Assessment of some homogeneous methods for the regional analysis of suspended sediment yield in the south and southeast of the Caspian Sea

Hossein Kheirfam and Mehdi Vafakhah*

Department of Watershed Management Engineering, Faculty of Natural Resources, Tarbiat Modares University,
P.O. Box 46417-76489, Noor, Mazandaran Province, Iran.
**Corresponding author. e-mail: vafakhah@modares.ac.ir*

Regional analysis of suspended sediment yield (SSY) is commonly used to estimate sediment at a particular site where little or no information is available on sediment yield. In this research, the efficiency of three input selection and homogenization methods were evaluated in the estimation of SSY. Therefore, 42 sediment measurement stations and their upstream watersheds were selected and sediment rating curve was estimated by using regression models for each station. Mean annual SSY was estimated by using sediment rating curve and daily discharge. In the present study, in order to determine the independent variables in sediment yield, 11 physiographical, one climatic and two hydrologic variables of whole study watersheds were selected. Then the most effective independent variables were selected by using principal component analysis (PCA), Gamma test (GT) and stepwise regression (SR) techniques. After reducing 14 input variables to five (using PCA and GT) and two (using SR techniques), they are divided into homogeneous areas by Andrew curve (AC), cluster analysis (CA) and canonical discriminate function (CDFs) techniques. The watersheds were divided into two (using PCA-AC), three (using PCA-CA, PCA-CDFs and GT-CDFs), four (using GT-CA, GT-AC and SR-CA) and five (using SR-AC) homogenous regions. Multiple regression models to estimate mean annual SSY as a function of five (using PCA and GT) and two (using SR techniques) watershed characteristics were built in each homogeneous region, and compared to actual mean annual SSY in each station using relative error (RE), efficiency coefficients (CE) and relative root mean square error (RRMSE). The results showed that preprocessing the input variables by means of PCA and GT techniques has improved the homogeneous stations determination and the development models. According to the results, the best technique for determining homogeneous watersheds was AC technique with RE=49.24%, RRMSE=43.75% and CE=71.04%.

1. Introduction

The prediction of river sediment load is an important issue in hydraulic and sanitary engineering (Alp and Cigizoglu 2007). Sediment yield in a watershed occurs by some physical processes such as soil detachment, transportation and deposition. The concentration of sediment depends on rainfall intensity and volume, topography, land cover, management activities, soil texture and potential decomposition of soil fragments (Sadeghi and Singh 2005; Melesse *et al.* 2011). On the other hand, sediment transportation by river flow caused some problems such as reservoirs filling, soil erosion, streamflow direction change, river carrying reduction, hydrological facilities destruction,

Keywords. Suspended sediment; regional analysis; input selection techniques; watershed classification; Caspian Sea coastal watersheds.

water quality reduction and fish and wildlife habitats impairment (Cigizoglu 2004; Melesse *et al.* 2011). Therefore, monitoring of sediment load at the watershed outlets is helpful in reservoirs and facilities management (Kişi 2010), but establishing sediment-monitoring gaging stations in all watersheds is not economically feasible. For this reason, sediment modelling has been considered further. Estimation of suspended sediment concentrations (SSC) is often done with rating curves that establish an empirical relation between sediment concentration and discharge (Tramblay *et al.* 2010). Several studies have showed that rating curves' models tend to underestimate high SSC value (Asselman 2000) even when a bias correction method is applied (Walling and Webb 1988). Weak correlations between SSC and discharge have been reported in several studies. Tramblay *et al.* (2008) indicated that correlation between annual maximum SSC and discharge was significant in only 92 out of 208 stations in North America. Therefore, just discharge was not sufficient to estimate magnitude of extreme SSC in numerous rivers. On the other hand, several studies have showed that there is significant relationship between physical characteristics such as topography, geological, land use (vegetation) and climatic properties (precipitation volume and intensity) and suspended sediment load (Bray and Huixi 1993; Ludwig and Probst 1998; Restrepo *et al.* 2006) and mean SSC value (Jarvie *et al.* 2002; Siakeu *et al.* 2004; Syvitski and Milliman 2007; Nadal-Romero *et al.* 2011) or extreme SSC (Tramblay *et al.* 2007). Regional suspended sediment yield analysis involves the formation of a region that is required to affect the spatial transfer of information. In this context, a region means a collection of catchments, not necessarily geographically contiguous, which can be considered to be similar in terms of hydrological response. The goal of the regionalization process is the identification of groupings of catchments that are sufficiently similar to warrant the combination of suspended sediment yield information from sites within the region. The regions defined should thus be homogeneous with respect to suspended sediment yield characteristics. There are different methods for determining the similarity and homogeneity between watersheds such as Andrew curves (AC) (Nathan and McMahon 1990), cluster analysis (CA) (Zhang *et al.* 2011), canonical discriminate functions (CDFs) (Wilson 2002; Detenbeck *et al.* 2005; Carroll *et al.* 2009; Lin and Chen 2009) and other techniques. Several approaches for various purposes have been used in watersheds homogenization since 1981 (Hall and Minns 1999) up to now. In previous studies, the AC, CA and CDFs approaches were used separately and recommended as appropriate approaches. But still a comprehensive and perfect approach for desired purposes has not been approved by all researchers (Ilorme and Griffis 2011). Therefore, in this research, in order to compare the results of numerical and non-numerical homogenization approaches and also for more comprehensive conclusions, three homogeneous techniques, i.e., AC, CA and CDFs were used. Having large number of input variables is one of the main common problems for the modelling process. It is recommended to reduce the number of input variables even though this causes some information to be omitted (Lin and Wang 2006; Noori *et al.* 2011). There are different methods for reducing the number of input variables such as principal component analysis (PCA) (Ouarda *et al.* 2006; Zhang *et al.* 2006; Zhang 2007), Gamma test (GT) (Corcoran *et al.* 2003; Moghaddamnia *et al.* 2009), stepwise regression (SR) (Groupe de recherche en hydrologie statistique (GREHYS) 1996; Wang *et al.* 2006; Vafakhah 2007; Noori *et al.* 2010) and other techniques. In the research, three input selection techniques, i.e., PCA, GT and SR are used. Input selection and homogenization techniques have been successfully used by many researchers for various purposes such as prediction of stream flow, water quality parameters, organic material, sediment yield, evaporation, minimum flows and floods and also urban areas classifying (Nathan and McMahon 1990; Cavadias *et al.* 2001; Choi and Park 2001; Corcoran *et al.* 2003; Olden and Poff 2003; Caratti *et al.* 2004; Chen *et al.* 2004; Detenbeck *et al.* 2005; Eksioglu *et al.* 2005; Sohaili and Vafakhah 2005; Lin and Wang 2006; Owen *et al.* 2006; Ramachandra Rao and Srinivas 2006; Robertson *et al.* 2006; Wang *et al.* 2006; Khan *et al.* 2007; Vafakhah 2007; Yadav *et al.* 2007; Moghaddamnia *et al.* 2009; Noori *et al.* 2010; Tramblay *et al.* 2010; Noori *et al.* 2011). Caratti *et al.* (2004) have used climatic and morphometric data and classified sub-watersheds of Colombia river watershed based on environmental identifiers with CDFs and CA methods. They reported that the results of these two methods do not have significant differences and these techniques worked randomly for selecting effective parameters to classify sub-watersheds in homogeneous groups. Therefore, applying these methods must be carefully done. Lin and Wang (2006) used self-organizing map-based cluster and discrimination analysis (SOMCD) to identify the homogeneity of hydrogeological factors affecting low-flow characteristics in southern Taiwan. 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. Vafakhah (2007) used FA method to reduce the number of input variables and CA to homogenize the watersheds in south of Caspian Sea and classified the watersheds into two groups and also the three-part regression was determined as the best method for estimating suspended sediment yield. Tramblay *et al.* (2010) tested the feasibility of regionalizing extreme SSC for ungauged watersheds for 72 rivers in Canada and USA. Two approaches were compared, using either physiographic characteristics or seasonality of extreme SSC to delineate the regions. Multiple regression models to estimate SSC quantiles as a function of watershed characteristics were built in each region, and compared to a global model including all sites. Regional estimates of SSC quantiles were compared with the local values. The results showed that regional estimation of extreme SSC was more efficient than a global regression model including all sites. The most important variables for predicting extreme SSC are the percentage of clay in the soils, precipitation intensity and forest cover. To our knowledge, no studies have used these approaches (three-input selection and homogeneous techniques) for regional analysis of suspended sediment yield. The study aims to assess input variables (PCA, GT and SR techniques) and homogeneous watersheds (AC, CA and CDFs techniques) determination in SSY modelling.

2. Materials and methods

2.1 *Study area*

The watersheds in south and southeast of the Caspian Sea were selected as the study area due to the availability of independent variables such as maps, climatic and hydrological data and also the author's background knowledge of these watersheds. In addition, due to the fact that most of the upstream watersheds lack suspended sediment measuring stations, it is very important to estimate the sediment load entering into the Caspian Sea from the mentioned watersheds through the similar watersheds. Caspian Sea coastal watersheds which include Lahijan–Nour (code: 16), Haraz–Neka (code: 15 and 14) and Gorganroud (code: 13 and 12) rivers were selected for this research (figure 10). Lahijan–Nour and Haraz–Neka rivers are located in the longitude of 49°48′–54°41′E and latitude of 35°36′–37°19′N and comprise nearly 28463 km^2 and Gorganroud watershed located in the southeast of Caspian Sea coastal with longitude of 54°02′–56°16′E and latitude 36°34′–37°47′N comprises 13,170 km^2. The climate of these watersheds are semi-moderate arid to semi-humid by Amboreje method and warm Mediterranean by Gaussian method (Water Resources Research Center (WRRC) 1996).

2.2 *Selection of suitable stations*

Stations in south and southeast of the Caspian Sea with 20 years or more of SSC samples and daily discharge data were selected from Iranian Water Resources Research Organization. Stations on watersheds with major dams or reservoirs were excluded since sediment transport can be greatly affected by regulation in rivers. Thus 42 stations were selected in the analysis (table 1, figure 10).

2.3 *Mean annual suspended sediment yield*

A considerable part of sediment in rivers is transported during floods. SSC and water discharge were measured manually once a month and more samples were taken during floods for each sediment-gauging station by Iranian Water Resources Research Organization. Available years and the number of SSC samples were reported in table 1. Suspended sediment discharge was estimated by multiplying the water discharge by the SSC. Suspended sediment flux estimates were calculated using a relation of discharge to suspended sediment discharge known as a sediment rating curve (SRC) (Sadeghi and Mahdavi 2004). The most common SRC is a power function.

$$Q_s = aQ_w^b, \tag{1}$$

in which Q_w is discharge and Q_s is suspended sediment discharge. Values of a and b for a particular stream are determined from data via a linear regression between (log Q_s) and (log Q_w) (de Vente *et al.* 2007). The daily discharges subsequently were used to calculate the mean annual suspended sediment yield (SSY) in SRC for each station (Vanmaercke *et al.* 2011).

2.4 *Catchment characteristics*

In this study, 14 variables such as physiographic, climatic and hydrologic characteristics for all watersheds were obtained using Arc/GIS watershed. The topography maps from geographical organization in Iran at a scale of 1:50,000 were used for topography characteristics. In each station, 11 topography characteristics (size of area (km^2), perimeter of drainage area (m), weighed average height (m), main stream length (m), main stream slope (%), drainage density (%), compactness coefficient, watershed slope (%), average stream slope (%), watershed length (km), form factor and time of concentration obtained using Kirpich method) were extracted. Rainfall and peak discharge data were obtained from Iranian Water Resources Research Organization. Climatic (annual mean rainfall) and hydrologic (peak discharges with two-year return period and daily

Table 1. *List of stations used in the analysis.*

Station ID	Code	Name	Lat.	Long.	Available years	SSC sample no.	Area (km^2)
1	12-001	Tangrah	55°44′	37°27′	43	423	1672.43
2	12-005	Tamar	55°29′	37°28′	41	982	1541.12
3	12-007	Galikesh	55°27′	37°15′	42	469	404
4	12-009	Gholi Tappeh	55°25′	37°14′	26	112	81.47
5	12-011	Gonbad	55°08′	37°14′	43	3164	5438.93
6	12-013	Lazoureh	55°23′	37°13′	42	394	277.27
7	12-015	Pasposhteh	55°21′	37°10′	42	483	143.71
8	12-017	Novdeh	55°16′	37°03′	42	498	790.99
9	12-019	Arazkouseh	55°08′	37°13′	43	250	1562.34
10	12-021	Ramian	55°08′	37°01′	43	460	245.48
11	12-023	Ghazaghli	55°00′	37°13′	38	6150	7125.61
12	12-033	Taghiabad	54°38′	36°52′	15	158	109.08
13	12-035	Sadegorgan	54°36′	36°48′	5	52	65.47
14	12-043	Naharkhoran	54°28′	36°46′	22	260	97.07
15	12-045	Shastkalateh	54°20′	36°45′	5	48	91.72
16	12-049	Poleh Jaddeh	54°05′	36°47′	22	163	60.90
17	12-053	Vatana	53°57′	36°42′	16	147	21.5
18	12-071	Zaringol	54°54′	36°52′	41	521	342.82
19	12-083	Saramoo	54°49′	36°49′	33	346	390.41
20	12-085	Poleh ordougah	54°34′	36°46′	22	280	200.35
21	12-097	Siah Ab	54°03′	36°49′	37	318	1457.22
22	13-005	Sefidchah	53°54′	36°35′	39	1031	1043
23	13-006	Novzarabad	53°15′	36°49′	16	97	1992
24	13-013	Abaloo	53°19′	36°38′	40	1313	1962
25	13-019	Soleimantangeh	53°14′	36°15′	44	1699	92238
26	13-023	Varand	53°12′	36°21′	27	172	1194.75
27	13-025	Rigcheshmeh	53°10′	36°21′	44	540	2668.24
28	13-029	Kordkhil	53°07′	36°43′	26	275	4038.36
29	14-001	Shirgah	52°53′	36°17′	44	639	1825.18
30	14-005	Kasilian	52°53′	36°18′	44	535	346.2
31	14-007	Kiakola	52°48′	36°34′	44	510	2527.36
32	14-011	Qoraantalar	52°74′	36°17′	44	623	417.06
33	14-017	Koshtargah	52°39′	36°32′	27	618	1654.35
34	15-015	Razan	52°11′	36°11′	27	287	119.15
35	15-017	Karesang	52°22′	36°16′	44	2245	3967
36	16-003	Aghuzk Kati	52°03′	36°24′	27	796	112.26
37	16-009	Kheiroud Kenar	51°35′	36°38′	8	91	209
38	16-011	Novshahr	51°28′	36°40′	27	354	75.67
39	16-041	Haratbar	50°35′	36°45′	39	588	778
40	16-049	Gangsar	50°44′	36°50′	27	248	416.2
41	16-051	Ramsar	50°38′	36°55′	12	94	117.42
42	16-089	Dinasara	51°00′	36°39′	27	283	206

maximum discharge with two-year return period) characteristics of 42 selected stations were determined (Chow *et al.* 1988) (table 2).

3. Procedures

3.1 *Input selection techniques*

3.1.1 *Principal component analysis (PCA)*

Principal component analysis (PCA) can be used to identify the most relevant watershed characteristics when there are large volumes of information and it is intended to have a better interpretation of variables. In this method, the information of input variables will present with minimum losses in PCs (Noori *et al.* 2011). As the correlation between variables increases, the number of factors decreases (Abrahams 1972; White 1975; Çamdevýren *et al.* 2005). Since the variables have their main effect in equations, data should be standardized (Liu *et al.* 2003):

$$Z = \frac{X - \bar{X}}{\mathrm{SD}}. \qquad (2)$$

In this equation, $\bar{X}$ is mean, SD represents standard deviation of each series of data (X) and Z is

Table 2. *Watershed characteristics used in the study.*

No.	Symbol	Definition	Unit	Mean	Minimum	Maximum
1	A	Area	km^2	3333.98	21.5	92238
2	$Q_{\max}$	Daily maximum discharge	m^3 s^{-1}	33.46	1.2	177.79
3	QP	Peak discharge	m^3 s^{-1}	46.28	262.71	1.62
4	R	Mean annual rainfall	mm	636.99	446.36	1018.24
5	CC	Compactness coefficient	–	33.16	5.23	95.25
6	Tc	Time of concentration	hr	363.69	29.7	1313.98
7	P	Perimeter	km	151.86	16.2	449.65
8	H	Weighed average height	m	1445.93	666.9	2777
9	Sw	Watershed slope	Percent	22.81	9.3	45.8
10	Dd	Drainage density	km km^{-2}	0.28	0.1	1.13
11	Sr	Average stream slope	Percent	6.21	0.1	16.97
12	Sl	Main stream length	m	57586.9	5500	165000
13	Lw	Watershed length	km	45.85	6.56	140
14	FF	Form factor	–	1.49	0.22	2.82

standardized data. When a PCA is performed, the correlation between two or more variables is summarized in a scatter-plot which is called the factor load. A regression line with the maximum variance is then fit to represent a linear relationship between the variables. After this first factor has been identified, additional lines are drawn to maximize the remaining variability in a consecutive step-by-step extraction of factors. Variance maximizing rotation was used in the process of extracting each consecutive factor. The Kaiser criterion is the most widely used method to evaluate the maximum number of factors (i.e., linear combinations) to extract from the dataset (Kaiser 1960). This criterion requires that factors are retained only if their associated eigenvalues are greater than one. The variables with the highest factor loadings within each separate factor will likely share a common characteristic or combination of characteristics. These factor loadings are the correlation coefficients between the variables. The efficiency of PCA method was validated using KMO (Kaiser–Meyer–Olkin) and Bartlett tests. KMO should be one in the ideal case. High KMO values indicate a PCA with few errors, overall. If KMO is more than 0.5, PCA could be used (Shrestha and Kazama 2007).

3.1.2 *Gamma test*

Gamma test (GT) estimates the minimum mean square error (MSE) that can be achieved when modelling the unseen data using any continuous non-linear models (Moghaddamnia *et al.* 2009). GT was first reported by Koncar (1997) and Adoalbjörn *et al.* (1997) and later enhanced and discussed in detail by many researchers (Durrant 2001; Tsui *et al.* 2002). The interested readers should refer to the aforementioned papers for further details on GT. The GT can be achieved through winGammaTM software implementation (Durrant 2001).

3.1.3 *Stepwise regression*

Stepwise regression (SR) is a general statistical method to select variables (Faraway 2002). When the number of candidate covariates (N) is small, one can choose a prediction model by computing a reasonable criterion (e.g., root mean square error (RMSE), sum of square error (SSE) or cross-validation error) for all possible subsets of the predictors. However, as N increases, the computational burden of this approach increases very quickly. This is one of the main reasons why step-by-step algorithms like SR are popular. In this approach, which is based on linear regression model, first step is ordering of the explanatory variables according to their correlation with the dependent variable (from the most to the least correlated variable). Then, the explanatory variable, which is best correlated with the dependent variable, is selected as the first input. All remaining variables are then added one by one as the second input according to their correlation with the output and the variable which most significantly increases the correlation coefficient (R^2), is selected as the second input. This step is repeated N–1 times for evaluating the effect of each variable on model output. Finally, among N obtained subsets, the subset with optimum R^2 is selected as the model input subset. The optimum R^2 is integral to a set of variables after which adding a new variable does not significantly increase the R^2. SR is a combination of forward and backward methods.

3.2 *Homogeneous areas determination techniques*

3.2.1 *Andrew curves*

Andrew curve (AC) was developed by Andrews (1972) for detection of human fossils from monkey fossils. AC is a graphical and non-numeric method to illustrate continuous multivariable data (equation 3). Also, it is a useful method for detecting hidden structures in a fairly small dataset. These structures include clusters, outliers, correlated record and so on (Moustafa 2011). AC displays each three-dimensional point in two or three dimensions using Fourier interpolation functions (Horhota and Aitken 1991). In equation (3), x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, ... show variables used for homogeneous watershed determination. These variables have the most correlation with dependent variable (Moustafa 2011). The Andrew's plot (Fourier curve) is plot of (t, y_{it}) in the range of $t\varepsilon[-\pi, \pi]$, where y_{it} is given by

$$\begin{aligned} y_{it} &= \frac{X_i 2}{\sqrt{2}} + x_{i2}\cos(\lambda_{1t}) + x_{i3}\sin(\lambda_{1t}) \\ &\quad + x_{i4}\cos(\lambda_{2t}) + x_{i5}\sin(\lambda_{2t}) + \cdots \end{aligned} \tag{3}$$

where $\lambda = i, i = 1, 2, 3, ...$

3.2.2 *Cluster analysis*

Cluster analysis (CA) was first reported by Tryon (1939). CA is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters (Jobson 1992; Lin and Wang 2006). Many algorithms have been proposed for CA divided into two main groups including hierarchical and repeated discriminates. Most researchers use hierarchical

Table 3. *The pros and cons methods used in this study.*

Techniques	Pros	Cons
Input selection techniques		
SR	• Widely used • Appealing and easy to compute/use	• Less possibility to achieve the optimal model because of drop and add variables one at a time • Use of degrees of freedom This method may overstate the significance of results
PCA	• Suitable for large datasets • Applicable for all datasets • Variables are ordered in terms of standard error • Can be used to standardize data • Reduces dimensionality of data: easier to learn	• Non linear structure is hard to model with PCA Black box: data vectors become hard to interpret
GT	• A reliable analysis method for non-linear time series • The noise estimate provides a stopping criteria for training • Can determine the minimum number of data points required to build a good model	• The efficiency of this method is dependent on the sample size (unsuitable for small size samples) Less user friendly software packages
Homogenization techniques		
CDFs	• Simple, easy to understand • Require little data preparation nominal or categorical data	• Not a unique solution • Univariate, ignores redundancy among variables • Statistical inference difficult
CA	• Simple, clear and popular • Decently efficient • Guaranteed convergence • Can accommodate any shape (with enough clusters)	• Initialization and local optimal • Need to define the number of clusters • Convex clusters of roughly the same size and densities • Tends to create unbalanced cells • Highly sensitive to noise and outliers • Not have a generally accepted method to choose cut off point
AC	• Allows to draw virtually two and three dimensions • Based on Fourier series where the coefficients are the observation's values • Based on the Parseval's identity (energy norm)	• Hard to interpret • Separation of homogeneous samples based on expert opinion • Inability to represent high dimensions, and the cluttering effect with large datasets

method as a suitable algorithm (Ramos 2001). In hierarchical method, data grouping is done by two methods – cumulative and distribution. In cumulative method, at first each data makes a single group. Then similar groups gradually combine to an individual group. In distribution method, grouping is in inverse way of cumulative method. Output of hierarchical method is in dendrogamatic form and the relation between each class of data is displayed by their similarity (Tryon 1939). A number of cluster analyses were carried out, combining different distance measures and linkage methods for a range of number of clusters. In each case, the regression model was developed to assess performance of the model (Tramblay *et al.* 2010). The Ward linkage algorithm was chosen because it tends to form spherical clusters of equal size and gives the best results for identification of homogeneous regions in several regional flood frequency studies (Nathan and McMahon 1990; Ouarda *et al.* 2006).

3.2.3 *Canonical discriminate functions*

Canonical discriminate functions (CDFs) are defined as linear combinations that separate groups of observations, and canonical variables are defined as linear combinations associated with canonical correlations between two sets of variables. Objects, which have similar variances in the analyzed parameters, will have similar discriminant scores. After the plotting process they will group together. Also, relationships between variables can be easily identified by the respective coefficients. Strongly correlated variables will generally have the same magnitude and orientation when plotted, whilst uncorrelated variables are typically orthogonal to each other (Carroll *et al.* 2009). The overall summary of the pros and cons of the techniques used in this study is presented in table 3.

3.3 *Performance evaluation*

In each homogeneous region, a relation between the mean annual SSY and the watershed characteristics was established using multiple linear regressions. 70% stations were used randomly for modelling and the remaining 30% stations were used for testing. Estimation of reliability for different approaches was performed using a Jack–Knife re-sampling procedure to calculate error statistics. In each region, every site was in turn considered ungauged and removed from the database. The remaining sites were then used to build a regression model to estimate the mean annual SSY at the station that had been removed. Then, by using the difference between the local mean annual SSY and the Jack–Knife estimation, it is possible to compute relative error (RE), efficiency coefficients (CE) and relative root mean squared error (RRMSE) for each site. The three performance evaluation criteria used in this study can be calculated utilizing the following equations.

$$\mathrm{RE} = \left|\frac{Q_O - Q_E}{Q_O}\right| \times 100 \tag{4}$$

$$\mathrm{CE} = \frac{1/n\sum_{i=1}^{n}(Q_O - \overline{Q}_O)^2 - \frac{1}{n}\sum_{i=1}^{n}(Q_O - Q_E)^2}{1/n\sum_{i=1}^{n}(Q_O - \overline{Q}_O)^2} \times 100 \tag{5}$$

Table 4. *KMO and Bartlett's test.*

KMO	0.77
Bartlett's test	520.48
Significance level	0.00

Table 5. *Correlation matrix of the factors selected by PCA.*

	Initial eigen values		
Factor	Total	% of variance	Cumulative %
1	**6.40**	**45.74**	**45.74**
2	**1.68**	**12.02**	**57.76**
3	**1.38**	**9.89**	**67.65**
4	**1.07**	**7.66**	**73.31**
5	**1.02**	**7.32**	**82.63**
6	0.856	6.112	88.751
7	0.752	5.374	94.125
8	0.312	2.225	96.350
9	0.162	1.159	97.509
10	0.118	0.844	98.353
11	0.105	0.748	99.102
12	0.057	0.407	99.509
13	0.041	0.293	99.802
14	0.028	0.198	100.000

More effective factors are marked as bold.

Table 6. *Component loading matrix.*

	Component				
Variable	1	2	3	4	5
$Q_{\max}$	0.44	0.85*	−0.05	0.01	0.001
Qp	0.41	0.81*	−0.10	0.09	0.02
R	−0.12	−0.002	0.06	0.13	0.89*
CC	0.85*	0.38	0.03	−0.10	−0.05
Tc	0.84*	0.19	−0.13	−0.12	−0.05
A	0.10	−0.06	0.032	−0.82*	−0.05
P	0.93*	0.16	−0.07	0.03*	−0.004
H	0.25	−0.10	0.88*	−0.13	0.04
Sw	−0.45	0.04	0.81*	0.09	−0.06
Dd	−0.33	−0.04	0.31	0.30	−0.48
Sr	−0.69	−0.18	0.31	0.30	−0.16
Sl	0.95*	0.17	−0.01	−0.06	0.005
Lw	0.93*	0.14	0.07	−0.01	0.04
FF	0.53	−0.49	−0.22	0.40	−0.007

*Values greater than the acceptable threshold (>0.7).

$$\mathrm{RRMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(\frac{Q_O - Q_E}{Q_O}\right)^2}{n}} \times 100 \qquad (6)$$

where Q_O and Q_E are respectively, the observed and the predicted mean annual SSY, $\overline{Q}_O$ is the average observed mean annual SSY and n is the number of observed data.

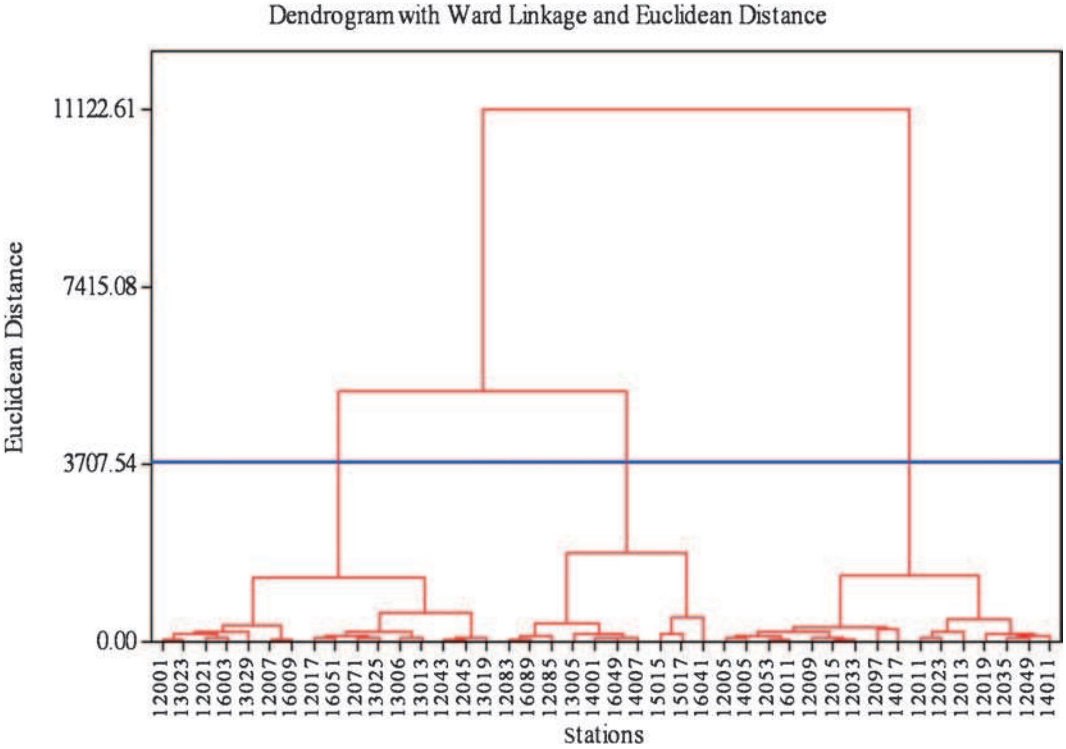

Figure 1. Dendrogram of the hierarchical clustering on watershed characteristics using PCA.

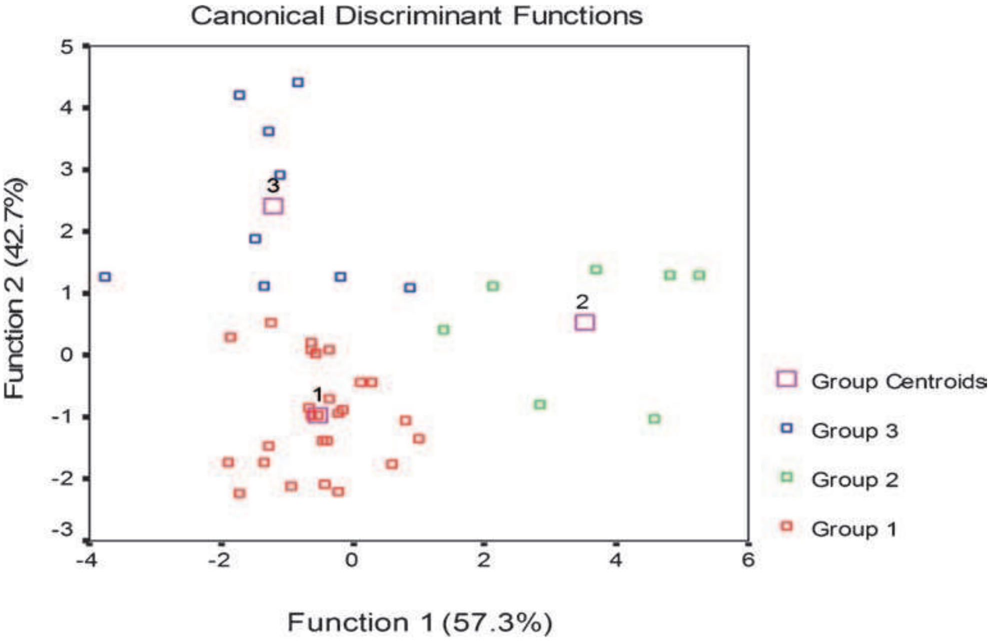

Figure 2. A plot of the first and second discriminante functions on watershed characteristics using PCA. Variables closely related to the first two functions were main stream length (Sl), daily maximum discharge with 2-year return period ($Q_{\max}$), weighed average height (H), area (A) and mean annual rainfall (R).

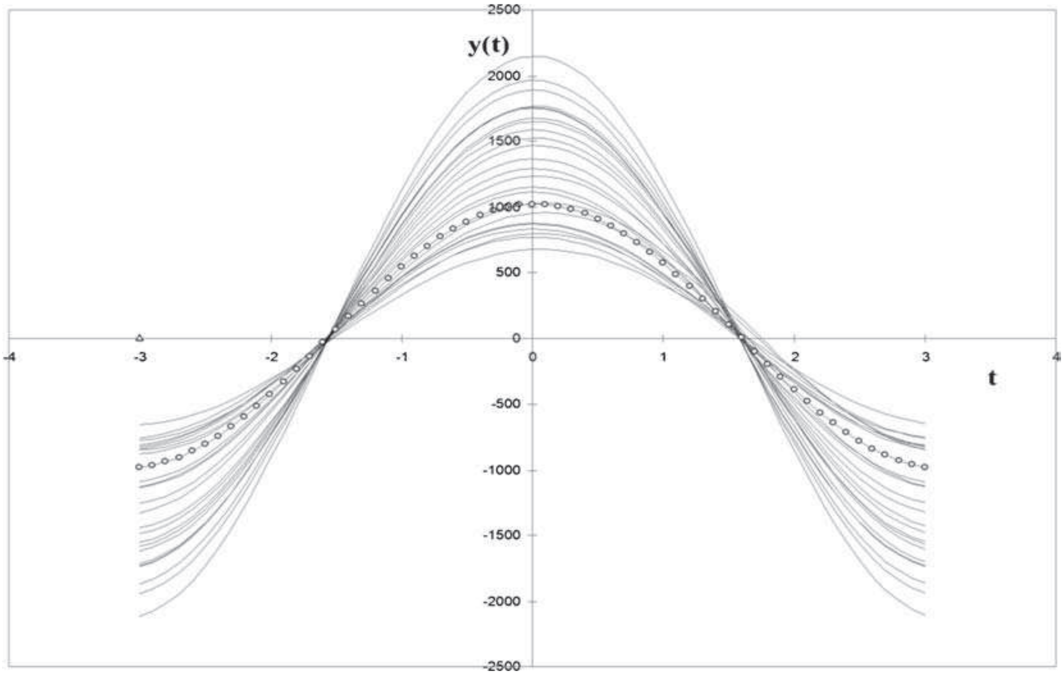

Figure 3. Andrews curves on catchment characteristics using PCA.

4. Results and discussion

4.1 *Principal component analysis*

The PCA for selecting the best set of catchment characteristics are listed in tables 4, 5 and 6. According to table 4, the Kaiser–Meyer–Olkin (KMO) value and significance level of Bartlett's test were 0.775 and 0.00, respectively, that confirmed application of PCA for input variables (Shrestha and Kazama 2007). For PCA application, after standardization of input variables, correlation symmetrical matrix R was formed with dimensions 14×14 (equivalent to the number of input variables). The results of table 5 shows that among the 14 variables, first five variables which had eigenvalues more than one (Shrestha and Kazama 2007) were selected as independent variables and 82.639% of variability is explained. Also, the amount of eigenvalues and variance values of the other variables are shown in table 5. As per table 6, the main stream length (Sl) variable was selected as the most appropriate independent variable in component 1, which can explain others. Other variables, such as the daily maximum discharge with two-year period ($Q_{\max}$), the weighed average height (H), the area (A) and the mean annual rainfall (R) were selected as independent variables in components 1, 2, 3, 4 and 5, respectively.

After selecting relevant variables using PCA, the CA, CDF and AC were used for determining homogeneous watersheds. Figures 1, 2 and 3 show the results of CA, CDFs and AC grouping.

Table 7. *The RE, RRMSE and CE criteria of models in testing periods.*

Allocation of ungauged site via	CA	CDFs	AC
RE%	69.27	36.05	58.19
RRMSE%	56.83	44.09	46.85
CE%	46.47	45.83	64.39

Table 8. *The Gamma test results for the 14 catchment characteristics.*

Input variables	Gamma value
all	0.13166
all-Qp	0.61746
all-FF	0.13562
all-Sr	0.13546
all-Lw	0.13328
all-CC	0.13271
all-H	0.13265
all-Tc	0.13254
all-Dd	0.13174
all-P	0.13127
all-R	0.13127
all-A	0.13115
all-Sl	0.13097
all-$Q_{\max}$	0.12456
all-Sw	0.12085

As per figure 1, the watersheds were divided into three homogeneous groups. It is obvious from figure 2 that the watersheds were divided into three homogenous groups.

Based on figure 3, the watersheds were divided into two homogenous groups using AC. For each homogeneous group, a regression model between catchment characteristics and the mean annual SSY was fitted independently. The RE, RRMSE and CE criteria of models in testing periods are given in table 7.

As per table 7, the AC was selected as the best method for region homogenization, because it had the lowest value of RE and RRMSE (with

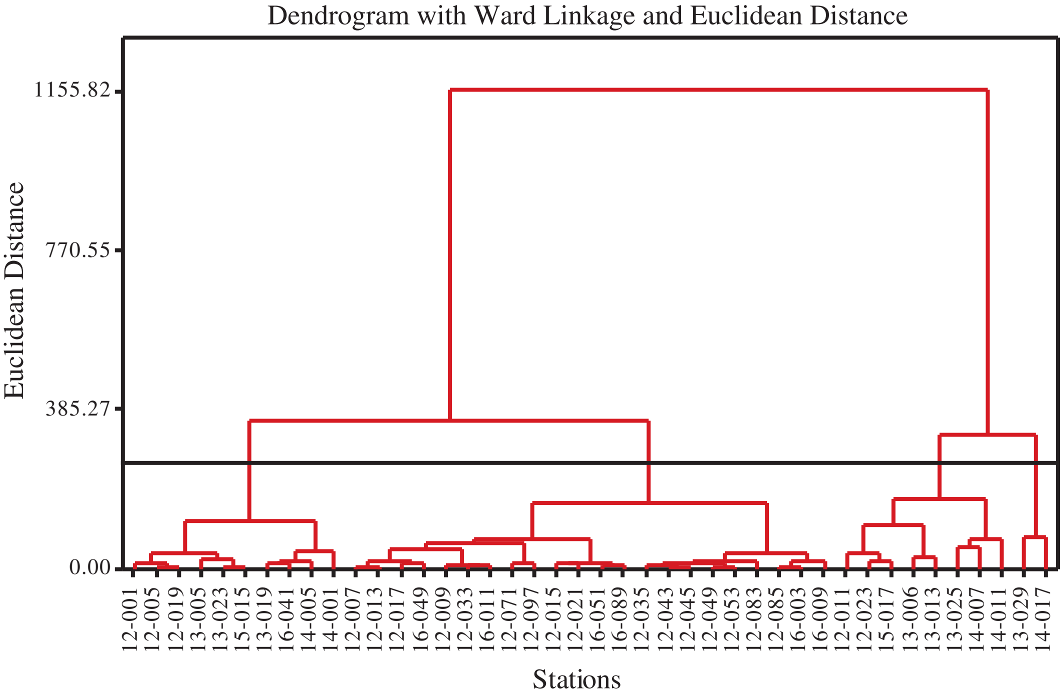

Figure 4. Dendrogram of the hierarchical clustering on watershed characteristics using GT.

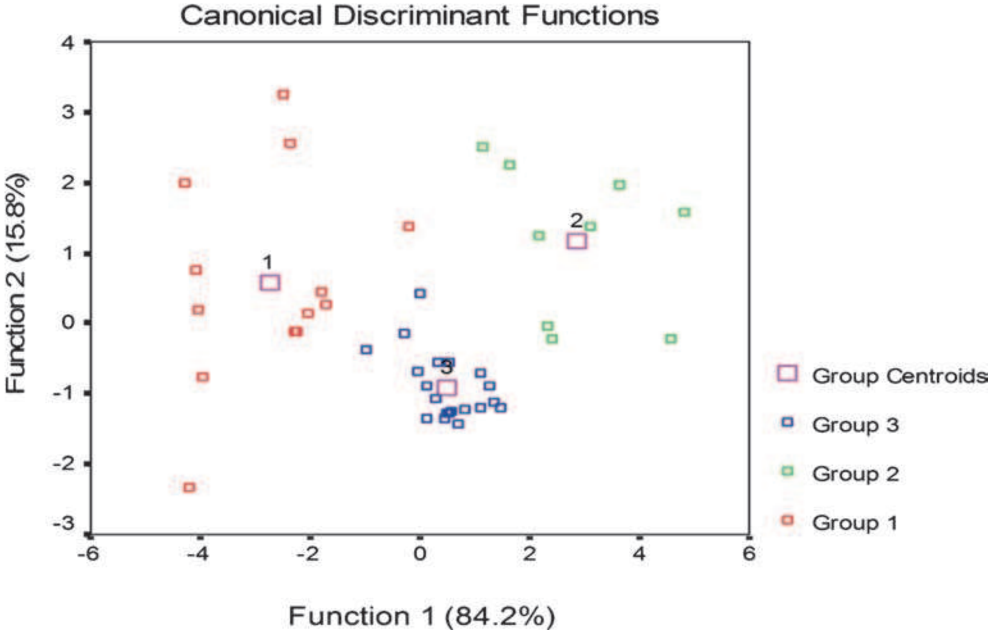

Figure 5. A plot of the first and second discriminante functions on watershed characteristics using GT. Variables closely related to the first two functions were peak discharge with two-year return period (Qp), form factor (FF), average stream slope (Sr), watershed length (Lw) and compactness coefficient (CC).

58.19 and 46.85%, respectively) and highest value of CE (with 64.39%). This result was different from Zhang *et al.* (2011), as they suggest PCA integrated with CA and GIS methods as the most powerful approach for large-scale regionalization. This disagreement may be due to small watershed areas in this study. In figure 3, the watersheds with different area and mean annual SSY were located in homogeneous groups. This result was in compliance with results of Tramblay *et al.* (2010).

4.2 *Gamma test*

In the present study for determining the more important variables, as a first step, the Gamma value was calculated for a combination of all variables (14 input variables). Next, one of the variables was omitted and Gamma value was calculated for a combination of the other variables (13). Then the omitted variable in the previous step was returned and another variables was omitted from the original combination (14) and Gamma value was calculated for the new combination that contained 13 variables. This process continued for all variables, one by one, and Gamma value was computed in each step. In this method, omitting important variables result in increasing Gamma value in comparison with the Gamma value of original combination (Noori *et al.* 2010). The results for different combinations were shown in table 8. This table indicates that peak discharge with two-year return period (Q_{p}) was the most important variable because it had the highest Gamma value. The other principal variables were the form factor (FF), the average stream slope (Sr), the watershed length (Lw) and the compactness coefficient (CC), respectively. Therefore, Qp, FF, Sr, Lw and CC were selected as the optimum input variables in homogeneous watershed determination method in order to predict mean annual SSY.

Table 9. *The RE, RRMSE and CE statistics of models in testing period.*

Allocation of ungauged site via	CA	CDFs	AC
RE%	57.17	41.08	45.91
RRMSE%	48.15	41.08	55.27
CE%	57.75	33.47	75.53

Table 10. *Coefficients of determination (R^2) of variables influence on SSY.*

Input variables	R^2
Qp	0.679
$Q_{\max}$	0.617
H	0.169
Sw	0.146
CC	0.114
Sr	0.08
P	0.077
Sl	0.071
FF	0.053
R	0.034
Lw	0.031
A	0.03
Dd	0.011
Tc	0.008

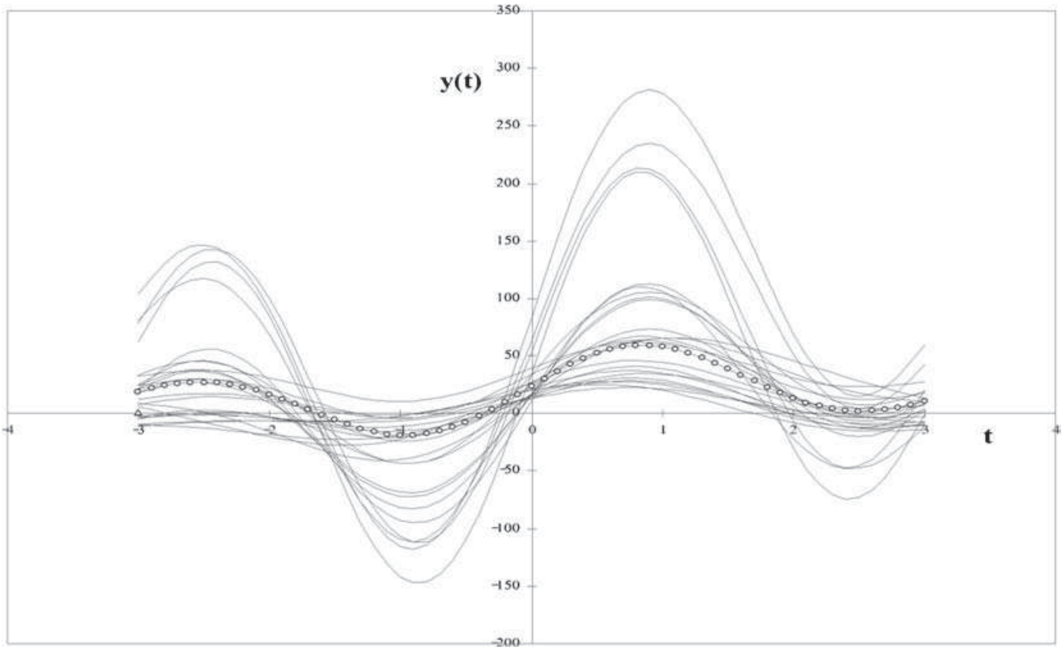

Figure 6. Andrews curves on catchment characteristics using GT.

After reducing the variables by GT, homogeneous watersheds were determined using CA, CDF and AC based on the selected variables. The watersheds were divided into four homogeneous groups by using CA (figure 4). Figure 5 indicates the results of CDFs. Figure 5 shows that the watersheds were divided into three homogeneous groups.

AC was used for determining homogeneous watersheds. In this method, the watersheds were divided into four homogenous groups (figure 6). For each homogeneous group, a regression model between catchment characteristics and the mean annual SSY was fitted independently. The RE, RRMSE and CE criteria of models in testing period are

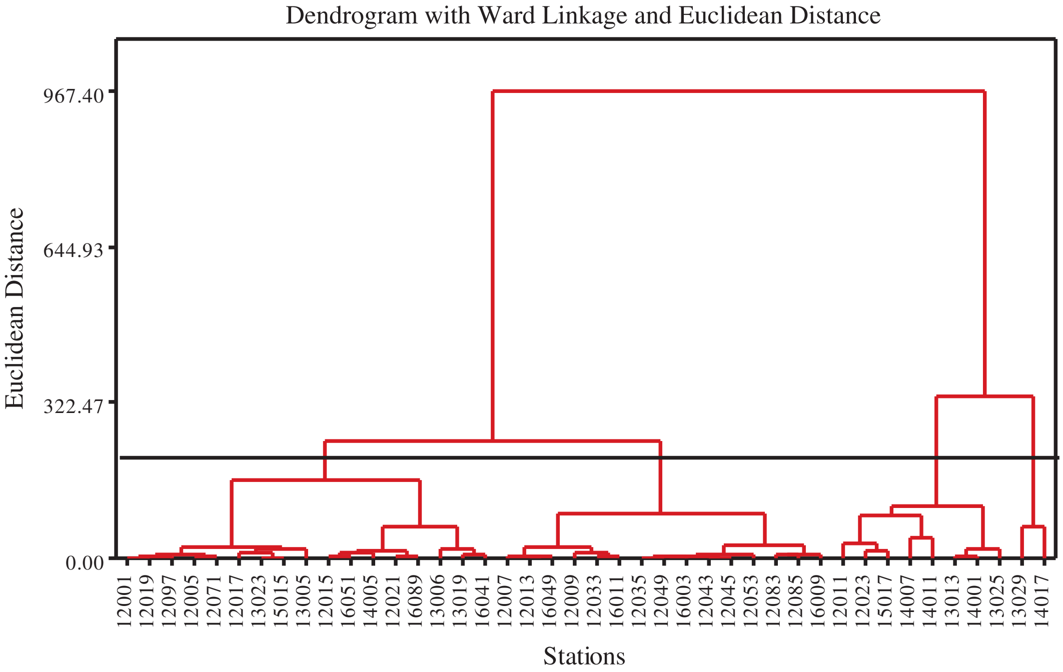

Figure 7. Dendrogram of the hierarchical clustering on watershed characteristics using SR.

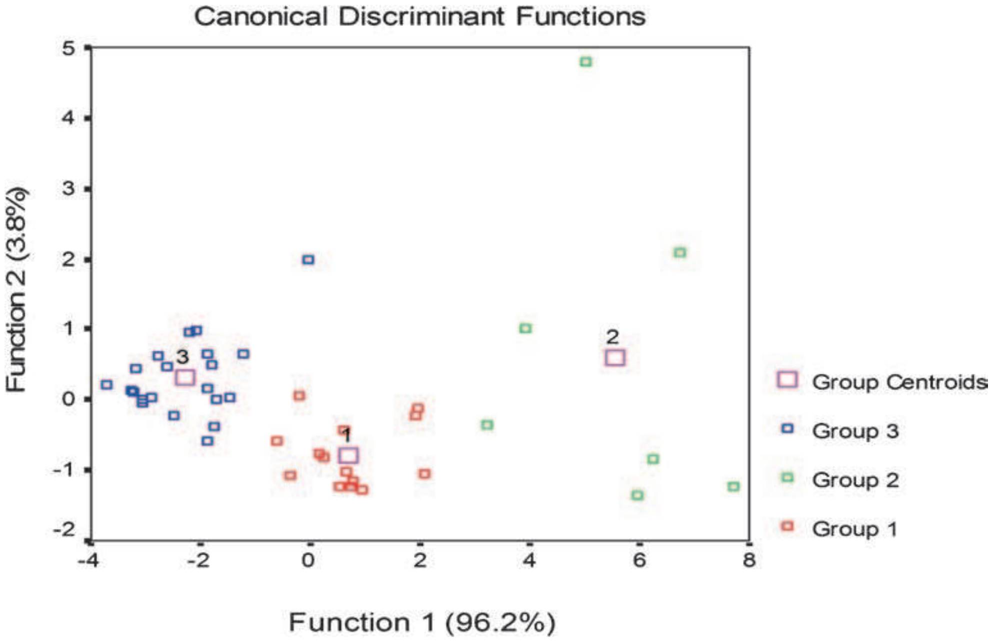

Figure 8. A plot of the first and second discriminante functions on watershed characteristics using SR. Variables closely related to the first two functions were daily maximum discharge with two-year return period ($Q_{\max}$) and compactness coefficient (CC).

given in table 9. The best estimation results were obtained when AC was considered for homogeneous watersheds because CE value was high (with 75.53%), while RE and RRMSE values were low (with 45.91 and 55.27%, respectively).

4.3 *Stepwise regression*

In the present study, the SR method was used as a linear input selection technique to select the best subset of 14 input candidates. Different variables have different influence on amount of sediment yield in watersheds. On the other hand, it can be possible to predict the sediment amount as dependent variable with high accuracy using effective variables reduction. SR is one of the regression modelling methods and in this method all independent variables are used as model inputs one by one and this operation continues until the error reaches to the desired level of significance (Faraway 2002). In this study, stepwise regression was used to determine the relation between variables and their importance, and finally peak discharges with two-year return period (Qp) and compactness coefficient (CC) were selected as the most effective factors. Also, table 10 shows the effect of each variable on SSY based on coefficients of determination (R). After reducing the variables by SR, homogeneous watersheds were determined using CA, CDFs and AC based on the selected variables (figures 7, 8 and 9). Mentioned figures show that the watersheds were divided into four, three and five homogeneous groups by using CA, CDFs and AC methods, respectively.

Similar to previous section, multiple regressions were used to establish a relation between watershed characteristics and the mean annual SSY for each homogeneous group. Table 11 shows RE, RRMSE and CE criteria values of models in testing period. It is obvious from table 11 that the best estimation results were obtained when AC was considered for homogeneous watersheds because CE value was high (with 73.2%), while RE and RRMSE values were low (with 43.64 and 55.42%, respectively). On the other hand, the worst estimation results were obtained when CA was considered for homogeneous watersheds. It should be noted that none of 14 variables were directly omitted using PCA

Table 11. *The RE, RRMSE and CE statistics of models in testing period.*

	CA	CDFs	AC
RE%	86.16	52.38	43.64
RRMSE%	81.21	55.65	55.42
CE%	23.54	58.41	73.2

Table 12. *Average RE, RRMSE and CE for three homogenization methods.*

	CA	CDFs	AC
RE%	70.87	43.89	49.24
RRMSE%	62.06	48.69	43.75
CE%	42.58	56.30	71.04

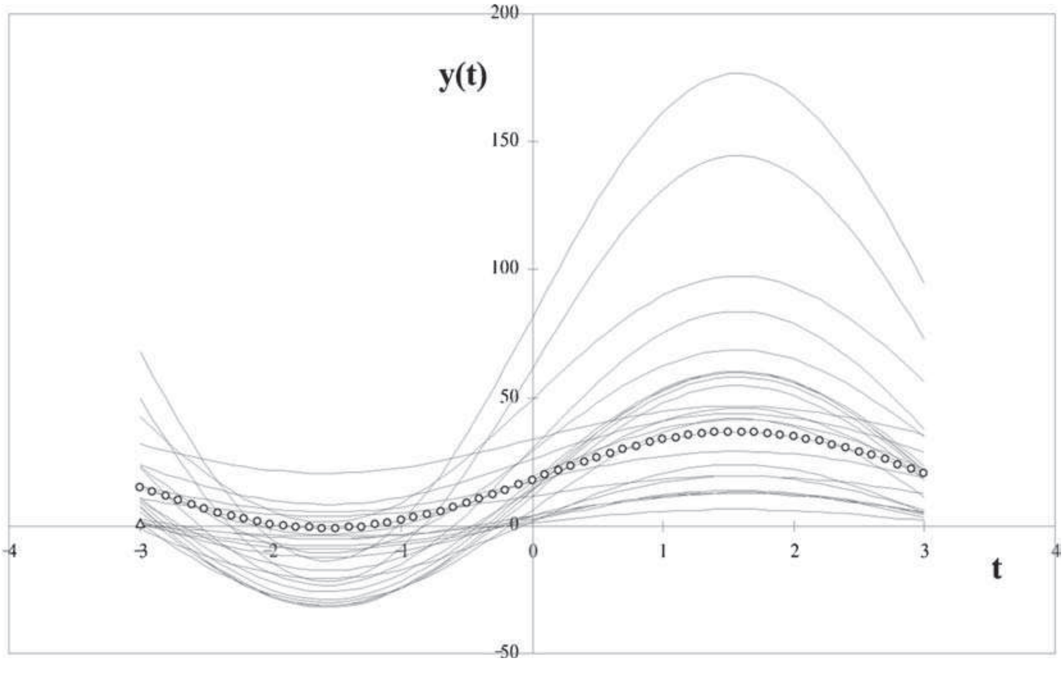

Figure 9. Andrews curves on catchment characteristics using SR.

method (in contrast with GT and SR methods). Therefore, the effects of all variables were considered in PCA method, while the number of variables was reduced from 14 to 2 and 5 in SR and GT methods, respectively. Finally, according to table 11, the CA has the worst performance due to the low selected variables using SR (Yadav *et al.* 2007). In addition, the input variables reduced

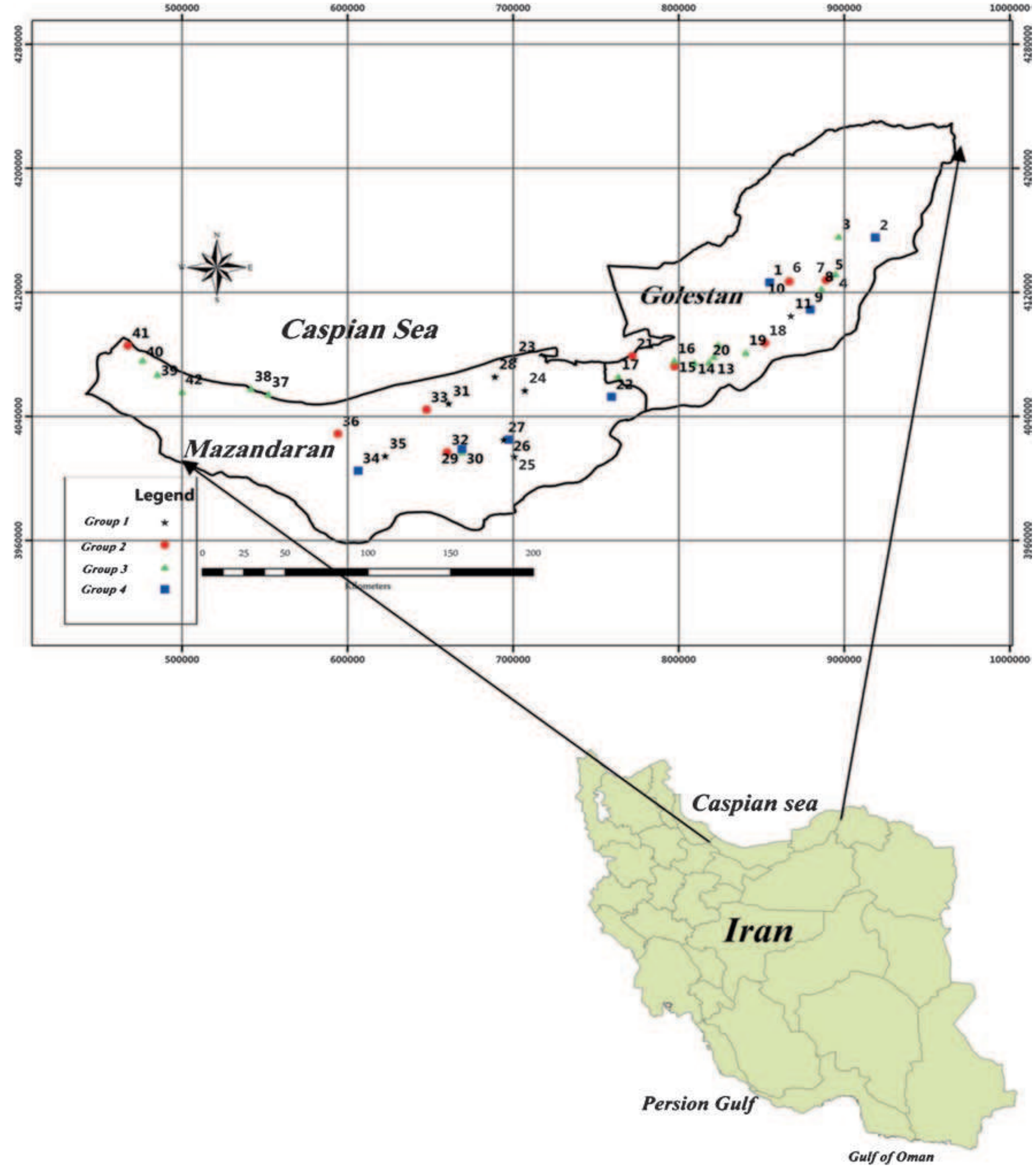

Figure 10. Location of study stations and homogeneous regions using combination of GT and AC.

Table 13. *The obtained models for each homogenized groups using combination of GT and AC.*

Homogeneous group	Model
1	–
2	$Q_s = 27.235\text{Qp} + 3.203\text{FF} + 235.336\text{Sr} + 8\text{Lw} - 177.923\text{CC} - 1800$
3	$Q_s = -20.944\text{Qp} + 8.014\text{FF} + 32.237\text{Sr} + 42.195\text{Lw} + 10.279\text{CC} + 11.086$
4	$Q_s = 85.913\text{Qp} + 91.867\text{FF} + 364.556\text{Sr} + 52.205\text{Lw} + 1.302\text{CC} - 8600.483$

from 14 to 5 in the PCA and GT methods. Therefore, PCA and GT methods had a positive effect compared to the regression methods. These results were consistent with Noori *et al.* (2011) results.

To determine the best homogenization method, the average three performance evaluation criteria (RE, RRMES and CE) were computed (table 12) using the results of tables 6, 8 and 10. The obtained results indicate that AC was the best method with RE=49.24, RRMSE=43.75 and CE=71.04 while CA was the worst technique with RE=70.87, RRMSE=62.06 and CE=42.58. These results were consistent with Ramachandra Rao and Srinivas (2006) results.

Location of the homogenized watersheds were determined by using the best method of reducing the number of variables (GT) and homogenization method (AC) as shown in figure 10. The obtained models for each homogenized groups were shown in table 13.

5. Conclusions

This study investigated the adequacy of using three reduction and homogenization techniques for regional analysis of mean annual SSY in south and southeast of Caspian Sea, Iran. Three reduction methods, i.e., PCA, GT and SR were used with three homogeneous techniques, i.e., CA, CDFs and AC for mean annual SSY regionalization. In this research, 14 variables were used to select effective variables on mean annual SSY. Five variables were introduced as effective variables on mean annual SSY which include: main stream length (Sl), daily maximum discharge with two-year period ($Q_{\max}$), weighed average height (H), area (A) and mean annual rainfall (R) using PCA, peak discharge with two-year return period (Qp), form factor (FF), average stream slope (Sr), watershed length (Lw) and compactness coefficient (CC) using GT and peak discharges with two-year return periods (Qp) and compactness coefficient (CC) using SR. According to these results, daily maximum discharge with two-year period ($Q_{\max}$) and peak discharge with two-year return period (Qp) were selected by two mentioned methods. This result shows the importance of floods in sediment transport. The PCA and GT were selected near effective variables. Generally, it is clear that the input selection variables (by means of the PCA and GT) has a positive effect on the catchment grouping methods and regional analysis of mean annual SSY models performance. In addition, PCA method reduces the number of input variables without eliminating them (against GT and other input reduction methods). Finally, input selection variables by the PCA and GT techniques are recommended for increasing the catchment grouping methods and regional analysis of mean annual SSY models performance especially in cases where lack of knowledge about the input variables exists. Finally, for each homogeneous group, their variables were selected by input selection techniques (PCA, SR and GT), a regression model between catchment characteristics and the mean annual SSY was fitted independently and evaluation criteria to determine the best homogenization technique. The results of the average three performance evaluation criteria for three homogenization techniques indicated that AC was the better method for watershed homogenization methods for the regional analysis of SSY (with RE=49.24, RRMSE=43.75 and CE=71.04) than CDFs (with RE=43.89, RRMSE=48.69 and CE=56.30) and CA (with RE=70.87, RRMSE=62.06 and CE=42.58).

References

Abrahams A D 1972 Factor analysis of drainage basin properties: Evidence for stream abstraction accompanying the degradation of relief; *Water Resour. Res.* **8(3)** 624–633.

Adoalbjörn S, Končar N and Jones A J 1997 A note on the Gamma test; *Neural Comput. Appl.* **5(3)** 131–133.

Alp M and Cigizoglu H K 2007 Suspended sediment load simulation by two artificial neural network methods using hydrometeorological data; *Environ. Model. Softw.* **22(1)** 2–13.

Andrews D 1972 Plots of high-dimensional data; *Biometrics* **28** 125–136.

Asselman N E M 2000 Fitting and interpretation of sediment rating curves; *J. Hydrol.* **234(3–4)** 228–248.

Bray D I and Huixi X 1993 A regression method for estimating suspended sediment yields for ungauged watersheds in Atlantic Canada; *Can. J. Civil Eng.* **20(1)** 82–87.

Çamdevýren H, Demýr N, Kanik A and Keskýn S 2005 Use of principal component scores in multiple linear regression models for prediction of Chlorophyll-*a* in reservoir; *Ecol. Model.* **181(4)** 581–589.

Caratti J F, Nesser J A and Lee Maynard C 2004 Watershed classification using canonical correspondence analysis and clustering techniques: A cautionary note 1; *JAWRA J. Am. Water Resour. Assoc.* **40(5)** 1257–1268.

Carroll S P, Dawes L, Hargreaves M and Goonetilleke A 2009 Faecal pollution source identification in an urbanising catchment using antibiotic resistance profiling discriminant analysis and partial least squares regression; *Water Res.* **43(5)** 1237–1246.

Cavadias G S, Ouarda T B M J, Bobée B and Girard C 2001 A canonical correlation approach to the determination of homogeneous regions for regional flood estimation of ungauged basins; *Hydrol. Sci. J.* **46(4)** 499–512.

Chen S, Hong X, Harris C J and Sharkey P M 2004 Sparse modelling using orthogonal forward regression with PRESS statistic and regularization; *IEEE Trans. Syst. Man Cybern. B Cybern.* **34(2)** 898–911.

Choi D J and Park H 2001 A hybrid artificial neural network as a software sensor for optimal control of a wastewater treatment process; *Water Res.* **35(16)** 3959–3967.

Chow V T, Maidment D R and Mays L W 1988 *Applied Hydrology*; McGraw-Hill, New York (SW/2778).

Cigizoglu H K 2004 Estimation and forecasting of daily suspended sediment data by multi-layer perceptrons; *Adv. Water Resour.* **27(2)** 185–195.

Corcoran J J, Wilson I D and Ware J A 2003 Predicting the geo-temporal variations of crime and disorder; *Int. J. Forecasting* **19(4)** 623–634.

de Vente J, Poesen J, Arabkhedri M and Verstraeten G 2007 The sediment delivery problem revisited; *Prog. Phys. Geog.* **31** 155–178.

Detenbeck N E, Brady V J, Taylor D L, Snarski V M and Batterman S L 2005 Relationship of stream flow regime in the western Lake Superior basin to watershed type characteristics; *J. Hydrol.* **309** 258–276.

Durrant P J 2001 Wingamma: A non-linear data analysis and modeling tool with applications to flood prediction; PhD thesis, Cardiff University, Wales, UK.

Eksioglu B, Demirer R and Capar I 2005 Subset selection in multiple linear regression: A new mathematical programming approach; *Comput. Ind. Eng.* **49(1)** 155–167.

Faraway J 2002 Practical regression and ANOVA in R; http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf.

Groupe de recherche en hydrologie statistique (GREHYS) 1996 Presentation and review of some methods for regional flood frequency analysis; *J. Hydrol.* **186(1–4)** 63–84.

Hall M J and Minns A W 1999 The classification of hydrologically homogeneous regions; *Hydrol. Sci. J.* **44(5)** 693–704.

Horhota S T and Aitken C L 1991 Multivariate cluster analysis of pharmaceutical formulation data using Andrews plots; *J. Pharm. Sci.* **80(1)** 85–90.

Ilorme F and Griffis V W 2011 A novel procedure for delineation of hydrologically homogeneous regions and the classification of ungauged sites for design flood estimation; *J. Hydrol.* **492** 151–162.

Jarvie H, Oguchi T and Neal C 2002 Exploring the linkages between river water chemistry and watershed characteristics using GIS-based catchment and locality analyses; *Reg. Environ. Change* **3(1)** 36–50.

Jobson J D 1992 *Applied multivariate data analysis: Vol. II, Catagorical and multivariate methods*; Springer-Verlag.

Kaiser H F 1960 The application of electronic computers to factor analysis; *Edu. Psychol. Meas.* **20(1)** 141–151.

Khan J A, Van Aelst S and Zamar R H 2007 Building a robust linear model with forward selection and stepwise procedures; *Comput. Stat. Data Anal.* **59** 239–248.

Kişi Ö 2010 River suspended sediment concentration modeling using a neural differential evolution approach; *J. Hydrol.* **389(1–2)** 227–235.

Koncar N 1997 Optimisation methodologies for direct inverse neurocontrol; PhD thesis, Imperial College of Science, Technology and Medicine, University of London.

Lin G F and Wang C M 2006 Performing cluster analysis and discrimination analysis of hydrological factors in one step; *Adv. Water Resour.* **29(11)** 1573–1585.

Lin S W and Chen S C 2009 PSOLDA: A particle swarm optimization approach for enhancing classification accuracy rate of linear discriminant analysis; *Appl. Soft Comput.* **9(3)** 1008–1015.

Liu C W, Lin K H and Kuo Y M 2003 Application of factor analysis in the assessment of groundwater quality in a blackfoot disease area in Taiwan; *Sci. Total Environ.* **313(1–3)** 77–89.

Ludwig W and Probst J L 1998 River sediment discharge to the oceans: Present-day controls and global budgets; *Am. J. Sci.* **298(4)** 265–295.

Melesse A M, Ahmad S, McClain M E, Wang X and Lim Y H 2011 Suspended sediment load prediction of river systems: An artificial neural network approach; *Agri. Water Manag.* **98(5)** 855–866.

Moghaddamnia A, Ghafari Gousheh M, Piri J, Amin S and Han D 2009 Evaporation estimation using artificial neural networks and adaptive neuro-fuzzy inference system techniques; *Adv. Water Resour.* **32(1)** 88–97.

Moustafa R E 2011 Andrews curves; *Wiley Interdisciplinary Reviews: Computational Statistics* **3(4)** 373–382.

Nadal-Romero E, Martínez-Murillo J F, Vanmaercke M and Poesen J 2011 Scale-dependency of sediment yield from badland areas in Mediterranean environments; *Prog. Phys. Geog.* **38** 381–386.

Nathan R J and McMahon T A 1990 Identification of homogeneous regions for the purposes of regionalization; *J. Hydrol.* **121(1–4)** 217–238.

Noori R, Hoshyaripour G, Ashrafi K and Araabi B N 2010 Uncertainty analysis of developed ANN and ANFIS models in prediction of carbon monoxide daily concentration; *Atmos. Environ.* **44(4)** 476–482.

Noori R, Karbassi A R, Moghaddamnia A, Han D, Zokaei-Ashtiani M H, Farokhnia A and Gousheh M G 2011 Assessment of input variables determination on the SVM model performance using PCA Gamma test and forward selection techniques for monthly stream flow prediction; *J. Hydrol.* **401(3–4)** 177–189.

Olden J D and Poff N L 2003 Redundancy and the choice of hydrologic indices for characterizing streamflow regimes; *River Res. Appl.* **19(2)** 101–121.

Ouarda T B M J, Cunderlik J M, St-Hilaire A, Barbet M, Bruneau P and Bobée B 2006 Data-based comparison of seasonality-based regional flood frequency methods; *J. Hydrol.* **330(1–2)** 329–339.

Owen S M, MacKenzie A R, Bunce R G H, Stewart H E, Donovan R G, Stark G and Hewitt C N 2006 Urban land classification and its uncertainties using principal component and cluster analyses: A case study for the UK West Midlands; *Landscape Urban Plan* **78(4)** 311–321.

Ramachandra Rao A and Srinivas V V 2006 Regionalization of watersheds by hybrid-cluster analysis; *J. Hydrol.* **318(1–4)** 37–56.

Ramos M C 2001 Divisive and hierarchical clustering techniques to analyse variability of rainfall distribution patterns in a Mediterranean region; *Atmos. Res.* **57(2)** 123–138.

Restrepo J D, Kjerfve B, Hermelin M and Restrepo J C 2006 Factors controlling sediment yield in a major South American drainage basin: The Magdalena River Colombia; *J. Hydrol.* **316(1–4)** 213–232.

Robertson D, Saad D and Heisey D 2006 A regional classification scheme for estimating reference water quality in streams using land-use-adjusted spatial regression-tree analysis; *Environ. Manag.* **37(2)** 209–229.

Sadeghi S H R and Mahdavi M 2004 Applicability of SEDIMOT II model in flood and sediment yield estimation; *J. Agri. Sci. Tech. (JAST)* **6** 147–154.

Sadeghi S H R and Singh J K 2005 Development of a synthetic sediment graph using hydrological data; *J. Agri. Sci. Tech. (JAST)* **7** 69–77.

Shrestha S and Kazama F 2007 Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin Japan; *Environ. Model. Softw.* **22(4)** 464–475.

Siakeu J, Oguchi T, Aoki T, Esaki Y and Jarvie H P 2004 Change in riverine suspended sediment concentration in central Japan in response to late 20th century human activities; *Catena* **55(2)** 231–254.

Sohaili S and Vafakhah M 2005 Study on efficiency of non-numerical (Andrew curve) method in determination of homogeneous area in flood estimation; *Pajouhesh & Sazandegi* **68** 73–81 (in Persian).

Syvitski J P M and Milliman J D 2007 Geology, geography and humans battle for dominance over the delivery of fluvial sediment to the coastal ocean; *J. Geol.* **115** 1–19.

Tramblay Y, Ouarda T B M J, St-Hilaire A and Poulin J 2010 Regional estimation of extreme suspended sediment concentrations using watershed characteristics; *J. Hydrol.* **380(3–4)** 305–317.

Tramblay Y, St-Hilaire A and Ouarda T B M J 2007 Modelling extreme suspended sediment concentrations in North America: Frequency analysis and correlations with watershed characteristics; In: Water Quality and Sediment Behaviour of the Future: Predictions for the 21st Century Proceedings of Symposium HS2005 at IUGG2007, Vol Perugia, Italy, July 2007; *IAHS Publ.* **314** 20–27.

Tramblay Y, St-Hilaire A and Ouarda T B M J 2008 Frequency analysis of maximum annual suspended sediment concentrations in North America/Analyse fréquentielle des maximums annuels de concentration en sédiments en suspension en Amérique du Nord; *Hydrol. Sci. J.* **53(1)** 236–252.

Tryon R C 1939 *Cluster analysis*; McGraw–Hill, New York.

Tsui A P M, Jones A J and Guedes de Oliveira A 2002 The construction of smooth models using irregular embeddings determined by a Gamma Test analysis; *Neural Comput. Appl.* **10(4)** 318–329.

Vafakhah M 2007 Regional analysis of sediment yield in the part of Caspian Sea coastal basin; *Iranian Journal of Agricultural Sciences and Natural Resources* **13(6)** 121–132 (in Persian).

Vanmaercke M, Poesen J, Verstraeten G, de Vente J and Ocakoglu F 2011 Sediment yield in Europe: Spatial patterns and scale dependency; *Geomorphology* **130(3–4)** 142–161.

Walling D E and Webb B W 1988 The reliability of rating curve estimates of suspended yield: Some further comments on sediment budgets; In: Proceedings of the Porto Alegre Symposium, December 1988; *IAHS Publ.* **174** 337–350.

Wang X X, Chen S, Lowe D and Harris C J 2006 Sparse support vector regression based on orthogonal forward selection for the generalised kernel model; *Neurocomputing* **70(1–3)** 462–474.

Water Resources Research Center (WRRC) 1996 Water Year Report, Water Resources Management Organization, Ministry of Energy Iran.

White E L 1975 Factor analysis of drainage basin properties: Classification of flood behavior in terms of basin geomorphology; *JAWRA J. Am. Water Resour. Assoc.* **11(4)** 676–687.

Wilson D I 2002 Derivation of the chalk superficial deposits of the North Downs England: An application of discriminant analysis; *Geomorphology* **42(3–4)** 343–364.

Yadav M, Wagener T and Gupta H 2007 Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins; *Adv. Water Resour.* **30(8)** 1756–1774.

Zhang Q, Wu F, Wang L, Yuan L and Zhao L 2011 Application of PCA integrated with CA and GIS in eco-economic regionalization of Chinese Loess Plateau; *Ecol. Econ.* **70(6)** 1051–1056.

Zhang Y, Li H, Hou A and Havel J 2006 Artificial neural networks based on principal component analysis input selection for quantification in overlapped capillary electrophoresis peaks; *Chemomet. Intell. Lab. Syst.* **82(1–2)** 165–175.

Zhang Y X 2007 Artificial neural networks based on principal component analysis input selection for clinical pattern recognition analysis; *Talanta* **73(1)** 68–75.

MS received 18 November 2014; revised 16 April 2015; accepted 17 April 2015