2019 年（第六届）全国大学生统计建模大赛

基于长短期记忆模型的入室盗窃犯罪预测研究

参赛单位:湖北经济学院

参赛者姓名:朱艳敏 王同俊 蔡黎

目　录

摘要..... I

Abstract..... II

一、引言.....1

（一）研究背景与意义.....1

（二）国内外研究进展.....1

（三）研究思路与方法.....4

二、数据的来源与预处理..... 5

（一）数据的来源.....5

（二）数据预处理.....5

三、数据的理解与展示..... 8

（一）基于时间维度的分析.....8

1.1 日期变化分析.....8

1.2 节假日变化分析.....9

1.3 天气变化分析.....9

（二）基于空间维度的分析.....10

2.1 基于聚类分析的案件分布模式识别.....10

2.2 基于行政区域划分的案件分布模式识别.....11

2.3 基于网格划分的案件分布模式识别.....13

四、基于 LSTM 模型的犯罪预测..... 14

（一）基于 LSTM 的犯罪预测模型构建..... 14

（二）犯罪预测模型结果与分析.....16

（三）模型的评估与比较.....20

五、结论与建议..... 20

参考文献.....22

表目录

表 1　110 接警数据相关变量.....5
表 2　不同天气情况下案件数量日均值.....9
表 3　最邻近指数法计算结果.....11
表 4　武汉市各区人口、生产总值、土地面积.....12
表 5　武汉市各区各行业 POI 数量.....13
表 6　BD、RD 数据集在工作日、休息日上案件发生区域数和数量的方差分析.....19
表 7 BD、RD 数据集在工作日、休息日上预测准确率方差分析19
表 8　武汉市入室盗窃模型预测结果比较表.....20

图目录

图 1　技术路线图.....4
图 2　接警数据预处理流程图.....7
图 3　2015 年至 2019 年案件数量日历热力图.....8
图 4　2015 年至 2019 年各星期入户盗窃案件数量.....9
图 5　各政区发生入室盗窃案件数量热力图.....11
图 6　武汉市入室盗窃案件 RD,BD 数据分布图.....13
图 7　空间相关度示意图.....15
图 8　时空相关条件下不同数据形式对应的犯罪预测模型.....16
图 9　RD-LSTM 模型在不同超参数取值下 MSE 变化.....18
图 10　RD-LSTM 模型不同失活率下 MSE 箱线图.....18
图 11　BD-LSTM、RD-LSTM 测试集上预测准确率.....19

摘　要

随着警务信息化的快速发展，警务大数据逐渐形成，这为利用警务大数据和统计建模方法助力智慧警务建设提供了可能。该背景下，利用报警数据开展犯罪预测成为了新的研究热点。本文首先从时间和空间两个维度对 2015 年 1 月 3 日至 2019 年 4 月 30 日共 1579 天武汉市 110 入室盗窃接警数据进行了统计分析。在此基础上，根据入户盗窃犯罪事件的时空分布特征，分别构建了二值化长短期记忆犯罪预测模型(BD-LSTM)和频数统计长短期记忆犯罪预测模型(RD-LSTM)。接下来，通过深度学习训练得到模型中迭代次数、回看天数、空间依赖度、失活率等参数的最佳性能值。最后，利用最优参数下的预测模型对武汉市每天案件发生概率以及案件发生数量进行了预测测试，并与其他模型进行了对比。结果表明，相比多种机器学习模型和自激点过程模型，本文构建的长短期记忆犯罪预测模型具有更好的预测精度和稳健性。

关键词：犯罪预测；警务大数据；入室盗窃；RNN；LSTM

Abstract

With the rapid development of police informationization, police big data has gradually formed, which provides the possibility of using police big data and statistical modeling methods to help smart police construction.In the background, the use of alarm data to carry out crime prediction has become a new research hotspot.This paper firstly analyzes the data of 110 burglary alarms in Wuhan City from January 3, 2015 to April 30, 2019, from two dimensions of time and space.On this basis, according to the spatial and temporal distribution characteristics of the crime of burglary, a binarized long short-term memory crime prediction model (BD-LSTM) and a frequency statistics long short-term memory crime prediction model (RD-LSTM) were constructed respectively.Then, through the deep learning training, the optimal performance values of the parameters such as iteration times, look-back days, spatial dependence and dropout rates are obtained.Finally, using the prediction model under optimal parameters, the probability of occurrence of daily cases in Wuhan and the number of cases were predicted and tested, and compared with other models.The results show that compared with the multi-machine learning model and the self-excited point process model, the long short-term memory crime prediction model constructed in this paper has better prediction accuracy and robustness.

Key words: crime prediction ;police big data ;burglary ; RNN ; LSTM

一、引言

（一）研究背景与意义

近年来，国内很多地区陆续出台了智慧产业、智慧城市建设等方面的政策和规划，力争在新一轮发展竞争中占据制高点。2017 年 9 月公安部推出《关于深入开展“大数据+网上督察”工作的意见》，要求到 2020 年底，建成基于公安云计算平台的全国公安机关警务督察一体化应用平台，相关运行机制进一步健全完善，警务督察部门的动态监督和预警预测能力进一步提升。这些政策的出台既全面推动了智慧城市、智慧警务在中国的建设发展，也进一步凸显了发展智慧警务的重要意义。

构建智慧警务是新一轮信息技术变革下的时代潮流。新的技术往往孕育着新的重大突破，大数据将成为公安战斗力生成的核心要素。拥有对海量数据占有、控制、分析、处理的主导权，将大数据优势转化为公安决策优势，继而转化为治安优势，对于整合警务资源、改造警务流程、创新警务模式、降低警务成本、实现警务效能的最优化具有重要推动作用。

该背景下，通过“事后”大数据分析提升警务系统战斗力得到了快速发展，但是“事先”研判却少有应用。而大数据真正的价值在于预测，因此犯罪预测作为公安系统“打、防、管、控”四个环节中“防”这一环节，意义格外重要。首先，预测是预防的基础。公安工作的中心已从严厉打击与处罚转移至犯罪预防方面。预防必须建立在对事物发生发展规律认识的基础之上。犯罪预测既是建立在对犯罪案件发生规律认识的基础上，是犯罪规律体系的重要组成部分，成为犯罪预防的基础[1]；其次，犯罪预测是制定防范对策的重要依据。各机构可通过历史的犯罪数据来对犯罪活动进行预测性建模，以便提前预测某地区的犯罪分布情况，这样就可以提前部署好警力，且有针对性的制定好防范措施；最后，犯罪预测是目标管理的依据。目标管理的核心是目标的设定。按照主观意识进行指标设定，往往会偏离实际，违背科学，结果则不尽人意。预测是以科学的手段为基础，减少了人为因素的干扰，有利于目标的实现[2]。因此，科学的进行犯罪预测极为重要，它是城市安全管理目标制定的依据。

（二）国内外研究进展

就国内来看，由于我国的犯罪学预测起步较晚，初期国内的犯罪预测研究多为经验预测，随着时间的发展，进入 21 世纪之后，慢慢才开始采用了数理分析的方法，主要有回归分析法、灰色系统理论分析法和最优组合分析法等[1]。如刘

小娟、高连生采用灰色系统理论的分析方法对刑事案件进行了动态分析[3]；杜益虹、刘世华则采用 Logistic 回归分析法来构建了犯罪嫌疑概率预测模型[4]；李明等采用了优化组合预测方法对犯罪量进行了动态预测[5]；另外也有利用自回归滑动平均（ARMA，Auto-Regressive and Moving Average Model）模型对我国财产类犯罪人数进行预测研究[6]，也有的通过马尔科夫链来进行民族地区的毒品犯罪预测研究[7]。虽然这些方法在当时都取得了不错的预测检验结果，但这些模型多偏于宏观上的预测，即主要是定性预测而非定量分析，预测结果对基层民警的指导意义不大。随着信息技术的不断发展，数据储存与数据共享技术也得到了不断的应用和发展，犯罪数据呈指数形式的增长。于是在大数据的时代背景下，为了使犯罪数据能够得到充分的利用，国内慢慢开始将机器学习的方法应用到了犯罪预测。具体的有利用支持向量机（SVM，Support Vector Machine）的方法对犯罪嫌疑人进行特征预测[8]，还有的运用随机森林分类器的方法来解决犯罪预测问题[9]；另外也有的利用神经网络来进行犯罪预测，如李卫红等将改进的 GA-BP 神经网络模型用于财产犯罪预测，并验证了 GA-BP（Genetic Algorithm-Back Propagation）模型的可靠性[10]；于红志等利用改进的模糊 BP（Back Propagation）神经网络进行犯罪预测，也取得了较好的预测准确率[11]。相比于传统的统计建模方法，机器学习的算法不仅能够实现犯罪活动的微观预测，并且其预测的准确率也较高。

就国外来看，国外初期对于犯罪预测的研究主要采取实证的方法，通过调查、数据收集、分析、归纳，得出重要的相关因子，从而揭示犯罪发生的规律。这种研究一般以影响犯罪的基本因子为基础，属于微观预测范畴，揭示了基本因子与犯罪形成的内在关联度[2]。由于国外研究起步较早，目前在这方面已有一定的经验，相关的参考文献较之于国内也更多一点。目前有较多的学者们利用移动网络数据和社交媒体数据来进行犯罪预测。比如 Andrey 等将移动网络上得到的汇总和匿名的人类行为数据与基本的人口统计信息相结合来预测欧洲大都市的犯罪热点[12]；另外也有专家们将核密度估计的方法与推特数据相结合来进行犯罪预测。比如 TAM 等将社会因素与推特上面的情绪因素相结合，并利用核密度估计（KDE，Kernel Density Estimation）的方法来进行犯罪预测[13]，该方法使得犯罪预测具有更大的意义。虽然核密度估计的这种预测方法其准确率较高，但该方法缺乏可移植性，即不能简单的推广到其他城市。另外在传统预测方法中，也有采用逻辑回归、时间序列以及自激点过程等模型来进行犯罪预测。比如 Antolos 等应用逻辑回归模型来检验犯罪预测因素与盗窃事件发生概率之间的关系，虽然检验结果比较显著，但该方法的局限性在于很难确定盗窃活动和具体地点的可能性[14]；Shrivastav 等应用模糊时间序列，以发现社区的犯罪模式。但是此方法仅适用于二进制事务数据，例如 0 或 1 这样的数据[15]；MOHLER 等则将地震学中的

自激点过程（SEPP，Self-Exciting Point Process）应用到了洛杉矶入室盗窃事件的预测当中，发现该方法对于犯罪预测也同样适用[16]。上述方法中，其中大多数都是利用一些传统的统计建模方法来进行犯罪预测，虽然在当时的情况下有着较高的预测准确率，但是随着时间的发展，这些方法也慢慢的出现了一些局限性。比如难以发现高度非线性关系，冗余和多个数据集之间的依赖关系，且传统的统计建模方法不能处理非结构化数据等等。于是在这样的情况下，国外的一些学者便开始利用机器学习和深度学习的方法来进行犯罪预测。比如 Kianmehr 等利用支持向量机进行热点预测，以预先确定犯罪率水平并给出数据的百分比。然而此方法的计算速度慢且计算费用高[17]；Nasridinov 等利用决策树帮助警方发现犯罪模式并预测未来的犯罪趋势[18]。随着时间的发展，神经网络慢慢也被运用到犯罪预测中。如 Chitsazan 等利用人工神经网络（ANN，Artificial Neural Network）来预测犯罪，该方法的优点是模型预测的准确率较高，但是模型训练的时间较长[19]；Wang 等人应用深度时空残差网络（ST-ResNet，Spatio-Temporal Residual Network）和数据增强技术在精细空间尺度下进行了实时犯罪预测[20]。总的来说，相比较于传统的统计建模方法来说，机器学习和深度学习的方法在犯罪预测方面有着更广泛的适用性和更高的预测准确率。

上述研究从诸多视角进行了犯罪分析和预测，但对于国内犯罪预测而言，依然存在一些不足之处。一是已有文献大多使用国外数据或国内的宏观统计数据，国内微观数据相对较少。二是使用的方法大多是基于传统的机器学习方法或统计学，较少使用深度学习。三是研究视角大多属于长期、宏观层面，犯罪短期研究还需进一步丰富与改进。

针对以上不足，本文以微观数据 110 接警数据为基础，使用深度学习算法进行犯罪事件的预测，进一步探索适合国内环境的犯罪预测模型。采用深度学习算法而不是传统机器学习算法主要基于以下两点考虑：首先，110 接警数据库的数据规模庞大，相比传统的机器学习算法，深度学习的预测效果会随着测试数据集规模的扩大而提升，从而能够充分发挥深度神经网络因深度和广度的大量增加带来的预测能力。深度学习算法在处理大规模 110 接警数据时具有更强的泛化能力，即训练模型的预测效果比传统机器学习更好；其次，犯罪事件具有时间和空间上的关联性，这种关联性在学习算法中表现为对数据时空特征的记忆性。深度学习之所以比传统机器学习具有更强的泛化能力，就是因为其良好的记忆能力，尤其长短期记忆模型(Long short-term memory , LSTM)算法[21]能够有效记住犯罪事件在发生时间和空间上的长期或短期特征，这是传统机器学习算法所不具备的。

本文先结合研究目的及研究方法的科学性、可行性原则构建犯罪预测体系；

通过时间和空间维度对已有数据进行描述统计分析，再将时空数据特性与神经网络模型相结合，构建基于长短期记忆模型的犯罪预测模型；最后，基于湖北省武汉市 2015 至 2019 年的 110 接警数据对文中提出的犯罪预测模型进行验证。结果表明，该模型可靠性强，能够充分利用犯罪事件间的时空特征，有效、充分地提取事件在空间上的邻近性、距离性、层次性，时间上的邻近性、趋势性。有助于基层警务人员合理安排警务资源、科学采取安保措施，及时、有效的将犯罪事件扼杀在摇篮中。

（三）研究思路与方法

本文的研究思路以接警数据的案件类别、时空属性、地理编码信息为基础，利用神经网络考虑犯罪事件的长短期记忆、并发性以及空间关系，构建科学、有效的犯罪预测体系。犯罪预测的技术路线(见图 1)由四部分组成。第一部分：数据获取。从 110 接警平台查询并提取相应的接警数据主要包括案件类别信息、案件时空信息、相应行政区域的地理编码，并对部分涉密信息进行脱敏处理。第二部分：数据预处理。首先，包括案件类别的热点编码、案件时空信息的去重和补全、相关行政区域地图的网格划分；接下来依次对案件进行空间聚类、时间分割，形成时空维度的犯罪数据窗口。第三部分：构建预测模型。根据输入数据的不同种类进行 LSTM 模型的设计、训练以及最优模型的选取，最终动态选取自适应阈值，输出相应预测结果。第四部分：模型的评估与比较。比较了本文构建的 LSTM 犯罪预测模型与机器学习、自激点过程等模型在犯罪预测准确率方面的优势。

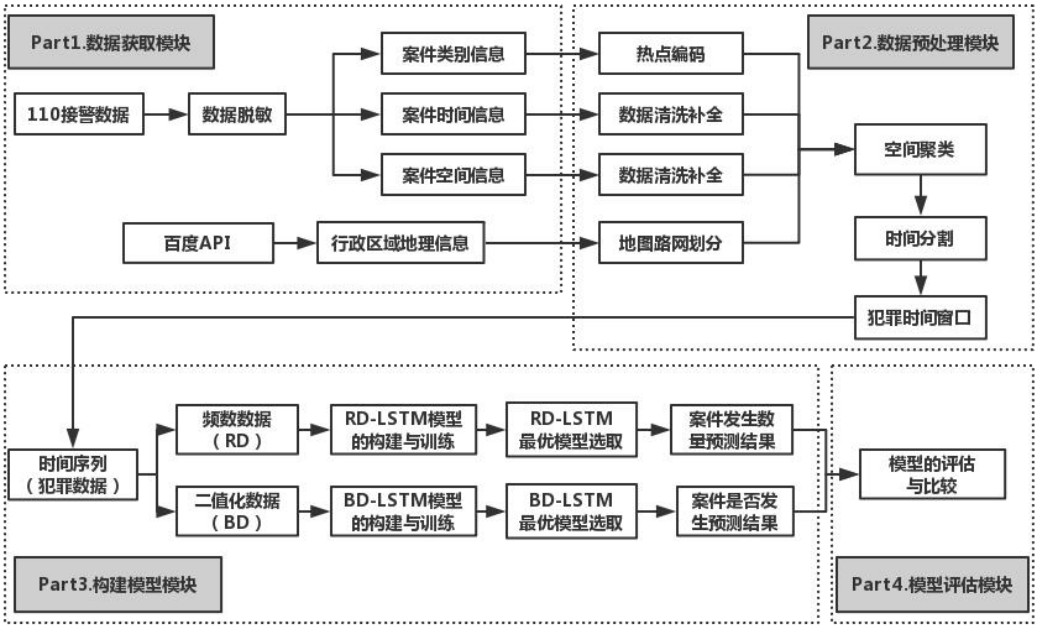

图 1　技术路线图

二、数据的来源与预处理

（一）数据的来源

本文的数据来源为武汉市公安局大数据实战应用中心 110 接警平台，数据包含 2015 年 1 月 3 日至 2019 年 4 月 30 日共 1579 天，原始数据接警数据包含报案人、接警单位、案件等相关信息。首先针对其中涉及隐私的相关信息进行了脱敏处理，然后从中提取了与案件相关的时间、空间、案件类别等相关信息。本文使用了接警数据中的部分相关变量，如表 1 所示。另外，相关行政区域地图的编码信息通过百度地图的 API 获取，主要包括地图以及行政区域的边界经纬度信息。

表 1　110 接警数据相关变量

属性	描述	数据形式
CaseType	案件类别	String
Latitude	地理坐标纬度	Float
Longitude	地理坐标经度	Float
Location	案发地点	String
DateTime	报案时间（精确到分）	Date

（二）数据预处理

数据预处理数据挖掘和建模分析必不可少的先行步骤。数据预处理的质量会直接决定数据分析与预测的准确性。本文主要是利用接警平台的时空大数据和百度地图地理信息数据，采用 LSTM 模型对指定地理范围内指定时间的指定犯罪种类进行预测。结合研究目的和算法模型对数据质量以及输入格式的要求，需对已有数据进行科学、合理的预处理。具体的预处理包括以下五个方面：第一，考虑到不同案件类型的时空异质性，本文对不同种类的案件分类进行预测，因此，需提前对案件进行类别的独热编码；第二，接警平台中经常存在报案人重复报警，偶尔也会出现案件时空信息缺失的情形，因此，需对原始数据进行去重和补全；第三，过大研究范围得到的预测结果对于基层警务人员实际工作的开展意义不大，因此，预测范围不是省、市、区、县等大的行政区域，而是对相应研究区域进行一定大小的网格划分，以小区、街道等更小的区域作为研究单位；第四，案件空间信息是精确的经纬度坐标数据即坐标点信息，研究的地理范围是一定经纬度区间的面信息，因此，需通过空间聚类将每个案件映射到对应的地理网格块之中；第五，案件时间信息的处理与空间信息处理类似，由于每个案件的时间信息是精确到分钟的，而案件预测是针对某天、某周或某月等某个时间段而言的，因此，需要对案件的时间信息做聚集处理，然后按照犯罪预测的相应时间单位进行

分割及合并。整体数据预处理过程如图 2 所示。下面分别对相应预处理过程进行详细说明。

第一，案件类别信息独热编码。独热编码，又称为 One-Hot 编码或一位有效编码，主要是通过 N 位状态寄存器来对 N 种状态编码，每个状态对应一个独立的寄存器位，并且每次只有一位有效。独热编码多用于分类变量的表示，先将分类值映射到整数值，再将每个整数值表示为二进制向量。其中，整数的索引记为 1，其它均为 0。编码后，离散特征被数字化，不仅便于特征相似性的度量、距离的计算，同时便于特征的分离及合并。如，案件类别=[“入室盗窃”，“涉毒”，“扒窃”，“扰乱秩序”，“打架斗殴”，“抢劫”]，此处 N=6，则“入室盗窃”=>[1,0,0,0,0,0]，“打架斗殴”=>[0,0,0,0,1,0]。

第二，时空信息去重、补全。去重主要是针对重复报警信息进行过滤，补全主要是补全缺乏经纬度信息的案件记录。数据清洗过程中，首先，利用百度 API 采用爬虫技术对其进行信息补全；然后，对重复多次报警以及跨区域报警数据进行清洗过滤。最终保留的案件属性变量包括案件类别、报案时间、报案地点（纬度、经度、地名）。

第三，地图网格划分。事件通过经纬度信息与指定区域在地图上的经纬度信息对应，反应事件的发生位置。针对某市、某区或某县这样的行政区域进行犯罪事件的预测对于基层警务人员工作安排没有实质性的参考意义，而对每个指定大小的小区域在指定时段的犯罪情况做预测，能够为对应街道或是小区的相关工作人员的巡逻路线或是人员分配提供参考。因此，需要先将目标区域划分为一定面积的小网格块。以前的研究中很多直接采用经纬度等分后的地图网格进行后期的研究，如：将区域 S(lng_{min},lat_{min},lng_{max},lat_{max})分为 500*500 的网格，则分别用 Δlng 和 Δlat 对区域 S 进行分割，则 $\Delta lng = (lng_{max} - lng_{min})/500$，$\Delta lat = (lat_{max} - lat_{min})/500$，这种分割方式在经纬度上实现了等间隔分割。但经纬度等间隔划分后其对应的球面区域面积 S 可能存在很大差异，因为 A(lat_A,lng_A),B(lat_B,lng_B)两点间球面距离 d (单位:m)为:

$$d = 2\pi R * \arcsin\left(\frac{1}{2}\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2} * \frac{1}{180}\right) * 1000 \qquad (1)$$

其中，
$$x_1 = \cos lat_A * \cos \ln g_A \; ; \; x_2 = \cos lat_B * \cos \ln g_B$$
$$y_1 = \cos lat_A * \sin \ln g_A \; ; \; y_2 = \cos lat_B * \sin \ln g_B$$
$$z_1 = \sin lat_A \; ; \; z_2 = \sin lat_B$$

可看出球面距离 d 与两点间经纬度间隔并非简单的线性关系，因此直接均分

经纬度这种划分网格的方式会形成面积差异很大的网格。本文针对每个网格设定与其面积 s 成正比的权重 w。其他条件相同时，网格面积越大对应的预测准确率会越高，难以与不同方法对应不同大小网格的预测准确率进行对比。因此，本文对经纬度实现等间隔分割后，得到 $R \times C$ 网格数。由于地球近似为球体，等间隔划分经纬度后的网格投射到平面近似为等腰梯形。因此，对于网格g_i利用两经纬度间的距离公式计算出对应等腰梯形的四边边长分别为d_1、d_2、d_3、d_4，则网格g_i对应的面积s_i为：

$$s_i = \frac{1}{2}(d_1 + d_3)[d_2^2 - (\frac{1}{2}|d_1 - d_3|)^2]^{\frac{1}{2}} \quad (2)$$

其中$d_2 \approx d_4$,则$w_{s_i} = \mu s_i$。与其他不同大小网格对比时，以参照网格的面积大小s_0为单位，对本文网格大小设置相应的权重，实现准确率的可比性。

第四，空间聚类。结合已分割好的地图网格g_1,g_2,...,g_n分别将原始数据对应到相应网格中，实现原始数据的空间聚类。由于空间具有距离性、邻近性、层次性，因此可结合实际研究需要，对原始数据进行二次空间聚类。

第五，时间分割。时间分割过程是为 LSTM 模型输入数据做准备。假设犯罪预测模型是针对指定区域每天的犯罪情况做预测，则时间维度上需要将原始数据按天进行分割并对相同网格内的案件记录数进行合并。得到二值分类数据(BD)和频数回归数据(RD)，并分别利用 BD-LSTM 和 RD-LSTM 进行犯罪预测。

图 2　接警数据预处理流程图

三、数据的理解与展示

已有研究表明，犯罪案件的发生往往与空间[30]和时间[31]具有密切的关系，因此本文首先针对清洗过滤后得到 83952 条入室盗窃案件信息，分别基于时间和空间维度来对数据进行了理解和分析，为下一步构建符合犯罪事件时空特征的预测模型奠定基础。

（一）基于时间维度的分析

首先对入室盗窃案件在时间尺度上寻找规律，统计得出入室盗窃案件在各个细分的时间尺度上特点。以日期、节假日、雨雪天等时间尺度进行细分，深入剖析入室盗窃案件在不同时间尺度上的分布规律，找出案件发生的热点时间段，并对相关结论进行说明解释。

1.1 日期变化分析

我们按照日期的维度统计整理 2015 年 1 月 3 日至 2019 年 4 月 30 日武汉市每天入户盗窃案件数量，并做出日历热力图（见图 3），其中良表示发生案件数量为 0~30 起，轻度表示发生案件数量为 30~60 起，发生案件数量为 60~90 起显示为中度，90~120 起显示为重度。直观观察来看 2015 年主要为红色和橙色，即案件发生数量主要为重度和中度，2016 年主要为橙色和黄色，即案件发生数量主要为中度和轻度，2016 年主要为黄色，即案件发生数量主要为轻度，2018 年至 2019 年 4 月主要为黄色和蓝色，即案件发生数量主要为轻度和良。说明从总体趋势上看，武汉市的入户盗窃数量是下降的，即逐渐减少。

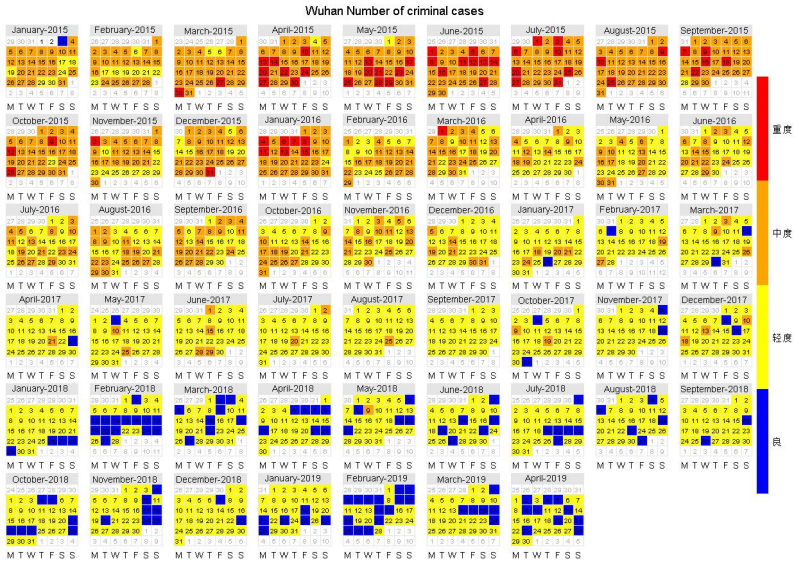

图 3　2015 年至 2019 年案件数量日历热力图

1.2 节假日变化分析

为了能提取出更准确时间尺度影响因素，以星期为时间尺度对案件的数量进行统计。从图 4 可以看出，对于入室盗窃案件，星期一案发数量最高，星期六和星期日的案发数量最少，其余的星期案件数量变化不大。因为一般周末人们在家休息，工作日外出工作，所以在工作日犯罪分子比较便于作案，数据显示情况也与实际情况像吻合。

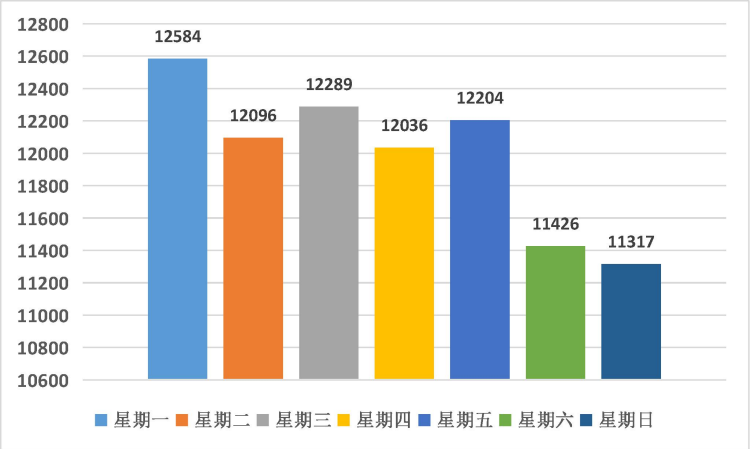

图 4　2015 年至 2019 年各星期入户盗窃案件数量

我们又进一步考虑了我国法定节假日的信息，2015 年 1 月 3 日至 2019 年 4 月 30 日一共包含 499 个休息日和 1080 个工作日，将入室盗窃案件的发生时间按照工作日、休息日打上标签（工作日标为 1，休息日标为 0）。其中，工作日包括国家规定的周末补班时间，不包括周一至周五的国家法定节假日；同样，休息日不包括国家规定的周末补班时间，包括周一至周五的国家法定节假日。统计得出工作日的案件数量日均值为 55 件，休息日的案件数量日均值为 50 件，工作日比休息日的案件数量日均值高，同上述星期的结论一致。

1.3 天气变化分析

我们爬取了中国天气网上武汉市的历史天气信息，并将其分类打三种标签（非雨非雪标记为 0，雨天标记为 1，雪及雨雪天气标记为 2）。统计得出非雨非雪天气的案件数量日均值为 55 件，下雨天的案件数量日均值为 51 件，雪及雨雪天气的案件数量日均值为 45 件。

表 2　不同天气情况下案件数量日均值

天气	案件数量日均值
非雨非雪天	55
雨天	51
雪及雨雪天	45

由上表得出在不同天气情况下，非雨非雪天气发生案件的数量最高，雪及雨雪天气发生案件数量最低，因为雪及雨雪天气，人们外出概率大大降低，犯罪机会也相应降低。

（二）基于空间维度的分析

按照数据的来源，犯罪数据可分为通过区域统计方法获得的面数据和采用GPS 等方式精确定位的点数据[21]。本文中使用的数据是精确定位的点数据。对犯罪区位论等相关理论成果研究可知，犯罪案件在空间上并不是常常是随机分布的，而是在某些区域有明显的集中性并呈现出一定的规律性。不同的地理位置以及周边环境会影响到人口密度的分布，从而影响到犯罪案件的发生。如商场、小区等人口密度大、人类活动频繁的场所将更有利于犯罪分子实施犯罪。对于犯罪案件时空特征的研究可以预防犯罪行为的发生，达到降低犯罪率的目的。

2.1 基于聚类分析的案件分布模式识别

最邻近指数法是计算与当前事件点最近点的距离来描述分布模式。该方法首先假设案件在空间上是随机分布无聚集现象，通过计算与最邻近的点之间的距离，与期望的距离相比较，判断事件点之间是否存在聚集现象[22]。当案件点的分布均匀，则最邻近距离均值与期望值之比大于 1；点的分布随机时此比值等于 1；点分布聚集时此比值小于 1，且此值越小则表示聚集程度越高。步骤如下：

（1）计算每两个最近案件点之间的距离，并求其平均值：

$$\bar{d}_{\min} = \frac{1}{n}\sum_{i=1}^{n} d_{\min}(p_i)$$

其中，n 是研究范围内点的数量，p_i 表示区域内的事件点。

（2）求出最邻近距离期望值：

$$E(d) = \frac{1}{2\sqrt{n/A}}$$

（3）计算最邻近距离与期望值的比值，得出最邻近指数指标值 R:

$$R = \frac{\bar{d}_{\min}}{E(d)}$$

入室盗窃案件在空间上的分布模式是聚集的还是随机的，还需要进一步在案件的空间维度进行分析。用最邻近指数法检验入室盗窃案件在空间维度上是否存

在聚集现象以及是否存在犯罪热点。最终计算结果如表 3 所示：

表 3　最邻近指数法计算结果

最邻近值 R	检验值	P 值
0.3008	-52.686	0.0006

检验结果中 P 值为 0，小于 0.05，拒绝空间分布非聚集的假设，R 值为 0.3008 小于 1，检验结果表明入室盗窃案件在空间上是呈现聚分布的。P 值以及 R 值的值都很小，说明研究区域中的入室盗窃案件确实存在空间聚集现象，并且聚集程度较高。

2.2 基于行政区域划分的案件分布模式识别

本文的研究区域包括武汉市及其市下的十三个辖区（江岸区、江汉区、硚口区、汉阳区、武昌区、洪山区、青山区、东西湖区、蔡甸区、江夏区、黄陂区、新洲区、汉南区）。我们根据犯罪案件的经纬度信息利用 python 代码将案件归属于不同的政区，统计出武汉市各行政区域在我们研究的历史期间总共发生过入室盗窃案件的数量。

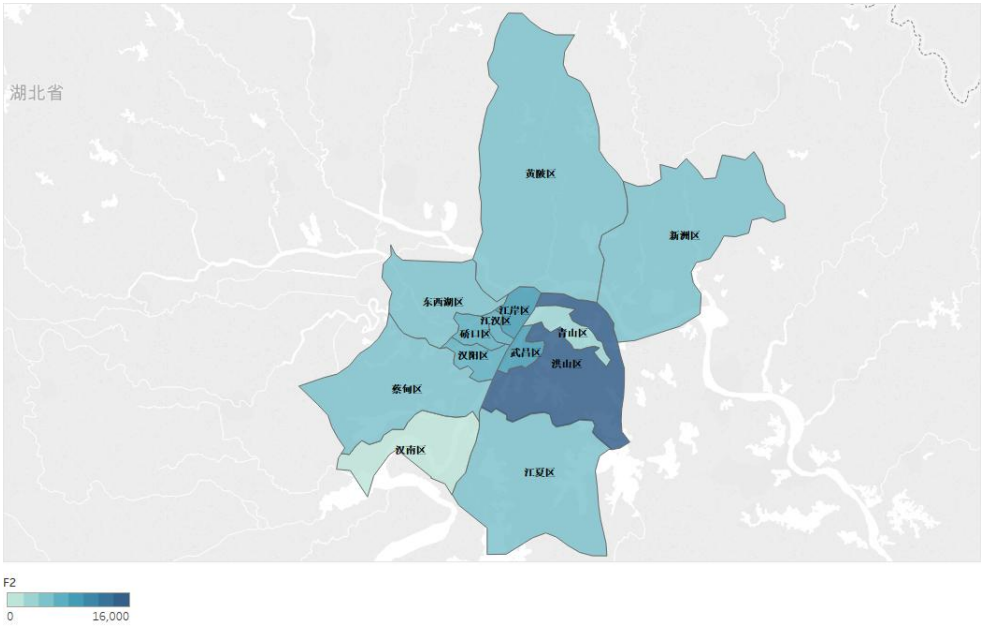

图 5　各政区发生入室盗窃案件数量热力图

图 5 为武汉市入室盗窃类案件案发地点所在政区的热力图。从图中可以看出洪山区的颜色最深，即该区域的入室盗窃犯罪案件数量最多、其次是武昌区、江岸区、硚口区，边远城区汉南区、青山区颜色最浅，即该区域犯罪案件数量最少，从而可以看出案件数量与地理位置有很大的关系。

为了进一步探索各个行政区域发生案件数量差异悬殊的原因，我们又做了两项工作，一是查询最新的武汉市 2018 年统计年鉴，得到武汉市武汉市各区人口总数、人口密度、生产总值、土地面积数据，二是爬取了武汉市全市的 POI（point of information）信息。

表 4　武汉市各区人口、生产总值、土地面积

地区	常住人口(万人）	人口密度（人/平方公里）	生产总值（百万元）	土地面积（平方公里）
全市	1089.29	1271	9692.74	8569.15
洪山区	163.75	2851	957.73	573.28
武昌区	127.63	19793	1102.5	64.58
江岸区	96.24	11988	1073.6	80.28
硚口区	86.85	21679	668.07	40.06
江汉区	72.96	25790	1142.58	28.29
汉阳区	65.27	5898	948.31	111.54
东西湖区	56.25	1133	730.63	495.34
蔡甸区	73.5	667	397.65	1093.17
江夏区	91.37	453	770.98	2018.31
黄陂区	98.83	437	702.49	2256.7
新洲区	90.21	616	676.32	1463.43
汉南区	13.55	467	----	287.05
青山区	52.88	9349	521.88	57.12

数据来源：2018 年武汉市统计年鉴

从上表 4 可以看出洪山区、武昌区、江岸区、硚口区、江汉区常住人口总数较多且人口密度也较大（洪山区人口密度小是因为其土地面积大，学生人数多），生产总值最高的为江汉区，其次是武昌区、江岸区、洪山区，而汉南区、青山区与上述区域相反，常住人口较少且生产总值较低，从而可以看出发生入户盗窃案件的数量与人口总数、人口密度、生产总值等有较为密切的关系。

在地图表达中，一个 POI（point of information）可代表一栋大厦、一家商铺、一处景点等等，通过 POI 搜索，完成找餐馆、找景点、找大学等等的功能。我们首先爬取武汉市全市的 POI 数据，接着将爬取下来的 POI 数据首先进行去重，最后将武汉 13 个区的 18 个行业 POI 分别存放，并统计出各个区的各行业 POI 个数（见表 5）。

表 5　武汉市各区各行业 POI 数量

POI / 政区	洪山区	武昌区	江岸区	硚口区	汉阳区	江汉区	江夏区	黄陂区	东西湖区	蔡甸区	新洲区	青山区	汉南区
公交站	907	330	358	216	398	247	764	1278	883	753	504	214	23
公司	12478	5889	5652	5196	4102	6212	4202	3642	5545	4029	1668	2048	49
地铁站	38	22	32	16	22	16	12	11	16	6	6	0	0
幼儿园	361	126	141	85	149	91	181	91	131	130	31	83	1
政府机构	2499	2108	1739	1198	990	1374	1417	1974	1008	1478	1270	820	78
百货商场	12	15	10	6	9	8	5	6	1	3	5	5	0
网吧	289	112	102	73	74	67	119	56	72	71	42	40	0
购物中心	47	45	20	20	26	40	15	30	13	10	38	7	0
酒吧	81	51	63	9	7	40	13	6	5	10	5	7	0
酒店	1754	813	441	306	366	633	618	464	302	363	356	150	4
长途汽车站	4	10	2	4	1	2	3	15	3	6	9	3	2
中学	77	43	51	27	30	35	42	53	26	47	50	28	1
住宅区	1514	1073	1051	612	661	739	512	470	443	516	362	361	6
写字楼	793	472	405	295	146	454	218	105	129	125	64	83	0
大学	73	17	3	5	6	0	34	3	5	5	6	3	0
小学	88	72	51	49	37	36	72	153	41	81	97	29	4
火车站	1	1	0	0	0	1	0	0	0	0	0	0	0
飞机场	0	0	0	0	0	0	0	1	0	0	0	0	0
总计	21016	11199	10121	8117	7024	9995	8227	8358	8623	7633	4513	3881	168

从上表可以看出洪山区、武昌区、江岸区 POI 总数较多，其中住宅区、大学、百货商场、公交站等远多于其他政区，这些地方人口较为聚集、流动性强，所以犯罪案件高发且多发。而青山区、汉南区这些政区人口流动性较弱，所以犯罪案件数量低于其他区域。

2.3 基于网格划分的案件分布模式识别

我们根据犯罪案件的经纬度信息利用 python 代码将案件归属于不同的网格，统计出 2015 年 1 月 3 日至 2019 年 4 月 30 日期间武汉市各网格总共发生过入室盗窃案件的频数。又进一步将数据二值化，即发生过案件的网格显示为 1，未发生过案件的网格显示为 0。

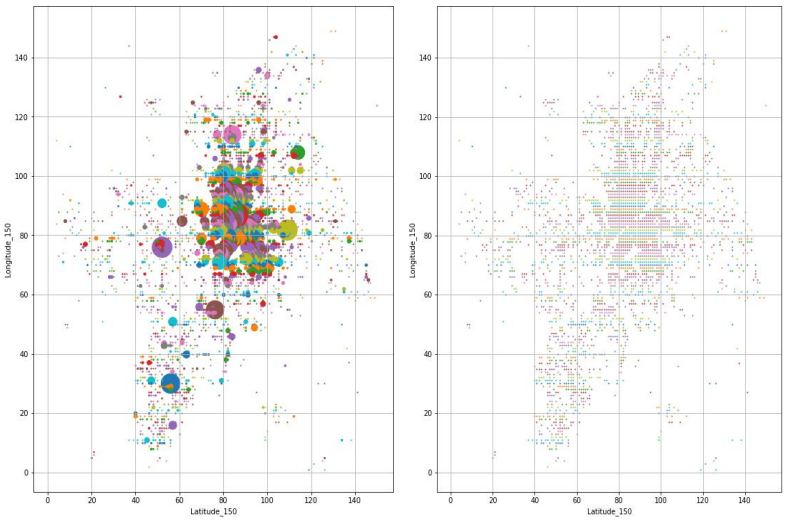

图 6　武汉市入室盗窃案件 RD,BD 数据分布图

图 6 分别基于回归数据和二值数据显示了 2015 年 1 月 3 日至 2019 年 4 月 30 日期间，武汉市各网格总共发生过多少起入室盗窃案件以及武汉市发生过入室盗窃案件的网格区域。不同散点大小表示不同数值大小。

四、基于 LSTM 模型的犯罪预测

犯罪预测的目的是利用地理信息及 110 接警的时空数据尽量提前准确预测潜在的违法犯罪行为，其本质是时空大数据的预测问题。国内微观警务数据不易获得，因此，相关研究较少，且大都通过传统机器学习方法针对时空特征进行分析，未能更进一步更精确的预测未来短时间的犯罪趋势。针对某地区小范围短时间的犯罪预测，实际是针对量大、稀疏且正负样本极不平衡的历史数据进行犯罪预测。传统机器学习方法对于解决高维、稀疏、倾斜数据效果不佳。近年来，计算机性能的提升，大数据时代的来临，深层神经网络模型强大的学习能力能够很好的解决上述问题，并且具有良好的学习效率和泛化能力。循环神经网络(Recurrent Neural Network ,RNN)是一种改进的多层感知器网络，用于处理序列数据。但犯罪事件通常呈现近期重复，即存在时空范围上的依赖，RNN 对于长期序列依赖会出现梯度消失问题，因此，不适合用于犯罪预测研究。LSTM 是在 RNN 基础上 1997 年由 Hochreiter 等提出[24]，2014 年由 Alex Graves 等进行改进的一种循环神经网络[25]。LSTM 通过确定新的输入是否被存储、遗忘或作为输出存储在记忆单元中，可以学习序列数据间的长短期依赖信息。

根据第 3 节的分析，入户盗窃事件与时间维度和空间维度都有相关性，时空数据的后期输出与前期的输入、输出相关，即输出依赖于输入及前期“记忆”。因此，本文主要运用 LSTM 构建犯罪预测模型。同时结合时间序列预测模型选取适当的阈值输出可读、易用的预测结果。

（一）基于 LSTM 的犯罪预测模型构建

假设已发生的犯罪事件在临近时间(以“天”为单位)及空间上均彼此相关，根据不同的输入数据类型，本文将创建两种形式的预测模型：BD-LSTM、RD-LSTM。设时间序列长度为 T,地图被分为 R 行 C 列，网格总数为 G。考虑空间相关，假设其与周边 q 环网格对应的犯罪数量相关，则其与周边的$(2q+1)^2$-1 个网格相关，图 7(a)(b)(c)中阴影网格分别表示 q=1,2,3 时与网格 g 相关的邻近区域。若周边近邻网格所在行 $i<0$ 或 $i>R$ 或者所在列 $j<0$ 或 $j>C$，均视为越界，对应$x_{tg}=0$,$g=i\times j$。

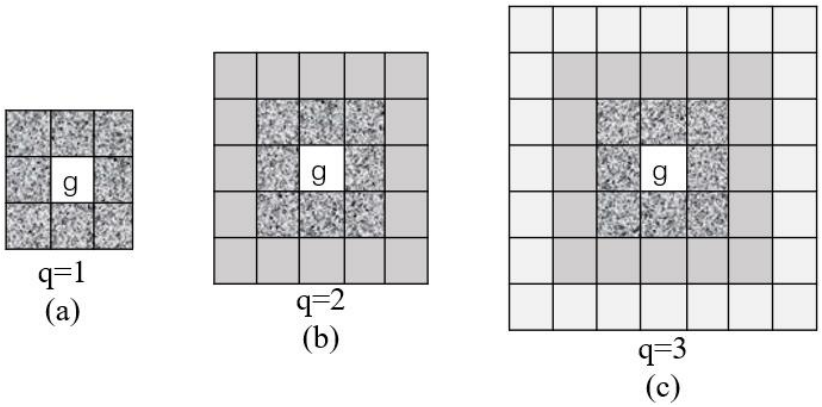

图 7　空间相关度示意图

假设 q=1 即与周边 8 个近邻网格区域相关，则v_{tg}^{B}为第 t 天,格子编号 $g = r * c$ 对应的近邻网格G_g^1= $[g,g_1,g_2,g_3,g_4,g_5,g_6,g_7,g_8]$犯罪事件二值化向量:

$$v_{tg}^{B} = [I(x_{tg}),I(x_{tg_1}),I(x_{tg_2}),I(x_{tg_3}),I(x_{tg_4}),I(x_{tg_5}),I(x_{tg_6}),I(x_{tg_7}),I(x_{tg_8})]^T$$

且 $I(x_{(tg_i)}) = \begin{cases} 1，当 x_{(tg_i)} \geq 1 \\ 0，当 x_{(tg_i)} = 0 \end{cases}$，$x_{(tg_i)}$表示地图网格第 r 行第 c 列对应区域在第 t 天发生的案件数量。v_{tg}^{R}为第 t 天,格子编号 $g = r * c$ 对应的犯罪事件加权频数统计向量:

$$v_{tg}^{R} = [w_{s_0}x_{tg},w_{s_1}x_{tg_1},w_{s_2}x_{tg_2},w_{s_3}x_{tg_3},w_{s_4}x_{tg_4},w_{s_5}x_{tg_5},w_{s_6}x_{tg_6},w_{s_7}x_{tg_7},w_{s_8}x_{tg_8}]^T$$

考虑时间相关，假设第 t+1 天的案件数量与之前 b 天的犯罪数量相关，称 b 为回看天数。第 t+1 天,格子编号 $g = r * c$ 对应的犯罪事件二值化矩阵样本$m_{t+1,g}^{B}$,标签值$y_{t+1,g}^{B}$分别为：

$$m_{t+1,g}^{B} = \left[v_{tg}^{B},v_{t-1,g}^{B},v_{t-2,g}^{B},\ldots,v_{t-b+1,g}^{B}\right]^T$$

$$y_{t+1,g}^{B} = [v_{t+1,g}^{B}]^T$$

第 t+1 天,格子编号 $g = r * c$ 对应的犯罪事件频数统计矩阵样本$m_{t+1,g}^{R}$,标签值$y_{t+1,g}^{R}$分别为：

$$m_{t+1,g}^{R} = \left[v_{tg}^{R},v_{t-1,g}^{R},v_{t-2,g}^{R},\ldots,v_{t-b+1,g}^{R}\right]^T$$

$$y_{t+1,g}^{R} = [v_{t+1,g}^{R}]^{T}$$

综合所有接警数据的时空信息，犯罪事件的二值化数据矩阵M^B及频数统计数据矩阵M^R分别为：

$$M^{B} = \begin{bmatrix} [m_{11}^{B}, m_{21}^{B}, \cdots, m_{T-b,1}^{B}] \\ [m_{12}^{B}, m_{22}^{B}, \cdots, m_{T-b,2}^{B}] \\ \cdots \\ [m_{1G}^{B}, m_{2G}^{B}, \cdots, m_{T-b,G}^{B}] \end{bmatrix}_{G\times(T-b)}, \quad M^{R} = \begin{bmatrix} [m_{11}^{R}, m_{21}^{R}, \cdots, m_{T-b,1}^{R}] \\ [m_{12}^{R}, m_{22}^{R}, \cdots, m_{T-b,2}^{R}] \\ \cdots \\ [m_{1G}^{R}, m_{2G}^{R}, \cdots, m_{T-b,G}^{R}] \end{bmatrix}_{G\times(T-b)}$$

假设回看长度为 b，空间相关度 q=1，本文提出基于 BD-LSTM、RD-LSTM 的犯罪预测模型如图 8 所示。本文所使用的 BD-LSTM、RD-LSTM 模型均由一个输入层、一个隐藏层和一个输出层组成。LSTM 模型中的权重在[0,1]间随机初始化。利用反向传播计算在学习过程中的每个阶段更新权重，不断优化网络产生的输出。不同之处在于，BD-LSTM 以M^B作为输入数据，二值交叉熵(binary_crossentropy)作为损失函数，Softmax 作为激活函数，而 RD-LSTM 以M^R作为输入数据，均方误差(mean_squared_error)作为损失函数，Linear 作为激活函数。

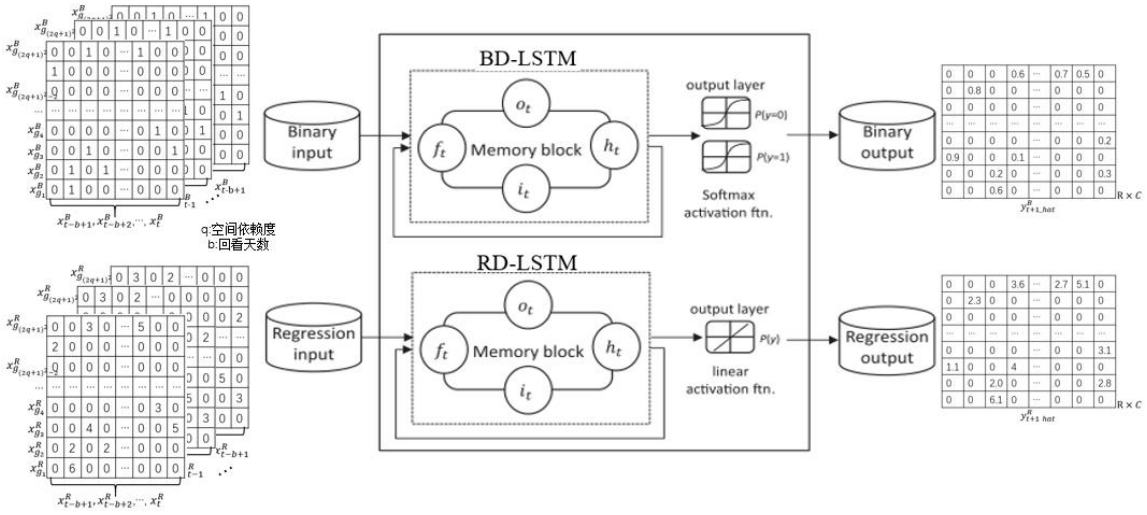

图 8　时空相关条件下不同数据形式对应的犯罪预测模型

（二）犯罪预测模型结果与分析

实验过程中 BD-LSTM、RD-LSTM 两模型分别取不同的迭代次数 epochs=50,100,500, 1000, 5000；不同的回看天数 lookBack=1,7,30,90,365；不同的空间依赖程度 q=0,1,2,3。组成多种不同参数的组合模型。为防止过拟合，分别设置不同的失活率 dropout =0.0,0.1, 0.2,…,0.9。此外，为度量不同参数模型的学习

能力及稳健性，分别对每种模型对应的不同参数组合运行 10 次。对所有样本数据按时间升序排列并进行编号，每次采用有放回的等概率抽样随机抽取一个样本作为测试集，直至抽取的所有不重复样本数量达到总样本所占比例 p 停止抽样且将这部分样本数据标记为测试数据集，余下 *1-p* 作为训练数据集。本文分别对 BD 和 RD 两类数据比较了 p=0.1,0.2,…,0.9 不同比例下测试集的平均预测准确率，结果显示在其它条件相同时，p 取 0.2，0.3，0.4 时预测准确率波动不大均在 BD 数据约为 60%，RD 数据约为 57%，由于总样本量不算很大 $p \geq 0.5$ 时基于 BD、RD 数据的预测性能急剧下降。而已有深度学习研究中，测试集与训练集常按照 3:7 的比例划分[29]，本研究后面的实验均是抽取总样本量的 30% 用于测试，余下 70% 用于训练。模型训练基于 tensorflow 背景下的 kares 框架实现。

关于犯罪预测模型的性能，主要关注模型的预测精度和鲁棒性。BD-LSTM、RD-LSTM 模型输出分别是 0-1 之间的概率取值以及非负回归数值，两模型分别利用二值交叉熵和均方误差计算实际数值与预测结果的距离，以此来衡量预测精度，其值越小，对应的预测精度越高。鲁棒性主要体现在模型对于每天预测精度的平稳性，能够适应一定范围内的波动。

图 9(a)(b)(c)(d)分别展示了迭代次数、回看期数、空间依赖度以及失活率取不同值时，RD-LSTM 模型对应的均方误差的变化。图 9(a)显示，随着迭代次数的增加，MSE 的变化。可看出，当迭代次数在 0-200 之间时 MSE 直线下降，200-800 期间依然保持很快的下降速度，800-1000 过程中 MSE 下降速度逐渐降低，1000-5000 过程中 MSE 略微有所增大且中间出现过几次波动，可能是因为迭代次数相对于样本量来说过多而产生了一定程度的过拟合导致的，因此本研究认为针对当前样本数据 epochs=1000 性能最优。图 9(b)显示了不同回看次数随着迭代次数的增加对应 MSE 的变化，考虑到整体来看入室盗窃在时间上可能具有某种周期规律比如，天、周、月、季度、年或是工作日、节假日等，所以预计执行的回看天数为 1,7,30,90,365。实际执行中发现 lookBack=30 时，每次迭代输入的样本量所占内存大大增加，导致模型训练时间很长。考虑到硬件设备的限制以及较长的训练时间，并且随着 lookBack 取值增加，损失的样本信息也越多，对于 lookBack=90,365 的情况未进行实验。图(b)表明 lookBack=7 时，模型预测精度更高。图 9(c)展示了不同空间依赖度下，测试集 MSE 的变化。不同取值 q 代表了不同地理范围的相互影响，q 的大小与最初网格的划分粒度有关，最初网格划分越小可能 q 对应的取值会越大。图(c)结果表明，对于当前网格大小，q=1 时性能最佳。图 9(d)比较了不同失活率，整体对应的预测精度差异。图(d)显示，失活率越大模型初始学习效率越高但预测精度提升缓慢，dropout 取 0.0,0.1,0.2,0.3 均表现出较好的性能，无法通过图(d)确定最佳失活率。为进一步比较 dropout=0.0,0.1,0.2, 0.3 的实验性能，本研究对不同失活率对应的整体预测精度

MSE 的均值、方差、最大最小等进行了综合比较。对比结果见图 10，其中 epochs=1000, q=1, lookBack =7,dropout 依次为 0.0,0.1,0.2,0.3。结果表明 dropout=0.1 时 MSE 波动范围最小，中位数也最低，因此，最佳失活率为 0.1。BD-LSTM 结果与 RD-LSTM 结果一致，因此，不做多余的说明与展示。综上所述，基于 LSTM 模型的最优超参数组合为 epochs=1000,lookBack=7,q=1, dropout=0.1，最初样本数 1572，其中训练样本 1200，测试样本 372。

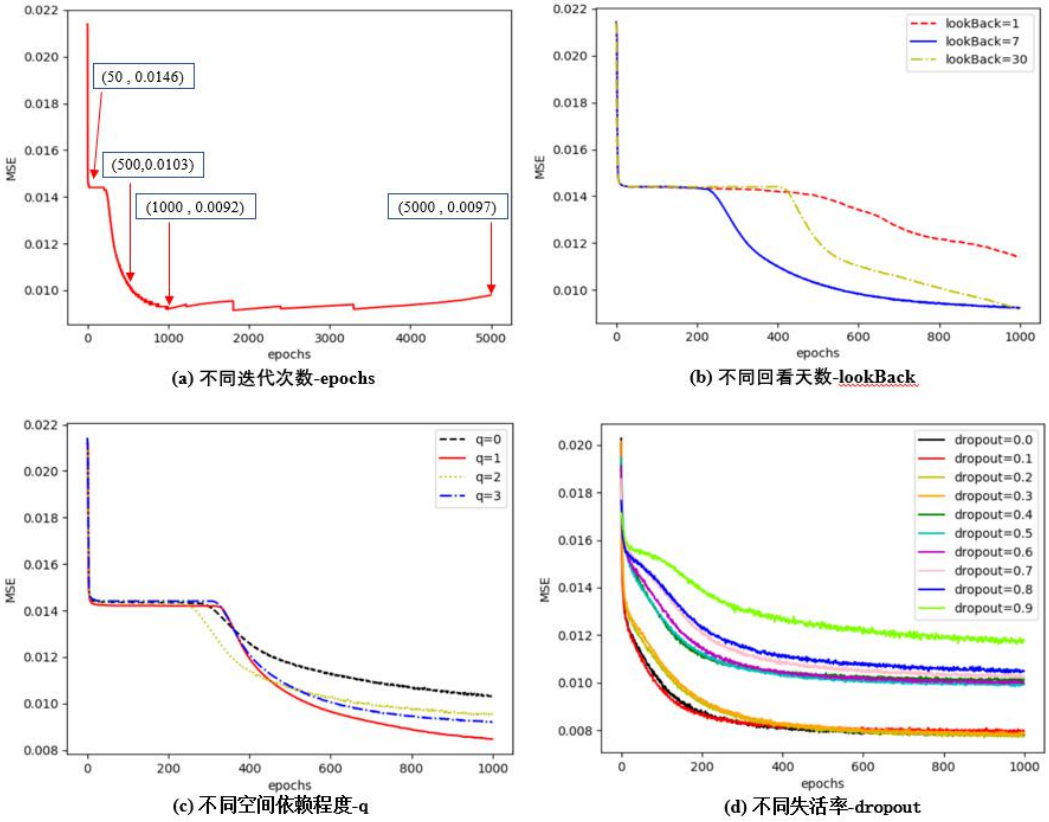

图 9　RD-LSTM 模型在不同超参数取值下 MSE 变化

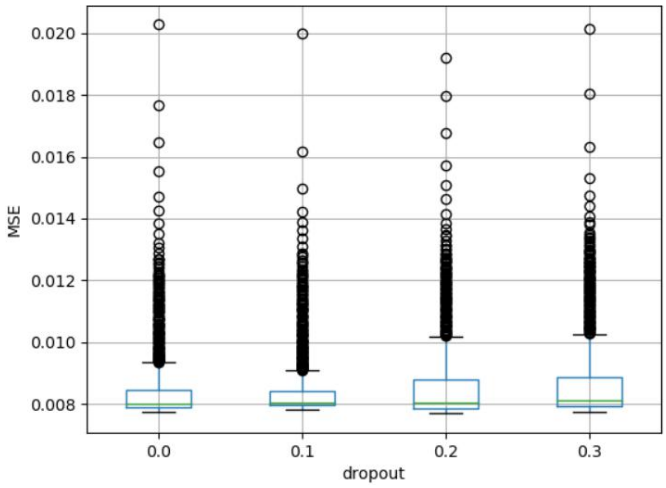

图 10　RD-LSTM 模型不同失活率下 MSE 箱线图

考虑到最佳回看天数为 7，表明入室盗窃可能存在以周为单位的时间特征。因此，本文将入室盗窃案件的发生时间按照工作日、休息日打上标签。通过方差分析发现工作日和休息日入室盗窃的区域发生频率和数量发生频率均存在显著差异(见表 6)。同样利用方差分析发现工作日、休息日对应准确率无显著性差异(见表 7)。再次证明本文提出的犯罪预测模型对于入室盗窃案发时间的差异具有鲁棒性。

表 6　BD、RD 数据集在工作日、休息日上案件发生区域数和数量的方差分析

数据类型		自由度	总平方和	均方平方和	F 值	P 值
BD	DATE	1	4655.7546	4655.7546	36.2167	7.97E-11
	Residual	1577	177043.1671	112.2658		
RD	DATE	1	8268.8523	8268.8523	41.5797	6.05E-10
	Residual	1577	293335.5623	186.0086		

表 7　BD、RD 数据集在工作日、休息日上预测准确率方差分析

数据类型		自由度	总平方和	均方平方和	F 值	P 值
BD	DATE	1	18.0461	18.0461	2.5678	0.1148
	Residual	370	2504.0126	6.7676		
RD	DATE	1	15.0556	15.0556	1.5569	0.2126
	Residual	370	3241.9452	8.762		

BD-LSTM、RD-LSTM 模型在包含工作日和休息日的 335 个样本的测试集上利用动态网格搜索法对各样本自适应设置阈值后分别得到 BD-LSTM、RD-LSTM 在测试集上准确率(见图 11)。整体来看，同一批测试集样本，二值化数据对应的预测准确率高于频数统计数据。表明预测指定时间段、指定区域、指定类别案件是否发生易于预测其具体的案件发生数量。

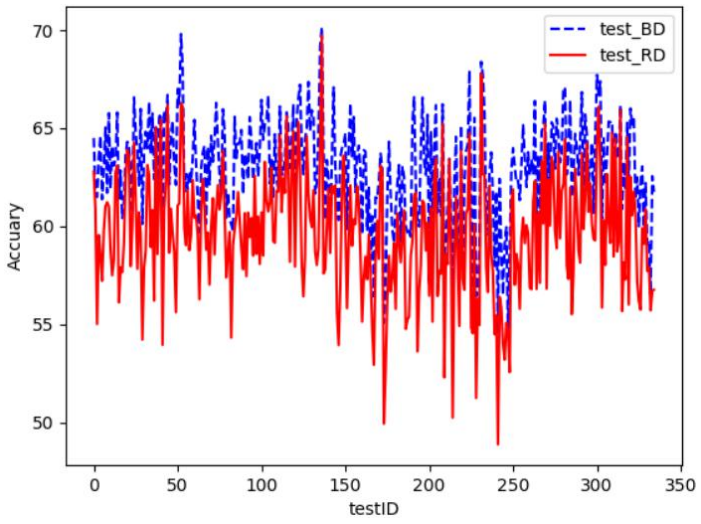

图 11　BD-LSTM、RD-LSTM 测试集上预测准确率

（三）模型的评估与比较

目前国内外针对犯罪预测研究广泛采用的模型包括决策树(DT)[26]、朴素贝叶斯(NB)[26]、随机森林(RF)[28]、自激点模型(SE)[16,27]，其公布的准确率在 35%至60%之间，由于实际使用的数据集存在差异，无法实现完全同一标准的对比。为检验基于 LSTM 模型的入室盗窃犯罪预测性能，本文比较了武汉市数据集在决策树、朴素贝叶斯、随机森林、自激点模型上的预测效果。通过实证比较研究，结果发现，相对于传统的机器学习方法，本文提出的基于 LSTM 的犯罪预测模型可靠性更强。本文自适应阈值选取条件下，BD-LSTM、RD-LSTM 对应的平均准确率分别为 63.52%和 59.68%。结果统计见表 8。

表 8　武汉市入室盗窃模型预测结果比较表

方法	DT	NB	RF	SE	RD-LSTM	BD-LSTM
准确率(%)	36.18	40.85	42.33	35.96	59.68	63.52

五、结论与建议

本文提出了一种基于 LSTM 模型的犯罪预测系统。在犯罪预测体系结构中，假设犯罪事件存在空间依赖的前提下，对地图实行网格划分，并将相邻地理位置的案件归于同一网格，为犯罪预测对应的地理范围提供了基本单元。时间分割和窗口化提供了以天为单位的时间序列数据集，可用于训练和测试犯罪预测模型。基于接警时空大数据利用地理空间依赖性及发生时间关联性的二值分类数据和频数回归数据分别构建了 BD-LSTM 和 RD-LSTM 模型。以武汉市的接警数据为例，分别学习了两模型的最优超参数取值、评估了两模型的预测精度和模型鲁棒性，证明了该预测系统的可靠性。

本文的主要贡献及发现有以下两点：第一，丰富了犯罪预测研究文献同时扩宽了深度学习应用领域。结合 LSTM 模型，构建入室盗窃犯罪预测预警系统，准确评价、科学预测和防范犯罪风险，既有益于法治软环境的营造，也有利于促进社区环境的安定、和谐。在一定时空范围内对入室盗窃犯罪进行趋势预测和风险预警，实现对犯罪行为的源头预防和有效控制。第二，开发了一种新的、可靠性强的犯罪预测系统。通过对比同一超参数多种不同取值对应的预测性能，学到了适用于武汉市入室盗窃类最佳性能的犯罪预测模型。通过最优回看天数以及空间依赖程度取值，证明入室盗窃类犯罪存在一定程度的时空模仿预期效应，与地震的余震效应类似。另外，将武汉市数据集在目前犯罪预测广泛采用的传统机器学习方法上进行了对比研究，结果表明，本文提出的基于 LSTM 犯罪预测系统

具有更好的预测效果。

由于不同犯罪种类在时间或空间上可能具有不同的发生模式，因此，未来针对不同种类的犯罪预测有必要结合实际的犯罪数据再次进行评估。在今后的工作中，由于犯罪预测的精度和速度在犯罪管理系统中起着至关重要的作用，因此可以根据不同的行政区域和时间尺度设计犯罪预测系统，使其更准确、更快速地进行预测。为了提高预测质量，一方面，本研究所提出的架构也可以采用最近开发的其它深度学习模型；另一方面，未来的研究可加入一些影响犯罪的其他因素，同时可以基于网格及路网综合信息进行犯罪预测。

参考文献

[1] 黄超, 李继红. 犯罪预测的方法[J]. 江苏警官学院学报, 2011, 26(1):107-110.

[2] 李继红,黄超.中外犯罪预测比较研究[J]. 学理论, 2010,29(8):155-156.

[3] 刘小娟,高连生.灰色系统理论在犯罪动态预测中的应用[J].中国人民公安大学学报(社会科学版),2005,66(1):44-48.

[4] 杜益虹, 刘世华. 基于 Logistic 回归的犯罪概率预测研究[J]. 绍兴文理学院学报, 2016,36(8):24-30.

[5] 李明,薛安荣,王富强,吴正寅.犯罪量动态优化组合预测方法[J].计算机工程,2011,37(17):274-275+278.

[6] 屈茂辉, 郝士铭. 基于 ARMA 模型的我国财产类犯罪人数预测研究[J]. 中国刑事法杂志, 2013,59(4):100-106.

[7] 吴绍兵,王昌梅.基于马尔科夫链的民族地区毒品犯罪预测研究[J].计算机与数字工程,2015,43(07):1270-1273.

[8] 李荣岗,孙春华,姬建睿.基于支持向量机的嫌疑人特征预测[J].计算机工程,2017,43(11):198-203.

[9] 孙菲菲,曹卓,肖晓雷.基于随机森林的分类器在犯罪预测中的应用研究[J].情报杂志,2014,33(10):148-152.

[10] 李卫红,闻磊,陈业滨.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),2017,42(08):1110-1116+1171.

[11] 于红志, 刘凤鑫, 邹开其. 改进的模糊 BP 神经网络及在犯罪预测中的应用[J]. 辽宁工程技术大学学报(自然科学版), 2012,31(02):244-247.

[12] Bogomolov A, Lepri B, Staiano J, et al. Once upon a crime: towards crime prediction from demographics and mobile data[C]//Proceedings of the 16th international conference on multimodal interaction. ACM, 2014: 427-434.

[13] TAM S, TANRIÖVER Ö Ö. CRIME PREDICTION USING SOCIAL SENTIMENT AND SOCIO-FACTOR[J]. Communications Faculty of Sciences University of Ankara Series A2-A3 Physical Sciences and Engineering, 2018, 60(1): 11-20.

[14] Antolos D, Liu D, Ludu A, et al. Burglary crime analysis using logistic

regression[C]//International Conference on Human Interface and the Management of Information. Springer, Berlin, Heidelberg, 2013: 549-558.

[15]Shrivastav A K, Ekata D. Applicability of soft computing technique for crime forecasting: A preliminary investigation[J]. International Journal of Computer Science & Engineering Technology, 2012, 9(9): 415-421.

[16]Mohler G O, Short M B, Brantingham P J, et al. Self-exciting point process modeling of crime[J]. Journal of the American Statistical Association, 2011, 106(493): 100-108.

[17]Kianmehr K, Alhajj R. Crime Hot-spots prediction using support vector machine[C]//IEEE International Conference on Computer Systems and Applications, 2006. IEEE, 2006: 952-959.

[18]Nasridinov A, Ihm S Y, Park Y H. A decision tree-based classification model for crime prediction[M]//Information Technology Convergence. Springer, Dordrecht, 2013: 531-538.

[19]Chitsazan M, Rahmani G, Neyamadpour A. Groundwater level simulation using artificial neural network: a case study from Aghili plain, urban area of Gotvand, south-west Iran[J]. Geopersia, 2013, 3(1): 35-46.

[20]Wang B, Yin P, Bertozzi A L, et al. Deep learning for real-time crime forecasting and its ternarization[J]. arXiv preprint arXiv:1711.08833, 2017 26(1): 84-90.

[21]Hochreiter S, Schmidhuber J. Long short-term memory.[J]. Neural Computation, 1997, 9(8):1735-1780.

[22]于嘉.长春市犯罪空间分析[D]．东北师范大学，2010.

[23]程薇,吴健平.国外犯罪时空分布研究综述[J].世界地理研究,2013,22(2):151-15.

[24]Graves A. Long Short-Term Memory[M].Supervised Sequence Labelling with Recurrent Neural Networks. 2012.

[25]Sak, Hasim, Senior, Andrew, Beaufays F. Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition[J]. Computer Science, 2014,6(5):338-342.

[26]Almanie T, Mirza R, Lor E. Crime Prediction Based On Crime Types And Using Spatial And Temporal Criminal Hotspots[J]. Computer Science, 2015, 5(4):235-239.

[27]Rosser G , Cheng T . Improving the Robustness and Accuracy of Crime Prediction with the Self-Exciting Point Process Through Isotropic Triggering[J]. Applied Spatial Analysis and Policy, 2019, 12(1):5-25.

[28]柳林, 刘文娟, 廖薇薇,等. 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, 2018, 37(6):761-771.

[29]周志华. 机器学习[M]. 北京：清华大学出版社，2016：25.

[30]毛媛媛, 戴慎志. 犯罪空间分布与环境特征--以上海市为例[J]. 城市规划学刊, 2006,12(3):85-93.

[31]Sagovsky A, Johnson S D. When does repeat burglary victimization occur[J]. Australian & New Zealand Journal of Criminology, 2007, 40(1):1-26.

致 谢

作为湖北经济学院湖北数据与分析中心的科研助理，我们有幸参与了中心与武汉市公安局大数据实战应用中心合作的智慧警务项目。项目研究过程也是学习过程，有艰辛，但更多的是快乐。这篇论文既是一个研究总结，也是这段快乐日子的一份记录。

首先要感谢指导教师张耀峰教授，在论文的思路设计和撰写方面，给了我们很多有益的指导。张教授是一位眼界开阔、思想深远的优秀导师。因为他，我们才得以接触到这么前沿大数据应用项目，每当我们“跑偏”时，他总能将我们及时拉回并帮助我们找到正确的道路，还会陪着我们一起“傻笑”。大数据研究没有数据是万万不行的，因此，我们还要感谢为我们提供数据的武汉市公安局大数据实战应用中心，尤其感谢该中心的胡欣警官，不仅为我们提供了犯罪领域知识的讲解，还不厌其烦的接受我们的各种咨询，穿着警服的胡警官“最帅”。最后要感谢湖北数据与分析中心智慧警务团队的所有小伙伴们，真心的感谢你们，因为只有“陪伴才是最长情的告白”，这篇论文“有我们的一半也有你们的一半”。

最后，我们想用张耀峰教授常和我们说的一句话来结束，“幸福就是做着快乐而有意义的事，梦想就是做着坚持就能感到幸福的事”！