

基于降雨类型直方图分析的降雨站点相似性研究

章龙飞¹, 朱跃龙¹, 李士进¹, 高祥涛²

(1. 河海大学计算机与信息学院, 江苏 南京 210098; 2. 江苏省水文水资源勘测局, 江苏 南京 210029)

摘要:为研究降雨观测站点间的相似性,提出了基于单场降雨类型直方图分析的降雨站点相似性比较模型。首先提取降雨时间序列中的雨量项,经过数据预处理,选择和提取单场降雨特征,并将这些特征统计量进行数据标准化处理;采用基于 Davies-Bouldin 指数聚类有效性指数的 K-means 方法进行聚类,最后对聚类结果进行基于单场降雨的降雨类型直方图相似性分析。实验结果符合相关区域的站点相似性实际情况,表明该相似性模型是可靠的,具有较好的应用价值。

关键词:降雨序列;相似性分析;特征提取;Davies-Bouldin 指数;K-means 聚类;降雨类型直方图

中图分类号: P334

文献标识码: A

文章编号: 1000-0852(2013)03-0010-08

1 引言

时间序列是一种在科学研究、商业应用中普遍存在的数据形式^[1-2],如灾变预测^[3-4],降雨预测^[5],金融数据和电力负载数据预测^[6]等。在水文领域,时间序列的研究由来已久^[7],通过研究水文时间序列的时空规律,对洪水暴雨等自然灾害进行预测,为流域水利建设及农业发展提供决策依据。降雨是水资源的主要来源之一,水利、农业以及生态系统都受其影响,水资源是社会经济发展的重要支撑和保障条件。因此,研究降雨时间序列对社会经济可持续发展和水资源管理具有重要的意义。

为了进行流域水资源科学管理和调度,需要全面了解流域水资源现状,特别是降雨信息^[8]。但是,目前我国大部分地区雨量站分布稀疏,虽然有 3 400 多个国家级水文站点,但布局不甚合理,大多部署在经济发达地区。随着水利信息化的不断推进,新增设了许多雨量站,但由于许多雨量站设站时间较晚,部分站点降雨资料只有短短几年^[9]。

为了更准确地获取基础水文数据,科学管理水资源,需要解决以下问题:缺失资料站点的数据补齐,如 A 站点有 20a 的降雨数据,B 站点有 40a 的降雨数据,若它们在近 20a 内的降雨类型相似,则 B 站可以补齐 A

站前 20a 的数据;去除旧的冗余站点,即地理位置相近的两个站点若相似度极高,则可以考虑去除其中的一个冗余站点(若两站点相似度很高但地理位置较远,一般不考虑去除冗余站点的问题)。因此有必要对降雨站点进行相似性研究,为水资源管理提供新的技术支撑。

以往对降雨时间序列的研究主要包括降雨预测、极值分布、周期分析和降水场的时空分布等。陈晓宏等把聚类分析理论引入降雨空间特征分类研究,进行降雨特征空间分布模式识别,预测无记录的站点归类,为降雨资料的差补延长提供了一种新的途径^[9];张国建等利用最大与最小年降水量比值来分析降水量的年际变化特征^[10];苏布达等利用 Weibull 分布分析了长江流域降水极值时间序列的分布特征^[11];刘德地等将云模型引入到降雨量时空分布特性的研究中^[12];巴金福等根据黄河上游逐月降水资料分析了汛期降雨量的变化趋势和空间分布特征及其变化^[13]。

降雨序列实质是不连续的时间序列,此前大多研究都是基于单纯日月或年降雨量累积的宏观统计分析,没有从更细粒度分析单场降雨的统计特征,而单场降雨序列恰恰包含了能反映降雨序列本质的更丰富的降雨过程信息。例如同样年降雨量总和为 1 000mm 的降雨,可能是 8 次单场降雨为 125mm,也可能是 20 次单场降雨为 50mm,还可能是若干场不同降雨量,不同

收稿日期:2012-10-16

基金项目:国家自然科学基金资助项目(51079040);水利部 948 项目(201016)

作者简介:章龙飞(1987-),男,浙江诸暨人,硕士,主要从事水文数据挖掘工作。E-mail:zhanglongfeiie@gmail.com.cn

通讯作者:朱跃龙(1959-),男,江苏建湖人,博士,教授,主要从事信息学、智能信息处理研究。E-mail:ylzhu@hhu.edu.cn

持续时间的降雨。若单纯利用年降雨量的累计而不考虑单场降雨的具体情况,则无法区分一年内两个站点降雨类型的差异,更无法区分降雨时空分布等信息;若考虑单场降雨,则可分析更多具有区分度的降雨信息,包括单场降雨和,单场降雨日均值,大于或小于某个临界值的日降雨量,单场降雨天数等。

目前国内基于单场降雨序列的相似性研究基本是空白,因此本文针对这种降雨时间序列不连续的特点,基于单场降雨对各种特征量统计分析。通过研究基于单场降雨的降雨序列相似性,继而探索区域降雨的相似性方法。从数据挖掘的角度,首先对降雨数据进行预处理,特征选择与提取,然后进行基于 Davies-Bouldin 指数^[14]的 K-means 聚类^[15]得到单场降雨类型,再结合信息检索领域的词袋统计模型进行基于单场降雨类型直方图的降雨站点相似性比较,最后采用江苏省境内过去 10 多年的降雨观测数据对本文提出模型进行验证。实验结果最终符合现实状况,表明该模型是有效的,具有较好的实际应用价值。

2 基于单场降雨类型直方图的降雨站点相似性模型

为建立降雨站点相似性模型,本文首先对雨量站点数据进行预处理,划分单场降雨,对特征进行选择与提取,采用基于 Davies-Bouldin 指数的 K-means 聚类,然后通过基于单场降雨类型直方图计算不同降雨站点的相似性,最后结合区域的降雨实际情况和站点地理空间上的分布因素,在一定误差范围内验证提出模型是否符合预定相似度评价要求,是否与实际情况相符。

整体相似性模型构建流程图如图 1 所示。

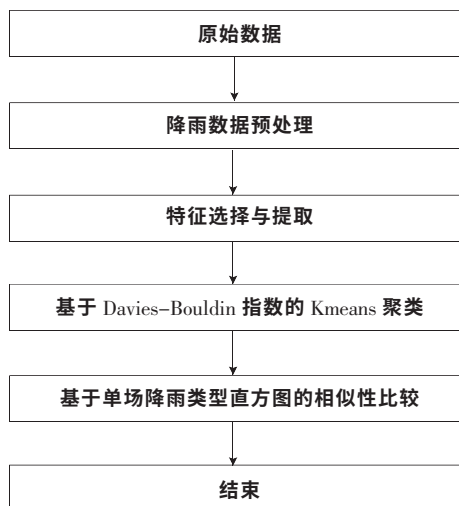


图1 降雨类型直方图相似性比较流程图

Fig1 Flow chart for similarity comparison of precipitation type histogram

2.1 特征选择与提取

特征选择是非常关键的步骤,它将直接影响聚类结果的合理性,进一步影响相似性比较的有效性。经过相关文献的总结分析,本文选取以下 9 个基于单场降雨的特征统计量:雨量和 r_{sum} 及天数 n ^[16],日平均降雨量 r_{ave} ^[17],日降雨量最大值 r_{max} 和最小值 r_{min} ^[18],日降雨量小于 1.27mm 雨量和 r_{127sum} 与降雨天数 r_{127} ,日降雨量大于 50mm 的雨量和 r_{50sum} 与降雨天数 r_{50} ^[19]。

单场降雨雨量和与降雨天数可以反映单场降雨总量和持续时间;日平均值能够反映单场降雨的强弱情况;最大值和最小值能够反映单场降雨的极值情况,故作为特征统计量;我国气象部门规定,暴雨为 24h 雨量超过 50mm 的降雨,所以选择 50mm/d 作为日降雨极大值的统计量;根据 Munger 指数^[20],日降雨量小于 1.27mm 可以判定干旱天气,所以选择 1.27mm/d 作为降雨的极小值统计量。

2.2 基于 Davies-Bouldin 指数的 K-means 聚类

K-means (K-均值) 聚类算法由于其算法简单,计算速度快,通常被作为大样本聚类分析的首选方法,是最常用的聚类方法之一,广泛应用于数据挖掘研究中。但 K-means 算法存在一定局限性,如聚类数目 K 需事先指定。本文提出采用聚类有效性指数进行聚类数目的自动确定。

传统的 K-means 算法是将含有 n 个数据点 (实体) 的集合 $X=\{x_1, x_2, \dots, x_n\}$ 划分为 K 个类簇 C_j 的问题 ($j=1, 2, \dots, K$)。首先随机选取 K 个数据点作为 K 个类簇的初始簇中心,集合中每个数据点被划分到与其距离最近的簇中心所在的类簇中,形成初始的 K 个聚类分布。对已分配的每一个类簇计算新的簇中心,然后继续进行数据分配的过程,这样迭代若干次后,若簇中心不再发生变化,说明数据对象已经全部分配到各自所在的类簇中。

基于 Davies-Bouldin 指数 (聚类有效性指数) 的 K-means 聚类算法能自动确定最佳聚类个数,它解决了传统的 K-means 算法需要事先确定 K 值的缺点。本文采用基于 Davies-Bouldin 指数的 K-means 聚类方法,算法步骤如下:

(1) 准备聚类所需的样本数据及设定 K 的大致范围,本文设 K 初始值为 6, $\text{Max } K=20$;

(2) 判断 K 是否小于等于 $\text{Max } K$, 如果是,则随机取 K 个聚类中心,转向(3),否则转向(7);

(3) 分别计算各个样本点到 K 个簇中心的距离,

并且将该样本点归到与其距离最近的簇中;

(4)将所有样本点归类结束之后,重新计算 K 个簇的聚类中心;

(5)比较新计算的聚类中心与原聚类中心,若聚类中心不同,则转向(3),否则转向(6);

(6) 计算当前聚类数目 K 对应的 Davies-Bouldin 的值,然后 $K=K+1$,转向(2);

(7)最后比较所有的 Davies-Bouldin 值,将最小的 Davies-Bouldin 值所对应的 K 作为最后输出。

2.3 基于单场降雨的降雨类型直方图

在信息检索技术中,有一个词袋模型 (Bag of words model)的概念,它是自然语言处理和信息检索中的一种有效模型。在该模型中,文本(段落或者文档)被看作是无序的词汇集合,忽略语法甚至是单词的顺序,每个文档被表示为一个用词汇出现次数所组成的矢量。

本文结合词袋模型的特点,将基于 Davies-Bouldin 指数的 K-means 聚类结果获得的单场降雨类别看作一个单词,相同类别号的单场降雨落入相同的降雨类型袋中,这种新的模型被命名为降雨类型词袋模型 (Bag of precipitation types)。由于可用类似图像分类中直方图的形式表示,即每一个类别号所对应的降雨场次为 bin 的频度值,每一个站点对应的各个类别降雨场次的组合构成降雨类型直方图。图 2 和图 3 给出了江苏省北部 1 号(N1)站点和 2 号(N2)站点的降雨类型直方图,其中 14 表示聚类结果有 14 种降雨类型,可以看出由于这两个站点地理位置比较接近,其降雨类型直方图分布也比较相似,下面给出其相似性数值计算方法。

2.4 基于降雨类型直方图的相似性度量

获得降雨类型直方图后,需进一步建立降雨站点间的相似性模型。本文先将每一个站点各类别对应的

降雨场次进行归一化计算处理,尽可能减小同一站点不同特征量由于数量上不同对相似性结果产生贡献率不同的现象,然后利用直方图相交(Histogram intersection)的方法获取两站点之间的相似性。

本文中两站点直方图相交计算方法如公式 (1)所示:

$$\text{sim}(M, N) = \sum_{i=1}^k \min(a_i/A, b_i/B) \quad (1)$$

式中: M, N 为两站点; a_i 和 b_i 为两站点第 i 类降雨类型号所对应的降雨场数; A, B 分别为 M, N 两站点降雨场次总数; k 取值为 14, 代表 14 种降雨类型。

采用公式(1),计算得到图 2 和图 3 对应降雨类型直方图相似度为 0.903,可见该度量确实反映出上述两个站点的降雨类型相似性。根据基于单场降雨的降雨类型直方图相似性原理,约定两站点相似度最高值为 1,最小值为 0。影响降雨站点的相似性因素有很多:如水库,江河湖海等水源,全球气候变化,地理空间位置等。在空间位置上距离相近的降雨站点相似度越高,距离较远的则相似度越小。从区域划分的角度来说,同一区域内两站点相似度一般要大于不同区域站点的相似度。

3 实验结果与分析

本文实验数据为江苏省水文水资源勘测局提供的江苏省各地区雨量站点 2000~2011 年单日不同观测时刻的降雨数据。

数据预处理阶段,首先挑选三块区域站点,如图 4~6 所示,标明为北部(N)、中部(M),南部(S)。

剔除区域中有站名但无数据的站点,如南部的天王寺站点。将这三个区域中有降雨数据的站点数据提取出来,其对应的站点代号和地名如表 1 所示,其中北部站点 7 个,中部 6 个,南部 8 个,这些站点数据相对较完整

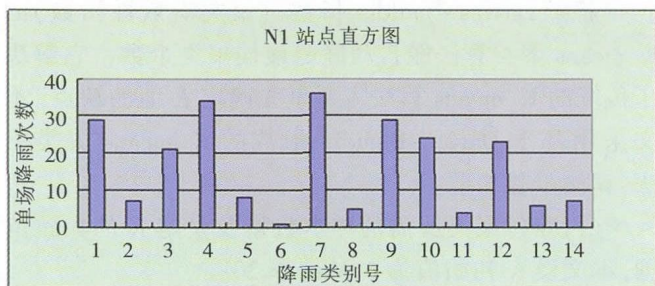


图 2 N1 站点基于 Davies-Bouldin 指数的 K-means 聚类结果的降雨类型直方图

Fig.2 Precipitation type histogram of N1 sites based on the K-means clustering results with the Davies-Bouldin index

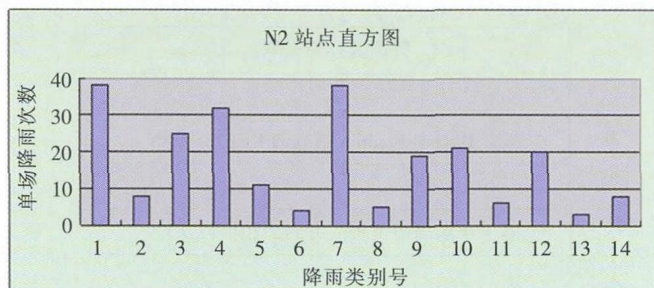


图 3 N2 站点基于 Davies-Bouldin 指数的 K-means 聚类结果的降雨类型直方图

Fig.3 Precipitation type histogram of N2 sites based on the K-means clustering results with the Davies-Bouldin index



图 4 北部(N)站点图
Fig.4 Distribution of the stations in the north area (N)



图 5 中部(M)站点图
Fig.5 Distribution of the stations in the middle area (M)



图 6 南部(S)站点图
Fig.6 Distribution of the stations in the south area (S)

表1 三块降雨区域所对应的站点
Table1 The stations corresponding to the three pieces of precipitation area

北部站点名	北部站号	代号	中部站点名	中部站号	代号	南部站点名	南部站号	代号
小塔山水库	54902	N1	盱眙	53000	M1	赤山新闸	69446	S1
青口	54740	N2	三河闸	53010	M2	大溪水库	69902	S2
石梁河水库	54900	N3	龙王山水库	53259	M3	高淳	69448	S3
大兴镇	54733	N4	桂五水库	53251	M4	句容	69439	S4
横沟水库	54930	N5	化农水库	53255	M5	溧阳	69451	S5
临洪	54812	N6	红旗水库	53906	M6	沙河水库	69900	S6
小许庄	54731	N7				天生桥闸	69442	S7
						王母观	69602	S8

且能代表江苏省南、北、中不同区域的降雨特点。

提取每个站点近 10 年数据中的雨量项,去除那些重复冗余的日降雨信息。考虑到某些场次的降雨中间有暂时不下雨,但仍应是同一场雨的情况,约定降雨间隔小于两天为同一场降雨,对单场降雨进行划分。再根据划分好的降雨场次计算 2.1 节的 9 个特征统计量。将这 9 个特征统计量作为单场降雨的特征向量,并且对三个区域的 21 个站点的所有特征向量进行 z-score 数据标准化,可以保证每一个特征值对聚类结果的贡

献率不会因为数值的大小而有显著的差异。再对这些特征向量进行基于 Davies-Bouldin 指数的 K-means 聚类,单场降雨类型最终被聚成 14 类。

统计聚类结果,得到站点的降雨类型直方图,如北部 N2 站点各降雨类型数据如下: 238,38,8,25,32,11,4,38,5,19,21,6,20,3,8,得到的降雨类型直方图如图 3 所示。

再利用 2.4 节模型对站点相似性进行分析,包括相同和不同区域站点相似性分析,同一区域站点相似

度如表 2~4 所示。

如表 2 为北部的 7 个站点间的相似度。模型的评价标准如下:若两站点相似度大于 0.8,则表明相似度较高;若两站点小于 0.8,则表明相似度较低;同一区域内站点的相似度一般要高于不同区域内站点的相似度,结合表 2 中数据分析如下:

在同一区域内,站点的相似度应该较高,N3 与 N1~4 之间由于空间上较近,相似度都高于 0.9,符合模型的评价标准。N1 与 N2~4、N6 站点的相似度也都接近 0.9,表明 N1 与大部分区域内站点的相似度符合评价标准。由于 N7 在空间上偏离其他站点,所以与其他站点的相似度相对较低,但 N7 与其他站点相似度仍

然接近 0.8。

此外个别站点如 N5 与其他站点的相似度出现偏差,表明模型存在一定的误差,进一步的研究将从特征选择和地理空间差异因素来分析这种偏差产生的原因。但是大部分区域内站点的相似度在评价范围之内,北部站点相似度正确率在 70% 以上。

如表 3 为中部的 6 个站点间的相似度,M6 站点由于空间上相对 M1~4 站点较远,所以这部分相似度较低,略低于 0.8,而与 M6 较近的 M5 站点,相似度仍然达到了 0.8 以上。大部分站点相似性都在 0.8 以上,正确率几乎达到了 100%。

表2 北部区域内部站点相似度
Table2 Similarity of the stations in the north area

站点相似度	N1	N2	N3	N4	N5	N6	N7
N1	1	0.903074	0.904808	0.883736	0.694924	0.866606	0.787833
N2	0.903074	1	0.90049	0.892857	0.696679	0.923903	0.826891
N3	0.904808	0.90049	1	0.914951	0.669473	0.901389	0.808824
N4	0.883736	0.892857	0.914951	1	0.689576	0.912932	0.79902
N5	0.694924	0.696679	0.669473	0.689576	1	0.674036	0.794718
N6	0.866606	0.923903	0.901389	0.912932	0.674036	1	0.807096
N7	0.787833	0.826891	0.808824	0.79902	0.794718	0.807096	1

表3 中部区域内部站点相似度
Table3 Similarity of the stations in the middle area

站点相似度	M1	M2	M3	M4	M5	M6
M1	1	0.88308	0.833835	0.834359	0.839372	0.719073
M2	0.88308	1	0.835478	0.81697	0.85122	0.754372
M3	0.833835	0.835478	1	0.84787	0.867531	0.720929
M4	0.834359	0.81697	0.84787	1	0.852214	0.770502
M5	0.839372	0.85122	0.867531	0.852214	1	0.836296
M6	0.719073	0.754372	0.720929	0.770502	0.836296	1

表4 南部区域内部站点相似度
Table4 Similarity of the stations in the south area

站点相似度	S1	S2	S3	S4	S5	S6	S7	S8
S1	1	0.920763	0.884571	0.890932	0.875543	0.924789	0.907639	0.897134
S2	0.920763	1	0.859095	0.865851	0.850829	0.904446	0.882869	0.879204
S3	0.884571	0.859095	1	0.890278	0.921974	0.890929	0.933477	0.826083
S4	0.890932	0.865851	0.890278	1	0.911865	0.892214	0.914057	0.850413
S5	0.875543	0.850829	0.921974	0.911865	1	0.901122	0.94101	0.84481
S6	0.924789	0.904446	0.890929	0.892214	0.901122	1	0.89213	0.896791
S7	0.907639	0.882869	0.933477	0.914057	0.94101	0.89213	1	0.849597
S8	0.897134	0.879204	0.826083	0.850413	0.84481	0.896791	0.849597	1

如表 4 中为南部区域站点间的相似度, 数据显示南部站点的相似度普遍在 0.85 以上, 相似度正确率也几乎达到了 100%。这是由于南部站点靠近长江, 又有

众多湖泊, 区域降雨普遍较多。南部区域站点的高度相似性验证了模型在区域内部相似性度量的可靠性。不同区域站点相似性比较如表 5~7 所示。

表5 北部站点与中部站点相似性度

Table5 Similarity of the stations between the north and middle areas

站点相似度	M1	M2	M3	M4	M5	M6
N1	0.902992	0.837353	0.827357	0.82547	0.824003	0.709109
N2	0.882676	0.870861	0.886555	0.843642	0.847008	0.710314
N3	0.877564	0.825823	0.826942	0.864167	0.825509	0.711473
N4	0.862519	0.817446	0.807718	0.853725	0.832098	0.723207
N5	0.630952	0.622059	0.732904	0.678639	0.753856	0.669975
N6	0.862576	0.824236	0.839599	0.816825	0.825518	0.678734
N7	0.790498	0.816485	0.843919	0.832157	0.900449	0.79452

表 5 中为北部与中部站点间的相似度。两区域都位于长江以北, 所以在降雨类型上呈现出相似性的一致性,

但也反映出模型所存在的问题: 不同区域的站点, 考虑到其他因素, 即使地理位置较远, 也可能有较大相似性。

表6 北部站点与南部站点相似度

Table6 Similarity of the stations between the north and southern areas

站点相似度	S1	S2	S3	S4	S5	S6	S7	S8
N1	0.729325	0.742772	0.762788	0.78929	0.758561	0.74336	0.766783	0.664185
N2	0.755944	0.765987	0.79899	0.811591	0.797151	0.757487	0.797634	0.689482
N3	0.764918	0.766844	0.743071	0.807398	0.768543	0.763703	0.775881	0.711156
N4	0.747859	0.745411	0.747522	0.806823	0.764641	0.790696	0.743306	0.701138
N5	0.657519	0.637719	0.66023	0.650794	0.636482	0.622107	0.648191	0.608588
N6	0.76567	0.753377	0.785245	0.837869	0.834148	0.782092	0.814349	0.731695
N7	0.789299	0.779099	0.778208	0.802734	0.787651	0.76668	0.796761	0.720472

表 6 中为北部与南部站点相似度, 由于南部和北部空间上相距相对较远, 而且相隔长江, 降雨有较大的差异, 站点的相似性普遍在 0.8 以下甚至到达 0.6, 相似度正确性接近 90%。

表 7 中为南部与中部站点相似度, 由于不同区域空间上的差异, 超过 85%以上的站点相似性在 0.8 以下, 验证了模型在不同区域内相似度的正确性。

从以上站点相似度比较的表格中, 我们不难发现, 站点间地理位置和长江流域南北这两个因素是构成降雨相似性差异的主要原因。除个别站点外, 同一区域内站点相似度普遍在 0.8 甚至 0.9 以上, 不同区域内站点的相似度普遍在 0.8 甚至接近 0.6; 地理空间相对较远的北部与南部站点相似度小于地理空间相对较近的北部与中部站点相似度; 而且中部与南部站点相似性普遍小于中部与北部的站点相似度, 这是由于以长江为分水岭, 南北区域的降雨有较大差异, 这 and 实际情况比较吻合, 从而验证了本文基于单场降雨类型直方图分析相似性模型的正确性。

表7 南部与中部站点相似度

Table7 Similarity of the stations between the south and middle areas

站点相似度	S1	S2	S3	S4	S5	S6	S7	S8
M1	0.751155	0.76593	0.803038	0.814678	0.800017	0.775839	0.806066	0.691108
M2	0.759015	0.758697	0.82703	0.797989	0.817501	0.780566	0.814873	0.693484
M3	0.741128	0.747001	0.770553	0.783148	0.779698	0.73611	0.789507	0.686717
M4	0.754152	0.760851	0.718512	0.787483	0.742252	0.748456	0.756946	0.71957
M5	0.783863	0.759731	0.808943	0.810009	0.791062	0.794841	0.797037	0.719664
M6	0.779466	0.731079	0.751476	0.73401	0.749433	0.76933	0.754903	0.698974

4 结语

此前对降雨数据的特征提取和分析大多比较宏观,陈晓宏等在东江流域对于降雨时空分布的研究较多,将模式分类引入降雨分析,从气象学的角度,结合年降雨量,以高程、气温和蒸发量作为特征进行分析。本文从包含更多信息的单场降雨过程入手,基于降雨数据本身的特性分析,综合降雨时间序列的多个统计特征,利用基于 Davies-Bouldin 指数的 K-means 聚类算法,首次建立了基于单场降雨类型直方图分析的降雨站点相似性比较模型。

实验结果比较符合实际的降雨情况,反映了区域降雨的特点,具有较好的实际应用价值。但个别站点相似性仍然有一定的误差,通过进一步分析区域降雨站点的实际地理与水文情况,完善数据标准化过程,选取更合适的统计特征,模型的正确性将得到提高。下一步研究将通过把站点降雨的相似性进一步推广到区域降雨的相似性,建立区域降雨相似性度量方法。

参考文献:

- [1] Maimon O, Rokach L. Data Mining and Knowledge Discovery Handbook[M]. Springer, 2010.
- [2] 李士进, 朱跃龙, 张晓花, 等. 基于 BORDA 计数法的多元水文时间序列相似性分析[J]. 水利学报, 2009, 40 (3): 378-384. (LI Shijin, ZHU Yuelong, ZHANG Xiaohua, et al. BORDA counting method based similarity analysis of multivariate hydrological time series[J]. Journal of Hydraulic Engineering, 2009, 40 (3): 378-384. (in Chinese))
- [3] ZHANG Ye, TIAN Wen, LIU, Shengpeng. Fire time series forecasting based on Markov-SVR model [A]. Proceedings of Multimedia and Signal Processing (CMSP) of 2011 International Conference on Guilin, China [C]. 2011:278-281.
- [4] ZHAO Chengping, LIANG Chuan, GUO Haiwei. Research on hydrology time series prediction based on grey theory and support vector regression [A]. Proceedings of Computer Distributed Control and Intelligent Environmental Monitoring (CDCIEM) of 2011 International Conference on Chengdu, China [C]. 2011:1673-1676.
- [5] 王永县, 詹一辉, 张少屹. 降水时间序列的聚类分析和预测[J]. 系统工程理论与实践, 1994, 11: 67-71. (WANG Yongxian, ZHAN Yihui, ZHANG Shaohong. Clustering analysis and prediction for precipitation time series [J]. Systems Engineering -Theory & Practice, 1994, 11: 67-71. (in Chinese))
- [6] Sapankevych, N., Sankar, R. Time series prediction using support vector machines: a survey [J]. Computational Intelligence Magazine, IEEE, 2009, 4(2): 24-38.
- [7] 朱跃龙, 王咏梅, 万定生, 等. 基于语义相似的水文时间序列相似性挖掘[J]. 水文, 2011, 31(1): 35-40. (ZHU Yuelong, WANG Yongmei, WAN Dingsheng, et al. Similarity mining of hydrological time series based on semantic similarity measures[J]. Journal of China Hydrology, 2011, 31(1): 35-40. (in Chinese))
- [8] 谈戈, 夏军, 李新. 无资料地区水文预报研究的方法和出路[J]. 冰川冻土, 2004, 26(2): 192-196. (TAN Ge, XIA Jun, LI Xin. Hydrological prediction in ungauged basins [J]. Journal of Glaciology and Geocryology, 2004, 26(2): 192-196. (in Chinese))
- [9] 陈晓宏, 刘德地, 王兆礼. 降雨空间分布模式识别 [J]. 水利学报, 2006, 37 (6): 711-716. (CHEN Xiaohong, LIU Dedi, WANG Zhaoli. Recognition method for spatial distribution pattern of precipitation [J]. Journal of Hydraulic Engineering, 2006, 37 (6): 711-716. (in Chinese))
- [10] 张国建, 尹红美, 李珂凌. 大气降水时空分布的数理统计与分析[J]. 水资源保护, 2005, 21(5): 34-36. (ZHANG Guojian, YIN Hongmei, LI Keling. Analysis of spatial-temporal distribution of atmospheric precipitation using statistics [J]. Water Resources Protection, 2005, 21(5): 34-36. (in Chinese))
- [11] 苏布达, 姜彤. 长江流域降水极值时间序列的分布特征[J]. 湖泊科学, 2008, 20 (1): 123-128. (SU Buda, JIANG Tong. Distribution feature of time series of extreme precipitation over the Yangtze River basin [J]. Journal of Lake Sciences, 2008, 20 (1): 123-128. (in Chinese))
- [12] 刘德地, 陈晓宏, 楼章华. 基于云模型的降雨时空分布特性分析[J]. 水利学报, 2009, 40 (7): 850-857. (LIU Dedi, CHEN Xiaohong, LOU Zhanghua. Analysis on characteristics of spatial-temporal precipitation distribution based on cloud model [J]. Journal of Hydraulic Engineering, 2009, 40(7): 850-857. (in Chinese))
- [13] 巴金福, 祁万明, 汪青春, 等. 黄河上游降水时空分布特征及变化趋势[J]. 青海电力, 2000, (1): 10-12. (BA Jinfu, QI Wanming, WANG Qingchun, et al. Features and fluctuating tendency of the precipitation spatial distribution of Yellow River upstream [J]. Qinghai Electric Power, 2000, (1): 10-12. (in Chinese))
- [14] Davies D.L., Bouldin D.W. A cluster separation measure [J]. IEEE Trans. Pattern Anal. Machine Intell, 1979, 1 (4): 224-227.
- [15] TAN Pangning, Michael Steinbach, Vipin Kumar. 数据挖掘导论[M]. 北京: 人民邮电出版社, 2010. (TAN Pangning, Michael Steinbach, Vipin Kumar. Introduction to Data Mining [M]. Beijing: Posts & Telecom Press, 2010. (in Chinese))
- [16] 唐晶晶. 1952~2007 年间长江流域降水时空分布变化-全球变暖下降水重新分配实例分析 [D]. 上海: 华东师范大学硕士论文, 2010. (TANG Jingjing. The Temporal and Spatial Distribution of Precipitation over Yangtze River Catchment from 1952 to 2007: Case Study of Precipitation Redistribution under Global Warming [D]. Shanghai: Master's Dissertation of East China Normal University, 2010. (in Chinese))
- [17] 邵晓梅, 严昌荣, 魏红兵. 基于 Kriging 插值的黄河流域降水时空分布格局 [J]. 中国农业气象, 2006, 27 (2): 65-69. (SHAO Xiaomei, YAN Changrong, WEI Hongbing. Spatial and temporal structure of precipitation in the Yellow River basin based on Kriging method [J]. Chinese Journal of Agrometeorology, 2006, 27 (2): 65-69. (in Chinese))

- Chinese))
- [18] 丁永建,叶佑生,周文娟. 黑河流域过去 40a 来降水时空分布特征 [J]. 冰川冻土, 1999, 21 (1):42-48. (DING Yongjian, YE Baisheng, ZHOU Wenjuan. Temporal and spatial precipitation distribution in the Heihe catchment, northwest China, during the past [J]. Journal of Glaciology and Geocryology, 1999, 21(1):42-48. (in Chinese))
- [19] Richard R. Heim Jr, 周跃武, 冯建英. 美国 20 世纪干旱指数评述 [J]. 干旱气象, 2006, 24 (1):79-89. (Richard R. Heim Jr, ZHOU Yuewu, FENG Jianying, A Review of twentieth-century drought indices used in the United States [J]. Arid Meteorology, 2006, 24 (1):79-89. (in Chinese))
- [20] 焦菊英, 王万忠. 黄土高原降雨空间分布的不均匀性研究 [J]. 水文, 2001, 21 (2):20-24. (JIAO Juying, WANG Wanzhong. Non-uniformity of spatial distribution of rainfall on the loess plateau [J]. Journal of China Hydrology, 2001, 21(2):20-24. (in Chinese))

Study on Similarity Model of Precipitation Series Based on Precipitation Type Histogram

ZHANG Longfei¹, ZHU Yuelong¹, LI Shijin¹, GAO Xiangtao²

(1. School of Computer & Information, Hohai University, Nanjing 210098, China;

2. Hydrology and Water Resources Survey Bureau of Jiangsu Province, Nanjing 210029, China)

Abstract: Similarity model for precipitation stations based on analysis of the single precipitation type histogram was proposed to study the similarity between the precipitation stations. Firstly, the precipitation item was extracted from the precipitation time series, the features of single rainfall were chosen and extracted after data preprocessing. Then, the data standardization was conducted on the feature vectors. Further more, the standardized feature data were clustered using K-means clustering method based on the Davies-Bouldin index cluster validity index. Finally, the clustering results were analyzed using the single precipitation type histogram similarity model. The experimental results show that the similarity for the stations in the related regions is consistent with the underlying fact, which verifies that the proposed model is of great value for water resources management.

Key words: precipitation time series; similarity analysis; feature extraction; Davies-Bouldin index; K-means clustering; histogram of precipitation type

(上接第 4 页)

- [7] 陈元芳,沙志贵,顾圣华,等. 可考虑历史洪水对数正态分布线性矩法的研究[J]. 河海大学学报(自然科学版), 2003, (01):80-83. (CHEN Yuanfang, SHA Zhigui, GU Shenghua, et al, Study on L-moment estimation method with consideration of historical floods for log-normal distribution [J]. Journal of Hehai University (Natural Sciences), 2003, (01):80-83. (in Chinese))
- [8] Institute of Hydrology. Flood Estimation Handbook (Volumes 1-5) [M]. Institute of Hydrology, Wallingford, UK, 1999.
- [9] Maidment D R. Handbook of Hydrology [M]. McGraw-Hill, INC., 1992.
- [10] 周芬,郭生练. P-III 型分布参数估计方法的比较研究[J]. 水电能源科学, 2003, 21 (3):10-13. (ZHOU Fen, GUO Shenglian. Comparative study on parameter estimation methods of Pearson type III distribution [J]. Water Resources and Power, 2003, 21(3):10-13.(in Chinese))

Comparison among Theoretical Frequency Distributions of P-III, Log P-III and GL

LIU Jie¹, CHEN Hua², WANG Jinxing³, CHEN Xing², LI Tianyuan²

(1. School of Urban and Rural Construction, Chengdu University, Chengdu 610106, China;

2. State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan 430072, China;

3. Bureau of Hydrology, MWR, Beijing 100053, China)

Abstract: The theoretical frequency distributions of P-III, Log P-III and GL (Generalized Logistic Distribution) were compared and analyzed by using annual maximum sampling and L-moment method in the river basins of the Huaihe and Pearl, where there are 18 and 16 hydrometry stations with data series (1951-2010) collected, respectively. The results show that the design flood from the curve of P-III is less than that from the other two curves and the errors of P-III are less than these of log P-III by analyzing the errors of the curve fitting in these two basins.

Key words: Huaihe River; Pearl River; P-III; Log P-III; GL