Wiadomości Statystyczne. The Polish Statistician, 2022, vol. 67, 2, 1–20 DOI: 10.5604/01.3001.0015.7717
Statystyka w praktyce / Statistics in practice
Zgłoszony/submitted: 1.08.2021, poprawiony/revised: 9.01.2022, zaakceptowany/accepted: 4.02.2022

The selection of areas for case study research in socio-economic geography with the application of *k*-means clustering

Agata Warchalska-Troll,[a] Tomasz Warchalski[b]

Abstract. The grouping techniques which are known in statistics are rarely used by geographers to select a research area. The aim of the paper is to examine the potential use of the *k*-means clustering (partitioning) method for the selection of spatial units (here: gminas, i.e. the lowest administrative units in Poland) for case studies in socio-economic geography. We explored this topic by solving a practical problem consisting in the optimal designation of gminas for in-depth research on the interaction between nature protection and local and regional development in the Polish Carpathians. Particular attention was devoted to defining an appropriate number of clusters by means of the elbow method as well as the pseudo-*F* statistic (the Calinski-Harabasz index). The data for the analysis were mostly provided by Statistics Poland and covered the period of 1999–2012. The multi-stage procedure resulted in the selection of the following gminas: Cisna, Lipinki, Ochotnica Dolna, Sękowa, Szczawnica and Zawoja.

The example described in the paper demonstrates that the *k*-means technique, despite its certain deficiencies, may prove useful for creating classifications and typologies leading to the selection of case study sites, as it is relatively time-effective, intuitive and available in open-source software. At the same time, due to the complexity of the socio-economic characteristics of the areas, the application of this method in socio-economic geography may require support in terms of the interpretation of the results through the analysis of additional data sources and expert knowledge.

Keywords: case study, *k*-means partitioning, elbow method, pseudo-*F* statistic, Calinski-Harabasz index

JEL: C38, O18, R58

Wybór obszarów do studiów przypadku w geografii społeczno-ekonomicznej z zastosowaniem metody grupowania *k*-średnich

Streszczenie. Znane w statystyce techniki grupowania są rzadko wykorzystywane przez geografów do wyboru obszaru badań. Celem analiz opisanych w artykule było sprawdzenie możliwości zastosowania metody podziału *k*-średnich do wyboru jednostek przestrzennych

[a] Instytut Rozwoju Miast i Regionów, Warszawa, Polska; Uniwersytet Jagielloński w Krakowie, Instytut Geografii i Gospodarki Przestrzennej, Polska / Institute of Urban and Regional Development, Warsaw, Poland; Jagiellonian University in Krakow, Institute of Geography and Spatial Management, Poland. ORCID: https://orcid.org/0000-0003-1314-3206. Autor korespondencyjny / Corresponding author, e-mail: agata.warchalska.troll@gmail.com.

[b] Badacz niezależny, Polska / Independent researcher, Poland. ORCID: https://orcid.org/0000-0002-2894-2265. E-mail: twarchalski@gmail.com.

(w tym przypadku gmin) do studiów przypadku. Dokonano tego poprzez rozwiązanie problemu metodycznego polegającego na optymalnym wyznaczeniu gmin do pogłębionych badań nad relacją między ochroną przyrody a rozwojem lokalnym i regionalnym w polskich Karpatach. Szczególną uwagę zwrócono na określenie odpowiedniej liczby skupień za pomocą metody łokcia (ang. *elbow method*) oraz statystyki pseudo-*F* (wskaźnika Calińskiego-Harabasza). Dane wykorzystane w analizach pochodziły z Głównego Urzędu Statystycznego i obejmowały okres 1999–2012. W rezultacie kilkustopniowej procedury wytypowano gminy: Cisna, Lipinki, Ochotnica Dolna, Sękowa, Szczawnica i Zawoja.

Opisany w artykule przykład pokazuje, że metoda *k*-średnich, pomimo pewnych słabości, może być przydatna do tworzenia klasyfikacji i typologii prowadzących do wyboru obszarów do studiów przypadku ze względu na jej użyteczność oraz dostępność w oprogramowaniu typu *open source*. Zarazem jednak – z uwagi na stopień złożoności społeczno-ekonomicznych cech obszarów – zastosowanie tej metody w geografii społeczno-ekonomicznej może wymagać wsparcia interpretacji jej wyników analizą dodatkowych źródeł informacji oraz wiedzą ekspercką.

Słowa kluczowe: studium przypadku, grupowanie metodą *k*-średnich, metoda łokcia, statystyka pseudo-*F*, wskaźnik Calińskiego-Harabasza

1. Introduction

The selection of a study area is a fundamental stage in geographic research, regardless of the specialisation involved. By its very nature, research in geography is often based on case studies of areas selected for their uniqueness or, in contrast, for being typical in terms of certain features. This basic dichotomy of the nature of the case study as a research strategy underlies its various classifications (see for instance Taylor, 2016, p. 583). In most cases the choice of a study unit(s) needs to be presented against a broader spatial background and sometimes additionally justified by a more detailed comparative analysis. The latter gains particular importance when the line of argument is based on the representativeness of the studied area.

This paper addresses the practical issues relating to the statistical techniques applied to official statistical data. The aim of the paper is to examine the potential use of the *k*-means partitioning method for the selection of spatial units (here: gminas, i.e. the lowest administrative units in Poland) for case studies in socio-economic geography. In its substantial part, the analyses presented in this paper were prepared within a research project investigating selected aspects of the influence of national parks and NATURA 2000 sites on local and regional development in the Polish Carpathians.[1] The study focused on three aspects: (1) technical infrastructure, (2) economic activity and its structure, and (3) social development. The project was initiated in 2013, with its fundamental part completed in 2019, and consisted of stages including desk and field research. When the research reached its summarising phase, it was noted that *k*-means clustering (partitioning) is still rather rarely used in socio-economic geography in the procedure of selecting the case study area (for more details please see the overview presented later in this

[1] Some results of this project were already published (Warchalska-Troll, 2018, 2019).

section). Considering the above, the authors decided that the applied research scheme is worth exploring, thus the paper focuses on discussing its strengths and weaknesses on the basis of actual research carried out in socio-economic geography.

As we reviewed the 2015–2019 issues of four Polish geographical journals of a well-established position at a national level and a growing share of foreign contributions (*Bulletin of Geography. Socio-economic series*, *Geographia Polonica*, *Polish Geographical Review*, *Regional and Local Studies*), we found that in only three cases statistical techniques were used in the process of selecting the study areas in socio-economic geography. In one case, *k*-means partitioning was applied for this purpose (Mikuš et al., 2016), although not by the authors themselves, but through a reference to another publication which was not available to us (Šebová, L. (2013). *Identifikácia marginálnych regiónov na Slovensku* [Doctoral disertation]. Comenius University in Bratislava).

From a total of 475 papers on socio-economic geography published in the above-mentioned journals, 181 included case studies. The vast majority of these articles (150 out of 181, i.e. 83%) included only a descriptive (although sometimes not even a comparative) comment on how and why a particular area(s) was selected for the case study analysis. Out of the 181 papers, 22 (i.e. 12%) provided no justification for the choice of the studied area and in the six remaining cases these justifications were reduced to one or two general phrases.[2]

Although the paper focuses on Poland, we also investigated how often '*k*-means' and 'case study' appear together in the scientific content of two high-impact databases:[3] Taylor & Francis online and Science Direct (by Elsevier). To remain discipline-specific, we narrowed our search to journals on socio-economic geography. In the Taylor & Francis online database, we found only 43 records meeting the criteria mentioned above in the field of 'Geography', and out of these 26 were in the 'Human Geography' section (which included transport geography, social geography, economic geography, regional geography and more). In none of these 26 papers *k*-means was applied in the selection of the studied sites. In the Science Direct database, our inquiry provided 914 research papers published in journals belonging to the social sciences discipline, out of which we verified the content of three journals which are often chosen by geographers and which had the highest number of items responding to the above-mentioned query: *Applied Geography* (28 results), *Landscape and Urban Planning* (26) and *Land Use Policy* (23). Here again, we found no example of the *k*-means technique used for the designation of the studied areas. Clustering was evidently broadly applied for classifications, typologies and regionalisations, but not

[2] However, it is worth noting that in numerous cases applying statistical techniques to study site selection is unnecessary or even useless, as many socio-economic geographers investigate very 'soft', qualitative matters, unique places or rarely represented features. Moreover, the understanding of case study as a research approach may vary among scientists (Babbie, 2007; Taylor, 2016). Finally, it should be taken into account that the research papers usually present only a part or a summary of broader research projects, which in consequence may also relate to the description of the applied methods.

[3] The query was executed in mid-December 2021.

in the study site selection procedure itself. This query is for sure limited in scope and gives only an idea of how popular the *k*-means technique is in socio-economic geography. However, the answer to the query at least suggests that the *k*-means and similar clustering techniques[4] are not a 'must-have' in human /socio-economic geographers' toolkit when it comes to selecting sites for their case studies.

2. Research method

2.1. Data and study area

The fields of interest described in the previous part of this paper implied our choice of variables from within the data provided by the Local Data Bank of Statistics Poland. The initial database covered 79 socio-economic variables downloaded directly from the Statistics Poland website, and further 44 variables were our own calculations based on the official statistics data. Four supplementary variables such as the list of spa cities/towns, the list of Natura 2000 sites, etc. were also taken from official governmental sources. All in all, the total database covered 127 variables (features), many of which were overlapping in terms of their substantive content. With the purpose of narrowing the set of the variables to be covered in our detailed analyses, we experimented with different subsets of 'feature-representatives' in infrastructure, economic activity, social development and environment. We then investigated the correlation (*r* Pearson) between the variables and evaluated the outcome using Student's *t*-test at a 0.05 significance level. At this point, it should be mentioned that the fact of a strong correlation existing between the variables posed no technical problem in the *k*-means analysis, which was at the core of our procedure; however, we did prefer to avoid strong correlations whenever possible. Despite this investigation and given the complexity of the research matter, the final choice of the variables could not be made without an expert decision and included: (1) the total number of tourist accommodation establishments; (2) the occupancy of bed places; (3) bed places per 1,000 citizens; (4) the share of entities in manufacturing (section C, divisions 10, 11, 15 and 16 of NACE Rev. 1.2 / PKD 2007) in the total number of the entities in the REGON register of economic units; (5) the share of entities in retail trade (except for sales of motor vehicles and motorcycles; section G, division 47 of NACE Rev. 1.2 / PKD 2007) in the total number of the entities in the REGON register; (6) the share of entities in selected services[5] (sections/divisions I56, M72, 74, P85, R90, 91, 93 and S94 of NACE Rev. 1.2 / PKD 2007) in the total number of the entities in the REGON register; (7) the share of

[4] The searching algorithm sorted the outcome by how well (thus not always in 100%) it matched the inquiry.

[5] Food and beverage service activities; scientific research and development; other professional, scientific and technical activities; education, creative, arts and entertainment activities; libraries, archives, museums and other cultural activities; sports activities; activities of membership-based organisations.

persons using the water supply system; (8) the share of persons using the sewage system; (9) outlays on fixed assets serving environmental protection and water management; (10) non-profit organisations and associations per 10,000 citizens; (11) natural persons conducting economic activity per 100 citizens; (12) entities newly included in the REGON register per 10,000 citizens (Table 1).[6]

Table 1. Basic characteristics of the variables chosen to create a classification of the Polish Carpathian gminas with the aim to select case study sites

Variable (description and unit)	Years	Value		
		min	max	mean
Total tourist accommodation establishments	2012	0.00	307.00	8.27
Occupancy of bed places in tourist accommodation establishments (total number of visitors in a given year per number of bed places in tourist accommodation establishments)[a]		0.00	158.07	16.97
Bed places in tourist accommodation establishments per 1,000 citizens[a]		0.00	825.79	44.82
Share (in %) of entities operating in manufacturing (section C, divisions 10, 11, 15 and 16 of NACE Rev. 1.2 / PKD 2007) in the total number of entities included in the REGON register[a]		0.01	0.33	0.06
Share (in %) of entities operating in retail trade except for the sales of motor vehicles and motorcycles (section G, division 47 of NACE Rev. 1.2 / PKD 2007) in the total number of entities included in the REGON register[a]		0.05	0.27	0.16
Share (in %) of entities operating in services (sections/divisions I56, M72, 74, P85, R90, 91, 93 and S94 of NACE Rev. 1.2 / PKD 2007) in the total number of entities included in the REGON register (selected services: food and beverage service activities; scientific research and development; other professional, scientific and technical activities; education, creative, arts and entertainment activities; libraries, archives, museums and other cultural activities; sports activities; activities of membership-based organisations)[a]		0.05	0.24	0.13
Entities newly added to the REGON register per 10,000 citizens (the total for the whole available period, i.e. 2009–2012)[a]	2009–2012	114.00	990.00	302.30

a Authors' calculations.

[6] In contrast to e.g. Kraszewska (2016), who investigated the regional differentiation of the risk of poverty occurring in Poland using the hierarchical Ward clustering method, in this study the variables cannot be classified as 'enhancers/boosters' and 'dampers', since the analysis aims at classifying gminas according to a selection of features and not at investigating the influence of these features on any phenomenon.

Table 1. Basic characteristics of the variables chosen to create a classification of the Polish Carpathian gminas with the aim to select case study sites (cont.)

Variable (description and unit)	Years	Value		
		min	max	mean
Natural persons conducting economic activity per 100 citizens	2012	4.10	25.40	9.68
Non-profit organisations and associations per 10,000 citizens		7.00	93.00	26.84
Persons using the water supply system in % of the total population		0.00	97.70	42.31
Persons using the sewage system in % of the total population		0.00	1.00	0.43
Outlays on fixed assets serving environmental protection and water management (the total for the whole available period, i.e. 1999–2008), *per capita* (average population for the period 1999–2008) in thousand PLN[a]	1999–2008	0.00	12.94	1.18

a Authors' calculations.

Source: authors' work based on data from the Local Data Bank of Statistics Poland.

After verifying that the variables to be used for the calculations have an acceptable coefficient of variation (in all cases it was above 20%) a min-max normalisation was performed using the following formula (Larose & Larose, 2014, pp. 26–27):

$$X^*(i) = \frac{X\ (i) - \min\ (X)}{max\ (X) - \min\ (X)},$$

where:

$X(i)$ – the value of the *ith* observation of variable X,

$X^*(i)$ – the normalised value of the *ith* observation of variable X,

$\min(X)$ – the minimum value of variable X,

$\max\ (X)$ – the maximum value of variable X.

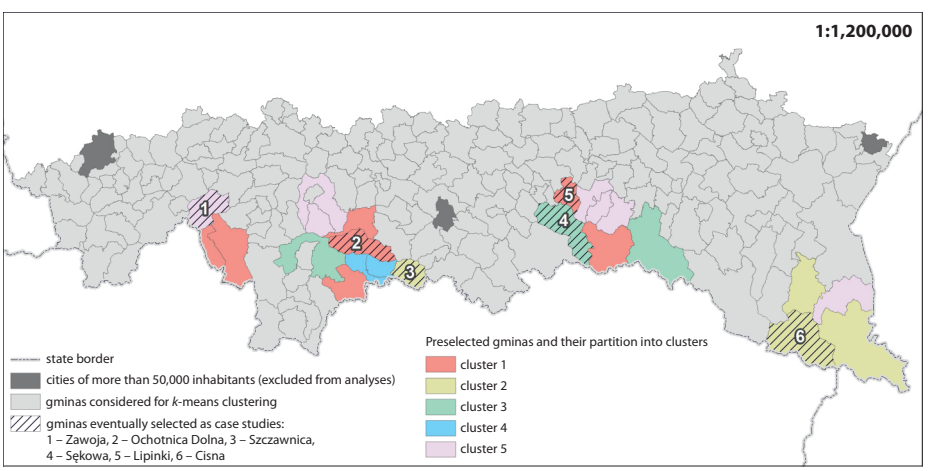
The basic analyses of data were performed in 2013 and early 2014, preceding the field-based part of the research project which was mostly carried out between 2014 and 2015. In the later stages of the case study selection procedure, several additional sources of information were used, i.a. official records (protocols) of public consultation meetings concerning the Natura 2000 sites, as well as an examination of the local press concerning gminas that were within the scope of our interest.

As regards the study area, the criterion of a gmina being listed under the Carpathian Convention (as of late 2013, when the study project was prepared) was adopted. After excluding 3 towns/cities exceeding 50,000 inhabitants, the first stage of our analysis covered 192 gminas (Figure 1).

2.2. Case study selection procedure

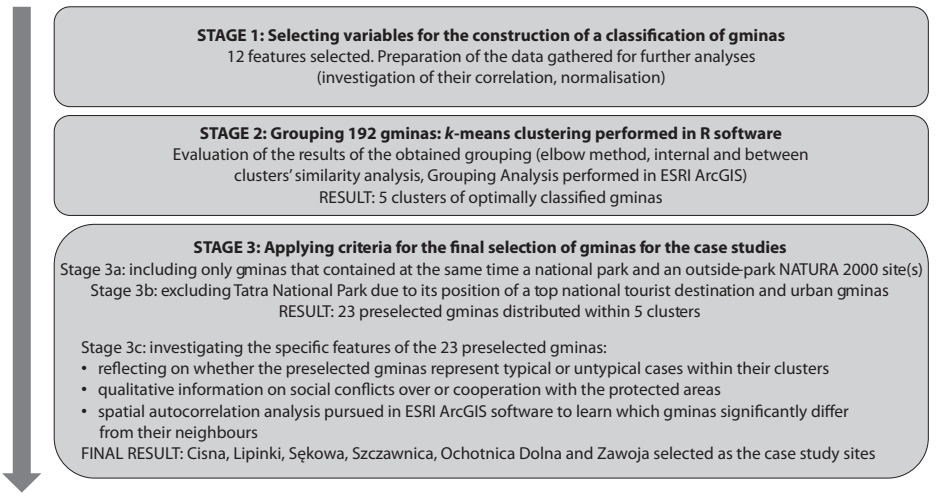
The case study selection procedure could be divided into three main stages (Figure 2). Experimenting with different variants of the clustering of a highly heterogeneous group of gminas located in the Polish Carpathians was the key part of the procedure. Once the final number of classes was chosen, we proceeded to select gminas representing each group and characterised by some additional features. This matter involved preselecting 23 gminas for closer investigation before the final six (Figure 1) were chosen, which was to guarantee that the topic of the research was covered to the greatest possible extent, and also that the number of cases to be investigated through in-depth field studies is feasible, given the limited time and budget.

Figure 1. Study area: gminas of the Polish Carpathians and the spatial scope of the analysis at different stages of the case study choice procedure



Note. Gminas selected for the case studies against the background of the whole set of Polish gminas covered by the Carpathian Convention as of late 2013 (N=195).
Source: authors' work based on the list of the Carpathian Convention gminas as of late 2013 and authors' analysis using data from the Local Data Bank of Statistics Poland; basemap: National Register of Boundaries (data of the National Geodetic and Cartographic Resource (Pol. Państwowy Zasób Geodezyjny i Kartograficzny).

Figure 2. A general framework of the case study selection procedure



Source: authors' work.

2.3. *k*-means clustering used for grouping gminas

The grouping was performed on a set of 192 gminas (defined in the previous sections of the article), using a non-hierarchical grouping method in the form of the *k*-means algorithm. This is one of the most widespread and well-established techniques whose core part was developed from the late 1950s to the late 1970s (Hartigan & Wong, 1979; Lloyd, 1982; MacQueen, 1967; Steinhaus, 1957[7]). This technique is applied in the field of cluster analysis – a branch of data analysis which involves dividing a particular set of observations into a given number of homogeneous subsets called clusters or groups. The synthetic measure of the level of the clusters' homogeneity, called the 'within groups sum of squared error' (R Core Team, n.d.) or the 'within-cluster sum of square errors' (WSS), represents the dispersion of the attribute values of individual objects within the clusters they belong to (Zhang et al., 2016). The procedure starts with an initial partitioning (usually created randomly) and then continues with modifying it iteratively until the best possible division into groups is found, i.e. the smallest WSS possible is achieved. In other words, at each step, increasingly more homogeneous groups are created. As a result of the algorithm implementation, we obtain a division of the set into a specified number of clusters and the value of WSS, which characterises the final

[7] A draft manuscript of this paper first circulated for comments in 1957 at Bell Telephone Laboratories, but was submitted for publication only in 1981.

level of the clustering quality (the lower the WSS, the better the clustering). Intuitively, it may be expected that the larger the number of clusters, the lower the WSS value. However, from the practical point of view, large numbers of clusters (groups) may not always be useful for researchers. They may cope with cluster abundance by assuming a maximum number of groups that would suit their research purposes. The challenge of selecting the optimal number of clusters will be addressed later in this section. Calculations for the purpose of the analysis were made with the kmeans ('stats' package) function of the R software (R Core Team, n.d.), implementing the *k*-means algorithm by default in the Hartigan and Wong version (Hartigan & Wong, 1979).

The *k*-means has been recently applied in various research topics in regional studies (Table 2), serving to create regionalisations as well as classifications and typologies.

Table 2. Examples of the application of *k*-means clustering in regional studies

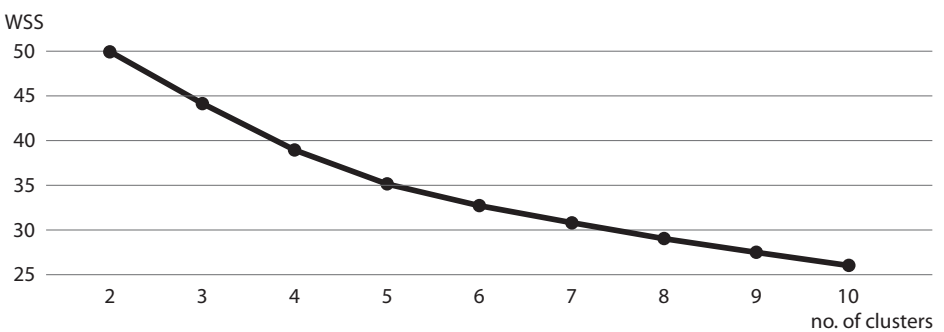
Paper	Purpose of the grouping
Crone (2005)	regionalisation of the US states (48 contiguous states) in terms of business cycles – comparing own clustering to the one performed by the Bureau of Economic Analysis (BEA)
Brauksa (2013)	exploration of socio-economic differentiation of Latvian municipalities following an administrative territorial reform
Šebová (2013), as cited by Mikuš (2016)	regionalisation of Slovakia in terms of socio-economic marginality
Li et al. (2016)	*k*-means used to support a geographically weighted regression (GWR) analysis of community resilience to seismic hazards in Southwest China, in particular to divide the study area into subregions based on characteristics defined by regression coefficients
Malinowski (2016)	classification of Polish regions in what concerns a relationship between the human potential and the economic effectiveness of enterprises – partitioning around medoids (PAM, *k*-medoids) method instead of *k*-means (a similar non-hierarchical clustering algorithm)
Novotná et al. (2016)	socioeconomic spatial typology of the Pilsen region, the Czech Republic
Stukalo & Simakhova (2018)	investigation of social economy patterns for 40 selected countries
Bole et al. (2019)	typology of small industrial towns in Slovenia
Dawidowicz (2020)	assessment of the financial situation of Polish gminas
Bayisa et al. (2020)	modelling and forecasting ambulance calls in Northern Sweden: *k*-means clustering used for the estimation of one of the parameters of the model
Kong et al. (2021)	exploring energy consumption patterns within the Qinghai Province, China

Source: authors' work based on a literature review.

Not having any initial assumption as to the desired number of clusters, we tested grouping variants from 2 to 10 clusters, and then evaluated the outcome based on the WSS value. Then we decided on how many clusters to consider using the elbow

method (Everitt et al., 2011; Peeples, 2011; Thorndike, 1953; Tibshirani et al., 2001; Zhang et al., 2016). This approach consists in the visual analysis of a graph showing the dependence between the WSS value and the number of clusters: we examine the course of the WSS (k) function (the WSS value for the division of the studied group of objects into k clusters) and search for such k (number of clusters) for which the graph clearly breaks down. The very idea of this method dates back to the 1950s (Thorndike, 1953) and is based on theoretical considerations related to the 'Gap statistic' (Tibshirani et al., 2001). As a result of its application, taking into account the graph obtained on the basis of our calculations (Figure 3), we came to the conclusion that the optimal division of our set involves five clusters.

Figure 3. The WSS of the k-means cluster analysis performed on a set of 192 gminas of the Polish Carpathians



Source: authors' work based on an analysis performed in R software involving data from the Local Data Bank of Statistics Poland.

While the elbow method can be seen as a bit arbitrary approach, we compared its outcome with the Calinski-Harabasz index (also called the pseudo-F statistic), which is a ratio reflecting within-group similarity and between-group difference (Caliński & Harabasz, 1974; Milligan & Cooper, 1985), and obtained the same results.

3. Results: the final selection of gminas for case studies

In the process of the final selection of the gminas for case studies, we adopted further criteria to narrow the potential set of the entities. Firstly, we decided to consider only those gminas which have national parks (or at least their buffer zone) and Natura 2000 site(s) other than parks[8] on their territory in order to identify the similarities and differences between these types of protected areas. Subsequently, we excluded

[8] In Poland, all national parks are by default included in the Natura 2000 network, so only analysing the Natura 2000 sites besides parks made sense in this research.

smaller towns and all gminas of the Tatra Mountains. The latter would be a too specific case as the area forms the top tourist destination in Poland, while we were interested in discovering any subtle interrelationships between nature protection and development; additionally, we aimed to focus on more peripheral areas. Next, we looked at how the 23 gminas which met the criteria above were distributed among the five groups determined on the basis of *k*-means clustering.

The default approach involved selecting one case per group. For this purpose, we examined the positions of the gminas within their groups, understood as the Euclidean distance from the groups' centres (Figure 4). We preferred to choose those cases which were close to their clusters' centres when one case was to be selected or, by contrast, to be not too close to each other (but also not too far from the cluster centre) when two cases in one group were considered. These analyses (Figure 4B) also led us to refrain from selecting a representative of cluster 4, which contained only two gminas that met our additional criteria described above – Czorsztyn was here a big outlier, while Krościenko nad Dunajcem appeared to us a too similar case to Szczawnica in terms of its location (Pieniny Mts), and by contrast, not having any additional features (Table 3) which were interesting from the perspective of the purpose of the study. The principle component analysis (PCA) demonstrated that clusters 2 and 4 became clearly separated only after the PC3 was used (the PC1 to PC2 diagram in Figure 4A shows that the two groups were still quite mixed), which suggests that cluster 4 (which incidentally formed the smallest group) can be to some extent treated as a subgroup of cluster 2. The latter circumstance additionally justified Szczawnica being chosen instead of any of the other two Pieniny gminas which fell into cluster 4.

The choice described above was further confirmed by the results of an analysis of the clusters' characteristics, involving both an investigation of the distribution of the variables considered for the *k*-means procedure (Figure 5) and a general, more general knowledge about the studied area, including qualitative and difficult to measure information.

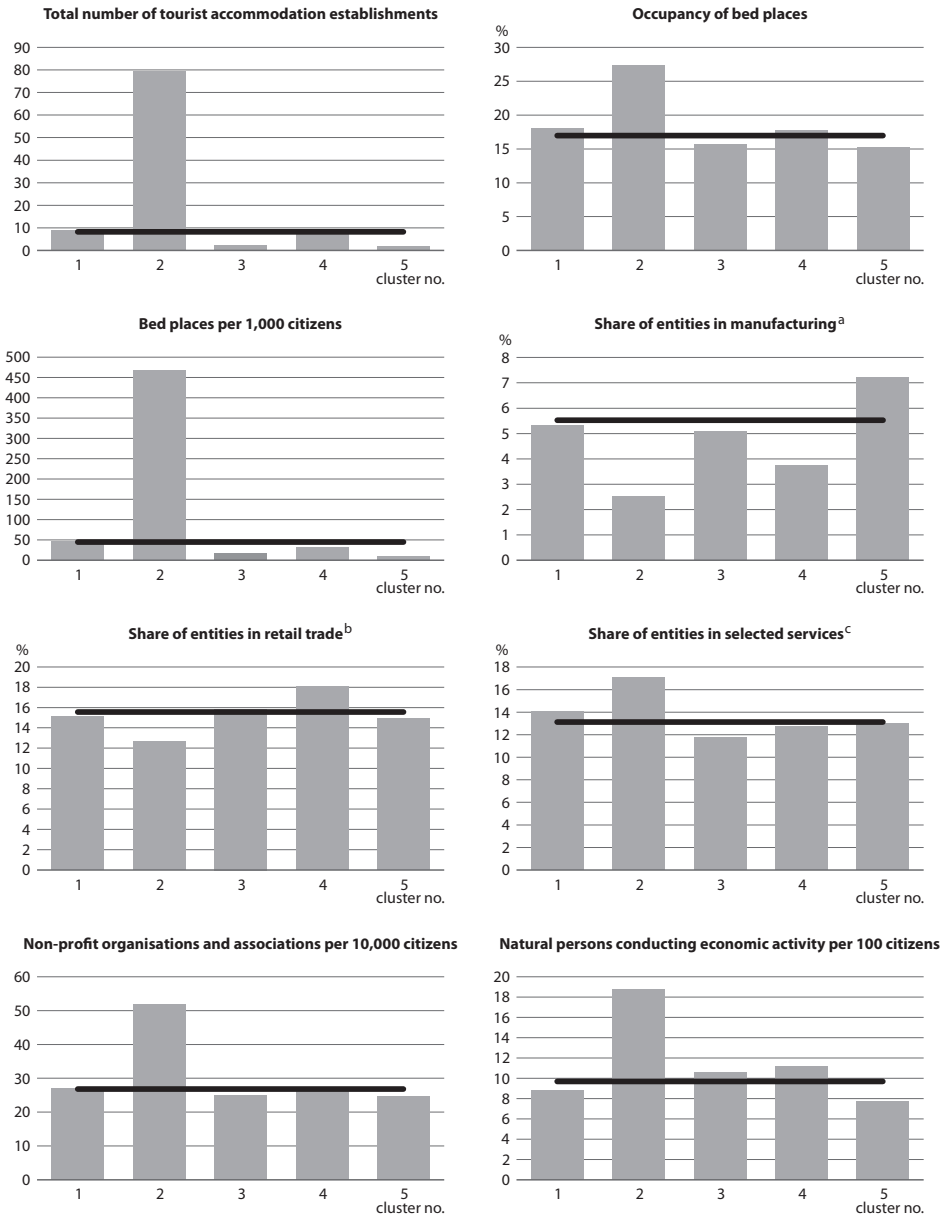
The latter included such aspects as the social reception of the protected areas (examples of both conflicts and good cooperation practices, resulting from the analysis of public consultation documents and from the local press[9]). We aimed at creating a well-balanced representation of different Carpathian physical and cultural landscapes and strived to include at least one gmina specialising in spa treatment (Table 3) in order to verify whether this kind of 'gmina profile' goes hand in hand with nature protection, as simple logic would suggest.

[9] For more details on the sources used at this stage of the project, please see Warchalska-Troll (2018, pp. 51–52).

Figure 4. Preselected gminas of the Polish Carpathians distributed among five clusters

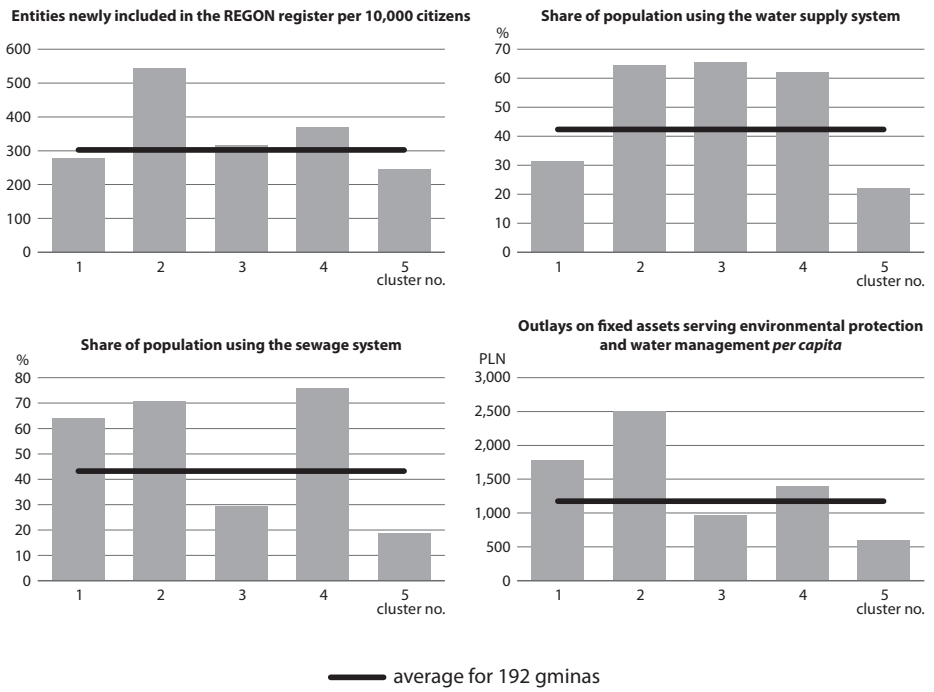
Note. The eventually selected gminas are shown in boxes.
Source: authors' work based on the analysis performed in R software using data from the Local Data Bank of Statistics Poland.

Figure 5. Average values for the variables considered in *k*-means clustering of 192 gminas of the Polish Carpathians for the five selected clusters, compared to the average values for the whole set



a Share of entities in section C, divisions 10, 11, 15 and 16 of NACE Rev. 1.2 / PKD 2007 in the total number of entities in the REGON register. b Share of entities in section G, division 47 of NACE Rev. 1.2 / PKD 2007 (except for the sales of motor vehicles and motorcycles) in the total number of entities in the REGON register. c Share of entities in sections/divisions I56, M72, 74, P85, R90, 91, 93 and S94 of NACE Rev. 1.2 / PKD 2007 in the total number of entities in the REGON register.

Figure 5. Average values for the variables considered in *k*-means clustering of 192 gminas of the Polish Carpathians for the five selected clusters, compared to the average values for the whole set (cont.)



Note. The data for the analysis were obtained in early 2014 and show the situation as of 2012, except for entities newly added to the REGON register per 10,000 citizens (which include the total for the whole available period, i.e. 2009–2012), and outlays on fixed assets serving environmental protection and water management *per capita* (the total for the whole available period, i.e. 1999–2008). For more information on the data used, please see Table 1.
Source: authors' work based on data from the Local Data Bank of Statistics Poland.

Table 3. A brief description of the main characteristics of the five clusters obtained through *k*-means clustering

Cluster no.	Cluster characteristics	Distribution of the 23 gminas selected at earlier stages	Long-lasting and significant conflicts over nature protection	Grounded and wide cooperation between the local people and nature protection stakeholders
1	Urban-rural or rural gminas retaining an agricultural component and traditional activities/lifestyles/folklore	Jabłonka	no	no
		Kamienica	no	no
		Krempna	yes	no
		Lipinki	no	no
		Lipnica Wielka	no	yes
		Łapsze Niżne	no	no
		Ochotnica Dolna	yes	yes

Table 3. A brief description of the main characteristics of the five clusters obtained through *k*-means clustering (cont.)

Cluster no.	Cluster characteristics	Distribution of the 23 gminas selected at earlier stages	Long-lasting and significant conflicts over nature protection	Grounded and wide cooperation between the local people and nature protection stakeholders
2	Largest tourist centres	**Cisna**	yes	no
		Lutowiska	no	yes
		Solina	no	no
		Szczawnica[a]	yes	no
3	Local trade and services centres with no single specialisation	Dukla	no	no
		Nowy Targ	no	no
		Sękowa[a]	yes	yes
4	Urban gminas and secondary tourist centres	Czorsztyn	no	no
		Krościenko nad Dunajcem	no	no
5	Peripheral, rural gminas with an important production (mostly forestry and logging) component, looking for their opportunity in winter sports with moderate success	Czarna	no	no
		Dębowiec	no	no
		Mszana Dolna	no	no
		Niedźwiedź	no	no
		Nowy Żmigród	no	no
		Osiek Jasielski	no	no
		Zawoja	yes	no

a Spa resorts.

Note. The table illustrates the situation as of 2014.

Source: authors' work based on an analysis of the official documents relating to public consultations (concerning plans issued by nature protection institutions) and the local press.

We not only investigated how the 23 preselected gminas were distributed within the clusters, but also their within-group positions (Figure 4). Then the spatial autocorrelation was examined to reveal which gminas significantly differ from their neighbourhood, and as such may provide some additional input. For this purpose the Anselin Local Morans I tool available in the ArcGIS software was used. We counted how many times each gmina stood out from the entire considered set at this stage, taking into account some basic features (population density, tourism infrastructure, transport accessibility, environmental infrastructure, economic and civic/social activity of the population, poverty). As a result, the Cisna gmina joined the already selected Szczawnica gmina in cluster 2. Cisna was the only gmina which appeared different from its neighbours in terms of five inquiries, five times showing a 'high-high' autocorrelation. To justify two cases appearing also in cluster 1, given the general profile of this cluster (urban-rural or rural gminas retaining an agricultural component and traditional activities/lifestyles/folklore (Table 3), the

authors chose Ochotnica Dolna and Lipinki, which represent two different regions (Gorce Mts and Beskid Niski Mts, respectively), two different physical features (Ochotnica Dolna is located in a mountain valley with steep slopes and harsh climate while Lipinki are situated in the foothills, with much more favourable conditions for agriculture), and, as a result, two different agricultural specialisations (pastoralism with mostly sheep herding and crops growing and cattle farms, respectively). Moreover, Ochotnica Dolna has a history of intensive and yet mixed experiences in cooperating with nature protection institutions. On the one hand, the gmina had a relatively good relationship with a nearby national park, based on common projects relating to the promotion of local cultural heritage and traditional summer farming revival, but on the other hand, numerous conflicts occurred in the process of implementing the Natura 2000 network on the gmina's territory. In turn, Lipinki had the lowest in the whole set values of variables relating to the development of tourism and as such effectively balanced the tourism-oriented gminas of Szczawnica, Cisna, Zawoja and (to a certain extent) Sękowa in our final case study set.

Among the numerous clustering techniques (Everitt et al., 2011), the *k*-means algorithm can be perceived as relatively time-effective and intuitive and its results can be easily interpreted; therefore, it may be a good choice when a quick outcome is anticipated. In this research, the overall scheme (summarised in Figure 2), whose central element was *k*-means partitioning, not only confirmed the authors' initial intuition of which gminas could be interesting to take a closer look at (e.g. Szczawnica, Zawoja), but also suggested the less obvious choices of Sękowa or Lipinki, and indicated Cisna as the most specific case in the generally 'standing out' region of the Bieszczady Mts. Moreover, as noted by Zhang et al. (2016), in contrast to many other methods (e.g. hierarchical ones), this one is less vulnerable to noisy data and was found to produce more stable results (i.e. a slight change in the scope of the data does not significantly affect the clustering result). Its certain inconvenience – arbitrarily assigning the number of clusters as its parameter – is in most cases easy to overcome by repeating the calculations for other values (it is usually quite clear for the researcher to what extent the number of clusters may vary). In these kinds of instances, the best option can be determined on the basis of one of the methods of evaluating the quality of clustering (Everitt et al., 2011; Kodinariya & Makwana, 2013; Migdał-Najman, 2011; Milligan & Cooper, 1985), like in our case the elbow method (Thorndike, 1953; Tibshirani et al., 2001) and the pseudo-*F* statistic, also called the Calinski-Harabasz index (Caliński & Harabasz, 1974; Everitt et al., 2011; Larose & Larose, 2014; Milligan & Cooper, 1985).

When reviewing studies similar to ours, we have observed that the number of clusters was sometimes based on an expert decision preceded by a qualitative or

quantitative analysis of the study area (e.g. Novotná et al., 2016), apart from cases when the amount of clusters was defined in advance due to a research hypothesis or prerequisites (e.g. Brauksa, 2013; Crone, 2005; Malinowski, 2016). In contrast, some authors propose to derive the optimal number of clusters from hierarchical clustering (e.g. Bole et al., 2019; Šebová, 2013 as cited by Mikuš et al., 2016; Stukalo & Simakhova, 2018) or from the Akaike information criterion (AIC) method (Li et al., 2016). Not undermining any of these approaches, similarly to e.g. Dawidowicz (2020), Kong et al. (2021), Nicholson et al. (2019)[10] and Zhang et al. (2016),[11] we opted for the elbow method which is simple to use and intuitive. Its possible drawback is that the curve which is supposed to indicate the optimal number of clusters may sometimes appear too 'smooth' (not having a distinct 'elbow'). This can happen when a large number of observations is considered and a relatively small number of clusters may be accepted. A possible solution to this problem could be replacing the WSS parameter by the difference in minWSS (Zhang et al., 2016). Another option is to verify the results obtained through the elbow method with an alternative technique, as we did by means of the pseudo-*F* statistic.

While using the *k*-means algorithm one should bear in mind that it may produce slightly different outcomes depending on which initial centroids are selected, as explained by Everitt et al. (2011) and noticed by many researchers (Crone, 2005; Novotná et al., 2016; Zhang et al., 2016). So, in case of doubts and if a particular implementation allows it, calculations may be repeated with other initial centroids (which is possible using the R software as we did). Gao & Kupfer (2018) also emphasise the fact that the *k*-means algorithm is 'non-spatial', as it does not seek to define by default the spatially coherent regions. This was not a problem in our case, as we only needed to classify gminas, not find clusters in the spatial sense. However, in some instances the latter may be achieved by adding extra attributes which numerically define the 'levels of neighbourhood' between the objects of research (Crone, 2005). Thanks to this possibility, the *k*-means method is likely to become more broadly appreciated by geographers. Finally, we would like to mention that apart from the classic GIS software, algorithms dedicated to solving tasks relating to data analysis in a spatial context are also available in the R software we used (e.g. the skater function of the spdep package as shown by Nicholson et al., 2019). In turn, the popular GIS software, ESRI ArcGIS, applies the *k*-means algorithm for its Grouping Analysis tool with a 'no spatial constraint' parameter (ESRI, n.d.).

[10] Nicholson et al. (2019) use the skater function in the spdep R package (instead of the *k*-means), which also performs spatial clustering and the value of the *k* parameter is to be defined by the researcher.
[11] Zhang et al. (2016) use a modified variant of the elbow method.

4. Conclusions

To sum up, in the procedure we followed, we combined a statistical analysis with qualitative criteria to narrow and finalise a selection of gminas for in-depth studies, as we believe that statistics should verify and support, not depreciate or dominate the 'spatial intuition' of geographers. The fact that we operated on qualitative and quantitative data of a diverse quality, and broadly used official statistics data, proves that our observations are not detached from the real challenges faced by scholars in this field. Our calculations were made using the open source R software, which does not have any specific requirements concerning the hardware, so our scheme may be copied by everyone. We did not want our research to be consistent with the belief that although statistics reveal a lot, at the same time they tend to conceal what is vital, therefore we considered a wide variety of non-measurable factors at the final stage of our research framework.

In conclusion, we have noticed a considerable potential of *k*-means clustering for case study selection in geographical research, which so far seems to have been only superficially explored in this application. In contrast to what is sometimes practised, we would not recommend this technique for regionalisation and other analyses aiming to find spatial clusters (or at least not this technique alone, as by definition it is non-spatial) but rather for classifications and typologies. When deciding which area to study, this method is likely to remain objective and not limited by measurability when supplemented by qualitative information concerning the local and regional spatial contexts. A practicable improvement to the analysis scheme described in this paper may concern the use of principal components instead of raw data for less correlated and more refined variables, similarly for instance to what Bole et al. (2019) described. Future studies relating to the topic may also include further research as to what extent and under what conditions clusters obtained from different partitioning methods converge.

References

Babbie, E. (2007). *Badania społeczne w praktyce*. Wydawnictwo Naukowe PWN.

Bayisa, F. L., Ådahl, M., Rydén, P, & Cronie, O. (2020). Large-scale modelling and forecasting of ambulance calls in northern Sweden using spatio-temporal log-Gaussian Cox processes. *Spatial Statistics*, *39*, 1–22. https://doi.org/10.1016/j.spasta.2020.100471.

Bole, D., Kozina, J, & Tiran, J. (2019). The variety of industrial towns in Slovenia: a typology of their economic performance. *Bulletin of Geography. Socio-economic Series*, *46*(46), 71–83. http://doi.org/10.2478/bog-2019-0035.

Brauksa, I. (2013). Use of Cluster Analysis in Exploring Economic Indicator Differences among Regions: The Case of Latvia. *Journal of Economics, Business and Management*, *1*(1), 42–45. http://doi.org/10.7763/JOEBM.2013.V1.10.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, *3*(1), 1–27.

Crone, T. M. (2005). An alternative definition of economic regions in the United States based on similarities in state business cycles. *The Review of Economics and Statistics*, *87*(4), 617–626. https://doi.org/10.1162/003465305775098224.

Dawidowicz, D. (2020). Ocena sytuacji finansowej gmin z wykorzystaniem metody *k*-średnich. *Wiadomości Statystyczne. The Polish Statistician*, *65*(7), 26–46. https://doi.org/10.5604/01.3001.0014.3284.

ESRI. (n.d.). *Grouping Analysis (Spatial Statistics) 8.* Retrieved June 24, 2021, from https://pro.arcgis.com/en/pro-app/2.8/tool-reference/spatial-statistics/grouping-analysis.htm.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis* (5th edition). John Wiley & Sons. https://doi.org/10.1002/9780470977811.

Gao, P., & Kupfer, J. A. (2018). Capitalizing on a wealth of spatial information: Improving biogeographic regionalization through the use of spatial clustering. *Applied Geography*, *99*, 98–108. https://doi.org/10.1016/j.apgeog.2018.08.002.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A *K*-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100–108. https://doi.org/10.2307/2346830.

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies*, *1*(6), 90–95.

Kong, W., Wang, Y., Dai, H., Zhao, L., & Wang, C. (2021). Analysis of energy consumption structure based on K-means clustering algorithm. *E3S Web of Conferences*, *267*, 1–5. https://doi.org/10.1051/e3sconf/202126701054.

Kraszewska, B. (2016). Wykorzystanie analizy skupień w ocenie zróżnicowania zagrożenia ubóstwem w podregionach Polski. *Wiadomości Statystyczne. The Polish Statistician*, *61*(5), 17–36. https://doi.org/10.5604/01.3001.0014.0993.

Larose, D. T., & Larose, C. D. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining* (2nd edition). John Wiley & Sons. https://doi.org/10.1002/9781118874059.

Li, X., Wang, L., & Liu, S. (2016). Geographical Analysis of Community Resilience to Seismic Hazard in Southwest China. *International Journal of Disaster Risk Science*, *7*(3), 257–276. https://doi.org/10.1007/s13753-016-0091-8.

Lloyd, S. P. (1982). Least Squares Quantization in PCM. *IEEE Transactions on Information Theory, 28*(2), 129–137. https://doi.org/10.1109/TIT.1982.1056489.

MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281–297). University of California Press. https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Some-methods-for-classification-and-analysis-of-multivariate-observations/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992.

Malinowski, M. (2016). Potencjał ludzki a efektywność ekonomiczna przedsiębiorstw – wykorzystanie metod taksonomicznych w ujęciu regionalnym. *Studia Regionalne i Lokalne*, (2), 87–109. https://doi.org/10.7366/1509499526405.

Migdał-Najman, K. (2011). Ocena jakości wyników grupowania – przegląd bibliografii. *Przegląd Statystyczny*, *58*(3–4), 281–299.

Mikuš, R., Máliková, L., & Lauko, V. (2016). An introductory study of perceptual marginality in Slovakia. *Bulletin of Geography. Socio-economic Series*, (34), 47–62. http://dx.doi.org/10.1515/bog-2016-0034.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, *50*(2), 159–179. https://doi.org/10.1007/BF02294245.

Nicholson, D., Vanli, O. A., Jung, S., & Ozguven, E. E. (2019). A spatial regression and clustering method for developing place-specific social vulnerability indices using census and social media data. *International Journal of Disaster Risk Reduction*, *38*, 101–224 https://doi.org/10.1016/j.ijdrr.2019.101224.

Novotná, M., Šlehoferová, M., & Matušková, A. (2016). Evaluation of spatial differentiation in the Pilsen region from a socio-economic perspective. *Bulletin of Geography. Socio-economic Series*, (34), 73–90. https://doi.org/10.1515/bog-2016-0036.

Peeples, M. A. (2011). *R Script for K-Means Cluster Analysis*. Retrieved May 27, 2021, from http://www.mattpeeples.net/kmeans.html.

R Core Team. (n.d.). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Retrieved August 30, 2020, from https://www.R-project.org/.

Steinhaus, H. (1957). Sur la division des corps matériels en parties. *Bulletin L'Académie Polonaise des Sciences*, *4*(12), 801–804. http://www.laurent-duval.eu/Documents/Steinhaus_H_1956_j-bull-acad-polon-sci_division_cmp-k-means.pdf.

Stukalo, N., & Simakhova, A. (2018). Global parameters of social economy clustering. *Problems and Perspectives in Management*, *16*(1), 36-47. https://doi.org/10.21511/ppm.16(1).2018.04.

Taylor, L. (2016). Case Study Methodology. In N. Clifford, M. Cope, T. Gillespie & S. French (Eds.), *Key Methods in Geography* (3rd edition; pp. 581–595). SAGE Publications.

Thorndike, R. L. (1953). Who belongs in the family?. *Psychometrika*, *18*(4), 267–276. https://doi.org/10.1007/BF02289263.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *63*(2), 411–423. https://doi.org/10.1111/1467-9868.00293.

Warchalska-Troll, A. (2018). Natura 2000 sites in the Polish Carpathians vs local development: inevitable conflict?. *eco.mont Journal of Mountain Protected Areas Research and Management, 10*(2), 50–58. https://doi.org/10.1553/eco.mont-10-2s50.

Warchalska-Troll, A. (2019). Do Economic Opportunities Offered by National Parks Affect Social Perceptions of Parks? A Study from the Polish Carpathians. *Mountain Research and Development, 39*(1), 37–46. https://doi.org/10.1659/MRD-JOURNAL-D-18-00055.1.

Zhang, Y., Moges, S., & Block, P. (2016). Optimal Cluster Analysis for Objective Regionalization of Seasonal Precipitation in Regions of High Spatial–Temporal Variability: Application to Western Ethiopia. *Journal of Climate*, *29*(10), 3697–3717. https://doi.org/10.1175/JCLI-D-15-0582.1.