DOI:10.3876/j.issn.1000-1980.2016.04.012

降雨空间分布的模糊熵聚类分析

张继国[1],管耀宗[2],朱永忠[2]

(1. 河海大学水利信息统计与管理研究所,江苏 常州 213022; 2. 河海大学理学院,江苏 南京 210098)

摘要:为了提高降雨量插值精度,充分挖掘降雨变量信息,利用模糊熵聚类分析算法,对流域内雨量站进行模糊熵聚类研究,通过基于可能性分布和距离判定的聚类有效性函数确定模糊熵系数和聚类数,从而得到模糊聚类结果,改进原有的插值方法。以淮河流域蚌埠站以上区域99个雨量站雨量数据,分别在一般情况下和模糊熵聚类情况下做交叉验证,结果显示,模糊熵聚类分析在反距离平方插值法中对降雨精度有所提升。

关键词:降雨空间分布;降雨数据精度;模糊熵聚类分析;聚类有效性分析;降雨量插值

中图分类号:P467 **文献标志码**:A **文章编号**:1000-1980(2016)04-0353-05

Fuzzy entropy clustering analysis of spatial distribution of precipitation

ZHANG Jiguo[1], **GUAN Yaozong**[2], **ZHU Yongzhong**[2]

(1. *Institute of Information Statistics and Management of Water Resources, Hohai University, Changzhou* 213022, *China*;
2. *College of Sciences, Hohai University, Nanjing* 210098, *China*)

Abstract: In order to improve the accuracy of precipitation interpolation and fully explore the information regarding precipitation variables, fuzzy entropy clustering (FEC) was carried out at rain gauge stations in a basin. A clustering validity function, based on possibility distribution and distance determination, was used to determine the fuzzy entropy coefficient and the number of clusters, so as to obtain the fuzzy clustering results and improve the original interpolation method. Based on data from 99 rain gauge stations located above the Bengbu Station in the Huaihe River Basin, cross validation was conducted under non-clustering and FEC conditions. The results demonstrate that FEC improves the precipitation accuracy in the inverse distance squared interpolation method.

Key words: spatial distribution of precipitation; accuracy of precipitation data; fuzzy entropy clustering (FEC) analysis; clustering validity analysis; precipitation interpolation

降雨量是水文模型中径流模拟最基本、最主要的一个输入项,是研究其他水文问题的基础。其空间分布特征是影响产汇流模拟及其他一系列水文问题的重要控制因素[1]。随着研究的深入,水文模型对降雨数据精度和广度的要求越来越高。理论上获取高精度降雨数据的方法是建立高密度的雨量站网,但是由于经济条件和技术手段的约束,大部分地区气象观测站点数量不足,分布密度有限。因此,利用现有气象观测站的数据,通过空间插值对观测数据进行补充尤为重要,孔云峰等[2]通过多种插值方法探究了美国德州的空间雨量数据。然而,大尺度流域上的降雨空间具有很强的时空分布不均匀性和复杂性,对此,在区域内对已有站点作聚类分区处理,即将复杂的降雨测量站点系统划分成不同的子系统,减少不确定性因素的影响,是一种切实有效的研究方法[3]。李生辰等[4]在2007年研究了青藏高原降雨分区问题。杨绚等[5]在2008年通过降雨变化特征对中国干旱地区进行了聚类划分;郑永宏等[6]2012年研究了湖北省的降雨分区问题。

聚类分区主要分硬聚类和模糊聚类。相对于硬聚类,模糊聚类方法能够对类与类之间有交叉的数据样

收稿日期:2015-10-17

基金项目:江苏省自然科学基金(BK20131135);江苏省自然科学青年基金(BK20130242)

作者简介:张继国(1956—),男,湖北汉川人,教授,主要从事水文不确定性分析、信息熵理论与方法研究。E-mail:zhangjg@hhuc.edu.cn

本集进行有效的聚类,所得的聚类结果明显优于硬聚类方法。由于模糊聚类建立了数据样本对于类别的不确定性的描述,表达了样本类属的模糊性,因此能够更客观地反映实际情况[7],并被广泛应用于水文研究中[8-9]。本文根据模糊熵聚类算法,将淮河流域蚌埠站以上的99个雨量站进行模糊划分,并研究模糊聚类分析在降雨量插值精度中的应用,为流域内雨量建模分析、水文循环研究、灾害预报等提供理论依据。

1 模糊熵聚类

1.1 模糊熵目标函数

Tran等提出的模糊熵聚类算法(fuzzy entropy clustering)[10]是在模糊 C 均值聚类算法(fuzzy C-means clustering)基础上,引入熵的概念,对隶属度值分布进行了算法优化。

对有 T 个成员的 D 维空间中的数据集 $X=\{x_1,x_2,\cdots,x_T\}\in\mathbb{R}^D$ 进行聚类分析,得到隶属度矩阵 $\boldsymbol{U}=[u_{it}]$ 和聚类 C_i 中心 $\theta_i=\{\theta_{i1},\theta_{i2},\cdots,\theta_{iD}\}$,$i=1,2,\cdots,C$,其中 u_{it} 表示成员 $X_t=\{x_{t1},x_{t1},\cdots,x_{tD}\}$ 对于聚类 C_i 的隶属度,满足:

$$0<u_{it}<1 \qquad \sum_{i=1}^{C}u_{it}=1 \qquad 0<\sum_{t=1}^{T}u_{it}<1 \tag{1}$$

在聚类过程中,数据隶属度表征了聚类的模糊程度。数据对聚类的隶属度差异越大即聚类的信息熵越大,聚类的效果越好,由此,在目标函数中引入熵函数 $E(\boldsymbol{U})=-\sum_{i=1}^{C}\sum_{t=1}^{T}u_{it}\lg u_{it}$,得到新的聚类目标函数:

$$H_n(\boldsymbol{U},\theta,X)=\sum_{i=1}^{C}\sum_{t=1}^{T}u_{it}\,d^2(x_t,\theta_i)+n\sum_{i=1}^{C}\sum_{t=1}^{T}u_{it}\lg u_{it} \tag{2}$$

式中:C——聚类数;T——样本数;u_{it}——成员 x_t 对聚类 C_i 的隶属度;n——模糊熵系数;θ_i——聚类中心;$d(x_t,\theta_i)$——样本 x_t 与 θ_i 的差异距离。

利用拉格朗日算子对目标函数求极值可以得到模糊熵聚类算法隶属度矩阵和聚类中心的更新方程(推导过程见文献[10]):

$$\begin{cases} u_{it}=\left\{\sum_{j=1}^{C}\left[\mathrm{e}^{d^2(x_t,\theta_i)}\,\middle|\,\mathrm{e}^{d^2(x_t,\theta_j)}\right]^{\frac{1}{n}}\right\}^{-1} & (3)\\ \theta_i=\sum_{t=1}^{T}u_{it}\,x_t\,\Big|\,\sum_{t=1}^{T}u_{it} & (4)\end{cases}$$

1.2 聚类数的确定

Bezdek[11]基于香农信息熵公式定义了模糊划分的划分熵:$H(\boldsymbol{U};C)=-\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{C}u_{it}\ln u_{it}$,用于表示每个样本点对 C 个聚类中心可能性分布(即隶属度函数)对应的香农信息熵平均值。

相应地,每一个聚类中心对 T 个样本构成的可能性分布也有一个香农信息熵,由此定义可能性划分熵为

$$H^*(\boldsymbol{U};C)=-\frac{1}{C}\sum_{i=1}^{C}\sum_{t=1}^{T}\frac{u_{it}}{\sum_{t=1}^{T}u_{it}}\ln\frac{u_{it}}{\sum_{t=1}^{T}u_{it}} \tag{5}$$

范九伦[12]根据Bezdek划分熵和可能性划分熵定义了基于可能性分布的聚类有效性函数(式(6)),并提出当 $H_{\mathrm{p}}(\boldsymbol{U};C)$ 取得最大值的时候,$\boldsymbol{U}$ 为最佳聚类隶属度矩阵、C 为最佳聚类数。

$$H_{\mathrm{p}}(\boldsymbol{U};C)=H(\boldsymbol{U};C)-H^*(\boldsymbol{U};C)+\ln n-\ln C \tag{6}$$

因此,在模糊熵聚类算法过程中,可以拟定多个聚类数和模糊熵系数,当$(\boldsymbol{U}^*,C^*)$满足 $H_{\mathrm{p}}(\boldsymbol{U}^*;C^*)=\max_{c}(\max_{\Omega c}(\boldsymbol{U},C))$时,$C^*$ 则是最优聚类数。

1.3 模糊熵系数的确定

聚类分析是按照一定要求对事物进行分类的过程,其核心是差异越大的样本区分程度越大,即对于样本 $X_i=\{x_i(1),x_i(2),\cdots,x_i(D)\}$,$X_j=\{x_j(1),x_j(2),\cdots,x_j(D)\}\in\mathbb{R}^D$,样本距离 $d(X_i,X_j)$ 越大,则希望其隶属度向量 $\boldsymbol{u}_i$、$\boldsymbol{u}_j$ 的差异越大。其中 $\boldsymbol{u}_i=(u_{1i},u_{2i},\cdots,u_{Ci})$,$(i=1,2,\cdots,T)$ 表示样本 X_i 对各聚类的隶属度。

在已知数据样本中选取 X_i、X_j、X_k、X_l(i、j、k、$l=1,2,\cdots,T$)4个样本作为样本空间 Ω,其隶属度向量为 $\boldsymbol{u}_i$、

$\boldsymbol{u}_j$、$\boldsymbol{u}_k$、$\boldsymbol{u}_l$,定义基于距离判定的模糊聚类有效函数为

$$E(\boldsymbol{U}) = P\left\{ [d(X_i,X_j) - d(X_k,X_l)][d(\boldsymbol{u}_i,\boldsymbol{u}_j) - d(\boldsymbol{u}_k,\boldsymbol{u}_l)] > 0 \right\} \tag{7}$$

式中:$P\{*\}$——事件 $*$ 发生占总样本空间 Ω 的比例。

在聚类数为 C^* 的情况下,根据模糊熵聚类算法,模糊熵系数 n 决定了隶属度矩阵。因此,$E(\boldsymbol{U}) = E'(n)$。当 n^* 满足 $E'(n^*) = \max\limits_{n}(E'(n))$ 时,n^* 为最优模糊熵系数。

1.4　研究步骤

根据实际数据规模与研究要求,拟定多个聚类数 C、模糊熵系数 n 和停止阈值 ε。随机选择 C 个样本数据作为初始聚类中心 $\theta^0 = (\theta_1,\cdots,\theta_C)$。根据式(3)、式(4)更新样本隶属度矩阵 $\boldsymbol{U}_{C\times D}^m$ 和聚类中心 θ^m(m 为迭代次数)。当满足迭代停止条件($|\theta^m - \theta^{m-1}| \leqslant \varepsilon$)时,停止迭代,得到隶属度矩阵 $\boldsymbol{U}_{C\times D}^m$。对多个聚类数和模糊系数组合进行模糊熵聚类算法得到相应的雨量站聚类隶属度矩阵,通过式(6)、式(7)计算相应的聚类有效性值 H_{p} 和距离判定有效性值 E,并以此确定最优聚类数和最优模糊熵系数,从而得到最优模糊聚类结果。

1.5　聚类合理性验证

对于 T 个已知站点,每个站点有 D 个月降雨量数据,其降雨量矩阵为 $\boldsymbol{Z} = (Z_{ij})_{T\times D}$。分别假设某个站点为未知站点,通过反距离平方插值法得到反距离平方插值下的站点计算降雨量矩阵 $\boldsymbol{Z}' = (Z'_{ij})_{T\times D}$,其中 $Z'_{ij} = \sum\limits_{k=1}^{T-1}\frac{z_{kj}}{d_k^2}\Big/\sum\limits_{k=1}^{T-1}\frac{1}{d_k^2}$。

同样分别假设一个站点 p 为未知站点,对其他站点进行模糊熵聚类算法分析,得到剩余站点的隶属度矩阵 $\boldsymbol{U}_{\mathrm{p}} = (u_1,u_2,\cdots,u_{T-1}) = (u_{ij})_{C\times(T-1)}$。通过反距离平方插值得到假设的未知站点对各聚类的隶属度向量 $\boldsymbol{u}'_{\mathrm{p}} = (u_{1\mathrm{p}},u_{2\mathrm{p}},\cdots,u_{C\mathrm{p}})$,$u_{i\mathrm{p}} = \sum\limits_{k=1}^{T-1}\frac{u_{kj}}{d_k^2}\Big/\sum\limits_{k=1}^{T-1}\frac{1}{d_k^2}$。做归一化处理:$\boldsymbol{u}_{\mathrm{p}}^* = (u_{1\mathrm{p}}^*,u_{2\mathrm{p}}^*,\cdots,u_{C\mathrm{p}}^*)^{\mathrm{T}}$,$u_{i\mathrm{p}}^* = u_{i\mathrm{p}}^*\Big/\sum\limits_{j=1}^{C}u_{j\mathrm{p}}$,$i = 1,2,\cdots,C$。利用隶属度和反距离平方插值法,计算模糊熵聚类后的计算降雨量矩阵 $\boldsymbol{Z}^* = (z_{ij}^*)_{T\times D}$,$z_{ij}^* = \sum\limits_{k=1}^{T-1}\frac{\boldsymbol{u}'_i\boldsymbol{u}_{\mathrm{p}}^* z_{kj}}{d_k^2}\Big/\sum\limits_{k=1}^{T-1}\frac{\boldsymbol{u}'_i\boldsymbol{u}_{\mathrm{p}}^*}{d_k^2}$。

通过真实降雨量矩阵 $\boldsymbol{Z}$、一般情况下的插值雨量矩阵 $\boldsymbol{Z}'$ 和模糊熵聚类情况下的插值雨量矩阵 $\boldsymbol{Z}^*$ 计算多个交叉统计量[13]。记 X_{o},X_{e} 分别是已知降雨量数据和交叉检验的插值计算降雨量数据,N 为数据个数,各统计量(相关系数 R、平均相对误差 R_{MAE}、均方根误差 R_{MSE}、复合相对误差 C_{RE})计算公式如下:

$$\begin{cases} R = \dfrac{\sum\limits_{t=1}^{N}\left[X_{\mathrm{o}}(t) - \overline{X}_{\mathrm{o}}\right]\left[X_{\mathrm{e}}(t) - \overline{X}_{\mathrm{e}}\right]}{\sqrt{\left[X_{\mathrm{o}}(t) - \overline{X}_{\mathrm{o}}\right]^2}\sqrt{\left[X_{\mathrm{e}}(t) - \overline{X}_{\mathrm{e}}\right]^2}} & R_{\mathrm{MAE}} = \dfrac{\sum\limits_{t=1}^{N}\left|X_{\mathrm{e}}(t) - X_{\mathrm{o}}(t)\right|}{\sum\limits_{t=1}^{N}X_{\mathrm{o}}(t)} \\ R_{\mathrm{MSE}} = \sqrt{\dfrac{1}{N}\sum\limits_{t=1}^{N}\left[X_{\mathrm{e}}(t) - X_{\mathrm{o}}(t)\right]^2} & C_{\mathrm{RE}} = \dfrac{\sum\limits_{t=1}^{N}\left[X_{\mathrm{e}}(t) - X_{\mathrm{o}}(t)\right]^2}{\sum\limits_{t=1}^{N}\left[X_{\mathrm{o}}(t) - \overline{X}_{\mathrm{o}}\right]^2}\times 100\% \end{cases} \tag{8}$$

相关系数除去了偏差和方差的影响,考虑了插值估计数据与实际数据变化的同步性,表示了插值估计序列替代实际观测序列的潜在能力。平均相对误差和均方根误差反映了插值估计序列与实际序列比较得到的误差平均情况。复合相对误差是描述插值序列与实际序列的相似性指标,该统计量对大误差数据十分敏感。

2　研 究 实 例

2.1　数据来源及相似度计算

研究数据来自淮河流域蚌埠站以上区域 99 个雨量站 1953—2013 年 732 个月的降雨量数据,站点基本情况见文献[14]。

淮河流域位于东经 112°~118°、北纬 31°~35°的区域内,介于长江和黄河两大流域之间。在气候上,它处于南北气候过渡带,降水时空分布严重不均,差异较大。淮河又是我国南北方的一条自然分界线。因此,

研究淮河流域的降水时空不确定性具有较高的科学价值。

对于降雨量数据 $X_i=\{x_{i1},x_{i2},\cdots,x_i\ 732\}$,$X_j=\{x_{j1},x_{j2},\cdots,x_{j732}\}$,其欧氏距离为

$$d_{ij}=\sqrt{\sum_{k=1}^{732}(x_{ik}-x_{jk})^2} \tag{9}$$

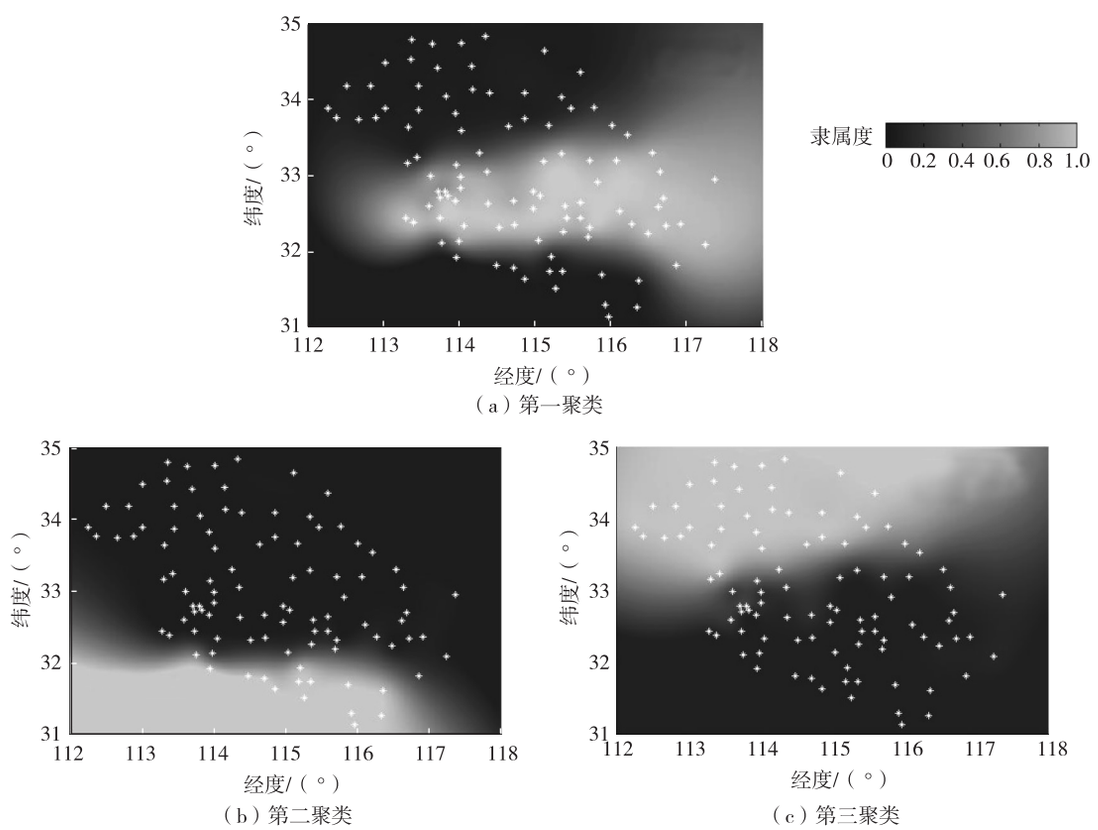
计算所有已知站点间的距离,获取最大值 $d_{\max}=\max\limits_{l\leqslant i,j\leqslant D}(d_{ij})$,并定义标准欧氏差异距离 d_{ij}^* 作为站点 X_i、X_j 的差异距离。

$$d_{ij}^*=\frac{d_{ij}}{d_{\max}} \tag{10}$$

2.2 聚类结果

研究发现,当聚类数过大时,出现聚类中心彼此靠近并有重合现象;模糊熵系数小于0.02时聚类划分过于分明,近似硬聚类,当系数大于0.03时聚类间过于模糊。因此,拟定聚类数 $C=2$、3、4、5,模糊熵系数 $n=0.02$、0.022、0.024、0.026、0.028、0.03和停止阈值 $\varepsilon=0.0001$。计算相应的聚类有效性值 H_P,当 $C=2$、3、4、5时,$H_P=0.0004$、0.1011、0.0584、0.0326。比较可得,当 $C=3$ 时,H_P 取得最大值,因此选取最优聚类数 $C^*=3$。

在 $C^*=3$ 的情况下分别计算各模糊熵系数对应隶属度矩阵的距离判定有效性值 E,当 $n=0.02$、0.022、0.024、0.026、0.028、0.03时,$E=0.7618$、0.7632、0.7640、0.7642、0.7625、0.7641。比较可得,在 $C^*=3$ 的情况下,当 $n^*=0.026$ 时 E 取得最大值。从而得到相应的最优隶属度矩阵 $\boldsymbol{U}^*$。聚类结果如图1所示(图中所示星点表示已知雨量站位置)。



(a)第一聚类 (b)第二聚类 (c)第三聚类

图1 聚类隶属度分布

Fig. 1 Distribution of degree of membership in EFC

图1中3幅图分别表示3个聚类的隶属度分布,其中颜色越浅表示该区域对此聚类的隶属度越高,反之越低。中部雨量站集中在第一聚类中,南部雨量站集中在第二聚类中,北部雨量站集中在第三聚类中。可以看出,算法得到的聚类结果有明显的地理位置聚拢性,十分符合实际情况,即地理位置较近的地方,降雨量差异性较小。此外,分区多以纬度划分为主,说明降雨量在纬度变化中的差异变化大,符合我国淮河流域基本

自然情况,即南北降雨量差异大、东西降雨量差异小。

2.3　验证结果

模糊熵聚类情况下和一般情况下反距离平方加权插值法各交叉统计量如表 1 所示。

表 1　交叉验证统计量对比

Table 1　Comparison of statistics in cross-validation

统计方法	R	R_{MAE}	R_{MSE}	C_{RE}/%
模糊熵聚类	0.803 8	0.399 4	54.245 2	37.57
一般情况	0.781 4	0.416 2	55.283 1	39.02

可以看出,与一般情况相比,模糊熵聚类分析情况下,站点雨量插值与测量值相关系数有所提高,各类误差都有所下降,表明模糊熵聚类方法具有一定的优越性。

3　结　　语

通过模糊熵聚类分析,可以深入挖掘降雨信息在流域内的分布,有利于更加深入地探究降雨系统内部的关系,以及多种不确定因素。本文对淮河流域蚌埠站以上区域降雨量数据进行了模糊熵聚类分析,获得模糊熵聚类结果。同时,通过交叉验证法,说明了模糊熵聚类算法在反距离平方加权插值中的实用性。

参考文献:

[1] 石朋,芮孝芳. 降雨空间插值方法的比较与改进[J]. 河海大学学报(自然科学版),2005,33(4):361-365. (SHI Peng, RUI Xiaofang. Comparison and improvement of spatial rainfall interpolation methods. [J]. Journal of Hehai University (Natural Sciences), 2005,33(4):361-365. (in Chinese))

[2] 孔云峰,仝文伟. 降雨量地面观测数据空间探索与插值方法探讨[J]. 地理研究,2008,27(5):1097-1108.(KONG Yunfeng, TONG Wenwei. Spatial exploration and interpolation of the surface precipitation data[J]. Geographical Research, 2008, 27(5): 1097-1108. (in Chinese))

[3] 张继国, 谢平, 龚艳冰, 等.降雨信息空间插值研究评述与展望[J].水资源与水工程学报, 2012, 23(1):6-9. (ZHANG Jiguo, XIE Ping, GONG Yanbing, et al. Review and perspectives of the research on spatial interpolation of rainfall information [J]. Journal of Water Resources and Water Engineering, 2012,23(1):6-9.(in Chinese))

[4] 李生辰, 徐亮, 郭英香, 等.近 34a 青藏高原年降水变化及其分区[J].中国沙漠, 2007, 27(2):307-314. (LI Shengchen, XU Liang, GUO Yingxiang, et al. Change of annual precipitation over QinghaiXizang Plateau and subregions in recent 34 years [J]. Journal of Desert Research, 2007, 27(2): 307-314. (in Chinese))

[5] 杨绚, 李栋梁.中国干旱气候分区及其降水量变化特征[J].干旱气象, 2008, 26(2):17-24. (YANG Xuan, LI Dongliang. Precipitation variation characteristics and arid climate division in China [J]. Arid Meteorology, 2008, 26(2):17-24.(in Chinese))

[6] 郑永宏, 林爱文, 代侦勇.湖北省降水分区研究[J].长江流域资源与环境, 2012, 21(7):859-863. (ZHENG Yonghong, LIN Aiwen, DAI Zhenyong. Research on precipitation regionalization in Hubei Province [J]. Resources and Environment in the Yangtze Basin, 2012,21(7):859-863.(in Chinese))

[7] 雷鸣. 模糊聚类新算法的研究[D].天津:天津大学,2007.

[8] 冀鸿兰,卞雪军,徐晶. 黄河内蒙古段流凌预报可变模糊聚类循环迭代模型[J]. 水利水电科技进展,2013,33(4):14-17. (JI Honglan, BIAN Xuejun, XU Jing. Variable fuzzy clustering loop iteration model for ice-run forecast in Inner Mongolia reach of Yellow River [J]. Advances in Science and Technology of Water Resources, 2013, 33(4):14-17.(in Chinese))

[9] 樊哲超,陈建生,董海洲,等. 应用环境同位素和模糊聚类方法研究堤防渗漏[J]. 水利水电科技进展,2005,25(2):8-10,57. (FAN Zhechao, CHEN Jiansheng, DONG Haizhou, et al. Application of environmental isotope and fuzzy clustering method to study of seepage from dykes [J]. Advances In Science and Technology of Water Resources, 2005, 25(2):8-10,57.(in Chinese))

[10] WU Xiaohong, ZHOU Jianjiang. Possibilistic fuzzy entropy clustering[J]. Journal of Computational Information Systems, 2007, 3(1):25-33.

[11] BEZDEK J C. Pattem recognition with fuzzy objective function algorithms [M]. New York: Plenum, 1981.

[12] 范九伦. 模糊聚类新算法与聚类有效性问题研究[D].西安:西安电子科技大学,1998.

[13] 熊秋芬,黄玫,熊敏诠,等. 基于国家气象观测站逐日降水格点数据的交叉检验误差分析[J]. 高原气象,2011,30(6):1615-1625.(XIONG Qiufen, HUANG Mei, XIONG Minquan, et al. Cross-validation error analysis of daily gridded precipitation based on China meteorological observation [J]. Plateau Meteorology,2011,30(6):1615-1625. (in Chinese))

[14] 张继国.降雨时空分布不均匀性信息熵研究[D].南京:河海大学,2004.