ISSN 1644-0765
DOI: http://dx.doi.org/10.15576/ASP.FC/2017.16.1.53
www.formatiocircumiectus.actapol.net/pl/

ACTA SCIENTIARUM POLONORUM

Acta Sci. Pol. Formatio Circumiectus 16 (1) 2017, 53–63

CLUSTER ANALYSIS IN DETERMINATION OF HYDROLOGICALLY HOMOGENEOUS REGIONS WITH LOW FLOW

Agnieszka Cupak, Andrzej Wałęga, Bogusław Michalec
University of Agriculture in Krakow

Abstract. In the paper results of evaluation of using statistical methods for data agglomeration were made. For determination of hydrological homogeneous regions characterized by low flow and selected physiographic and meteorological features of the catchments cluster analysis was used. The study material included daily flows from the multi-year period of 1963–1983 collected for 15 catchments located in the upper Vistula basin. 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. Then, $Q_{95\%}$ was subsequently standardized by the catchment area and resulting specific low flow discharges q_{95} $dm^3 \cdot s^{-1} \cdot km^{-2}$. The calculations and their analysis showed that, statistical methods of data agglomeration may be used for determination of hydrologically homogeneous regions. And with use Ward's method with cut-off level at 40% two homogeneous regions were identified.

Key words: low flow, cluster analysis, regionalization

INTRODUCTION

Low flows, their regime and effects on biological functioning and stability of aquatic ecosystems constitute an important issue in hydrology [Števková et al. 2012]. River flow is a result of complex natural processes occurring in a catchment. Supply of the entire system depends mainly on precipitation, while storage and flow depend on physiographic parameters of a catchment. Natural factors that affect the system of low river flow include the type of soils, their infiltration and hydraulic properties, aquifer level, rate, frequency and amount of supply, evaporation rate within the area of the basin, distribution of vegeta-

Corresponding authors – Adres do korespondencji: Agnieszka Cupak, Andrzej Wałęga, Katedra Inżynierii Sanitarnej i Gospodarki Wodnej, Bogusław Michalec, Katedra Inżynierii Wodnej i Geotechniki, Uniwersytet Rolniczy w Krakowie, al. Mickiewicza 24-28, 30-059 Kraków; e-mail: a.cupak@ur.krakow.pl, a.walega@ur.krakow.pl, rmmichbo@cyf-kr.edu.pl.

tion, topography, and climate [Smakhtin 2001, Ziernicka-Wojtaszek and Kaczor 2013]. Information on low flow with specific guarantee is indispensable for the management of water resources, e.g. for determination of energy production in hydroelectric power stations, design of water intakes, fish farming systems, irrigation systems, biological flow and for determination of the amount of contamination introduced into a watercourse. According to the provisions of Water Framework Directive, determination of the ecological status of a watercourse requires evaluation of its low flow [Bodulski et al. 2005, Laaha and Blöschl 2006, Kaznowska and Banasik 2009, Mamun et al. 2010]. The methods for low flow evaluation depend on available catchment documentation. In uncontrolled catchments with no documentation, low flow may be determined by applying such methods as regional regression, regional curves, regional maps and other methods of spatial interpolation of low flow parameters or by calculating the low flow using a series of flows obtained e.g. from a modeling method [Smakhtin 2001].

Regionalization of flow parameters usually assumes, that catchments similar in terms of climate, geology, topography, plant cover and soils respond in a similar way to the flow, e.g. to unit runoff from the catchment, mean monthly distribution of flows, duration of some flow periods, frequency and rate of high and low flow in catchments of similar size, etc. The investigated areas may be delineated based on geographical, administrative of physiographic boundaries. Homogeneous regions may be treated as collections of catchments similar in terms of their hydrological response but not necessarily in geographical features. Developing separate regression models for each homogeneous region may improve the predictive power of the final regression equation. Catchment grouping may be based on standard flow parameters determined from available documentation. Catchment grouping is often performed by means of multivariate statistical analysis [Smakhtin 2001]. One of the methods commonly used in hydrology is cluster analysis. Jain and Dubes [1988] divided cluster analysis methods into two basic categories, i.e. hierarchical and partitional methods. Hierarchical methods allow for the generation of agglomeration groups, while partitional methods may generate any number of agglomerations. Algorithms represented by the agglomeration method include: single-link (nearest-neighbor) cluster, complete-link (furthest neighbor) cluster, average-link, and Ward's method. The Ward's method is the data agglomeration method most often used in hydrology and climatology. It is different from other methods as the distance between clusters is estimated by analysis of variance. This method helps to minimize the sum of squared deviations of two clusters that can be formed at any stage [Rao and Srinivas 2008].

The aim of the research was to evaluate the possibility of using statistical methods for data agglomeration, i.e. cluster analysis for determination of hydrological homogeneous regions characterized by low flow and 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 15 catchments (Fig. 1), located in the upper Vistula basin and characterized by specific physiographic and meteorological parameters (Table 1). The following physiographic and meteorological parameters of the catchment were investigated: length of the watercourse (L), catchment area (A), mean annual air temperature (T), mean annual precipitation (P), mean catchment slope (i), median catchment altitude (H_{me}), watercourse staff gauge zero ($P.z.$), land use (U), and soil use (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.

The research included 15 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 absin 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 upper Vistula basin is characterized by diverse geological structure. The main regional units include Carpathians, part of the pre-Carpathian Depression, parts of Lublin and Upper Silesia basins, Cracovian Monocline, Nida Basin and Holy Cross Mountains [Więcławik 1991]. The Tatra Mountains, the Podhale and flisch Carpathians are dominated by weathering-exposed clay soils (less often sandy loam soils), of more or less skeletal character. In the eastern Tatra Mountains the soil cover is formed by moderately and highly skeletal loamy podzols. Western parts of the Tatra Mountains are covered with highly skeletal limestone soils. The Bieszczady Mountains and the Low Beskids are dominated by more compact silty loam and clay-loam soils. They are usually moderately and poorly skeletal eutric cambisols (sometimes silt calcisols), with usually gleyic deeper layers. The Carpathian Foothills are covered mainly by silty-clay luvisols and ecologically similar eutric cambisols. The areas of Sandomierz Basin and Oświęcim Basin include mainly Pleistocene sands and loamy sands and accompanying sands over loam and tills. The flood plains of the Vistula and its tributaries are dominated by alluvial deposits [Adamczyk 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 Pszczynka catchment at Międzyrzecze gauge and the Łęg at Kępa 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, Kępa 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 (Table 1).

Fig. 1. Location of the analyzed catchments within the upper Vistula basin
Ryc. 1. Lokalizacja analizowanych zlewni na obszarze dorzecza górnej Wisły

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.

The next step involved determination of homogeneous areas (cluster agglomeration) based on the correlation between the catchment parameters and unit runoff q_{95}. This was performed using a hierarchical cluster analysis method, i.e. Ward's method. In this method, estimation of the distance between clusters is carried out using analysis of variance. It consists in minimizing the sum of squared deviations for any two clusters that may be formed at any stage. This method is deemed to be very effective, even though it yields small clusters. It is a commonly used regionalization technique in hydrology and meteorology [Stanisz 2007].

There is no clear formal criterion for determining clusters based on a dendrogram. Therefore, a chart illustrating agglomeration course was used as a criterion for stopping the agglomeration process and final verification of the objects (individual catchments) on a dendrogram.

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

Table 1. Statistical summary of catchment characteristics
Tabela 1. Charakterystyki statystyczne analizowanych zlewni

Variable Zmienna	Variable description Charakterystyka	Units Jednostka	Min. Minimum	Mean Średnia	Max. Maksimum
A	Catchment area – Powierzchnia zlewni	km^2	70.3	495.6	2034.0
L	Watercourse length – Długość cieku	km	8.8	43.9	95.2
T	Mean annual air temperature Średnia roczna temperatura powietrza	°C	5.8	7.2	8.0
P	Mean annual precipitation Średni roczny opad	mm	603.8	770.4	1113.0
I	Mean catchment slope Średni spadek zlewni	–	0.002	0.020	0.091
H_{me}	Median catchment altitude Średnia wysokość zlewni	m a.s.l.	202.0	344.6	720.0
P_z	Gauge zero level Poziom zera wodowskazu	m a.s.l.	141.4	229.0	393.6
U_{Li}	Coniferous forests – Lasy iglaste	%	0.0	18.9	59.0
U_{LM}	Mixed forests – Lasy mieszane	%	0.0	8.8	34.0
U_{UZ}	Grassland – Łąki	%	0.0	8.6	30.0
U_{Go}	Arable land – Grunty orne	%	39.0	60.1	87.0
U_{OZ}	Built-up areas – Obszary zabudowane	%	0.0	3.5	29.0
S_M	Fluvisols – Mady	%	0.0	15.9	34.0
S_B	Cambisols – Gleby brunatne właściwe	%	0.0	13.3	83.0
S_{BW}	Eutric Cambisols Gleby brunatne wyługowane	%	0.0	18.3	76.0
S_P	Luvisols – Gleby płowe	%	0.0	21.8	67.0

(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:

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

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.

STUDY RESULTS

The use of Ward's method (Fig. 2) for determination of hydrologically homogeneous regions resulted in agglomeration of the investigated catchments into clusters.

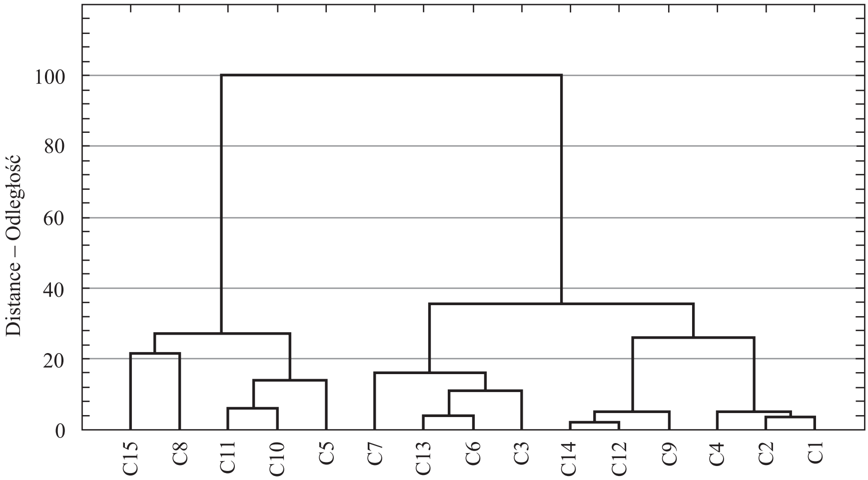

Fig. 2. Agglomeration with Ward's method
Ryc. 2. Aglomeracje metodą Warda

The cut-off level was determined from the graph depicting the course of the agglomeration by assuming that it falls at the point of the highest pitch between the bonds, at 40% of the bond distance.

Localization of the investigated catchments forming the clusters identified with Ward's method is presented in Figure 3. The first cluster was formed by the catchments of the Wieprzówka, the Biała, the Skawa, the Łososinka, and the Soła, and it was characterized by 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 > 230 m). The catchments in this cluster were also similar in terms of the watercourse length, as they group the shortest rivers not exceeding 28 km. Unit runoff q_{95} ranged from 3 to 5 $dm^3 \cdot s^{-1} \cdot km^{-2}$. Cambisols and gravel-based eutric cambisols were absent in the catchments belonging to this cluster. The soil cover included fluvisols of different permeability (about 22%), and acidic cambisols and eutric cambisols based on sandy noncarbonate decomposed sedimentary rocks (> 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%). The second cluster grouped the other 10 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} < 350 m a.s.l.), lowest mean catchment slope (I < 0.014), watercourse staff gauge zero (P_z < 230 m), area covered by coniferous forests (< 20%), and precipitation height of up to 800 mm. The catchments were also similar in terms of watercourse length that ranged from 28 to 95.2 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. Cambisols and gravel-based eutric cambisols covered about 20% of the cluster catchments, fluvisols of different permeability accounted for < 20%, and there were no acidic cambisols or eutric cambisols based on sandy noncarbonate decomposed sedimentary rocks.

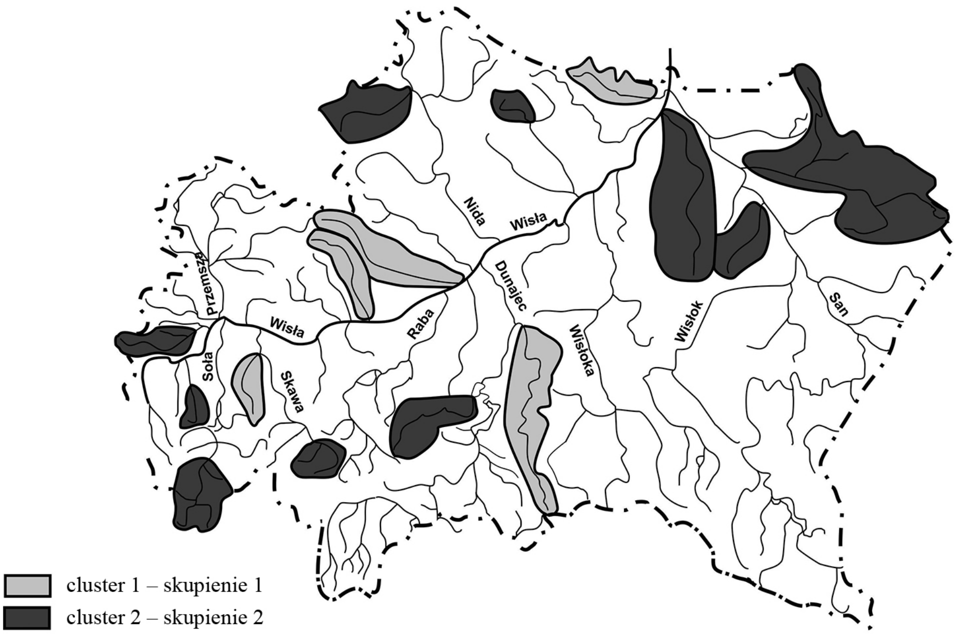

Fig. 3. Localization of the investigated catchments forming the clusters identified with Ward's method

Ryc. 3. Rozmieszczenie analizowanych zlewni tworzących skupienia określonych za pomocą metody Warda

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 cambisols and gravel-based eutric cambisols, and so they were not accounted for in the correlation model. No significant correlations were found for significance level 0.05. However, when the significance level was increased to 0.10, a significant correlation was found for the percentage of the area covered by fluvisols and arable lands. The best fit of the correlation model for the significance level of 0.05 was achieved for the relationship between q_{95} and percentage of the area covered by fluvisols of different permeability and arable lands ($R^2 > 0.7$). Negative correlation achieved in both cases indicated that the smaller the area covered by fluvisols and arable lands the higher the low flow. Correlation coefficient was also high for mean catchment slope ($r > 0.7$).

Table 2. Correlations in the first cluster [Zych 2014]
Tabela 2. Zależności korelacyjne w pierwszej grupie [Zych 2014]

Correlation model – Model korelacji	r	p	R^2
$q_{95} = 0.14281 - 0.6596 \cdot A$	–0.13	0.83	0.02
$q_{95} = -1.472 - 2.779 \cdot L$	–0.69	0.19	0.48
$q_{95} = -0.0269 + 0.43414 \cdot P$	0.20	0.74	0.04
$q_{95} = 0.60534 + 0.16806 \cdot T$	0.16	0.80	0.03
$q_{95} = -0.3430 + 0.72399 \cdot H_{me}$	0.48	0.42	0.23
$q_{95} = -0.7004 + 1.0566 \cdot I$	0.79	0.11	0.63
$q_{95} = -0.5135 + 0.91477 \cdot P_z$	0.63	0.26	0.39
$q_{95} = 1.5148 - 1.545 \cdot S_M$	–0.87	0.05	0.76
$q_{95} = 0.32305 + 0.13692 \cdot S_{BW}$	0.09	0.88	0.01
$q_{95} = 0.24309 + 0.40396 \cdot U_{Li}$	0.36	0.55	0.13
$q_{95} = 0.20585 - 1.496 \cdot U_{Go}$	–0.86	0.06	0.74

In the second cluster comprising 10 catchments, the multiple regression model looked as follows:

$$q_{95} = -1{,}726 + 19{,}063 \cdot I - 0{,}052 \cdot S_M + 0{,}0083 \cdot P - 0{,}0067 \cdot H_{Me} + 0{,}0044 \cdot P_z \quad (2)$$

In the formula (2), the parameters significant for the significance level 0.05 included mean catchment slope, fluvisols, precipitation, and median altitude of the catchment. Positive correlations were obtained for mean catchment slope, watercourse staff gauge zero and precipitation, which indicated that the greater value of these parameters the greater the low flow. The other morphoclimatic parameters included in the formula had negative sign, which means that their decrease was accompanied by an increase in the low flow. Additionally, the model fit was verified against an incorporation of an uncontrolled catchment into the region (Fig. 4).

Values of q_{95} calculated from the regional regression equation revealed a 96% match with the observed values (R^2 = 0.96) for the significance level 0.05. The model fit was the best for the catchments closest to the diagonal line. They included the catchments of the Tanew, the Biała, the Pszczynka, the Skawa, the Biała Nida and the Soła, where the flows calculated from the equation (2) were the same as those observed. The lowest fit was achieved for the catchments of the Łososinka and the Trzebośnica, but the difference between calculated and observed flow was small and amounted to about 0.2 $dm^3 \cdot s^{-1} \cdot km^{-2}$.

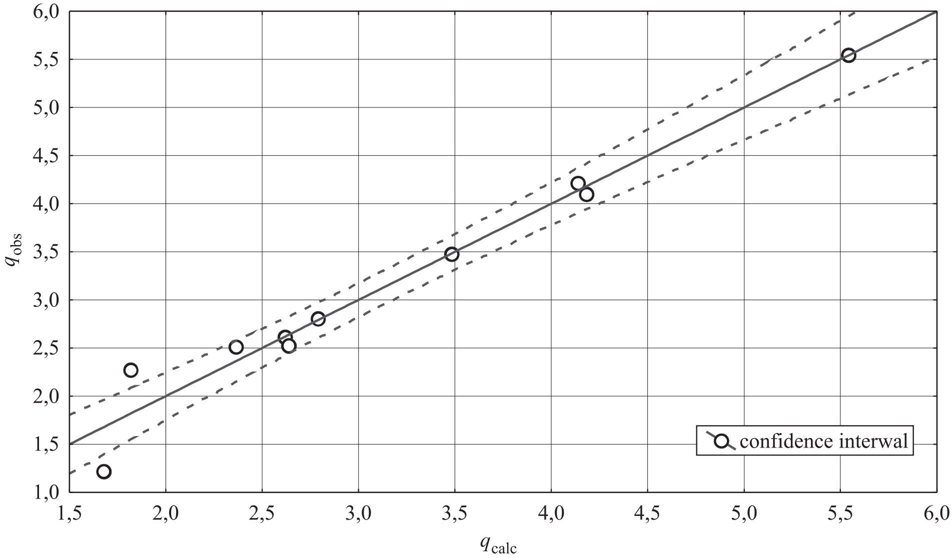

Fig. 4. Differences between observed and calculated q_{95}, $dm^3 \cdot s^{-1} \cdot km^{-2}$

Ryc. 4. Wykres rozrzutu wartości q_{95}, $dm^3 \cdot s^{-1} \cdot km^{-2}$, obserwowanych do obliczonych

CONCLUSIONS

The presented calculations and their analysis constituted the basis for drawing the following conclusions:

- Statistical methods of data agglomeration may be used for determination of hydrologically homogeneous regions.
- The use of Ward's method with cut-off level at 40% allowed for identification of two homogeneous regions.
- 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.
- The best fit of the correlation model for the significance level 0.05 in the first cluster was achieved for the relationship between q_{95} and percentage of the area covered by fluvisols of different permeability and arable lands ($R^2 > 0.7$).
- The regression equation for the second cluster included such parameters as mean catchment slope, fluvisols, precipitation, median catchment height, and watercourse staff gauge zero.
- The values of the unit runoff calculated from the regional regression equation matched in 96% with the observed values, at a significance level 0.05 ($R^2 = 0.96$).

REFERENCES

Adamczyk, B. (1991). Gleby. [W:] Dorzecze górnej Wisły, Red. I. Dynowska, M. Maciejewski. PWN, Kraków.

Bodulski, J., Ciepielowski, A., Dąbkowski, S.L., Głogowska, E. (2005). Dynamika zmian natężenia przepływu wód powierzchniowych w zlewniach użytkowanych rolniczo i leśnie. Acta Sci. Pol., Formatio Circumiectus, 4(1), 3–24.

Chełmicki, W. (1991). Położenie, podział i cechy dorzecza. [W:] Dorzecze górnej Wisły. Red. I. Dynowska, M. Maciejewski. PWN, Kraków.

Jain, A.K., Dubes, R.C. (1988). Algorithms for clustering data. Prentice Hall, Englewood Cliff (NJ).

Laaha, G., Blöschl, G. (2006). A comparison of flow regionalization methods – catchment grouping. J. Hydrol., 323, 193–214.

Mamun, A.A., Hashim, A., Daoud, J.I. (2010). Regionalization of low flow frequency curves for Peninsular Malaysia. J. Hydrol., 381,174–180.

Kaznowska, E., Banasik, K. (2009). Ocena intensywności niżówek rzecznych w małej zlewni rolniczej Niziny Mazowieckiej w ostatnich 45 latach. Acta Sci. Pol., Formatio Circumiectus, 8(3–4), 5–16.

Rao, A.R., Srinivas, V.V. (2008). Regionalization of Watersheds. An approach based on cluster analysis. Springer, New York.

Roczniki Hydrologiczne Wód Powierzchniowych. Dorzecze Wisły i Rzeki Przymorza na Wschód od Wisły (1968). Wydawnictwo Komunikacji i Łączności, IMGW, Warszawa.

Smakhtin, V.U. (2001). Low flow hydrology: a review. J. Hydrol., 240, 147–186.

Stachý, J. (red.) (1987). Atlas Hydrologiczny Polski. T. I. IMiGW, Wydawnictwo Geologiczne, Warszawa.

Stanisz, A. (2007). Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny.T. 3. Analizy wielowymiarowe. StatSoft, Kraków.

Števková, A., Sabo, M., Kohnová, S. (2012). Pooling of low flow regimes using cluster and principal component analysis. Slovak J. Civil Engin., 20, 19.

Więcławik, S. (1991). Budowa geologiczna. [W:] Dorzecze górnej Wisły. Red. I. Dynowski, M. Maciejewski. PWN, Kraków.

Ziernicka-Wojtaszek, A., Kaczor, G. (2013). Wysokość i natężenie opadów atmosferycznych w Krakowie i okolicach podczas powodzi w okresie maj-czerwiec 2010. Acta Sci. Pol., Formatio Circumiectus, 12(2), 143–151.

Zych, N. (2014). Zastosowanie metod analizy skupień do wyznaczenia regionów hydrologicznie homogenicznych. Maszynopis, UR, Kraków.

ZASTOSOWANIE METOD ANALIZY SKUPIEŃ DO WYZNACZENIA REGIONÓW HYDROLOGICZNIE HOMOGENICZNYCH ZE WZGLĘDU NA PRZEPŁYW NISKI

Streszczenie. W pracy określono możliwość zastosowania statystycznej metody grupowania danych – analizy skupień, w celu określenia regionów hydrologicznie homogenicznych ze względu na wielkość przepływu niskiego oraz parametry meteorologiczne i fizjograficzne zlewni. Założony cel zrealizowano w oparciu o przepływy dobowe pochodzące z wielolecia 1963–1983, dla 15 wybranych zlewni zlokalizowanych w dorzeczu górnej Wisły. W analizie uwzględniono przepływ jednostkowy q_{95}, jest przepływ który jest osiągnięty prze 95% dni w rozpatrywanym okresie. Przepływ ten jest powszechnie wykorzystywaną w Europie charakterystyką i zostały wybrany ze względu na jego wielorakie zastosowanie w zarządzaniu zasobami wodnymi. Następnie został odniesiony do powierzchni zlewni i określono przepływ niski przypadający na powierzchnię zlewni q_{95} $dm^3 \cdot s^{-1} \cdot km^{-2}$. Na podstawie przeprowadzonej analizy wyodrębniono dwa skupienia. Przeprowadzone obliczenia wykazały, że statystyczna metoda aglomeracji danych może być wykorzystana do wyznaczenia obszarów hydrologicznie homogenicznych. Na podstawie metody Warda, na poziomie odcięcia 40%, wyznaczono dwa homogeniczne regiony.

Słowa kluczowe: przepływ niski, analiza skupień, regionalizacja

Accepted for print – Zaakceptowano do druku: 1.02.2017

For citation: Cupak, A., Wałęga, A., Michalec, B. (2017). Cluster analysis in determination of hydrologically homogeneous regions with low flow. Acta Sci. Pol., Formatio Circumiectus, 16(1), 53–63.