基于聚类归因的水文数据筛分方法研究

刘 畅[1],孙雪蓉[2],张 健[2]

(1.黄河水利委员会山东水文水资源局,山东 济南 250100;2.山东省水文中心,山东 济南 250002)

【摘 要】为提升水文规律理论研究的效率,有必要对数量多、维度广的长系列水文数据进行筛分处理,以更好地适配研究重点和实际需求。文章以泺口水文站为例,基于多种水文特征指标,利用K-均值聚类与归因分析的水文数据筛分方法,对输沙率精测数据进行筛分。结果表明:基于聚类归因的水文数据筛分方法可充分考虑各类水文要素及特征指标之间的相关关系,使水文数据筛分成果更加符合自然科学原理、更具实际应用价值。

【关键词】数据筛分;水文数据;K-均值聚类;独立性检验;归因分析

【中图分类号】P333 【文献标志码】A 【文章编号】1009-6159(2025)-10-0060-04

Research on Hydrological Data Screening Method Based on Cluster Attribution

LIU Chang[1], SUN Xuerong[2], ZHANG Jian[2]

(1. Hydrology and Water Resources Bureau of Shandong Province, Yellow River Conservancy Commission, Jinan 250100, China;
2. Hydrology Center of Shandong Province, Jinan, Shandong 250002, China)

Abstract: In order to improve the efficiency of theoretical research on hydrological laws, it is necessary to screen long-series hydrological data with large quantity and wide dimensions, so as to better adapt to research priorities and practical needs. Taking Luokou Hydrological Station as an example, this paper uses the hydrological data screening method combining K-means clustering and attribution analysis to screen the accurate measurement data of sediment transport rate based on a variety of hydrological characteristic indicators. The results show that the hydrological data screening method based on cluster attribution can fully consider the correlation between various hydrological elements and characteristic indicators, making the hydrological data screening results more in line with the principles of natural science and more practically applicable.

Key words: Data Screening; Hydrological data; K-means clustering; Independence test; Attribution analysis

水文数据是开展水文基础规律研究的基石,具有重要的经济效益和社会效益。随着水文监测技术的发展[1]和信息技术的进步[2],水文数据呈现出海量、多源、异构、时变的特点,形成了水文大数据[3]。鉴于水文数据种类繁多、总量庞大,尤其是长系列历史资料蕴含了海量的多维度信息,因此在开展水文基础规律研究时,有必要根据研究重点和实际需求,对水文数据进行筛分处理。

文章利用聚类方法,对水文数据进行不同维度的初步分类,再根据辅助研究数据与水文数据的归因分析结果,选择相关性较强的特征维度,完成水文数据筛分。基于聚类归因完成筛分的水文数据,相较于原始数据更便于分类别、分维度、分层次深入挖掘和归纳水文基础特性,提升水文规律理论研究的效率。

1 筛分流程

基于聚类归因的水文数据筛分流程如下:

1)基于多种水文特征指标,通过聚类分析,对目标水文数据进行不同维度的初步分类;

2)根据实际需求,选取辅助研究数据并对其完成分类;

3)结合目标水文数据多维度分类结果,完成辅助研究数据与各特征指标的归因分析[4];

收稿日期:2025-04-23

作者简介:刘畅(1992—),女,工程师

4)选出与辅助研究数据相关性较强的特征指标,采用其分类结果,完成目标水文数据的筛分。

基于聚类归因的水文数据筛分技术路线如图 1 所示,其中,聚类分析采用 K-均值聚类法,归因分析依据独立性检验原理。

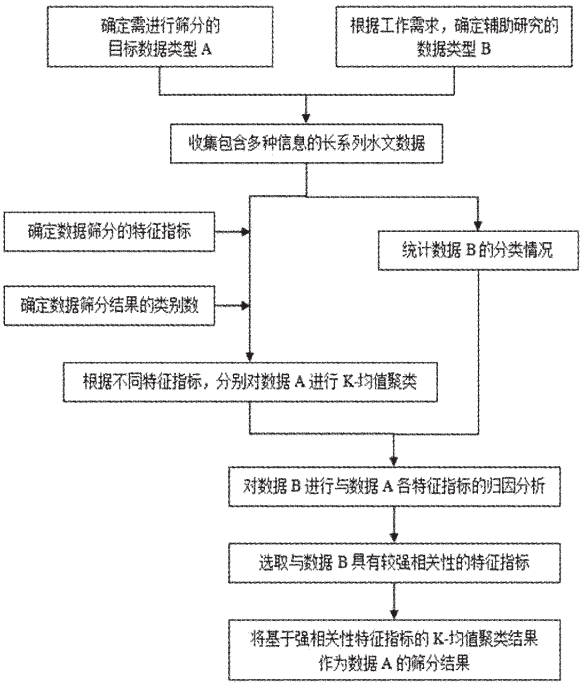

图 1　基于聚类归因的水文数据筛分技术路线图

2　水文数据 K-均值聚类

对于需进行筛分的目标水文数据,选取 n 个水文特征指标,以其中一个水文特征指标为例,采用 K-均值聚类法[5],将目标水文数据分为 k 类。

2.1　初始聚类

随机选择 k 个初始聚类中心,将目标水文数据中的每组数据划分至与之距离最近的聚类中心所在的类。计算距离时,将两组水文数据视为 n 维空间中两点 A_i 与 A_j,其空间距离计算公式如下:

$$d_{ij}=\sqrt{\sum_{P=1}^{n}(x_{ip}-x_{jp})^2} \tag{1}$$

式中:d_{ij} 表示 A_i 与 A_j 两点间的距离,即两组水文数据的距离;n 表示数据点的特征维度数量,即水文数据的特征指标数量;x_{ip} 和 x_{jp} 分别表示 A_i 与 A_j 两点在第 p 个维度的坐标,即两组水文数据第 p 项水文特征指标的数值。

若采用根据单个特征指标进行聚类的方式,则在计算距离时仅需考虑一个维度,公式(1)可简化为:

$$d_{ij}=\sqrt{(x_{i1}-x_{j1})^2}=|x_{i1}-x_{j1}| \tag{2}$$

式中:x_{i1} 和 x_{j1} 分别表示 A_i 与 A_j 两点在某一维度的坐标,即两组水文数据某一项水文特征指标的数值;其余变量的意义同上。

2.2　聚类更新迭代

基于每次聚类的结果,重新计算下一次的聚类中心,公式如下:

$$y_{j,i}=\frac{1}{N_j}\sum_{t=1}^{N_j}x_{jt,i}\quad(i=1,2,\cdots,n;j=1,2,\cdots,k) \tag{3}$$

式中:y_j 表示第 j 个类的聚类中心,$y_{j,i}$ 表示其第 i 维的值,即第 i 项水文特征指标的数值;N_j 表示第 j 个类中的水文数据组数;x_{jt} 表示第 j 个类中的第 t 组水文数据,$x_{jt,i}$ 表示其第 i 维的值,即第 i 项水文特征指标的数值;k 表示类的数目;n 的意义同上。

若采用根据单个特征指标进行聚类的方式,则聚类中心仅需进行一个维度的计算,公式(3)可简化为:

$$y_j=\frac{1}{N_j}\sum_{t=1}^{N_j}x_{jt}\quad(j=1,2,\cdots,k) \tag{4}$$

式中:y_j 表示第 j 个类的聚类中心的水文特征指标数值;x_{jt} 表示第 j 个类中的第 t 组水文数据的水文特征指标数值;其余变量的意义同上。

在每次计算出新的聚类中心后,重新将目标水文数据中的每组数据划分至与之距离最近的新聚类中心所在的类。以此类推,不断进行聚类更新迭代,直至聚类结果收敛,即新一轮的聚类结果与上一轮完全一致。将此时停止更新迭代的聚类结果作为 K-均值聚类结果,目标水文数据被分为 k 类。

3　基于独立性检验的归因分析

采用 K-均值聚类法完成对目标水文数据 A 的初步分类(根据特征指标的不同、类的数目不同,得到多套不同的分类结果),再结合辅助研究数据 B 的分类情况,通过独立性检验[6]的方法完成数据 B 与数据 A 的归因分析。

设目标水文数据 A 的 n 个水文特征指标为 a_1、a_2、a_3……a_n,根据每个特征指标,分别将数据 A 分为 k_{min} 至 k_{max} 类;设辅助研究数据 B 的指定特

征指标为b,据此将数据B划分为r类。根据目标水文数据A的各种聚类分析结果,依次将其与数据B的r类分类结果进行独立性检验列联表的制作,并计算统计量χ^2,通过查找《χ^2分布表》,定量判断此时的两个特征指标存在相关性或独立性的显著性水平。

例如,根据第m个水文特征指标a_m将目标水文数据A分为k类时,独立性检验的原理和步骤如下:

1)基于此时目标水文数据A的聚类分析结果,将a_m与b视为分类变量,将a_m与b的取值范围分别划分为k个和r个互不相交的区间。

2)统计在k×r个区间中出现的数据数量,制作独立性检验列联表,见表1。

表1 独立性检验列联表

	a_m						
b	1	2	…	j	…	k	Σ
1	N_{11}	N_{12}	…	N_{1j}	…	N_{1k}	$N_{1\cdot}$
2	N_{21}	N_{22}	…	N_{2j}	…	N_{2k}	$N_{2\cdot}$
…	…	…	…	…	…	…	…
i	N_{i1}	N_{i2}	…	N_{ij}	…	N_{ik}	$N_{i\cdot}$
…	…	…	…	…	…	…	…
r	N_{r1}	N_{r2}	…	N_{rj}	…	N_{rk}	$N_{r\cdot}$
Σ	$N_{\cdot 1}$	$N_{\cdot 2}$	…	$N_{\cdot j}$	…	$N_{\cdot k}$	N

表1中,N_{ij}为b∈第i区间、a_m∈第j区间的水文数据数量,$N_{i\cdot}$为b∈第i区间的水文数据数量,$N_{\cdot j}$为a_m∈第j区间的水文数据数量,N为目标水文数据A的数据总量。关系如下:

$$N_{i\cdot}=\sum_{j=1}^{k}N_{ij}\quad(i=1,2,\cdots,r) \tag{5}$$

$$N_{\cdot j}=\sum_{i=1}^{r}N_{ij}\quad(j=1,2,\cdots,k) \tag{6}$$

$$N=\sum_{i=1}^{r}\sum_{j=1}^{k}N_{ij} \tag{7}$$

式中各变量的意义同上。

3)计算当前的统计量χ^2,判断水文特征指标a_m与b存在相关性的显著性水平。

假设a_m与b相互独立,则存在如下概率关系:

$$P_{ij}=P_{i\cdot}P_{\cdot j} \tag{8}$$

式中:P_{ij}表示b∈第i区间且a_m∈第j区间的概率;$P_{i\cdot}$表示b∈第i区间的概率;$P_{\cdot j}$表示a_m∈第j区间的概率。

按下式计算统计量χ^2:

$$\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(N_{ij}-Np_{i\cdot}p_{\cdot j})^2}{Np_{i\cdot}p_{\cdot j}} \tag{9}$$

式中各变量的意义同上。

当$n\to\infty$时,统计量χ^2服从自由度为rk-1的χ^2分布。将$p_{i\cdot}$和$p_{\cdot j}$以极大似然估计值$N_{i\cdot}/N$和$N_{\cdot j}/N$代替,代入式(9),可得统计量χ^2的简化计算式:

$$\chi^2=N\left(\sum_{i=1}^{r}\sum_{j=1}^{k}\frac{N_{ij}^2}{N_{i\cdot}N_{\cdot j}}-1\right) \tag{10}$$

式中各变量的意义同上。

此时,统计量χ^2服从自由度为(r-1)×(k-1)的χ^2分布。根据自由度(r-1)×(k-1)和给定的显著性水平α,查找《χ^2分布表》,得到临界值χ_α^2。若式(10)计算χ^2的值满足$\chi^2\geqslant\chi_\alpha^2$,则拒绝$a_m$与b相互独立的假设,即水文特征指标$a_m$与b具有相关性;若$\chi^2<\chi_\alpha^2$,则接受假设,即水文特征指标$a_m$与b是相互独立的。

按照上述步骤,根据目标水文数据A在逐个特征指标(a_1、a_2、a_3……a_n)、逐个分类数目(k_{min}至k_{max})下的分类情况,依次进行与指定特征指标b的独立性检验。最终选出显著性水平最低的一种情况,此时该水文特征指标与指定特征指标b的相关性最强,即可得出辅助研究数据B与目标水文数据A之间的归因分析结论。

4 算例分析

以黄河山东测区泺口水文站的输沙率精测资料为例,采用基于K-均值聚类与归因分析的水文数据筛分方法,通过分析输沙率测次的关键水文特征指标与断面泥沙分布特征之间的相关关系,对该站的目标输沙率测次进行筛分。

4.1 问题提出

泺口水文站位于山东省济南市,为泥沙测验一类站,其泥沙测验资料在黄河下游治理开发中发挥着重要作用,对该站开展泥沙测线精简优化分析具有重要意义。收集泺口水文站2003—2022年间不同含沙量级的23个测次的输沙率选点法精测资料,经初步分析发现,不同输沙率测次的断面泥沙分布特征不尽相同,若将上述长系列资料中的全部测次作为一个整体进行泥沙测线优

化分析研究,效果并不理想,未达预期精度。

因此,需对泺口水文站输沙率精测的长系列测次数据进行筛分。将输沙率精测数据作为目标水文数据,将断面泥沙分布特征作为辅助研究数据,通过聚类与归因,选取出输沙率测次资料中与断面泥沙分布特征具有较强相关性的水文特征指标,以此为依据进行输沙率测次的筛分。

4.2　基于不同特征指标的聚类

选取各输沙率测次的断面平均流速、断面面积、水面宽、平均水深、断面平均流量、断面平均含沙量共6个指标作为水文特征指标,分别根据每个指标的数值对输沙率测次进行K-均值聚类。将类的数目设为2类和3类,分别统计23个输沙率测次的聚类结果,见表2、表3。

表2　泺口水文站各输沙率测次K-均值聚类结果统计表(2类)

水文特征指标	类型	水文特征指标取值范围	属于此类型的输沙率测次编号
断面平均流速/(m·s^{-1})	小	1.22~1.83	4、7、9、15、16、19、20、23
	大	1.85~2.23	1、2、3、5、6、8、10、11、12、13、14、17、18、21、22
断面面积/m^2	小	540~1 330	1、2、7、9、12、14、15、16、17、18、19、20、22、23
	大	1 440~1 980	3、4、5、6、8、10、11、13、21
水面宽/m	小	188~239	1、2、3、4、5、6、7、8、9、10、11、15
	大	243~287	12、13、14、16、17、18、19、20、21、22、23
平均水深/m	小	2.26~5.5	1、7、9、12、14、15、16、17、18、19、20、22、23
	大	5.6~8.4	2、3、4、5、6、8、10、11、13、21
断面平均流量/(m^3·s^{-1})	小	825~2 220	7、9、15、16、19、20、22、23
	大	2 400~4 030	1、2、3、4、5、6、8、10、11、12、13、14、17、18、21
断面平均含沙量/(kg·m^{-3})	小	2.06~17.4	3、4、5、6、8、11、12、13、14、15、18、19、20、21
	大	26.9~68.6	1、2、7、9、10、16、17、22、23

注:测次编号根据施测时间顺序由前到后排列。

4.3　断面泥沙分布特征归因分析

将各输沙率测次的断面泥沙分布特征概括为4种类型,统计结果见表4。

分别进行6个水文特征指标与断面泥沙分布特征的独立性检验,当水文特征指标分为2类时,独立性检验的自由度为(2-1)×(4-1)=3;当水文特征指标分为3类时,独立性检验的自由度为(3-1)×(4-1)=6。不同水文特征指标、不同分类区间个数时的独立性检验结果见表5。

表3　泺口水文站各输沙率测次K-均值聚类结果统计表(3类)

水文特征指标	类型	水文特征指标取值范围	属于此类型的输沙率测次编号
断面平均流速/(m·s^{-1})	小	1.22~1.53	7、9、15
	中	1.69~1.94	2、3、4、5、6、13、14、16、17、19、20、23
	大	1.98~2.23	1、8、10、11、12、18、21、22
断面面积/m^2	小	540~1 000	7、9、15、19、20、22
	中	1 080~1 440	1、2、10、12、14、16、17、18、23
	大	1 490~1 980	3、4、5、6、8、11、13、21
水面宽/m	小	188~221	6、7、8、9、10
	中	231~249	1、2、3、4、5、11、12、13、14、15、16、17
	大	263~287	18、19、20、21、22、23
平均水深/m	小	2.26~4.06	15、19、20、22、23
	中	4.45~5.6	1、2、7、9、12、14、16、17、18
	大	6.3~8.4	3、4、5、6、8、10、11、13、21
断面平均流量/(m^3·s^{-1})	小	825~1 830	7、9、15、19、20、22、23
	中	2 220~2 680	1、2、4、12、14、16、17
	大	2 800~4 030	3、5、6、8、10、11、13、18、21
断面平均含沙量/(kg·m^{-3})	小	2.06~8.83	5、6、8、11、12、13、14、15、18、21
	中	13.1~26.9	3、4、16、19、20
	大	28.8~68.6	1、2、7、9、10、17、22、23

表4　泺口水文站各输沙率测次断面泥沙分布特征统计表

特征类型	率测次编号
Ⅰ型:右岸深槽不明显,高含沙区集中分布于断面左中侧底部	1、2、3、4、8、18、19
Ⅱ型:右岸深槽不明显,高含沙区集中分布于断面右中侧底部	5、6、7、22
Ⅲ型:右岸深槽明显,高含沙区集中分布于深槽内	9、10、11、12、13、14、17
Ⅳ型:右岸深槽明显,高含沙区集中分布于全断面底部	15、16、20、21、23

由表5可得出断面泥沙分布特征的归因分析结论:断面平均流量分为2类时,与断面泥沙分布特征存在相关性的显著性水平$\alpha<0.1$,即二者存在相关性的概率超过90%,为所有水文特征指标中最高的。

4.4　筛分结果

根据聚类归因的结论,对泺口水文站输沙率精测数据按照断面平均流量指标进行筛分。将全部输沙率测次分为2类,分别为"小流量类型"的测次和"大流量类型"的测次,各类(下转第73页)