XIX International Conference on Water Resources
CMWR 2012
University of Illinois at Urbana-Champaign
June 17-22, 2012

FLOOD FEATURE IDENTIFICATION AND CLUSTERING IN WUJIANG RIVER, SOUTH CHINA

CHEN Xiao-hong[1,2]，WANG Li-na[3]

1. SunYat-sen University
Center for Water Resources and Environment
Guangzhou 510275, China
e-mail: eescxh@mail.sysu.edu.cn

2. Key Laboratory of Water cycle and water security in Southern China of Guangdong Higher Education Institutes
Guangzhou 510275, China

3. South China Normal University
School of Geography
Guangzhou 510631, China
email: linawang2004@163.com）

Keywords: Fuzzy C-Means model, Simulated annealing algorithm, Flood clustering

Summary. *Wujiang River, one of the main branches of the Beijiang River in South China, is frequently suffered from flood disasters. Flood clustering becomes one of the critical sub-issues for realizing the different types of flood features. This paper attempts to put forward a flood clustering approach for flood feature identification which is important to the flood risk management and flood forecasting. In this paper we proposed the simulated annealing algorithm to solve the shortcomings of the Fuzzy C-Means which is generally used for flood clustering. The specific design of the clustering model was given. FCM and FCM-SA algorithms are used for clustering the flood features covering a period of 42 years from 1965 to 2006 in Wujiang River. The result showed that both FCM and FCM-SA can be successfully applied to cluster flood features, however, the FCM-SA has more advantages than the FCM. The flood clustering results can help us to identify very quickly the flood characteristics in the previously year. The clustering result showed that the flood situation of Wujiang River is much more serious.*

1 INTRODUCTION

Floods, the very complex natural system, are governed by large number of unpredictability and uncertainty variables. Following the recent conclusions of the fourth IPPC climate assessment report[1] that, enhanced meteorological extremes are to be expected during 21st century[2]. The flood frequency with which they occur is on the up-rise in many regions of the world. This phenomena is mainly due to population growth, urbanization process speeding up, fast industry development, especially irrational exploitation of land, water resources, and forest[3]. The social economy damage caused by floods was enormous. It has been reported[4] that floods each year affect more than 3,046,770,000 people in the world and 300,040,000 people in China. In Wujiang River in 2006, floods took 52 lives and caused more than $5.8 billion in economic damages. Flood control and flood risk management, the very complex subject, arouse a widespread interest. Many scholars have achieved a lot in the fields of flood detecting and monitoring[5], flood management[6], flood risk management[7], and flood forecasting[8].

The classification of floods is an important issue for flood control. It is closely connected with flood management and flood forecasting. Although classification and clustering are often mentioned in the same breath, they are different analytical approaches. Classification is similar to clustering in that it also segments customer records into distinct segments called classes. But unlike clustering, a classification analysis requires that the end-user knows ahead of time how classes are defined. At present, there is no unification standard for flood classification. Clustering analysis, which is a classification technique for forming homogeneous group within complex data sets, is useful in flood classification and flood features grouping. Specifically, clustering is a method for sorting out scattered data sets into a kind of groups with rules. It refers to the process of grouping a collection of objects into classes. The clustering results display that objects within the same class are similar, and objects from different classes are dissimilar. Cluster analysis can be used as a form of descriptive characteristics among different floods. The purpose of flood clustering is to sort floods into groups, so that the flood characteristics is similar between members of the same cluster and dissimilar between members of different clusters. That is, flood cluster analysis aims at the identification of groups of flood with a common characteristic.

Fuzzy c-means (FCM), presently the most well known and powerful method in cluster analysis, is the extension of hard c-means (HCM). FCM is an unsupervised technique that has been successfully applied to clustering analysis, such as target recognition, soil clustering[9] and image segmentation[10]. In those literatures, there are numerous studies carried out using fuzzy c-means (FCM) clustering technique for solving several engineering problems. However, the fuzzy c-means algorithm suffers from the serious drawbacks that its performance heavily depends on the initial starting conditions and it may get stuck in sub-optimal solutions. In order to solve this problem and improve the performance of Fuzzy C-Means algorithm, the simulated annealing algorithm is presented. Simulated annealing algorithm employs a random search which not only accepts changes that decrease objective function, but also tolerates some changes that increase it. The major advantage of simulated annealing algorithm over other methods is its ability to avoid becoming trapped at local minima.

The aim of this study is to test the FCM-SA model for flood clustering. At the same time, FCM and FCM-SA and their applications will be introduced in detail in the following sections.

2 ALGORITHMS

2.1 The fuzzy C-means algorithm

The fuzzy C-means algorithm (FCM) is one of the most widely used fuzzy clustering algorithms. FCM was originally introduced by Dunn in 1973[11], and improved by Jim Bezdek[12] in 1981 who used it for fuzzy clustering.

The algorithm is based on minimization of the following objective function:

$$J_m(U,V)=\sum_{i=1}^{n}\sum_{j=1}^{k}u_{ij}{}^{m}d_{ij}^{2},1\le m\le\infty \qquad (1)$$

Where, $d_{ij}^2=\left\|x_j-v_i\right\|^2$, $U=\{u_{ij}\};v=(v_1,v_2,\cdots,v_n)$, u_{ij} is the degree of membership of x_j in the i th cluster, v_i is the i th cluster center, ||*|| is a norm expressing the similarity between any measured data and the cluster center, m is any real number greater than 1.

Fuzzy partition is carried out through an iterative optimization of Equation (1) with the update of membership u_{ij} and the cluster center v_i by

$$u_{ij} = [\sum_{j=1}^{k} (\frac{d_{ik}}{d_{jk}})^{\frac{2}{m-1}}]^{-1} (i = 1, 2, \cdots, n; j = 1, 2, \cdots, k) \tag{2}$$

$$v_i = \frac{\sum_{i=1}^{n} u_{ij}{}^m x_j}{\sum_{i=1}^{n} u_{ij}{}^m} \tag{3}$$

The creation in this iteration will stop when $\max_{ij} \{ \left| u_{ij}^{k+1} - u_{ij}^{k} \right| \} < \varepsilon$, here ε is a termination criterion between 0 and 1, whereas k is the iteration step. This procedure converges to a local minimum or a saddle point of J_m.

2.2 Simulation annealing algorithm

Simulated annealing (SA) is one heuristic process which has successfully solved difficult problems on a consistent basis[13]. SA[14], which presents an optimization technique, originates from the analogy of a physical process in which the matter is moving from a high-energy state to low-energy state. The SA algorithm mainly anneals a physical process, which is the best way to lower the material in order to bring it to a low-energy state. As far back as 1953, Metropolis *et al*[15] proposed an algorithm for the efficient simulation of the evolution of a solid to thermal equilibrium. The procedure of FCM-SA is as follows: 1) Initialization. The algorithm starts with the beginning values of T_0 (temperature), ω (initial solution) and responsible for objective function (J_m, as Equation (1)). 2) A random perturbation is applied to produce a new value (ω') of objective function, and $\omega' = \omega_k$. According to the Metopolis criterion[15], the new value is current energy level. 3) If $J_m > J_{m'}$, then the new energy level is accepted, otherwise, the new energy level is accepted with probability $\mathrm{p} = \mathrm{e}^{-(J_{m'} - J_m)/T}$, here T is the temperature at that equilibrium. 4) By gradually decreasing the current temperature T, repeat the procedure from step two to four. 5) If the annealing process is finished, then continue the next step, otherwise, go back to step two. 6) The current value (ω_k) will be achieved until no more improvements are possible.

3 CASE STUDY

3.1 Descriptions of study area and data

Beijiang River, with the basin area of 46710 km^2, is one of the main branches of the Pearl River which is the largest river in south China. The study area chosen for this research is the Wujiang River, one of the important branches of the Beijiang River. Geographic coordinate of Wujiang River is at latitude of 24°46′ to 25°41′ N and longitude of 112° 23′ to 113°36′E. The Wujiang River begins at the three Mountain Ridges of Linwu County in Hunan province, and traverses two provinces, Hunan and Guangdong provinces. With a river length about 260 km, the drainage area of the Wujiang River is 7097 km^2. The climate in the basin is East Asian monsoon, with an average precipitation between 1300mm to 1500 mm per year. The annual discharge of the Wujiang River at its mouth average over 147×$10^8$$m^3$ per year. The normal natural maximum flows occur during the rainy season from April to September, average approximately 110.25×$10^8$$m^3$ at Lishi (second) Hydrologic Station, which is located near the mouth of the river basin and controls a drainage area of 6976 km^2.

The available flood data of Wujiang River cover a period of 42 years between 1965 and 2006. This paper applies FCM-SA and FCM for flood classification. In addition, this paper examines the feasibility of FCM-SA in flood clustering by comparing it with FCM. This paper reveals the changing regularity of floods in Wujiang River and provides theory reference for valid flood management, watershed comprehensive control and optimizing allocation of water resources.

3.2 Data pretreatment

A flood event is mainly characterized by peak flow, volume and duration, which might be mutually correlated, as illustrated in Figure 1. The flood duration, determined by rainfall duration and land use and land cover of the basin, could be contained within the maximum 3-day and 7-day flood discharges. Flood intensity is one of the main flood characters. In this paper, all flood events are described in terms of flood intensity. The flood intensity index is accordingly computed by peak flow（Q）, water level (H, corresponding to flood peak), the maximum of the 3-day volume (V_{3D}) and 7-day volume (V_{7D}). In this section, we will explain how to compute the maximum 3-day and 7-day flood discharges.

From the physical point of view, generally, there has close correlation among those characteristic values. The correlation coefficient can be expressed by

$$\rho = \frac{E[(X-\mu_x)(Y-\mu_y)]}{\sigma_x \sigma_y} \tag{4}$$

The correlation coefficients between flood peaks and water level (ρ_1), flood peak and the maximum 3-day volume (ρ_2), flood peak and the maximum 7-day volume (ρ_3), water level and the maximum 3-day volume (ρ_4), water level and the maximum 7-day volume (ρ_5) are computed using Equation (4) respectively to be $\rho_1 = 0.9773$, $\rho_2 = 0.9173$, $\rho_3 = 0.8404$, $\rho_4 = 0.8928$ and $\rho_5 = 0.8289$. These correlation coefficients show those characteristic values have close physical relationship, and they determine the flood intensity index.

According Figure 1, the maximum 3-day flood discharge is determined as

$$Volume_{3-day} = \int_{i_{SH}}^{i_{EH}} Q_i - (Q_{i_{SH}} + Q_{i_{EH}})/2 \times (i_{EH} - i_{SH}) \tag{5}$$

Where Q_i is the observed volume of the *i*th hour for a flood event, $Q_{i_{SH}}$ and $Q_{i_{EH}}$ are observed hourly volumes which are responsible for the maximum 3-day volume at the starting and ending hour time, respectively. In Equation (5), i_{EH} subtract i_{SH} get 72 hours, namely three days.

The 3-day flood discharge is expressed as

$$\begin{aligned} Volume_{3-day} &= \int_{i_{SH}}^{i_{EH}} Q_i - (Q_{i_{SH}} + Q_{i_{EH}})/144 \\ &= \sum_{i=i_{SH}}^{i_{EH}} Q_i - \frac{Q_{i_{SH}} + Q_{i_{EH}}}{144} \end{aligned} \tag{6}$$

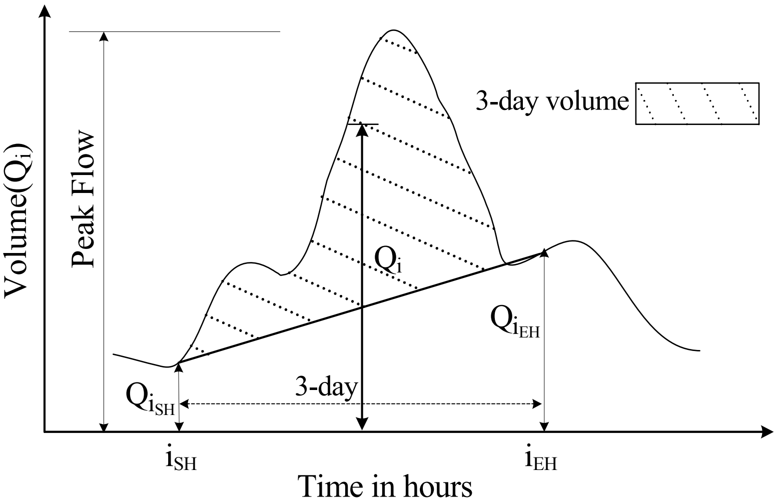

Figure 1: Characteristics of a flood event

The calculation of the 7-day volume has a similar form:

$$\begin{aligned} Volume_{7-day} &= \int_{i'_{SH}}^{i'_{EH}} Q'_i - (Q'_{i_{SH}} + Q'_{i_{EH}})/336 \\ &= \sum_{i=i'_{SH}}^{i'_{EH}} Q'_i - \frac{Q'_{i_{SH}} + Q'_{i_{EH}}}{336} \end{aligned} \tag{7}$$

Using Equations (6) and (7), the maximum 3-day and 7-day flood discharges can be calculated.

The four characteristic values have different physical meaning and different quantity grades. In order to eliminate the influence caused by different units of characteristic and different dimensions of process variables on the predictive outcome, data pretreatment must be put into force. The four types of data pretreatment considered in this work are: 1) mean centering, 2) differentiation, 3) normalization, and4) auto-scaling.

In this paper, normalization has been used for data pretreatment. Giving a flood data matrix x_{ij}^0 which represents the *i*th observation for the *j*th variable, the construction function of normalization is

$$x_{ij} = \frac{x_{ij}^0 - x_{j_{\min}}^0}{x_{j_{\max}}^0 - x_{j_{\min}}^0} \tag{8}$$

Where, x_{ij} is the normalization value. The original flood data matrix x_{ij}^0 is transformed into data matrix x_{ij} ($x_{ij} \in [0,1]$) with the data pretreatment normalization.

3.3 Categories of best classification for flood features

According to the FCM-SA algorithm, the categories of the best classification can be determined by the value of mean objective function ($\overline{J}$) which is defined as

$$\overline{J} = \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{n} u_{ij}^{m} \left\| x_j - v_i \right\|^2 \tag{9}$$

The value $\overline{J}$ displays the mean distance between the individual sample and the samples center. The number of centers is the categories of classification. The best classification should correspond to the minimum $\overline{J}$ value. The relationship between number of centers and mean objective function value $\overline{J}$ is shown in Figure 1.

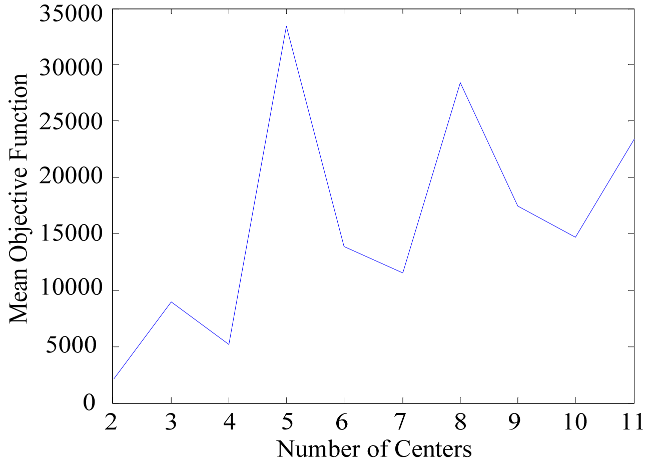

Figure 2: The relationship between number of centers and mean objective function

Figure 2 shows that the minimum mean objective function value corresponds to two sample centers. That means the flood features should be divided into two different types. However, the practical flood features in a river basin are too different to be clustered by two categories. From Figure 2, the second smaller value of mean objective function corresponds to four sample centers. Based on this, the flood features can be divided into four different types which matches the real flood situation in Wujiang River Basin: (I) catastrophic flood, (II) big flood, (III) generally flood and (IV) small flood. Therefore, here the best classification for flood features is four categories.

3.4 Results of flood clustering

The primary purpose of flood cluster analysis is to assemble objects based on flood characteristics. The clustering results of objects should exhibit high internal (within-cluster) homogeneity of flood features to detect similar groups among the sampling floods. The clustering results obtained using the FCM-SA and FCM algorithms for floods in Wujiang River are listed in Table 1.

Items	The clustering results	
	FCM-SA	FCM
I	2006	2006
II	2002，1994，1968	2002，1994，1968
III	2005，2001，1999，1998，1995，1993，1992，1985，1983，1982，1980，1978，1977，1976，1975，1973，1972，1969	2005，2001，1999，1998，1997，1995，1993，1992，1985，1983，1982，1981，1980，1978，1977，1976，1975，1973，1972，1969
IV	2004，2003，2000，1997，1996，1991，1990，1989，1988，1987，1986，1984，1981，1979，1974，1971，1970，1968，1967，1966，1965	2004，2003，2000，1996，1991，1990，1989，1988，1987，1986，1984，1979，1974，1971，1970，1968，1967，1966，1965

Table 1: The flood clustering results of Wujiang River

Table 1 showed that the two approaches, FCM and FCM-SA, obtained basically the same floods clustering results. Only little difference existed for the floods in 1997 and in 1981, which are small floods (IV) by FCM-SA, but belong to generally floods (III) by FCM. Generally speaking, clustered results of both FCM-SA and FCM are accordance with the actual situation of floods in Wujiang River. FCM-SA algorithm has been successfully applied for flood clustering. Fuzzy c-means clustering analysis produced clear membership patterns and clustered flood types effectively.

3.5 Evaluation of FCM and FCM-SA

Here we run FCM-SA and FCM independently ten times to compare their effectiveness in flood clustering. The number of iterations and the corresponding iterative times of FCM-SA and FCM are listed in Table 2. It can been seen that all the ten times operation give the results that, both the number of iterations and the iterative time (s) of the calculation run by FCM-SA are smaller than FCM. This means that the clusters obtained by FCM-SA have, on average, better running performance than FCM. Put in another way, it may also indicate that FCM-SA is able to escape sub-optimal solutions better than FCM.

Dataset	The compare projects			
	The number of iterations		The iterative time (s)	
	FCM-SA	FCM	FCM-SA	FCM
1	4246. 73	8013.45	2.9844	4.0915
2	4613. 59	7195.75	3.0402	3.8951
3	4435.13	8211.43	3.0220	3.9457
4	4216.08	7819.48	3.1184	4.2704
5	4623.54	8046.52	3.0938	3.8473
6	4349.73	7952.09	3. 0703	3.9431
7	4773.18	7974.44	3.0665	4.2404
8	4909.24	8017.64	3.0875	4.1124
9	4178.24	7909.13	3.1674	3.8914
10	4876.08	7875. 62	3.1295	3.9700
The mean value	4545.15	7904.44	3.0789	4.0207

Table 2: Performance comparison between FCM and FCM-SA

4 CONCLUSIONS

Using the yearly flood data of Lishi Station in Wujiang River, hierarchical cluster analysis grouped the 41 sampling floods into four categories of flood characteristics by FCM-SA and FCM respectively. The categories indentified for the 41 sampling floods by both FCM-SA and FCM are basically the same and match the practical flood features in Wujiang River, which indicated that the classification of flood features by both FCM-SA and FCM are reasonable.

The biggest advantage of FCM algorithm in flood features clustering is its higher efficiency in clustering large data sets. However, its use is often restricted to numeric data because this algorithm minimizes a cost function by calculating the means of clusters. The SA algorithm presented in this paper removes this limitation and keeps the high efficiency at the same time. A flood clustering model has been created with the application of FCM-SA, in which SA algorithm is used for its faster convergence speed and less iterations to overcome the shortcomings of traditional FCM algorithm. The clustering performance of the two algorithms, FCM-SA and FCM, has been evaluated using flood features of Wujiang River. Both the number of iterations and the iterative time (s) of the calculation run by FCM-SA are smaller than FCM. The clusters obtained by FCM-SA have, on average, better running performance than FCM. The satisfactory results have demonstrated the effectiveness of the FCM-SA for flood clustering. FCM-SA can improve the results of FCM for an unsupervised classification. The use of FCM-SA algorithm in flood clustering analysis will have very good prospects.

ACKNOWLEDGEMENTS

This research is support by The research is financially supported by the Key National Natural Science Foundation of China (Grant No. 50839005), the project from Guangdong Science and Technology Department (Grant No.2010B050300010), the project from Guangdong Water Resources Department (Grant No.2009-39), the program for Guangdong Climate Change (Grant No.ACCC/20100705-1) and the key program cultivating research from Sun Yat-sen University (Grant No.10lgzd11).

REFERENCES

[1]IPCC: Fourth Climate Assessment Report, IPCC, Geneva. (2007).

[2]Reggiani P, Weerts A H. A Bayesian approach to decision-making under uncertainty: An application to real-time forecasting in the river Rhine. Journal of Hydrology, 356:56-69(2008).

[3]Drogue G, Pfister L, Leviandier T, et al.. Simulating the spatio-temporal variability of streamflow response to climate change scenarios in a mesoscale basin. Journal of Hydrology, 293: 255-269(2004).

[4]EM-DAT. The OFDA/CRED (Office of Foreign Disaster Assistance /Centre for Research on the Epidemiology of Disasters) International Disaster Database, Université Catholique de Louvain, Brussels, Belgium. Downloaded from http://www.em-dat.net/disasters/profiles.php. (2002).

[5]Felipe Ip, Dohm J M, Baker VR, et al. Flood detection and monitoring with the Autonomous Sciencecraft Experiment onboard EO-1. Remote Sensering of Environment, 2006, 101: 463-481.

[6]Plate E J. Flood risk and flood management. Journal of Hydrology. 267: 2-11(2002).

[7]Johnson R. Flood Planner: A manual for the natural management of river floods. WWF Scotland, Edinburgh. (2007).

[8]Kim G, Barros A P. Quantitive flood forecasting using multisensor data and neural network. Journal of Hydrology, 246: 45-62(2001).

[9]Goktepe A B, Altun S, Sezer A. Soil clustering by fuzzy c-means algorithm. Advances in Engineering Software, 36: 691-698(2005).

[10]Chuang K S, Tzeng H L, Chen S, et al. Fuzzy c-means clustering with spatial information for image segmentation. Computerized Medical Imaging and Graphics, 30: 9-15(2006).

[11]Dunn J C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. Journal of Cybernetics, 3: 32-57(1973).

[12]Bezdek J C. Pattern Recognition with Fuzzy Objective Function Algoritms. Plenum Press, New York, (1981).

[13]Murray A T, Richard L. Church. Applying Simulated Annealing to Location-Planning Models. Journal of Heuristics, 2: 31-53(1996).

[14]Kirkpatrick S, Gelatt C D, Vecchi M P. Optimization by Simulated Annealing. Science, 220(4598): 671-680(1983).

[15]Metropolis N, Resenbluth A W, Resenbluth M N, et al. Equations of state calculation by fast computing machines. J.Chem. Phys., 21:1087-1091(1953).