水　利　学　报
2019年6月　SHUILI　XUEBAO　第50卷　第6期

文章编号:0559-9350(2019)06-0773-07

基于机器学习短历时暴雨时空分布规律研究

刘媛媛[1]，刘洪伟[2]，霍风霖[2]，刘业森[1]
（1. 中国水利水电科学研究院，北京　100038；2. 北京市水务局，北京　100038）

摘要：城市内涝风险的精细化管理和防洪排涝市政工程的科学设计，需要对当地降雨的时空分布特征有深入的了解。而传统以单站雨型代表整个区域降雨特征的分析方法，不能满足这一要求。本文尝试将机器算法引入到暴雨时空分布特征研究中，以北京城区2004—2016年降雨资料为研究样本，利用动态聚类算法，提取北京城区短历时暴雨时空分布的动态特征。经分析，北京汛期的短历时暴雨时空分布特征，可以分为3种类型：(1)降雨自西北部山区移动到城中心区，逐渐扩散到城区；(2)降雨集中在城区西南部地区，逐渐向北部和城中心区扩散；(3)降雨集中在城区中心区和东部地区，基本不发生移动。研究结果表明，基于机器学习算法提取的暴雨时空分布特征，与实际暴雨时空动态发展趋势相符，并且有各自对应的降雨形成的不同物理机制，可为城区降雨设计、城市内涝风险管理等工作提供借鉴与参考。

关键词：机器学习；时空分布；特征提取；短时强降雨
中图分类号：P338　　**文献标识码：**A　　**doi：**10.13243/j.cnki.slxb.20180943

1　研究背景

近年来，在我国由于暴雨引发的内涝已经成为影响城市生产生活运转的重要问题[1]。尤其是随着城市规模的快速扩大，人口和经济活动变得越来越集中，内涝灾害所造成的损失也被成倍放大，暴雨内涝对我国城市建设和运行的影响也越来越明显[2]。据不完全统计，2010年1—8月底，我国遭受洪涝灾害的县级以上城市已经超过了200座，其中大多数为暴雨内涝[3]。住建部2010年对32个省的351个城市内涝情况的调研结果显示，自2008年，62%的被调研城市都发生过不同程度的积水内涝[4]，“逢大雨必涝”已成为目前我国城市的通病。在极端天气多发、“逢大雨必涝”的背景下，城市内涝风险管理工作的水平亟待进一步提高。

城市内涝的风险管理依赖于准确的降雨预报，当前降雨预报的主要工具是数值天气预报模式，然而由于受整个学科水平的限制，要想在暴雨来临前对城市雨量及其时空分布进行精准预报十分困难，因此在实际工作中，降雨的数值预报可用性并不强。目前城市内涝的风险管理，更多的是依赖科学的市政工程措施和基于预案的应急管理手段。科学的工程措施与管理手段离不开对当地暴雨时空分布特征的深入了解，尤其是对降雨动态变化特征的深入了解。掌握短历时强降水时空分布特征和规律，在暴雨到来之前，根据历史规律提前预估降雨的动态发展趋势，即可根据历史上已出现过的内涝灾情预判不同地区的内涝风险，从而对内涝的风险管理具有重要意义。

城市内涝积水的具体情况与暴雨雨强和时空变化特征有直接关系[5]，在汇流历时内平均雨强相同的条件下，雨峰在中部或后部的雨型比均匀形雨型的洪峰大30%以上[6]。传统的暴雨时空分布研究，

收稿日期：2018-10-24
基金项目：国家重点研发计划项目(2016YFC0803109，2017YFC1502703)；2018年度国际水利水电科技发展动态调研专项(JZ0145C132018)
作者简介：刘媛媛(1978-)，高级工程师，主要从事防洪减灾相关研究。E-mail：50237479@qq.com
通讯作者：刘业森(1980-)，工程师，主要从事空间信息技术及防灾减灾研究。E-mail：liuys@iwhr.com

主要针对单站(代表站)的雨型开展。Pilgrim等[7]提出的雨型与实际降雨过程较为相似，其将雨峰时段放在出现可能性最大的位置上，雨峰时段在总雨量中的比例取各场降雨雨峰所占比例的平均值，再根据平均值由小到大的次序确定降雨强度从大到小的顺序，最后计算各时段内各场次降雨量占总降雨量百分比的平均值，代入确定的顺序中，由此得出雨型。Keifer等[8]根据强度-历时-频率关系得到了一种不均匀的设计雨型——芝加哥雨型，该雨型中任何历时内的雨量等于设计雨量。Huff[9]通过研究美国伊里诺斯州的暴雨，将整个降雨历时平均分成4部分，依据降雨峰值出现的4个时段位置，将降雨划分为4种雨型。对每一类雨型，设计多种不同频率的分配过程。这些针对单站(代表站)的降雨雨型长期以来得到了广泛的应用，但令人遗憾的是，它们不能体现降雨过程在时间、空间上动态变化的综合特征。尤其大城市地区，热力、动力环境均存在明显的空间分异，传统的单站(代表站)暴雨雨型就更不能满足分析降雨时空分布特征的要求，很难对城市内涝积水风险的精细化管理形成有效支撑。

近十几年来，人工智能(Artificial Intelligence，AI)技术在计算机视觉、自然语言处理、机器翻译、医学成像、医疗信息处理、机器人与控制生物信息等领域已取得长足进展[10]，尤其在医疗诊断[11]、无人驾驶等方面，表现不俗，而在图像识别、语音识别等方面，AI更是超过人脑，识别的准确性更高。机器学习(Machine Learning，ML)是AI的核心组成部分，是实现人工智能的主要途径。机器学习的核心是让计算机自动“学习”的各类算法，可以帮助计算机对大量样本数据进行分析并获得规律，然后利用规律对未知数据进行分类或预测。机器学习已在许多领域引发了历史性的革命，在气象灾害识别预测上也已得到了较好的应用[12]。同时，随着降雨数据采集和传输的发展，降雨数据更加丰富。基于大量降雨样本数据，利用机器学习技术对降雨的时空分布特征进行提取，成为可能。

本文将机器学习引入到暴雨时空分布研究中，提出针对超大城市短历时强降雨时空分布模式研究方法。区别于传统的针对单站降雨雨型分析的方法，本文以整个研究范围内所有雨量站为研究对象，利用机器学习算法，提取出研究区域短时强降雨时空动态分布特征，可为城市降雨方案设计、城市内涝风险分析等提供借鉴与参考。

2　研究方法

2.1　方法流程　强降雨范围、暴雨主要落区移动路径、降雨量变化、最强时段出现时间等，都是暴雨的时空分布特征，了解这些特征，对于有针对性地防范暴雨对城市的冲击有重要意义。本文将这些降雨的时空特征用高维数组进行表达，利用机器学习算法，提取短时强降雨的时空分布特征。主要步骤如下：(1)数据采集和质量控制：对历史短历时强降雨数据进行质量控制，剔除不合理的数据。(2)降雨过程结构化：首先划分降雨场次，将连续的降雨资料划分为若干暴雨场次，并从中筛选出暴雨过程，作为研究样本；其次从时间维度和空间维度构建数组，将一场暴雨过程数字化。(3)聚类分析，特征提取：利用机器学习算法，对历史降雨样本集进行聚类分析和特征提取。

2.2　降雨过程结构化　对各场次降雨，从时间维度和空间维度构建高维数组。将一场降雨过程中每时段的降雨，用矩阵的方式描述。一场降雨有n个时段，就有n个矩阵。历史暴雨样本集中就有N场雨，那就有N个这样的高维矩阵。该高维矩阵包含这次降雨过程的时间特征和空间特征。

基于该方法，建立降雨过程样本集Ω，实现了多场次降雨的结构化描述，见下式：

$$\Omega=\{Q_1,\ Q_2,\ \cdots,\ Q_N\} \tag{1}$$

$$Q_i=\begin{bmatrix} x_{1t1} & x_{2t1}\cdots & x_{st1} \\ x_{1t2} & x_{2t2}\cdots & x_{st2} \\ \cdots & \cdots & \cdots \\ x_{1tm} & x_{2tm}\cdots & x_{stm} \end{bmatrix} \tag{2}$$

式中：Ω为历史暴雨样本集，包括N个暴雨场次；Q_i为第i次降雨过程；x_{stm}为第s个雨量站tm时刻的降雨量，s=1，2，3，…，S，tm=1，2，3，…，m，S为雨量站个数，m为时段数。

2.3 动态聚类分析 本文所用的动态聚类算法属于人工智能技术中常见的经典聚类算法，是无监督的学习算法。动态聚类分析[13]的基本思想是：通过迭代寻找r个聚类的一种划分方案，使得用这r个聚类的均值来代表相应各类样本时，所得的总体误差最小。即通过该算法，将总体样本集划分为r个子集，使得各子集内的样本近似，而各子集之间的样本不同。再提取各子集的均值，得到属于该子集的降雨特征。

分析时，先随机选择r个样本点，也就是r个降雨过程作为r个子集的初始聚类中心，计算所有样本与这r个初始聚类中心的距离，并把样本划分到与之距离最近的那个中心所在的子集中，使所有的样本根据距离，自动聚集到各个子集中，从而得到初始分类类型数以及初始子集。计算各子集所有样本的均值，得到新一代的聚类中心，再次计算所有样本与新的聚类中心的距离，自动聚集，得到新的聚类中心，计算各子集所有样本的均值。不断迭代，并比较第p代和第p+1代聚类中心，如果相差在范围之内，则认为计算收敛，从而得到最终的子集及各子集的聚类中心。

该聚类方法收敛速度快，容易解释，聚类效果较好，适用于高维数据。但是该方法的聚类结果受初始聚类中心选择的影响较大。因此本文在迭代收敛后，通过与实际雨量空间特征不断的比较分析，判断子集数和初始子集中心是否合理，调整子集数以及子集的初始中心，以此反复进行聚类的迭代运算，直至确定合理的雨量空间分布特征类型数和聚类中心。计算步骤如下：

(1)分析的样本集为$\Omega=\{Q_1,\ Q_2,\ \cdots,\ Q_N\}$，$M$为最大迭代次数，$r$为初始划分的子集数，$C=\{C_1,\ C_2,\ \cdots,\ C_r\}$为$r$个子集。初始时$C_j=\varnothing$，$j$=1，2，…，$r$；

(2)从Ω中随机选取r个样本，作为初始r个子集的各中心向量$Z_j^{\,0}=\{z_1,\ z_2,\ \cdots,\ z_r\}$(0为迭代计算次数)；

(3)对于n=1，2，…，N，计算样本Q_i($Q_i\in\Omega$)与每个聚类中心$Z_j=\{z_1,\ z_2,\ \cdots,\ z_r\}$的距离$d_{ij}=\left\|Q_i-z_j\right\|_2^2$，如果$d_{ij}=\min\{d_{ij}\}$ $i=1$，2，…，N，则$Q_i\in C_j$。更新$C_j=C_j\cup Q_i$；

(4)对于j=1，2，…，r，对C_j中的所有样本点，重新计算中心向量$Z_j^{\,1}=\dfrac{1}{C_j}\sum_{Q_{i\in C_j}}Q_i$；

(5)不断重复迭代，如果$Z_j^{\,p+1}\neq Z_j^{\,p}$，$j$=1，2，…，$r$，则回到步骤(2)，重复迭代计算。如果$Z_j^{\,p+1}=Z_j^{\,p}$，$j$=1，2，…，$r$，运算结束；

(6)输出各子集$C=\{C_1,\ C_2,\ \cdots,\ C_r\}$以及各子集的均值$Z_j^{\,p+1}=\{z_1,\ z_2,\ \cdots,\ z_r\}$。

3 计算实例分析

3.1 资料处理 北京作为超大规模城市，最近30年经历了快速的城市化进程，同时内涝问题加剧。北京市主汛期(6—8月)降雨集中，汛期突发的强降雨基本占了全年降水量的大部分[14]。这种集中型的极端降雨，是北京严重城市内涝灾害的主要诱因。尤其2011年“6·23”与2012年“7·21”特大暴雨，北京全市发生了严重的洪涝灾害，并造成了重大人员伤亡和财产损失。2016年“7·20”暴雨，虽然降雨量与“7·21”暴雨相当，但历时较长，雨强减半，危害较轻，可见雨型时空分布有显著影响。

本文选取北京市水文总站提供的2004—2016年北京城区14个气象站(如图1所示)，夏季(6—8月)5 min间隔降雨监测数据。在实际的监测降雨数据中，可能会受监测设备或者人为影响，出现非常不合理的数据，导致不能分析出合理的结果，或者根本不会出结果。因此在做分析之前，需要对实际的监测降雨数据进行质量控制，剔除不合理的数据。质量控制的标准如下：(1)单站5 min降雨量超过10 mm，且孤立存在，该站前后时间都没有降雨，则认为是不合理的记录；(2)某站5 km×5 km范围内的雨量站监测数据为0，而该站5 min降雨量超过10 mm，则认为是不合理的记录。

对于单站不合理的记录，利用该站5 km×5 km范围内雨量站差值结果，代替该站不合理的记录。

将质量控制后的数据进行场次划分，筛选出短历时强降雨过程作为样本。首先，根据北京城区降雨的特点，连续超过2h5min降雨量小于0.1 mm，则认为无有效降雨，按照这个规则划分降雨场次。根据北京城区的暴雨蓝色预警指标：1h降雨量达30 mm以上或6h降雨量达50 mm以上[15]，筛选出暴雨过程。按照上述标准和流程，在2004—2016年间，共筛选出强降雨过程89场，将场次降雨划分为12个时段，构建高维数组样本库。

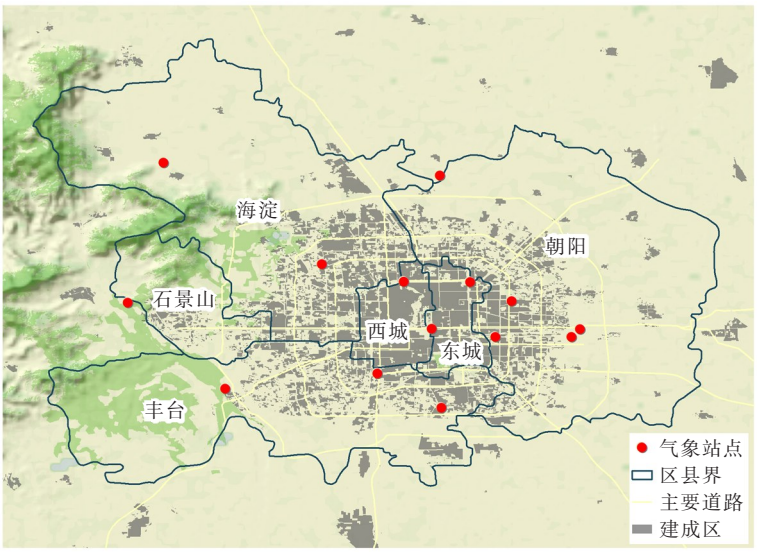

图1　气象站分布

3.2　结果分析　本文基于机器学习算法，提取北京城区夏季短历时强降雨的时空分布特征。经分析，北京夏季短历时强降雨过程可分为3类：第一类，暴雨中心自西北山区向城中心区移动；第二类，暴雨中心自西南经城区，向东北方向移动；第三类，暴雨集中在城区，基本不发生移动。

这3类降雨模式的雨量格局及发展过程存在明显差异，为了进行直观对比，以各站点各时段的降雨量为权重，计算了这3类降雨各时段的雨量重心点，结果见图2。图2中，分别用不同颜色的正方形、圆形、三角形代表这3类降雨$t1$～$t12$时段的雨量重心点。从总体位置来看，第一类、第二类、第三类的重心点分别偏西北、西南、东南。从暴雨中心移动过程来看，第一类明显由西北向城区移动，第二类主要自西南地区向北部和城中心区移动，第三类基本集中在城中心区。其中第一类的暴雨过程，暴雨中心移动的距离较长，而第三类的暴雨过程，降雨比较集中，暴雨中心基本不发生移动。

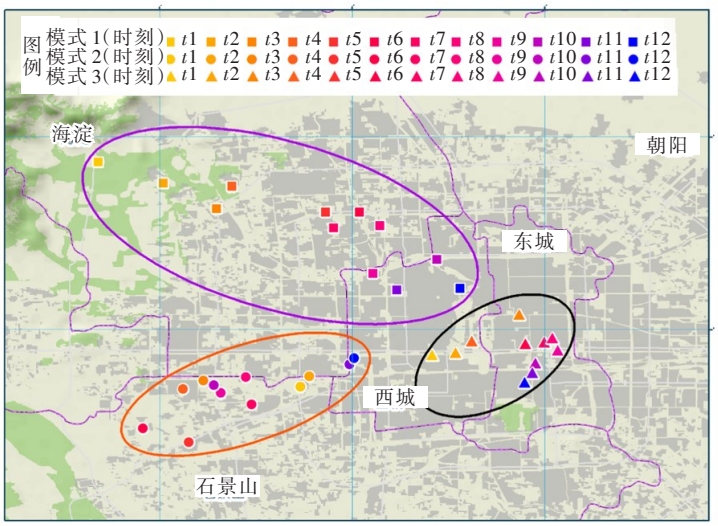

图2　不同类型降雨重心点时空变化对比

各类型降雨模式不同时刻的雨量分布格局，见图3—图5。图中色块表示各时刻降雨空间分布情况。

（1）第一类降雨。如图3所示。该类型降雨的时空分布特征为：降雨从西北部山区向城区中心区和城区东部扩散。降雨一开始都集中在西北部山区，城区其余各站基本没有降雨。而后降雨量逐渐分散，各站都有降雨发生，属于该类型的暴雨占统计样本数的31.8%。出现这类过程时，北京通常位于东移低涡前部的动力不稳定区域，高空大尺度低涡自西北向东移动，当近地层水汽条件配合时，即会自西北向东南出现强降水。2011年“6·23”大暴雨就属于该类型，该场大暴雨自西北部山区逐渐向城中心区移动，西北部地区降雨量较大[16]，符合第一类降雨的时空分布规律。

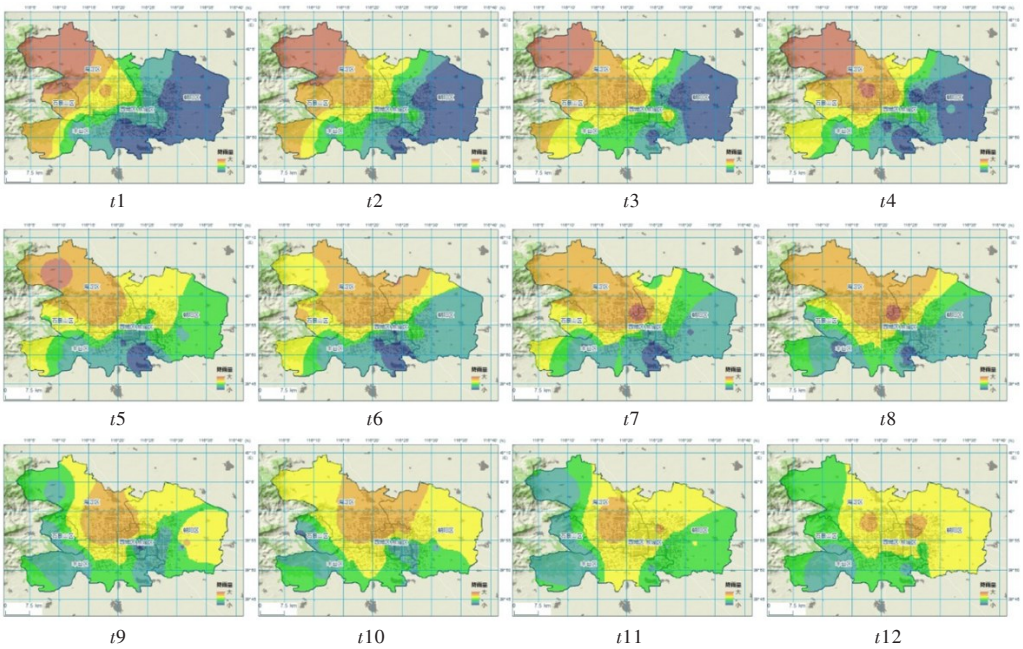

图3　第一类降雨时空分布

（2）第二类降雨。如图4所示。该类型降雨的时空分布特征为：主要降雨都集中在城区南部和西南部地区，逐渐向北部和城中心区扩散。属于该类型降雨占样本总数的13.7%，出现这种降雨时，北京通常处于槽前的西南暖湿气流里，或是位于偏南低空急流前部激发的强对流中，其本质是暖区强降水，因为湿度和对流不稳定条件好，降水强度大。

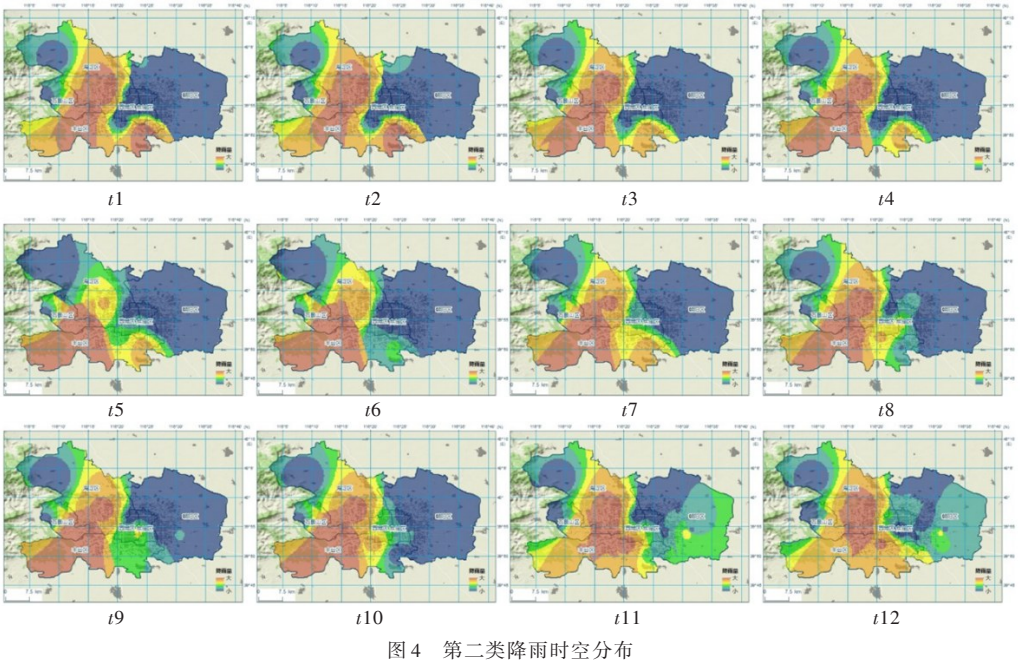

图4　第二类降雨时空分布

(3)第三类降雨。如图5所示。该类型降雨的时空分布特征为：主要降雨集中在城区中心区和东部、南部地区，基本不发生移动，降雨集中。该类型降雨占样本总数的54.6%。为北京城区夏季主要降雨类型，基本发生在午后到傍晚。2004年“7·10”特大暴雨就是属于该类型的典型暴雨过程。本次暴雨中心位于城区中部和西部。城区平均降雨量为81 mm，而全市平均降雨量23 mm。市级报汛站中最大点天安门站降雨量达到106 mm[17]，符合该类型降雨的时空分布规律。属于该类型的暴雨过程，主要是受城市“雨岛效应”影响造成的，多为局地性降水。

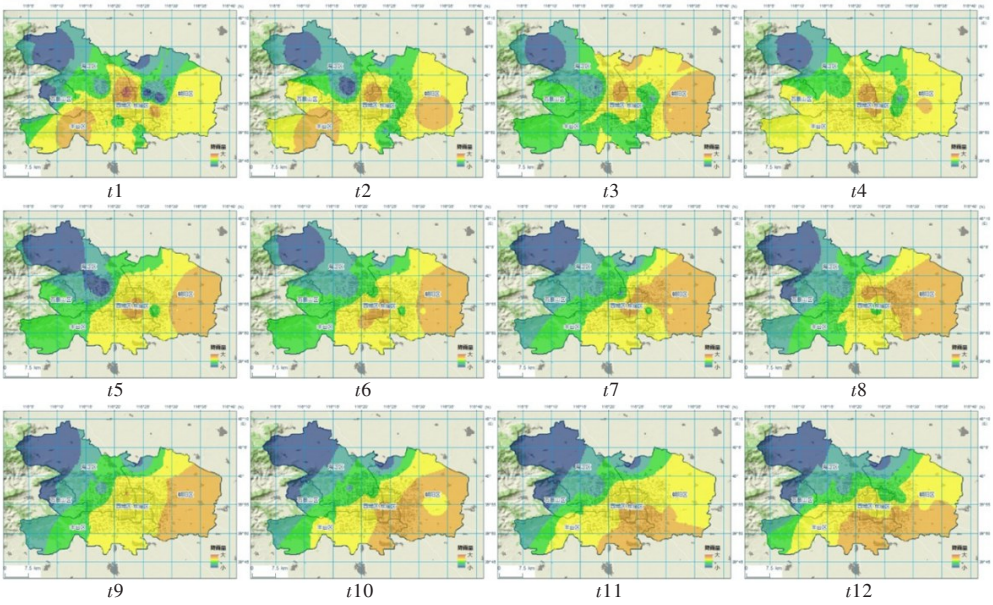

图5　第三类降雨时空分布

综上所述，北京城区汛期短时强降雨过程时空分布特征可以归为三种类型，这三种类型的降雨时空分布特征与属于该类的实际降雨过程基本符合，而且这三种类型降雨时空分布特征，均有相应物理机制的解释。

4　结论与展望

深入了解城市暴雨动态变化的特征规律，掌握短历时强降水时空分布特征，是科学应对暴雨引发的积涝等次生灾害的先决条件，而传统的以点带面的方法，不能很好地分析暴雨动态变化特征，无法满足提高城市内涝风险管控水平的需求。本文将机器学习算法引入到降雨时空分布特征的研究中，实现了暴雨时空分布的特征提取，将北京城区的汛期短时强降雨过程分为3种类型：(1)降雨自西北部山区移动到城中心区，逐渐扩散到城区；(2)降雨集中在城区西南部地区，逐渐向北部和城中心区扩散，北部山区基本无降雨；(3)降雨集中在城区中心区和东部地区，基本不发生移动。经机器学习，提取出的各降雨模式的时空分布规律，与属于各模式的实际降雨过程基本相符，并且有各自对应的降雨形成的不同物理机制。当然也应看到，本文仅以北京城区14个站2004—2016年的降雨资料为分析样本提取的降雨时空分布特征，可能具有一定的随机性和不确定性，对于全北京市域的特点还未很好把握，有待进一步深入开展。未来可收集整理更大范围、更长序列的降雨资料，进一步补充完善数据样本，从更大范围探讨降雨过程的变化特征，以得到更加全面完整的分析结果。

参　考　文　献：

[1] 王浩，梅超，刘家宏．海绵城市系统构建模式[J]．水利学报，2017，48(9)：1009-1014，1022．

[2] 张炜，李思敏，时真男．我国城市暴雨内涝的成因及其应对策略[J]．自然灾害学报，2012，21(5)：

180-184.
[3] 程晓陶.城市型水灾害及其综合治水方略[J].灾害学, 2010, 25(S1): 10-15.
[4] XIA Jun, SHI Wei. Perspective on water security issue of changing environment in China[J]. Journal of Hydraulic Engineering, 2016, 47(3): 292-301.
[5] RAFIEEINASAB A, NOROUZI A. Toward high-resolution flash flood prediction in large urban areas-Analysis of sensitivity to spatiotemporal resolution of rainfall input and hydrologic modeling[J]. Journal of Hydrology, 2015, 531(2): 370-388.
[6] 吴彰春, 岑国平, 安智敏.坡面汇流的试验研究[J].水利学报, 1995(7): 84-89.
[7] PILGRIM D H, CORDERY I. Rainfall temporal patterns for design floods[J]. Journal of the Hydraulics Division, 1975, 101(1): 81-95.
[8] KEIFER C J. CHU Henry Hsien. Synthetic storm pattern for drainage design[J]. Journal of the Hydraulics Division, 1975, 101(10): 1371-1371.
[9] HUFF F A. Time distribution of rainfall in heavy storms[J]. Water Resources Research, 1967, 3(4): 1007-1019.
[10] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521: 436-444.
[11] 王威, 李郁, 张文娟, 等.深度学习技术在疾病诊断中的应用[J].第二军医大学学报, 2018, 39(8): 852-858.
[12] 张文海, 李磊.人工智能在冰雹识别及临近预报中的初步应用[J].气象学报, DOI:10.11676/qxxb2019.014.
[13] 杨明极, 马池, 王娅, 等.一种改进的K-means聚类的FCMM算法[J].计算机应用研究, DOI:10.3969/j.issn.1001-3695.2017.12.0851.
[14] 刘熙明, 胡非, 李磊, 等.北京地区夏季城市气候趋势和环境效应的分析研究[J].地球物理学报, 2006, 49(3): 689-697.
[15] 北京市气象局.北京市气象灾害预警信号与防御指南[Z].2016.
[16] 陈静, 刘琳.2011年汛期北京城市暴雨特征及其灾害成因初步分析[J].暴雨灾害, 2011, 30(3): 282-287.
[17] 化全利, 吴海山, 白国营.2004年7月10日北京城区暴雨分析及减灾措施[J].水文, 2005, 25(3): 63-64.

An application of machine learning on examining spatial and temporal distribution of short duration rainstorm

LIU Yuanyuan[1], LIU Hongwei[2], HUO Fenglin[2], LIU Yesen[1]

(1. *China Institute of Water Resources and Hydropower Research, Beijing* 100038, *China*;
2. *Beijing Water Authority, Beijing* 100038, *China*)

Abstract: It is necessary and important to understand the spatial and temporal distribution characteristics of rainfall in order to improve management of urban waterlogging risk and scientific design of municipal works for flood control and drainage. However, it is hard to derive accurate prediction based on monitoring data of a single station or the average of multiple stations. Availing of machine learning algorithm, this paper contributes to investigating rainfall spatial and temporal distribution characteristics. Specifically, the technique of dynamic clustering algorithm examines the spatial and temporal distribution characteristics of rainfall in Beijing urban area during recent 10 years. The result shows that there are three types of rainstorms in Beijing urban area. Rainstorms of type 1 move from the northwest to the center of Beijing, then spreads to the eastern part of the urban area; rainstorms of type 2 occurs in the southwestern region of the urban area, and gradually northward, but there is no rainfall in the mountainous northwest; rainstorms of type 3 are concentrated in the central and eastern regions, and basically does not move. The results are consistent with the actual rainstorm process. It provides references for urban rainfall design and urban waterlogging risk management.

Keywords: machine learning; spatiotemporal patterns; feature selection; short duration rainstorm

(责任编辑：王成丽)