Comparing Partition of Clustering

Václavíková Štefánia

SLOVAK UNIVERSITY OF TECHNOLOGY
IN
BRATISLAVA
FACULTY OF CIVIL ENGINEERING

Comparing Partition of Clustering

- Cluster Analysis
- Similarity
- Entropy
- Application in hydrology

Cluster Analysis- applications

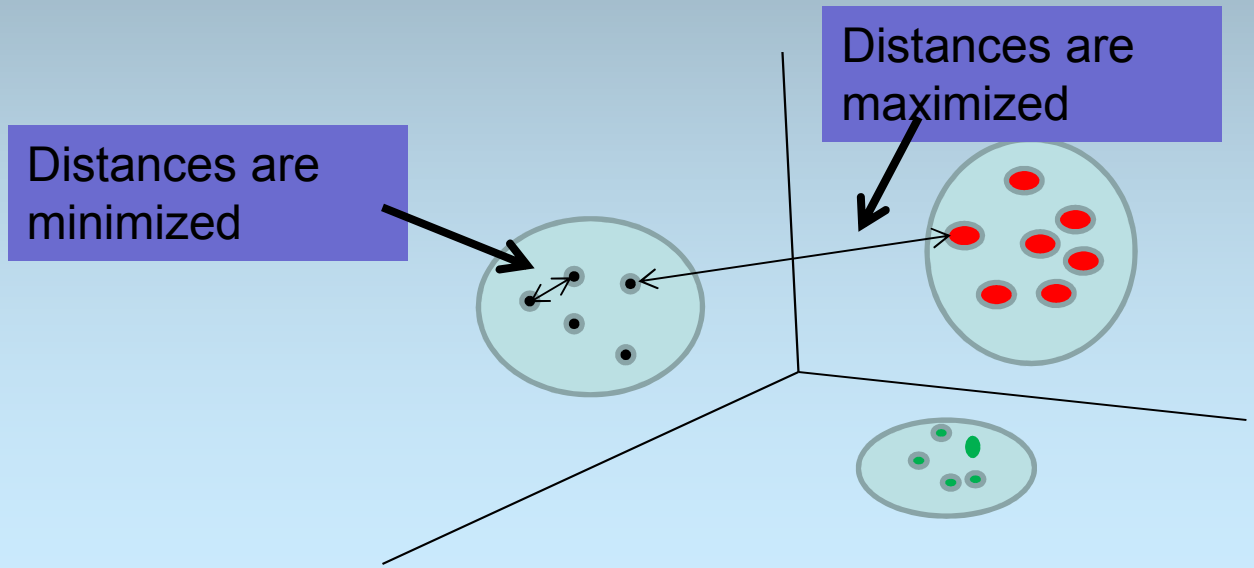
- biology
- medicine
- economy
- hydrology
- psychology…

numerical taxonomy, mathematic, taxonomy,categorisation, clasification...

R.C.Tyron (1939) was the first who worked up the methods of cluster analysis and it's usage in psychology

Cluster Analysis

- the objects inside the cluster have the greatest degree of similarity while the objects from different clusters show the highest rate of dissimilarity.

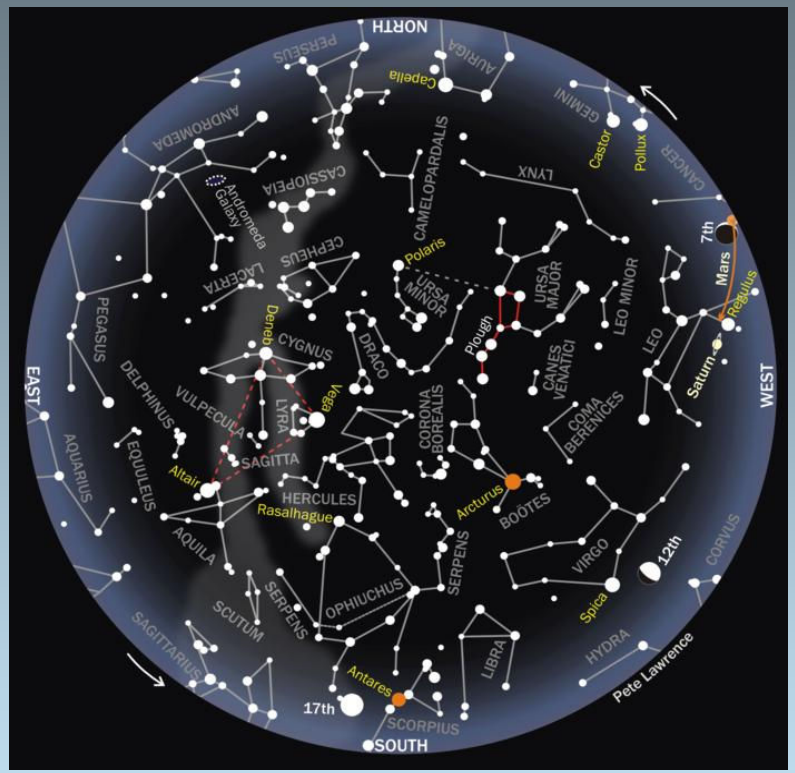


Cluster Analysis- purpose

- Understanding-

groups/clusters-forming in set of all objects allows clearer view and better orientation among the objects.

- find the similar genes
- group words in documents
- life conditions
- flow of the rivers
- similar behavior of animals …



- Summarization - reduction of large data sets

Similarity-dissimilarty

There exist many coefficients suitable for measurement of similarity. The measurement usage depends on what are we going to compare.

- Objects
- Attributes
- Category
- Clusters
- Results of clusterings

Similarity measure

Let E be a nonempty set . A function s

$$s: E^2 \to R$$

with properties

1. $s(X,Y) \geq 0$
2. $s(X,X) = 1$
3. $s(X,Y) = s(Y,X)$

will be called *similarity measure*

Dissimilarity measure

Let E be a nonempty set. Then the functions

$$\mathrm{s}, d: E^2 \to R$$

with properties

1.	$\mathrm{s}(X,Y) \geq 0$	$\mathrm{d}(X,Y) \geq 0$
2.	$\mathrm{s}(X,X) = 1$	$\mathrm{d}(X,X) = 0$
3.	$\mathrm{s}(X,Y) = \mathrm{s}(Y,X)$	$(X,Y) = \mathrm{d}(Y,X)$

will be called *similarity measure* *dissimilarity measure*

$$s = 1 - d \qquad \mathrm{d}(X,Y) \leq \mathrm{d}(X,Z) + d(Z,Y)$$

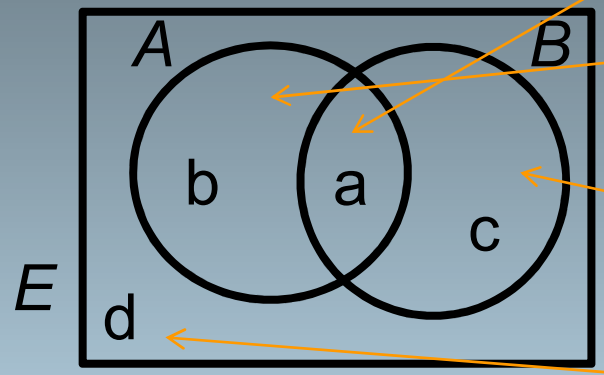
Distance measure

Measurement of distances between objects in the space

- Euklides $D_E(X_i, X_j) = \sqrt{\sum_1^m (x_{il} - x_{jl})^2}$
- Manhatan $D_B(X_i, X_j) = \sum_1^m (x_{il} - x_{jl})^2$
- Minkovski $D_M(X_i, X_j) = \sqrt[q]{\sum_1^m (x_{il} - x_{jl})^q}$
- ….

Symmetric difference

Let A, B are



a- number of objects which belong currently to cluster *A* and *B*

b- number of objects which belong only to cluster *A*

c-number of objects which belong to cluster *B*

d-number of objects which don't belong to any of *A* and *B*

$$d(A,B) = P(A\Delta B)$$

$$A\Delta B = (A \cup B) - (A \cap B) = (A \cap B^C) \cup (A^C \cap B)$$

$$P(A\Delta B) \in \langle 0,1 \rangle$$

Asociation coeficients

- Sokal-Michener $\frac{a+d}{a+b+c+d}$ $1-P(A\Delta B)$
- Dice coefficient $\frac{2a}{2a+b+c}$ $\frac{2(1-P(A\Delta B/A\cup B))}{2-P(A\Delta B/A\cup B)}$

In general we can define the function

$$f(k,q,t,x)=\frac{k(1-x)}{t+qx},$$

where $k,t,q\in R^{+}\quad x\in\langle 0,1\rangle$

$$x=P(A\Delta B)$$

$$x=P(A\Delta B/A\cup B)$$

Entropy

The concept entropy was originally derived in thermodynamics in1865 by **Rudolf Clausius**.

Claude E. Shanon developed the modern concept of ‘information’ and ‘logical’ *entropy* as a part of information theory in the late 1940s (1947). The second notion of information was *mutual information*. There is a measure of te information contained in one process about another process.

Entropy

Let $E \neq \emptyset$,

A is a partitions of E , $A = \{a_1, \ldots, a_n\}$, $a_i \cap a_j = \emptyset$,

$a_i \cup a_i = E$

then $H(A) = -\sum p_k \ln p_k$,

where p – is probability $p(a_i) = \frac{|a_i|}{|E|}$

Let us denote $0_E = \{E\}$

$1_E = \{\{x\}; x \in E\}$

$$0 = H(0_E) \leq H(A) \leq H(1_E) = ln|E|$$

Entropy

Let *A* and *B* are two partitions of *E*

$$A = \{a_1, \dots, a_n\}$$
$$B = \{b_1, \dots, b_k\}$$

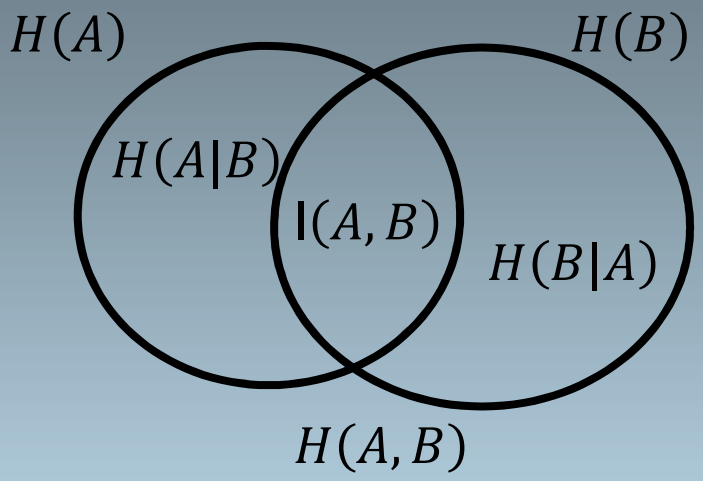
refinement $A \vee B = \{a_i \cap b_j \neq \emptyset, i \neq j\}$

Join entropy

$$H(A,B) = H(A \vee B) = -\sum_{ij} p(a_i \cap b_j) \ln p(a_i \cap b_j)$$

Basic properties

A, B are partitions of E



$$H(A,B) \leq H(A) + H(B)$$

Mutual information

$$I(A,B) = H(A) + H(B) - H(A,B)$$

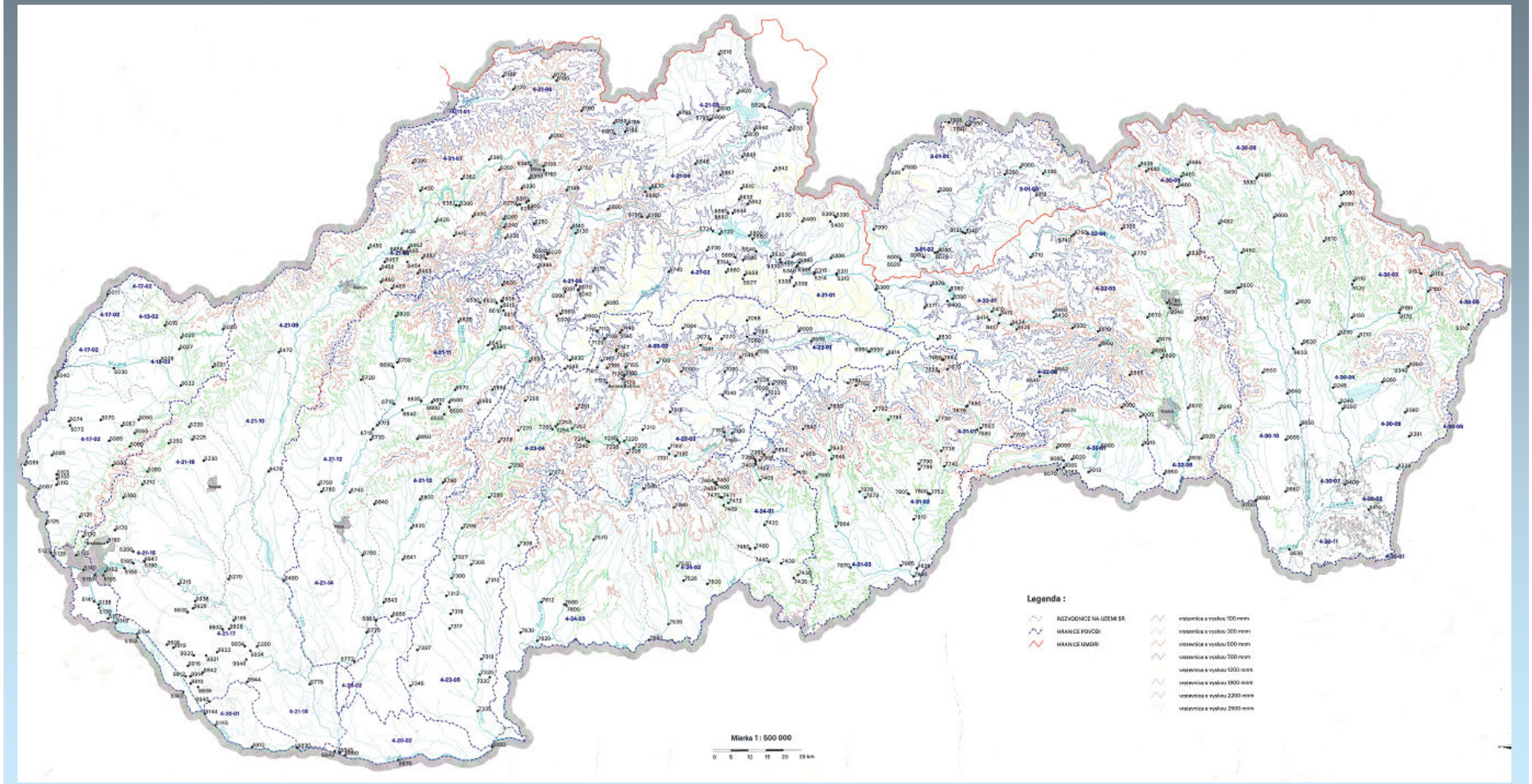
$$I(A,A) = H(A); \quad I(0_E, A) = 0$$

$$d(A,B) = H(A,B) - I(A,B)$$

is a metric

$$d(A,B) \leq d(A,C) + d(C,B)$$

Catchments


Legenda :
ROZVODNICE NA UZEMÍ SR
HRANICE POVODÍ
HRANICE UMÓRI
Mierka 1 : 500 000

Applications

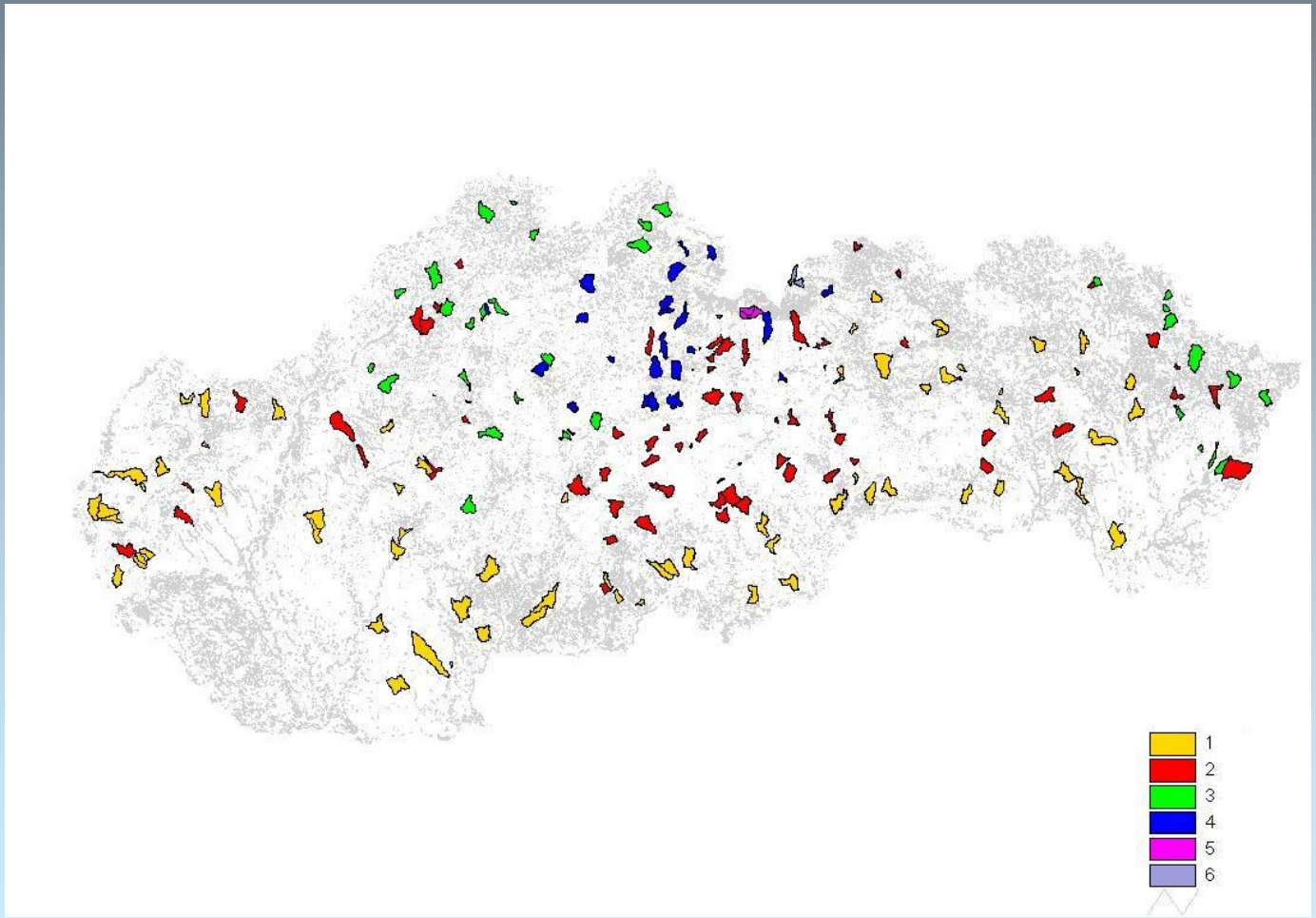
		B						
A		b1	b2	b3	b4	b5	b6	
	a1	18	21	24	0	0	0	63
	a2	9	25	7	11	0	0	52
	a3	6	12	0	12	4	2	36
	a4	21	8	3	0	0	0	32
	a5	13	2	1	0	0	0	16
	a6	5	1	0	4	0	0	10
		72	69	35	27	4	2	209

Applications

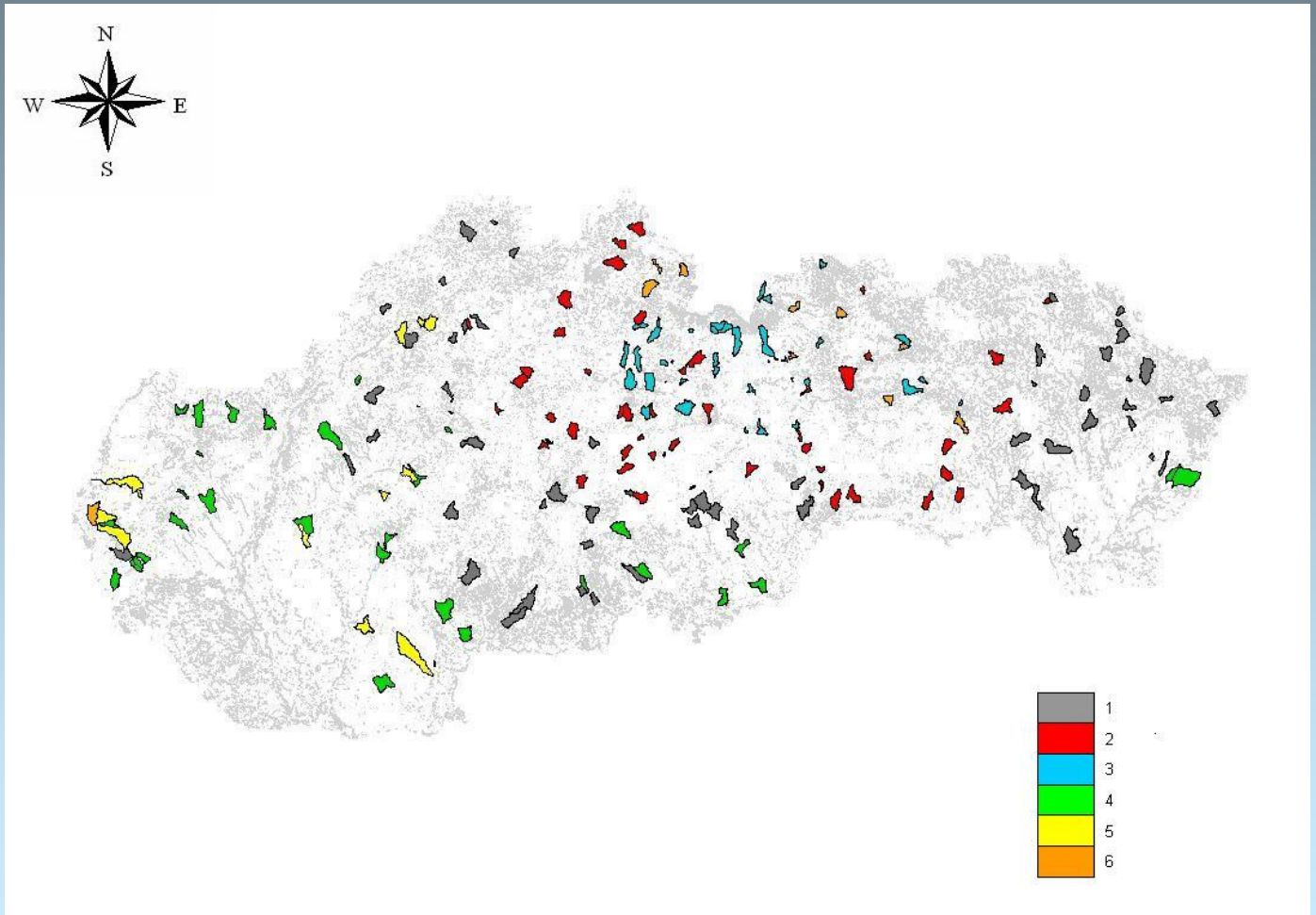
	Value	V_{norm}
H(A)	1.64004	0.59658
H(B)	1.41683	0.51538
H(A,B)	2.74908	1
I(A,B)	0.30779	0.11196
d(A,B)	2.4413	0.88804

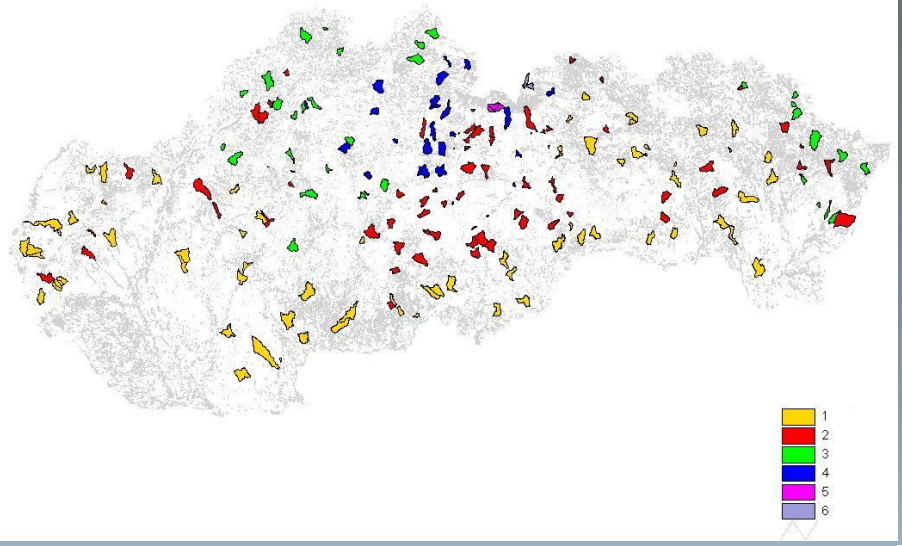
$$V_{NORM} = Value/H(A,B)$$

Partition A



Partition B





Thanks for your attention