Journal of Ecological Engineering
Volume 18, Issue 2, March 2017, pages 44–50
DOI: 10.12911/22998993/67504

Research Article

INITIAL RESULTS OF NONHIERARCHICAL CLUSTER METHODS USE FOR LOW FLOW GROUPING

Agnieszka Cupak[1]

[1] Department of Sanitary Engineering and Water Management, Faculty of Environmental Engineering and Land Surveying, al. Mickiewicza 24-28, 30-059 Krakow, Poland, e-mail: a.cupak@ur.krakow.pl

Received: 2016.11.28
Accepted: 2017.01.02
Published: 2017.03.01

ABSTRACT
In the paper the possibility of using statistical method for data agglomeration, i.e. nonhierarchical cluster analysis for low flow grouping was made. The study material included daily flows from the multi-year period of 1963–1983 collected for 19 catchments, located in the upper Vistula basin. Regions with the same flow were determined with the use of nonhierarchical cluster analysis (K-means). Groups were characterized by low flow and selected physiographic and meteorological features of the catchments. The procedure of catchments assigning to the clusters was started from two clusters and finished at five. The next moving and assigning of catchments into clusters resulted in a cluster in which there was only one catchment (for five clusters). Another objects' delineation did not give an objective effects, based on which it was difficult to determine a clear criterion of assigning each catchments into the clusters. The last step involved development of the models reflecting correlation and regression relationships. The identified clusters comprised catchments similar in terms of unit runoff, watercourse length, mean precipitation, median altitude, mean catchment slope, watercourse staff gauge zero, area covered by coniferous forests, arable lands, and soils.

Keywords: low flow, K-means method, nonhierarchical cluster analysis

INTRODUCTION

Determination of river discharges during floods and droughts is a necessary precondition for many topics related to water resources management. For example, usage of water resources, the qualitative and quantitative protection of water courses and flood and drought protection are very important issues. In case of an assessment of water resources, low flows need to be known. Determination of low flows are necessary for aquatic ecosystems, what depends on the concentration of pollutants and water temperature. Quantity and seasonality of low flows have to be considered when determining design values of acceptable sewage loads. Next, the consistent determination of low flow characteristics should take different physical processes of summer and winter droughts into account [Laaha 2002]. In case of gauged watersheds, direct – statistical methods can be used for low flows evaluation. But in case of the ungauged ones, different methods must be used.

Hydrologic regionalization is a typical example where statistical techniques are needed in order to group river flow series with similar behavior. The analysis is aimed at delineating homogeneous regions, where watersheds are similar with respect to several attributes such as physical, climatic, and hydrologic features in order to transfer information or models from gauged watershed to ungauged ones. In Poland hydrological division was made for the sake of maximum flow, e.g. Dębski [1961], Dynowska [1971], Ziemońska [1973] or Pociask-Karteczka [1995]. Cluster analysis has been used in many studies e.g. Srinivas [2009], Rao and Srinivas [2006a, b], Števková et al. [2012], Kahya and Demirel [2007], Kahya et al. [2008], Gottschalk [1985], Lin and Chen [2006], Laaha and Blöschl [2006], Wałega et al. [2009], Gutry-Korycka et al. [2009] or Bryndal [2011]. Rutkowska et al. [2016]

divided the Upper Vistula River basin into pooling groups with similar dimensionless frequency distributions of annual maximum river discharge with the use of cluster analysis and the Hosking and Wallis L-moment.

Ungauged sites, sharing similar features with one of the identified clusters, are then treated in an analogous manner as any other member of the group [Corduas 2011]. Among the aforementioned studies, the K-means method and Ward's method are the most frequently used. The K-means method is the best known of the nonhierarchical clustering methods. The values of seeds have a great influence on the quality of clustering using a K-means clustering or a related technique. When the number of clusters is too large, there is probably no training data in the cluster. In addition, no objective method to determine the number of cluster is another disadvantage [Lin and Chen 2006]. Burn [1989] used the K-means clustering algorithm to determine appropriate grouping of a network of streamflow gauging stations in southern Manitoba, Canada. Kowalczak [1986] used nonhierarchic cluster analysis for particle catchments classification for the Upper Noteć basin (Poland). Lecce [2000] also used the K-means method to examine spatial variations in the timing of flooding in the southeastern United States [Lin and Chen 2006]. Burn and Goel [2000] applied the K-means algorithm to site characteristics (catchment area, length and slope of the main stream of river) of a collection of catchments in India to derive regions for flood frequency analysis [Rao and Srinivas 2008].

In nonhierarchical – partitional clustering algorithms a single partition of the data to recover the natural grouping present is generated. The partitional clustering algorithms require an initial guess of the number of clusters and cluster centers. They can be classified based on the technique used to initiate clusters, clustering criteria, and the type of data for which they are applicable [Rao and Srinivas 2008].

The aim of the research was to evaluate the possibility of using statistical methods for data agglomeration, i.e. nonhierarchical cluster analysis for low flows grouping connection with selected physiographic and meteorological features of the catchments.

MATERIALS AND METHODS

The study material included daily flows from the multi-year period of 1963–1983 collected for 19 catchments (Fig. 1). At least 10 – years long observational daily flows was a criterion for catchments chosen. located in the upper Vistula basin and characterized by specific physiographic and meteorological parameters (Table 1). For analysis 11 physiographic and meteorological parameters of the catchment were investigated: length of the watercourse (L), catchment area (A), mean annu-

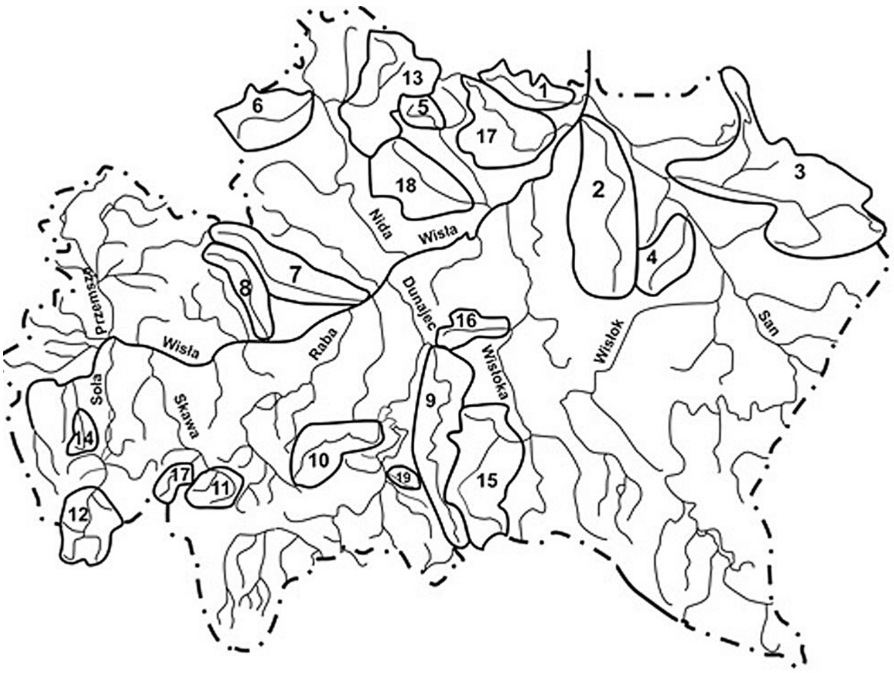

Figure 1. Location of analyzed catchments in Upper Vistula basin

al air temperature (T), mean annual precipitation (P), mean catchment slope (i), median catchment altitude (H_{me}), watercourse staff gauge zero (P_z), land cover (U) and soils (S).

Low flows were quantified by $Q_{95\%}$, i.e. the discharge that is exceeded on 95% of all days of the measurement period. This low flow characteristic is widely used in Europe and was chosen due to its relevance for multiple choices of water management, e.g. for the design of water supply systems. Then, $Q_{95\%}$ was subsequently standardized by the catchment area and resulting specific low flow discharges q_{95} $dm^3 \cdot s^{-1} \cdot km^{-2}$.

Physiographic parameters, land cover and air temperature were determined as specified in the Hydrological Atlas of Poland [Stachý 1987] and by Chełmicki [1991]. Data on daily flows and mean precipitation were taken from the Hydrologic Yearbook for the Vistula basin [1963–1983].

The research included 19 selected catchments located in the upper Vistula basin (Fig. 1). This area is spread within three great Carpathian physiographic units: the Carpathians (40% of the basin area), the Subcarpathian valleys (about 35% of the basin area) and the Małopolska Upland (about 25% of the basin area). The Carpathians and the Upland are the source areas for most of the upper Vistula tributaries, while the Subcarpathian valleys are a transit area for the Vistula and an estuary area for the rivers and streams formed in the Carpathians and Subcarpathian Uplands [Chełmicki 1991]. The investigated catchments differ in terms of their area, from 70.3 km^2 to 2034 km^2, and mean catchment slope – from 0.002 for the Łęg at Kępie Zaleszańskie gauge to 0.091 for the Biała at Bielsko gauge (Table 1).

Interpretation of hydrological events was also based on median catchment altitude that ranged from 720 m a.s.l. (the Soła, at Cięcina gauge) to 202.0 m a.s.l. (the Łęg, Kępie Zaleszańska gauge). The area was characterized by diverse land use. The basin was dominated by small patches of arable land that covered from 39% of the lands around the Sarzyna at Trzebośnica gauge to 87% of lands surrounding the Szreniawa (Biskupice gauge). Dominant type of forest was coniferous one, covering over 20% of land in seven of all the investigated catchments. In case of catchments chosen for the analysis the fluvisols, euteric cambisols and luvisols are dominated. Fluvisols dominates in catchments of Dłubnia, Biała Tarnowska, Szreniawa, Łęg and Trzebośnica and covers about 20% of its area. In case of catchments of Skawa, Łososina and Soła euteric cambisols dominates. Luvisols covers about 70% of the catchments area of Czarna catchment.

For the purpose of cluster analysis determination, all catchments characteristics were standarized, in case of which a mean is equal 0 and variation is equal 1.

Table 1. Statistical summary of basic catchment characteristics

Lp.	Stream gauge	River	Water-course length [km]	Catchment area [km^2]	q_{95} [$dm^3 \cdot s^1 \cdot km^2$]	Gauge zero level [m a.s.l.]
1	Dwikozy	Opatówka	38.2	256.0	1.32	141.39
2	Kępie Zaleszańskie	Łęg	51.0	822.0	1.21	144.36
3	Harasiuki	Tanew	72.0	2034.0	2.80	165.54
4	Sarzyna	Trzebośnica	24.8	249.0	2.27	164.06
5	Raków	Czarna	17.0	221.0	2.52	219.05
6	Mniszek	Biała Nida	34.8	439.0	2.61	217.72
7	Biskupice	Szreniawa	56.8	682.0	3.00	180.61
8	Zesławice	Dłubnia	42.4	264.0	2.65	208.10
9	Koszyce	Biała Tarnowska	64.4	957.0	1.89	190.73
10	Jakubkowice	Łososina	33.2	343.0	2.51	247.31
11	Osielec	Skawa	17.6	244.0	3.48	393.64
12	Cięcina	Soła	28.4	413.0	4.10	381.63
13	Morawica	Czarna Nida	32.4	755.0	1.91	223.33
14	Bielsko	Biała	8.8	70.3	5.54	316.46
15	Topoliny	Ropa	40.6	970.0	2.16	224.79
16	Grabiny	Grabinianka	24.8	180.0	2.44	187.68
17	Koprzywnica	Koprzywianka	38.6	499.0	1.50	259.10
18	Wilkowa	Wschodnia	40.6	650.0	1.08	166.32
19	Nowy Sącz	Łubinka	11.2	66.3	2.10	281.33

The measure of similarity adopted for the calculations was squared Euclidean distance that allows for evaluation of a distance between objects described by selected parameters. This measure is often used for object classification due to simple mathematical properties and convenient graphical interpretation.

In the next step, homogenous regions were determined. For this purpose nonhierarchical cluster analysis (K-means) was used. In case of this method, the procedure is running as follow: at first, objects (catchments) at random are assigned to cluster, and then iterative are moved between clusters, so that to minimized intra-group changeability and maximized intergroup changeability.

The last step involved development of the models reflecting correlation and regression relationships. Regional regression is built as a multiple regression (1), representing the relationship between low flow (dependent variable) and morphoclimatic parameters (independent variables). It is used to identify the parameters that most strongly shape the low flow. To determine the power of regression equation, coefficient of determination R^2 for the level of significance 0.05 was calculated. The best results were obtained while using stepwise regression:

$$q_{95} = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_{p-1} \cdot x_{p-1} \quad (1)$$

where: x_i – morphoclimatic parameters of a catchment,
β_i – regression coefficient.

Additionally, the regression model fit was tested against including an uncontrolled catchment into the investigated region.

RESULTS

In case of K-means method, the procedure of catchments assign to the clusters was started from two clusters and finished at five. The next moving and assigning of catchments into clusters resulted in a cluster in which there was only one catchment (for five clusters). Another object's delineation did not give an objective effects, based on which it was difficult to determine a clear criterion of assigning each catchment into clusters. The best fit was obtained in case of choosing two clusters. In the first cluster there were five catchments (Biała, Skawa, Łososina, Łubinka and Soła), and in the second, another eleven (Fig. 2).

The first cluster was formed according to the greatest values of mean precipitation (P>845 mm), median altitude (H_{me}>400 m a.s.l.), mean catchment slope (I>0.028) and watercourse staff gauge zero (P_z>240 m). The catchments in this cluster were also similar in terms of the watercourse length, as they group the shortest rivers not exceeding 35 km. Unit runoff q_{95} ranged from 3 to 6 $dm^3 \cdot s^{-1} \cdot km^{-2}$. The soil cover included fluvisols of different permeability (about 22%), and acidic cambisols and eutric cambisols based on sandy noncarbonate decomposed sedimentary rocks (>

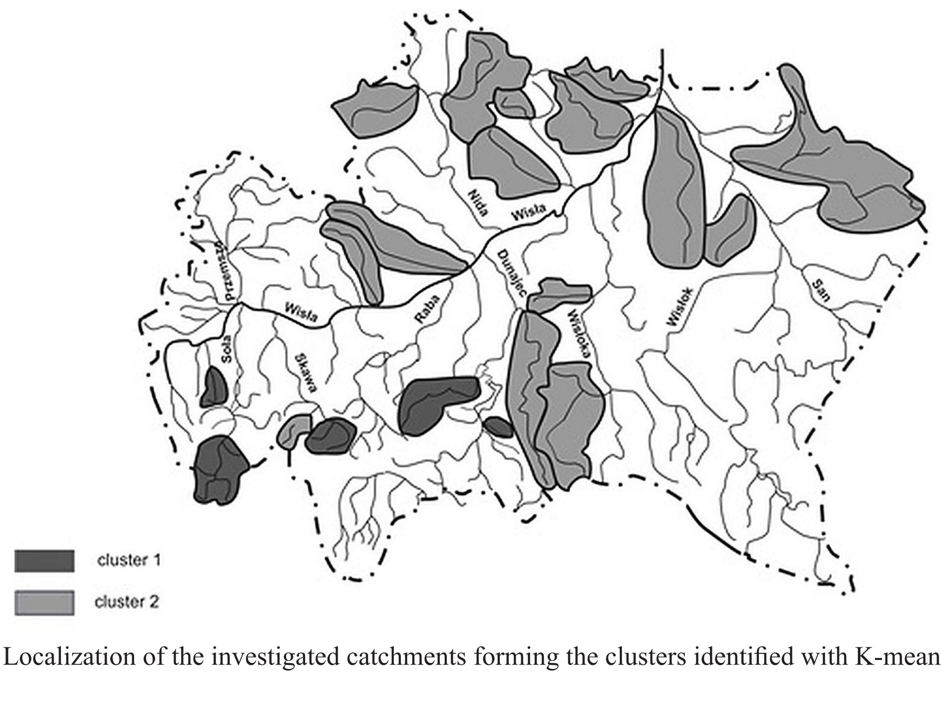

Figure 2. Localization of the investigated catchments forming the clusters identified with K-means method

20%). Another parameter considered for the analyzed cluster was land use. The catchments were similar in terms of the area covered by coniferous forests (>20%) and arable lands (about 55%). In case of catchments in this group, there are high differences of area and small permeability of the ground, what is not favorable for water retention and intensive precipitation cause flash floods. In dry season there are very small flows and often change in water stages during a year. Although, the right side of Carpathian part of the basin is characterized by low unit outflow, but for these catchments it was not stated. It confirms the statement, that for catchments located in the basin area of Skawa, Dunajec the values of low flow are higher, than for catchments located on the East from Dunajec river. The Soła and Skawa rivers are one of the most rich in water rivers in the Upper Vistula basin [Osuch 1991].

The second cluster grouped the other 14 catchments. It included the rivers with the lowest unit runoff q_{95} below 3 $dm^3 \cdot s^{-1} \cdot km^{-2}$, median catchment altitude (H_{me}<370 m a.s.l.), lowest mean catchment slope (I<0.018), watercourse staff gauge zero (P_z<260 m), area covered by coniferous forests (<20%), and precipitation height of up to 850 mm. The catchments were also similar in terms of watercourse length that ranged from 17 to 72 km. Dominant land use in these catchments was arable land that accounted for about 62% of the catchment area, much more than in the first cluster. These catchments are located on area of the Małopolska Upland (on the left side of Upper Vistula basin) and the Carpathian Foothills. The Małopolska Upland has a huge possibility of water infiltration, what compared with smaller amount of precipitation and low river slope allowed for water storage and slower outflow, which the mostly is underground. The Carpathian Foothills (Łęg, Tanew and Trzebośnica) has an indirect character, with dominant role of drainage areas [Osuch 1991]. In this group are catchments with lower diversity of unit outflow.

Correlation and regression relationships were determined separately for each cluster. In the first cluster, comprising a small number of catchments, only the correlations between q_{95} and individual independent variables were determined (Table 2).

The catchments of this cluster were devoid of grassland (U_z), and so they were not accounted for in the correlation model. No significant correlations were found for significance level 0.05. Even though in case of such parameters as median catchment altitude and soils (fluvisols and eutric cambisols) the correlation was weak (correlation coefficient under 0.3), it was decided to show both the correlation model and the correlation coefficient.

In the second cluster, including 14 catchments, a multiple regression model was made as follow:

$$\begin{aligned} q_{95} = -1.096 + 0.016 \cdot U_z + \\ + 0.052 \cdot L_i - 0.02 \cdot S_P + \\ + 0.029 \cdot H_{me} + 0.532 \cdot T - \\ - 106.01 \cdot i - 0.013 \cdot P + 0.036 \cdot S_{BW} \end{aligned} \tag{2}$$

In the equation (2), at a significance level 0.05 statically important was only coniferous forests. In case of grassland, coniferous forests, median catchment altitude, mean annual air temperature and eutric cambisols the relationship had a positive character, what means, that the greater value

Table 2. Correlations in the first cluster

Correlation model	r	p	r^2
q_{95} = 4.40 – 0.01 * A	-0.32	0.10	0.01
q_{95} = 4.40 – 0.04 * L	-0.32	0.60	0.10
q_{95} = -3.59 + 0.008 * P	0.64	0.25	0.40
q_{95} = -1.14 + 0.72 * T	0.49	0.39	0.24
q_{95} = 2.19 + 0.003 * H_{me}	0.21	0.73	0.04
q_{95} = 1.37 + 43.36 * i	0.79	0.11	0.63
q_{95} = 0.42 + 0.01 * P_z	0.45	0.45	0.20
q_{95} = 3.15 + 0.02 * S_M	0.10	0.87	0.01
q_{95} = 3.65 – 0.002 * S_{BW}	-0.04	0.95	0.002
q_{95} = 3.81 – 0.26 * S_P	-0.42	0.95	0.18
q_{95} = 2.73 + 0.04 * U_{Li}	0.61	0.27	0.38
q_{95} = 4.56 – 0.07 * U_{LM}	-0.87	0.06	0.75
q_{95} = 8.68n- 0.09 * U_{Go}	-0.73	0.09	0.68

of each analyzed parameters the greater value of low flow. Other parameters included in the equation had negative character, what means that the value of low flow increase as the value of these parameters decrease. Additionally, fitting of the model in case of including an uncontrolled catchment into region was checked (Fig. 3).

Values of q_{95} calculated on the basis of the regional regression correspond in 81% to the observed values (R^2=0.81), for the significance level 0.05. Catchments, which were the nearest to the diagonal line characterized the best fit of the model. It was catchments Opatówka, Wschodnia, Koprzywianka, Biała Tarnowska, Ropa, Grabinianka, Czarna, Biała Nida and Dłubnia, in case of which flows, calculated on the basis of equation (2), had almost the same value as the observed one. The least fitting was in case of other rivers, but the difference between observed and calculated flows was small and was about 0.5 $dm^3 \cdot s^{-1} \cdot km^{-2}$.

CONCLUSIONS

The presented calculations and their analysis are the initial research for regional regression model with use K-means method, in spite of low flow. The analysis was made for 19 catchments, located in the Upper Vistula basin, for which at least 10-years long observational daily flows were obtained. Because different results could be obtained, when catchments with different physiographic characteristics will be taken, so in the further research many more catchments needs to be taken under consideration, which are diversified according to their location, land use and morphoclimatic parameters influenced on low flow. Selected catchments differ in regards of physiographic and meteorological parameters. The use of K-means method allowed for identification of two groups of catchment with similar value of low flow. The identified clusters comprised catchments similar in terms of unit runoff, watercourse length, mean precipitation, median altitude, mean catchment slope, watercourse staff gauge zero, the area covered by coniferous forests, arable lands, and soils. With the use of physiographic and meteorological features the approximate value of specific low flow discharge can be calculated in case of ungauged catchments. The K-mean method also could be used for grouping of watersheds, according to hydrological characteristics.

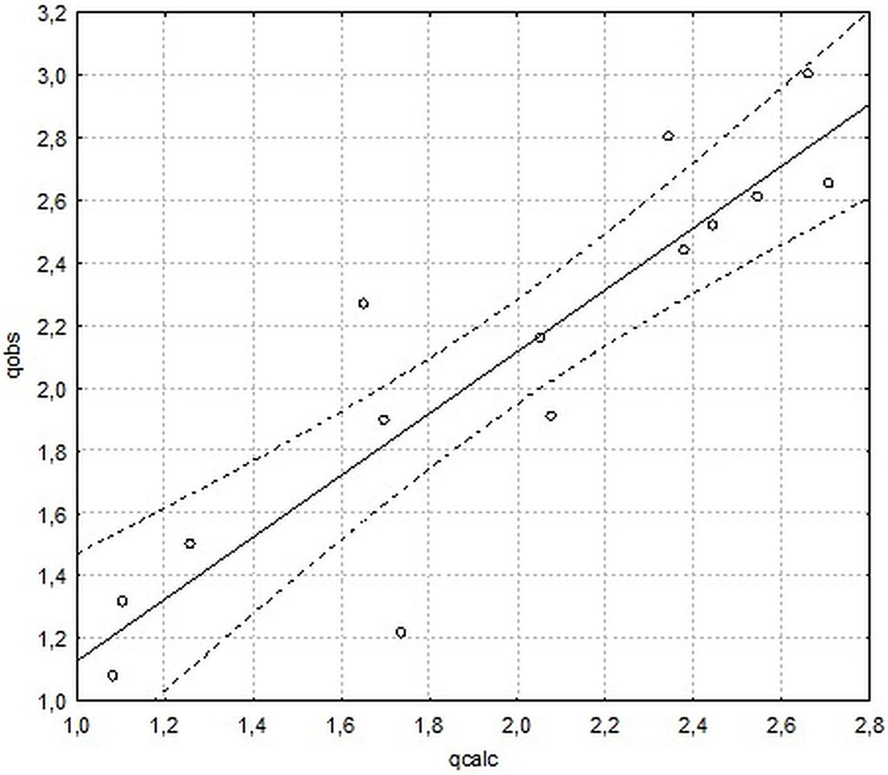

Figure 3. The scatter graph of observed and calculated q_{95} value

REFERENCES

1. Bryndal T. 2011. Identyfikacja małych zlewni podatnych na formowanie gwałtownych wezbrań (na przykładzie Pogórza Dynowskiego, Strzyżowskiego i Przemyskiego). Przegląd Geograficzny 83, 1, s. 27–49.
2. Burn D.H. 1989. Cluster analysis as applied to regional flood frequency – analysis. Journal of Hydrology, 104 (1–4), 345–361.
3. Burn, D. H., Goel, N. K. 2000. The formation of groups for regional flood frequency analysis. Hydrol. Sci. J.45(1), 97–112.
4. Chełmicki W. 1991. Położenie, podział i cechy dorzecza [W:] Dorzecze górnej Wisły pod red. Dynowskiej I. i Maciejewskiego M., PWN Kraków.
5. Corduas M. 2011. Clustering streamflow time series for regional classification. Journal of Hydrology 407, 73–80
6. Dębski K. 1961. Charakterystyka hydrologiczna Polski. Łódź-Warszawa, PWN.
7. Dynowska I. 1971. Typy reżimów rzecznych w Polsce, Zeszyty Naukowe UJ, Prace Geogr., 28, Prace IG UJ 50, Kraków, 155.
8. Gottschalk, L. 1985. Hydrological regionalization of Sweden. Hydrological Sciences Journal, 30:1, 65–83.
9. Gutry-Korycka M., Woronko D., Suchożebrski J. 2009. Uwarunkowanie regionalne maksymalnych prawdopodobnych przepływów rzek polskich. Prace i Studia Geograficzne 43, 25–48.
10. Hydrologic Yearbook for the Vistula basin. Vistula basins and rivers of Przymorze on the east from Vistula river, 1963–1983, Publishing of Transport and Communication IMGW, Warsaw.
11. 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.
12. Kahya E., Demirel M.C., Bég O.A. 2008. Hydrologic homogeneous regions using monthly streamflow in Turkey. Earth Sci. Res. J. vol. 12, No. 2, 181–193.
13. Kowalczak P. 1986. Metoda typologii hydrograficznej niehierarchiczną analizą skupień (na przykładzie dorzecza górnej Noteci) [W:] Niektóre problemy metodyczne w hydrologii pod red. Z. Mikulskiego. Dokumentacja Geograficzna. Instytut Geografii i Przestrzennnego Zagospodarowania, 2, PAN, 38049.
14. Laaha G. 2002. Modelling summer and winter droughts as a basis for estimating river low flows. FRIEND 2002–Regional Hydrology: Bridging the Gap between Research and Practice (Proceedings of the Fourth International FMIiND Conference held at Cape Town. South Africa. Mardi 2002). IAI IS Publ. no. 274.
15. Laaha G. and Blöschl G. 2006. A comparison of low flow regionalization methods-catchment grouping. Journal of Hydrology 323, 193–214.
16. Lecce S.A. 2000. Spatial variations in the timing of annual floods in the southeastern United States. Journal of Hydrology. 235, 151–169.
17. Lin G.F., Chen L.H. 2006. Identification of homogenous regions for regional frequency analysis using the self-organizing map. Journal of Hydrology 324, 1–9.
18. Osuch B. 1991. Reżim odpływu powierzchniowego [W:] Dorzecze górnej Wisły pod red. Dynowskiej I. i Maciejewskiego M., PWN Kraków.
19. Pociask-Karteczka J. 1995. Założenia metodyczne regionalizacji hydrologicznej na przykładzie dorzecza górnej Wisły. Rozpr. habil. UJ, nr 291.
20. Rao A.R., Srinivas V.V. 2006a. Regionalization of watersheds by fuzzy cluster analysis. Journal of Hydrology, 31(1–4), 37–56.
21. Rao A.R., Srinivas V.V. 2006b. Regionalization of watersheds by hybrid cluster analysis. Journal of Hydrology, 31(1–4), 57–79.
22. Rao A.R., Srinivas V.V. 2008. Regionalization of Watersheds. An approach based on cluster analysis, Springer.
23. Rutkowska, A., Żelazny, M., Kohnová, S. et al. 2016. Regional L-Moment-Based Flood Frequency Analysis in the Upper Vistula River Basin, Poland. Pure Appl. Geophys. pp 1–21, doi:10.1007/s00024–016–1298–8.
24. Srinivas V.V. 2009. Regionalization of watersheds using soft computing techniques. ISH Journal of Hydraulic Engineering, vol. 15, No SP. 1:170–193.
25. Stachý J. (red). 1987. Atlas Hydrologiczny Polski, t. I, Instytut Meteorologii i Gospodarki Wodnej, Wydawnictwo Geologiczne, Warszawa.
26. Števková A., Sabo M., Kohnová S. 2012. Pooling of low flow regimes using cluster and principal component analysis, Slovak Journal of Civil Engineering, vol. XX, 19.
27. Wałęga A., Krzanowski S., Chmielowski K. 2009. Wykorzystanie metody analizy skupień do identyfikacji jednorodnych zlewni pod względem indeksów powodziowości i wybranych charakterystyk fizjograficznych. Infrastruktura i Ekologia Terenów Wiejskich 6, 67–81.
28. Ziemońska Z. 1973. Stosunki wodne w Polskich Karpatach Zachodnich, Prace Geogr. PAN, 103, pp. 124.