Available online at www.sciencedirect.com
ScienceDirect
Advances in Water Resources 29 (2006) 1573–1585
Advances in Water Resources
www.elsevier.com/locate/advwatres

Performing cluster analysis and discrimination analysis of hydrological factors in one step

Gwo-Fong Lin *, Chun-Ming Wang

Department of Civil Engineering, National Taiwan University, Taipei 10617, Taiwan

Received 19 January 2004; received in revised form 14 November 2005; accepted 21 November 2005
Available online 19 January 2006

Abstract

Based on self-organizing map, a method that can perform cluster analysis and discrimination analysis in one step is proposed in this paper. Using the proposed method, one can view the relative topological relationships of input patterns, determine the proper number of clusters, and assign unknown patterns to known clusters without losing any information of input patterns. Regarding the capability of determining the proper number of clusters, the proposed method is superior to conventional cluster analysis. The discrimination results also show that the assignments of unknown patterns to known clusters are reasonable using the proposed method. The advantages of the proposed method are also demonstrated by an application to the hydrological factors affecting low-flow duration curves in southern Taiwan.

Keywords: Cluster analysis; Discrimination analysis; Self-organizing map; Neural network; Low-flow characteristics; Hydrogeological factors

1. Introduction

Hydrologists often encounter a problem that certain hydrological information is required but unavailable at an ungauged site. Such a problem can be solved using regionalization. Regionalization is a useful tool to extrapolate certain hydrological information at an ungauged site using the information of gauged sites [13]. The substances of regionalization are a set of hydrological factors that can describe certain hydrological information properly. For different applications, regionalization may use different sets of hydrological factors describing different hydrological information. Regionalization contains two tasks: delineating the hydrologically homogeneous regions and developing a regional estimation method. Delineating hydrologically homogeneous regions is to discover the similar hydrological characteristics among gauged sites so that the accuracy of the extrapolation can be improved. The identification of the homogeneities of hydrological factors is the key point of delineating hydrologically homogeneous regions. The regional estimation method is often a set of regression models based on several different classes of certain hydrological information at gauged sites. Then the extrapolation of certain hydrological information at an ungauged site can be performed using the corresponding regression model. Therefore, for well extrapolations, one needs an appropriate method to help develop and choose the corresponding regression model for an ungauged site.

Delineating hydrologically homogeneous regions may combine several procedures [8,13–16], including principal component analysis, cluster analysis, and discrimination analysis (abbreviated as PCD herein). It is necessary to explain briefly the objectives of the three multivariate statistical techniques of PCD used in regionalization. In conventional regionalization, principal component analysis is usually used to reduce

* Corresponding author. Tel.: +886 2 33664368; fax: +886 2 23631558.
E-mail address: gflin@ntu.edu.tw (G.-F. Lin).

0309-1708/$ - see front matter
doi:10.1016/j.advwatres.2005.11.008

the dimension of the input data. For example, Yu et al. [16] employed principal component analysis to select several predominating components, so the dimensions of the data that was linearly formulated using the original hydrological factors were reduced. With the reduced dimension of input data, the computation complexities of the subsequent analysis are lowered, but important information is lost [1]. Nevertheless, the complete information contained in the raw hydrological factors should be preserved. If principal component analysis is only used to reduce the dimension of input data, principal component analysis will not be an appropriate procedure for regionalization.

The second procedure of conventional regionalization, cluster analysis, is to explore the relationships of hydrological factors at gauged sites. After cluster analysis, the grouping of hydrological factors can be derived and the hydrological homogeneous areas can then be delineated. Furthermore, the regional estimation method that is often a set of regression models is developed. The set of regression models is derived using regression analysis. Each cluster has its corresponding regression model that is derived based on the data members of the cluster.

The third procedure of conventional regionalization, discrimination analysis, is to build a model to assign an ungauged site to a known cluster, so that a proper regression model can be selected to extrapolate the specific hydrological information. It should be noted that the hydrological factors of ungauged sites are given but are not analyzed by cluster analysis. The objective of discrimination analysis is accomplished using two steps. The first step is to find out the discriminants of the hydrological factors of gauged sites whose grouping is already known. Based on the discriminants, the second step is to develop the model to appropriately assign ungauged sites to known clusters. The cluster that an ungauged site belongs to can then be determined using the model developed by discrimination analysis. Therefore, one can choose the proper regression model for the ungauged site.

For a clear description of cluster analysis, the following mapping equation is used:

$$\Phi : \mathbf{P} \rightarrow \mathbf{Q} \tag{1}$$

where $\mathbf{P}$ and $\mathbf{Q}$ are finite sets, and Φ is the mapping. In regionalization, $\mathbf{P}$ represents the set of hydrological factors at gauged sites. $\mathbf{Q}$ is the set of clusters that $\mathbf{P}$ is classified into. The objective of cluster analysis is to find an appropriate mapping Φ without the prior knowledge of $\mathbf{P}$ (i.e., the grouping $\mathbf{Q}$ of the data set $\mathbf{P}$). Thus, the term "unsupervised learning" is applied to cluster analysis [1,4,7]. After the cluster analysis, the grouping of the input data is detected. The number of clusters and the members belonging to the corresponding cluster are both determined. According to the results of cluster analysis, the hydrological homogeneous regions are delineated. Then the regional estimation method can be developed and the accuracy of the extrapolation can be improved.

It is clearer to explain discrimination analysis using Eq. (1). In discrimination analysis, the prior knowledge of data set $\mathbf{P}$ (i.e., the grouping $\mathbf{Q}$ of the data set $\mathbf{P}$) is known. The objective of discrimination analysis is to establish a proper mapping Φ so that the data set $\mathbf{P}$ can be classified into the appropriate clusters. Meanwhile, new data that is not analyzed by cluster analysis can be assigned to the proper cluster using the results of the above discrimination analysis. Since the prior knowledge of $\mathbf{P}$ is known, the term "supervised learning" is applied to discrimination analysis. In conventional regionalization, discrimination analysis develops a classification model based on the results of cluster analysis. The classification model is used to assign an ungauged site to a known cluster according to properties of the hydrological factors of the ungauged site.

A sufficient method of assigning ungauged sites to known clusters is necessary for regionalization. Some studies use only cluster analysis for regionalization [2,11]. However, choosing a proper regression model for an ungauged site is difficult using only cluster analysis. It is realized that PCD comprises three complicated statistical techniques. In most cases of regionalization, the amount of hydrological information is rather small due to the limitation of the data acquiring techniques. Hence, PCD is not an efficient method for regionalization.

One problem of the conventional cluster analysis is to determine the number of clusters. For hierarchical cluster analysis methods, the dendrogram is used to show the relative topological relationships of input patterns and the number of clusters is determined using a certain complicated statistic computed from the dendrogram [1,3]. Different hierarchical cluster analysis methods (e.g., single linkage and complete linkage methods) often lead to different dedrograms [12], so the clustering results derived from the different dendrograms are inconsistent. For non-hierarchical cluster analysis methods such as K-means [6], the topological relationships of input patterns cannot be easily obtained. The number of clusters should be determined in advance although the actual grouping of input patterns is unknown. Thus, the determination of the proper number of clusters is also a problem for non-hierarchical cluster analysis methods. In conclusion, the determination of the proper number of clusters is a subtle problem for conventional cluster analysis methods.

When the conventional cluster analysis is applied to a data set, the relationships among the data set are discovered. That is the mapping Φ (Eq. (1)) is found. Intuitively, researchers may think that they can assign a pattern (known or unknown) to a known cluster using the mapping Φ, since the mapping Φ is found. However, the assignments of patterns cannot be done only with

the conventional cluster analysis. Because the conventional cluster analysis just provides algorithms to analyze the data set, it does not provide a facility to store the knowledge (i.e. the mapping Φ) of data set. After the conventional cluster analysis is performed, the knowledge of data is discarded. Only the relationships of the data are discovered. Thus, for the assignments of known or unknown patterns, discrimination analysis is necessary. In short, PCD is not an intuitional method for hydrologists.

Artificial neural network is now a popular tool to deal with massive and complex data to derive useful information. There are many kinds of artificial neural networks categorized by its learning process. The artificial neural network used herein is the Self-Organizing Map (SOM) proposed by Kohonen [7]. SOM is a competitive and unsupervised network. The term, "unsupervised", means that the knowledge of environment is not learned from the specific input–output examples. Instead, it learns the knowledge of environment only from the input patterns and then stores the knowledge in the network. An attractive capability of SOM is to map high dimensional input patterns onto a lower dimensional output space and to preserve the topological relations of input patterns. The characteristic, coupled with the unsupervised nature of its learning algorithm, has rendered the SOM an attractive alternative for solving various problems that traditionally have been the domain of conventional statistical and operational research techniques. SOM is often used to extract the specific features and to discover the statistical distribution of a complex phenomenon. Mangiameli et al. [9] compared SOM with other seven hierarchical cluster analysis methods. Their result shows that the performance of SOM in clustering messy data is better than that of the other seven hierarchical clustering methods. Michaelides et al. [10] adopted the SOM to classify the rainfall variability to provide prototype classes of weather variability. Their results show that SOM can detect much more detail of rainfall variability than hierarchical cluster analysis methods can. For the regionalization of the flood frequency, Zhang and Hall [18] compared SOM, Ward's method and the Fuzzy C-means approach. Their results indicate that SOM is preferable over the other two methods.

The nature of the unsupervised network, SOM, is similar to the conventional cluster analysis. However, unlike the conventional cluster analysis, the knowledge of input patterns can be stored in network itself. This is a fascinating advantage of SOM over the conventional cluster analysis. The purpose of this paper is to propose a simple method for delineating hydrological homogeneous regions. As aforementioned, principal component analysis may not be appropriate for regionalization. For facilitating the delineation of hydrological homogeneous regions in an intuitional way, the objective of this paper is to develop a method that can perform cluster analysis and discrimination analysis in one step. First, the algorithm and architecture of the SOM are presented. On the basis of the nature of "unsupervised learning", SOM is a good tool for clustering. Then, based on SOM, a method combining cluster analysis and discrimination analysis is developed. Finally, the low-flow characteristics in southern Taiwan are analyzed using the proposed method to identify their homogeneity and to verify the capability of assigning unknown patterns to known clusters.

2. Method

2.1. Algorithm of SOM

The essential mechanism of SOM is the competitive and unsupervised learning process in which the neurons of the network compete each other to be activated. The output space of SOM can be one- or two-dimensional. Higher dimensions of the output space are acceptable but not common. SOM has two layers: the input layer containing the input nodes, and the Kohonen layer with numerous neurons fully connected by every input node in the input layer. A SOM with two input nodes and twenty-four neurons is shown in Fig. 1. SOM is an iterative algorithm containing three processes: the competitive process, the cooperative process and the adaptive process.

2.2. The competitive process

Let an input pattern denoted by

$$\mathbf{x} = [x_1, x_2, \ldots, x_m]^{\mathrm{T}} \tag{2}$$

where m is the dimension of the input pattern $\mathbf{x}$. The synaptic weights vector of each neuron has the same dimension as input patterns. Let the synaptic weights vector denoted by

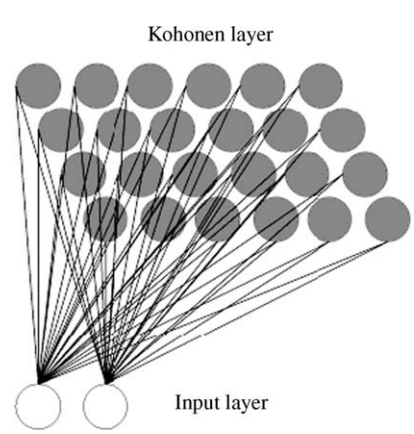

Fig. 1. Architecture of SOM.

$$\mathbf{w}_j = [w_{j1}, w_{j2}, \ldots, w_{jm}]^{\mathrm{T}}, \quad j = 1, 2, \ldots, l \tag{3}$$

where l is the total number of neurons in the network. The synaptic weights are initialized as small random numbers. In the competitive process, the neurons of the network compete each other to determine which one to be activated. The neuron that is activated is called the winning neuron. The way to determine which neuron is the winning neuron is to find the neuron that best matches the current input pattern feeding to the SOM. The measure of the similarity between neurons and input patterns is the Euclidean distance. Hence, we may determine the winning neuron by applying the condition [7]:

$$i(\mathbf{x}) = \arg\min_j \|\mathbf{x} - \mathbf{w}_j\|, \quad j = 1, 2, \ldots, l \tag{4}$$

where $i(\mathbf{x})$ is the neuron that best matches the corresponding input pattern $\mathbf{x}$ and the $\|\cdot\|$means the Euclidean distance.

2.3. The cooperative process

In the cooperative process, the influence (i.e. lateral interaction) of the winning neuron is delivered to its neighboring neurons. The location of the winning neuron is the center of the topological neighborhood of cooperating neurons. The topological neighborhood implies the lateral interactions between the winning neuron and its neighborhood. The amplitude of the topological neighborhood $h_{j,i(\mathbf{x})}(n)$ should decreases monotonically with the lateral distance. Thus, a typical $h_{j,i(\mathbf{x})}(n)$ is defined by the Gaussian function [7]:

$$h_{j,i(\mathbf{x})}(n) = \exp\left(-\frac{d_{j,i}^2}{2\sigma^2(n)}\right), \quad n = 0, 1, 2, \ldots \tag{5}$$

where $d_{j,i}$ is the Euclidean distance between the winning neuron i and the activated neuron j in the output space, $h_{j,i(\mathbf{x})}(n)$ is the topological neighborhood at time n between the winning neuron i and the excited neuron j, and $\sigma(n)$ is the effective width which corresponds to the radius around neuron j at time n. The $\sigma(n)$ should decrease monotonically with time. Readers can obtain more details from Kohonen [7].

2.4. The adaptive process

In adaptive process, the synaptic weights are adjusted according to input patterns. The adjustment of synaptic weights is based on the Hebbian hypothesis [7]. The algorithm that adjusts the synaptic weights is defined as follows [7]:

$$\mathbf{w}_j(n+1) = \mathbf{w}_j(n) + \eta(n) h_{j,i(\mathbf{x})}(n)(\mathbf{x} - \mathbf{w}_j(n)) \tag{6}$$

where $\eta(n)$ is the learning rate at time n, $\mathbf{w}_j(n+1)$ is the synaptic weights of neuron j at time $n+1$, $\mathbf{w}_j(n)$ is the synaptic weights of neuron j at time n, and $h_{j,i(\mathbf{x})}(n)$ is the topological neighborhood as defined in Eq. (5). The learning rate should shrink with time monotonically as shown in the following equation [7]:

$$\eta(n) = \eta(0) \exp\left(-\frac{n}{1000}\right) \tag{7}$$

where $\eta(0)$ is the initial learning rate.

As the iteration of SOM proceeds, the winning neuron and the neighboring neurons become more and more similar to the corresponding input pattern. Thus, the synaptic weights of the winning neuron and the neighboring neurons move closer to input patterns.

2.5. SOM-based cluster and discrimination analysis (SOMCD)

In this section, since the relationships of input patterns can be stored, a SOM-based cluster and discrimination analysis (referred to as SOMCD hereafter) method is proposed. SOM is adopted herein to perform a transformation of input patterns into a two-dimensional discrete lattice to reveal the topological locations and statistical distributions of input patterns.

After the SOM training is done, feeding the SOM with all input patterns that have learned by the SOM can lead to the feature map. Applying Eq. (3) to a trained SOM with all its input patterns yields the feature map. The feature map is a two dimensional lattice and each grid represents one neuron. The way to obtain the feature map is to label all winning neurons (some specific grids) in the output space (the lattice) with the identities of corresponding input patterns. With the feature map, the relative topological relationships between input patterns can be identified. The location of a winning neuron in the feature map shows the topological location of a corresponding input pattern in the input space, and the density of neurons shows the statistical distribution of input patterns.

If a neuron responds to a specific input pattern, the neuron is called the image of the specific input pattern or the neuron is "imaged" by the specific input pattern. The density map can be obtained by applying the following equation to a trained SOM:

$$N = \mathrm{Num}(i_j), \quad j = 1, 2, \ldots, l \tag{8}$$

where N is an integer, i is the neuron of the trained SOM, and Num( ) is a function counting the number of the neuron i "imaged" by certain input patterns. Every pattern in the input space has only one image, but one neuron can be the image of many input patterns. Patterns that are close in the input space tend to crowd their images in a certain place of the feature map [19]. The void in the input space where there is no input pattern is also shown in the feature map. According to this property, the density map can reveal the grouping of in-

put patterns. From Eq. (8), the density map can be obtained easily by labeling each grid of the map with the integer N. Suppose that the number of each grid is the "elevation" of the density map. Then the grouping of input patterns is shown by certain isolated "plateaus" separated by "valleys" on the density map. The "plateaus" imply the aggregations of input patterns that are mapped onto the density map. The "valleys" imply the void of input patterns that are mapped onto the density map. Thus, according to the variation of the "elevation" on the density map, the grouping of the input patterns can be easily identified. The valleys are the clusters' boundaries. Therefore a proper number of clusters can be determined.

There are some neurons "imaged" by certain input patterns in the feature map. According to the aforementioned feature map and density map, labeling the "imaged" neurons with the identity of the corresponding cluster forms a part of a discrimination map. There are still some blank neurons not labeled. A complete discrimination map can be obtained by applying the following equation to the blank neurons:

$$I = C(i) = \arg\min_{j} \|\mathbf{w}(i) - \mathbf{x}_j\|, \quad j = 1, 2, \ldots, p \tag{9}$$

where I is the identity of a specific cluster, $C(\ )$ is a function indicating the cluster to which the blank neuron i belongs, $\mathbf{w}(i)$ is the weights of the blank neuron i, $\mathbf{x}$ is a input pattern, and p is the number of input patterns. The identity of a blank neuron is the identity of the input pattern which best matches the synaptic weights of the blank neuron. A complete discrimination map can then be obtained by labeling all blank neurons with their corresponding identities I as shown in Eq. (9). The discrimination map can be divided into a certain number of regions. Each region represents a specific cluster. If a pattern is mapped onto a specific region of the discrimination map, the pattern belongs to this specific cluster.

All the three maps (feature map, density map and discrimination map) derived from SOMCD are for visual inspection to discover the relationships of input patterns. For visual inspection, the hexagonal lattice of SOM is more ideal than the square lattice, since the hexagonal lattice does not only emphasize horizontal and vertical directions [7]. Thus in this paper, the lattice of SOM is chosen to be hexagonal. The dimension of SOM is another critical item. Kohonen [7] has suggested the rectangular dimension is better than square one. A well choice of the dimension of a SOM is $q \times r$ where q and r are different positive integers. A SOM of this type is well oriented along with its input patterns and be more stable during the adaption process [7]. The rectangular SOM is adopted in this paper.

The adaption process of SOM involves two phases: an ordering phase followed by a tuning phase [5]. During the first phase, the topological ordering of the weights of the neurons occurs. And then during the second phase, the map is fine tuned so that the statistical properties of the input patterns are captured accurately. The parameters (i.e. learning rate, radius, dimension and training epochs) of SOM during these two phases may significantly influence the results of SOMCD.

2.6. Components of SOMCD

As shown in Fig. 2, the feature map, the density map and the discrimination map are the essentials of SOMCD. The flowchart of SOMCD is shown in Fig. 2. The initialization stage is to determine the architecture of SOM, and the sampling stage is to draw an input pattern to SOM. In the adaption stage, the aforementioned competitive, cooperative and adaptive processes are performed sequentially. The stop condition is when SOM does not change anymore through adaption. If the stop condition is not satisfied, the analysis returns to the sampling stage. Once the stop condition is satisfied, a SOM can be obtained from which the feature map can be derived. Furthermore, the density map can be derived from the feature map. Finally, based on both the feature map and the density map, the discrimination map can be obtained. It should be noted that the feature map, the density map and the discrimination map are all derived from the same SOM.

The functions of the feature map, the density map and the discrimination map are stated below. The relative topological relationships of input patterns can be identified using the feature map, and the proper number of clusters can be easily determined using the density map. Comparing the feature map and the density map, one can obtain the members of each cluster. A pattern can be properly assigned to a known cluster using the discrimination map.

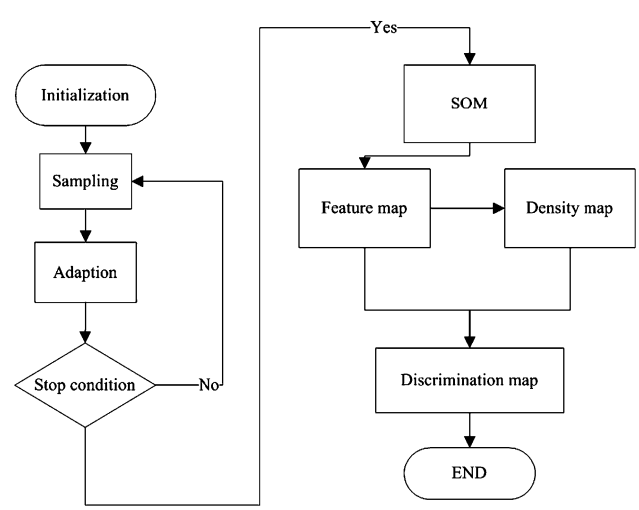

Fig. 2. Flowchart of SOMCD.

3. A case study

In this section, the hydrological factors affecting low-flow characteristics in southern Taiwan are analyzed using SOMCD. The procedures of the analysis are presented and the results are then discussed.

3.1. Study area and data description

Fig. 3 shows the study area and the locations of 33 streamflow gauges. The study area has an area of about 6000 km^2. The flow-duration curves for rivers at these 33 streamflow gauges and the hydrogeological formations of the corresponding watersheds are the data set analyzed in this paper. It should be noted that the gauge number also serves as the identity of the watershed. Watersheds 3, 16 and 17 are randomly chosen for the validations of the discrimination capabilities of SOMCD. Both the spatial and temporal distributions of rainfall in the study area are highly non-uniform. Thus the low-flow characteristics are important information for the planning and management of the water resource in this area.

In this paper, the term "low-flow" is defined as the portion of the daily flow duration curve with more than 30% time flow exceeding indicated value. For three reasons, the nine hydrogeological parameters of the 30 watersheds are used to delimit the homogeneous regions in this paper. First, the low-flow characteristics are significantly influenced by the hydrogeological formations of a watershed [15,16]. Second, these parameters vary little in time so that the delineations are stable. Third, these hydrogeological parameters can be estimated at ungauged watersheds, so it is possible to assign an ungauged site into the proper homogeneous region. According to the hydrogeological map [17], there are nine hydrogeological formations (lake and eight surficial rocks). The properties of these nine hydrogeological formations are listed in Table 1. Ratios of each hydrogeological formation's area to the whole watershed area are used as the hydrogeological factors herein.

3.2. Results

The three maps of SOMCD are all derived from the same SOM, but the interpretations are different. Regarding these three maps, it should be noted that the horizontal and vertical directions have no explicit meanings, and only the relative topological relationships and densities of the images of the input patterns are concerned.

Our experiences on SOMCD show that the radius (i.e. $\sigma(n)$ of Eq. (5)) is the dominant parameter. The influence due to the radius on SOMCD is given in this section. For demonstrating the influence of the radius, different settings of the radius are used to perform SOMCD while other parameters are fixed. The values of the radius are quite different during the two phases of SOM. During the ordering phase, the radius should be set to cover all neurons at the beginning of the phase and then shrinks to the distance between two neighboring neurons [7]. All SOMs used in this paper follow the principle. During the tuning phase, the initial value of the radius should be set to a smaller value than that during the ordering phase. For convenience, the initial value of the radius during the tuning phase is defined as $\sigma_t(0)$. Different $\sigma_t(0)$ may cause different results of SOMCD. Also, the radius during the tuning phase shrinks to the distance between two neighboring neurons.

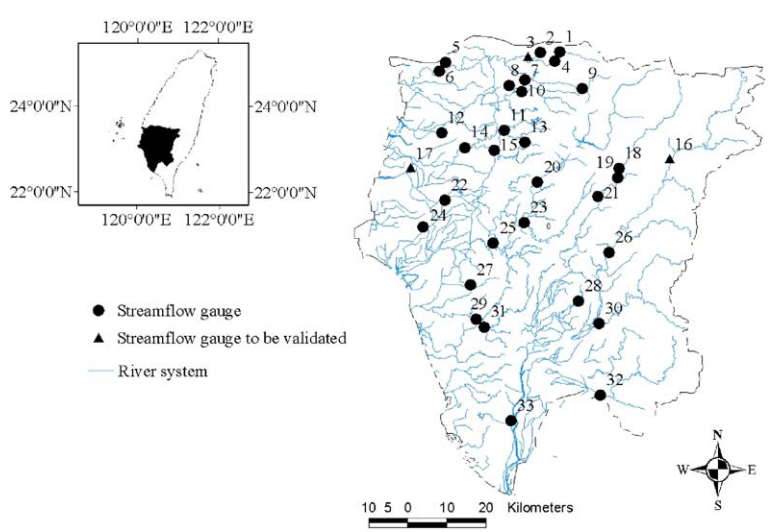

Fig. 3. The study area and the locations of 33 streamflow gauges.

Table 1
Properties of lake and different types of rocks

Symbol	Formation	Property
Rock 1	Conglomerate and pyroclastics	Parts of the area of this rock have aquifers
Rock 2	Lateritic terrace deposits	Parts of the area of this rock have high yielding aquifers
Rock 3	Terrace deposits	Most parts of the area of this rock have high yielding aquifers
Rock 4	Recent alluvium	High yielding aquifers distributing over downstream area
Rock 5	Mudstone	No aquifer in the area of this rock
Rock 6	Shale and argillite	Area of this rock has poor yielding aquifers
Rock 7	Sandstone	Area of this rock has poor yielding aquifers
Rock 8	Limestone and coral reef	Parts of the area of this rock have few poor yielding aquifers
Lake	Lake	Providing recharge of groundwater and supply of baseflow

The input patterns of SOMCD are the 30 sets of nine hydrogeological factors for the 30 watersheds in the study area. Both the percentages and the spatial distributions of the nine hydrogeological parameters of the studied watersheds are different. As a consequence, the weightings that imply the influences of the nine hydrogeological parameters on the low-flow characteristics are not easily assessed. There is no prior information about the degree of the respective hydrogeological factors affecting the low-flow characteristics. Thus, assuming these nine hydrogeological parameters have the same influence on the low-flow characteristics is a choice. In this paper, we assume these nine hydrogeological factors have the same weight. If the evidence of giving different weights to these nine hydrogeological parameters is found, different weights shall be assigned to these parameters.

Two SOMs of different dimensions are used in this paper. One is 5×3 and the other is 7×4. Two different $\sigma_t(0)$ are used to train the two SOMs. Other parameters (i.e., training epochs and learning rate) of the SOMs are all the same. The training epochs of the ordering phase are 1000 and those of the tuning phase are 4000. The initial learning rates during the ordering phase and the tuning phase are respectively 0.9 and 0.02. The grids of the three maps (feature maps, density maps and discrimination maps) represent the neurons in the output space. The feature maps derived from the two SOMs with

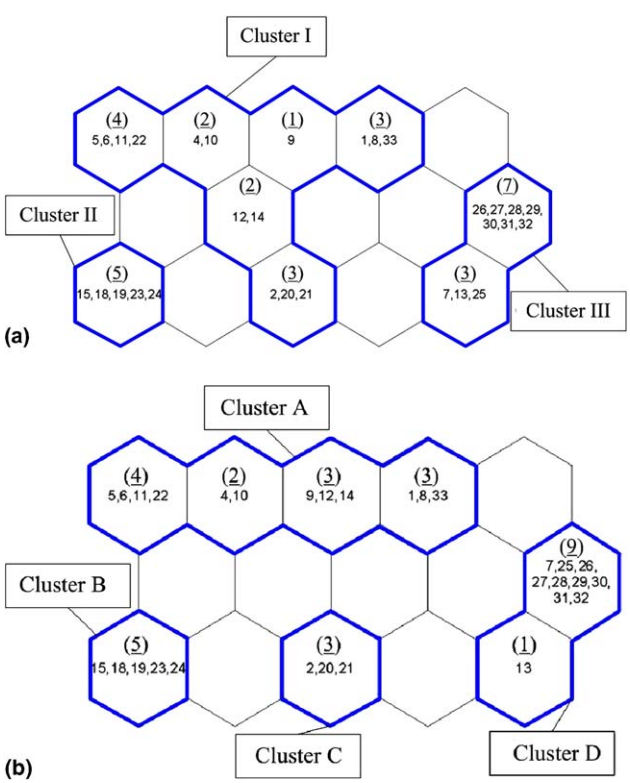

Fig. 4. Combinations of the feature maps and the density maps derived from the SOM of dimension 5×3 with (a) $\sigma_t(0) = 1$ and (b) $\sigma_t(0) = 2$.

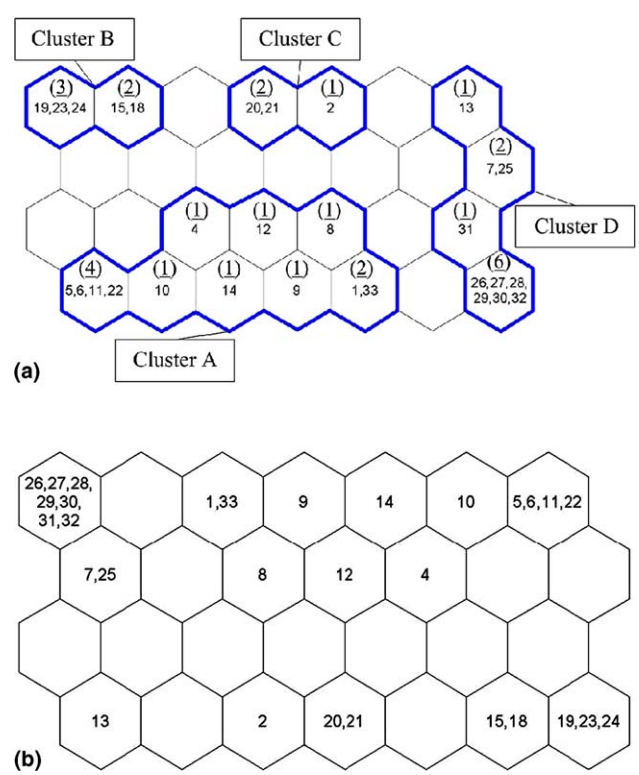

Fig. 5. Combinations of the feature maps and the density maps derived from the SOM of dimension 7×4 with (a) $\sigma_t(0) = 1$ and (b) $\sigma_t(0) = 2$.

different $\sigma_t(0)$ are combined with the corresponding density maps in Figs. 4 and 5, respectively. The numbers in each grid of feature maps refer to the identities of watersheds (the gauge numbers). The numbers underlined in the parentheses are the "elevation" of the grids. The "elevation" of blank grids is zero. Regions surrounded with bold lines are isolated plateaus. According to the density map of Fig. 4(a), the 30 sets of hydrological factors are classified into three clusters. According to the density maps of Figs. 4(b) and 5(a), the 30 sets of hydrological factors are classified into four clusters. The members of each cluster can be identified by comparing the feature map and the corresponding density map. One can find that the clustering results of the SOM of dimension 5×3 with $\sigma_t(0) = 2$ (Fig. 4(b)) and the SOM of dimension 7×4 with $\sigma_t(0) = 1$ (Fig. 5 (a)) are identical. Cluster II (Fig. 4(a)) is identical to Cluster B (Figs. 4(b) and 5(a)). Cluster III (Fig. 4(a)) is identical to Cluster D (Figs. 4(b) and 5(a)). The discrimination maps derived from the two SOMs with different $\sigma_t(0)$ are given in Figs. 6 and 7. In the discrimination maps, symbols (3), (16) and (17) label the images of watersheds 3, 16 and 17 which are for validations of discrimination capability of SOMCD. In Fig. 6(a), watersheds 3, 16 and 17 are assigned to clusters **I**, **II** and **III**, respectively. The discrimination maps in Figs. 6(b) and 7 both indicate that watersheds 3 and 16 are assigned to cluster A and watershed 17 to cluster D. The results of discrimination in Figs. 6(b) and 7 are the same.

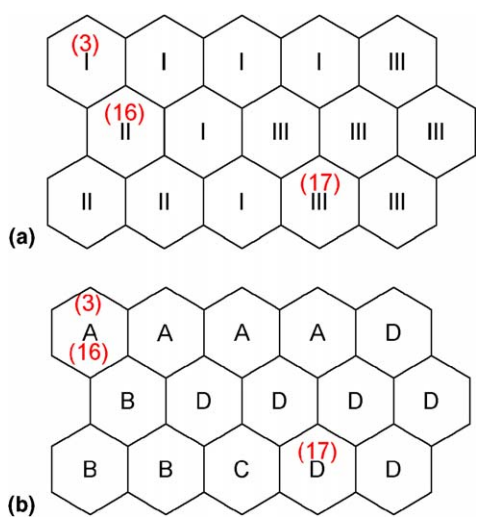

Fig. 6. The discrimination maps derived from the SOM of dimension 5×3 with (a) $\sigma_t(0) = 1$ and (b) $\sigma_t(0) = 2$.

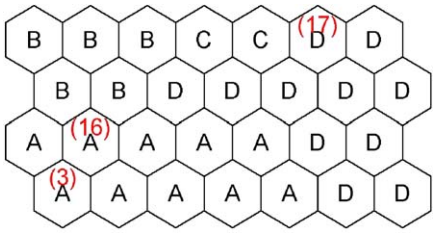

Fig. 7. The discrimination map derived from the SOM of dimension 7×4 with $\sigma_t(0) = 1$.

4. Results discussions

4.1. Demonstrations of the relative topological relations of input patterns using the feature map

The feature map provides topological relationships of 30 sets of hydrogeological factors. In Figs. 4 and 5, the images of watersheds 5 and 6 all fall in the same grid. The image of watershed 25 is in the opposing side of the image of watersheds 5 and 6. This phenomenon implies that the hydrogeological factors of watersheds 5 and 6 are similar. And the hydrogeological factors of watershed 25 significantly differ from those of watersheds 5 and 6. The comparison of the hydrogeological formations of watersheds 5, 6 and 25 is provided in Fig. 8. An advantage of SOMCD over conventional cluster analysis methods is that the relative topological relationships of input patterns can be easily identified from the locations of the corresponding images in the feature map. Similar input patterns are mapped onto the vicinity regions of the feature map; on the other hand, dissimilar input patterns are mapped onto different regions in the feature map. When images are distant from each other in the feature map, it implies that the corresponding input patterns are far from each other in the input space. In contrast, when images are close in the feature map, it implies that the corresponding input patterns are close to each other in the input space. However, the distance between two neurons on the feature map is not an absolute measure of the corresponding input patterns in the original input space. It should be noted that only the relative topological rela-

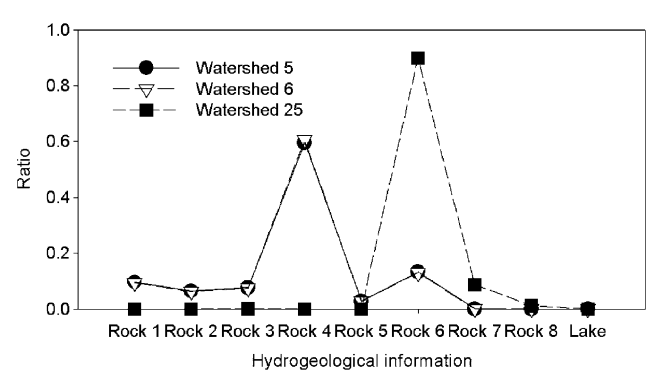

Fig. 8. Ratio of each hydrogeological formation's area to the whole watershed area.

tionships of input patterns are shown in the output space of SOM.

4.2. The number of clusters

The clustering result deduced from Fig. 4(a) is referred to Result I. Result II refers to the clustering result deduced from Fig. 4(b), and Result III corresponds to that from Fig. 5(a). As aforementioned, the Result II is identical to Result III.

The initial value of the radius during the tuning phase, $\sigma_t(0)$, plays a vital role in SOMCD. A significant crowding effect results from a larger value of $\sigma_t(0)$. On the contrary, the crowding effect is slight when a smaller value of $\sigma_t(0)$ is used. The crowding effect refers to the degree of the images of input patterns crowding in some regions of the output space. For example, the $\sigma_t(0)$ used to obtain Fig. 4(a) is less than that used to obtain Fig. 4(b). The distribution of the images of input patterns in Fig. 4(a) is more even than the distribution of the images of input patterns in Fig. 4(b). In other words, the images of the input patterns in Fig. 4(b) are more crowding in some regions of the output space so that the number of the blank neurons in Fig. 4(b) is more than that in Fig. 4(a). From Fig. 4(a) and (b), the image of watershed 7 and 25 moves to the up-right neuron from their original neuron, while $\sigma_t(0)$ alters from 1 to 2. The similar phenomenon can be found in Fig. 5(a) and (b). A large $\sigma_t(0)$ enlarges the ability of SOM distinguishing the clusters of input patterns, but reduce the capability for displaying details of the topological relationships of input patterns.

The dimension of SOM also influences SOMCD. For example, the dimension of SOM of Result I is smaller than that of Result III. The $\sigma_t(0)$ used in Result I is identical to Result III. The fraction of the blank neurons in Fig. 5(a) is more than that of Fig. 4(a). One interesting thing can be observed from Results I and III. Watershed 2, 20 and 21 are drawn out from Cluster **I**. They form Cluster C. If watershed 2, 20 and 21 is eliminated, Results I, II, and III are all the same. From the above discussions, it can be concluded that Cluster **I** is a nested cluster. Clusters A and C are small clusters within Cluster **I**. A SOM of a large dimension shows more details of the topological relationships of input patterns, but makes the identification of clusters boundaries (i.e., the determination of the proper number of clusters) more difficult.

From Fig. 5(b), one can find that there is a cluster that has only one member. The purpose of the cluster analysis in regionalization is to provide a clustered data set for developing a regional estimation method. The regional estimation method is a set of regression models. The confidence of a regression model highly depends upon the number of available data. Only one data point is not sufficient to construct a regression model, especially a nonlinear regression model. The clustering result deduced from Fig. 5(b) is not suitable for regionalization. Thus, the bold line of isolated plateaus in Fig. 5(b) is not drawn. The corresponding discrimination map is skipped. The discussions of the clustering results of Fig. 5(b) are also ignored in this paper. The confidence of a regression model may increase as the number of available data contained in the corresponding cluster increases. The accuracy of extrapolations may increase with increasing number of clusters. However, the number of clusters increases as the number of available data that are for developing regression models decreases. There is the trade-off between the confidence of regression models and the accuracy of extrapolations. The proper number of the hydrogeological factors can be 3 or 4 depending on the requirements and judgments of the analysts. The average specific low-flow duration curves deduced from Results I are depicted in Fig. 9. The average specific low-flow duration curves deduced from Results II and III are depicted in Fig. 10.

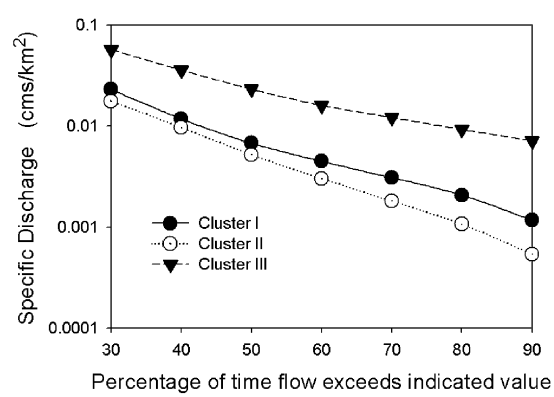

Fig. 9. The average specific low-flow duration curves of the three clusters.

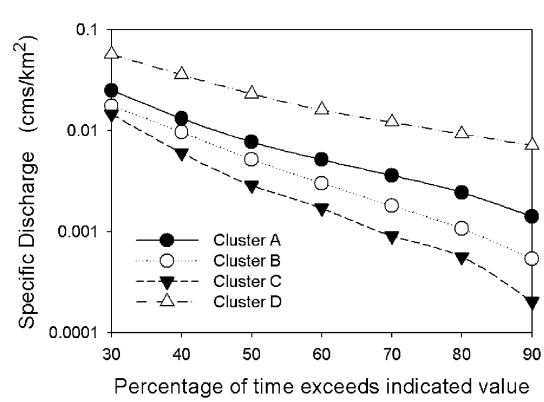

Fig. 10. The average specific low-flow duration curves of the four clusters.

4.3. Validations of the discrimination maps

Since the knowledge of input patterns is stored in the SOM itself, unlike convention cluster analysis, one does not need another method to perform discrimination analysis. The assignments of unknown patterns to known clusters are one kind of "generalization" of SOM. The capability of SOM for "generalization" is guaranteed [7]. The basis of the regionalization of the low-flow characteristics is that watersheds with similar hydrogeological factors should have similar low-flow characteristics, whereas watersheds with distinct hydrogeological factors should have different low-flow characteristics. The validations of the capability of discrimination of SOMCD are shown in Figs. 11 and 12. In Figs. 11 and 12, the upper and lower limits indicate one time of standard deviation of the data points in each cluster. In Fig. 6(a), the discrimination analysis is performed based on the Result I. Since watersheds 3, 16 and 17 are respectively assigned to clusters **I**, **II** and **III**, the specific low-flow duration curves of watersheds 3, 16 and 17 should be similar to the average ones of clusters **I**, **II** and **III**, respectively. The results shown in Fig. 11 conform to the inference. In Figs. 6(b) and 7, the discrimination analysis is performed based on Results II and III. Watersheds 3 and 16 are assigned to cluster A, and watershed 17 to cluster D. Hence, the specific low-flow duration curves of watersheds 3 and 16 should be similar to the average one of cluster A, and the curve of watershed 17 similar to the average one of cluster D. The results shown in Fig. 12 conform to the inference. The discrimination results of SOMCD are reasonable. The locations of streamflow gauges of the three clusters derived form Result I are shown in

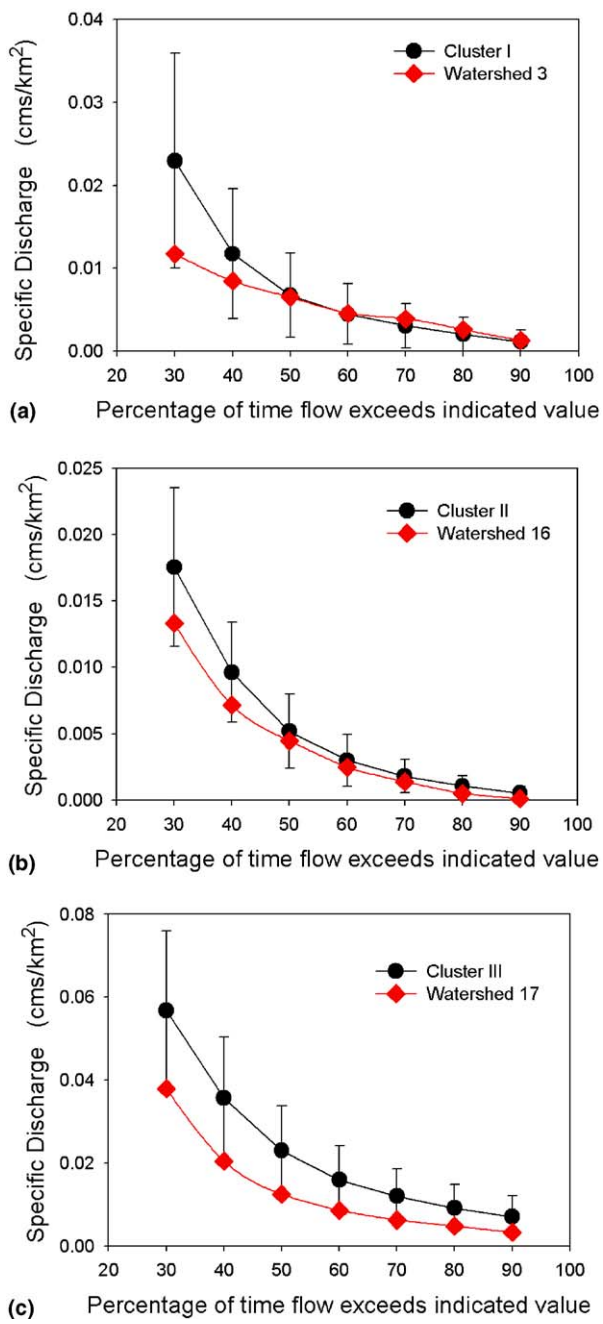

Fig. 11. Validations of the discrimination map (three-cluster case): (a) watershed 3, (b) watershed 16, and (c) watershed 17.

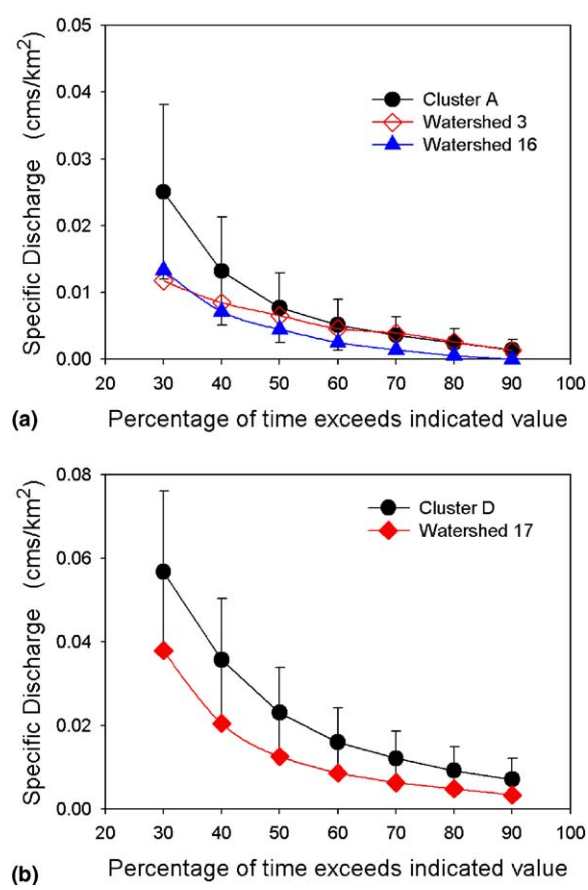

Fig. 12. Validations of the discrimination map (four-cluster case): (a) watersheds 3 and 16, and (b) watershed 17.

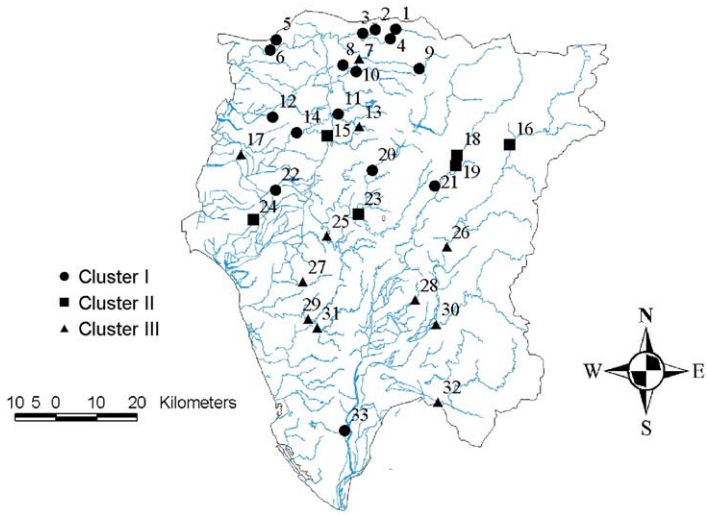

Fig. 13. Locations of streamflow gauges (three-cluster case).

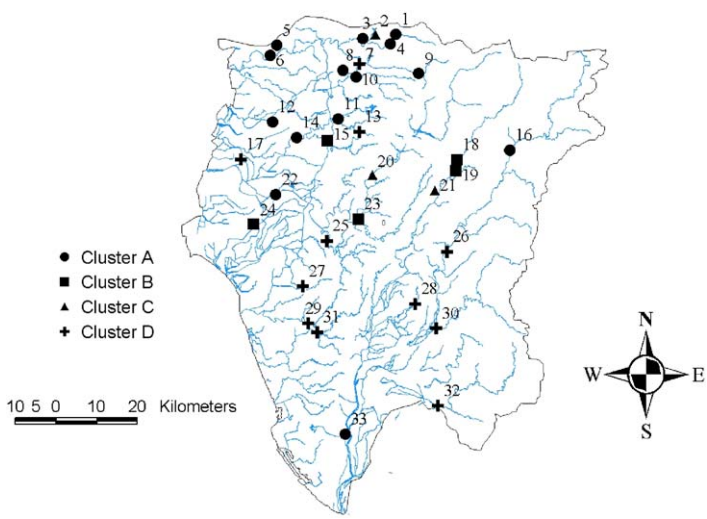

Fig. 14. Locations of streamflow gauges (four-cluster case).

Fig. 13. The locations of streamflow gauges of the four clusters derived form Results II and III are shown in Fig. 14.

4.4. Parameters settings

The other two parameters of SOMCD (training epochs and learning rate) influence the results of SOMCD, but the effects are not as significant as that of the $\sigma_t(0)$. However, the two parameters should be selected carefully. The number of the training epochs should be large enough so that SOM matures. SOM should be trained as many epochs as no significant changes of the result of SOMCD can be found. This is a strategy to determine the number of training epochs and is adopted in this paper. Another thing we found is that when the number of training epochs is sufficient, learning rate would not influence the results of SOMCD. However, learning rate during the ordering phase should be significantly larger than that during the tuning phase [7]. There is one more thing should be stated clearly. The initialization method of SOM in this paper is random initialization. Kohonen [7] indicated this is an inefficient method. Other more elaborate initialization methods may be considered. However, in our applications, the whole procedures of SOMCD performed on the data set in hand are completed in no more than 2 seconds. Methods that improve the efficiency of SOMCD are not considered in this paper.

Using SOMCD, analysts should first use a SOM of a small dimension and then use several different $\sigma_t(0)$ to train the SOM. $\sigma_t(0)$ should not be larger than half of the value of the initial radius during the ordering phase. Several results can be obtained by using different $\sigma_t(0)$. If

the results were reasonable and satisfied, then analysis procedures stop. Otherwise, a SOM of a larger dimension is chosen, and then above procedures are performed on the SOM again until a reasonable and satisfied result is obtained. Because the efficiency of SOM is rather high, the strategy of using SOMCD is simple and feasible.

4.5. Comparisons of SOMCD with conventional cluster analysis

There are various statistics in conventional cluster analysis that help hydrologists select the proper number of clusters [3]. However, they are independent from the methods for conventional clusters analysis, so that additional works are required to calculate the statistics. The numbers of clusters that are determined using these statistics are also likely inconsistent. Hydrologists who use conventional cluster analysis to determine the appropriate number of clusters should select the proper combination of the clustering method and the statistic among the various options. Therefore, we think that the determination of the proper number of clusters is a difficult task for conventional cluster analysis.

One advantage of SOMCD over conventional cluster analysis methods is that it does not need to determine the number of clusters in advance. Like hierarchical cluster analysis, hydrologists can inspect the input patterns in various views with SOMCD. The grouping of input patterns can be displayed by the "topography" of the density map and then the proper number of clusters can be easily determined by dividing the density map into several regions. That is SOMCD can display the boundaries between clusters explicitly on the maps, even when different resolutions of SOMCD are adopted. Consequently, the proper number of clusters can be easily determined using SOMCD without additional works. The assignments of unknown patterns to known clusters can be appropriately achieved by using the discrimination map. In conclusion, SOMCD combines cluster analysis and discrimination analysis in only one step. SOMCD is less complex than PCD. The comparison of SOMCD and PCD is provided in Table 2.

Table 2
Comparison of SOMCD and PCD

Item	SOMCD[a]	PCD[b]
Complexity	Low	High
Revealing the relative topological relationships of input patterns	Yes	*
Determining a proper number of clusters	Easy	Difficult
Assigning unknown patterns to known clusters	Yes	Yes
Performing cluster analysis and discrimination analysis in one step	Yes	No

[a] SOMCD: SOM-based Cluster and Discrimination analysis.
[b] PCD: Principal component analysis + Cluster analysis + Discrimination analysis.
* Depends on the cluster analysis method used.

5. Conclusions

A method (SOMCD) that can perform cluster analysis and discrimination analysis in one step is proposed in this paper. First the algorithms of SOMCD are developed and presented. A case study is performed using SOMCD to identify the homogeneity of hydrogeological factors affecting low-flow characteristics in southern Taiwan. The clustering results show that low-flow duration curves of the study area can be classified into three or four clusters. Analysts should choose the results according to the requirements of applications and their own judgments. The assignments of unknown watersheds to known clusters are also performed with the corresponding discrimination map. The discrimination results show that the assignments of ungauged watersheds to known clusters are reasonable. It is concluded that the proposed SOMCD is an efficient and effective method for identifying the homogeneity of hydrological factors and assigning unknown patterns to known clusters.

References

[1] Arabie P, Hubert LJ, Soete Gd. Clustering and classification. Singapore: World Scientific; 1996.
[2] Burn DH. Cluster analysis as applied to regional flood frequency. J Water Resour Plann Manage 1989;115(5):567–82.
[3] Everitt B. Cluster analysis. New York: Halsted Press; 1980.
[4] Gordon AD. Classification. Boca Raton: Chapman & Hall/CRC; 1999.
[5] Haykin S. Neural Networks: A comprehensive foundation. New Jersey: Prentice-Hall; 1999.
[6] Johnson RA, Wichern DW. Applied multivariate statistical analysis. New Jersey: Prentice-Hall; 2002.
[7] Kohonen T. Self-organizing maps. New York: Springer; 2001.
[8] Lin GF, Chen LH, Kao SC. Development of regional design hyetographs. Hydrol Process 2005;19(4):937–46.
[9] Mangiameli P, Chen SK, West D. A comparison of SOM neural network and hierarchical clustering methods. Eur J Oper Res 1996;93:402–17.
[10] Michaelides SC, Pattichis CS, Kleovoulou G. Classification of rainfall variability by using artificial neural networks. Int J Climatol 2001;21:1401–14.
[11] Mosley MP. Delimitation of New Zealand hydrological regions. J Hydrol 1981;49:173–92.
[12] Nathan RJ, McMahon TA. Identification of homogenous regions for the purpose of regionalization. J Hydrol 1990;121:217–38.
[13] Riggs HC. Regional analysis of streamflow characteristics. Techniques of Water Resources Investigations Book 4, Chapter B3. Washington, DC: USGS; 1973.
[14] Schreiber P, Demuth S. Regionalization of low flows in southwest Germany. Hydrol Sci 1997;42(6):845–58.
[15] Vogel RM, Kroll CN. Regional hydrogeologic–geomorphic relationship for the estimation of low-flow statistics. Water Resour Res 1992;28(9):2451–8.
[16] Yu PS, Yang TC, Liu CW. A regional model of low flow for southern Taiwan. Hydrol Process 2002;16:2017–34.

[17] WRPC. Hydrogeological map of Taiwan. Taipei: Water Resource Planning Commission (WRPC); 1986.
[18] Zhang J, Hall M. Regional flood frequency analysis for the Gan-Ming River basin in China. J Hydrol 2004;296(1-4):98–117.
[19] Zhang X, Li Y. Self-Organizing Map as a new method for clustering and data analysis. In: Proceedings of international joint conference on neural networks. Nogoya; 1993. p. 2448–51.