Use of Cluster Analysis for Classification of Tourism Potential

Petr Chalupa, Martin Prokop, Jaromír Rux
College of Polytechnics Jihlava

Abstract

The aim of the research was to prove the methods of the cluster analysis, which have already been used by a number of scientific disciplines, are usable also in tourism. The cluster analysis is a mathematical statistical method which allows a set of objects of input data matrix to be divided into several clusters. Measures of distance are used to evaluate similarity of the objects. Euclidean distance can be used for quantitative variables. Tourism potential has been studied by a number of authors here and abroad, see bibliography. A district (Czech: okres) has been chosen as basic spatial unit in our research. The advantage of this approach is the possibility of using a public database of CZSO (Czech Statistical Office) which is the most important source of input data. Furthermore the data published by the Institute for Spatial Development in Brno have been used, the tourism potential of individual districts has been specified by the methods of cluster analysis, and the districts have been divided into six groups. Conformity of these groups with reality proves the appropriateness of the method used.

Keywords: Regionalization, Cluster Analysis, Tourism Potential, Qualitative Differentiation of the Czech Republic

Introduction

The aim of the research was to prove whether the methods of cluster analysis, which have been commonly used by a number of social and natural sciences, are also applicable to territory regionalization in terms of tourism potential. Different regions can be abstracted from a specific very complex geographical spatial object by different experts, not only if they approach the issue from different perspectives but even if they define it in the same perspective. However, this does not mean such subjectively defined regions do not exist objectively (Demek 1980). After all, for such regions we have chosen objectively existing parts, properties or relations from the given particular

geographical object. If we proceeded earnestly, from that perspective the defined regions would exist objectively. However, they do not exist with a finite number of elements and separately but only in a specific geographical object we have abstracted them from. Yet if necessary, we can define the exact boundaries of each region thanks to the fact we have chosen only a finite number of constituents, elements and links into it. The definition also applies in the case that the region has been abstracted from such geographical spatial object whose specific boundaries are completely featureless (the majority of cases). Of course, in the case that different people have abstracted different regions from the same specific geographical object, different boundaries may be defined (For example: Bauhalis 2000; Fisher 1982; 1987; Hübelová 2010; Kober 2006; Lauko 1982).

Materials and Methodology

Tourism Potential

Tourism potential has been studied by a number of authors here and abroad, see bibliography. (For example: Horník, Chalupa and Rux 1992; Klapka, Nováková and Frantál 2013; Vepřek 2002). For our purposes we consider works of Institute of Spatial Development in Brno (see Bína 2002; 2010a; 2010b) as a basis. For the first time the tourism potential for all municipalities in the Czech Republic was determined in 2001 using 24 constituents. An update has been made in April 2010.

Segments of the tourism potential are grouped into two following sub-potentials:

a) Tourism attractions potential

b) Potential of areas and lines affecting tourism

Ad a) point

Tourism attractions are e.g. châteaux, castles, botanical gardens, golf courses, spa locations and others, as well as natural sights such as caves, rock formations, etc. Other attractions consist of a certain reputation (e.g. municipalities renowned for winegrowing or brewing). Inclusion of a site into UNESCO World Heritage List creates significant added value of attraction (Chalupa, Janoušková and Hübelová 2013).

Ad b) point

This sector expresses broader territorial preconditions for tourism development. It is based on the fact that different areas have different general importance for tourism.

What is a part of the tourism potential and what is its attribute is discussed in the Institute of Spatial Development report. Statement that it is not possible to quantify all the constituents of tourism potential (e.g. genius loci) is correct

and essential. Authors of the report claim that when determining the breadth of the potential they took a golden mean. As an example they state that the existence of the castle is a part of the given site's potential, but services and attractions for visitors in the castle vicinity does not belong there. They do not consider also the existence of accommodation facilities, observation towers, nor downhill ski resorts as a part of the potential. We think such concept is too narrow. Surely nobody visits certain site in order to spend the night in a hotel. However, quality and price of accommodation can play a decisive role in the choice of tourist's destination.

The aim of our work is not in examining the definition of tourism potential at all but verification of suitability of the cluster analysis methods for its examination.

In 2010 (see Bína 2010b) an evaluation of utilizing the tourism potential by municipalities with extended powers (MEP) has been made. From Bína's work we took over numerical evaluation of the attractions potential recalculated for the districts. His evaluation of areas and lines is not appropriate for us and in this section we used data from the public database of CZSO.

As a criterion for utilizing the potential we used data about accommodation facilities visit rate from the public database of CZSO. We think that it is a basic factor which the tourism industry is interested in.

At the same time we once again remind that our work is not about a new detection of potential or evaluation of its utilization, but only about evidence that the cluster analysis methods that have been used in a number of disciplines are also suitable for tourism.

This method allows distribution of input data matrix objects set into several clusters. Both objects and variables can be clustered. We start from $n \times p$ X data matrix where n is the number of objects and p is the number of variables. We mark the number of clusters as k. We consider different explosions of n objects sets into k clusters. The aim is to achieve a situation where the objects within a cluster are similar as possible and objects from different clusters are least similar as possible (Jarkovský and Littnerová 2002).

To evaluate the quality of decomposition (see Hebák 2007) various criteria are used, e.g. Ward's criterion:

$$G = stE = \sum_{h=1}^{k} \sum_{i=1}^{n_h} \sum_{j=1}^{p} \left(x_{hij} - \bar{x}_{hj}\right)^2,$$

where E is the matrix of intra-cluster variability

$$E = \sum_{h=1}^{k} \sum_{i=1}^{n_h} \left(x_{hi} - \bar{x}_h\right) \left(x_{hi} - \bar{x}_h\right)^T,$$

x_{hi} is the observation vector for ith object in hth cluster, $\bar{x}_h$ is the average vector for hth cluster, and n_k is the number of objects in hth cluster. Creation of the most distant compact clusters occurs at minimum total sum of squared deviations of all the values from the respective cluster averages.

Measures of distance are used to evaluate similarity of the objects. Mutual distances for all the n object pairs are calculated, $n \times n$ type symmetrical square matrix of distance is created.

Euclidean distance can be used for quantitative variables.

$$D_E\left(x_i, x_{i'}\right) = \sqrt{\sum_{i=1}^{p} \left(x_{ij} - x_{i'j}\right)^2}.$$

The most common method of cluster analysis is **hierarchical clustering**, i.e. creation of hierarchical sequence of decompositions. Result of the hierarchical clustering can be best shown in a tree diagram, dendrogram. The distances between clusters are derived from the distances between objects. For example the smallest, the largest, or average distance can be used. Therefore there are several agglomerative methods. The Ward's method based on the aforementioned Ward's criterion G of decomposition quality is often considered the best of these methods.

The criterion for joining clusters is the increment of total intergroup sum of observations squared deviations from the clusters average as follows:

$$\Delta G = \sum_{i=1}^{g} \sum_{j=1}^{p} \left(x_{gij} - \bar{x}_{gj}\right) - \sum_{i=1}^{h} \sum_{j=1}^{p} \left(x_{hij} - \bar{x}_{hj}\right)^2 - \sum_{i=1}^{h'} \sum_{j=1}^{p} \left(x_{h'ij} - \bar{x}_{h'j}\right)^2,$$

where this increment is expressed as the sum of squares in the emerging cluster reduced by the sum of squares in the dissolving clusters. The expression can be simplified to the following form:

$$\Delta G = \frac{n_h n_{h'}}{n_h + n_{h'}} \sum_{j=1}^{p} \left(\bar{x}_{hj} - \bar{x}_{h'j}\right)^2.$$

Increment is expressed as the product of the Euclidean distance between the centroids of clusters considered for coupling and a coefficient dependent on the size of clusters. The coefficient value increases with increasing cluster size and for a fixed $n_h + n_{h'}$ is the maximum at clusters with the same size of $n_h = n_{h'}$. A coupling which ensures minimization of the ΔG criterion is performed; Ward's method therefore tends to remove small clusters. When using Euclidean distance, the following relationship can be used to convert distance matrix at each step of the clustering algorithm:

$$D_{gg'} = \frac{1}{n_h + n_{h'} + n_{g'}} \left[\left(n_h + n_{g'}\right) D_{hg'} + \left(n_{h'} + n_{g'}\right) D_{h'g'} - n_{g'} D_{hh'}\right],$$

where $D_{gg'}$ is considered the distance between gth and g'th cluster.

Results

We consider necessary to emphasize the importance and status of certain tourism region within hierarchically higher regional system, in our case the macro space of the Czech Republic.

The area is examined either as a unit mainly taxonomic, rather homogeneous, or as an area mainly complex with a complicated internal structure of the core (nodal) concentration and its background, which more accurately reflects the present geographical reality.

The tourism region can therefore be defined using the following system theory:

- The system (*"region"*) is a set consisting of a finite set of P elements and k final set – R local predicates.
- One sided predicates are called properties and/or characteristics. Two sided and multi-sided predicates are called relations and/or relationships.
- The set of all elements that lie outside the system but affect the system, forms the system surroundings.

So the open and closed systems are distinguished. In open systems there are relations between the system and its surroundings.

The tourism regions are always open systems. The set of elements consists of territorial units. One sided predicates are e.g. size of the area, number of workers in tourism, precipitation, and/or temperature. Two sided predicates are spatial interactions between territorial units.

Different regions can be abstracted from a specific very complex geographical spatial object by different experts, not only if they approach the issue from different perspectives but even if they define it in the same perspective (Mičian 1982). However, this does not mean such subjectively defined regions do not exist objectively. After all, for such regions we have chosen objectively existing parts, properties or relations from the given particular geographical object. If we proceeded earnestly, from that perspective the defined regions would exist objectively. However, they do not exist with a finite number of elements and separately but only in a specific geographical object we have abstracted them from. Yet if necessary, we can define the exact boundaries of each region thanks to the fact we have chosen only a finite number of constituents, elements and links into it. The definition also applies in the case that the region has been abstracted from such geographical spatial object whose specific boundaries are completely featureless (the majority of cases) (Lauko 1982).

Of course, in the case that different people have abstracted different regions from the same specific geographical object, different boundaries may be defined. We have chosen the cluster analysis method which seeks to identify the object clusters in a multidimensional space and subsequently reduce the multidimensional issue by categorizing the objects into identified clusters. Research data of D. Hübelová (2010) was used in the evaluation procedure.

The input matrix contains a total of 46 data representing potential attractiveness, potential infrastructure (capacity and occupancy rate of accommodation, transport availability), landscape potential, (national parks, protected landscape areas, national nature monuments, nature monuments, nature reserves, national nature reserves), basic information on population and settlements, pollution and crime rate.

Characteristics of Defined Groups

- **Strongly below average type of potential** in this group is based on the low number of monuments attractions and also the low attraction of the landscape. This is related to low level of infrastructure, particularly few accommodation facilities. The group has the lowest visit rate of tourists. Even these territories have tourism potential which must be utilized by perfect work of destination management.

- **A group with slightly below average potential** which is based (as in the first group) by small number of attractions, low potential of landscape, and weak infrastructure. In comparison with the first group there are districts with a larger number of inhabitants, population density twice as high, and higher air pollution. Visit rate is also below average but higher than in the first group.

- **A group with high tourism potential**. It is pleasant that this group includes most of the territory of the Czech Republic, proving that the republic as a whole has really high potential which has not been sufficiently utilized yet.

- **A group with the highest tourism potential** in the Czech Republic. The group has the highest attractions potential, landscape potential, and the best infrastructure facilities. Number of accommodation facilities is more than three times higher than national average. The national average for the number of visitors is exceeded even slightly higher (3.6 times). Trutnov District excels in both summer and winter tourism, Český Krumlov registered on the UNESCO List has a number of cultural attractions (Český Krumlov Château, Rožmberk Castle, Zlatá Koruna and Vyšší Brod Monasteries), water sports resort on Lipno Dam and Vltava River. The high potential of Karlovy Vary and Cheb Districts (Františkovy Lázně) is based on spa industry and it forms the following separate group.

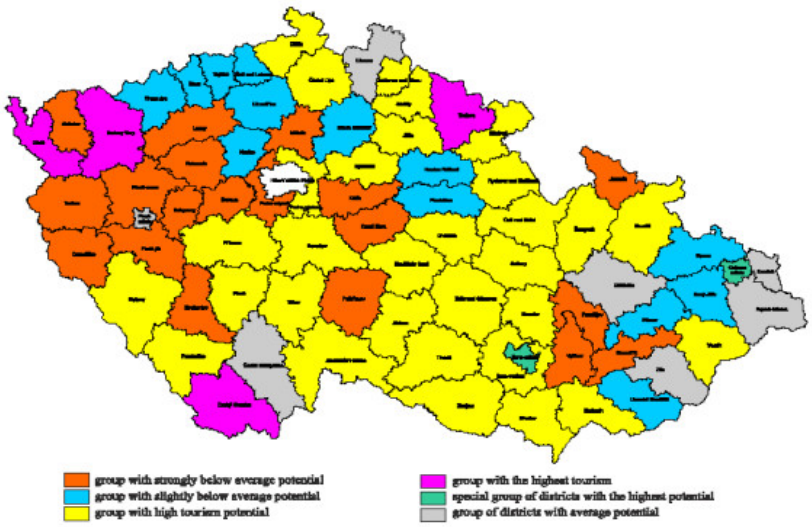
- **A special group of districts with the highest potential** are the urban districts of Brno and Ostrava. Most tourists coming belong to a group called Business Tourists. These are the participants of congresses, fairs, workshops, etc. Such tourists are very welcome. Their expenditures are above average, estimated as three times greater than expenditures of ordinary tourists who travel because of recreation or knowledge. This kind

of tourism is necessary to expand into other regions. The problem is the highest crime rate in the Czech Republic and high air pollution.

- **A group of districts with average potential**. Coincides with the previous group in high air pollution (almost twice the national average) and high crime rate. It has a lower level of infrastructure. The number of visitors is slightly above average indicating a good utilization of the potential.

In the Czech Republic there is no territory without tourism potential. However, there are considerable differences in its type, size, and utilization.

Map No. 1 Tourism Potential in the Czech Republic



Discussion and Conclusion

We are aware of the fact that there are many different methods for data clustering which differ by:

- Distance measurement between objects
- Algorithm for linking objects into clusters
- Interpretation of outputs

The meaningfulness of the clustering results depends both on the objective existence of clusters in the data and on arbitrarily set criteria of clusters definition Mundt 2001).The aim of the analysis can be either to determine links

between objects (dendrogram is a sufficient output) or to identify clusters in the data that will be used in further analysis as a simplification of multidimensional issue. In the process of clustering the most similar objects are gradually clustered until all the objects are merged into one cluster connecting all the objects in the analyzed file.

- The aim of this work was **to prove whether the methods of the cluster analysis, which have already been used by in a whole number of scientific disciplines for many years, are usable also in tourism.**
- The work was not preceded by a field research, all the data used are taken from official sources – the public database of the Czech Statistical Office and the data of the Institute of Spatial Development in Brno.
- The district has been selected as the basic unit because there are sufficient data in the CZSO for this unit. The high amount of input data could not be objectively processed and evaluated by classical methods.
- According to the user needs the input data matrix can be adjusted for other territorial units and also for other input data. Publicly available software has been used for the evaluation.
- Our method of cluster analysis allowed identifying that the republic can be divided into 6 groups of districts differing in characteristics of tourism potential. The achieved distribution is in accordance with reality which proves the suitability of using the cluster analysis.

References

BÍNA, J., 2002. Hodnocení potenciálu cestovního ruchu v obcích České republiky. *Urbanismus a územní rozvoj.* **5**(1), 2–11. ISSN 1212-0855.

BÍNA, J., 2010a. *Aktualizace potenciálu cestovního ruchu. 10.9/CR Task Final Report.* Brno: The Institute of Spatial Development in Brno.

BÍNA, J., 2010b. *Využití potenciálu cestovního ruchu. B.9/CR Task Final Report* [online]. Brno: The Institute of Spatial Development in Brno, [cit. 09-15-2013]. Available from: http://www.uur.cz/images/3-cestovni-ruch/vyuziti-potencialu-cestovniho-ruchu/vyuziti-potencialu-cr-zprava-2010.pdf

BUHALIS, D., 2000. Marketing the Competitive Destination of the Future. *Tourism Management: Research Policies Practice.* **21**(1), 97–116. ISSN 0261-5177.

CHALUPA, P., E. JANOUŠKOVÁ a D. HÜBELOVÁ, 2013. *Geografie obyvatelstva a sídel pro cestovní ruch.* Jihlava: VŠPJ. ISBN 978-80-87035-81-8.

DEMEK, J., 1980. Teorie regionální geografie. *Acta UC, Geographica.* **XV**, 43–52. ISSN 0300-5402.

FISHER, M. M., 1982. *Eine Methodologie der Regionaltaxomie: Probleme und Verfaren der Klassifikation und Regionalisierung in der Geographie und Regionalforchung.* Bremen: Bremer Beiträge zur Geographie und Raumplanug, Heft 3.

FISHER, M. M., 1987. *Some Fundamental Problems in Homogeneous and Functional Regional Taxomy.* Bremen: Bremer Beiträge zur Geographie und Raumplanug. Heft 11.

HEBÁK, P. et al., 2007. *Vícerozměrné statistické metody 3.* Praha: Informatorium. ISBN 80-7333-039-3.

HORNÍK, S., P. CHALUPA a J. RUX, 1992. *Rajonizace na základě hodnocení závislostí přírodních a socioekonomických faktorů životního prostředí.* Brno: Masaryk University. **42**. Writings of the Faculty of Education, Masaryk University. ISBN 8021003979.

HÜBELOVÁ, D., 2010. Specifika demografického vývoje České republiky. Sborník příspěvků z mezinárodní vědecké konference Region v rozvoji společnosti 2010 [CD-ROM]. *Sborník příspěvků z 2. konference konané 21. 10. 2010.* Brno: MeU, FRRMS, 71–75. ISBN 978-80-7375-435-8.

JARKOVSKÝ, J. a S. LITTNEROVÁ, 2002. *Vícerozměrné statistické metody: Shluková analýza. Preparation of Teaching Materials for the Mathematical Biology Course.* /ESF Project No. CZ.1.07/2.2.00/07.0318 , 2002 "VÍCEOBOROVÁ INOVACE STUDIA MATEMATICKÉ BIOLOGIE" VSM-O5-pdf-Adobe Reader/

KLAPKA, P., E. NOVÁKOVÁ a B. FRANTÁL, 2013. *Metodologické přístupy k hodnocení potenciálu cestovního ruchu území* [online]. [cit. 2013-07-17]. Available from http://geography.upol.cz/soubory/lide/klapka/klapka-novakova-frantal.pdf

KOBER, J., 2006. *Touristisches Zukunftskonzept: Harz 2015. Tourismus-Studien Sachsen-Anhalt.* Erfurt, Goslar, Hannover, Magdeburg: Niedersachsisches Ministerium für Wirtschaft, Arbeit und Verkehr [online]. [cit. 2013-07-17]. Available from http://www.sachsen-anhalt.de/fileadmin/Elementbibliothek/Bibliothek_Politik_und_Verwaltung/Bibliothek_Wirtschaftsministerium/Dokumente_MW/reisen_und_erholen/Handbuch_Harz_WEB-1.pdf

LAUKO, V., 1982. Podstata regionálnej geografie a jej postavenie v systéme geografických vied. *Geografický časopis.* **34**(3). 265–276. ISSN 1335-1257.

MIČIAN, L., 1982. Niektoré všeobecnogeografické problémy. *Zeměpis pre stredné školy.* Bratislava: SPN.

MUNDT, J. W., 2001. *Einführung in den Tourismus.* 1. Issue. Munich, Vienna: Oldenbourg Verlag. ISBN 3-486-25639-4.

VEPŘEK, K., 2002. Hodnocení potenciálu cestovního ruchu a jeho využití v územních plánech VÚC. In: *Urbanismus a územní rozvoj* [online]. 2002 [cit. 2013-07-17]. Available from www.uur.cz/images/publikace/uur/2002/2002-03/05.pdf

Contact address:
prof. PhDr. Petr Chalupa, CSc., College of Polytechnics, Tolstého 16, 586 01 Jihlava, *e-mail: chalupapet@seznam.cz*

Mgr. Martin Prokop, College of Polytechnics, Tolstého 16, 586 01 Jihlava, *e-mail: prokopm@vspj.cz*

RNDr. et PaedDr. Jaromír Rux, CSc., College of Polytechnics, Tolstého 16, 586 01 Jihlava, *e-mail: rux@vspj.cz*

CHALUPA, P., M. PROKOP and J. RUX. Use of Cluster Analysis for Classification of Tourism Potential. *Littera Scripta*. 2013, **6**(2), 59–68. ISSN 1805-9112.
