_____

ON SPATIAL ENTROPY IN GEOGRAPHICAL DATA

Vít PÁSZTO[1], Pavel TUČEK[1,2], Vít VOŽENÍLEK[1]

[1]*Department of Geoinformatics, Faculty of Science, Palacký University in Olomouc,,
Tř. Svobody 26, 771 46 Olomouc, Czech Republic*

[2]*Department of Mathematical Analysis and Applied Mathematics, Faculty of Science, Palacký University in
Olomouc, Tomkova 40, 779 00 Olomouc - Hejčín*

vit_p@atlas.cz, pavel.tucek@upol.cz, vit.vozenilek@upol.cz

Abstract: The aim of the article is to make one keep that the using of the computation of the entropy and spatial entropy in geographical sciences should be performed in every occasions in order to obtain better results. There is always a loss of information during the process of visualization and the amount of information has never been taken into account in the visualization and modelling of the spatial phenomena. Abstract concept of the entropy has penetrated into many scientific fields and thus has got many definitions. The entropy in the processing of geographical data represents some kind of uncertainty of the upcoming result. The spatial distribution of an investigated phenomenon that should be visualized or modelled is related to many factors. An understanding of the essence requires to select the optimum distribution of the measured data in order to obtain the maximum information. In theory, a well-designed data network can accurately represent and provide the needed information of a phenomenon. Another question which should be asked is the consecutive prediction of the given problem. This situation is much more complicated and should be taken into account also.

Keywords: GIS, Spatial Entropy, Information Theory, Visualization

1. Introduction

The concept of the entropy is relatively new, defined after the World War II. As one of elemental components of the Information theory, which was the basis of designing computers and for the informatics at all. Entropy expresses the amount of information in a message and quantifies the abstract concept of information, which is the basis of computer science.

Entropy was applied in many other fields of science (medicine, biology, chemistry, etc.) throughout the time, including geography and geoinformatics. The spatial entropy, entropy with spatial aspects, was used in the 1970s, and took into account the spatial distribution of geo-phenomenon.

The entry refers what are new possibilities of using the entropy supported by geoinformatics and GIS. Robust analysis and visualization tools provided in GIS give a new meaning of the entropy and enable studying and displaying geographic phenomenon and processes in a new relationship.

This article contains the overview that should serve as a motivation to the analysis of the entropy in the cartography, spatial modelling and visualization. We made also a particular processing of the given geographical data in order to show the necesarity of the entropy calculation.

2. Information theory

Basis on which the Information theory is built were found by Claude E. Shannon, a famous mathematician and electronic engineer. He formed an idea of describing the information using mathematics. Mathematic theory of information describes the information with statistical apparatus, such as probability, and according to [7] it investigates a mathematical representation of conditions and parameters, which influence transmission and receiving the signal and information, respectively.

There is always loss of information during the process of sending, transmitting and receiving a message (This message can be a chart, because it is a document providing a great sum of information [13]). Or if the information is coming through the system it can be enhanced by the noise of external source. This enhancement is, however, undesirable. Basic scheme of the information flow and components of a communication system is shown in Fig. 1.

Either the loss or enhancement of the information causes that the amount of original information is lower. Since Shannon developed the theory, we have been able to measure quantity of expected information.
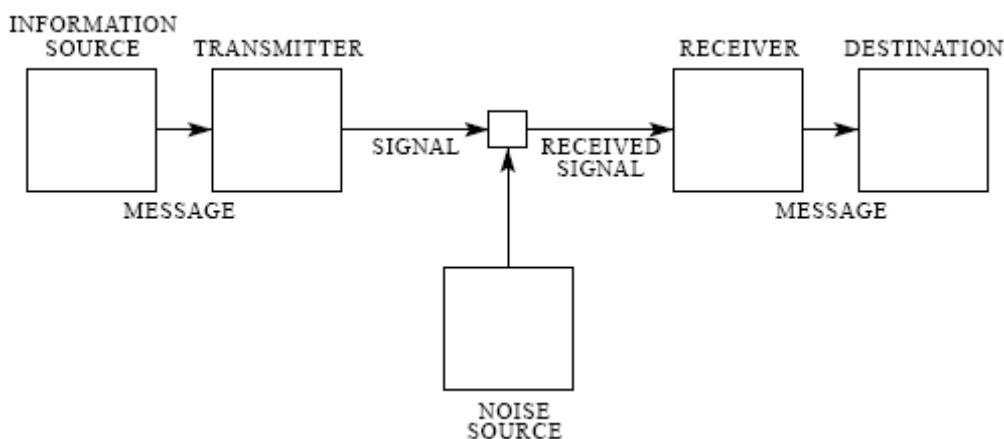

Fig. 1: General communication system [11].

In a geographical context, the information source could be tyflo-chart (chart for blind people), the transmitter is, in this case, a human touch. The transmitted signal could be disturbed by the noise source, e.g. damaged Braille's alphabet on the chart. The receiver and destination is then a blind person. How much geographic information was communicated? Does the reader have the information he needed?

There exist plenty of factors that affect these questions. Language of the chart, its quality, amount of the noise, experiences of the reader, conditions in which the chart is read, etc. All these factors have their probability of occurrence, which are one of the elements that impact measure of the information. In the other words, it can be evaluated, if (and how much) the information is useful for the receiver and helps to remove uncertainty and nascence of a certain phenomenon. One can answer the question how rich was the information?

This question could be turned upside down. How much uncertainty was removed? This uncertainty is generally called the entropy. As it is in [8], it is exactly the entropy, which satisfies demands for the uncertainty rate of the situation.

Another examples of using the information theory in cartography are described in [9], and thus can be used in geoinformatics and GIS as well [13].

_____

3. Entropy

Abstract concept of the entropy has penetrated into many scientific fields and thus has got many definitions. The entropy represents some kind of uncertainty of the system before receiving information. As it is mentioned in [8], the uncertainty rate associated with relevant situation is also the amount of information needed to eliminate this uncertainty. It means that the entropy determines how much information must be added to understand the problem.

Pechanec [10] provides an example with a person as a receiver of the information. Removing the uncertainty (entropy) can be insignificant, if this person has either a lack of knowledge or, on the contrary, a great experience with the problem of his/her interest. In the first case, the person would not be able to understand provided information; in the second, provided information brings no facts that the person has already known. Then, the suitable person would be an intermediate one.

The entropy is generally defined as *an average value of information rate to eliminate uncertainty, which is given by finite number of alternative events.* Shannon [11] defines the entropy as

$$H(S) = -\sum_{i=1}^{n} P(s_i) \log_2 P(s_i) \tag{1}$$

Where S is the system with finite number of possible events s_i, $P(s_i)$ represents the probability of event s_i occurring and the summation is over the range s_i = 1, 2, ..., n.

It is interesting to note (according to [11]), that $H = 0$ (entropy is minimal and no information is needed) if and only if all the $P(s_i)$ but one are zero, then this one having the value unity and it is the one and only case when H vanishes. Otherwise H is positive (because of the minus sign). Entropy H is maximum and then equal to $\log_2 n$ when all the $P(s_i)$ are equal. This means that the most uncertain situation occurs.

As an example can serve a classical fair coin flip. It is known that the probability of showing a head or tail is the same – 50 %. Then the uncertainty (entropy) is maximal. In the other case if there is a tennis match and on of the players becomes injured, it is evident, who is going to win the match by default. The probability is of one event is 100 % and the entropy is minimal. Fig. 2 shows the entropy in the case of two alternatives.
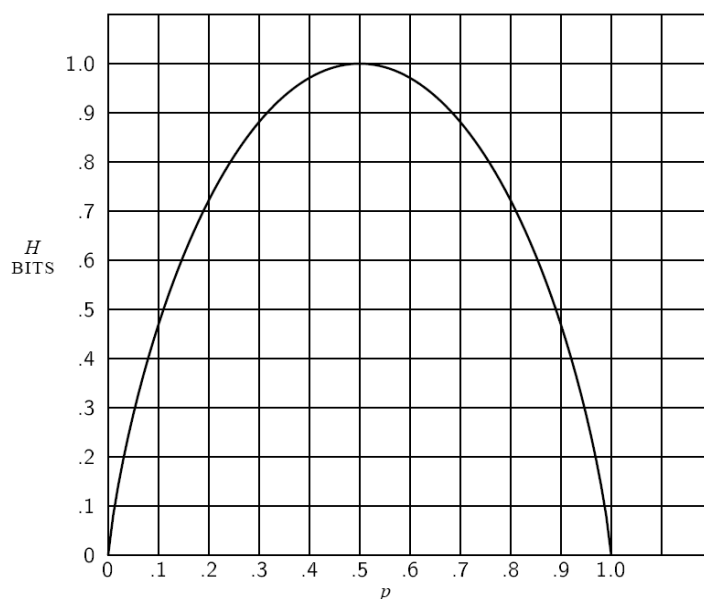

Fig. 2: Entropy in the case of two possibilities with probabilities *p* and (1 - *p*). [11]

The entropy units in Fig. 2 are in *bits* and the binary logarithm is used (log$_2$). If the common logarithm (log) is used, then units are *dits*, and units for natural logarithm (ln) are called *nits*.

Schematic diagram of eliminating the entropy (uncertainty) from a various systems *S* is shown in Fig. 3. The system (e.g. students in school) is at the beginning fully uncertain (disarranged); the entropy is at maximum and equal to 1. The information is added passing the time and the entropy is subsequently removed. After the whole relevant information is gained, the system becomes fully organized and the entropy is equal to 0. No further information is needed.
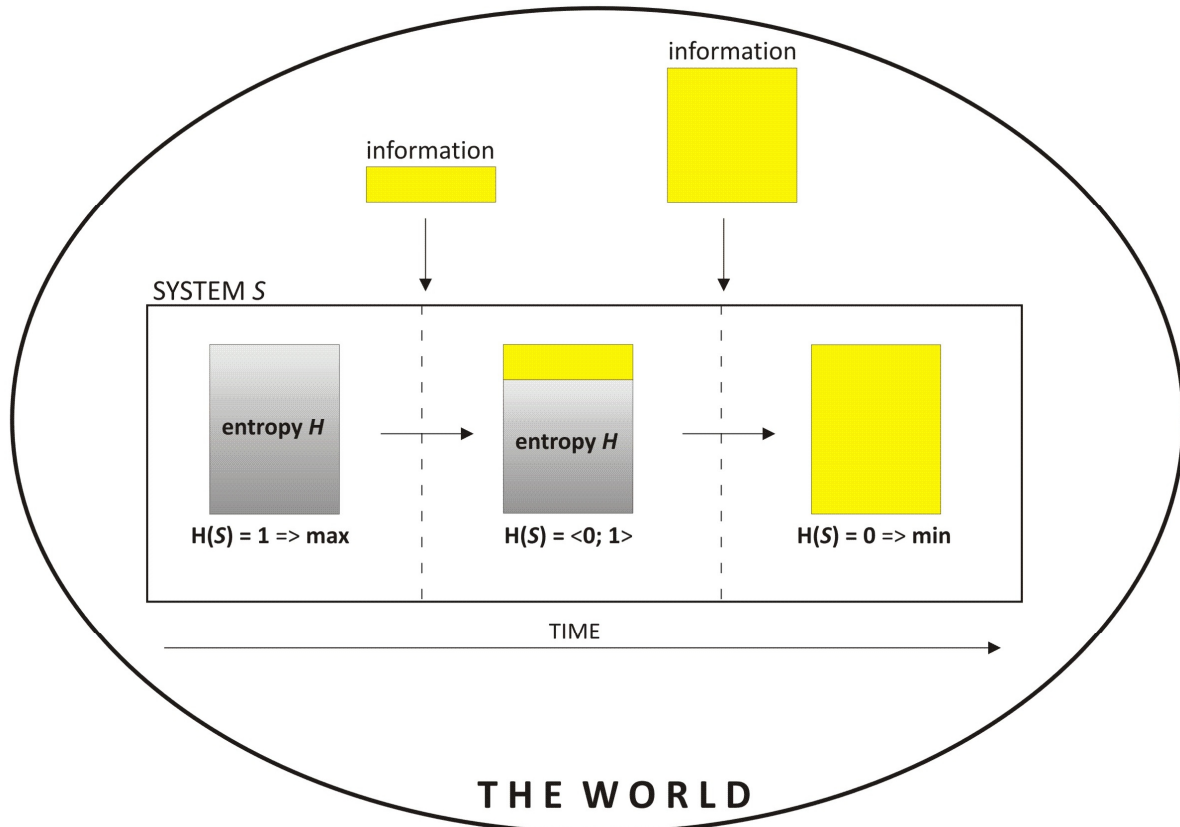

Fig. 3: Schematic diagram of eliminating the entropy.

Geography and geoinformatics are working with spatial information. Equation (1) has got no spatial pattern that is why an additional research was investigated. Geographers used entropy function and statistics in deriving the most probable profile of a river system, to build the distribution models of settlement, to analyze migration flows, to analyze spatial series and so on [3].

But it was Batty, who defined the *spatial entropy* function from the information theory basis previously presented by Shannon. In [3] the author derives equation (2) from (1) and shows the application in spatial zoning of the city. His later papers [e.g. 1, 2] are dealing with usage of (2) and its modifications in spatial aggregation, spatial hierarchy and zoning systems as well as the spatial geometry in general. Formula for *spatial entropy* in which Δx_i represents the spatial interval size is:

$$H = -\sum_i p_i \ln\left(\frac{p_i}{\Delta x_i}\right)$$

(2)

where p_i (as in equation (1)) represents the probability of event *i* occurring. The spatial component is implemented in equation (2) and this equation is thus more applicable in spatial analysis, such as comparisons between different regions [3].

The application of this formula, and its modifications, to some geographical problems are in great investigations of Department of Geoinformatics on Palacky University in Olomouc at the moment. Research is aimed at the usage of GIS and its visualization tools in this problem and implementation of the entropy function in a process of defining the optimal number of intervals of various phenomenon, classification of mutuality of different phenomenon [9], setting the optimal information preserved in a chart considering the size, measure, purpose (and other cartographic rules as in [14, 6, 4]) of the chart and its reader, and many more.

4. Entropy as a useful part of a spatial visualization and modelling

The spatial distribution of an investigated phenomenon that should be visualized is related to many factors. An understanding of the essence requires to select the optimum distribution of the measured data in order to obtain the maximum information. In theory, a well-designed data network can accurately represent and provide the needed information of a phenomenon.

More complicated is the situation in spatial prediction of the given problem. Kriging as an interpolator, which performs linear averaging on the basis of the observed data, is often used to compute the spatial variations [5]. A method composed of kriging and entropy can determine the optimum number and spatial distribution. Thus, the data at the locations can be reconstructed. The information entropy reveals the information of the each point. By calculating the entropy and the transmitted information, the candidate area for the measurement are prioritized.

Thus, the optimum spatial distribution and the minimum number observation in the network can be determined. The result shows that there exists the optimum number and spatial distribution of the data in order to provide the necessary information.
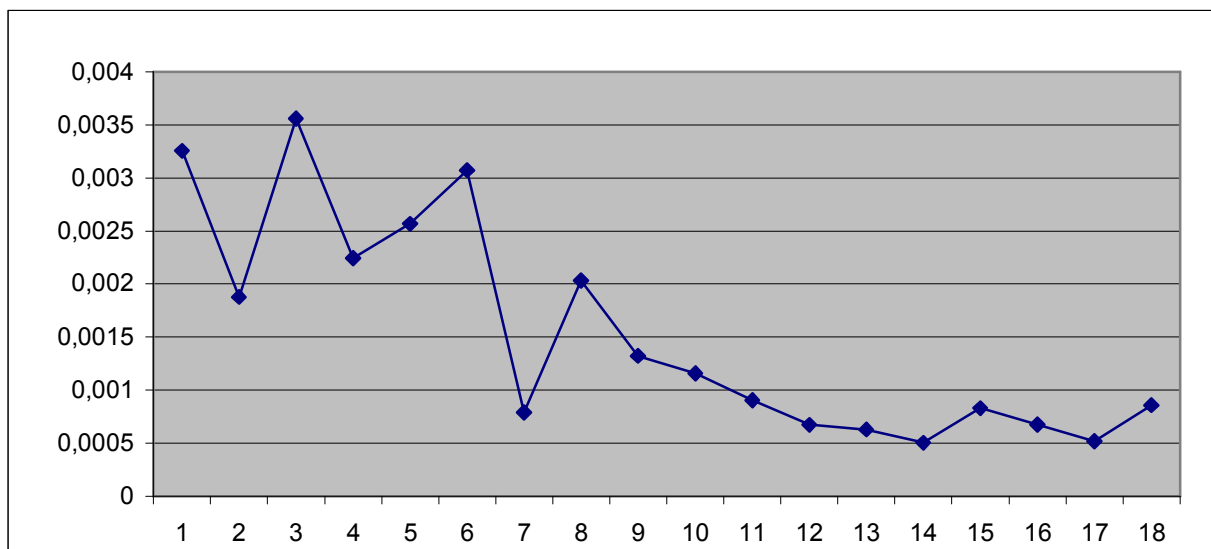

Fig. 4: Relative growth of the entropy function.

Figure 4 shows the dependency between the number of observation and the out coming kriging estimators of the spatial phenomenon. The number of observations is in thousands. Here we can clearly see that there exists the number of observations that gives us no more useful information for the out coming spatial modelling.
The out coming spatial model of the phenomenon gives us the next amount of information, which is shown on the next figure.
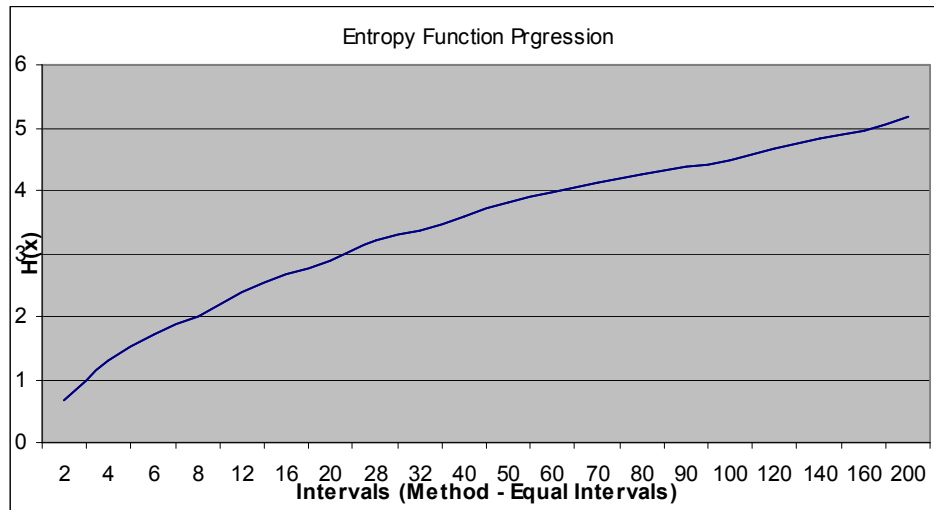
Fig. 5: Progression of the entropy function

From figure 5 one can see, that the out coming amount of the information, measured with the help of the entropy function, depends on the number of modelled points. Our study gives the answer that every spatial modelling has its own limits, where the obtained information is over the expectation. The ration between the relative growth of the entropy and number of modelled points is too small.

The last part of our study concern with the spatial visualization of natural phenomena. There have always exists discussions about the need of the precise and expert determination of the number of intervals which would be used for the out coming map or digital model of the given phenomenon. There exist series of methods for dividing of the investigated area, but we have proved, that each dividing of the investigated area and the information value contained in the resulting figure do not depend only on the number of intervals. Each method has different value of the out coming entropy (Fig. 6)
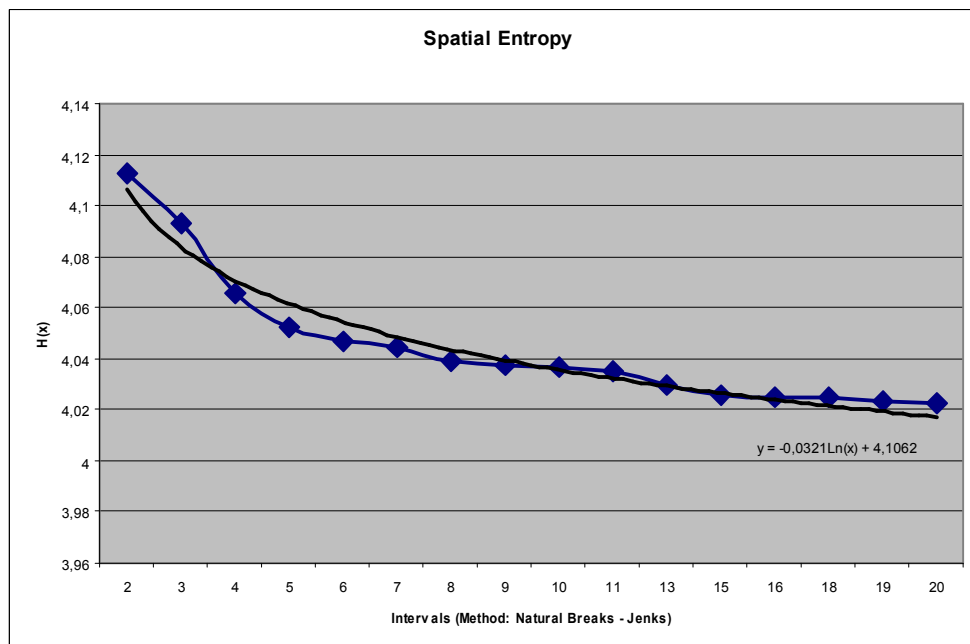

$$y = -0,0321 \ln(x) + 4,1062$$

Fig. 6: Progression of the spatial entropy function.

_____

We can see from fig. 6 that the function of the spatial entropy is the convex function and therefore there exists the point, where the next added interval to the visualization of the spatial phenomena gives us much less of information like the previous one. There always exists a factor of organization which is important to take into account, because different organization of the intervals (the same number of intervals) gives different amount of information. The expert for the investigated phenomenon should say the necessary amount of information that should be visualized and the GIS expert should create the appropriate visualization method. As an example, we have computed the optimal number of intervals for the climatological phenomenon (fig. 7). For more details see [12], where this problem is described in depth.
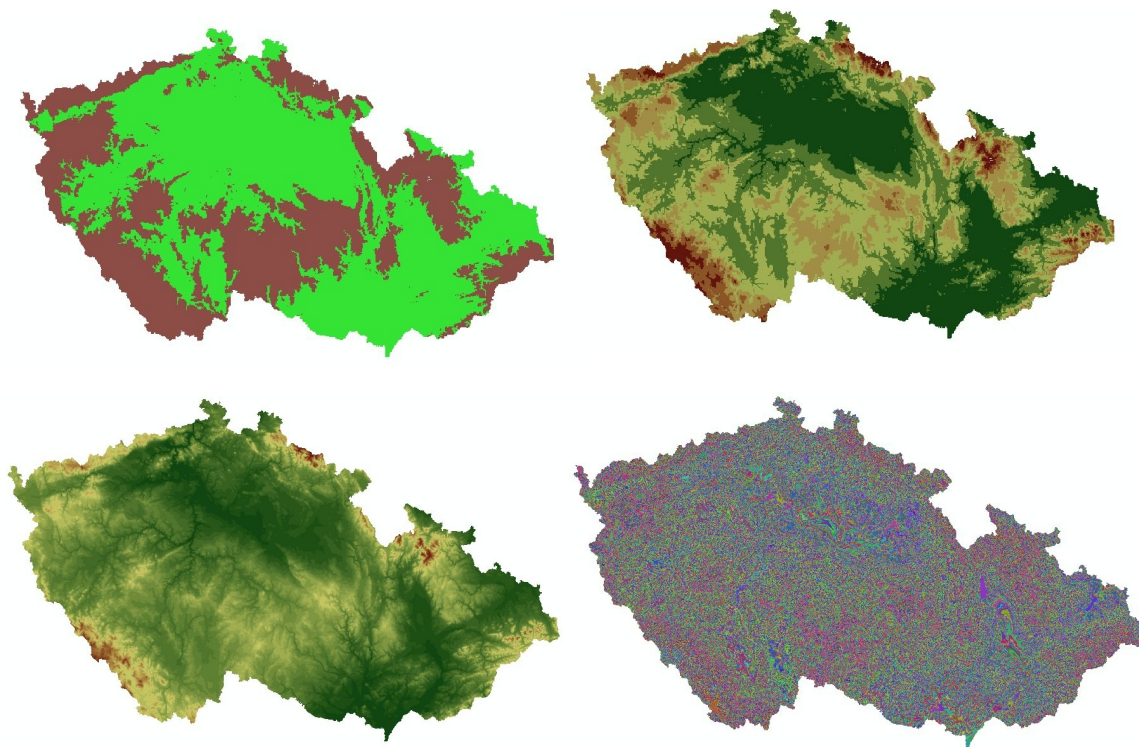

Fig. 7: GRID layers of the single one climatic phenomenon with different numbers of intervals.

In Fig. 7 are four GRID layers of the same climatic phenomenon, but with different numbers of intervals; increasing from upper left to lower right picture. Each one layers contains various amount of information, and it depends on many factors, which one is informational the richest.

5. Summary and Conclusion Remarks

The information theory, entropy and its spatial form is widely used in geographical research. Hand in hand with development in informatics, geoinformatics and GIS must also grow. As it is known, that geoinformatics deals with spatial information, it is easier to analyze this information using methods such as described in the information theory, namely the entropy function and spatial entropy function, respectively.

Nowadays, robust vizualization and analysis tools are provided in GIS software, so there is opened a new approach to analyze and visualize any phenomenon effectively and precisely. The entropy function brings new possibilities in understanding a nature around us. Further and deeper research in this problem is being hold on Department of Geoinformatics on Palacky University in Olomouc and new findings are awaited.

References

[1] BATTY, M. *Cost, Accessibility and Weighted Entropy*, Geographical Analysis, 15, 1983, 256-267.

[2] BATTY, M. *Entropy in Spatial Aggregation*, Geographical Analysis, 8, 1976, 1-21.

[3] BATTY, M. *Spatial Entropy*. Geographical Analysis, 6, 1974, 1-31.

[4] ČAPEK, J., et al. *Geografická kartografie.* Praha : SPN, 1992. 373 s.

[5] HOULDING, S. W.: *Practical Geostatistics: Modelling and Spatial Analysis*, New York: Springer. 2000

[6] KAŇOK, J. *Tematická kartografie.* Ostrava : Ostravská univerzita, 1999. 318 s.

[7] KUČEROVÁ. *1.1 Definice informace. : Data - informace - znalosti* [online]. 2006 [cit. 2008-04-20]. Dostupný z WWW: <http://info.sks.cz/users/ku/UIS/inform1.htm>.

[8] KOMENDA, Stanislav. *Základy statistiky ve zdravotnictví*. [s.l.] : [s.n.], 1991. 53 s.

[9] MURDYCH, Zdeněk. *Tematická kartografie*. [s.l.] : [s.n.], 1988. 248 s.

[10] PECHANEC, Vilém. *Nástroje podpory rozhodování v GIS*. [s.l.] : [s.n.], 2006. 104 s.

[11] SHANNON, C.E. A *Mathematical Theory of Communication*. Bell System Technical Journal. 1948, no. 27, s. 379-423, 623-656.

[12] TUČEK, P., PÁSZTO, V., VOŽENÍLEK, V.: *Entropie v kartografii*, In Sborník ČGS, in print

[13] VOŽENÍLEK, V. *Cartography for GIS – geovisualisation and map communication*. Univerzita Palackého v Olomouci, Olomouc, 2005, 140 s.

[14] VOŽENÍLEK, V. *Aplikovaná kartografie I – tematické mapy*. Univerzita Palackého v Olomouci, Olomouc, 2001, 187 s.