THEORETICAL VIEW OF THE SHANNON INDEX IN THE EVALUATION OF LANDSCAPE DIVERSITY

RADEK DUŠEK, RENATA POPELKOVÁ

University of Ostrava, Faculty of Science, Department of Physical Geography and Geoecology

ABSTRACT

Shannon's diversity index is frequently used in the determination of landscape diversity. Its indisputable advantage is a possibility to obtain numeric values that can subsequently be easily compared. However, accurate evaluation of landscape diversity from obtained results is rather complicated. The aim of the article is (i) to take a closer look at the theoretical origin of the formula that stems from the principles of the calculation of information entropy and (ii) to draw attention to several issues connected to the Shannon index application in landscape diversity assessment.

Numeric value of the Shannon's index depends on applied logarithm base that is not precisely specified by the formula. Presenting the resulting Shannon index value without stating the logarithm base is not very suitable. Nevertheless, a bigger problem is the dependence of the resulting Shannon's diversity index value on two parameters, namely the number of studied categories and evenness of spatial distribution of individual categories. The resulting value may be identical for different types of the division of the study area. Therefore, the number of categories and the evenness of spatial distribution need to be taken into consideration in the very assessment of the Shannon index result. The number of categories could also be presented along with the resulting Shannon's index value. A major drawback of the Shannon index is its inability to express spatial distribution of patches within the area; it only presents the total extent of each category. Out of existing modifications of the index that try to take spatial distribution into consideration, the most convenient is the coefficient of the distance between the extent of identical and different categories.

Based on arguments deriving from theoretical basis of the Shannon index formula and its practical application, a new view of landscape diversity maximum is presented. The application of the Shannon index disregards the fact that the original relation required for entropy calculation presupposes independence of the existing state (e.g. land cover categories in case of landscape assessment). With regard to the fact that commonly defined categories of patches are independent; the index calculation should make use of the relation considering conditional probabilities of the occurrence of a certain category.

Key words: Shannon index, entropy, landscape diversity, maximum diversity

1. Introduction

Landscape assessment represents a complex activity. Landscape can be described from the point of quality and quantity. Quality description of landscape mosaic focuses particularly on its content. On the base of qualitative characteristics all structures of landscape mosaic are ranged into individual categories, e.g. land cover categories. Quantitative approach deals with quantitative assessment, i.e. possibilities to measure and calculate various values of the landscape structure. Quantitative landscape characteristics are determined by means of landscape metrics describing landscape structure and evolution. A benefit of quantitative values consists in obtaining exact numeric data on the landscape structure that can be compared, e.g. various years within one locality or various localities in individual years (Popelková 2009). The number of landscape metrics is high. For example, McGarigal and Marks (1995) present 100 landscape metrics, many of which are mutually dependent (Cushman et al. 2008). Existing metrics are often modified and completely new metrics occur as well. One of them is a newly created coefficient of mining-based landscape transformation (Mulková, Popelková 2008), which is defined as the ratio of the area that originated due to mining to the area representing original cultural landscape. The coefficient represents the transformation process of original cultural landscape into mining landscape (Mulková 2007).

The use of landscape metrics for the assessment of landscape structure and evolution is, however, connected with many problems. The metrics cannot always be applied to all data. Special attention thus needs to be paid to the interpretation of quantitative data.

Landscape can be characterised by diversity expressing the extent of heterogeneity and variety of landscape structure. In ecology, landscape diversity indices are, for example, the Simpson's diversity index, which is particularly sensitive to species richness, and the Shannon index, which is sensitive to rare species (Farina 2006). Shannon's and Simpson's diversity indices can also be applied to landscape (UMass Landscape Ecology Lab 2012).

If landscape diversity needs to be assessed, it is Shannon's diversity index that is used most frequently, as stated by McGarigal and Marks (1995). The aim of the article is to draw attention to difficulties related to the application of this landscape assessment index.

To be quoted as:
Dušek, R. – Popelková, R. (2012): Theoretical View of the Shannon Index in the Evaluation of Landscape Diversity
AUC Geographica, 47, No. 2, pp. 5–13

2. Theoretic background

2.1 Origin

The author of Shannon's diversity index equation is Claude Elwood Shannon (1916–2001), an American electronic engineer and mathematician, known as the father of information theory. The equation, which was published in *A Mathematical Theory of Communication* in 1948 (Shannon 1948), was derived within information theory and the quantity that it expresses was named *entropy*. Shannon adopted this denomination from Boltzmann's thermodynamic entropy. Although there is formal congruence of the relations within information and thermodynamic entropy, it took many years until their mutual relation was proved. At present, the relation is applied in many scientific fields; among others, in geography and biology for diversity determination.

2.2 Information and entropy

In order to understand the importance of the relation that is used in connection with landscape diversity, its origin within the information theory needs to be shown.

Basic concepts:

– **alphabet** is a set of symbols – the number of symbols *s*, e.g. for the alphabet {a, b, c, d, e} is $s = 5$.

– **message** is a sequence of symbols, e.g. 'babaccdbeae', of the length *n*. In the given example $n = 11$.

The number of possible messages *N* of the length *n* over the alphabet of symbols *s* is calculated as a variation with repetition:

$N = s^n$.

In the above case $N = 5^{11} = 48{,}828{,}125$ possible messages.

We look for a function that can express information extent (*I*) contained in a single message. This function must comply with two requirements:

1. the amount of information in a message depends on the number of possibilities (alternatives) *N* – the higher *N is*, the more information the message contains

$I = f(N) = f(s^n)$,

2. if one message originates as a compilation of two messages ($n = n_1 + n_2$), the amount of information in the resulting message equals the sum of information contained in individual messages:

$I = f(s^{n_1+n_2})$ … the message originates as a compilation of two messages,

$I = I_1 + I_2 = f(s^{n_1}) + f(s^{n_2})$ … the resulting information is the sum of all information,

$f(s^{n_1+n_2}) = f(s^{n_1}) + f(s^{n_2})$ … we look for a function that complies with this equality relation.

The mathematical solution (Shannon 1948) is the equation:

$$I = k \cdot \log(s^n) = k \cdot n \cdot \log s, \tag{1}$$

where *k* is any constant – the issues of the constant are dealt with in the section 'Logarithm base'. The equation expresses the information extent within one message. If we want to express the average information extent for one symbol of the message, then we get

$$H = \frac{1}{n} = k \log s. \tag{2}$$

This relation is valid in the case of equal probability (frequency) of the occurrence of individual symbols in one message. If the symbols appear with dissimilar probability $p_i \in (p_1, p_2 \ldots p_s)$, where $0 \le p_i \le 1$, then after making modifications and using Stirling's formula (Shannon 1948) we get the relation expressing the amount of information for one symbol of a message

$$H = \frac{1}{n} = -k\sum_{i=1}^{s} p_i \log p_i . \tag{3}$$

As stated above, the quantity *H* was denominated entropy (information entropy). The relation (3) presents the features of entropy:

1. entropy only depends on probabilities (not on values of symbols in a message),

2. entropy is invariant to the sequence of symbols,

3. for a concrete *s*, maximum entropy occurs for $p_1 = p_2 = \ldots = p_s = 1/s$, i.e. for steady representation of symbols,

4. minimum entropy occurs for $s = 1$, then $p_1 = p = 1$ and $p \log p = 1 \cdot \log 1 = 1 \cdot 0 = 0 \rightarrow H = 0$.

2.3 Application for landscape diversity

The original equation contained probabilities of phenomena (symbols) that in landscape assessment were substituted with proportional representation of areas of individual categories. Proportional representation of individuals within individual categories is used in order to determine the diversity of species in biology. Shannon's diversity index (*ShI*) used in landscape assessment is defined as follows:

$$ShI = -\sum_{i=1}^{m} P_i \log P_i, \tag{4}$$

where *m* is the number of the studied categories (e.g. land cover categories), P_i is proportional representation of *i*-th category in the total area:

$$P_i = \frac{B_i}{\sum_{i=1}^{m} B_i}, \; B_i \text{ is surface area of } i\text{-th category.}$$

ShI expresses uncertainty with which we are able to predict which category a randomly selected point within the studied area will belong to.

The calculation of *ShI* derives from the fact that a higher value of *ShI* points to higher landscape diversity. Although absolute *ShI* amount can be interpreted with

difficulties, *ShI* is frequently used as a relative index in the comparison of various areas or the same locality in several years (McGarigal and Marks 1995).

The calculation of diversity by means of *ShI* is not the only use case the logarithmic function in geography. The other applications are presented for example in Thomas and Huggett (1980).

3. Practical problems related to the Shannon index application

Equation (1) and subsequently equation (3) were derived from basic requirements of the function *I* or *H*. The equations do not define the values *I* and *H* absolutely clearly because they contain a constant *k* and a logarithmic function. As for the logarithmic function, a parameter is logarithm base *a* which the equations do not determine. With regard to the fact that the logarithmic function generally complies with the requirements of the equations, it is possible to select a random value complying with the condition $a \in R, 0 < a \neq 1$ as the logarithm base. The logarithm base is also related to the constant *k*. In the calculation, the constant is a criterion that decides about the units in which the result will be expressed. The constant can be selected based on the requirements of a concrete equation value for given values *n* and *s*. For example, if the average information extent for one symbol of binary alphabet ($s = 2$) is required to equal one, we can start from equation (2):

$$H = I/n = k \cdot \log_a s,$$
$$H = k \cdot \log_a 2 = 1,$$
$$k = 1/\log_a 2.$$

Using the relation (Rektorys 1963)

$$\log_b x = \frac{\log_a x}{\log_a b} \tag{5}$$

we get: $H = \frac{1}{\log_a 2} \log_a s = \log_2 s$.

The selection of the constant thus determined the logarithm base.

It is the binary logarithm (logarithm to the base 2) that is exclusively used in the information science because each symbol in a message brings one unit of information and the information of the whole message is given by the number of symbols. As it is generally known, a unit of information expressed by means of binary logarithm is the bit. Other commonly used logarithm bases are $e = 2.7182\ldots$ (Euler number) – the so-called natural logarithm and 10 – decadic logarithm.

The question of logarithm base in *ShI* remains often unresolved. Consequently, it is merely the *ShI* value that is presented regardless the concrete logarithm base. This fact is accompanied by disunity in logarithm denotation. Different countries and different fields of science use different logarithm denotations: log, lg, ln, lb, and ld. If the logarithm base used in the calculation is unknown, no comparison of calculated values with the values of other authors is possible. This problem could be solved by conventional modification of the equation in which the logarithm base is presented as an index, as in the case $ShI_2 = 1.12$, which means that the index was calculated by means of a binary logarithm.

3.1 Numeric perspective

For $P_i = 0$, the logarithm reaches the value minus infinity and the expression $P_i \cdot \log P_i$ is indefinite ($0 \cdot -\infty$). On the basis of the limit it is determined that $0 \cdot \log 0 = 0$. There is a drawback in the practical calculation that $\log 0 = -\infty$ can cause calculation collapse. Therefore, there are two methods:

a) make sure zero stays away from the calculation – this is feasible using manual calculation in which zero values are left out,

b) include a condition into the calculation which says that for $P_i = 0$ the calculation disregards the equation and directly substitutes the expression $P_i \cdot \log P_i$ ($0 \cdot \log 0 = 0$) with zero. This way it is necessary to treat all bulk calculations in e.g. spreadsheet or GIS software.

ShI can acquire values coming from real numbers from 0 to $+\infty$, while 0 is the result for $P_i = 1$. The value of P_i decreases with increasing number of categories, namely for $P_i \to 0$ there is, in theory, $\log P_i \to +\infty$. Practically, however, *ShI* never reaches high values – e.g. for a million of evenly represented categories and in case of the binary logarithm $ShI_2 = 20$. As for practical calculations, we are always limited by the number of categories which is always finite. Even theoretically, with an infinite number of categories, we are limited by the accuracy of processed data. If we suppose the accuracy of 1 m^2, in case of e.g. the United Kingdom (243,610 km^2) we are able to distinguish $2.44 \cdot 10^{11}$ of individual patches for which $ShI_2 = 37.8$. Thus, it is impossible to even remotely approximate infinitely high values in practical calculations.

In calculations and subsequent assessment of landscape diversity, it is essential that maximum diversity value is precisely given for a concrete number of patch categories. Maximum diversity value occurs with even representation of categories.

4. Theoretical problems of the application of the Shannon index

4.1 Dependence on two parameters

Equation (4) shows that *ShI* value depends on two parameters. The first parameter is the number of studied

categories *m*; the *ShI* value increases with an increasing number of categories. The other parameter is the evenness of the spatial representation of individual categories; maximum *ShI* value is reached with fully even spatial representation of categories. The dependence of the only value of the result on two parameters is a drawback of the equation for *ShI* calculation. Equal diversity value is thus observed in the territory whose 90% is dominated by one category and the rest 10% is evenly divided into 10 parts (total number of categories in the area is thus 11) as well as in the territory that contains two categories in a ratio of 24.4 : 75.6. In both cases ShI_2 makes 0.8. The situation is graphically represented in Figure 1. The figure shows that *ShI* prefers large areas to small areas.

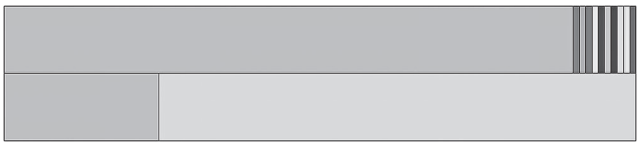

Fig. 1 Example of two areas with a different number of categories but equal *ShI* value

The dependence of *ShI* on two variables is given in Figure 2. The figure shows an isoline graph of *ShI* values. The calculation was carried out as follows:

1. the so-called basic part was singled out of the whole studied area (100%). The size of this basic part ranged from 1% (1/100 of the area) up to 50% (1/2 of the area) – the size of the basic part is presented on the horizontal axis in percentage,

2. the rest of the area (100% minus the basic part) was gradually divided into one, two, three up to fifty equal segments – the number of the segments is presented on the vertical axis,

3. providing each segment represents a different category, for the area divided in this way, *ShI* was calculated in the network of coordinates 1 × 1 and isolines were drawn.

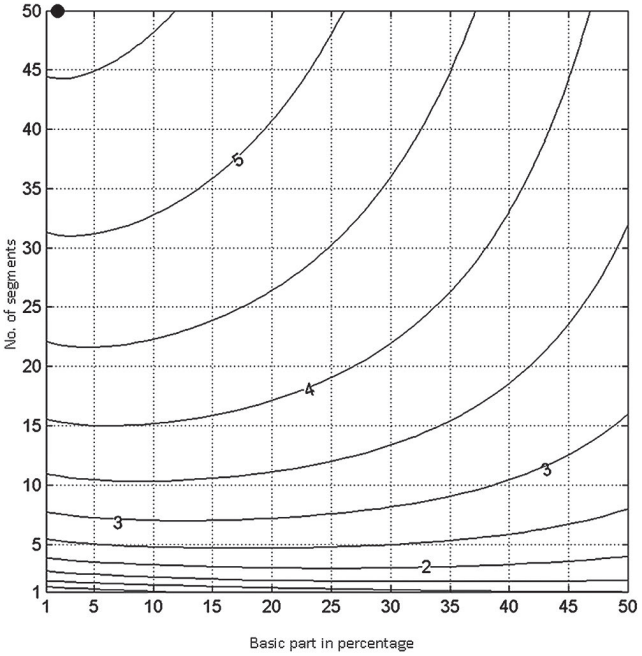

Fig. 2 *ShI* isoline graph for different types of area division – explanation in the text

As the graph shows, the highest *ShI* values are reached in the upper left corner (the upper limit is marked with a black point). A clear perspective view of the graph is given in Figure 3. The exact position of the upper limit occupies the coordinates $x = 1.96078 \doteq 1.96$; y = 50. In this case, the size of the basic part is 1.96% and the rest (98.04%) is divided into 50 segments. The size of one segment is 98.04/50 = 1.96. The whole area is therefore divided into 51 equal segments and ShI_2 = 5.672. The lower limit lies on the coordinates [1; 1], which correspond to the basic part equal to 1%, and the rest is divided into one segment – which means it is not further divided. The territory consists of two parts whose areas are in a ratio of 1 : 99 ShI_2 = 0.081. Then, ShI_2 = 1 for the coordinates [50; 1].

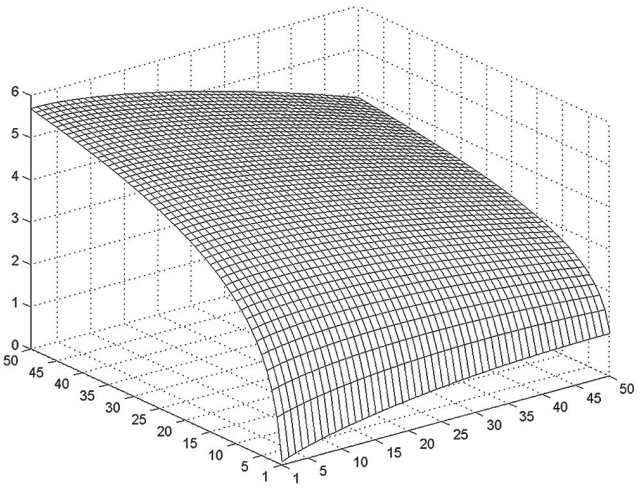

Fig. 3 Perspective view of the graph from Figure 2

Figure 2 demonstrates that *ShI* isolines are curves and the same index value thus holds true for more (in theory, infinitely many) variants of the area division. For this reason, the *ShI* interpretation is difficult.

Taking a formal view of diversity, the above described drawback of *ShI* can turn into a seeming asset. If we make little effort to understand diversity and simplify its definition to e.g. the Shannon index, then making an easy calculation we obtain a single number with which we can further work relatively blithely.

Similarly to the logarithm base, also in the case of the number of categories, it would be suitable to present the information together with *ShI* value. As for the example in Figure 1, it is convenient to distinguish $ShI_2 = 0.8\{11\}$ and $ShI_2 = 0.8\{2\}$, i.e. *ShI* calculated from eleven categories and two categories respectively. Gallego et al. (2000) use the denomination 'SH9, SH23' for the index calculated for 9 and 23 categories respectively. As in the case of logarithm base, concrete graphic form of the presentation of the number of categories is a question of convention. There is a multitude of possibilities (e.g. $ShI7_2 = 0.91$; $ShI(2) = 0.91[7]$; $ShI = 0.91[7/2]$; ${}^{7}ShI_2 = 0.91\ldots$) and the main problem is not how to visualise but to start visualising.

4.2 Diversity solving in the plane

Similarly to many other landscape metrics, a fundamental problem of *ShI* is that it is not explicitly spatial. Within the studied area, no attention is paid to spatial distribution of individual patch types with different land cover. This problem is demonstrated on two different territories in Figure 4. Both visualised territories contain three categories with spatial representation in a ratio of 36 : 34 : 30. Both the areas thus present identical $ShI_2 = 1.58\{3\}$. With the naked eye we can see in the figure that diversity within the first territory (Situation I) is higher than diversity within the other territory (Situation II), which, however, *ShI* results do not disclose. That is a fundamental problem of *ShI* related to its application. Landscape diversity should reflect not only the number of patches and their total extent but also the spatial distribution of individual landscape elements. *ShI*, however, is unable to express this information.

Although a majority of *ShI* users fail to occupy themselves with the issues of spatial distribution, some authors are evidently trying to deal with this disadvantage. One of the first authors to deal with the problem was Batty (1974) who used entropy for regionalization. Applying entropy in human geography, he completed the calculation with a parameter of the zone extent. Subsequently using the iterative technique, he clustered zones in order to achieve maximum entropy. One of his thoughts presented in the conclusion of his work can be mentioned in connection with the *ShI* application: 'It is unbelievable how many [geographic analyses] ignore the question of space' (sad but true this statement is even in this time of advanced geoinformation technologies). Batty's method was employed by e.g. Paszto et al. (2009) in cartographic applications.

A different approach was selected by Gorelick (2006) who introduced a matrix analogue of the Shannon's and Simpson's indices. This procedure is an interesting transfer of *ShI* into 2D space. It can be used in theoretical solving of some problems; however, it is not suitable for practical application related to a territory of a general shape.

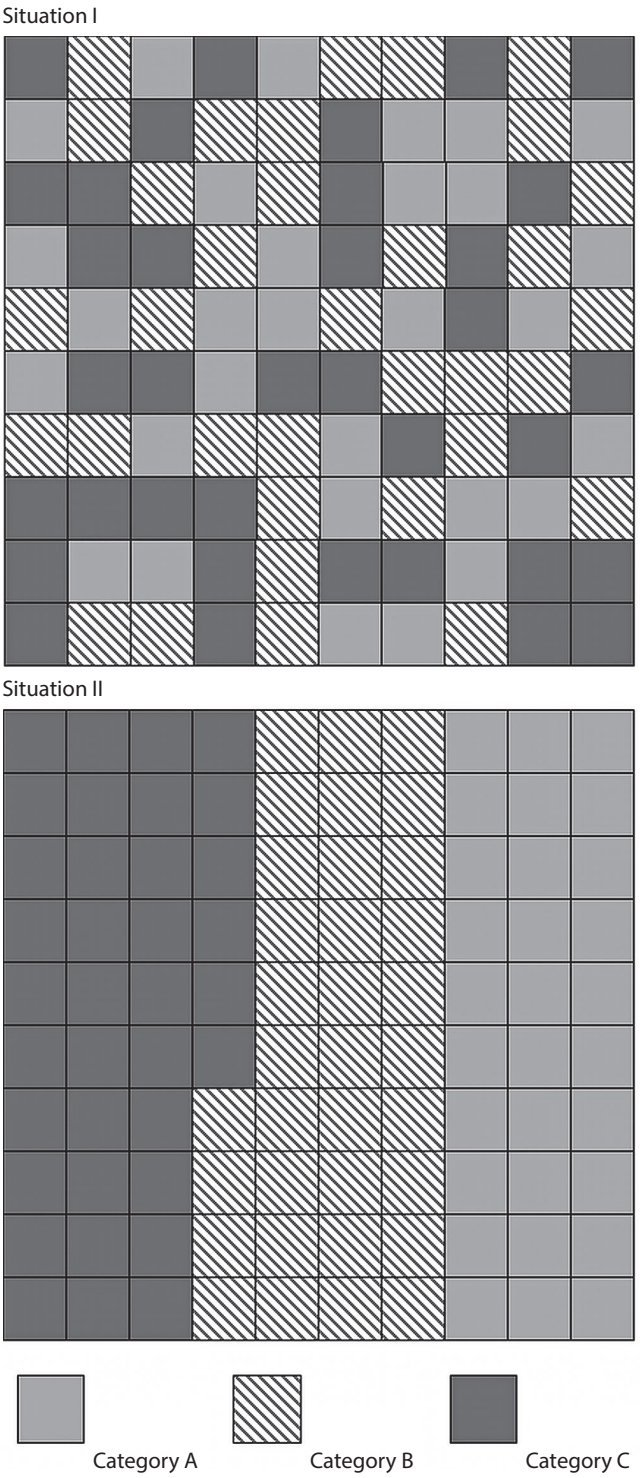

Fig. 4 Two examples of distribution of 3 categories in the area

Spatial distribution was successfully taken into consideration by Claramunt (2005). His method was applied by Li and Claramunt (2006) and Wang and Wang (2011). Its principle is the calculation of average distances between patches of the same category and patches of different categories. The equation (4) was modified into the form

$$ShI = -\sum_{i=1}^{m} \frac{d_i^{int}}{d_i^{ext}}(P_i \log P_i) , \quad (6)$$

where d^{int} is an average distance between areas of the same category ('inner distances'), d^{ext} is an average distance to areas of other categories (the distances relate to gravity centres of areas). The fraction is practically a balance which takes into consideration whether the areas of individual categories are clustered or 'dispersed' all around the territory. Examples of values calculated for the situations in Figure 4 are presented in Tab. 1 and Tab. 2.

Resulting *ShI* values after spatial distribution of individual areas has been taken into consideration:

Situation I: *ShI* = 1.5597.

Situation II: *ShI* = 0.9836.

The original value is equal for both situations: *ShI* = 1.5809.

The results show that these values rather correspond to general notion of landscape diversity. Advantage of this method rests in the fact that it can be used with raster data format in which each raster cell represents a single patch (as shown in Figure 4, Situation II). With regard to regular cell shape, there is no need to determine the distance between areas of a general shape.

Despite interesting results, the method brings a new problem, which is the theoretic base (not connected only with this method, but all *ShI* modifications). *ShI* has an explicit theoretic base and even though its interpretation in landscape assessment is difficult, it is clear what base it stems from and what conditions it complies with. The modification of the *ShI* equation does not mean 'modified Shannon index' but new characteristics with a different theoretic starting point that does not necessarily have to have anything in common with the original Shannon entropy.

Another problem of the calculation according to (6) is the determination of maximum diversity. The complicated parameter of average distances practically makes it impossible to exactly determine maximum possible diversity.

If we use landscape diversity to infer the species diversity, another important indicator, apart from the number of patches and their distribution, is also boundary segmentation and boundary length. However, no information of this parameter is brought by *ShI*. Unlike the spatial distribution of landscape elements, the issue of the boundaries is not solved by means of *ShI* modification, but different landscape metrics such as the total boundary length, boundary density and fractal dimension (McGarigal and Marks 1995).

4.3 Independence of categories

Shannon's derivation of entropy starts from a premise of independent occurrence of individual symbols in a message. This premise is valid for a randomly generated message, but not for a majority of natural systems. An example of non-random occurrence of symbols is a text in English or another language in which some pairs (trios …) of letters are more probable than others. Shannon (1948) presents the following hierarchy:

– zeroth-order approximation … symbols are independent and characterised by equal probability,

– first-order approximation … symbols are independent and characterised by dissimilar probability – it is a model for which entropy or *ShI* are derived,

– second-order approximation … symbols are characterised by dissimilar probability which depends on the previous symbol (some pairs of symbols occur more often that others),

– third-order approximation … symbols are characterised by dissimilar probability which depends on previous two symbols (some trios of symbols occur more often than others),

– etc.

Tab. 1 Values of average distances between categories

	A–A	B–B	C–C	A–B	A–C	B–C
Situation I	2.4635	2.5194	2.7126	2.5085	2.6262	2.6549
Situation II	1.7989	1.8575	1.8343	2.5009	3.8647	2.5903

Tab. 2 Values of the numerator and the denominator of a fraction for individual categories

	Situation I			Situation II		
	Category A	Category B	Category C	Category A	Category B	Category C
Numerator	2.4635	2.5194	2.7126	1.7989	1.8575	1.8343
Denominator	2.5690	2.5884	2.6415	3.2023	2.5497	3.1877
Fraction	0.9589	0.9734	1.0269	0.5618	0.7285	0.5754

In case of the second-order approximation and higher-order approximation it is necessary to consider aggregate or conditional probabilities.

Mutual independence of individual categories in landscape diversity assessment is given by a concrete definition of the categories. Commonly used categories are not independent. For example, the categories of arable land, forest and water bodies can be considered independent. On the contrary, the categories of industrial and commercial units and road and rail networks are dependent categories because the occurrence of road and rail networks is closely related to the occurrence of industrial and commercial units. A similar situation holds good for water bodies, water streams and other categories.

If we wanted to preserve entropy as a principle of diversity calculation, we would have to deal with at least the second-order approximation. This would mean being acquainted with estimates of mutual dependences of pairs of categories and applying them in the calculation. Such a method is not unusual and it is equation (4) that is almost exclusively used, even though land cover characteristics do not comply with its demands.

4.4 Maximum diversity

In addition to direct *ShI* calculation used in landscape assessment, we can also use maximum diversity, which is calculated:

$$ShI_{\max} = \log m \tag{7}$$

The denomination indicates that it concerns the highest possible value of the Shannon index that can be reached within a given territory. *ShI* reaches the highest values for regular representation of individual categories if the relation (4) is simplified to (7). Almost all authors present *m* as a number of land cover categories. However, a question is if it represents a number of categories occurring in the study area or a theoretically possible number of all the categories. Maximum diversity is generally calculated for the number of categories occurring in a given territory. Although the resulting value has a certain reporting ability, it is not maximum diversity. Maximum diversity can be obtained if *m* represents a number of all potential categories – this number depends on how the categories are defined. For example, the EU CORINE Land Cover (CLC) methodology contains 44 land cover categories.

Arguments for the use of all categories in the calculation of maximum diversity:

1. If we really want to achieve maximum entropy, the use of all categories brings a higher value.

2. In a calculation made on the basis of existing categories we only work with one parameter, which is the distribution of categories. In reality, diversity is dependent on two parameters, neither of which needs to be favoured. Using the existing categories, one parameter is fixed and we try to reach the maximum by changing the other parameter. It is difficult to imagine that we fix the distribution of categories while changing the number of categories because the change in the number of categories inevitably causes changes in the distribution of categories. Still, we can imagine a situation in which the distribution of categories is given and the maximum is reached by means of the number of categories. An extreme but easily understandable example is regular representation of categories. How can maximum diversity be achieved with regular representation of categories? We need to increase the number of categories to its maximum; the maximum is all the categories within used methodology.

3. The unsuitability of calculating maximum diversity using the number of existing categories becomes disclosed for $m = 1$. Then $ShI = 0$ and also $ShI_{\max} = 0$. We get into an absurd situation in which the smallest possible diversity is at the same time the maximum diversity.

4. Within the information theory entropy is calculated as a sum made not only over symbols contained in a message but over all the symbols of the alphabet (all possible states). Maximum entropy is then logically calculated from the number of alphabet symbols. If the land area of categories is an analogue of the probability of the symbol (state) occurrence, diversity is calculated over all categories and in the same way maximum diversity should be calculated.

The calculation of maximum diversity is related to the comparison of the existing diversity with its maximum value. A different case is the calculation of Shannon's Evenness Index (*ShEI*) which also contains 'maximum' diversity in the denominator: *ShEI* is supposed to express the regularity in the representation of individual categories, which is why the denominator is logically calculated from the existing categories. In this case it is not suitable to use the term 'maximum' diversity for the denominator. Instead of the equations $ShEI = \frac{-\sum_{i=1}^{m}(P_i \log P_i)}{ShI_{\max}} = \frac{-\sum_{i=1}^{m}(P_i \log P_i)}{\log m}$ it would be more suitable to use:

$$ShEI = \frac{-\sum_{i=1}^{m}(P_i \log P_i)}{ShI'} = \frac{-\sum_{i=1}^{m}(P_i \log P_i)}{\log n},$$ where

n is a number of categories occurring in the territory, $n \leq m$.

5. Recommendation and conclusion

'A model of reality is created emphasizing certain connections, while (inevitably) neglecting others,' present Guiasu and Shenitzer (1985) in connection with the maximum entropy used for solving mathematical models of reality. This statement precisely depicts the situation in which *ShI* tries to express very complex landscape

characteristics using a single figure. Despite the fact that the Shannon index is the most popular diversity index (McGarigal and Marks 1995), its application is connected with a number of problems that a majority of users are unaware of or fail to deal with. Mechanical application of *ShI* is not suitable since the resulting values can bring distorted characterisation of the studied locality. From the above mentioned we can derive the following recommendations for the application of *ShI*:

1. be aware of the origin of the equation,
2. avoid overestimating the *ShI* value – it does not represent universal characteristics of landscape diversity,
3. avoid presenting the resulting numeric value only; include parameters that are essential for further comparison of the results – logarithm base and the number of categories,
4. be aware of insufficient information related to spatial distribution; use *ShI* modification that takes spatial distribution into consideration,
5. express maximum landscape diversity advisedly.

5.1 Conclusion

As it is necessary to approach the application of Shannon's diversity index critically, it is also necessary to pay close attention to other concepts and procedures related to landscape diversity. With regard to the fact that on one hand, diversity is a very complex feature and, on the other hand, it is a very popular phenomenon, there are many other improper simplifications. For example, the statement that higher diversity leads to higher landscape stability is completely misleading if the type and quality of patches are disregarded. Therefore, the understanding and characterisation of landscape diversity require a more complex approach which avoids using a single figure or a simple term but focuses on capturing diversity from a complex perspective, for example La Rosa, Martinico, Privitera (2011), Fahrig, Baudry, Brotons et al. (2011).

Acknowledgements

This research was supported by a grant project of the Grant Agency of the Czech Republic No. P410/12/0487: 'The process of industrialization and landscape changes in the Ostrava Industrial Zone in the 19th and 20th centuries'.

REFERENCES

BATTY, M. (1974): Spatial Entropy. Geographical Analysis, 6, pp. 1–31.

CLARAMUNT, C. (2005): A Spatial Form of Diversity. In: COHN, A. G., MARK, D. M. (eds.) COSIT 2005. Lecture Notes in Computer Science, 3693, pp. 218–231. Springer, Heidelberg.

CUSHMAN, S. A., McGARIGAL, K., NEEL, M. (2008): Parsimony in landscape metrics: strength, universality, and consistency. Ecological Indicators, 8, pp. 691–703.

FARINA, A. (2006): Principles and Methods in Landscape Ecology. Springer, The Netherlands, 537 p.

FAHRIG, L., BAUDRY, J., BROTONS, L. et al. (2011): Functional landscape heterogeneity and animal biodiversity in agricultural landscapes. Ecology Letters, 14, pp. 101–112

GALLEGO, J., ESCRIBANO, P., CHRISTENSEN, S. (2000): Comparability of landscape diversity indicators in the European Union, pp. 84–97.

GORELICK, R. (2006): Combining richness and abundance into a single diversity index using matrix analogues of Shannon's and Simpson's indices. Ecography, 29, pp. 525–530.

GUIASU, S., SHENITZER, A. (1985): The Principle of Maximum Entropy. The Mathematical Intelligencer, 7(1), pp. 42–48. Springer-Verlag New York.

LA ROSA, D., MARTINICO, F., PRIVITERA, R. (2011): Dimensions of landscape diversity: ecological indicators for landscape protection planning. RegioResources 21 Conference The European Land Use Institute. [Presentations], 25 p.

LI, X., CLARAMUNT, C. (2006): A Spatial Entropy-Based Decision Tree for Classification of Geographical Information. Transactions in GIS, 10(3), pp. 451–467.

McGARIGAL, K., MARKS, B. J. (1995): FRAGSTATS: spatial pattern analysis program for quantifying landscape structure. USDA For. Serv. Gen. Tech. Rep. PNW-351., 141 p., [online]. [cit. 31. 1. 2012]. Available at: http://www.umass.edu/landeco/pubs/mcgarigal.marks.1995.pdf .

MULKOVÁ, M. (2007): Využití konvenčních metod DPZ při sledování antropogenních změn krajiny v poddolovaných oblastech. [Ph.D. Thesis]. Masarykova univerzita v Brně, 166 p.

MULKOVÁ, M., POPELKOVÁ, R. (2008): Cultural and Mining Landscapes: Different Changes and Different Methods to Evaluate the Landscape Development. In: SVATOŇOVÁ, H. (ed.) Geography in Czechia and Slovakia. Theory and Practice at the Onset of 21st century, Brno, pp. 153–159.

PASZTO, V., TUČEK, P., VOŽENÍLEK, V. (2009): On Spatial Entropy in Geographical Data. Sborník příspěvků 16. konference GIS Ostrava.

POPELKOVÁ, R. (2009): Retrospektivní analýza vývoje krajiny s využitím geoinformačních technologií [Ph.D. Thesis] Vysoká škola báňská – Technická univerzita Ostrava, Hornicko-geologická fakulta, Institut geoinformatiky, 168 p.

REKTORYS, K. et al. (1963): Přehled užité matematiky. Česká matice technická, Praha 1963.

ROCCHINI, D. (2005): Resolution Problems in Calculating Landscape Metrics. Spatial Science, 50(2), pp. 25–36.

SHANNON, C. E. (1948): A Mathematical Theory of Communication. The Bell System Technical Journal, 27, pp. 379–423, 623–656, July, October.

THOMAS, R. W., HUGGETT, R. J. (1980): Modeling in geography: A mathematical approach. Methuen, London, 338 p.

UMass Landscape Ecology Lab [online]. [cit. 1. 2. 2012]. Available at: http://www.umass.edu/landeco/.

WANG, B., WANG X. (2011): Spatial Entropy-Based Clustering for Mining Data with Spatial Correlation. HUANG, J. Z., CAO, L., SRIVASTAVA, J. (eds.): PAKDD 2011, Part I, LNAI 6634, pp. 196–208, Springer-Verlag Berlin Heidelberg.

RÉSUMÉ

Teoretický pohled na Shannonův index při hodnocení diverzity krajiny

Příspěvek se zabývá teoretickými aspekty používání Shannonova idexu při hodnocení diverzity krajiny a rozebírá problémy, které s sebou přináší jeho formální aplikace. Základním východiskem je původní teoretický základ vzorce, který vychází z principů výpočtu informační entropie. Analyzována je závislost numerické hodnoty indexu na dvou parametrech, kterými jsou počet sledovaných kategorií a rovnoměrnost plošného zastoupení vyskytujících se kategorií. Vzhledem k tomu, že pro více způsobů rozdělení zájmového území může být výsledná hodnota Shannonova indexu stejná, je nutné při hodnocení výsledku přihlížet k počtu kategorií i k rovnoměrnosti plošného zastoupení jednotlivých kategorií. Proto je doporučeno uvádět u výsledné hodnoty Shannonova indexu také počet kategorií.

Základním nedostatkem Shannonova indexu je skutečnost, že nevyjadřuje prostorové uspořádání ploch kategorií v území, ale pracuje pouze s celkovými velikostmi kategorií. Je uveden přehled a provedeno zhodnocení existujících modifikací indexu, které se snaží prostorové uspořádání zohlednit. Jako nejvhodnější se jeví využití koeficientu, který zohledňuje vzdálenosti mezi plochami stejných a odlišných kategorií. Tato modifikace je podrobně představena na modelovém příkladu.

Na základě argumentů vycházejících z teoretického základu vzorce pro výpočet Shannonova indexu a z praktických aplikací tohoto vzorce je také představen nový pohled na pojem maximální diverzita krajiny.

Při užívání Shannonova indexu je zcela opomíjeno, že původní vztah pro výpočet entropie předpokládá nezávislost vyskytujících se stavů (v případě hodnocení krajiny např. kategorií krajinného pokryvu). Vzhledem k tomu, že běžně definované kategorie ploch nejsou nezávislé, měl by se pro výpočet Shannonova indexu užívat vztah zohledňující podmíněné pravděpodobnosti výskytu kategorií.

Radek Dušek
University of Ostrava
Faculty of Science
Department of Physical Geography and Geoecology
Chittussiho 10
710 00 Ostrava-Slezská Ostrava
Czech Republic
E-mail: radek.dusek@osu.cz

Renata Popelková
University of Ostrava
Faculty of Science
Department of Physical Geography and Geoecology
Chittussiho 10
710 00 Ostrava-Slezská Ostrava
Czech Republic
E-mail: renata.popelkova@osu.cz