DETERMINING THE SIGNIFICANCE LEVEL OF TOURIST REGIONS IN THE SLOVAK REPUBLIC BY CLUSTER ANALYSIS

JOZEF GÁLL[1]

Stanovenie úrovne významnosti regiónov cestovného ruchu v Slovenskej republike pomocou zhlukovej analýzy

***Abstract**: For the effective development and organization of tourism in any tourist region it is most important to assess its assumptions for tourism. Determining the significance of tourist regions in the Slovak Republic can help actors to better identify their priorities and further steps for the development of tourism in a particular region. In the Slovak Republic, there are few documents that classify tourist regions based on their potential and whose results are obtained from current data. The main aim of this article is to construct a model of cluster analysis, which will track the grouping of tourist regions into clusters based on their degree of significance in the tourism sector. Cluster analysis applied in the paper, which includes several different algorithms and methods for grouping objects with the most similarities to the appropriate groups. It deals with how the objects (statistical units) should be grouped in such a way as to make the most similarities within the groups and the greatest difference between the groups. Information on the variables examined, which describe the situation in year 2018, is drawn from the DATAcube database and the document Regionalization of tourism in the Slovak Republic issued by the Ministry of Economy of the Slovak Republic. Based on the results of the cluster analysis, the application of which has helped to obtain a picture of the division tourist regions into four clusters, the degree of significance each cluster and its members was determined by means of average variables. The present article and its results can be used as a basic material for future research in the tourism sector in the Slovak Republic.*

***Keywords**: cluster analysis, methods of hierarchical clustering, regionalization of tourism, tourist regions of Slovakia*

JEL Classification: C38, L83, R11

[1] Ing. Jozef Gáll, University of Economics in Bratislava**,** Slovak Republic, e-mail: jozef.gall@euba.sk

1. Introduction

Given the increasing perception of tourism as a catalyst in national and regional economic development, this sector becomes the subject of a particular interest between tourism researchers, practitioners and politicians. That is why, in recent years, several governments have focused on developing tourism in countries to attract visitors and invest in isolated regions [1].

For the region to be suitably developed in terms of tourism, it is necessary to examine the conditions for tourism. This investigation has resulted not only in the definition of tourist regions in the Slovak Republic, but also to their division into categories based on the established criteria for assessing their potential in a document issued by the Ministry of Economy of the Slovak Republic in year 2005 – *Regionalization of tourism in the Slovak Republic* [8].

It is now possible to determine the level of significance of tourist regions as a basic material for future publications and concepts regarding the economic development of a region in a competitive market by applying different stochastic methods, the results of which interpret the current state and (more often) more reliable information than non-updated publications. Cluster analysis is a widely used segmentation technique that forms groups of cases based on predefined variables, and group members (= cluster) should have the most similar variables (principle of homogeneity), while members of other groups are unlike (principle of heterogeneity) [2].

J. Han and M. Kamber [3] give examples of the use of cluster analysis in many social areas, such as:

- marketing – identification of groups of customers with similar interests;
- biology – identification of genes with similar functions;
- tourism – identifying groups of customers for holiday selection, identifying groups of countries surveyed for their competitiveness analysis, etc.

The application of the cluster analysis in practice in tourism sector occurs in publications around the world, such as the analysis of APEC's tourism competitiveness by J. C. Navarro Chávez et al. [6], analysis of tourism benefits on Hainan Island by H. Lifang et al. [5] or the classification of tourism potential in the Czech Republic by the authors of P. Chalupa et al. [4].

2. Data and Methodology

The main aim of this article is to identify a cluster model formed from the tourist regions which resemble each other but differ from other clusters of tourist regions based on the characteristics studied. The main tool used to define the tourist regions is represented by spatial geographical mapping –

cluster analysis. It allows for research to degrade, which a cluster of the tourist regions with the smallest distance between them is at the highest level [2].

Applied cluster analysis will be based on statistical data relating to the traffic of the tourist regions and the overall assessment of their potential. These variables have been selected to reflect the current situation of tourism in the regions (current facilities of the tourist region, attendance, etc.). Data describing the situation in year 2018 are obtained from the DATAcube database (2019) and document issued by Ministry of Economy of the Slovak Republic in year 2005 – *Regionalization of tourism in the Slovak Republic*.

Currently, there are several tools and statistical software available for the application of cluster analysis. The study will be conducted using a modern statistical program R, which is designed for statistical data analysis and their graphical display [7]. In the first stage of clustering, the distance between the objects is determined. I. Stankovičová and M. Vojtková [10] define several methods for measuring distances. This article will use one of the most widely used distance measures – *Euclidean distance* between objects *i* and *j* and a set of *n* variables according to the following formula [10]:

$$d_{ij} = \sqrt{\sum_{k=1}^{n} \left(X_{ik} - X_{jk}\right)^2} \tag{1}$$

where

d_{ij} ...Euclidean distance

n ...number of variables

X_{ik} ...the value of *k* variable for the *i* object

X_{jk} ...the value of *k* variable for the *j* object

Cluster analysis can be more subjective, and the result depends on the chosen method of grouping objects. In order to achieve the correct definition of the tourist regions, the following agglomerative methods of hierarchical clustering will be tested and compared to show whether there are significant differences in the results, or the tendency obtained are similar:

a) Complete Linkage is based on a maximum (longest) distance of object consisting of one sample from each cluster.

b) Single Linkage is one of the oldest methods of clustering. The determining character of this method is that the distance between the clusters is defined as the shortest pair of objects, considering only objects consisting of one sample from each cluster.

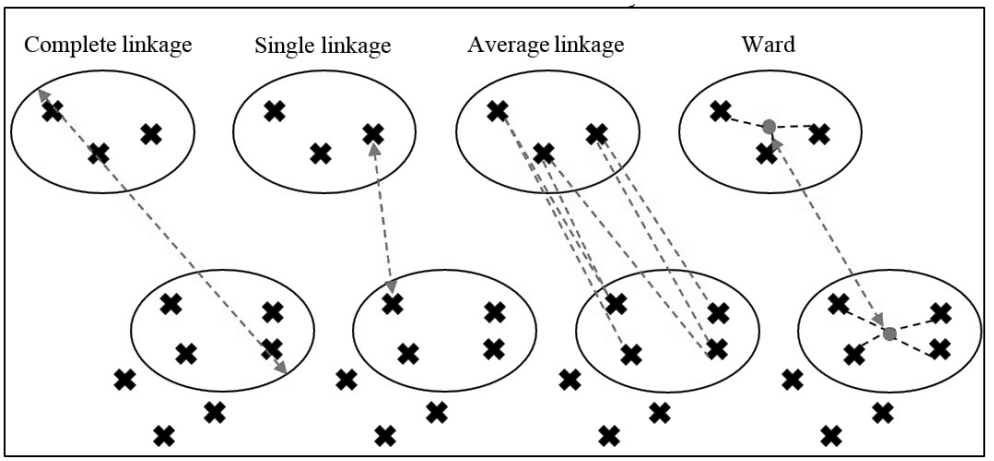
c) Average Linkage works like the previous two methods of clustering.

However, the distance between two clusters is the average of the distances between all the sample objects from each cluster.

d) Ward's method does not calculate the distance between clusters, but the joining of two clusters is based on the magnitude of the sum of the squared deviations in order to maximize their internal homogeneity. Figure 1 graphically illustrates these four approaches.

Figure 1

Graphical illustration of agglomerative methods of hierarchical clustering



Source: processed by the author according to Everitt et al., 2001 [2]

e) Centroid method combines groups into one cluster, among which is the smallest distance of their centroids, which represent the average of the samples of each cluster.

The results of the study will be represented through two-dimensional diagrams, which represent the process of creating clusters in each step of the analysis as a logical tree. Tree diagram – *dendrogram* also indicates the distance at which the object being examined were bound.

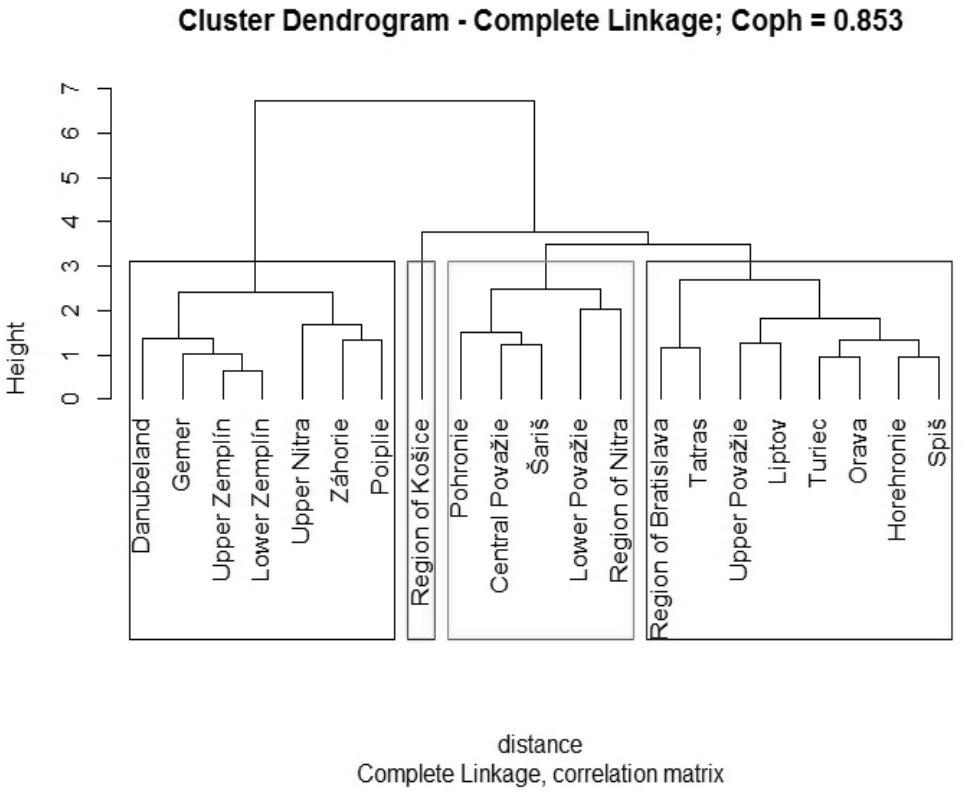
After applying selected clustering methods, a *cophenetic correlation coefficient* is used to validate the quality of the clustering and select the best dendrogram, which is generally used as a criterion for evaluating the effectiveness of the different methods of clustering. In the case of a cophenetic correlation coefficient, the principle is that the higher its value, the greater the credibility of the method and the obtained model of clusters better.

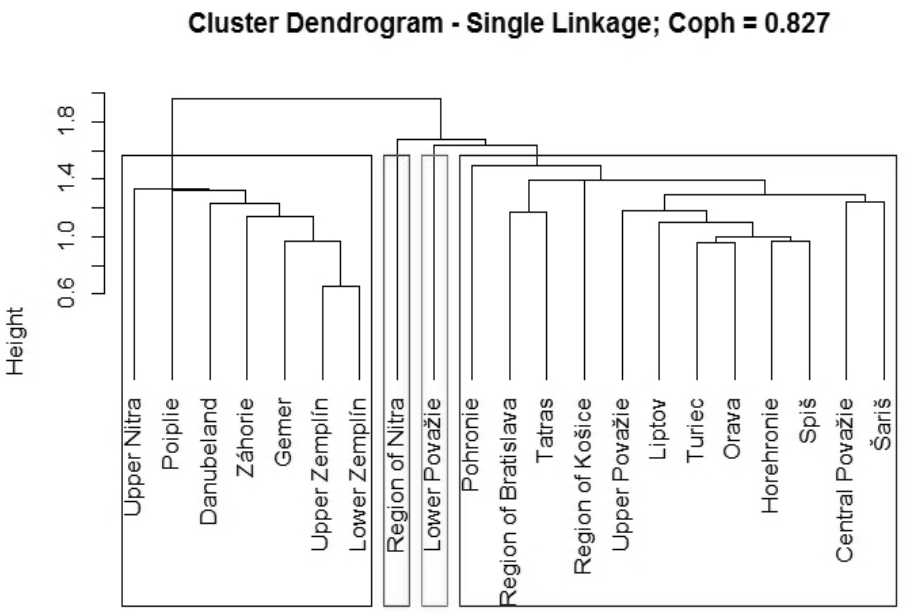
3. Results

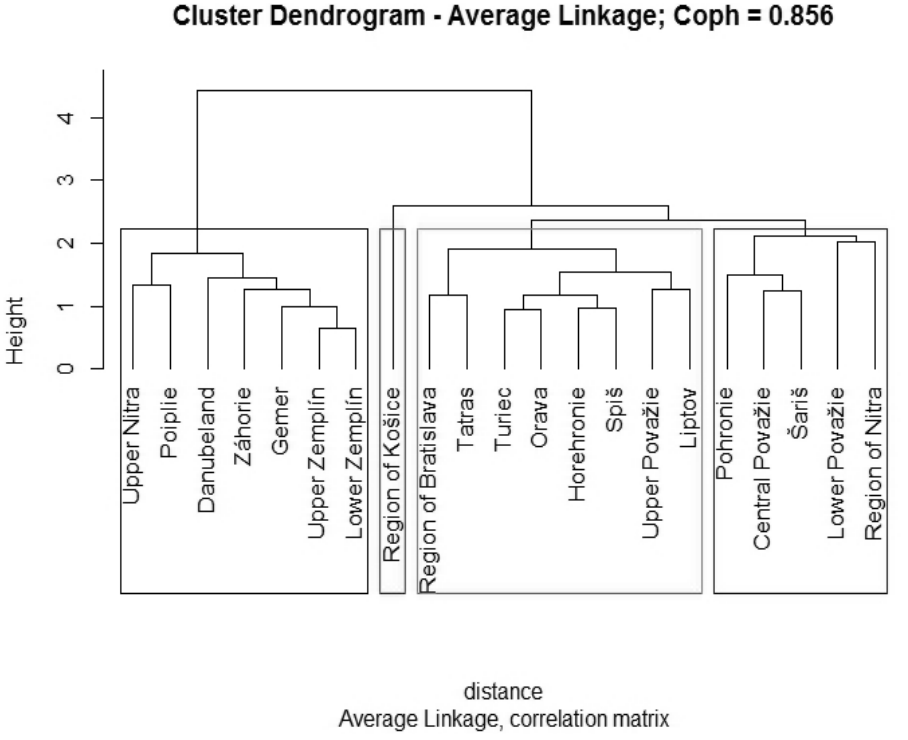
The following part of this article is focused on the application of the cluster analysis and the division of the tourist regions into clusters using an available database. At the beginning of the research, the same number of clusters (4) was determined to be used in the results of all agglomerative methods of hierarchical clustering. Due to the size of the objects examined, the number of resulting clusters is appropriate. On average, each cluster will contain 4 – 5 objects, or significantly larger or smaller groups may be created, up to individuals. At the same time, this step will eliminate possible problems related to the interpretation of the results obtained. Cluster analysis outputs are shown in Figure 2, which presents the results of selected agglomerative methods of hierarchical clustering.

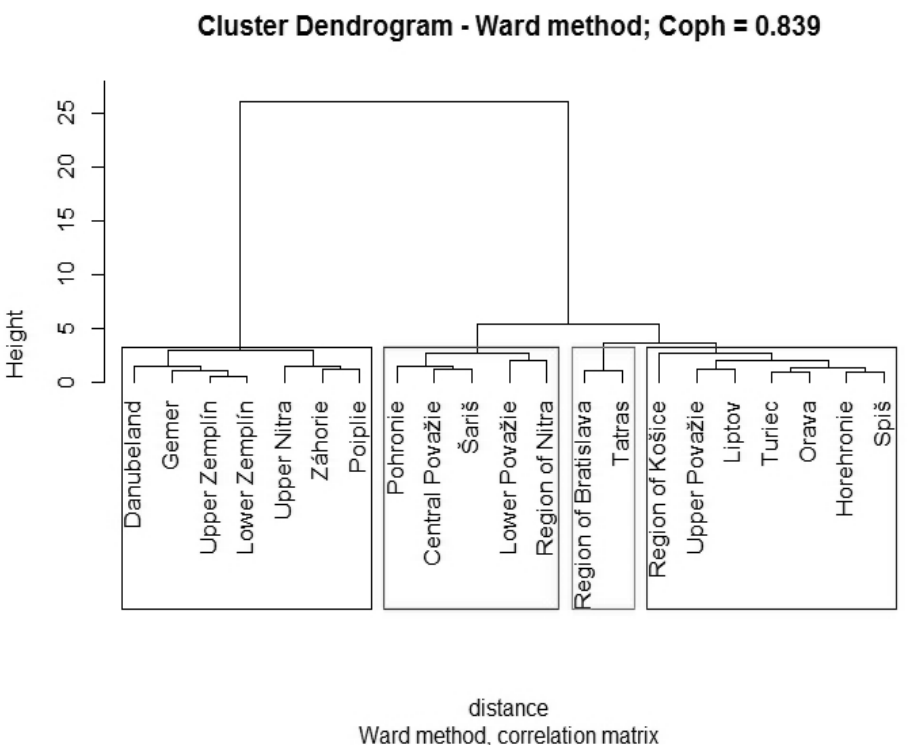
Figure 2

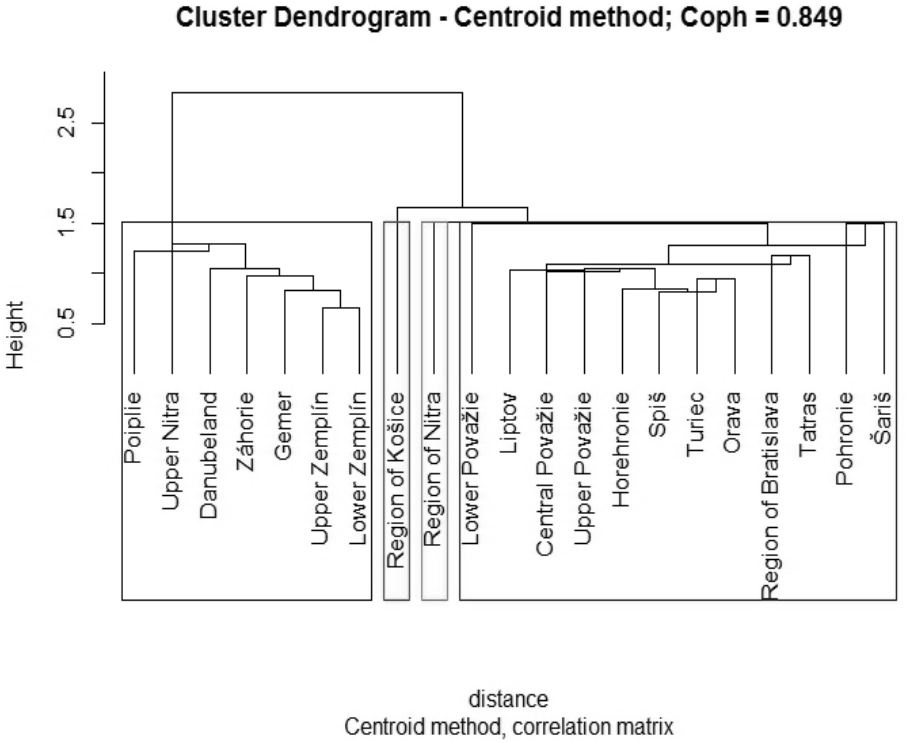
Cluster analysis – Dendrograms of agglomerative methods of hierarchical clustering




Cluster Dendrogram - Single Linkage; Coph = 0.827
Height
0.6
1.0
1.4
1.8
Upper Nitra
Poiplie
Danubeland
Záhorie
Gemer
Upper Zemplín
Lower Zemplín
Region of Nitra
Lower Považie
Pohronie
Region of Bratislava
Tatras
Region of Košice
Upper Považie
Liptov
Turiec
Orava
Horehronie
Spiš
Central Považie
Šariš
distance
Single Linkage, correlation matrix


Cluster Dendrogram - Average Linkage; Coph = 0.856
Height
0
1
2
3
4
Upper Nitra
Poiplie
Danubeland
Záhorie
Gemer
Upper Zemplín
Lower Zemplín
Region of Košice
Region of Bratislava
Tatras
Turiec
Orava
Horehronie
Spiš
Upper Považie
Liptov
Pohronie
Central Považie
Šariš
Lower Považie
Region of Nitra
distance
Average Linkage, correlation matrix





Source: processed by the author using statistical program R according to data by the Statistical Office of the Slovak Republic, 2019 [9]

Figure 2 illustrates five dendrograms, while the conclusions of each methods of hierarchical clustering are significantly different. The results obtained for each method are illustrated by four clusters, but their composition and members are not identical in either case. Although each dendrogram provides an interesting evaluation of input variables, at this stage of analysis it is not yet possible to determine the level of significance for individual clusters in dendrograms.

To verify and compare the methods of clustering used, a cophenetic correlation coefficient was used, which is calculated for the Euclidean distance according to the relation (1) to determine the adequacy of the clustering method and to determine by the methods the best cluster model. The resulting cophenetic correlation coefficient values are presented in table 1.

Table 1

Cophenetic correlation coefficient of applied agglomerative methods of hierarchical clustering

Euclidean distance	
Methods of hierarchical clustering	**Cophenetic correlation coefficient**
Complete Linkage	0,853
Single Linkage	0,827
Average Linkage	**0,856**
Ward's method	0,839
Centroid method	0,849

Source: processed by the author using statistical program R according to data by the Statistical Office of the Slovak Republic, 2019 [9]

From the table the most suitable method of clustering is the average linkage whose value of the cophenetic correlation coefficient is 0,856. The analysis results in four clusters composed of the following tourist regions:

Cluster I. Upper Nitra, Poiplie, Danubeland, Záhorie, Gemer, Upper Zemplín, Lower Zemplín;
Cluster II. Region of Košice;
Cluster III. Pohronie, Central Považie, Šariš, Lower Považie, Region of Nitra;
Cluster IV. Region of Bratislava, Tatras, Turiec, Orava, Horehronie, Spiš, Upper Považie, Liptov.

The next step of the cluster analysis is the technical implementation that identifies a cluster that provides the simplest and compact cluster. Table 2 shows the mean values of the input variables used in the cluster analysis. This information helps to characterize individual clusters and to determine their higher or lower significance.

Table 2

Mean values of input variables in clusters

Variable/ Cluster	V_1	V_2	V_3	V_4	V_5	V_6
	Means					
I.	-1,11489	-1,21793	-1,03763	-1,10489	-1,20637	-1,10914
II.	-0,95919	0,12316	-0,79304	0,46607	0,68244	0,91007
III.	0,46295	0,19259	0,88474	0,02771	0,14795	-0,10522
IV.	**0,80609**	**0,92993**	**0,45408**	**0,89120**	**0,87780**	**0,92251**

Source: processed by the author using statistical program R according to data by the Statistical Office of the Slovak Republic, 2019 [9]

The table above shows the average of the standard values of the input variables in each cluster. The results obtained allow the determination of four levels of significance for tourist regions:

- **Cluster I.**
  - The first cluster is made up of less developed tourist regions (especially the Southern and Eastern Slovak regions), which have the least suitable localization and implementation prerequisites for the development of tourism at national level and their use is rather regional.
- **Cluster II.**
  - The second cluster consists of the Region of Košice, which has a high level of localization and implementation preconditions for the development of tourism, but its potential is not sufficiently evaluated. The use the Region of Košice is qualified at regional and national level.
- **Cluster III.**
  - The third cluster consists of tourist regions, which have the same quality localization and implementation conditions, but their use is (as opposed to the fourth cluster) time limited. Like the fourth cluster, their importance is at national and international levels, and over time they are expected to achieve the highest level of significance in terms of tourism development.
- **Cluster IV.**
  - The fourth cluster, which has achieved the best average values, goes beyond its use and importance to national and international levels. Regions belonging to this cluster have high-quality localization and implementation prerequisites for the development of tourism, which is also confirmed by the fact that there are popular and visitors seeking destinations in the Slovak Republic (such as Region of Bratislava, Tatras, Orava, Liptov, etc.).

4. Conclusions

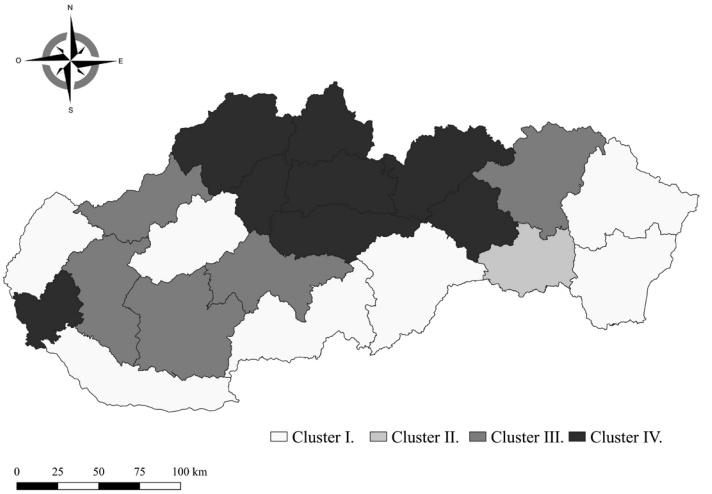
The main task of regionalization is a clear and systematic representation of the preconditions for the development of tourism. These assumptions include supply side factors (location and implementation conditions) and demand side factors (selective tourism conditions) [11]. An important step in the coordination and planning of the development of tourism in the Slovak Republic was the creation of a document Regionalization of tourism in the Slovak Republic prepared by the Ministry of Economy of the Slovak Republic in year 2005. Among other things, this act defined the tourist regions and evaluated their potential, including the localization and implementation assumptions of tourism. This concept also currently belongs to the base of the tourism policy tools in the Slovak Republic.

The use of a modern stochastic method makes it possible to categorize individual regions and create clusters, or also groups of regions that are most like each other, in terms of selected input variables. In contrast, between clusters of tourist regions (when comparing individual clusters to each other), these clusters are the most different from each other [2].

Applying a cluster analysis, the process of which is divided into several, successive steps, has shown a considerable difference in tourist regions. The analysis has divided the tourist regions into four clusters – categories (from lowest to highest significance) in terms of possible development and direction of investment in tourism. The visual map of division of tourist regions into clusters according to their significance is provided by the following map (Figure 3).

Figure 3

Map of tourist clusters according to their level of significance



Source: processed by the author using statistical program R, 2019

The most important cluster is the Region of Bratislava, Tatras, Turiec, Orava, Horehronie, Spiš, Upper Považie, and Liptov. All average values of the input variables were positive and significantly higher in this cluster than in the other clusters. Tourist regions belonging to this cluster have high-quality localization and implementation prerequisites for the development of tourism.

As already mentioned above, the Regionalization of tourism in the Slovak Republic is one of the tools of tourism policy at the national level. However, the problem is the fact that the document does not provide updated data concerning the potential of the regions and cannot be fully served as a basic pillar in the concept of strategic documents relating to development and investments directed to tourism at regional level. Given this fact and the efforts to pursue the issue of tourism development in the tourist regions in further scientific and research activities, the aim of this article was to offer a current picture of the division of tourist regions according to their significance into clusters using input indicators describing the situation in year 2018. The achieved results will be used in connection with tourism cluster mapping as factors of regional development in the Slovak Republic.

Acknowledgement

This contribution is the result of the internal grant project I-19-102-00 of the University of Economics in Bratislava for young pedagogical staff, scientific and PhD students entitled *Modern stochastic methods applied in tourism in the Slovak Republic*.

References

[1] DWYER, L. – KIM, CH. 2003. Destination Competitiveness: Determinants and Indicators. *Current Issues in Tourism.* 6:5, 369–414. London: Routledge, DOI 10.1080/13683500308667962.

[2] EVERITT, B. S. – LANDAU, S. – LEESE, M. 2001. *Cluster Analysis.* London: Arnold, 2001. ISBN 978-0470749913.

[3] HAN, J. – KAMBER, M. – PEI, J. 2001. *Data Mining – Concepts and Techniques.* London: Academic Press, 2001. ISBN 978-9380931913.

[4] CHALUPA, P. – PROKOP, M. – RUX, J. 2013. Use of Cluster Analysis for Classification of Tourism Potential. In *Littera Scripta*. Vol. 6, Issue 2, pp. 59–68. ISSN 1805-9112.

[5] LIFANG, H. – YIJUAN, CH. – CHENGYI, Z. 2014. The Cluster Analysis about Tourism Benefit in Hainan. In *Proceedings of the 2014 International Conference on Mechatronics, Electronic, Industrial and Control Engineering.* Atlantis Press, pp. 742–745. ISBN 978-94-62520-42-4.

[6] NAVARRO CHÁVEZ, J. C. – ZAMORA TORRES, A. I. – CANO TORRES, M. 2016. Hierarchical Cluster Analysis of Tourism for Mexico and the Asia-Pacific Economic Cooperation (APEC) Countries. In *Turismo em Análise.* Vol. 27, Issue 2, pp. 235–255. ISSN 1984-4867.

[7] R core team. 2019. *A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. [online]. Available at: <http://www.R-project.org/>.

[8] *Regionalization of tourism in the Slovak Republic.* Bratislava: Ministry of Economy of the Slovak Republic, 2005.

[9] Statistical Office of the Slovak Republic. 2019. *Kapacity a výkony ubytovacích zariadení podľa okresov - ročné údaje.* (Capacities and performances of accommodation facilities by districts – annual data) [online]. Available at: <http://datacube.statistics.sk/#!/view/sk/VBD_SK_WIN/cr3002rr/Kapacity%20a%C2%A0v%C3%BDkony%20ubytovac%C3%ADch%20zariaden%C3%AD%20pod%C4%BEa%20okresov%20-%20ro%C4%8Dn%C3%A9%20%C3%BAdaje%20%5Bcr3002rr%5D>. [accessed 24.06.2019].

[10] VOJTKOVÁ, M. – STANKOVIČOVÁ, I. 2007. *Viacrozmerné štatistické metódy s aplikáciami* (Multidimensional statistical methods with applications). Bratislava: Iura Edition spol s r.o., 2007. ISBN 978-80-8078-152-1.

[11] VYSTOUPIL, J. – ŠAUER, M. et al. 2011. *Geografie cestovního ruchu České republiky* (Geography of tourism of the Czech Republic). Plzeň: Aleš Čeněk, 2011. ISBN 978-80-7380-340-7.