# Are Shape Metrics Useful for a Geocomputation? CORINE Land-Cover Analysis Case Study

Vít Pászto<sup>1</sup>, Lukáš Marek<sup>1</sup>, Pavel Tuček<sup>1,2</sup>

<sup>1</sup>Department of Geoinformatics, Faculty of Science, Palacky University in Olomouc Tř. Svobody 26, Olomouc, 771 46, Czech Republic
<sup>2</sup>Department of Mathematical Analysis and Applied Mathematics, Faculty of Science, Palacky University in Olomouc
17. listopadu, 771 46 Olomouc, Czech Republic
{lukas.marek, pavel.tucek}@upol.cz, vit.paszto@gmail.com

Abstract. Since shape metrics emerged in the landscape ecology as a new tool for quantitative evaluation of a landscape, it has become easier for geocomputation methods in GIS to adopt theirs principles. Nevertheless, there are still different scientific opinions about the usefulness of shape metrics. The paper describes shape metrics application for Corine Land Cover 1990, 2000 and 2006 areas (CLC) analysis along with statistical methods and discusses its benefits and disadvantages. The main goal of the paper is to evaluate CLC dataset without including attribute or qualitative information into analysis using shape metrics calculation. Thus, only geometric part of the data has been processed. Twenty eight metrics have been used for more than 900 areas (patches) from CLC dataset covering Olomouc region. Metrics values have been calculated and consequently used for correlation analysis, principal component analysis and cluster analysis. The results of the study represent complex evaluation of CLC Level 1 classes using, fundamentally, only the shape of CLC areas (patches). The analysis results show that shape metrics are very useful to identify groups of landscape patches with similar shape.

Keywords: shape metrics, GIS, land-cover, geocomputation, clustering.

# **1** Introduction

Since landscape ecologists can use capabilities of computer calculations, they are able to apply numerous tools to quantify landscape patches in an effective way. For this purpose, various indexes and metrics based on a patch shape have been derived, because according to [16] landscape ecology is largely founded on the notion that environmental patterns strongly influence ecological processes. Authors in [8] mentioned that developing methods to quantify landscape patterns are considered as a prerequisite to the study of pattern-process relationships. Authors in [8] continue and claim that progress has been facilitated by recent advances in computer processing and geographic information technologies.

Shape metrics are exactly those methods used for quantitative description of a patch shape, which represents real world objects. Shape and spatial metrics was

recently used in various topics, e.g. city footprint and form evaluation ([5], [14]), measuring city sprawl [15], analysis of landscape ([3], [10], [17]), in remote sensing [9] and also in a land-use change modelling [4]. Metrics are now being implemented in GIS software or extensions for GIS software but still not widely used. With the use of multivariate statistics, it is possible to evaluate, cluster and classify patches only according to their quantitative characterization. Mentioned methods are considered as a geocomputational and are both stand-alone and integrated in GIS.

There are several approaches how to classify landscape patches, but none of these are using shape metrics in combination with multivariate statistics for complex quantitative description of a landscape. It is common to use only a limited number of metrics to evaluate one specific patch group (e.g. habitats of particular species, humid areas, urbanized areas etc.). It is important to note that appropriate use of chosen metric depends on what is under the scope of study. One metric is more suitable for a one type of analysis, another for a different type. Although the use of metrics is purpose-dependent, metrics for this paper were chosen with an intention to calculate the most available ones for consequent multivariate statistics and tested if they can be (altogether) a tool for semi-automatic landscape classification. Similarly, analyzed patches used in this paper cover every patch type defined in CLC Level 1 classification nomenclature.

Thus, the approach presented in this paper is quite unique and the aim is complex landscape analysis via geocomputational methods to evaluate their usability for a landscape classification. Classification and proposed clustering methods were done with the view of the fact that only landscape patch shapes (geometry) were evaluated. Resulting clusters refer about the similarity of patch shapes and group areas with similar geometry. It is then evaluated what is the ratio of CLC Level 1 patches within clusters created only with the respect of shape metrics.

#### 2 Data, Study Region and Methods

Analysis was performed on freely available CLC dataset from 1990, 2000 and 2006 using Level 1 nomenclature, which classifies a land cover into 5 main categories - artificial surfaces, agricultural areas, forest and semi-natural areas, wetlands and water bodies. Overall, for 944 landscape patches (sum from all years) from Level 1 shape metrics calculations were done. Landscape patches are elementary, further nondivisive units of a landscape and according to [2] are defined as a relatively homogeneous areas that differs from its surroundings. These basic units or areas represent a specific type of land cover and together form a landscape matrix [2]. It is possible to group fundamental landscape patches according to their common characteristics to obtain more general patch type in different scale level, e.g. using CLC nomenclature – artificial surfaces are composed of urban fabric; industrial, commercial and transport units; mine, dumps and construction sites; and artificial, non-agriculture vegetated areas [1]. Furthermore, industrial, commercial and transport units consist of industrial or commercial units, road and rail networks and associated land, port areas and airports, which represent the highest resolution units or patches in CLC nomenclature.

Olomouc region (Fig. 1) was chosen as a study area, with more than 300 patches of CLC nomenclature types in each reference year, in order to follow previous fractal analysis of this area [11]. Olomouc region has an area around 800 km<sup>2</sup> and lies in a valley almost 20 km wide in south-east direction. This part of the region is mostly covered with agricultural areas and artificial surfaces, which are villages and the centre of the region – Olomouc city. North east part of the region is represented by hilly landscape and is covered with forests and semi-natural areas.

Shape metrics are fundamentally based on an area of a shape and its perimeter (these two characteristics are itself considered as shape metrics and are very easy to obtain), but most of metrics are more complicated to calculate and are treated as shape indexes. Anyway, there are plenty of software tools to perform metrics calculation. In this study, FRAGSTATS 4.1 and Shape Metrics toolbox for ArcGIS 10.x for Desktop was used. Multivariate statistics was performed in RStudio environment using R Project programming language.

List of metrics calculated in this study are in Table 1 and their description is available in [8] and in [12]. Nevertheless, it is worth to mention, why it is useful to calculate shape metrics. Since shape metrics take into account only geometric properties of the patch, it is possible to eliminate expert subjectivity in landscape description process. There is no doubt that expert skills are crucial in decision making process, but shape metrics serve them as a "statement of fact" to support their expert knowledge.

Prior to the shape metrics computing, their selection needed to be done, because calculation of some metrics is time-consuming – Shape Metrics toolbox requires vector data and since vertexes are necessary for complicated formulas of some metrics, calculation time for one single patch takes more than 10 minutes – and therefore those were excluded from the analysis.



Fig. 1. Current Olomouc region with Corine Land Cover layer from 1990 and its position within Czech Republic

Shape metrics	Shape metrics			
Area index	Girth index			
Circumscribing index	Normalized Girth index			
Contiguity index	Gyrate index			
Core index	Perimeter-area ratio index			
Core Area Index	Perimeter index (FRAGSTATS 4.1)			
Number of Core Areas	Perimeter index (Shape Metrics Toolbox)			
Dispersion index	Normalized Perimeter index (Shape Metrics Toolbox)			
Normalized Dispersion index	Proximity index			
Depth index	Normalized Proximity index			
Normalized Depth index	Range index			
Detour index	Normalized Range index			
Normalized Detour index	Shape index			
Exchange index	Spin index			
Normalized Exchanged index	Normalized Spin index			

Table 1. Shape metrics used for geocomputation

Shape metrics in Table 1 were calculated for every single patch in CLC datasets. Next step was to perform Principal Component Analysis (PCA) of shape metrics to substitute the informational rich complete list and set main three components for consequent clustering. These components are in sum carrying 92 % of the original dataset variability and are composed of various metrics (main variance contribution from Gyrate index, Shape index, Core index, Normalized Core index, Proximity index, Exchange index, Spin index, Girth index, Dispersion index, Range index and Detour index). These and other metrics are forming the first, second and third component with different weights. Principal Component Analysis and the estimation of number of clusters could be depicted via graph of similarity of components within various numbers of clusters (Fig. 2).



Fig. 2. A plot of principal components intra-cluster similarity within the specific number of clusters

Figure 2 shows a similarity, based on properties of shape metrics, in given number of clusters according to the method of least squares. It is clear that the similarity within 5 clusters is the highest with the respect of cluster number minimalization. The similarity highly increases between 4 and 5 numbers of clusters and does not significantly increase further. Therefore, it is optimal to cluster the dataset into 5 groups which correspond with the CLC Level 1 nomenclature.

Next step was to perform a cluster analysis. To find the best cluster method, a cluster simulation was run. Overall, 840 combinations of methods and individual settings combinations were given. It is quite subjective phase which cluster method and its settings to chose. It depends on what the user desires to achieve. Nevertheless, the simulation of cluster method suitability was performed using silhouette index. The higher the silhouette index the more suitable a clustering method is. There were only marginal differences among silhouette index values of the best proposed methods and that is why the selection of methods was partly left on researcher subjectivity.

Because there are five categories in CLC Level 1 nomenclature and according to withiness of clusters (Fig. 2), only those cluster methods with highest rank in simulation that define five groups were selected.

The first one was hierarchical method (method which creates tree structure – dendrogram) called DIANA – DIvisive ANAlysis Clustering. The DIANA-algorithm constructs a hierarchy of clusters; starting with one large cluster containing all objects and then the cluster is divided until each cluster contains only a single object [6]. Then, the number of groups is defined, and according to that, values are clustered (Fig. 3). For better interpretation and visualization, colour bars were added. Upper bar is representing desired five target clusters, lower bar is depicting five groups of every single patch from CLC Level 1 nomenclature matching to upper bar.



Fig. 3: DIANA clustering dendrogram with five target clusters.

The second method was non-hierarchical, and partitioning, respectively, which means that dataset is broken up into desired number of groups using medoids (representative objects of a dataset, whose average dissimilarity to all surrounding objects is minimized) and is called PAM – Partitioning Around Medoids. This method is similar to the K-means clustering, but K-means uses means or centroids to cluster a dataset. The PAM is treated to be more robust than K-means because of minimizing dissimilarity instead of Euclidean distances ([7], [13]). Resulting clusters according to the two main components are depicted via 2D graph (Fig. 4).



Fig. 4: PAM clustering 2D graph with five target clusters.

# **3** Results and Comments

Both clustering methods were performed upon shape metrics and their principal components, respectively. Cluster groups were set only according to quantitative values and only non-spatial attribute space of the dataset was performed. Resulting groups are interpreted according to their patch type membership and shape characteristic. Clustering merge patches with the respect of their shape but not directly according to the patch CLC Level 1 type as formerly proposed. Thus, clusters are formed mostly of geometrically similar patches that are, for the most cases, partially patch CLC Level 1 type-independent. Anyway, there are some groups with a significant ratio of one specific patch type category.

First clustering (DIANA) delimitates 5 main clusters (Table 2). Main patch type in the first cluster is agriculture areas (49 %). In the second and third one, main patch type is artificial surfaces (59 %) and (42 %), respectively. Other patch types are not so dominant.

Cluster number	Total number of patches	Total number of patch type					
		Agriculture areas	Artificial surfaces	Forest and semi- natural areas	Water bodies	Wetlands	
1	560	275	124	136	3	22	
2	273	163	67	40	3	0	
3	105	27	44	31	3	0	
4	3	0	0	3	0	0	
5	3	3	0	0	0	0	

 Table 2.
 Number of patches in DIANA clustering.



**Fig. 5.** Olomouc region with Corine Land Cover layer from 1990 and cluster groups according to DIANA clustering method.

Figure 5 shows individual patches classified by DIANA method into 5 groups. For the group number 1 contains mostly very small patches that are close to a minimal size defined by CLC methodology (25 hectars) and are narrowly elongated. Group number 2 incorporates mainly incompact and complex patches (patches with gaps, complicated shapes etc.). Group number 3 is similar to the previous one but patches are more compact (excluding Olomouc city due to its area metrics values) and more regular in their shapes. Groups number 4 and 5 are very similar and are composed of forests and semi-natural areas (Group number 4) and agriculture areas (Group number 5). This is because these two landscape types are represented in GIS as a continuous layer and are extraordinary in all aspects of shape metrics values. The very same principle as in the previous case was used to CLC dataset using PAM method of clustering. Target clusters defined by PAM are in Table 3. It is evident from both Table 3 and Figure 6 that this non-hierarchichal method distributed patches into groups more equally (excluding cluster number 5).

Cluster number	Total number of patches	Total number of patch type					
		Agriculture areas	Artificial surfaces	Forest and semi- natural areas	Water bodies	Wetlands	
1	191	51	82	46	0	12	
2	255	115	62	67	3	8	
3	210	127	41	40	0	2	
4	282	76	146	54	6	0	
5	6	3	0	3	0	0	

 Table 3.
 Number of patches in PAM clustering.

Main patch type in the first and fourth cluster is artificial surfaces (43 %) and (52 %), respectively. In the second and third one, main patch type is agricultural areas (45 %) and (61 %), respectively. Other patch types are not so dominant.



Fig. 6. Olomouc region with Corine Land Cover layer from 1990 and cluster groups according to PAM clustering method

Excepting the group number 1, which is characteristic by containing rather small patches and those narrowly elongated, rest of the groups are the mix of various patches. Forests and semi-natural areas that made up self group using previous DIANA method (group number 4) are now joined with agricultural areas (in DIANA method group number 5) represented in this case by group number 5. Group number 3 contains mainly individual small patches. Barring the group number 1, it is very difficult to find some common characteristics for each group calculated by PAM clustering method. Therefore, it is more suitable in this case to perform analysis of the landscape using DIANA clustering method. However, it depends on the purpose what clustering method to use. If one want to have a complex view onto a landscape, DIANA could be used. On the other hand, PAM identified and pinpointed patches that are narrowly elongated more clearly, thus PAM could serve as a clustering method for elongated patches searching.

Aim of this analysis and calculation was to use clustering methods in order to create distinctive groups of landscape patches. Assumption was that CLC Level 1 patch type is directly influenced by their shape metrics, and vice versa. Ideally, if one of these clustering methods creates same clusters as original types of patches (e.g. artificial surfaces will form their own cluster), it will be very reliable to use them in future automatic classification of any patches. But none of cluster groups in both clustering methods were typical by containing one specific group of patch type in significant amount to claim that e.g. artificial surfaces has very unique shape and thus they form a special group. It is possible to use fuzzy words (e.g. it is more or less "agricultural" cluster) for concluding evaluation statements. Thus, it is needed to analyze patches individually and to search for contexts in detailed level in CLC nomenclature. On the other hand, maybe if larger area would be studied (e.g. entire Czech Republic), the similarity within the cluster would be greater due to the total number of patches involved into shape metrics computation. In other words, proportion of different patch types would not affect final results that much.

Hereby presented procedure could be also modified in the way that input clustering variables will not be principal components, but values of shape metrics themselves. Or another clustering method will be used, regardless to the cluster precision simulation.

Although previously presented results could not provide very convincing results at the first sight, the opposite is true because of the combination of strictly statistical methods together with spatial (visual) evaluation allowed new possibilities of data analysis to arise and unhide clusters of similar areas with similar properties.

Nevertheless, by using above mentioned methods, it is possible to group CLC patches according to their shape similarity, which is useful in a landscape evaluation. Consequent interpretation should take into account the knowledge of shape metrics and the geographic region for which landscape patches are analyzed.

Acknowledgments. The article was created within the project CZ.1.07/2.3.00/20.0170 and CZ.1.07/2.4.00/31.0010, supported by the European Social Fund and the state budget of the Czech Republic.

# References

- EUROPEAN ENVIRONMENTAL AGENCY (2007): EEA Technical report No. 17/2007, CLC2006 technical guidelines, Copenhagen, 2007, 70 p.
- [2] FORMAN, R.T.T (1995): Land Mosaics: The Ecology of Landscapes and Regions. Cambridge University Press, Cambridge, UK
- [3] GUSTAFSON, E. J. (1998): Quantifying Landscape Spatial Pattern: What Is the State of the Art?, Ecosystems, Vol. 1, No. 2., pp. 143–156.
- [4] HEROLD, M., COUCLELIS, H., CLARKE, K. C. (2003): The role of spatial metrics in the analysis and modeling of urban land use change, Computers, Environment and Urban Systems 29, pp. 369–399.
- [5] HUANG, J., LU, X. X., SELLERS, J. M. (2007): A global comparative analysis of urban form: Applying spatial metrics and remote sensing, Landscape and Urban Planning 82, pp. 184–197.
- [6] KAUFMAN, L., ROUSSEEUW, P. J. (1990): Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.
- [7] KUMAR, P., WASAN, S.K. (2011): Comparative Study of K-Means, Pam and Rough K-Means Algorithms Using Cancer Datasets. ISCCC 2009, Proc. of CSIT vol.1.
- [8] MCGARIGAL, K., MARKS, B. J. (2012): FRAGSTATS HELP, University of Massachusetts, 168 p., available from <a href="http://www.umass.edu/landeco>">http://www.umass.edu/landeco></a>.
- [9] MESEV, V. (2007): Integration of GIS and Remote Sensing, Wiley; 1 edition, 312 p.
- [10] NUNGESSER, M. K. (2011): Reading the landscape: temporal and spatial changes in a patterned peatland, Springer, Wetlands Ecological Management 19, pp. 475–493.
- [11] PÁSZTO, V., MAREK, L., TUČEK, P. (2011): Fractal Dimension Calculation for CORINE LandCover Evaluation in GIS – A Case Study., DATESO 2011, VŠB-TU Ostrava, pp.186-195.
- [12] PARENT, J., CIVCO, D., ANGEL, S. (2012): Shape Metrics (presentation). University of Connecticut, ESRI 2009 User Conference, available from < http://clear.uconn.edu/tools/Shape\_Metrics/pubs.htm>.
- [13] PARK, H. S., JUN, C. H. (2009): A simple and fast algorithm for K-medoids clustering, Expert Systems with Applications, 36, (2), pp. 3336–3341.
- [14] SHPUZA, E. (2007): Urban Shapes and Urban Grids: A Comparative Study of Adriatic and Ionian Coastal Cities, Proceedings, 6th International Space Syntax Symposium, İstanbul. 22 p.
- [15] TORRENS, P. M., ALBERTI, M. (2000): Measuring Sprawl. CASA Working Paper 27, UCL London, 34 p.
- [16] TURNER, M. G. (1989): Landscape ecology: the effect of pattern on process. Ann.Rev.Ec.Syst. 20: pp.171–197.
- [17] WU, J. G., JELINSKI, D. E., LUCK, M., TUELLER, P. T. (2000): Multiscale analysis of landscape heterogeneity: Scale variance and pattern metrics. Geographic Information Sciences 6. pp. 6–19.