人口损失评估,取得了较好效果。但影响暴雨灾害人口损失的因素很多,除了文中考虑的解釋变量,在孕灾环境方面还有河网密度、不透水地表组成等因素,承灾体方面还需考虑人口结构、人的主观能动性,如预警和转移安置,及强降雨的发生时间等(Boyd, 2010; 扈海波和张艳莉, 2014; 王秀荣等, 2016; 尹卫霞等, 2016)。智能网格产品具有高时空分辨率,可以反映短时强降水的致灾危险性,并给出县域尺度上的精细化分布。但由于智能网格业务化时间较短,其实况分析和预报产品的准确度还有待检验,这会对灾害损失预估模型的结果产生影响。

今后,可通过增加合理的解释变量,进一步完善



模型。在获取全国暴雨灾情资料的基础上可将此模型拓展到全国范围,并对接全国智能网格产品,进行精细化暴雨灾害人口损失预报,为防灾减灾提供依据。

## 参考文献

- 曹勇,刘凑华,宗志平,等,2016. 国家级格点化定量降水预报系统[J]. 气象,42(12):1476-1482. Cao Y, Liu C H, Zong Z P, et al, 2016. State-level gridded quantitative precipitation forecasting system[J]. Meteor Mon, 42(12):1476-1482(in Chinese).
- 付晶莹,江东,黄耀欢,2014. 中国公里网格人口分布数据集[J]. 地理学报, 69(S1):41-44. Fu J Y, Jiang D, Huang Y H, 2014. Chinese kilometer grid population distribution dataset[J]. Acta Geograph Sin, 69(S1):41-44(in Chinese).
- 扈海波,张艳莉,2014. 暴雨灾害人员损失风险快速预评估模型[J]. 灾害学, 29(1):30-36. Hu H B, Zhang Y L, 2014. Quick assessing model on casualty loss in rainstorms[J]. J Catastrophol, 29(1):30-36(in Chinese).
- 黄耀欢,江东,付晶莹,2014. 中国公里网格 GDP 分布数据集[J]. 地理学报, 69(S1):45-48. Huang Y H, Jiang D, Fu J Y, 2014. Chinese kilometer grid GDP distribution dataset[J]. Acta Geograph Sin, 69(S1):45-48(in Chinese).
- 金荣花,代刊,赵瑞霞,等,2019. 我国无缝隙精细化网格天气预报技术进展与挑战[J]. 气象, 45(4):445-457. Jin R H, Dai K, Zhao R X, et al, 2019. Progress and challenge of seamless fine gridded weather forecasting technology in China[J]. Meteor Mon, 45(4):445-457(in Chinese).
- 赖成光,陈晓宏,赵仕威,等,2015. 基于随机森林的洪灾风险评价模型及其应用[J]. 水利学报, 46(1):58-66. Lai C G, Chen X H, Zhao S W, et al, 2015. A flood risk assessment model based on random forest and its application[J]. J Hydr Eng, 46(1):58-66(in Chinese).
- 李亨,田原,邹伦,等,2014. 基于随机森林方法的滑坡灾害危险性区划[J]. 地理与地理信息科学, 30(6):25-30. Li T, Tian Y, Wu L, et al, 2014. Landslide susceptibility mapping using random forest[J]. Geogr Geo-Inform Sci, 30(6):25-30(in Chinese).
- 李文娟,赵放,赵敏杰,等,2018. 基于数值预报和随机森林算法的强对流天气分类预报技术[J]. 气象, 44(12):1555-1564. Li W J, Zhao F, Li M J, et al, 2018. Forecasting and classification of severe convective weather based on numerical forecast and random forest algorithm[J]. Meteor Mon, 44(12):1555-1564(in Chinese).
- 王盼,陆宝宏,张瀚文,等,2014. 基于随机森林模型的需水预测模型及其应用[J]. 水资源保护, 30(1):34-37, 89. Wang P, Lu B H, Zhang H W, et al, 2014. Water demand prediction model based on random forests model and its application[J]. Water Resour Protect, 30(1):34-37, 89(in Chinese).
- 王秀荣,吕终亮,王莉萍,等,2016. 一种简化的暴雨灾害风险及影响评估方法和应用研究——以京津冀“7·21”暴雨事件为例[J]. 气象, 42(2):213-220. Wang X R, Lü Z L, Wang L P, et al, 2016. Simplified assessment method and application research of rainstorm disaster risk and impact—using Jing-Jin-Ji “7. 21” heavy rain as an example[J]. Meteor Mon, 42(2):213-220(in Chinese).
- 王志军,顾冲时,娄一青,2008. 基于支持向量机的溃坝生命损失评估模型及应用[J]. 水力发电, 34(1):67-70. Wang Z J, Gu C S, Lou Y Q, 2008. Model for evaluation of loss-of-life caused by dam breach based on support vector machine[J]. Water Power, 34(1):67-70(in Chinese).
- 吴琼,李运田,郑献卫,2013. 面向非平衡训练集分类的随机森林算法优化[J]. 工业控制计算机, 26(7):89-90. Wu Q, Li Y T, Zheng X W, 2013. Optimized random forests algorithm for imbalanced training sets[J]. Ind Control Comput, 26(7):89-90(in Chinese).
- 吴小君,方秀琴,任立良,等,2018. 基于随机森林的山洪灾害风险评估——以江西省为例[J]. 水土保持研究, 25(3):142-149. Wu X J, Fang X Q, Ren L L, et al, 2018. Risk assessment of mountain torrents disaster based on random forest—a case study in Jiangxi Province[J]. Res Soil Water Conserv, 25(3):142-149(in Chinese).
- 杨柳,王钰,2015. 泛化误差的各种交叉验证估计方法综述[J]. 计算机应用研究, 32(5):1287-1290, 1297. Yang L, Wang Y, 2015. Survey for various cross-validation estimators of generalization error[J]. Appl Res Comput, 32(5):1287-1290, 1297(in Chinese).
- 杨帅,苏筠,2014. 县域暴雨洪涝灾害损失快速评估方法探讨:以湖南省为例[J]. 自然灾害学报, 23(5):156-163. Yang S, Su Y, 2014. Exploration of rapid evaluation method for rainstorm flood losses at county level; a case study of Hunan Province[J]. J Nat Dis, 23(5):156-163(in Chinese).
- 尹卫霞,余瀚,崔淑娟,等,2016. 暴雨洪水灾害人口损失评估方法研究进展[J]. 地理科学进展, 35(2):148-158. Yin W X, Yu H, Cui S J, et al, 2016. Review on methods for estimating the loss of life induced by heavy rain and floods[J]. Prog Geogr, 35(2):148-158(in Chinese).
- 张雪蕾,汪明,曹寅雪,等,2018. 3种典型机器学习方法在灾害敏感性评估中的对比[J]. 中国安全生产科学技术, 14(7):81-87. Zhang X L, Wang M, Cao Y X, et al, 2018. Comparison of three typical machine learning methods in susceptibility assessment of disasters[J]. J Saf Sci Technol, 14(7):81-87(in Chinese).
- 赵钢铁,杨大文,蔡喜明,等,2012. 基于随机森林模型的长江上游枯水期径流预报研究[J]. 水力发电学报, 31(3):18-24, 38. Zhao T T G, Yang D W, Cai X M, et al, 2012. Predict seasonal low flows in the upper Yangtze River using random forests model[J]. J Hydroelectr Eng, 31(3):18-24, 38(in Chinese).
- Boyd E C, 2010. Estimating and mapping the direct flood fatality rate for flooding in greater New Orleans due to Hurricane Katrina[J]. Risk Hazards Crisis Public Policy, 1(3):91-114.
- Breiman L, 1996. Bagging predictors[J]. Mach Learn, 24(2):123-140.
- Breiman L, 2001a. Statistical modeling: the two cultures[J]. Stat Sci,

- 16(3):199-215.
- Breiman L, 2001b. Random forests[J]. *Mach Learn*, 45(1):5-32.
- Dong L J, Li X B, Peng K, 2013. Prediction of rockburst classification using random forest[J]. *Trans Nonferr Met Soc China*, 23(2):472-477.
- Hanson S, Nicholls R, Ranger N, et al, 2011. A global ranking of port cities with high exposure to climate extremes [J]. *Climatic Change*, 104(1):89-111.
- Iverson L R, Prasad A M, Matthews S N, et al, 2008. Estimating potential habitat for 134 eastern US tree species under six climate scenarios[J]. *Forest Ecol Manage*, 254(3):390-406.
- McGovern A, Gagne II D J, Troutman N, et al, 2011. Using spatiotemporal relational random forests to improve our understanding of severe weather processes[J]. *Stat Anal Data Min*, 4(4):407-429.
- Peduzzi P, Chatenoux B, Dao H, et al, 2012. Global trends in tropical cyclone risk[J]. *Nat Climate Change*, 2(4):289-294.
- Peng M, Zhang L M, 2012. Analysis of human risks due to dam-break floods—part 1: a new model based on Bayesian networks[J]. *Nat Hazards*, 64(1):903-933.
- Penning-Rowsell E, Floyd P, Ramsbottom D, et al, 2005. Estimating injury and loss of life in floods; a deterministic framework[J]. *Nat Hazards*, 36(1/2):43-64.
- Williams J K, 2014. Using random forests to diagnose aviation turbulence[J]. *Mach Learn*, 95(1):51-70.