

基于随机森林和时空核密度方法的不同周期 犯罪热点预测对比

柳林^{1,2,3}, 刘文娟¹, 廖薇薇¹, 余洪杰¹, 姜超¹, 林荣平¹, 纪佳楷¹, 张政¹

(1. 中山大学地理科学与规划学院综合地理信息研究中心, 广州 510275; 2. 广州大学地理科学学院公共安全地理
信息分析中心, 广州 510006; 3. 辛辛那提大学地理系, 美国辛辛那提 OH45221-0131)

摘要: 犯罪预测对于制定警务策略、实施犯罪防控具有重要意义。机器学习和核密度是2类主流犯罪热点预测方法,然而目前还鲜有研究对这2类方法在不同时间周期下的犯罪预测效果进行系统比较,本文试图对此进行补充。本文以2013-2016年5月的公共盗窃犯罪历史数据作为输入,分别对比了在接下来2周、1个月、2个月、3个月4个不同时间周期随机森林方法与基于时空邻近性的核密度方法的犯罪热点预测效果,结果发现:在各时间周期上,随机森林分类热点预测方法的面积和案件量命中率均比时空核密度方法准确性高;并且2种方法均能有效地识别犯罪热点中的高发区域,其中在较小范围较短时间随机森林识别热点中的高发区效率更高,而在较大范围较长时间周期上时空核密度方法识别高发区更优。

关键词: 时空核密度; 随机森林算法; 犯罪热点预测; 犯罪高发区识别

1 引言

随着中国城市化进程的快速推进,城市犯罪问题日渐突出,探索如何有效地开展犯罪防控既是公安部门的工作重点,也是犯罪研究的热点和难点。其中,犯罪预测是进行有效犯罪预防的基础,对于打击犯罪、维护社会稳定具有重要意义。过去犯罪预测常以经验规律为基础,而在当前大数据时代背景下,随着机器学习方法在各研究领域的不断普及和应用,犯罪预测防控研究迎来了新的机遇。

国外在犯罪预测方面,以美国的实证研究最有代表性,涉及治安状况、热点时序、作案地居住地、作案类型、特定人群等多个方面。目前中国的犯罪

预测研究以国家治安形势预测(梁晓军, 2001)、区域犯罪趋势预测(王发曾, 1992)等宏观趋势理论探讨居多,有关犯罪热点预测的实证研究还较少(陈鹏等, 2011; 刘大千等, 2012; 阎耀军等, 2013; 李卫红等, 2017)。犯罪热点预测以日常活动理论(Cohen et al, 1979; Clarke et al, 2004)、环境犯罪学(Ratcliffe, 2004; Brantingham, 2008)为背景,认为犯罪的发生离不开时间和空间。研究发现,犯罪案件的发生在空间上不是随机均匀分布,而是呈现出一定的集聚和离散特征(Ratcliffe, 2004; Bowers et al, 2005; Grubestic et al, 2008)。对此,Weisburd(2015)总结了犯罪空间集聚规律,认为在地理空间单元上,大部分犯罪案件集聚在小部分区域,这表明犯

收稿日期:2018-02-02;修订日期:2018-03-29。

基金项目: 国家自然科学基金重点基金项目(41531178);广东省自然科学基金研究团队项目(2014A030312010);国家自然科学基金项目(41171140);广东省科技计划项目(2015A020217003) [Foundation: Key Program of National Natural Science Foundation of China, No.41531178; Research Team Program of Natural Science Foundation of Guangdong Province, China, No.2014A030312010; National Natural Science Foundation of China, No.41171140; Science and Technology Program of Guangdong Province, China, No.2015A020217003]。

作者简介: 柳林(1965-),男,湖南湘潭人,博士,教授,主要研究方向为地理信息科学、犯罪时空分析与模拟等, E-mail: lin.liu@uc.edu。

引用格式: 柳林, 刘文娟, 廖薇薇, 等. 2018. 基于随机森林和时空核密度方法的不同周期犯罪热点预测对比[J]. 地理科学进展, 37(6): 761-771.
[Liu L, Liu W J, Liao W W, et al. 2018. Comparison of random forest algorithm and space-time kernel density mapping for crime hot-spot prediction[J]. Progress in Geography, 37(6): 761-771.]. DOI: 10.18306/dlkxjz.2018.06.003

罪案件在空间上的分布存在热点和冷点区域(Brantingham et al, 1999; Groff et al, 2002),从而使得一定程度上预测犯罪高发的时间和地点成为可能。

传统的犯罪风险估算方法通常从犯罪案件历史分布中探测出犯罪热点区域,并假设这种规律将会持续到下一个时间周期(Gorr et al, 2003)。如地形风险模型(Risk Terrain Modeling, RTM)(Caplan et al, 2011),考虑到犯罪地的邻近性和犯罪要素的聚集性,采用犯罪相关环境因素的数据和犯罪历史数据等进行犯罪预测,并且对于长周期稳定的犯罪热点预测比较有效。常用的核密度估计方法利用犯罪案件的空间集聚规律进行犯罪风险制图,被公认为能有效地识别热点区域(Hirschfield et al, 2001; Chainey et al, 2002; Clarke et al, 2005; Chainey et al, 2008; Chainey et al, 2013)。Bowers等(2004)发明的犯罪风险地图(ProMap)本质上也是采用了考虑时间关联性的核密度方法。国内有研究发现,基于时间临近性的核密度方法对未来一年的犯罪热点预测效果比一般的核密度估计效果更好(徐冲等, 2016)。时空核密度方法能较好的可视化犯罪热点的形成,但挖掘数据信息的能力不强。

近年来,利用大数据和机器学习、深度学习的方法进行犯罪预测研究已成为热点,不少研究结合人口经济统计数据、土地利用数据、手机数据等和犯罪历史数据进行了不同时间周期的犯罪预测实证研究(李卫红等, 2017; Kianmehr et al, 2008; Bogomolov et al, 2014; Rummens et al, 2017),采用的方法有随机森林、支持向量机、神经网络、贝叶斯模型等多种算法。在各种机器学习算法中,随机森林算法已被证明在多个领域具有较强非线性关系数据处理能力和较高的预测准确率(Genuer et al, 2010; Kandaswamy et al, 2011; Rodriguez et al, 2012)。随机森林的方法虽然学习效率较高,但对于犯罪热点形成原理解释尚不足。

综上所述,时空核密度和随机森林2种算法在犯罪热点预测上均表现不俗,但2种预测方法在不同时间周期的犯罪热点预测时,到底哪种方法效果更好,在预测原理上各自的优势如何,尚无研究作过系统比较。而通过对比能加深对于犯罪时空热点分布规律的理解,同时更好地指导未来的犯罪预测研究和警务防控实践。因此,本文采用历史犯罪数据对比不同时间尺度上时空核密度和随机森林

算法2类犯罪热点预测方法的准确性和有效性。

2 研究区域和数据

2.1 研究区域

研究区(图1)隶属于中国东南沿海特大城市ZG市HT区。HT区位于ZG市老城区东部,辖内各种交通资源高度聚集,经济持续平稳较快发展。作为全市的中心城区,人口结构复杂,流动性大,社会经济活动繁杂多样。2015年行政区域总面积约137.38 km²,辖有21条行政街。全区户籍人口84.46万人,常住人口154.57万人,2015年全年地区生产总值3438.65亿元,比上年增长8.8%,总量连续9年位居全市首位。同时,该区警务信息化水平较高,数据记录准确全面,可靠性高。

为了减弱边缘效应,研究区以HT区4个派出所(QJ、ZJ、HC、CB)管辖区向外600 m的缓冲区为边界。基本现状数据分别来源于ZG市公安局2013-2016年110接警数据和P-GIS(警务地理信息系统)数据库。110接处警数据记录了每起案件的案件类型,案发时间、地点坐标及接警单位等信息。经自动匹配和人工校正后将110接警数据落到地理空间上,并剔除不在研究区内的案件点。其中公共盗窃警情相比于其他,案件量多,影响范围广,社会潜在危害影响大。

2.2 数据概况

从2013-2016年ZG市HT区公共盗窃案件数量来看,2016年明显少于其他年份,但分月统计显示(图2):4年都呈现出较为相似的趋势性规律,1-2月



图1 研究区

Fig.1 The study area

份案件量陡降,3月份案件量恢复到稳定水平,3-12月份案件量较为平稳,在一定范围内上下波动。各年份的分日统计(图3)特征也较为相似,案件量从1-30日均平缓的上下波动,受年际大小月份影响,2、4、6、9、11月无31日,31号的案件统计数量少于其他月份。

另外统计发现,若以50 m×50 m的网格单元划分研究区,各年全部案件都集聚在7%~8%的网格中。进而对2013-2016年公共盗窃案件按年进行热点分析(图4),发现4年间研究区内公共盗窃案件高发区主要集中在3个部分,热点范围有发生变化,空间位置几乎无太大转移。2013年、2014年、2015年与2016年的犯罪热点高发前20%的区域重叠度分别高达0.91,0.88,0.89,说明研究区内公共盗窃犯罪热点空间重叠度非常高,犯罪热点在年时间周期上相当稳定,变化很小。

3 研究方法

3.1 基于时空邻近性的核密度方法

地理学第一定律认为事物在空间上的分布相互联系,且邻近事物之间的联系更紧密,存在集聚、随机、规则分布。基于此原理,核密度(Kernel Density)方法对事物之间的空间联系进行量化计算来反映其分布规律,搜索半径用来划定事物之间的邻近阈值,选取特定的空间衰减函数来描述某事件点与搜索半径覆盖范围内事件点的局部空间关联,表示事物的空间联系紧密度与邻近距离的关系。基于时空邻近性考虑的核密度进一步认为,近期发生的事件相对发生越久远的事件对空间位置的影响更显著,因此加入时间远近大小作为犯罪风险估算的权重。这种方法充分考虑了事物的时空邻近关系与集聚特征形成的联系,从而能根据历史案件累积

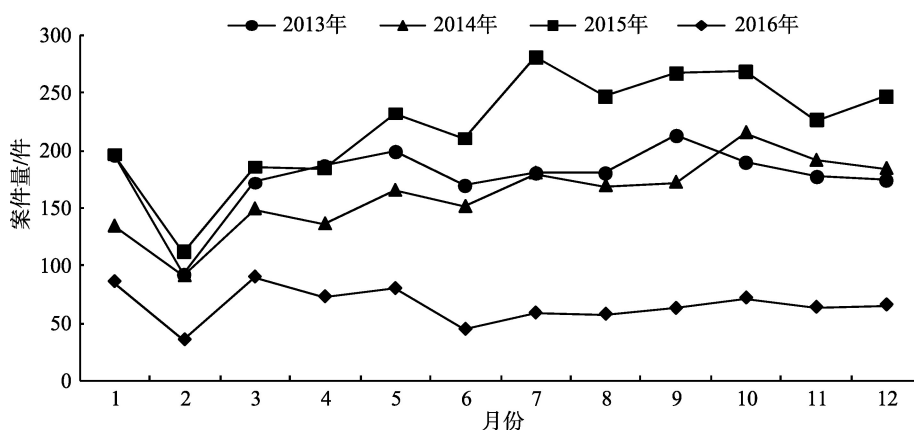


图2 2013-2016年公共盗窃案件分月统计图

Fig.2 Monthly counts of theft from 2013 to 2016

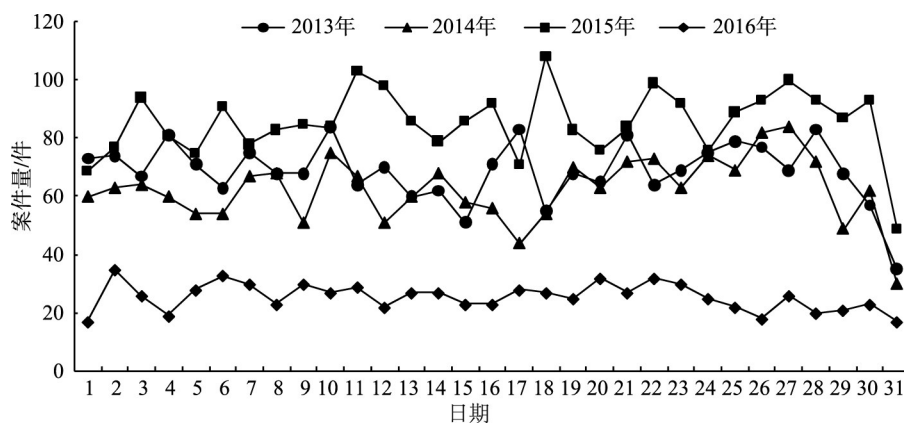


图3 2013-2016年公共盗窃案件分日统计图

Fig.3 Daily counts of theft from 2013 to 2016

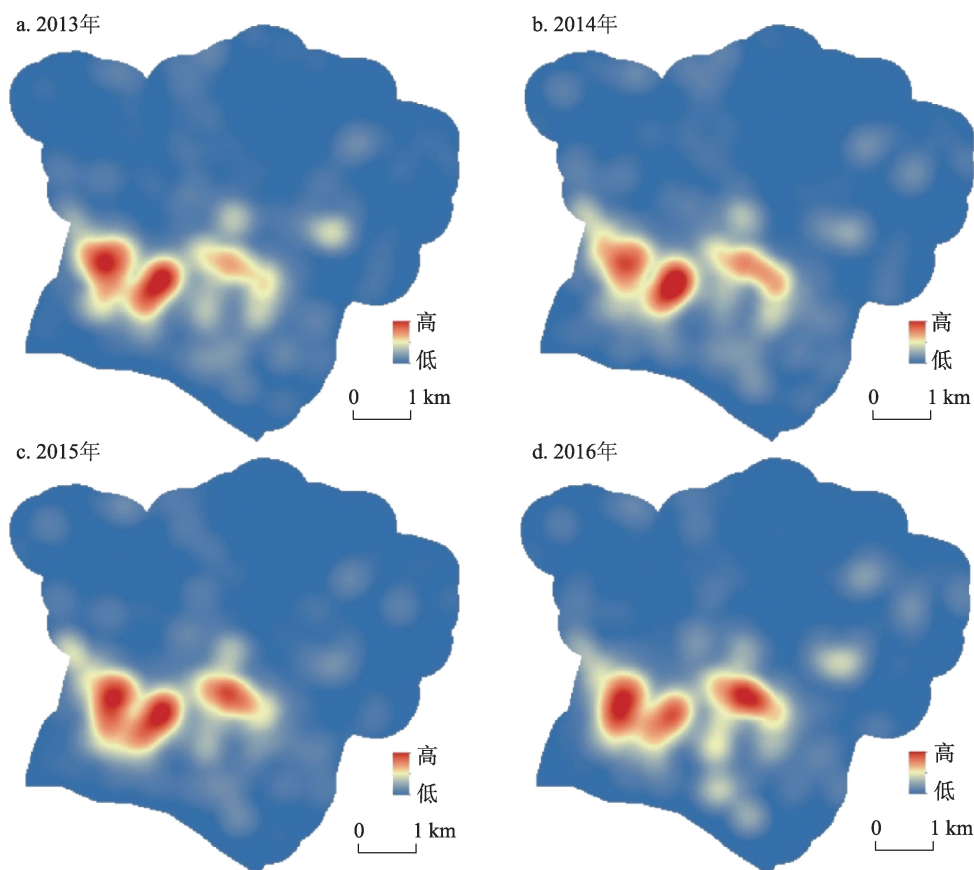


图4 2013-2016年公共盗窃案件热点分布图

Fig.4 Hotspot distribution of theft from 2013 to 2016

的空间分布集聚规律来定量估算风险大小,进而较好地识别整体上稳定存在的犯罪高发热点区域。缺点是时空核密度估算热点的高低受历史事件点邻近程度影响,与重复发生的事件点的时空位置可能错位。

3.2 随机森林分类算法

随机森林是由Leo Breiman于2001年提出的一种利用多棵树对样本进行训练并预测的分类算法(Breiman, 2001)。其基本原理是在决策树的基础上,综合多个决策的结果。单棵树的分类能力可能很小,但在随机产生大量的决策树后,一个测试样本可以通过每一棵树的分类结果经投票统计后选择最可能的分类。假设有 N 个样本,每个样本有 M 个特征,具体实现步骤如下:

(1) 原始训练集为 N ,应用bootstrap法有放回地随机抽取 k 个新的自助样本集,并由此构建 k 棵分类树,每次未被抽到的样本组成了 k 个袋外数据;

(2) 设有 M 个变量,则在每一棵树的每个节点处随机抽取 m 个变量($m \ll M$),然后在 m 中选择一个

最具有分类能力的变量,变量分类的阈值通过检查每一个分类点确定;

(3) 每棵树最大限度地生长,不做任何修剪;

(4) 将生成的多棵分类树组成随机森林,用随机森林分类器对新的数据进行判别与分类,分类结果按树分类器的投票多少而定。

随机森林分类算法中生成了多棵决策树,综合多个决策结果可提高分类的灵敏度和准确度。利用随机森林算法进行犯罪预测,能挖掘特征之间和样本之间的相互联系来学习犯罪时空分布的先验规律进行模型训练;此外,在决策树生成过程中引入了随机性,不易出现过拟合现象,对犯罪发生过程中存在的随机因素有较好的容忍度。这种方法具有较高的学习效率和能力,但只能进行热点和非热点的定性判断,不能直接量化估计犯罪发生的可能性大小。

3.3 评价指标

命中率是常见的用来评价犯罪预测准确性的指标之一。本文主要考虑预测面积的命中率、案件

量命中率、预测案件密度3个指标。预测面积的命中率即在一定预测面积比下预测为热点的区域中预测正确的面积比例;案件量命中率指的是一定预测面积比例下预测为热点的区域中实际发生案件量与总案件量之比;预测案件密度是一定预测面积比例下预测正确的热点区域内案件量与面积之比。

4 预测实验及对比

4.1 预测实验

基于时空邻近性的核密度进行犯罪风险估算在以往的核密度方法的基础上加上了事件点时间关联的计算,因此将2013-2016年5月的公共盗窃7773起案件点的时间和位置作为输入,采用在ArcGIS10.0软件中的核密度基础上进行二次开发的工具实现时空邻近相似性核密度犯罪风险估算。本文参照文献徐冲等(2013)选取Gaussian函数作为距离衰减函数及反时间距离权重计算公式构建模型进行实验。其中有两个重要参数需要注意,距离搜索半径参考文献(徐冲等, 2013; 徐冲等, 2016)设为150 m,输出密度风险图的栅格单元大小设为50 m×50 m。以2016年6、7、8月发生的公共盗窃案件点作为验证数据。

随机森林分类算法直接预测每一个空间单元是否为热点。与时空核密度保持一致,将研究区划分为50 m×50 m的网格,则研究区含网格数共计14720个,每一个单元网格格的犯罪热点预测可以看作一个二分类问题,在下一个预测时期内预测为热点或者非热点。选择每一个单元中2013年、2014年、2015年与预测目标时间同期的时间段,临近时间内均向前推算若干与预测目标时段长度相同的时间段的历史案件数据建立随机森林预测模型,建立决策树300棵,随机抽取变量数设为4。经过多次计算发现,选取4个临近步长时段,已能达到较好的预测效果。并且对4个临近时段内犯罪案件的发生地以150 m为搜索半径作核密度分析,取每

个网格的临近4个时间段的极差标准化核密度值。以每个网格中各时段的犯罪案件量及案件密度为样本,用前一期的犯罪案件作为训练标准设置热点和非热点分类标签,预测这一期是否为犯罪热点。

由于该区域内公共盗窃案件的空间分布集聚特征显著,有案件发生的区域占整个研究区的面积比例极小,并且在研究时段的2周内总共有76起案件发生在研究区的72个网格内,案件密度较小,为方便与其他时段进行比较,统一将有案件发生的区域均视为热点区域。本文对2016年6月份前2周、1个月(6月)、2个月(6、7月)、3个月(6、7、8月)均进行了预测,各时间周期的犯罪案件网格分布情况如表1,以1个月为例对实验操作说明如表2。

4.2 预测结果

时空核密度风险估算得到的是栅格化密度风险图,而随机森林算法得到的是矢量的矩形网格热点图。为便于对比二者的预测结果,先将栅格图转为矢量网格图,其中实际有案件发生的网格视为实际热点网格,若实际热点网格与预测热点网格重叠则视为预测正确。分别统计预测热点区域的面积和真实案件数量来比较2种方法的预测准确性。由于研究区内案件量少,集聚程度较高,大部分区域没有案件发生,在2016年6月份前2周、1个月(6月)、2个月(6、7月)、3个月(6、7、8月)共4个时间周期内有案件发生的网格单元数量分别为72、140、282、415,分别约占研究区面积的0.5%、1%、2%、3%。对于随机森林预测结果,按照热点得票数对随机森林预测网格进行排序,分别选取4个时间周期投票数靠前的72、140、282、415个网格作为热点区域。相应地,时空核密度算法抽取密度风险最大的前72、140、282、415个栅格单元作为热点区域。对比各时间段时空核密度和随机森林方法预测结果及真实案件分布情况(图5),红色的网格是2种方法均正确预测为热点的区域,绿色的网格是只有时空核密度方法预测正确的热点区域,蓝色的网格是只有随机森林预测方法预测正确的热点区域,底图

表1 不同周期案件量及有案件发生网格情况
Tab.1 Count of grids where crimes occurred and total crime numbers in different time periods

	热点网格数/个	总案件量/起	单个网格最大案件量	热点网格案件量中位数	热点网格案件量平均值
2016年6月份前2周	72	76	2	1	1.05
2016年6月份	140	164	6	1	1.17
2016年6、7月份	282	350	12	1	1.24
2016年6、7、8月份	415	554	20	1	1.33

表2 2016年6月份犯罪热点分类预测实验说明
Tab.2 Explanation of crime hotspot classification
prediction for June, 2016

	周期性	临近性
输入训练数据	2013年5月份案件量	2016年1月份案件核密度
	2014年5月份案件量	2016年2月份案件核密度
	2015年5月份案件量	2016年3月份案件核密度
	/	2016年4月份案件核密度
输入测试数据	2013年6月份案件量	2016年2月份案件核密度
	2014年6月份案件量	2016年3月份案件核密度
	2015年6月份案件量	2016年4月份案件核密度
	/	2016年5月份案件核密度
输入分类标签	2016年5月份热点/非热点	/
输出分类标签	/	2016年6月份热点/非热点

是对应预测时间段真实案件分布的核密度图。可见,随着预测时间周期的延长,2种方法预测正确的网格数量也随之增加,并且各预测周期内随机森林方法预测的正确热点网格数量均大于时空核密度估算方法。从空间上看,时空核密度方法预测正确

的结果空间上更为集聚,而随机森林预测正确的热点区域在整个研究区内分布较为分散。

4.3 对比分析

首先对比2种方法预测面积的命中率。如图6所示,横坐标是预测热点的面积与研究区总面积之比,纵坐标表示预测正确的热点网格数与预测热点网格数之比。总体上,2种方法预测面积的命中率随预测时间周期的延长逐渐提高,比如在与真实案件分布等面积预测时,时空核密度方法在2周、1个月、2个月、3个月的预测正确网格比例依次为6.94%、7.86%、11.35%、15.66%;而随机森林方法依次为9.72%、18.57%、25.53%、33.73%。不难看出在预测同等面积热点区域条件下,随机森林方法预测正确的网格面积在各测时间周期上均高于时空核密度方法。

将研究区2016年对应预测时间段内的案件点与预测结果进行叠加统计,进一步对预测案件量的命中率进行评估。落在预测热点区域内的视为正

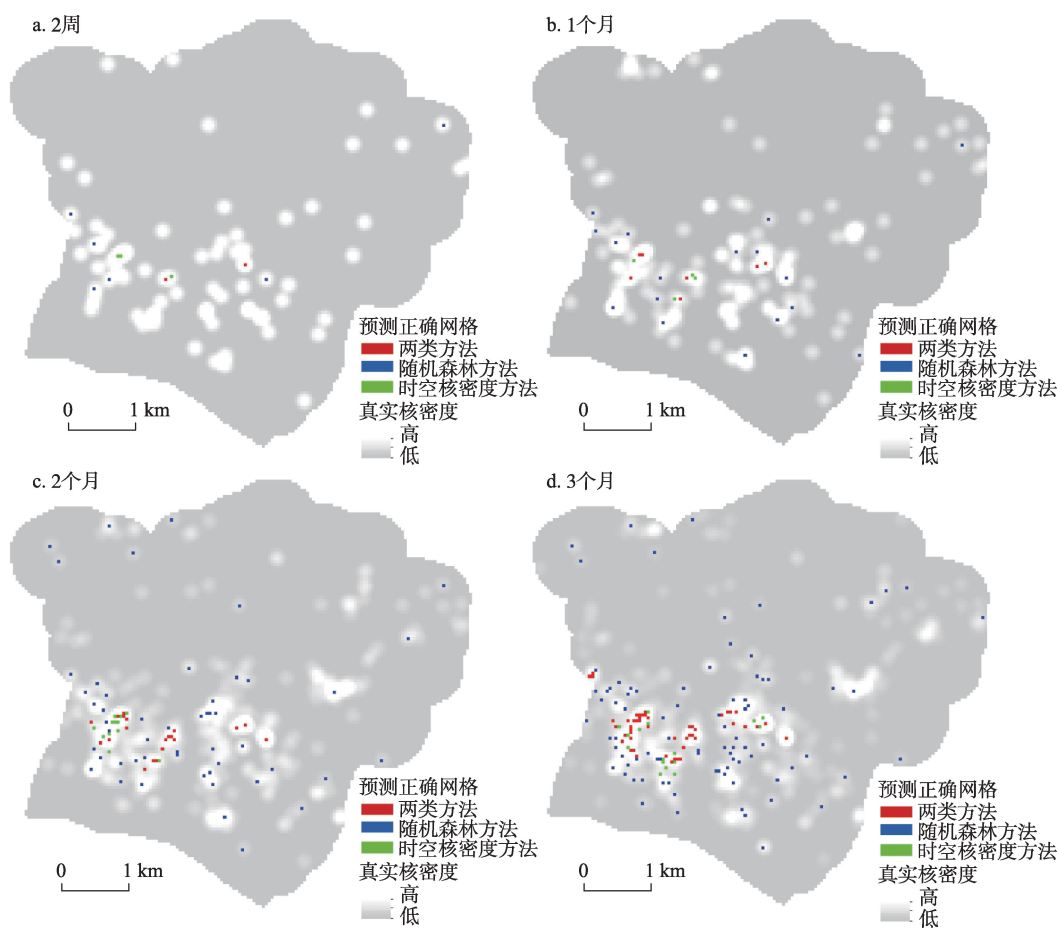


图5 时空核密度与随机森林预测热点对比图

Fig.5 Comparison of hotspot prediction results of space-time kernel density and random forest methods

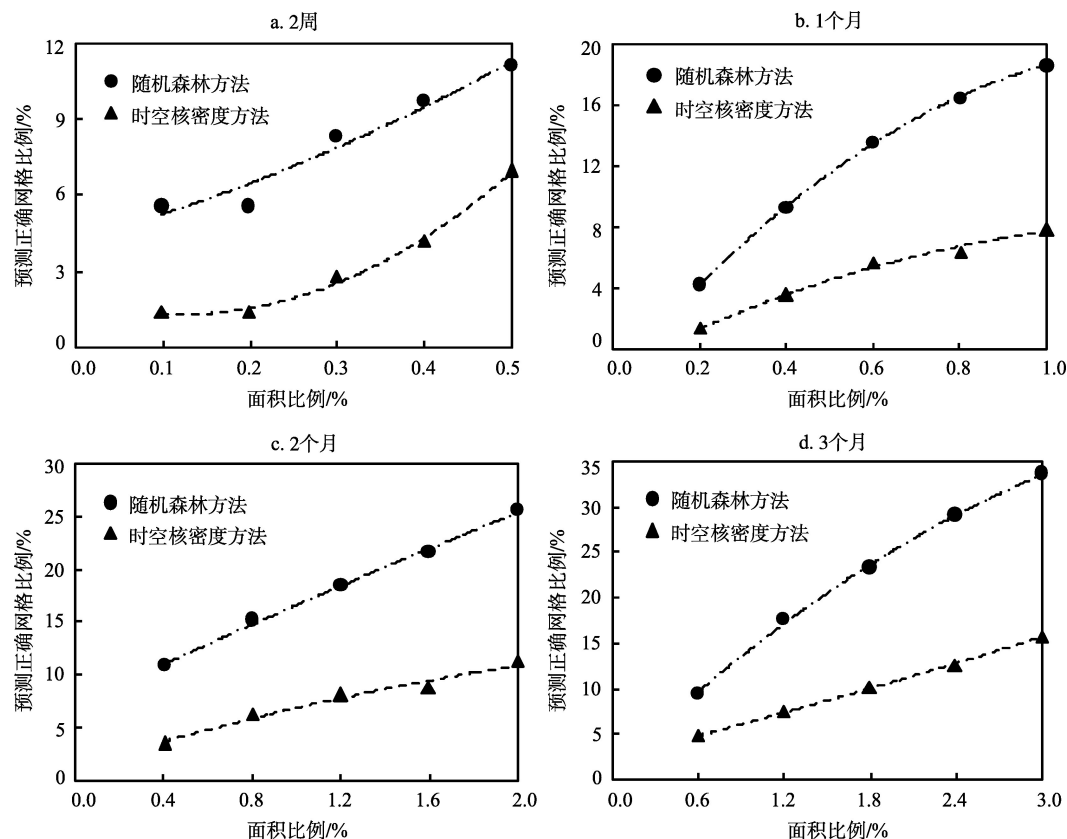


图6 各时间周期预测正确热点网格占比

Fig.6 The hit area ratio in different time periods

确预测的案件,统计预测正确的案件数占总案件数的比例,可以得到各预测时间周期下2种方法在不同预测面积下正确预测的案件比例。2种方法的案件量预测准确性对比如图7所示。与预测面积的命中率对比结果类似,2种方法在各时段上预测的案件量命中率在长周期时间上要优于短周期。同等预测面积条件下,随机森林方法预测案件量命中率在各个时间长度上均高于时空核密度方法。如在与真实案件分布等面积预测时,时空核密度方法在前2周、1个月、2个月、3个月的预测案件量比例依次9.21%、12.19%、18.28%、25.27%;而随机森林方法依次为10.53%、26.22%、35.14%、45.30%。

在同样的研究区范围内,受环境因素影响,案件的发生在空间上的分布是不均匀的,同样大小的空间单元内发生的案件量大小也存在差异。案件量较大的地方是热点区域中的高发区,社会治安形势严峻,也是犯罪防控和警务实践关注的重点之一。图8是各时间段预测正确热点区域案件密度与预测面积之间的关系。标准情况是指同等预测面积大小条件下,按案件量从高到低排序,选取研究

区内案件量靠前的网格中总案件量与预测面积之比。由图中可以看出,从2周、1个月、2个月到3个月4个时间周期上,在较小较短时间周期内随机森林方法识别案件高发区的效率要高于标准条件,时空核密度方法则要低于标准条件;而在较长时间周期较大预测范围内预测正确的热点区域内案件密度大小顺序依次是时空核密度方法、随机森林方法、标准条件。说明2种方法均能有效识别犯罪热点中的高发区域。其中,在较小范围较短时间范围内随机森林识别热点中的高发区效率更高,而在较大范围较长时间周期上时空核密度方法识别高发区更优。

5 讨论与结论

国内外有不少学者采用核密度、机器学习等方法进行犯罪热点预测,并采用不同指标对预测效果进行评估,本文2种方法的预测结果可与其他方法间接对比评价。尽管研究区和研究数据不同,与Bowers等(2004)经典的风险地图预测精度相比,其

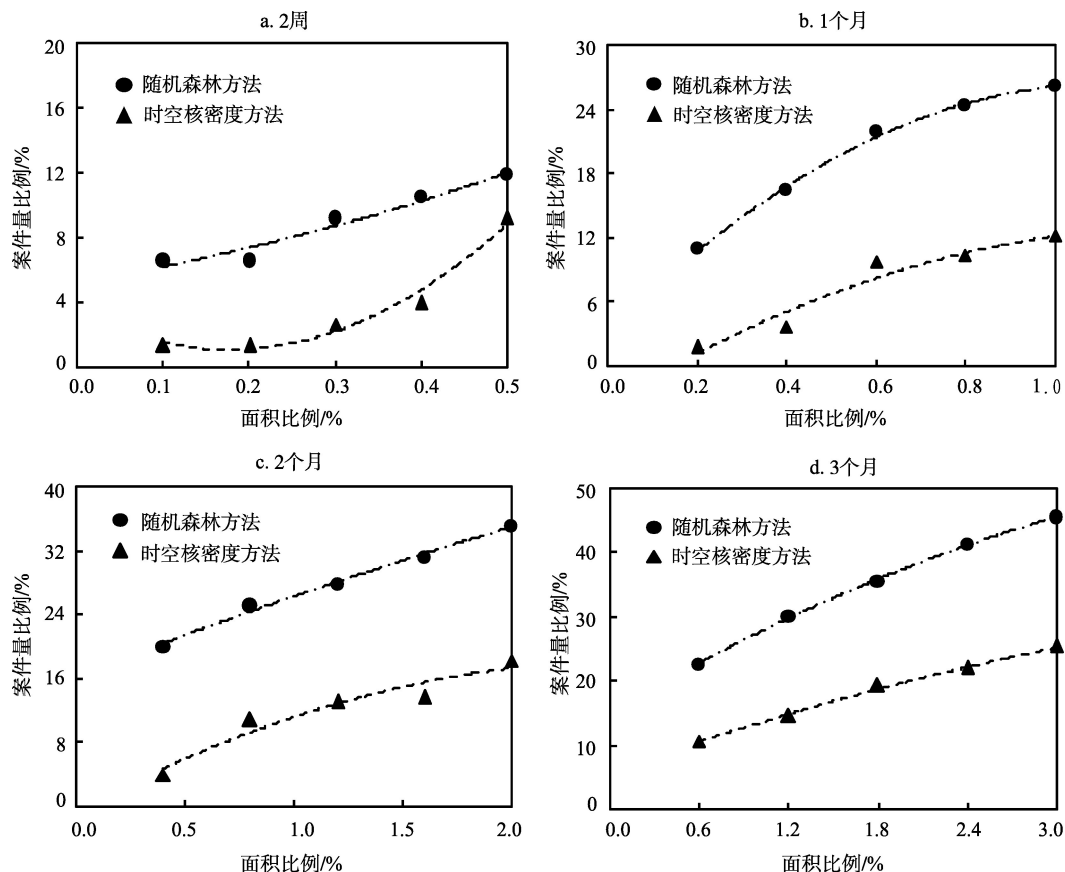


图7 各时间周期预测正确热点区域内案件量比例

Fig.7 The hit case ratio in different time periods

预测的未来一周发生的案件量为70起,栅格单元大小同样为 $50\text{ m} \times 50\text{ m}$,共计10816个网格,在风险值前20%的区域预测的案件量命中率为64%。本文实验中2周案件量与之相差无几,而时空核密度方法风险值前20%的区域的案件量命中率达到75%。与Rummens等(2017)的文章中利用人口统计、社会经济、土地利用等数据,采用逻辑回归、神经网络及综合2种方法的预测结果相比,在 $125\text{ m} \times 125\text{ m}$ 的网格单元尺度下1个月周期内,神经网络方法预测前20%的区域案件量命中率最高,为70.48%;同样以1个月为周期本文中随机森林方法以 $50\text{ m} \times 50\text{ m}$ 进行预测案件量命中率为76.83%,高于该值。Bogomolov等(2014)也提到采用了逻辑回归、支持向量机、神经网络、决策树等其他机器学习的方法进行犯罪热点分类预测,其中随机森林算法效果最好。上述预测结果对比,在一定程度上表明了时空核密度和随机森林是2种比较有代表性的犯罪预测方法。

本文以ZG市HT区4个派出所的公共盗窃案

件数据为案例,对比了不同时间尺度上的时空核密度和随机森林算法这2类主流方法的犯罪热点预测效果,结果发现:2种方法均能有效识别犯罪热点中的高发区域,随机森林分类热点预测方法的面积和案件量命中率均比时空核密度估算准确性高。并且在较短时间较小范围内,随机森林分类热点预测方法比时空核密度犯罪风险估算的方法识别热点中的高发区效率更高;而在较大范围较长时间周期上时空核密度犯罪风险估算的方法识别热点中的高发区更优。

长期预测主要服务于宏观政策的制定,而短期预测能为特定时间内的警力资源部署提供支持,对于犯罪防控均具有重要意义。不同警务策略的制定除了时间有效性考虑,还受到空间可及范围的制约,本文的犯罪热点预测对比结果对于长短周期的警务防控策略以及一定空间范围的警务管控均有一定的实践参考意义。但值得讨论的是,不同的研究区不同的犯罪案件类型的空间分布特征差异较大,采用何种原理的预测方法更好还难以直接定

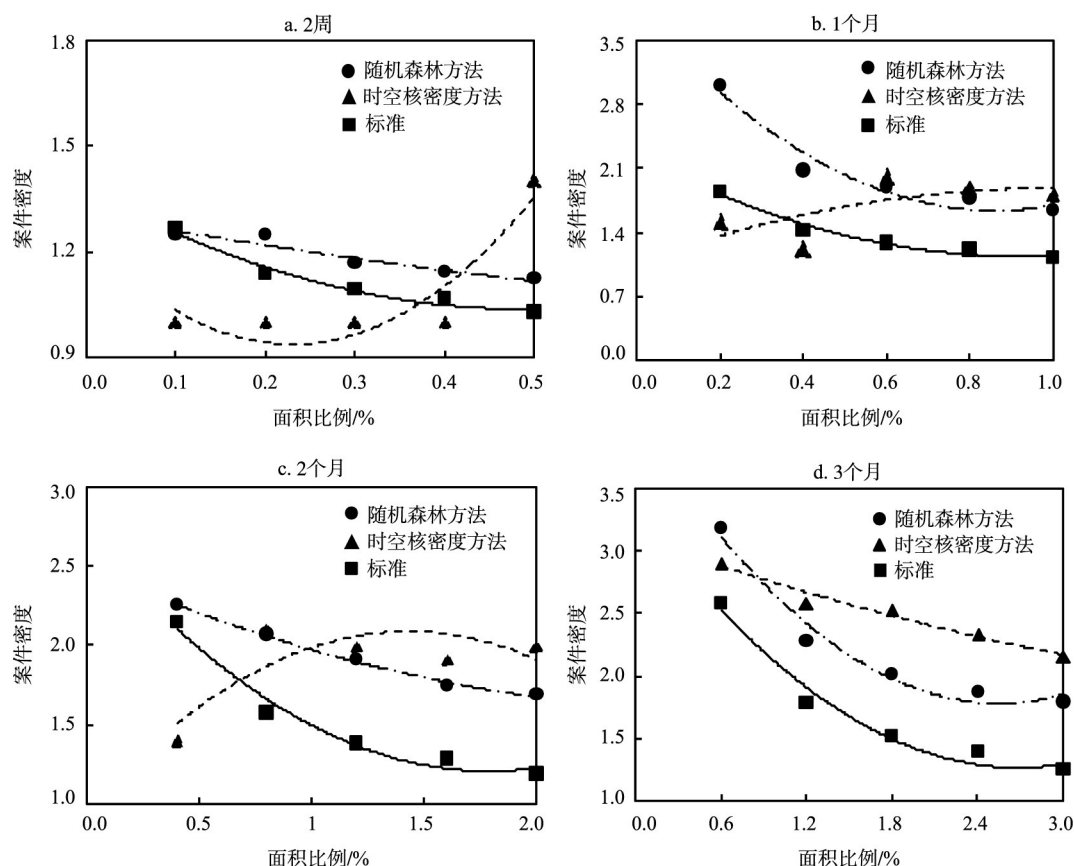


图8 各时间周期预测正确热点区域案件密度

Fig.8 Case density of the hit areas in different time periods

论。因此,本文的预测对比结果是否适用于其他研究区案件类型还需要采取进一步的实验验证,可能会在犯罪案件的发生及时空热点的形成规律研究等方面有新的发现。

参考文献(References)

- 陈鹏, 疏学明, 袁宏永, 等. 2011. 时空犯罪热点预测模型研究[J]. 系统仿真学报, 23(9): 1782-1786. [Chen P, Shu X M, Yuan H Y, et al. 2011. Research about spatial-temporal forecasting of crime hotspots[J]. Journal of System Simulation, 23(9): 1782-1786.]
- 李卫红, 闻磊, 陈业滨. 2017. 改进的GA-BP神经网络模型在财产犯罪预测中的应用[J]. 武汉大学学报: 信息科学版, 42(8): 1110-1116. [Li W H, Wen L, Chen Y B. 2017. Property crime forecast based on improved GA-BP Neural Network Model[J]. Geomatics and Information Science of Wuhan University, 42(8): 1110-1116.]
- 梁晓军, 高展. 2001. “十五”期间犯罪预测与侦查工作机制创新[J]. 公安研究, (7): 30-32. [Liang X J, Gao Z. 2001. "Shiwu" qijian fanzui yuce yu zhencha gongzuo jizhi ch-

uangxin[J]. Policing Studies, (7): 30-32.]

- 刘大千, 修春亮. 2012. 国内外犯罪地理学研究进展评析[J]. 人文地理, 27(2): 38-44. [Liu D Q, Xiu C L. 2012. Review of studies on criminal geography[J]. Human Geography, 27(2): 38-44.]
- 王发曾. 1992. 城市犯罪发展趋势预测[J]. 城市问题, (6): 15-18. [Wang F Z. 1992. Chengshi fanzui fazhan qushi yuce [J]. Urban Problems, (6): 15-18.]
- 徐冲, 柳林, 周素红. 2016. 基于临近相似性考虑的犯罪热点密度图预测准确性比较: 以DP半岛街头抢劫犯罪为例[J]. 地理科学, 36(1): 55-62. [Xu C, Liu L, Zhou S H. 2016. The comparison of predictive accuracy of crime hotspot density maps with the consideration of the near similarity: A case study of robberies at DP Peninsula[J]. Scientia Geographica Sinica, 36(1): 55-62.]
- 徐冲, 柳林, 周素红, 等. 2013. DP半岛街头抢劫犯罪案件热点时空模式[J]. 地理学报, 68(12): 1714-1723. [Xu C, Liu L, Zhou S H. 2013. The spatio-temporal patterns of street robbery in DP Peninsula[J]. Acta Geographica Sinica, 68(12): 1714-1723.]

- 阎耀军, 张明. 2013. 犯罪预测时空定位信息管理系统的构建[J]. 中国人民公安大学学报: 社会科学版, 29(4): 73-80.
- [Yan Y J, Zhang M. 2013. fanzui yuce shikong dingwei xinxi guanli xitong de goujian[J]. Journal of People's Public Security University of China: Social Sciences Edition, 29(4): 73-80.]
- Bogomolov A, Lepri B, Staiano J, et al. 2014. Once upon a crime: towards crime prediction from demographics and mobile data[C]//International conference on multimodal interaction. ACM, 427-434.
- Bowers K J, Johnson S D. 2005. Domestic burglary repeats and space-time clusters[J]. European Journal of Criminology, 2(1): 67-92.
- Bowers K J, Johnson S D, Pease K. 2004. Prospective hot-spotting: The future of crime mapping[J]. British Journal of Criminology, 44(5): 641-658.
- Brantingham P L, Brantingham P J. 1999. A theoretical model of crime hot spot generation[J]. Studies on Crime & Crime Prevention, 8(1): 7-26.
- Brantingham P L, Brantingham P J. 2008. The rules of crime pattern theory[M]. Devon, UK: Willan Publishing.
- Breiman L. 2001. Random forests[J]. Machine Learning, 45(1): 5-32.
- Caplan J M, Kennedy L W, Miller J. 2011. Risk terrain modeling: Brokering criminological theory and GIS methods for crime forecasting[J]. Justice Quarterly, 28(2): 360-381.
- Chainey S, Ratcliffe J. 2013. GIS and Crime Mapping[M]. London, UK: John Wiley & Sons.
- Chainey S, Reid S, Stuart N. 2002. When is a hotspot a hotspot? A procedure for creating statistically robust hotspot maps of crime[A]//Kidner D, Higgs G, White S. Innovations in GIS 9: Socio-economic applications of geographic information science. London, UK: Taylor and Francis: 21-36.
- Chainey S, Tompson L, Uhlig S. 2008. The utility of hotspot mapping for predicting spatial patterns of crime[J]. Security Journal, 21(1-2): 4-28.
- Clarke R V G, Eck J E. 2005. Crime analysis for problem solvers in 60 small steps[J]. Washington, DC: Center for Problem Oriented Policing.
- Clarke R V G, Felson M. 2004. Routine activity and rational choice[M]. New Brunswick, NJ: Transaction Publishers.
- Cohen L E, Felson M. 1979. Social change and crime rate trends: A routine activity approach[J]. American Sociological Review, 44(4): 588-608.
- Genuer R, Poggi J -M, Tuleau-Malot C. 2010. Variable selection using random forests[J]. Pattern Recognition Letters, 31(14): 2225-2236.
- Gorr W L, Harries R. 2003. Introduction to crime forecasting [J]. International Journal of Forecasting, 19(4): 551-555.
- Groff E R, Vigne N G L. 2002. Forecasting the future of predictive crime mapping[J]. Analysis for Crime Prevention, 29-57.
- Grubestic T H, Mack E A. 2008. Spatial-temporal interaction of urban crime[J]. Journal of Quantitative Criminology, 24(3): 285-306.
- Hirschfield A, Bowers K. 2001. Mapping and analyzing crime data: Lessons from research and practice[J]. Minerva Medica, 63(49): 2736-40.
- Kandaswamy K K, Chou K C, Martinetz T, et al. 2011. AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties[J]. Journal of Theoretical Biology, 270(1): 56-62.
- Kianmehr K, Alhajj R. 2008. Effectiveness of support vector machine for crime hot-spots prediction[J]. Applied Artificial Intelligence, 22(5): 433-458.
- Ratcliffe J H. 2004. Crime mapping and the training needs of law enforcement[J]. European Journal on Criminal Policy and Research, 10(1): 65-83.
- Ratcliffe J H. 2006. A temporal constraint theory to explain opportunity-based spatial offending patterns[J]. Journal of Research in Crime and Delinquency, 43(3): 261-291.
- Rodriguez Galiano V F, Ghimire B, Rogan J, et al. 2012. An assessment of the effectiveness of a random forest classifier for land-cover classification[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 67: 93-104.
- Rummens A, Hardyns W, Pauwels L. 2017. The use of predictive analysis in spatiotemporal crime forecasting: Building and testing a model in an urban context[J]. Applied Geography, 86: 255-261.
- Weisburd D. 2015. The law of crime concentration and the criminology of place[J]. Criminology, 53(2): 133-157.

Comparison of random forest algorithm and space-time kernel density mapping for crime hotspot prediction

LIU Lin^{1,2,3}, LIU Wenjuan¹, LIAO Weiwei¹, YU Hongjie¹, JIANG Chao¹,

LIN Rongping¹, JI Jiakai¹, ZHANG Zheng¹

(1. Center of Integrated Geographic Information Analysis, School of Geography and Planning, Sun Yat-Sen University, Guangzhou 510275, China; 2. Center of Geographic Information Analysis for Public Security, School of Geographic Sciences, Guangzhou University, Guangzhou 510006, China; 3. Department of Geography, University of Cincinnati, Cincinnati OH45221-0131, Ohio, USA)

Abstract: Crime prediction is of great significance for the formulation of police tactics and the implementation of crime prevention and control in different time periods. Machine learning and density mapping are two common approaches for crime hotspot prediction. However, there exists few published work that systematically compares the predicted results of these two approaches. This study aimed to fill the gap. With crime patterns uncovered from 2013 to May 2016, we predicted hot-spot distribution of theft crimes in the period of first two weeks of June, July, and August in 2016 by random forest algorithm and traditional space-time kernel density method and compared the two sets of predictions. The research area was divided into grid cells of 50 m×50 m. Each cell was predicted as either hot-spot or non-hot-spot area in the next predicting period. Then we overlaid the forecast results and location of real cases to evaluate the accuracy of the two methods. The results show that both the hit rate of area and cases of the random forest classification hot-spot prediction method are higher than that of the space-time kernel density within different periods. Both methods can effectively identify high-crime areas of crime hot spots in prediction. In a relatively short period of time and small area, the random forest classification hotspot prediction method is more effective than the space-time kernel density method. However, in a relatively long term and large area, the space-time kernel density crime risk estimation method yields better result in identifying high crime areas.

Key words: space-time kernel density; random forest algorithm; crime hotspot prediction; high crime areas identification