The use of hierarchical cluster analysis for grouping atmospheric precipitation in Poland

Agnieszka Cupak^{1,}, Agnieszka Ziernicka-Wojtaszek¹, Joanna Krużel¹, Krzysztof Józwiakowski², Andrzej Wałęga¹, Antoni Grzywna², Andrzej Bochniak², Radomir Obroślak², Andrzej Mazur², and Artur Serafin²*

¹University of Agriculture in Krakow, Al. Mickiewicza St. 24/28, 30-059 Kraków, Poland

²University of Life Science in Lublin, Leszczyńskiego St. 7, 20-069 Lublin, Poland

Abstract. The aim of this study is to present the application of statistical methods to assess the possibility of grouping precipitation according to their elevation above sea level and precipitation levels in the temperate climate of Poland. The country was divided into regions with similar levels of precipitation using cluster analysis by Ward's method. The study was performed with meteorological data on average monthly precipitation of 53 meteorological stations from 1981 to 2010. The selection of stations was dictated by the need to consider the variability in amounts of annual precipitation throughout the country. The data was used to calculate average annual precipitation total as well as average precipitation total in the summer half-year (from May to October) and the winter half-year (from November to April) for Poland, from 1981 to 2010. Statistical method was performed and six clusters were generated, for the elevation of the stations and average annual precipitation as well as average precipitation for hydrological half-years. The average annual precipitation for the clusters ranged from 530 mm (cluster 5) to 820 mm (cluster 2). The average precipitation for the winter hydrological half-year ranged from 190.1 mm (cluster 1) to 288.8 mm (cluster 5). The average precipitation for the summer hydrological half-year ranged from 326.3 mm (cluster 5) to 605 mm (cluster 2). The conclusions from the analysis carried out in the study are that: (1) grouping by means of Ward's method can be used to distinguish homogeneous areas with the same levels of precipitation; (2) both precipitation and the elevation, at which meteorological stations are located, are the basis for distinguishing clusters in Ward's method.

1 Introduction

Urbanization processes exert multiple pressures on the hydrologic cycle. Specifically, increases in impervious surface result in increased hydraulic efficiency in urban catchments, and can substantially decrease the capacity for a given landscape or region to infiltrate precipitation, with a concomitant increase in the production of runoff, shorter times of concentration or lag times and decreased recharge of water tables with a corresponding decline in base flows [1]. O'Driscoll et al. [2] have reported that peak flow in urbanized areas can be 400% higher than in a pre-development catchment.

Statistical methods are used to identify areas that are homogeneous in terms of some examined quality, especially cluster analysis, is commonly applied in hydrometeorology to determine relationships between circulation and water reserves [3-6]. Geographic Information Systems (GIS) may be used also in the concept of the "digital city" [7]. Increasingly, city managers are turning to the collection, archiving and analysis of data for their

urban areas, especially through facilities offered by advanced GIS and remote sensing. GIS maps of areas at risk are valuable information and communication facilities in their own right.

The aim of this work is to present the application of statistical methods to assess the possibility of grouping precipitation stations according to their elevation above sea level and precipitation levels, in the temperate climate of Poland. The country was divided into regions with similar levels of precipitation using cluster analysis by Ward's method.

2 Materials and methods

2.1 Characteristics of the research area

Poland is located in the central part of Europe, in the middle latitudes. Its location is defined by the following geographical coordinates: westernmost point – 14°08'E, easternmost point – 24°09'E, northernmost point – 54°50'N, southernmost point –

*Corresponding author: a.cupak@ur.krakow.pl

49°00'N [8]. It is situated in the temperate zone, which causes substantial variation in weather conditions from month to month and from year to

year. This is due to the frequent and active movement of air masses [9].

Table 1. Precipitation totals at meteorological stations in Poland in the years 1981-2010 and their locations

Number	Station	Elevation a.s.l.*	Precipitation (mm)		
			Annual	Summer half-year	Winter half-year
1.	Białystok	148	576	370	206
2.	Bielsko-Biała	398	970	651	319
3.	Chojnice	164	589	358	231
4.	Częstochowa	293	634	405	230
5.	Elbląg	40	690	428	263
6.	Gorzów Wielkopolski	72	542	313	230
7.	Jelenia Góra	342	689	447	242
8.	Kalisz	138	500	319	180
9.	Katowice	284	723	449	274
10.	Kętrzyn	108	593	380	213
11.	Kielce	260	618	389	230
12.	Kłodzko	356	585	407	178
13.	Koło	116	524	333	191
14.	Koszalin	32	738	448	290
15.	Krakow	237	672	437	235
16.	Legnica	122	515	352	163
17.	Lesko	420	823	559	264
18.	Leszno	91	545	326	219
19.	Lębork	38	700	419	281
20.	Lublin	238	575	365	210
21.	Łódź	187	569	344	226
22.	Mława	147	548	342	206
23.	Nowy Sącz	292	731	509	223
24.	Olsztyn	133	626	380	246
25.	Opole	165	604	394	210
26.	Ostrołęka	94	541	349	192
27.	Piła	72	556	331	225
28.	Płock	106	518	322	197
29.	Poznań	83	521	316	205
30.	Przemyśl	279	650	443	207
31.	Puławy	140	572	366	206
32.	Racibórz	205	602	408	194
33.	Resko	52	721	393	329
34.	Rzeszów	200	653	435	218
35.	Sandomierz	217	543	372	171
36.	Siedlce	152	526	345	181
37.	Słubice	21	574	329	245
38.	Sulejów	188	551	351	200
39.	Suwałki	184	598	375	223
40.	Szczecin	1	559	324	235
41.	Szczecinek	137	618	358	261
42.	Świnoujście	6	571	331	240
43.	Tarnów	209	714	480	234
44.	Terespol	133	520	338	182
45.	Toruń	69	538	345	193
46.	Ustka	6	694	412	281
47.	Warsaw	106	531	340	191
48.	Wieluń	200	590	371	218
49.	Włodawa	177	522	346	176
50.	Wrocław	120	538	351	186
51.	Zamość	212	551	379	172
52.	Zgorzelec	203	617	379	238
53.	Zielona Góra	192	585	338	246

*Elevation a. s. l. from Climate Atlas of Poland [10]

2.2 Materials

The study was carried out with meteorological data on average monthly precipitation of 53 meteorological stations from 1981 to 2010. The data was obtained from the Polish Institute of Meteorology and Water Management. The selection of stations was dictated by the need to consider the variability in amounts of annual precipitation throughout the country. The data was used to calculate average annual precipitation totals as well as average precipitation totals in the summer half-year (from May to October) and the winter half-year (from November to April) for Poland in the years from 1981 to 2010 (Tab. 1).

2.3 Characteristics of the statistical method

Cluster analysis makes possible to divide a set of observations into subsets of similar objects forming groups called clusters. It is assumed that the objects allocated to a given cluster are shown similar qualities and that their characteristics are determined by similar factors. Before the cluster analysis was begun, the data were standardized according to the following formula (1) [11]:

$$Z_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i} \quad (1)$$

$$\bar{x}_i = \frac{\sum_{j=1}^t x_{ij}}{t} \quad (2)$$

$$s_i = \frac{\sum_{j=1}^t (x_{ij} - \bar{x}_i)^2}{t-1} \quad (3)$$

where:

x_{ij} – value for i -th object and j -th characteristic,

$\bar{x}_i$ – mean value for j -th characteristic, determined by formula (2),

s_i – standard deviation for j -th characteristic (eq. 3).

The next step was agglomeration of clusters, taking into account the average precipitation (annual, from the summer half-year, and from the winter half-year) and the station's elevation above sea level. The analysis was performed by Ward's method. The objective function (W) of Ward's algorithm [12] minimizes the sum of squares of deviations of the feature vectors from the centroid of their respective cluster (Eq. 4) [13].

$$W = \sum_{k=1}^K \sum_{j=1}^m \sum_{i=1}^{N_k} (y_{ij}^k - y_j^k)^2 \quad (4)$$

where:

W – the objective function of Ward's algorithm,

K – number of clusters,

N_k – the number of feature vectors in cluster k ,

y_{ij}^k – the rescaled value of attribute j in the feature vector i assigned to cluster k ,

y_j^k – the mean value of feature j for cluster k .

Ward's algorithm begins with a singleton cluster. At this point the cluster centres is the same as the feature centres. Therefore, the value of the objective function is zero. At each step in the analysis, the union of every possible pair of clusters is considered and the two

clusters whose fusion results in the smallest increase in W are merged. The change in the value of the objective function, due to the merger depends only on the relationship between the two merged clusters, and not on the relationships with other clusters [13-15]. In this method, distances between clusters are determined based on the analysis of variance. This consists in minimizing sums of squares of deviations within the clusters. At each stage, from all combinable pairs of clusters the one is selected whose combination gives a cluster of minimum diversity.

Euclidean distance was adopted as a measure of distance between clusters. This metric is often used for classification of objects due to its convenient graphic interpretation and simple mathematical properties [16-18]. It is a simple geometric distance in a multidimensional space, calculated using the following formula (5):

$$distance(x, y) = \{\sum_i (x_i - y_i)^2\}^{1/2} \quad (5)$$

where:

$d(x, y)$ – distance between pair of clusters x and y ,

x_i, y_i – the i -th object for obtained x and y cluster.

The cut-off level adopted to determine the number of clusters was based on the agglomeration chart.

In the final step, maps of the locations of precipitation stations included in each cluster were created.

3 Results

In the case of station elevation and average annual precipitation, a cut-off level of 4 was adopted, which allowed for the generation of six clusters (Fig. 1).

The first cluster included 17 stations. They were similar in terms of the elevation of the stations, which for this cluster ranged from 100 to 240 m above sea level, and the average annual precipitation, was 585 mm. This cluster covers the southern part of the South Baltic Lakelands (Pojezierza Południowobałtyckie), the Central Poland Lowlands (Niziny Środkowopolskie) and Polesie [16] (Fig. 2). The second cluster was formed by two stations, Bielsko-Biała and Lesko, which have the highest elevations (above 390 m a.s.l.) and the highest average annual precipitation, amounting to 820 mm. This cluster covers the western part of the Outer Western Carpathians and the Eastern Beskids (Fig. 2).

The next cluster was formed by 10 stations, with elevations ranging from 200 to 356 m a.s.l. and precipitation of 660 mm. This cluster covers the Sudety Mountains together with Przedgórze Sudeckie, the Silesia-Krakow Upland (Wyżyna Śląsko-Krakowska) and Małopolska Upland (Wyżyna Małopolska), Northern Podkarpacie, and the central part of the Outer Western Carpathians (Fig. 2). The fourth cluster consisted of 5 stations, with elevations ranging from 6 to 52 m a.s.l. and average annual precipitation amounting to 708.6 mm. This cluster covers the South Baltic Coastland (Pobrzeża Południowobałtyckie) except for its western part (Fig. 2). The fifth cluster consisted of 16 stations, whose elevation ranged from 69 to 180 m a.s.l., with average annual precipitation of 530 mm. This cluster

encompasses the northern and southern part of the South Baltic Lakelands, the north-western part of the Sudety Mountains with Przedgórze Sudeckie, the East Baltic Lakelands (Pojezierza Wschodniobałtyckie), the Podlasie and Belarus Plateau (Wysoczyzny Podlasko-Białoruskie), the southern part of the Central Poland Lowlands, and the Lublin-Lviv Upland (Wyżyna

Lubelsko-Lwowska) (Fig.2). The last cluster was formed by three stations (Słubice, Szczecin and Świnoujście), located at the lowest elevations (less than 25 m a.s.l.) and with average annual precipitation of 560 mm. This cluster encompasses the western part of the South Baltic Coastland, lying at the Polish border, and the South Baltic Lakelands (Fig. 2).

Fig. 1. Agglomeration of precipitation stations according to the elevation of the station and average annual precipitation by Ward’s method

Fig. 2. Location of precipitation stations in individual clusters based on the elevation of the station and the average annual precipitation

In the case of agglomeration of stations according to the elevation of the stations and the average precipitation from the summer half-year, with a cut-off level of 4, again six clusters were obtained, as in the case of the average annual precipitation (Fig. 3).

The first cluster was formed by 16 stations whose elevations ranged from 94 to 170 m a.s.l., with average precipitation of 354.7 mm in the summer half-year. This

cluster included the northern part of the South Baltic Lakelands, the north-western part of the East Baltic Lakelands, the eastern part of the Saxon-Lusatian Lowlands (Niziny Sasko-Łużyckie), the Central Poland Lowlands, the Podlasie and Belarus Plateau, and the northern part of Polesie (Fig. 4).

Fig. 3. Agglomeration of precipitation stations according to the elevation of the station and average precipitation from the summer half-year by Ward’s method

Fig. 4. Location of precipitation stations in individual clusters based on the elevation of the station and the average precipitation from the summer half-year

The second, third and fourth clusters were formed by the same stations as in the analysis for the elevation of the stations and the average annual precipitation. The second cluster was formed by two stations, Bielsko-Biała and Lesko, which were located at the highest elevations (above 390 m a.s.l.) and had the highest average precipitation for the summer half-year, amounting to 605 mm. This cluster includes the western part of the Outer Western Carpathians and the Eastern Beskids (Fig. 4). The third cluster was formed by 10 stations, located at elevations from 200 to 356 m a.s.l. and with precipitation of 440.1 mm. This cluster encompasses the Sudety Mountains with Przedgórze Sudeckie, the Silesia-Krakow Upland and Małopolska Upland, Northern Podkarpacie, and the central part of the Outer Western Carpathians (Fig. 4). The fourth cluster was formed by 5 stations, at elevations from 6 to 52 m a.s.l. and with average precipitation for the summer half-year of 420 mm. This cluster covers the South Baltic Coastland except for its western part (Fig. 4). Cluster 5 consists of 9 stations, at elevations from 1 to 106 m a.s.l., and the

average precipitation from the summer half-year was 326.3 mm. This cluster covers the western part of the South Baltic Lakelands, lying at the Polish border, and the South Polish Lakelands except for their northern part (Fig. 4).

The sixth and last cluster was formed by 11 stations, at elevations ranging from 170 to 240 m a.s.l. and with average precipitation for the summer half-year amounting to about 370 mm. This cluster covers the south-western part of the South Baltic Lakelands, the north-western part of the Sudety Mountains with Przedgórze Sudeckie, the southern and central part of the Central Poland Lowlands, the Eastern Beskids, the central part of Polesie, and the eastern part of the East Baltic Lakelands, lying at the Polish border (Fig. 4).

Based on the grouping of stations according to the elevation of the station and the average precipitation for the winter half-year (with a cut-off level of 4), six clusters were again obtained, as in the case of the previous analyses (Fig. 5).

Fig. 5. Agglomeration of precipitation stations according to the elevation of the station and average precipitation for the winter half-year by Ward's method

The first cluster included 20 stations, at elevations from 69 to 217 m a.s.l. and with average precipitation for the winter half-year amounting to 190.1 mm. This cluster encompassed the eastern part of the Saxon-Lusatian Lowlands, the southern part of the South Baltic Lakelands, the Central Poland Lowlands, the Podlasie and Belarus Plateau, Polesie, the northern and eastern part of the Małopolska Upland, and the eastern part of the Lublin-Lviv Upland (Fig. 6). The second cluster was formed by 4 stations, with elevations ranging from 284 to 420 m a.s.l. and with average precipitation for the winter half-year amounting to 275 mm. This cluster

covers the central part of the Sudety Mountains with Przedgórze Sudeckie, the central part of the Silesia-Krakow Upland, the western part of the Outer Western Carpathians, and the Eastern Beskids (Fig. 6).

Another cluster was formed by 10 stations at elevations from 133 to 209 m a.s.l., where the precipitation was 234.1 mm. This cluster covers the north-western part of the Sudety Mountains with Przedgórze Sudeckie, the south-western part of the South Baltic Lakelands, the northern part of the Silesia-Krakow Upland, the central part of the Central Poland Lowlands, and Northern Podkarpacie (Fig. 6). The fourth cluster

was formed by 7 stations, at elevations ranging from 237 to 356 m a.s.l., and the mean precipitation for the winter half-year was 216 mm. This cluster covers the southern part of the Sudety Mountains with Przedgórze Sudeckie, the central part of the Silesia-Krakow Upland, the central and southern part of the Małopolska Upland, the central part of the Outer Western Carpathians, the northern part of the Lublin-Lviv Upland, and the south-eastern part of Northern Podkarpacie (Fig. 6). Then next two clusters consisted of 5 and 7 stations, respectively. In the case of cluster 5, these were stations at elevations ranging from

6 to 52 m a.s.l., with average precipitation for the winter half-year at a level of 288.8 mm. This cluster encompasses the South Baltic Coastland except for its western part (Fig. 6). In the case of cluster 6, the stations were located at the lowest elevations, from 1 to 108 m a.s.l., and the mean precipitation for the winter half-year was 230 mm. This cluster included the western part of the South Baltic Coastland, lying at the Polish border, the central-western part of the South Baltic Lakelands, and the northern part of the East Baltic Lakelands (Fig. 6).

Fig. 6. Location of precipitation stations in individual clusters based on the elevation of the stations and the average precipitation in the winter half-year

4 Conclusions

The following conclusions can be drawn from the analysis carried out in the study:

1. Both precipitation and the elevation at which meteorological stations are located are the basis for distinguishing clusters in Ward's method.
2. Six clusters were generated for the elevation (of the stations) and average annual precipitation as well as average precipitation for hydrological half-years.
3. The average annual precipitation for the clusters ranged from 530 mm (cluster 5) to 820 mm (cluster 2). The average precipitation in the winter hydrological half-year ranged from 190.1 mm (cluster 1) to 288.8 mm (cluster 5). The average precipitation in the summer hydrological half-year ranged from 326.3 mm (cluster 5) to 605 mm (cluster 2).
4. Grouping by means of Ward's method can be used to distinguish homogeneous areas with the same levels of precipitation.

Acknowledgements

Publication supported by the Polish Ministry of Science and Higher Education as a part of the program of activities disseminating science from the project „Organization of the First International Science Conference – Ecological and Environmental Engineering”, 26-29 June 2018, Kraków.

The authors would like to thank the Institute of Meteorology and Water Management for providing free access to monthly precipitation totals from 1981-2010 for this study.

References

1. WD Shuster, J. Bonta, H. Thurston, E. Warnemuende, DR Smith, Urban Water J. **2** (4), 13 (2005)
2. M. O'Driscoll, S. Clinton, A. Jefferson, A. Manda, S. Mcmillan, Water **2** (3), 43 (2010)
3. A. Wałęga, S. Krzanowski, K. Chmielowski, Infr. and Ecol. of Rur. Are. **6** (4), 14 (2009) (In Polish).

4. A. R. Rao, V. V. Srinivas, *Regionalization of Watersheds. An approach based on cluster analysis* (Springer, 2008)
5. G. Laaha, G. Blöschl, *Journal of Hydrology* **323**, 21 (2006)
6. G.W. Lin, C.M. Wang, *Advanced in Water Resources* **29**, 12 (2006)
7. RK Price, Z. Vojinovic, *Urban Water J.* **5** (3), pp 259-276 (2008)
8. A. Richling, W. Lewandowski, *Wykorzystanie krajobrazu*. In: *Geografia fizyczna Polski* (Ed. by A. Richling and K. Ostaszewska, PWN, Warszawa, 2009) (In Polish)
9. A. Woś, *Klimat Polski* (PWN Warszawa, 1999) (In Polish)
10. H. Lorenc, *Climate Atlas of Poland* (IMGW, Warszawa, 2005) (In Polish)
11. A. Stanis, *Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny t. 3. Analizy wielowymiarowe*, (StatSoft, Kraków, 2007) (In Polish)
12. J. H. Ward Jr., *Hydrological Sciences Journal* **31** (3), p 335 (1963)
13. A.R. Rao, V.V. Srinivas, *Journal of Hydrology* **31** (1-4), p 57 (2006)
14. D. Eszegár-Kiss, B. Ceasar, *Transportation Research Procedia* **22**, pp 25-34 (2017)
15. C.C. Aggarwal, C.K. Reddy *Data Clustering: Algorithms and Applications*, (CRC Press Taylor and Francis Group, 2014)
16. J. Brzeziński, *The multidimensional mathematical models in psychological research*, (PWN Warszawa, 1987) (In Polish).
17. J. Królczyk, M. Tukiendorf, *Acta Scientiarum Polonorum, Technica Agraria* **4** (2), pp 21-30 (2005) (In Polish)
18. J. Kondracki, *Krainy fizjograficzne Polski*, (2002) (In Polish)