DOI: 10.16198/j.cnki.1009-640X.2015.04.017
范梦歌, 刘九夫. 基于聚类分析的水文相似流域研究[J]. 水利水运工程学报, 2015(4): 106-111. (FAN Meng-ge, LIU Jiu-fu. Analysis of hydrologically similar basins based on clustering analysis[J]. Hydro-Science and Engineering, 2015(4): 106-111.)

基于聚类分析的水文相似流域研究

范梦歌, 刘九夫
(南京水利科学研究院, 江苏 南京 210029)

摘要: 在洪水预警预报中,一些中小流域常常由于水文资料短缺造成分析计算困难。随着流域下垫面地形、植被、土壤等数据的大量获取和数据挖掘技术的不断发展,应用聚类分析等无指导学习方法对流域下垫面大量数据进行分析,根据对象之间距离最近的原则对流域进行分组,得到水文相似流域,从而将有水文资料流域的水文参数应用于水文资料短缺的相似流域,使洪水预报得以实现。选取浙江省 118 个具有 20 年以上雨水情资料的流域进行研究,采用流域长度、流域宽度、河长、河流比降、流域平均坡度、流域形状系数、多年平均 1,3,6 和 12 h 面最大降水等 10 个指标,应用主成分分析对数据进行降维,进而对流域进行聚类分析,将浙江省流域分为 23 个相似组。在分组基础上,选取其中两组中径流资料大于 20 年的站点进行多年平均最大洪峰、洪量比对,验证水文相似流域分类合理性。结果表明相似流域组内年最大洪峰、年最大平均 1,3,6 和 12 h 洪量具有较大相似性,分类较为合理,从统计学角度为浙江省的洪水预警预报提供新的理论与思路。

关 键 词: 聚类分析; 主成分分析; 相似流域; 参数移植; 径流比对
中图分类号: TV12 **文献标志码:** A **文章编号:** 1009-640X(2015)04-0106-06

在中小流域的水文分析计算中,由于缺少气候和地理资料,使得设计洪水等计算产生困难,从而影响洪峰预测的准确性,因此常常需要应用相似流域的参数进行移植使用[1]。近年来,数据挖掘引起了信息产业界的极大关注,它可以将大量数据转换成有用的信息和知识,从而广泛用于各个领域[2]。随着水利行业信息化建设的发展,我国积累了大量宝贵的雨情、流域地理特征数据,运用数据挖掘的聚类分析技术对各水文站的大量数据进行无指导的分析、分类,能够对影响流域洪水洪峰特征的关键因素进行相似分组,从而指导水文资料短缺地区洪水预警预报。国内外学者伊璇[3], Y. He [4]和 J. M. Kileshye Onema[5]等均曾利用该方法进行相似流域分类来解决观测资料不完整导致的无法采用常规方法率定水文模型参数的问题。运用该方法时,由于流域许多指标之间的相关性,可能会造成信息的大量重叠,甚至掩盖其内在规律[6],因此本文在使用聚类分析法的基础上,首先应用主成分分析(principal components analysis,简称 PCA)对流域特征值进行预处理,进而识别浙江省相似流域,使该方法对水文分区的计算更符合实际应用的要求。

1 研究区概况

浙江省位于中国东海之滨,陆域面积 10.18 万 km^2,地形复杂,地势由西南向东北倾斜,呈阶梯下降,位于亚热带季风区,冬季低温少雨,夏季高温多雨,多年平均降水 1 600 mm,山区多年平均 1 800~2 200 mm,平原 1 100~1 300 mm,海岛 950~1 300 mm。年降水深年际变化较大,各站最大年降水深和最小年降水深的比值在 2~3 之间,常发生连续丰水、枯水的现象。本文采用浙江省 118 个水文站、787 个雨量站资料,水文站以上控制流域面积内雨量站个数最少 1 个,最多 51 个,水文站分布情况如图 1。

收稿日期: 2014-10-27
作者简介: 范梦歌(1990—), 女, 河南三门峡人, 硕士研究生, 主要从事洪水预报研究。E-mail:zdfanmengge@163.com

2　研究数据与方法

2.1　聚类指标的选取

1980 年后国内不少单位应用中小河流的降水和流量资料，开展了瞬时单位线的分析研究，为中小型水利水电工程的规划设计分析提供了新的途径[7]。由于其简便易操作且预报精度较高，得到了广泛应用。对流域应用瞬时单位线法时，需要用到两个参数，即 n 与 k。纳希把流域看作是 n 个等效线性水库的串联[8]，因此从概念上来讲，它是反映水流调节性能的一个指标，从定量上说，它与流域面积大小相适应[9]。k 为线性水库的蓄泄系数，相当于流域汇流时间的参数，具有时间因次，与流域坡度、河道比降、降水特性等参数相关性高。本文在聚类分析中选取 10 个参数，即反映流域地理特性的流域长度、流域宽度、河长、河流比降、流域平均坡度、流域形状系数，以及代表流域降水特性的多年平均 1，3，6 和 12 h 面最大降水。在计算流域平均长宽时，需先定义流域方向。流域方向指流域内干支流所有节点指向河口方向的平均值，流域方向上的外包矩形中与流域方向一致的边为流域长度，与流域长度垂直的另一边为流域宽度。

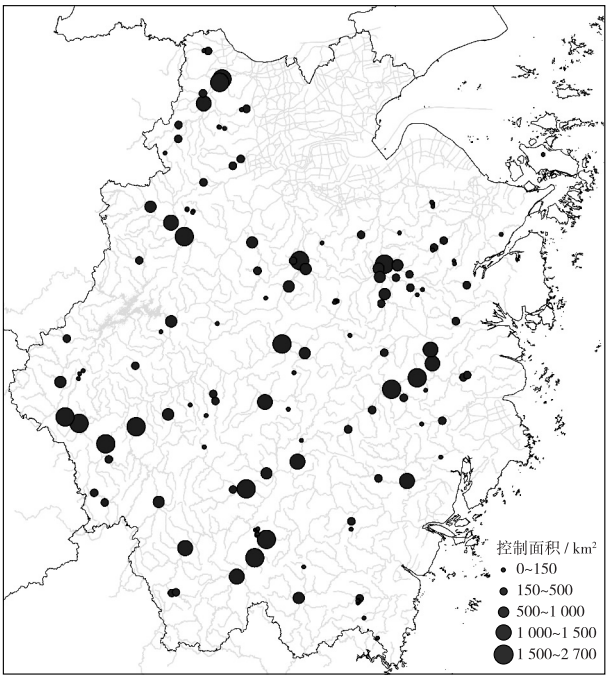

图 1　浙江省水文站分布

Fig. 1　Hydrological station distribution of Zhejiang Province

2.2　主成分分析

主成分分析是研究如何将多指标问题转化为较少的综合指标的一个重要统计方法。它能将高维空间的问题转化到低维空间去处理[10]，使问题变得简单直观，而且这些较少的综合指标之间互不相关，能提供原有指标的绝大部分信息。

主成分分析的基本算法和步骤如下：

(1)采集 p 维随机向量 $\boldsymbol{x}=(x_1,x_2,\cdots,x_p)^{\mathrm{T}}$ 的 n 个样品 $\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{ip})^{\mathrm{T}}$ 列出观察资料矩阵 $\boldsymbol{X}=(x_{ij})_{n\times p}$；

(2)对样本阵中原始数据进行预处理，即将原始数据转换为正指标，然后将所得数据标准化，得标准化阵：

$$\boldsymbol{Z}=\begin{bmatrix}\boldsymbol{z}_1\\ \boldsymbol{z}_2\\ \\ \boldsymbol{z}_n\end{bmatrix}=\begin{bmatrix}z_{11} & z_{12} & \cdots & z_{1p}\\ z_{21} & z_{22} & \cdots & z_{2p}\\ \vdots & \vdots & \ddots & \vdots\\ z_{n1} & z_{n2} & \cdots & z_{np}\end{bmatrix}$$

(3)计算上述矩阵的样本相关系数矩阵 $\boldsymbol{R}=[r_{ij}]_{p\times p}=\dfrac{\boldsymbol{Z}'\boldsymbol{Z}}{n-1}$；

(4)解释本相关系数阵 $\boldsymbol{R}$ 的特征方程，得 p 个特征值 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p$；

(5)确立主成分。通常可按累计方差贡献率 $\sum_{j=1}^{k}\lambda_j/\sum_{j=1}^{p}\lambda_j>85\%$ 的准则，确定 k，从而取前 k 个主成分：$Z_j=l_j'Y=l_{1j}Y_1+\cdots+l_{pj}Y_p\quad(j=1,2,\cdots,k)$；

(6)计算前 k 个主成分的样本值，从而可得到新指标(主成分)矩阵 $\boldsymbol{Z}=(z_{ij})_{n\times k}$，以其代替原流域特征观

测矩阵 $\boldsymbol{X}=(x_{ij})_{n\times p}$ 做聚类分析,可使问题简化及合理化。

2.3 系统聚类分析

聚类分析是数据挖掘最主要的方法之一,聚类就是将数据对象分组为多个类或簇,在同一个簇中的对象具有较高的相似度,而不同簇中的对象差别较大[11]。具体步骤如下:

(1)对于有 p 个变量的对象来说,n 个对象可以看作是 p 维空间的 n 个点,可以用点之间距离来度量对象间的接近程度[11]。其距离通常可以采用绝对值距离、欧氏距离、明考斯基距离和切比雪夫距离等。在本文中采用欧氏距离来度量其相异度,即 $d_2(x_i,x_j)=\|x_i-x_j\|_2=\left(\sum_{k=1}^{p}|x_{ik}-x_{jk}|^2\right)^{\frac{1}{2}}$。

(2)合并距离最近的 2 类为一新类,类间距离与上式距离不同,本文采用类平均法,即

$$\begin{cases} D_{pq}^2=\dfrac{1}{n_p n_q}\sum_i\sum_j d_{ij}^2 \\ D_{kr}^2=\dfrac{n_p}{n_r}D_{kp}^2+\dfrac{n_q}{n_r}D_{kp}^2 \end{cases}$$

式中:D_{pq} 为类 G_p,G_q 间的距离,D_{kr} 为任意其他类 G_q 到 G_r 间的距离;类 G_r 为类 G_p,G_q 合并而成的新类;n_p,n_q,n_r 分别为类 G_p,G_q,G_r 的流域样本个数[12]。

(3)计算新类与当前各类距离,合并距离最短的两项,若类的个数为 1,则聚类结束,否则返回步骤(2)。

(4)画聚类图。

3 结果与分析

3.1 基于主成分分析的浙江省流域聚类分析过程

采用浙江省具有 20 年以上降水资料的 118 个流域进行聚类研究,计算流域特征指标之间的相关性,其相关系数见表 1。表 1 可见,一些流域特征指标之间存在一定相关性,相关系数的范围为 0.14~0.98。相关程度最大的是流域平均长度与河长,为 0.98。多年平均最大 1 h 面雨量与流域长度、流域宽度、河长也具有较高相关性,分别为-0.83,-0.82 和-0.85。

表 1 相关系数矩阵

Tab. 1 Correlation coefficient matrix

参数	流域长度/km	流域宽度/km	河长/km	河流比降/‰	流域平均坡度/%	流域形状系数	M_{1h}/mm	M_{3h}/mm	M_{6h}/mm	M_{12h}/mm
流域长度/km	1.00	0.89	0.98	-0.59	-0.68	-0.48	-0.83	-0.70	-0.53	-0.40
流域宽度/km	0.89	1.00	0.92	-0.58	-0.67	-0.42	-0.82	-0.68	-0.51	-0.37
河长/km	0.98	0.92	1.00	-0.58	-0.68	-0.53	-0.85	-0.70	-0.52	-0.39
河流比降/‰	-0.59	-0.58	-0.58	1.00	0.94	0.58	0.54	0.42	0.27	0.15
流域平均坡度/%	-0.68	-0.67	-0.68	0.94	1.00	0.56	0.64	0.52	0.36	0.24
流域形状系数	-0.48	-0.42	-0.53	0.58	0.56	1.00	0.46	0.36	0.25	0.14
M_{1h}/mm	-0.83	-0.82	-0.85	0.54	0.64	0.46	1.00	0.93	0.79	0.68
M_{3h}/mm	-0.70	-0.68	-0.70	0.42	0.52	0.36	0.93	1.00	0.95	0.88
M_{6h}/mm	-0.53	-0.51	-0.52	0.27	0.36	0.25	0.79	0.95	1.00	0.98
M_{12h}/mm	-0.40	-0.37	-0.39	0.15	0.24	0.14	0.68	0.88	0.98	1.00

计算各成分特征值、贡献率及累积贡献率,前 3 个成分的方差累计贡献率已达到了 91.6%,超过了一般要求的 85%,所以选取 Z_1,Z_2,Z_3 为第 1,2,3 个主成分。表 2 为原 10 个流域指标在 3 个主成分上的荷载值。荷载值反映了所取主成分与各原始指标之间的关系,反映各指标对选取主成分所起的作用[6]。其中年最大 1h 面平均降水和河长对主成分 1 影响最大。

计算 3 个新的主成分在各流域上的得分，从而对新样本进行系统聚类分析，得到其系统聚类分析树状图，将浙江省 118 个流域划分为 23 个相似组，其中组内流域数最少的 1 个，最多的 16 个。下面将具体对结果进行组内分析及组间对比。

3.2　浙江省相似流域结果实例分析

浙江省两组相似流域 A，B 见表 3。根据聚类分析结果，结合图 1 省内水文站分布，可知在一个相似流域组内，部分站点具有地理位置邻近的特性，但同时也有站点地理位置相距较远，不具备地理相似性。表 3 中长风站与常山（二）站所控制的流域均位于浙江西部地区，两控制站仅相距 5.9 km。而柏枝岙（三）[13] 站所控制流域位于东部沿海地区，与长风站相距 226 km，地理位置相距较远。

表 2　主成分荷载矩阵

Tab. 2　Principal components load matrix

成分	1	2	3
流域长度/km	−0.90	0.17	0.34
流域宽度/km	−0.87	0.17	0.37
河长/km	−0.91	0.19	0.33
河流比降/‰	0.69	−0.54	0.35
流域平均坡度/%	0.78	−0.46	0.26
流域形状系数	0.57	−0.42	0.39
M_{1h}/mm	0.95	0.15	−0.08
M_{3h}/mm	0.90	0.42	0.07
M_{6h}/mm	0.77	0.61	0.18
M_{12h}/mm	0.65	0.71	0.22

表 3　浙江省相似流域 A 和 B

Tab. 3　Similar basins A and B in Zhejiang Province

流域	站名	流域长度/km	流域宽度/km	河长/km	河流比降/‰	流域平均坡度/%	流域形状系数	M_{1h}/mm	M_{3h}/mm	M_{6h}/mm	M_{12h}/mm
A	长风	66.3	57.0	111.3	2.3	4.1	0.17	16.8	38.7	61.3	86.9
	常山（二）	75.7	58.7	127.5	1.9	3.7	0.14	16.6	38.5	61.1	86.6
	下石埠	87.4	48.0	144.9	3.1	5.5	0.12	20.7	40.7	56.3	77.6
	义乌佛堂	71.6	64.9	110.4	2.0	2.7	0.19	15.2	32.8	47.6	66.2
	分水	67.7	71.4	124.0	3.8	5.0	0.17	15.8	35.5	54.9	75.7
	嵊县（二）	66.3	69.4	91.9	3.6	3.4	0.27	15.3	32.7	49.1	70.3
	柏枝岙（三）	88.3	51.6	124.9	2.0	4.1	0.16	16.7	36.5	57.4	85.1
	半阳	82.9	48.8	130.6	3.8	5.3	0.13	17.4	35.3	52.6	76.9
B	横塘村	55.1	40.4	73.5	1.5	5.8	0.24	18.7	39.2	56.9	80.3
	范家村	71.8	42.8	100.3	0.9	4.3	0.19	16.9	36.2	53.9	77.2
	港口	74.2	42.6	104.4	0.8	4.3	0.18	16.9	36.2	53.9	77.2
	双塔底	66.1	37.1	97.8	3.3	6.8	0.17	16.6	34.0	51.3	73.6
	青山殿	53.7	42.1	102.4	5.2	8.0	0.14	18.3	39.8	59.4	80.7
	诸暨	53.8	70.1	86.2	1.0	3.1	0.23	16.7	33.2	48.0	67.5
	仙居	62.3	45.8	97.2	2.9	5.0	0.17	17.7	37.4	56.8	83.0
	百步	67.2	45.8	94.7	2.9	4.5	0.15	18.6	40.1	59.7	85.4

由表 3 可知，相似流域组内均按照相近原则组成，因此在应用瞬时单位线等方法进行洪水预报时，可将组内已知参数的流域数据应用于未知参数的流域，增加其预报的准确度。同时，两组数据之间具有一定差别，观察 A、B 组两组数据，在多年最大 1，3，6 和 12 h 面平均降水基本相似情况下，A 组的流域长度、流域宽度、河长平均比 B 组大，流域形状系数比 B 组小。

选取相似组 A 和 B 中径流资料大于 20 年的站点进行多年最大洪峰、洪量分析，验证由水文相似流域决定因素进行分类的流域洪峰、洪量是否相似。由表 4 可知，A 组中长风站较常山（二）站的差距均在 5% 以内，柏枝岙（三）相较常山（二）站年最大洪峰平均值差距在 10% 以内。B 组中，横塘村较诸暨年最大洪峰平均值差距也在 10% 之内。范家村较诸暨差距较小，年最大 6 h 洪量仅为 0.2%。同时，将 A，B 组进行组间比较，组内站点的年最大洪峰流量、最大 1，3，6 和 12 h 洪量具有较大差别，A，B 组年最大洪峰流量平均值分别

为 2 810.7 和 760.0 m^3/s,B 组年最大 1,3,6 和 12 h 洪量相较 A 组差距分别为 73.1%, 73.0%,72.6%和 71.2%,均大于 70%,两组间具有较大差别,达到了水文相似流域分组的目的。

表 4 流域 A 和 B 组部分测站洪峰洪量对比

Tab. 4 Comparative analysis of flood peak and flood volume among some stations of basins A and B

流域	站名及站间对比	$Q_{max}/(m^3 \cdot s^{-1})$	$Q_{1h}/10^6 m^3$	$Q_{3h}/10^6 m^3$	$Q_{6h}/10^6 m^3$	$Q_{12h}/10^6 m^3$
A	长风	2 636.1	9.3	27.5	52.8	93.7
	常山(二)	2 768.8	9.7	28.6	54.5	98.2
	柏枝岙(三)	3 027.3	10.8	32.3	62.8	116.3
	长风较常山(二)差距(%)	4.8	3.3	3.7	3.2	4.6
	柏枝岙(三)较常山(二)差距(%)	9.3	12.2	13	15.3	18.5
B	横塘村	796.7	2.9	8.4	16.2	30.1
	诸暨	728.3	2.6	7.8	15.2	29.2
	范家村	755.1	2.6	7.7	15.2	29.5
	横塘村较诸暨差距(%)	9.4	9.2	8.2	6.8	2.9
	范家村较诸暨差距(%)	3.7	1.5	1.0	0.2	0.8

4 结 语

本文采用主成分分析对影响流域洪水预报的重要参数进行预处理,在此基础上对各流域进行系统聚类分析,并进行流域相似组组内、组间径流数据分析比对,结果表明分组效果良好。该方法操作简便,结果直观,从统计学的角度为相似流域的选择提供了一种有效的分析方法,为无资料地区的洪水预警预报的参数移植提供了新的理论和思路。本研究仍可在以下方面改进:

(1) 在考虑流域降雨中心的基础上,选择相似流域,分别按照降雨中心位于上、中、下游等情况分类研究,更具实用性。

(2) 对于某些特定地区可对影响流域降水较大的参数赋予一定的权重,以增强地区适用性。

(3) 在实际洪水预警预报中,可将相似组内已知流域的参数作为初始参数,在应用中进一步对其进行优化,以提高预报的精度。

参 考 文 献:

[1] 梁忠民, 钟平安, 华家鹏. 水文水利计算[M]. 北京: 中国水利水电出版社, 2006. (LIANG Zhong-min, ZHONG Ping-an, HUA Jia-peng. Hydrological and hydraulic calculation[M]. Beijing: China Water & Power Press, 2006. (in Chinese))

[2] 韩家炜. 数据挖掘概念与技术[M]. 北京: 机械工业出版社, 2007. (HAN Jia-wei. Data mining concepts and techniques [M]. Beijing: China Machine Press, 2007. (in Chinese))

[3] 伊璇, 周丰, 王心宇,等. 基于 SOM 的流域分类和无资料区径流模拟[J]. 地理科学进展, 2014, 33(8): 1109-1116. (YI Xuan, ZHOU Feng, WANG Xin-yu, et al. Classification and runoff simulation of data-scarce basins based on self-organizing maps[J]. Progress in Geography, 2014, 33(8): 1109-1116. (in Chinese))

[4] HE Y, BARDOSSY A, ZEHE E. A catchment classification scheme using local variance reduction method[J]. Journal of Hydrology, 2011, 411(1-2): 140-154.

[5] KILESHYE ONEMA J M, TAIGBENU A E, NDIRITU J. Classification and flow prediction in a data-scarce watershed of the equatorial Nile region[J]. Hydrology and Earth System Sciences, 2012, 16(5): 1435-1443.

[6] 包为民, 万新宇, 荆艳东,等. 基于主成分分析的河流洪水系统聚类法[J]. 河海大学学报: 自然科学版, 2008, 36(1): 1-5. (BAO Wei-min, WAN Xin-yu, JING Yan-dong, et al. Flood clustering method based on principal component analysis[J]. Journal of Hohai University (Natural Sciences), 2008, 36(1): 1-5. (in Chinese))

[7] 傅联森, 陈润, 周焕. 瞬时单位线法在浙江省应用的几个问题研究[J]. 水文, 2012, 32(3): 43-46. (FU Lian-sen, CHEN

Run, ZHOU Huan. Application of instantaneous unit hydrograph in Zhejiang Province [J]. Journal of China Hydrology, 2012, 32(3): 43-46. (in Chinese))

[8] 包为民. 水文预报[M]. 北京：中国水利水电出版社, 2009. (BAO Wei-min. Hydrologic forecasting[M]. Beijing: China Water & Power Press, 2009. (in Chinese))

[9] 钮泽宸，张佩琳，傅联森. 浙江省瞬时单位线法[J]. 浙江水利科技, 1990(1): 1-12. (NIU Ze-chen, ZHANG Pei-lin, FU Lian-sen. Instantaneous unit hydrograph in Zhejiang Province[J]. Zhejiang Hydrotechnics, 1990(1):1-12. (in Chinese))

[10] KEMSLEY E K. Discriminant analysis of high-dimensional data: a comparison of principal components analysis and partial least squares data reduction methods[J]. Chemometrics and Intelligent Laboratory Systems, 1996, 33(1): 47-61.

[11] 杨小兵. 聚类分析中若干关键技术的研究[D]. 杭州：浙江大学, 2005. (YANG Xiao-bing. Study on several key techniques in clustering analysis [D]. Hangzhou: Zhejiang University, 2005. (in Chinese))

[12] 张济世，刘立昱，程中山，等. 统计水文学[M]. 郑州：黄河水利出版社, 2006: 101-125. (ZHANG Ji-shi, LIU Li-yu, CHENG Zhong-shan, et al. Statistical hydrology[M]. Zhengzhou: The Yellow River Water Conservancy Press, 2006: 101-125. (in Chinese))

[13] 卢金利. 灵江永安溪始丰溪河段行洪能力分析及整治对策[J]. 东北水利水电, 2001(8): 29-30. (LU Jin-li. Flood capacity analysis and countermeasures of Lingjiang Yong-an Stream[J]. Water Resource & Hydropower of Northeast China, 2001(8): 29-30. (in Chinese))

Analysis of hydrologically similar basins based on clustering analysis

FAN Meng-ge, LIU Jiu-fu

(*Nanjing Hydraulic Research Institute*, *Nanjing* 210029, *China*)

Abstract: In the flood forecasting and warning, there are often difficulties in analyzing and calculating hydrological regime of some medium-small river basins due to the shortage of hydrological information. With data acquisition of underlying surface topography, vegetation, soil and sustainable development of the data mining method, it is possible to analyze the law in the hydrology data using the unsupervised learning technique such as a cluster analysis method. Thus the parameters of its similar basins can be used in the flood forecasting of one parameter-lacking basin. In this paper 118 river basins in Zhejiang Province, which have more than 20 years precipitation data, have been taken as the case studies. Using the basin length, basin width, river length, river slope, basin average slope, basin shape factor and the average maximum surface precipitation per 1 h, 3 h, 6 h and 12 h, the authors have first reduced the dimensionality using principal components analysis, and then have made the cluster analysis of the basins. The basins in Zhejiang Province are divided into 23 similar groups. On the basis of grouping, hydrological stations which have more than 20 years data of the maximum flood peak and volume are selected for comparison in order to verify whether the grouping is reasonable. The analysis results show that there is a great similarity of the maximum flood peak and volume in the similar basin groups. And the results can provide a new theory and thinking for the flood forecasting in Zhejiang Province from the point of view of statistics.

Key words: cluster analysis; principal component analysis; similar basin; parameter transplantation; runoff contrast