Kartografické listy, 2009, 17

Vít PÁSZTO, Pavel TUČEK

INFORMAČNÍ ZISK A ENTROPIE V KARTOGRAFICKÉ TVORBĚ

Pászto, V., Tuček, P.: Information Gain and Entropy in Cartographic Production. Kartografické listy 2009, 17, 6 figs., 2 tabs., 9 refs.

Abstract: The aim of this paper is to briefly present possibilities of using the information gain and information entropy in evaluation the similarity or the dependence between two maps and evaluate its basic patterns. Information gain (or loss) and information entropy are evaluated while generalization or categorization of geographical phenomena into groups is examined. Relatively old-fashioned concept of entropy is reused in the way of new perspective and powerful capabilities of Geographic Information System (GIS), its analytical tools and visualization engine. We give a new meaning of the entropy and allow all GIS users to study and display geographic phenomenon and processes in a new relationship with the help of robust analysis and visualization tools provided in GIS. The article contains the overview that should serve as a motivation to the analysis of the entropy in the cartography, spatial modeling and visualization. The whole theory is accompanied by of examples computed over the climatic raster datasets of Czech Republic.

Keywords: entropy, information gain, geoinformatics, cartography, information theory, geocomputation, classification

Úvod

Příspěvek představuje možnosti použití poznatků z obecné teorie informace, konkrétně uplatnění informačního zisku a entropie, v kartografii i geoinformatice. Důraz je mj. kladen na stanovení optimálního počtu kategorií zkoumaného jevu a také na zhodnocení interpolačních metod při tvorbě mapy. V neposlední řadě jsou nastíněny příklady použití uvedených metod při vizualizaci geografických objektů a jevů. U většiny kartografických děl se intuitivně předpokládá, že provedená interpolace, popř. jiná statistická metoda poskytuje, korektní výsledky, které se následně vizualizují. Proti tomuto tvrzení nelze nic namítat. Nicméně každý geografický jev, jenž je snaha kartograficky vyjádřit, vyžaduje jinou vizualizační metodu a odlišný přístup, aby dosažené výsledky byly korektně vizualizovány. Cílem tohoto článku je představit relativně nový způsob, jak přistupovat k tvorbě kartografických děl pomocí hodnocení informační entropie daného jevu a informačního zisku mapy. Právě díky moderním geotechnologiím a možnostem geografického informačního systému (GIS) lze na tento aspekt tvorby kartografických výstupů pohlížet v novém světle. Prozatímní metody založené na sofistikovaných statistických a dalších postupech ovšem neberou příliš v úvahu základní myšlenku kartografické tvorby, a tou je fakt, že uživatel mapového díla by měl být schopen během krátkého času absorbovat velké množství informace. Mnohdy je pozorováno to, že provedená vizualizace nedosahuje takové informační hodnoty, jakou by si kartograf představoval (s ohledem na uživatele díla). Tyto a další problémy mohou být řešeny pomocí metody informačního zisku a entropie, které tak znamenají inovativní přístup při hodnocení geografických jevů a jejich znázornění v mapě.

Teorie informace, entropie

Jak již bylo zmíněno v dalších příspěvcích autorů (Pászto et. al 2009, Tuček et. al 2009), tak je pojem informační zisk a informační entropie relativně nový a byl formalizován krátce po Druhé světové válce. V této době nastává rozmach výpočetní techniky a způsobů přenášení zpráv. Začíná se rodit informační společnost a také vědní obory jako je kybernetika. Součástí kybernetiky je i teorie informace, kterou zformuloval Claude E. Shannon.

Mgr. Vít PÁSZTO, Mgr. Pavel TUČEK, Přírodovědecká fakulta, Univerzita Palackého v Olomouci, Tř. Svobody 26, Olomouc, e-mail: vit_p@volny.cz, pavel.tucek@upol.cz

Informace, jakožto abstraktní nehmotný objekt, se stala cílem studia tehdejších matematiků. A právě ve své práci (Shannon 1948) mimo jiné definoval, jak tuto informaci měřit. Z práce je zřejmé, jak uvádí například i Pezlar (1998), že formalizace informace vychází ze pravděpodobnostně-statistického vyjádření informace.

Velice známá věta (Shannon 1948, Pezlar 1998) o informaci tvrdí následující: „*Informace je míra množství neurčitosti nebo nejistoty o nějakém náhodném ději, odstraněná realizací toho děje*".

Je potřeba tedy znát a umět hlavně kvantifikovat ono množství neurčitosti či nejistoty. K tomu definoval Shannon (1948) veličinu zvanou informační entropie. Vztah, podle kterého je entropie vypočítávána, je určen následovně:

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

kde *S* je systém s konečným počtem možných událostí s_i, $P(s_i)$ reprezentuje pravděpodobnost výskytu události s_i. Logaritmus se zpravidla používá se základem 2 (potom jednotka bit), nicméně lze použít dekadický i přirozený logaritmus (jednotka dit, resp. nit).

Jak uvádí Shannon (1948) i Komenda (1991), maximální entropie nastává při vyrovnaných šancích všech výskytů souboru, tedy nastává nejvíce nejistá (neurčitá) situace – $P(s_i)$ je pro všechna *i* stejná. Je tak potřeba získat maximum informace pro odstranění této neurčitosti.

Naopak minimální entropie nastane, pokud pravděpodobnost jedné a pouze jedné události $P(s_i)$ je rovna 1 a ostatních rovna 0. Tzn. že nastane jedna událost a nejistota je zcela odstraněna.

Odvozeným pojmem z výše uvedeného je tzv. informační zisk. Ten je získán ze vztahu:

$$I(S) = H(S)_{max} - H(S), \tag{2}$$

kde I(S) je informační zisk systému, $H(S)_{max}$ je maximální entropie (obecně rovna $\log_2 n$) a $H(S)$ je aktuální vypočtená entropie. Analogicky, čím je vyšší aktuální entropie, tím je nižší informační zisk, a naopak.

Je však důležité zmínit se, že výše zkoumaným systémem může být i mapa, datová vrstva v GIS, výsledek měření geografického jevu (socioekonomického i fyzicko-geografického), ale i vstupní parametry nejrůznějších prostorových analýz, stejně tak jako prostorová analýza samotná.

V následujících kapitolách bude stručně nastíněna možnost použití informačního zisku při hodnocení podobnosti a závislosti map (a jejich zákonitosti), bude zhodnocen informační zisk, resp. ztráta informace, při generalizaci či kategorizaci geografických jevů do skupin.

Entropie při hodnocení podobnosti map

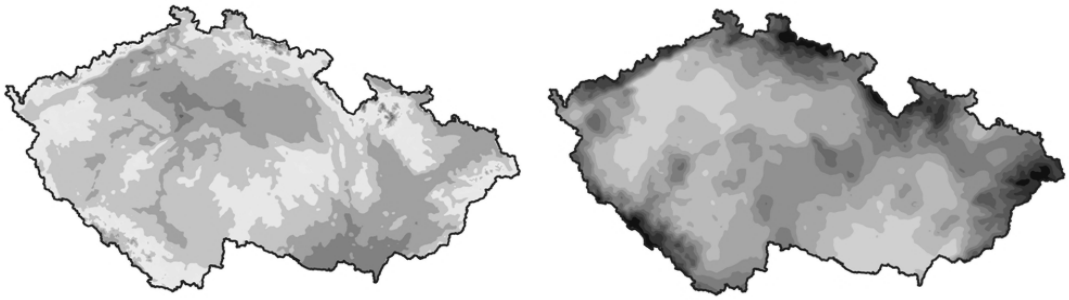
Inovovaný přístup hodnocení podobnosti, či vzájemné závislosti a podmíněnosti map, resp. jevu, jež mapa znázorňuje, vychází z dříve provedené studie (Tuček et al. 2009) a ukazuje nové aspekty použití entropie při tomto hodnocení.

V článku (Tuček et al. 2009) jsou hodnoceny dvě rastrové vrstvy z Atlasu podnebí Česka (Tolasz et al. 2007) a postup byl i v tomto případě obdobný. Nyní byly brány jako ukázková data dva rastry, jeden představoval průměrnou roční teplotu vzduchu a druhý průměrný roční úhrn srážek. Na obr. 1 jsou obě vrstvy znázorněny bez jakýchkoliv kartografických úprav.

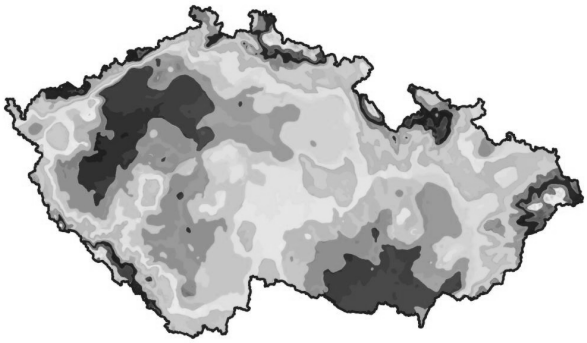
Obě rastrové vrstvy byly reklasifikovány do deseti kategorií kvůli stejnému počtu intervalů pro další analýzu. To bylo nutné, neboť rastr průměrné roční teploty vzduchu měl již danou desetidílnou stupnici, zatímco rastr průměrného ročního úhrnu srážek neměl. Reklasifikace byla také nutná kvůli následné kombinaci rastrů mapovou algebrou tak, aby vznikly nové unikátní kategorie (obr. 2). Předtím ale byla oběma rastrům vypočtena informační entropie, které byly srovnány s entropií kombinovaného rastru (tab. 1).

Tab. 1 Statistika entropie zkoumaných rastrů

	Skutečná entropie (nit)	Maximální entropie (nit)	Relativní entropie (%)	Informační zisk (nit)
Rastr teplot	1,480	2,303	64,266	0,823
Rastr srážek	2,025	2,303	87,938	0,278
Kombinace rastrů	3,185	4,078	78,101	0,893



Obr. 1 Vstupní rastrové vrstvy (vlevo průměrná roční teplota vzduchu, vpravo průměrný roční úhrn srážek)



Obr. 2 Kombinovaný rastr průměrných ročních teplot vzduchu a průměrného ročního úhrnu srážek

V tab. 1 jsou vypočteny charakteristiky entropie – skutečná (aktuální) entropie, maximální entropie, relativní entropie a informační zisk. Entropie v tomto případě ukazuje uspořádanost, resp. rovnoměrnost rozdělení hodnot do kategorií. Čím vyšší entropie (tím pádem i relativní entropie), tím jsou hodnoty jevu rovnoměrněji rozděleny do kategorií či intervalů. Lze podle toho usoudit, že daná mapa je vyváženější než jiná hodnocená (např. z hlediska použitých barev) a poskytuje uživateli více informací. Hodnocení entropie map je však pouze jedním z možných kritérií pro hodnocení náplně mapy. Čím je naopak entropie nižší, tak četnost určitého rozsahu hodnot se soustřeďuje více do jedné (nebo několika málo) kategorií či intervalů, a tak mapa neposkytuje uživateli dostatečnou informaci o rozložení hodnot určitého jevu v prostoru (pokud to však nebyl účel mapy). V tomto případě by pak bylo vhodnější danou kategorii či interval rozdělit na více dílčích tak, aby čtenář získal detailnější představu o daném jevu. Opět, pokud však účel mapového díla není odlišný.

Konečně podobnost či závislost jednotlivých map lze zhodnotit tzv. kombinační tabulkou (tab. 2), která ukazuje nově vzniklé kategorie kombinovaného rastru a lépe zobrazuje odvozenou závislost dvou jevů na sobě. Z tab. 2 je zřejmé, jaké jsou kombinace nově vytvořených unikátních kategorií.

Tab 2. Kombinační tabulka sloučeného rastru

		průměrná roční teplota vzduchu (℃)										relativní entropie daném intervalu (%)
		1	2	3	4	5	6	7	8	9	10	
roční úhrn srážek (mm)	415-529						X	X	X	X	X	70,962
	530-586				X	X	X	X	X	X		57,826
	587-640					X	X	X	X	X		65,879
	641-699				X	X	X	X	X	X		63,579
	700-766				X	X	X	X	X			76,143
	767-844			X	X	X	X	X	X			73,787
	845-941			X	X	X	X	X	X			79,880
	942-1061		X	X	X	X	X	X	X			78,119
	1062-1217	X	X	X	X	X	X	X				78,988
	1218-1498	X	X	X	X	X	X					75,193
relativní entropie daném intervalu (%)		95,908	82,932	69,804	74,301	89,798	81,488	82,065	76,985	46,120	---	

V tab. 2 je tedy znázorněna závislost výskytu určité hodnoty jevu na hodnotě jevu druhého. Lze tak jednoduše tvrdit např., že nejnižší průměrný úhrn srážek je vázán na průměrné roční teploty vzduchu nad 6 °C, kde kategorie s největší četností výskytu dané kombinace vyjadřuje oranžové pole – v případě intervalu nejméně intenzivních srážek je tak při 8 °C.

Čím nižší relativní entropie pro daný řádek či sloupec (tedy pro daný interval hodnot sledovaného jevu), tím v tomto intervalu hodnot převažuje závislost na intervalu hodnot (nebo více, ale ne mnoha intervalech) jevu druhého. Tedy například interval průměrného ročního úhrnu srážek od 530 do 586 mm vytvořil největší četnost výskytu v intervalu průměrných ročních teplot vzduchu od 7 do 8 °C včetně. A to více než 85 % hodnot srážek spadá do uvedeného intervalu teplot. V tomto případě je relativní entropie nízká a lze tedy tvrdit, že výskyt srážek v této kategorie je silně závislý na uvedených teplotách. Anebo naopak – při teplotách 7 až 8 °C včetně se vyskytuje zpravidla 530 až 586 mm srážek.

Naopak, čím vyšší relativní entropie, tím jsou jednotlivé hodnoty intervalu jednoho jevu rovnoměrně rozmístěny do intervalů hodnot druhého jevu. Například průměrná roční teplota vzduchu 5 °C s relativní entropií 89 % je víceméně rovnoměrně rozmístěna mezi všechny intervaly průměrného ročního úhrnu srážek od 530 mm výše. To znamená, že pokud bude mít určité území průměrnou roční teplotu 5 °C, tak se na tomto území mohou vyskytovat srážky v daných intervalech rovnoměrně.

Bez použití entropie pak lze i jednoduše hodnotit (jak již bylo zmíněno) hranice výskytu hodnot jednoho jevu v závislosti na druhém. Například lze tvrdit, že pokud se na daném území vyskytuje průměrná roční teplota vzduchu 1 °C, tak na tomto území spadne v průměru více než 1062 mm srážek za rok. Obdobný postup zmiňuje i Murdych (1988), avšak za použití sdružené entropie a koeficientu korelace.

Informační zisk při restrikci stupnice geografických jevů

Tento odstavec je zaměřen na kvantifikaci ztráty informace způsobenou restrikcí stupnice geografických jevů. Zjednodušeně řečeno, tento odstavec pojednává o možnosti korektního škálování a následné vizualizaci.

Každé měření musí vyústit v interpretaci měřeného, a tudíž je potřebné sofistikovaným přístupem vizualizovat získaná data. Pouze ten přístup, který zachová podstatnou část informace plynoucí z měřených dat, je přístupem vhodným pro vizualizaci. Údaje z prvotní škály ovšem standardní interpretaci neumožňují, což je důvodem k tomu, aby se základní škály transformovaly, popř. „restringovaly“. Teorie informace sice vznikla v souvislosti s technickými problémy přesunu

zpráv, ale myšlenky, které publikoval C. E. Shannon (Shannon 1948) mají obecnější charakter a lze jich tedy využít i v případě geografických informačních systémů.

Matematické prostředky pro odhad množství informace nemohou vždy zahrnout všechny aspekty kladené na výslednou podobu mapového výstupu. Ovšem restringovaná, popř. jiným způsobem modifikovaná základní škála, může při správném použití zvýraznit specifické cíle měření, jehož výslednou podobou má být vizualizace v podobě mapového výstupu.

Z výše uvedených faktů vyplývá, že korektní použití restrikce, popř. jiné modifikace základní škály vedoucí ke ztrátě informace, nemusí být vždy jevem negativním. Uvažujme klasický případ, kdy má být vizualizovaná náhodná veličina *X* (např. klimatický jev), která nabývá hodnot z množiny:

$$Z_n = \{1, 2, \ldots, n\} \tag{3}$$

Tato množina obsahuje originální hodnoty vizualizovaného jevu. Jsou to tedy takové hodnoty, pro které platí, že $\forall x_i \in Z_n$ může být nabyta opakovaně. V případě, kdy použijeme všechny originální hodnoty pro konstrukci škály vhodné pro následnou vizualizaci, lze stanovit, že informační zisk takového vizualizace je roven vztahu:

$$\Delta H_{Z_n - R_k} = H_{Z_n} - H_{R_k} \tag{4}$$

Ve vztahu (4) odpovídá $ln(n+1)$ maximální entropii a H_{Zn} odpovídá entropii dané základním škálováním Z_n (tento vztah upraven dle (2)).

Proveďme nyní určitou restrikci základní škály. Ta je určena rozkladem R_k množiny Z_n na k disjunktních podmnožin S_i , pro které platí:

$$k < n, R_k = \{S_i\}_{i=1}^{k}, \qquad S_i \cap S_j = 0 \quad \forall i \neq j, \quad \bigcup_{i=1}^{k} S_i = Z_n \tag{5}$$

Ze znalosti pravděpodobností $P(X \in S_i)$, i = 1,2,...,n můžeme určit entropii dle vztahu (1) a informační zisk z takto upravené škály lze stanovit dle vztahu:

$$I_{R_k} = ln(k) - H_{R_k} \tag{6}$$

Změnu informace vyplívající z takto provedené restrikce lze odhadnout pomocí:

$$\Delta I_{Z_n - R_k} = I_{Z_n} - I_{R_k} \tag{7}$$

Rovněž lze stanovit rozdíl v hodnotě míry neurčitosti způsobenou restrikcí dle vztahu:

$$\Delta H_{Z_n - R_k} = H_{Z_n} - H_{R_k} \tag{8}$$

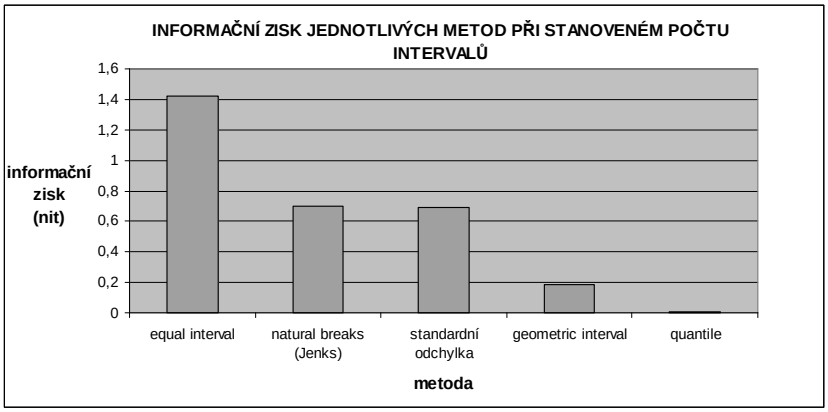
Nyní se naskýtají dvě možnosti, jak tohoto principu zachování množství informace využít pro korektní postup volby vhodné metody tvorby stupnic pro mapovou tvorbu, popř. pro jinou vizualizační metodu. Využití lze ale chápat v dvojím smyslu, a to:

1. Prostá generalizace informačním ziskem

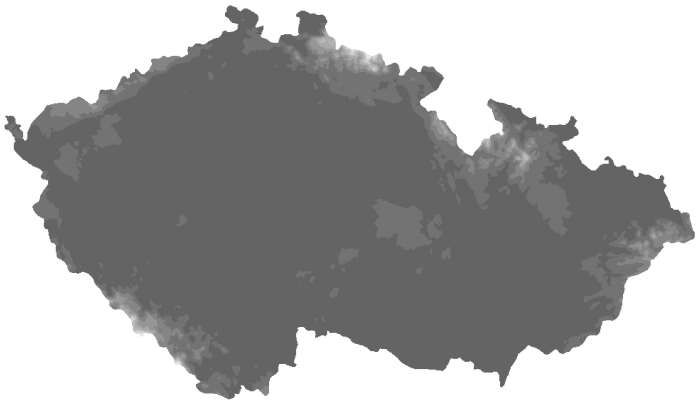
2. Rekurzivní generalizace informačním ziskem

1. Prostou generalizací informačním ziskem se rozumí použití výše uvedené metodiky určení informační ztráty v případě, kdy pro předem stanovený počet intervalů je hledána vhodná metoda dělení základní škály. Tento postup, založený na porovnání hodnot informačního zisku (popř. ztráty) plynoucího z provedené restrikce, lze do jisté míry chápat jako hodnotící algoritmus grafických výstupů využívajících různou metodiku dělení základní škály.

Na obr. 3 lze vidět, že pro předem stanovený počet intervalů, využitých pro vizualizaci jistého jevu, bylo odvozeno za pomocí navrženého postupu, že nejvhodnější metodou, vzhledem k zachování nejvyšší míry informace, je metoda *equal interval* (obr. 4).



Obr. 3 Graf hodnot informačního zisku pro stejný počet intervalů n=8 a pro různé metody



Obr. 4 Vizualizace zkoumaného jevu do předem stanoveného počtu intervalů metodou equal interval

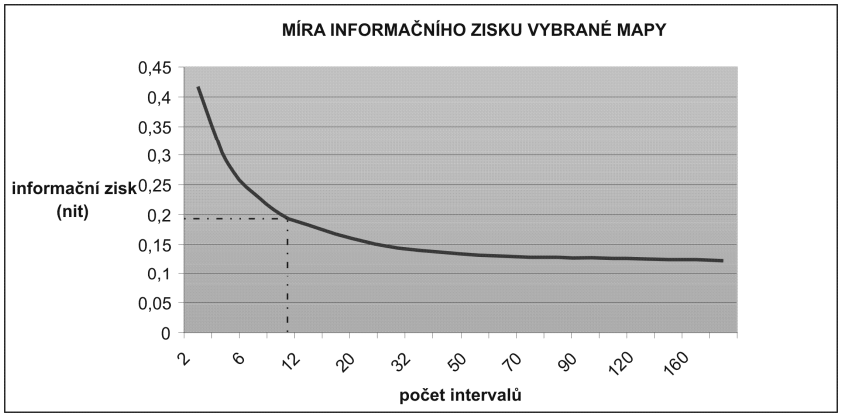
Toto je však pouze jedním z více kritérií pro správnou metodu dělení intervalů. Do uvedeného procesu zasahuje kartograf, specialista v oboru, ale i statistik. Někdy je totiž důležité zachovat určité extrémní hodnoty jevu (i když s malou četností výskytu ve výsledné mapě), ale přitom rozlišovat od chyb v měření či náhodných chyb, tedy tzv. outlierů (Voženílek et al. 2008).

2. Rekurzivní generalizaci informačním ziskem použijeme v případě, kdy chceme pro jednu konkrétní metodu dělení základní škály stanovit optimální počet intervalů, které mají být použity pro následnou vizualizaci. Tímto přístupem lze stanovit mez, za kterou již provedená generalizace nezpůsobuje vyšší míru ztráty informace, popř. lze tímto způsobem stanovit mez, ze kterou již nemá smysl více generalizovat, vzhledem k množství informace poskytované takovýmto výstupem.

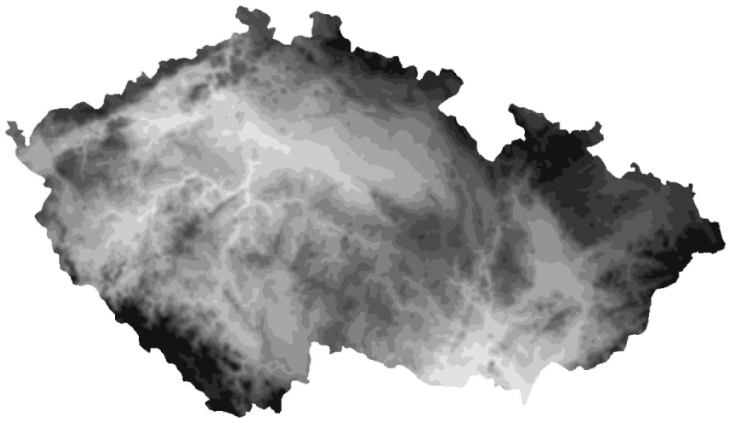
Na obr. 5 byla určena mez (hodnota, kdy relativní úbytek informačního zisku každého dalšího přidaného intervalu je nepatrný), která odpovídá počtu 11 intervalů, které byly následně využity pro vizualizaci zkoumaného jevu – obr. 6.

Z předložených výsledků lze tedy odvodit fakt, že pro každý grafický výstup existuje mez, kterou lze chápat jako hranici pro počet intervalů modifikované škály vhodných pro korektní vizualizaci. Každý další interval již ubírá v relativním poměru méně informace než kterýkoliv předchozí. Z toho rovněž plyne, že každý přidaný interval nad tuto hranici poskytuje relativně vyšší míru informace do výsledného grafického výstupu.

Při spojení těchto dosažených výsledků s ostatními pravidly tvorby grafických výstupů a pravidly kartografickými lze dosáhnout optimální podoby grafického výstupu (tedy mapy) obsahujícího optimální množství informace, které ani nebude vzhledem k danému formátu nedostatečné, ani přebytečné (vzhledem k danému účelu může být někdy moc podrobná vizualizace nadbytečná).



Obr. 5 Graf míry informačního zisku dosaženého postupnou generalizací mapového výstupu



Obr. 6 Vizualizace zkoumaného jevu při optimálním počtu 11 intervalů

Na závěr této problematiky lze podotknout, že kombinací dvou uvedených přístupů generalizace informačním ziskem lze docílit výběru optimální metody a následně optimálního počtu intervalů pro každý vizualizovaný jev.

Závěr

Příspěvek přináší inovativní postup hodnocení podobnosti či závislosti geografických jevů, a tedy i mapových výstupů, jež jev znázorňují. Teorie informace nabízí obecné znalosti, pojmy a vztahy k tomu, jak kvantifikovat množství informace ve zprávě. Takovou zprávou může být přirozeně i mapa sdělující efektivně a přesně velké množství prostorových informací (Voženílek 2005).

Pomocí nástrojů GIS je názorně ukázáno, jaký přínos může mít provedení statistiky informačního zisku a entropie při mapové tvorbě. Díky analytickým funkcím GIS lze zkoumat i skryté závislosti geografických jevů a jejich zákonitosti výskytu v prostoru, tedy i v rovině mapy.

Za využití matematického aparátu byl testován nový přístup při restrikci a kategorizaci původní naměřené stupnice hodnot právě použitím poznatků o informačním zisku a entropii. Byl navržen postup stanovení jak optimální metody dělení intervalů – tzv. prostá generalizace informačním ziskem, tak i stanovení optimálního počtu intervalů při znázornění geografického jevu vybranou metodou – tzv. rekurzivní generalizace informačním ziskem.

Literatura

KOMENDA, S. (1991). *Základy statistiky ve zdravotnictví*. Olomouc (Univerzita Palackého v Olomouci).

MURDYCH, Z. (1988). *Tematická kartografie*. Praha (Přírodovědecká fakulta Karlova Univerzita).

PÁSZTO, V., TUČEK, P., VOŽENÍLEK, V. (2009). On spatial entropy in geographical data. *GIS Ostrava 2009*, Ostrava. Dostupné na: http://gis.vsb.cz/GIS_Ostrava/GIS_Ova _2009/sbornik/Lists/Papers/017.pdf (30.5.2009)

PEZLAR, Z. (1998). *Základy teorie informace*. Brno (Konvoj).

SHANNON, C.E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, No. 27, s. 379-423, 623-656.

TOLASZ, R. ed. (2007). *Atlas podnebí Česka*. Praha a Olomouc (Český hydrometeorologický ústav a Univerzita Palackého v Olomouci).

TUČEK, P., PÁSZTO, V., VOŽENÍLEK, V. (2009). Entropie v kartografii. In *Sborník České geografické společnosti*, (v tisku).

VOŽENÍLEK, V. (2005). *Cartography for GIS – geovisualisation and map communication*. Olomouc (Univerzita Palackého v Olomouci).

VOŽENÍLEK, V., KAŇOK, J., TUČEK, P. (2008). Detekce, prokazatelnost a vizualizace extrémů demografických dat ve statistických souborech. *Kartografické listy*, 16, s. 113-125.

S u m m a r y

Information Gain and Entropy in Cartographic Production

The first part of the article refers about basic applications of using the information gain and entropy in cartography, supported by geoinformatics and GIS. Information theory provides a universal procedure for measuring information in general terms. Robust analysis and visualization tools provided in GIS give a new meaning of the entropy and allow us to study and display geographic phenomenon and processes in a new relationship as well as its expression via maps, atlases etc.

One can find here also important information about finding the optimal number of intervals for the visualization of the spatial phenomenon with the help of the entropy function computed for the resulting map. It is also possible to evaluate the orderliness of geographical phenomena or evaluate the correspondence between two phenomena shown in the map, and many more. All with the help of the entropy function.

Finally, two methods were introduced. They deal with the evaluating of the geographical phenomenon. First is focused on the selection of the most appropriate classification method – so-called Simple information gain generalization. And the second one is focused on setting the optimal number of intervals of described geographical phenomenon – so-called Recursive information gain generalization.

Fig. 1 Input raster layers (annual mean air temperature on left, annual mean total precipitation on right)

Fig. 2 Combined raster layer from raster in Fig. 1

Fig. 3 Information gain chart for n=8 intervals using different methods

Fig. 4 Visualization of n=8 interval map using equal interval method

Fig. 5 Information gain chart for in-sequence categorization of map

Fig. 6 Visualisation of given phenomenon divided into 11 optimal class intervals

Tab. 1 Entropy statistics of evaluated raster layers

Tab. 2 Combinational cross table of the combined raster layer

Lektoroval:

Doc. Ing. Miroslav MIKŠOVSKÝ, CSc.,
Praha, Česká republika