A. ŠTEVKOVÁ, M. SABO, S. KOHNOVÁ

POOLING OF LOW FLOW REGIMES USING CLUSTER AND PRINCIPAL COMPONENT ANALYSIS

Andrea ŠTEVKOVÁ

Department of Land and Water Resources Management
email: andrea.stevkova@stuba.sk
Research field: seasonality and regionalization of low flows

Miroslav SABO

Department of Mathematics and Constructive Geometry
email: miroslav.sabo@stuba.sk
Research field: Multivariate statistics, machine learning, data mining, programmingg

Silvia KOHNOVÁ

Department of Land and Water Resources Management
email: silvia.kohnova@stuba.sk
Research field: Design discharge estimation, regionalization, rainfall runoff modeling

Address:
Faculty of Civil Engineering,
Slovak University of Technology in Bratislava,
Radlinského 11, 813 68 Bratislava.

ABSTRACT

This article deals with the regionalization of low flow regimes lower than Q_{95} in Slovakia. For the regionalization of 219 small and medium-sized catchments, we used a catchment area running from 4 to 500 km² and observation periods longer than 20 years. The relative frequency of low flows lower than Q_{95} was calculated. For the regionalization, the non-hierarchical clustering K-means method was applied. The Silhouette coefficient was used to determine the right number of clusters. The principal components were found from the pooling variables on the principal components. The K-means clustering method was applied. Next, we compared the differences between the two methods of pooling data into regional types. The results were compared using an association coefficient.

KEY WORDS

- Low flow,
- Cluster analysis,
- Principal component analysis.

1. INTRODUCTION

Low flows, their regimes, and their influence on the biological functions of a river and its stability is an important discussion topic. Low flow is influenced by a number of aspects, mainly geological, hydrological and climatic factors (air temperature and precipitation). The occurrence of low flows also has a growing importance economically, because of the increased use of surface water sources for irrigation, fresh water production and power generation. Knowledge of low flows is also important for the biological balance of a river and the protection of an ecosystem and its biodiversity, as they have an impact on water temperature, lighting, speed, oxygen and changes in sedimentation. A limiting factor for organisms is also the duration of low flow periods, as a short duration period does not cause explicit damage to organisms.

One of the possible methods for assessing a low flow regime is to use a seasonality analysis. The most widely used seasonality analysis was presented by Burn (1997). This method was developed for

testing seasonality occurrences of floods on catchments in Canada. Also, this method has been used for the occurrence of seasonality for low flows in Austria (Laaha, 2006a, b). The authors used a seasonality cluster analysis to represent the monthly occurrence of low flows. They classified Austrian catchments into two clusters of winter and summer low flow types. Young (2000) applied a low-flow seasonality index for the estimation of low flow seasonality in the UK. Schreiber (1997) analysed the seasonality of mean annual 10-day minimums of total discharges measured in 169 catchments in southwest Germany. The results indicated the typical occurrence of low flows from September to October for large parts of the area studied, apart from the Pre-Alps, which are dominated by low winter flows.

Cluster analysis has been used in many studies. Rao and Srinivas (2006) used cluster analysis on the attributes and flow records from 245 gauging stations in Indiana. They used hierarchical and non-hierarchical analyses. Of the hierarchical clustering methods, they used single and complete linkage and Ward clustering methods, and for the non-hierarchical clustering analysis, they used

K-means clustering. The clustering algorithm yielded six acceptably homogeneous hydrologic regions that have been tested to be robust. Kahay (2007) used hierarchical clustering (single, complete and Ward methods) to create regional types of monthly low flows in Turkey. They found that the Ward method with Euclidean distance was more effective in the production of homogenous clusters compared to single linkage and complete linkage methods.

In Nathan and MacMahon (1990), the authors identified homogeneous subregions that can be considered to behave in a hydrologically similar fashion. The relative performance of several techniques is evaluated using the prediction of low-flow characteristics in a heterogeneous group of 184 catchments located in southeastern Australia. The authors offered approaches based on a combination of cluster analysis, multiple regression, principal component analysis, and the graphic representation of multi-dimensional data, when used for the identification of pooling groups. The techniques presented allow for the assessment of the suitability of applying the regional equations derived to an ungauged catchment.

Kohnová, et al. (2008) used the Burn vector and histograms of occurrences for a better understanding of low-flow seasonality in Slovakia. This article follows the studies of Kohnová, et al. (2009), Števková, et al. (2010), and Števková, et al. (2011) where the aim of the studies was to examine seasonality indices and their potential in the regionalisation of low flows in Slovakia.

We compared here the results of the K-means clustering method and a combination of the PCA and clustering methods by selecting the pooling variables. For reducing the high-dimensional data, Principal component analysis (PCA) was used and for regionalization, the non-hierarchical K-means clustering method was used. For determining the right number of clusters, the Silhouette coefficient was used. We compared the differences between the results of the pooling groups.

2. CLUSTER ANALYSIS

The information cluster analysis groups are based on data describing the objects or their relationships. The goal is to make the objects in a group similar (or related) to one another and different from the objects in other groups. The greater the similarity (or heterogeneity) within a group and the greater the differences between the groups, the better or more distinct the clustering (Tan, et al., 2006).

Data clustering (or just clustering), which is also called cluster analysis, is a method of creating groups of objects or clusters in such a way that objects in one cluster are very similar and objects in different clusters are quite distinct. Data clustering is often confused with classification, by which objects are assigned to predefined classes (Gan, et al., 2007).

2.1 Non-hierarchical cluster analysis – K – means clustering

The algorithm is very simple. Denote D as the set of all objects, i.e. $D = \{x_1, x_2, \dots, x_n\}$. Each object x_i has p features, i.e., $X_i = (X_{i1}, X_{i2}, \dots, X_{ip})$. Then, the K-means algorithm can be described as follows:

1. The user sets the number of clusters k . Denote these clusters $X_1, X_2, \dots, X_k$.
2. The algorithm randomly initializes the k cluster centers $\{\mu_1, \mu_2, \dots, \mu_k\}$ as random points in a p -dimensional space.
3. Then it repeats those two steps until the convergence criterion is satisfied (for example, until the clusters change)
 - $\forall i \in \{1, 2, \dots, n\}$ assign x_i to cluster X_j if

$$j = \arg \min_{l \in \{1, 2, \dots, k\}} d(x_i, \mu_l) \quad (1)$$

where $d(\cdot, \cdot)$ is some measure of distance (mostly Euclidean)

– find new cluster centers such as

$$\mu_l = \frac{\sum_{i: x_i \in X_l} x_i}{|X_l|} \quad (2)$$

where $l \in \{1, 2, \dots, k\}$ and $|\cdot|$ denotes the cardinality of the set (McQueen, 1967).

A silhouette coefficient was used to determine the correct number of clusters (Rousseeuw, 1987):

Let a_i denote the average distance between the i -th object and all the other objects in the same cluster, i.e.:

$$a_i = \frac{1}{|C(i)|} \sum_{j \in C(i)} \text{dist}(i, j) \quad (3)$$

Let b_i denote the average distance between the i -th object and all the other objects from the nearest cluster, i.e.:

$$b_i = \min_{C_k \in C \setminus C(i)} \sum_{j \in C_k} \frac{\text{dist}(i, j)}{n(C_k)} \quad (4)$$

where $C(i)$ is a cluster, where the i -th object has been assigned; $\text{dist}(i, j)$ is some measure of the distance of the i -th and the j -th object; and $|\cdot|$ denotes the cardinality of the set as before. Then we can define for the i -th object the width of its silhouette as:

$$S(i) = \frac{b_i - a_i}{\max(b_i, a_i)} \quad (5),$$

This measure takes the values from $<-1, 1>$; values close to 1 indicate that the object is well assigned, while negative values close to -1 indicate a poor assignment of the i -th object into its cluster. If we compute the width of the average silhouette for all the assumed objects, we will again obtain a measure with values from $<-1, 1>$, and again, values close to 1 mean a better partition.

According to (Meloun, 2005), the following rule of thumb is used to determine the correct number of clusters:

0.71 – 1.00	we can say that the i number of clusters is appropriate,
0.51 – 0.70	the structure has an average structure; the number of clusters is appropriate,
0.26 – 0.50	the structure is breakable, and we have to use another method to determine the correct number of clusters,
≤ 0.25	the structure is not good.

For determining the right number of clusters, we can define it as an average silhouette value. We used the highest average silhouette value for the correct number of clusters.

Dissimilarities in the measures between the clusters were determined with the Squared Euclidean distance.

Squared Euclidean distance

Let $x = (x_1, x_2, \dots, x_p)$ and $y = (y_1, y_2, \dots, y_p)$ are two points in p -dimensional space R^p (in our case, two observations of the p -dimensional random vector $X = (X_1, X_2, \dots, X_p)$). The Euclidean distance between them is defined as (Meloun, 2005):

$$d(x, y) = \sum_{i=1}^p (x_i - y_i)^2 \quad (6),$$

2.2 Principal component analysis (PCA)

Principal component analysis (“PCA“, also known as the Karhunen-Loeve transformation or Hotelling transformation or proper orthogonal decomposition) was originally proposed by Karl Pearson in 1901 (Pearson, 1901) and later generalized by Hotelling (Hotelling, 1935). This section is processed according to (Harman, 2011). Since the PCA is derived from matrix algebra, we first mention a theorem that is crucial to the idea of PCA.

Theorem 1: If Σ is a positively semidefinite $p \times p$ matrix, then there exists a system $u_1, u_2, \dots, u_p$ of orthogonal eigenvectors with a norm of 1. Matrix Σ can be expressed in the form

$$\Sigma = \sum_{i=1}^p \lambda_i u_i u_i^T = U \Lambda U^T \quad (7),$$

where λ_i is the eigenvalue of matrix Σ , which corresponds to eigenvector $u_i \forall i \in \{1, 2, \dots, p\}$; $U = u_1, u_2, \dots, u_p$ is an orthogonal matrix and $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$. If $\lambda_1 > \lambda_2 > \dots > \lambda_p$, then vectors $u_1, u_2, \dots, u_p$ are uniquely defined.

Now consider a p -dimensional random vector X with a mean μ and covariance matrix Σ , which is a positive semidefinite. Theorem 1 implies that a system of orthonormal eigenvectors $u_1, u_2, \dots, u_p$ of matrix Σ and $\lambda_1, \lambda_2, \dots, \lambda_p$ as corresponding eigenvalues have to exist. As before, we will denote as U the orthogonal matrix $(u_1, u_2, \dots, u_p)$. Without a loss in generality we can assume that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$, since we can easily reorder the eigenvalues with a suitable renumbering of the corresponding eigenvectors.

Theorem 2: Random vector $Y = U^T (X - \mu)$ has a zero mean and covariance matrix $D(Y) = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$, i.e., the components of this vector are uncorrelated with the variances $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$.

Eigenvector $Y = (Y_1, Y_2, \dots, Y_p)^T$ from the previous theorem is called the vector of the principal components of the random vector X . $\forall i \in \{1, 2, \dots, p\}$, random variable $Y_i = u_i^T (X - \mu)$ is called the i -th principal component of the random vector X . $\forall k \in \{1, 2, \dots, p\}$

$\alpha_k = \frac{\lambda_1 + \dots + \lambda_k}{\lambda_1 + \dots + \lambda_p}$ is called the total variance proportion, which is

explained by the first k principal components. In addition, the following theorem holds:

Theorem 3: The first principal component Y_1 of the random vector X has a maximal variance from all the normalized linear combinations of the components of vector X , i.e., $D(Y_1) = D(b^T X) \forall b \in R^p \|b\| = 1$. In addition, for $k \geq 2$, the k -th principal component Y_k of the random vector X has the largest variance from all such normalized linear combinations of the components of vector X that are uncorrelated with the principal components $Y_1, Y_2, \dots, Y_{k-1}$, i.e. $D(Y_k) = D(b^T X) \forall b \in R^p \|b\| = 1, b \perp u_1, \dots, b \perp u_{k-1}$ ($\perp$ is the symbol for orthogonality).

According to the previous calculations, we can perform the PCA analysis in the following way. We can decompose the covariance matrix to obtain the eigenvalues and corresponding eigenvectors. We chose such a number of eigenvalues that the proportion of the total variance will be at least 90 %. If this number is k , then each object will be described by the k dimensional vector (instead of the original vector that was p -dimensional).

Association coefficients were used to compare the various clustering techniques. A comparison of the similarities between the Jaccard, Sokal-Sneath (1), Dice and Russel Rao clusters were used, and the similarities between clustering techniques of the Rand coefficient were used (Nánásiová, et al. 2010; Jackson, et al. 1989).

$$\text{Jaccard coefficient } S_j = \frac{a}{a+b+c} \quad (8),$$

$$\text{Sokal-Sneath (1) } S_{SS1} = \frac{2(a+d)}{2(a+d)+b+c} \quad (9),$$

$$\text{Russel Rao coefficient } S_{RR} = \frac{a}{a+b+c+d} \quad (10),$$

$$\text{Dice coefficient } S_D = \frac{2a}{2a+b+c} \quad (11),$$

$$\text{Rand coefficient } S_{RAND} = \frac{a+d}{a+b+c+d} \quad (12)$$

The detailed code from the R language for determining the similarities between the clusters using similarity coefficients can be found in the appendix.

3. DATA

The seasonality analysis in this study was based on the data from 219 small and medium-sized catchments selected from the entire territory of Slovakia, with catchment areas ranging from 4 to

Fig. 1 Spatial distribution of the analyzed catchments in Slovakia.

500 km² and with observation periods longer than 20 years. Figure 1 illustrates the spatial distribution of the selected catchments in Slovakia.

The number of low-flow days with a discharge lower than the Q_{95} was calculated in each catchment. Subsequently, the relative frequency of low-flow occurrences in separate months during the year was estimated in all the observation stations. The relative frequency was estimated and used as a pooling variable in the clustering methods.

4. RESULTS

It is possible to see in Tab. 1 – the correlation matrix, that the strongest correlations are mostly between two succeeding months (i.e., between the diagonals). We had to reduce the correlations

Tab. 1 Correlation matrix of the relative frequency of low flows lower than Q_{95} .

	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
I	1											
II	0.795	1										
III	0.603	0.808	1									
IV	0.175	0.319	0.538	1								
V	-0.209	-0.207	-0.162	0.184	1							
VI	-0.520	-0.504	-0.428	-0.061	0.501	1						
VII	-0.717	-0.672	-0.553	-0.235	0.183	0.606	1					
VIII	-0.821	-0.790	-0.668	-0.317	0.102	0.430	0.796	1				
IX	-0.823	-0.835	-0.726	-0.333	0.121	0.395	0.573	0.757	1			
X	-0.415	-0.467	-0.475	-0.292	-0.028	0.0468	-0.038	0.1093	0.4463	1		
XI	0.170	-0.004	-0.124	-0.148	-0.089	-0.256	-0.435	-0.366	-0.195	0.412	1	
XII	0.695	0.405	0.194	0.017	-0.151	-0.380	-0.591	-0.642	-0.593	-0.148	0.440	1

Tab. 2 Table of principal components. The four principal components (PC) are bolded.

	1 PC	2 PC	3 PC	4 PC	5 PC	6 PC	7 PC	8 PC	9 PC	10 PC	11 PC	12 PC
I	0.39	-0.03	-0.10	-0.46	0.33	-0.06	0.45	-0.46	0.13	0.00	-0.04	0.29
II	0.45	0.24	0.28	0.10	0.03	-0.64	-0.35	0.13	-0.11	0.00	-0.03	0.29
III	0.29	0.28	0.35	0.36	-0.18	0.55	0.21	-0.10	-0.23	0.25	0.05	0.29
IV	0.03	0.05	0.05	0.08	-0.03	0.22	-0.02	0.12	0.17	-0.81	-0.39	0.29
V	-0.01	0.01	-0.02	0.02	0.02	0.08	-0.07	0.13	0.41	-0.17	0.83	0.29
VI	-0.06	0.05	-0.10	0.06	0.06	0.07	-0.08	0.26	0.66	0.49	-0.38	0.29
VII	-0.27	0.51	-0.65	0.25	0.10	-0.12	0.04	-0.12	-0.24	-0.01	0.02	0.29
VIII	-0.53	0.26	0.37	-0.47	-0.39	-0.16	0.16	-0.02	-0.06	0.03	-0.02	0.29
IX	-0.39	-0.23	0.32	0.09	0.67	0.10	-0.28	-0.16	-0.20	0.05	-0.02	0.29
X	-0.11	-0.54	-0.02	0.41	-0.15	-0.34	0.53	0.14	-0.08	0.02	0.00	0.29
XI	0.05	-0.37	-0.21	-0.02	-0.47	0.07	-0.48	-0.52	-0.01	0.06	-0.04	0.29
XII	0.16	-0.21	-0.27	-0.42	0.02	0.23	-0.10	0.58	-0.43	0.10	0.01	0.29

between the input data, so we reduced the values higher than 0.7. The reduced data for pooling into the pooling groups was used.

For the pooling of the catchments into pooling groups, we used the non-hierarchical K-means cluster method. A silhouette coefficient was used to determine the right number of pooling groups. The highest average silhouette of 0.383 occurred in two pooling groups. The analyzed catchments were pooled into two pooling groups, for which the average silhouette was the highest (0.383). These two groups represent two regimes. The occurrence of low flows in the first pooling group is during the winter (November – April), and the second pooling group has an occurrence of low flows during the summer – autumn (July – October).

Subsequently, we used the principal component analysis for the selection of the derived variables. Principal component analysis reduces the dimensionality of the data and creates new variables – principal components.

The reason why the principal components analysis was used is the strong correlation between the input data. In the correlation matrix (Tab. 1), you can see that the strongest correlations are mostly between two succeeding months (i.e. between the diagonals).

The percentage of the total variance is set at a horizon of 90% (Stankovičová, et al. 2007; Meloun, et al. 2005) for the selection of the principal component; for the analyzed data the horizon was set at 91.54%. The first principal component makes up 66.78% of the total variability of the data; the second principal component makes up 12.62%; the third principal component makes up 6.9%; and the fourth makes up 5.22%. For the next steps, we used four principal components. The first component explains almost 66.78% of the total variability of the data, and we can conclude that this

Tab. 3 Percent explained of the principal components for finding the principal components.

percent explained	
66.7864	91.54085 %
12.62353	
6.902576	
5.228347	
2.659351	
2.004881	
1.361264	
1.069859	
0.863966	
0.320155	
0.179673	
1.16E-29	

component is the most important. The first PC was most dominant (-0.53); the second PC was most dominant in October (-0.54); the third PC was most dominant in July (-0.65); and the last PC was most dominant in August (-0.47).

These four principal components were used for the cluster analysis, and the non-hierarchical K-means clustering method was used. The highest average silhouette of 0.703 occurred in two pooling groups. The analyzed catchments with the highest silhouette (0.703) were pooled into two pooling groups. The first pooling group represents

Fig. 2 Comparison of the low-flow regimes in the catchment following the derived pooling group's winter regime, and the summer – autumn regime.

a winter regime with the occurrence of low flows from December until March, and the second pooling group represents a summer – autumn regime, with the occurrence of low flows from July until November.

Subsequently, similarities between the PCA (K-means) and the K-means clustering methods using the Russel-Rao, Dice, Jaccard, Sokal-Sneath's (1) and Rand coefficients were determined. The Rand coefficient reached a value of 0.9539, which proved that these two clustering methods are comparable. Table 4 shows that the compared methods are very similar; the differences are in ten catchments. In fig. 2 the catchments are pooled into regional types, where catchments are marked when there are differences between

the pooling used. Here it is possible to see the differences are in the catchments between the regional types.

5. CONCLUSION

This article deals with comparing the classical clustering method and a combination of PCA and clustering by selecting the pooling variables. In the first step, the relative frequency of the occurrence of low flows lower than Q_{95} was used as the pooling variable for the clustering and created pooling groups using the non-hierarchical, K-means method. In the second step, we calculated the

Tab. 4 Comparison of the pooling catchments into pooling groups, PCA (K-means) and K-means.

K-means	PCA (K-means)	Russell and Rao	Dice	Jaccard	Sokal and Sneath's 1
1	1	0.374	0.943	0.891	0.804
1	2	0.014	0.028	0.014	0.007
2	1	0.032	0.063	0.032	0.016
2	2	0.58	0.962	0.927	0.864

Fig. 3 Pooled catchments into regional types, with marked catchments where there are differences.

principal components and created pooling groups from the principal components. Also, the non-hierarchical K-means clustering method was used.

The K-means method pooled the analyzed catchments into two pooling groups with a summer-autumn regime and a winter regime of occurrences of low flows lower than Q_{95} . The coefficient association which was used in this study also shows very small dissimilarities in the analyzed regional types. A difference was found in ten catchments which are situated on the boundary of the clusters. If there is a strong correlation between data, it is better to use PCA, because there were quite better results in our case.

The derived pooling groups can be subsequently used for the indirect estimation of design minimum discharges on ungauged catchments.

Acknowledgement

This work was supported by the Slovak Research and Development Agency under Contract No. APVV-0015-10 and the Slovak VEGA Grant Agency under Project No. 1/0908/11. The authors are grateful to RNDr. Beate Demeterovej, PhD., from SHMÚ Košice for her cooperation and for processing the input data.

APPENDIX 1

```

koeficienty<-function(x,y){
  n<-length(x)
  r<-(sort(x))[n]
  s<-(sort(y))[n]
  a1<-NULL
  for (i in 1:r){
    for(j in 1:s){
      u<-0
      v<-0
      z<-0
      w<-0
      for(k in 1:n) { if (x[k]==i&y[k]==j)u<-u+1
        if (x[k]==i&y[k]!=j)v<-v+1
        if (x[k]!=i&y[k]==j)z<-z+1
        if (x[k]!=i&y[k]!=j)w<-w+1
      }
      a<-c(i,j,round(u/n,3),round(((2*(u/n))/(2*(u/n+v/n+z/n)-v/n-
z/n)),3),round((1-
(v/n+z/n)/(u/n+v/n+z/n)),3),round((((u/n)/((u/n+v/n+z/n)+v/n+z/n)),3),round(((1-v/n-
z/n)/(1+v/n+z/n)),3),round(((2*(1-v/n-z/n))/(2-v/n-z/n)),3))
      a1<-rbind(a1,a)
    }
  }

  colnames(a1)<-c("Cislo          zhliku          1","Cislo          zhliku          2","Russel          and
Rao","G1:Dice","G1:Jaccard","G2:Sokal-Sneath`s          1","G2:Rogerr-Tanimoto","G2:Sokal-
Sneath`s 2")
  a1
}

```


REFERENCES

- [1] Burn D. H. (1997): Regionalisation of catchments for regional flood frequency analysis. *J. Hydrol. Engng.* 2 (2), 76-82.
- [2] Gan, G., Ma, Ch., Wu, J., (2007): Data clustering: theory, algorithms and application. Society for Industrial and Applied Mathematics and the Society for Industrial and Applied Mathematics, Philadelphia, 184.
- [3] Harman, R. (2011): Data mining (Mnohorozmerné štatistické analýzy), [in Slovak], 2011, manuscript, 8.
- [4] Hotelling, H., 1935: The most predictable criterion. *J. Ed. Psych.*, 26, 139-142.
- [5] Jackson, D., A., Somers, K., M., Harvey, H., H., (1989): Similarity coefficients: Measures of co-occurrence and association or simply measures of occurrence. *The American Naturalist*, Vol. 133, No. 3, March 1989, 436 – 453.
- [6] Kahya, E., Demirel, M. C., 2007: A Comparison of Low-Flow Clustering Methods: Streamflow Grouping. *Journal of Engineering and Applied Sciences* 2 (3), 524 – 530.
- [7] Kohnová, S., Hlavčová, K., Szolgay, J., Števková, A. (2009): Seasonality analysis of the occurrence of low flows in Slovakia: WMHE 2009. In: *Water Management and Hydraulic Engineering: Proceedings of 11th International Symposium*. Vol. II/Ohrid, Macedonia, 1. -5. 9. 2009. - Skopje : University of Sts. Cyril and Methodius, 711-720.
- [8] Kohnová, S., Szolgay, J., Hlavčová, K., Gaál, L., Števková, A., (2008): Assessment of low flows using pooling typification. (Posúdenie malej vodnosti metódou regionálnej typizácie), [in Slovak], Svf, STU Bratislava, 102.
- [9] Laaha, G., Blöschl, G. (2006a): A comparison of low flow regionalisation methods -catchments grouping. *Journal of Hydrology*, Vol. 323, Issues 1-4, (2006), 93-214,
- [10] Laaha, G., Blöschl, G., (2006b): Seasonality indices for regionalizing low flows. *Hydrological Processes* 20, 3851-3878,
- [11] MacQueen, J. B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 1:281-297,
- [12] Meloun, M., Miličák, J., Hill, M. (2005): Computer analysis of multivariate data in examples (Počítačová analýza vícerozměrných dat v příkladech), [in Czech] Prague: Academia. 2005. ISBN 80-200-1335-0, 449.
- [13] Nánásiová, O., Pastuchová E., Václavíková, Š. (2010): Classification of Association Coefficients and Its Application in Clustering Analysis (Asociačné koeficienty ako nástroj klasifikácie zhlukov). [in Slovak], *Magia - Mathematics, geometry and their applications*, Slovak University of Technology in Bratislava, Publishing House of STU, 51 – 55.
- [14] Nathan, R.J., McMahon, T.A. (1990): Identification of homogeneous regions for the purpose of regionalization. *Journal of Hydrology*, 121, 217-238.
- [15] Pearson, K. (1901): On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine* 2(6), 559–572.
- [16] Rao, A. R., Srinivas, V. V., (2006): Regionalization of watersheds by hybrid – cluster analysis. *Journal of Hydrology* 318, 37 – 56.
- [17] Rousseeuw, P. J., 1987: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computation and Applied Mathematics* 20, 53 – 65.
- [18] Schreiber P., Demuth S.: Regionalisation of low flows in southwest Germany. *Hydrol. Sci. J.* 42 (1997), 6:845-858.
- [19] Stankovičová, I., Vojtková, M. (2007): Multivariate data analysis with examples (Viacrozmerné štatistické metódy s aplikáciami). [in Slovak], Bratislava: IURA Edition, ISBN 978-80-8078-152-1, 261.
- [20] Števková, A., Kohnová, S., Hlavčová, K., (2011): Comparison of Summer Low Flow Regime Pooling Schemes in Slovakia. In: *Ecological methods in water management*. - Gdańsk: Wydawnictwo Politechniki Gdańskiej, 2011. - ISBN 978-83-7348-376-7, 148-153.
- [21] Števková, A., Kohnová, S., Hlavčová, K., Bohdalová, M. (2010): Pooling of seasonal occurrence of minimum annual discharges in Slovakia (Regionálna typizácia výskytu minimálnych priemerných denných prietokov na Slovensku) [in Slovak]. In: *Acta hydrologica Slovaca*. Roč. 11, č. 1, 36-45.
- [22] Tan, P. N., Steinbach, M., Kumar, V., (2006): Introduction to data mining. Addison-Wesley. 769.
- [23] Young AR, Round CE, Gustard A.: Spatial and temporal variations in the occurrence of low flow events in the UK. *Hydrology and Earth System Sciences* 4 (2000): 35–45.
- [24] R Development Core Team (2010): R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.