Kartografické listy, 2008, 16.

Vít VOŽENÍLEK, Jaromír KAŇOK, Pavel TUČEK

DETEKCE, PROKAZATELNOST A VIZUALIZACE EXTRÉMŮ DAT VE STATISTICKÝCH SOUBORECH

Voženílek, V., Kaňok, J., Tuček, P.: Extreme detection, provableness and visualization of statistical data files. Kartografické listy 2008, 16, 9 figs., 17 refs.

Abstract: The aim of the article is to make one known, that the extreme values of the observed geographical variable could be visualized in many ways. Authors introduce the optimal procedures for visualization of the extreme values in demographic datasets. The methodology is based on the statistical preprocessing of the data. The detection of the extreme values must be performed as first. The visualization of these values could be drawn after the significant prove. There is presented boundary which could be computed from the dataset and could serve as a signal point for the extreme values in the datasets. The extreme values are divided into two groups according to the point of occurrence. The first group is reserved for the extreme values in the frequency area. The second group describes the extreme values in the data area. The conclusion is dedicated to the measuring of the entropy in order to obtain the significant number of intervals for the creating of the maps with the respect to the occurrence of the extreme values.
Keywords: map, errors, outliers, extremes, visualization

Úvod

Matematizace se v geografii začala výrazněji projevovat v 60. letech minulého století. Tento proces představoval sblížení geografie a matematiky (Voženílek 2001), což vyústilo ve vytváření metodických postupů s přesnější argumentací a vyšší spolehlivostí. Začal se uplatňovat exaktnější a zobecněný výklad řady geografických teorií s důrazem na objasnění obecných vlastností prostorových struktur u vzhledově a předmětově různých jevů v krajině. Hlavním důsledkem však byla metodologická spojení s jinými vědními disciplínami pomocí formulací postupů a závěrů v obecně vědeckém jazyce matematiky. Statistické metody (též kvantitativní metody výzkumu) jsou dnes již nezbytnou součástí geografického výzkumu.

Příspěvek se zabývá rozborem problematiky extrémních hodnot, zejména jejich detekce, prokazatelnost a nalezení vhodné metody kartografické vizualizace extrémů demografických dat ve statistických souborech. Téma vizualizace extrémů v mapách je jedním ze stěžejních otázek při kvantitativním popisu a následné interpretaci vlastností zkoumaných jevů. Chybné určení extrémů a jejich následná vizualizace vede k chybným úsudkům o tvrzeních, která jsou uživatelům map sdělována. Korektní přístup matematického zpracování dat a využití odpovídajících metod vizualizace je vhodné u šetření statistických souborů, při kterých se upozorňuje především na ty hodnoty, jež jsou nějakým způsoben neobvyklé.

Prof. RNDr. Vít VOŽENÍLEK, CSc., Doc. RNDr. Jaromír KAŇOK, CSc., Mgr. Pavel TUČEK, Katedra geoinformatiky, Přírodovědecká fakulta, Univerzita Palackého v Olomouci, tř. Svobody 26, 771 46 Olomouc, Česká republika, e-mail: vit.vozenilek@upol.cz, jaromir.kanok@upol.cz, pavel.tucek@upol.cz

Pojem extrém

Extrémní hodnoty jsou součástí standardního popisu zkoumaného jevu, avšak při jejich nevhodné interpretaci mohou vést k podání zkreslené informace, která je obsažena v pojmu **extrém**[1]. Pojem extrém může v běžné řeči nabývat nejrůznějších významů. Většinou se jedná o hovorově označovaný jev, který dosáhl nějaké neočekávané hodnoty nebo nějakého neočekávaného stavu. Toto pojetí souvisí spíše s označováním hodnot extrémně vysokých, nežli jakkoli nápadně vybočujících.

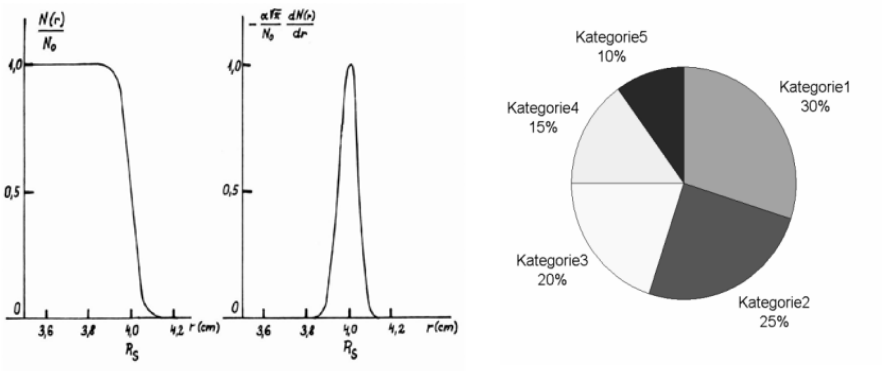
Pod pojem přírodní extrémy jsou zpravidla zahrnuty přírodní jevy vyznačující se významnými dopady na přírodu a lidskou společnost (Brázdil a kol. 2007). Zahrnují širokou škálu druhově pestrých jevů, spadajících tematicky do oblasti geofyziky (např. zemětřesení), geologie a geomorfologie (např. skalní řícení, sesuvy), meteorologie (např. vichřice, krupobití), hydrologie (např. povodně) či oceánografie (např. tsunami). Z hlediska jejich projevu lze rozlišovat mezi přírodním nebezpečím, přírodním rizikem a přírodní katastrofou (Glade a Dikau, 2001).

V matematice a dalších exaktně pracujících oborech existuje pro pojem extrém přesné vymezení, které jasně stanovuje význam extrému jako hodnoty, která je pro celý jev největší, resp. nejmenší na dané oblasti. Z hlediska statistického je však zkoumání extrému daleko složitější a rozsáhlejší, neboť doposud nebyla posuzována a plně implementována komplexní znalost o chování zkoumaného jevu. Uvažuje-li se zkoumaný jev jako statistická náhodná veličina, pak je studium extrémů složitější. Nejprve se rozhodování o extrémních hodnotách musí rozdělit na několik fází, a to na detekci extrémů, jejich prokazování, interpretaci a v případě kartografie i na jejich následnou vizualizaci.

Detekce extrémů

Při detekci (vymezování) extrémů je nejprve nutné stanovit, které hodnoty se za extrémy považují, a rozlišovat extrémy datové sady ve frekvenční oblasti a datové oblasti.

Extrémní hodnoty ve frekvenční oblasti **(frekvenční extrémy)** jsou takové hodnoty zkoumaného jevu, které svým výskytem, resp. četností výskytu, převyšují prokazatelně četnost výskytu ostatních hodnot zkoumaného jevu. Jedná se tedy výhradně o hodnoty s extrémním četnostním výskytem v datové sadě. Frekvenční extrémy lze graficky znázornit pomocí libovolného vyobrazení četností, popř. procentuálním vyjádřením podílu jednotlivých skupin na celkovém datovém souboru (obr. 1). Toto se týká i případů, kdy četnost této hodnoty je malá popř. procentuální podíl je malý.

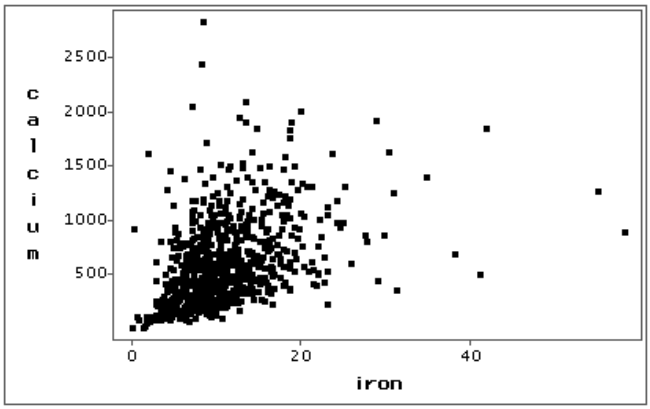


Obr. 1 Frekvenční extrémy zobrazené pomocí grafu hustoty pravděpodobností (vlevo) a procentuálního podílu (vpravo)

[1] Definice: Nechť je dán zkoumaný jev J(t), reprezentovaný datovou sadou D(t) s hodnotami, které odpovídají pozorováním $x_1(t)\ldots x_n(t)$, popř. $x(t_1)\ldots x(t_n)$, kde t reprezentuje čas. Potom extrémem jsou hodnoty x_i, popř. $x(t_i)$, které z hlediska výskytu nereprezentují přirozené chování zkoumaného jevu.

Extrémní hodnoty v datové oblasti **(datové extrémy)** jsou takové hodnoty zkoumaného jevu, které prokazatelně převyšují (nebo nedosahují) charakter zkoumaného jevu, což lze definitoricky zavést jako body ležící mimo majoritní část datového souboru. Při zavádění datového extrému je však třeba brát v úvahu, že datovou oblastí se rozumí obor hodnot, kterých může jev (zkoumaná náhodná veličina) nabývat. Extrémní hodnota v datové oblasti nemusí jednoznačně znamenat extrémní hodnotu zkoumaného a následné prokazování a následně i vizualizace jsou zavádějící a matoucí.

Fáze detekce extrémů se v datové sadě, kde se bere v úvahu frekvenční i datová oblast, provádí rozborem dat a statistickým zpracováním. Vlastní detekce extrémů je zaměřena na získání numerických hodnot extrémů z datové sady, respektive z modelu reprezentujícího zkoumaná data, čímž se rozumí výpočet extrémů pomocí aparátu matematické analýzy aplikovaného na numerický model zkoumaného jevu (Jarník 1984a). Takové extrémy se graficky vizualizují rovněž pomocí tzv. scatterplotu (obr. 2) nebo libovolného zobrazení outlierů – boxplot (obr. 3).



Obr. 2 Scatterplot pro znázornění extrémů v datové oblasti

Nelze obecně říci, že největší či nejmenší hodnoty jsou automaticky hodnotami extrémními. V této fázi studia extrémů je důležitá rozvaha o tom, zdali je pro posuzovatele důležitá extrémní hodnota v frekvenční oblasti nebo v datové oblasti (čili rozhodnutí zkoumají-li se frekvenční nebo datové extrémy) a souvisí-li tato hodnota se zkoumanou náhodnou veličinou nebo se jedná o hodnotu prokazatelně nesouvisející.

Prokázání extremity

Druhou fází studia extrémů je prokazování pravdivosti tvrzení o extremitě detekované hodnoty. Takové tvrzení však musí být podloženo sofistikovanou metodou, která na určité hladině významnosti zaručí, že se jedná o extrémní hodnotu zkoumaného jevu, nikoli o nápadně vybočující hodnotu, tzv. **outlier**[2]. Extrémy jsou přirozenou součástí chování jevu a mají k němu patřičný vztah, zatímco outliers žádný vztah k chování zkoumanému jevu nemají.

Outliers vznikají chybami v měření, špatným zápisem do počítačových systémů či pouhým nedopatřením. Pokud se soubory dat vyhodnocují včetně outliers, jsou získané výsledky zcela zkresleny. Proto musejí být outliers ze statistických souborů odstraněny. Metod k detekci a vyloučení outlierů ze statistických souborů a následnému odhalení skrytých extrémů je celá řada (např. Rousseeuw a Leroy 1987). Avšak některé metody jsou vzhledem ke své propracovanosti a nume-

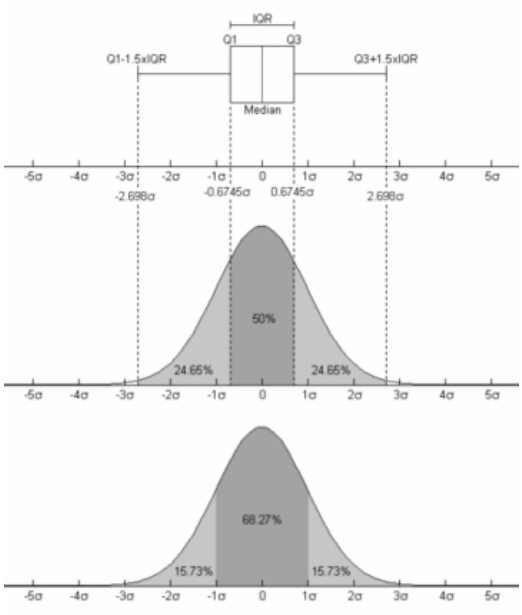
[2] Definice 1: Nechť je dán zkoumaný jev J(t), reprezentovaný datovou sadou D(t) s hodnotami, které odpovídají pozorováním x1(t)…xn(t), t reprezentuje čas. Hodnota pozorování xi se nazývá outlier, jestliže se jedná o nápadně vybočující pozorování způsobené chybným měřením, chybným zápisem nebo o signifikantně vybočující a nesouvisející datový údaj.

rické složitosti pro zpracování demografických souborů dat nevhodné (např. Rousseeuw a Van Driessen 1998). Ovšem jiné velice sofistikované a přitom numericky nenáročné metody (např. Filzmoser 2004) plně vyhovují jejich nasazení v demografii, čímž mnohonásobně zvyšují informativní hodnotu výsledných poznatků o zkoumaných jevech. Výjimečně může být po zhodnocení expertem outlier považován za extremní hodnotu.

Postup při analýze datových sad za účelem prokázání extrémů probíhá ve dvou krocích. Nejprve se identifikují body „podezřelé" z extrému, čímž se vymezí množina těch bodů, které se budou dále analyzovat. Ve druhém kroku se tato množina rozdělí na množinu extrémů a množinu outlierů.

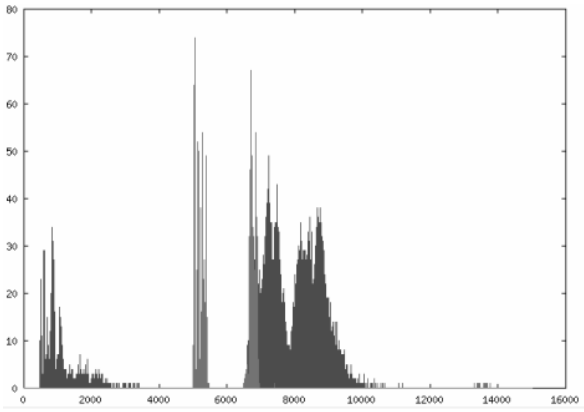
Jednoduché metody zaměřené na detekci outlierů jsou založeny na rozložení pravděpodobností zkoumané náhodné veličiny a dělí se na grafické a numerické (Filzmoser 2004, 2005). Do grafických metod zaměřených na rozpoznávání outlierů se řadí jednoduché vizualizační metody boxplot a histogram.

Boxplot je vizualizační metoda, která zkoumá rozložení hodnot kolem vypočteného mediánu (Rousseeuw a Leroy 1987). Tato metoda se neváže na aritmetický průměr, tedy není ani vázaná na tabulkové normální rozdělení. Podstatným faktem je stanovení tzv. mezikvartilového rozpětí (IQR), což je vzdálenost prvního a třetího kvartilu. Po stanovení IQR se snadno určuje hodnota „podezřelých" údajů. Tato hodnota (mezní hranice) se nachází ve vzdálenosti 1,5 násobku od prvního, resp. třetího kvartilu. Graficky je nejvhodnější tyto extrémy zobrazovat v boxplotu pomocí izolovaných bodů za hranicí 1,5 násobku IQR.



Obr. 3 Boxplot se zobrazenými hranicemi pro výskyt outlierů u normálního rozdělení

U metody histogramu je důležité dobře interpretovat rozložení hodnot. Histogram, který vykazuje na začátku nebo na konci výrazné výstupky (peaky), poukazuje na výskyty outlierů. Tyto outliery je ale nutno dobře identifikovat a zdůvodnit. Jedná-li se o několik málo hodnot, pak se zcela nepochybně interpretují jako outlier. V případě, že je hodnot více a jedná-li se o vícevrcholové rozdělení, pak se dané údaje nedají reprezentovat jako outliery a jedná se o extrémy. Bezprostředně potom je nutné podrobit dané hodnoty dalšímu zkoumání za účelem expertního vyhodnocování.



Obr.4 Ukázka tradičního histogramu

Stanovení počtu intervalů, jejich hranic a jejich pojmenování

Tradiční metody pro stanovení počtu intervalů založené na znalosti střední hodnoty a rozptylu náhodné veličiny jsou sice z pohledu statistiky vhodnou metodou, ale z pohledu kartografie většinou nezohledňují a detailně nepopisují zkoumaný jev. Metoda stanovení počtu intervalů při řešení extremity vychází ze znalosti možných výsledků náhodné veličiny. Tato oblast se postupně dělí na snižující se počet intervalů tak dlouho, dokud není poměr entropie a počtu intervalů maximální. Tato metoda je propojením statistického pohledu na data a kartografické vizualizace. Tímto postupem se stanoví optimální počet intervalů, který zaručí nejvyšší míru zachované informace a tím i vhodnou výchozí pozici pro kartografickou vizualizaci.

Při zpracování demografických dat se ve většině případů pracuje s diskrétní náhodnou veličinou, která nabývá hodnot z nějaké množiny. Tato množina je množinou možných realizací náhodné veličiny. Pro stanovení počtu intervalů se při řešení extremity používá zcela unikátní přístup založený na **míře entropie**, kterou daný počet intervalů ve výsledné kartografické vizualizaci vyjadřuje. Entropie je pojem vycházející z teorií pravděpodobnosti a informace a je součástí kybernetiky. Entropií[3] se rozumí míra informační vydatnosti, nebo také neurčitosti pokusu.

Informační entropie je také nazývána **Shannonovou entropií**, a to po Claude E. Shannonovi, který zformuloval mnoho klíčových poznatků teoretické informatiky a které se touto aplikací stávají klíčové také v oblasti geoinformatiky. Obecně pro systém (také demografický jev) $S \in \{s_1, s_2, \ldots, s_n\}, n \leq \infty$ s konečným počtem možných stavů a pravděpodobnostní distribucí $P(s_i)$ je informační entropie definována jako střední hodnota:

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

Entropie je maximální pro rovnoměrné rozložení

$$P(s_i) = \frac{1}{n} \text{ pro } \forall i \tag{2}$$

$$H(S) = -\sum_{i=1}^{n} \frac{1}{n} \log_2 \frac{1}{n} = -\log_2 \frac{1}{n} = \log_2 n \tag{3}$$

[3] Definice: Entropie je střední hodnota míry informace potřebné k odstranění neurčitosti, která je dána konečným počtem vzájemně se vylučujících jevů.

a minimální pro zcela deterministický systém

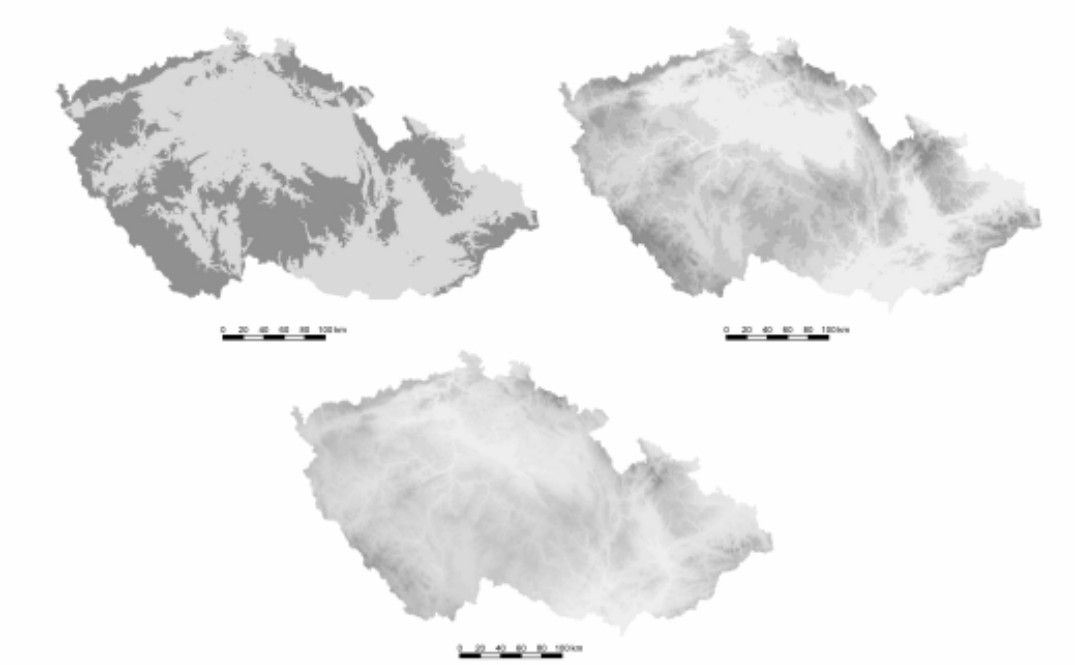
$$\exists P(s_k) = 1 \text{ a } P(s_i) = 0 \text{ pro } \forall i \neq k \tag{4}$$

$$H(S) = -\sum_{i=1}^{n} P(s_i) \log_2 P(s_i) = -\log_2 1 = 0 \tag{5}$$

Analogicky pro případ informací v mapě lze zformulovat, že informační entropie mapy je střední hodnota prostorové informace nesená v mapě. Z tohoto postupu je tedy patrné, že optimální počet intervalů je stanoven takovým způsobem, že:

$$\frac{H(i)}{N(i)} \rightarrow \max , \tag{6}$$

kde $N(i)$ je počet intervalů, $H(i)$ je entropie napočítaná na tomto dělení.



Obr. 5 Tři různé hodnoty relativní (v %) entropie (vlevo 32,3; vpravo 74,5; dole 78,08)

Dalším krokem při prokazování extremity jevu je stanovení nejnižší hodnoty extrému. Při tomto postupu je nutné brát v úvahu postupnou hierarchii hodnot náhodného jevu. Mezi hodnoty jevu patří všechny možné hodnoty, které lze při zkoumání náhodného jevu naměřit. Extrémem jevu je hodnota nevyhovující výše uvedené definici. Při určení hodnoty extrému se lze setkat se situací, kdy extrémem může být více hodnot v souboru. Proto je třeba stanovit nejnižší hodnotu, popř. nejvyšší hodnotu extrému. Nelze však za extrém vždy považovat pouze nejvyšší a nejnižší hodnoty náhodné veličiny. Tato situace může být dobře ilustrována na modelu získaného odhadem z dat. Model je vždy odhadem náhodného jevu a popisuje jej jako celek, tudíž můžeme získat i data, která jsou jinak neměřitelná. Při zkoumání modelu pomocí metod matematické analýzy (Jarník 1984a,b) lze dojít k poznání, že extrém se nachází nad, popř. pod, nejvyšší, popř. nejnižší naměřenou hodnotou.

Interpretace studia extrémů

Interpretací výsledků studia extrémů se rozumí sémantická interpretace, tedy slovní vyjádření extrémů zahrnující jejich formulaci ve vztahu ke zkoumanému jevu. Kartografická vizualizace je pak sdělení této formulace kartografickými prostředky/metodami tak, že interpretace kartografického vyjádření významu je identické s interpretací sémantickou.

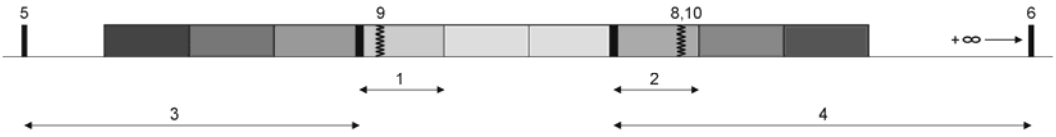
Při hledání extrému je nutné si uvědomit podstatnou skutečnost. Extrémní hodnotu lze chápat dvojím způsobem. Prvně jako hodnotu danou určitým předpisem, normou nebo zákonem. Za druhé jako hodnotu určenou statistickými metodami. Pokud jsou obě hodnoty identické nebo leží ve stejném intervalu odvozeném metodou postupného dělení, pak lze jejich kartografickou vizualizace provést stejným způsobem. Je-li ovšem každá z hodnot z jiného intervalu, pak je nezbytné použít různé kartografické přístupy. Názorně lze tuto situaci ukázat na ilustrativním příkladu, kdy náhodná veličina nabývá svých hodnot v intervalu <-1, 1>, kde extrémní hodnoty se nacházejí v intervalech <-1, -0,7> a <0,7, 1>. Tomuto faktu se podřídí i intervaly při tvorbě mezí intervalů stupnice pro vyjádření jevu v mapě. Pokud se ale při detekci outlierů zjistí, že extrémem muže být i hodnota 0,6, popř. -0,6 potom je potřeba zvážit, jak tuto situaci v mapě vyjádřit.

Kartografická vizualizace

Posledním krokem ve statistickém řešení extremity je její signifikantní vizualizace. Je více než zřejmé, že správná vizualizace je velmi citlivou částí celého postupu, kterou bez předchozích fází nelze realizovat. I v tomto kroku však vznikají problémy, a to jak z pohledu statistiky, tak z pohledu norem a zavedených postupů.

Nalezení optimální kartografické metody, která je schopna zachytit tuto obohacenou informaci, je úkolem tematických kartografů. Nejpoužívanější kartografické metody zanechávají informaci pouze o jedné předem definované hranici. Zahrnutí dlouhodobého vývoje nebo širšího statistického souboru může vést k posunu této definované hranice do jiných hodnot a znemožnění zjištění, zdali se jedná o skutečnou extremitu z pohledu statistiky nebo o „stanovenou“ extremitu normami.

Obr. 6 Grafické vyjádření extrémních hodnot ve stupnici pro kartogram, ve kterém nebyly zkoumány extrémní hodnoty



Obr. 7 Grafické vyjádření extrémních hodnot ve stupnici pro kartogram, ve kterém byly zkoumány extrémní hodnoty

Legenda k obrázkům 6 a 7:
1 – Oblast minimálních naměřených nebo zjištěných hodnot; 2 – Oblast maximálních naměřených nebo zjištěných hodnot; 3 – Oblast minimálních extrémních hodnot; 4 – Oblast maximálních extrémních hodnot; 5 – Teoreticky nejnižší dosažitelná hodnota; 6 – Teoreticky nejvyšší dosažitelná hodnota (příklad pro hodnotu v nekonečnu); 7 – Extrémní minimální hodnota (musí ležet v oblasti 3; příklad: kdy nebyla naměřena nebo zjištěna); 8 – Extrémní maximální hodnota (musí ležet v oblasti 4); 9 – Naměřená nebo zjištěna nejmenší hodnota (musí ležet v oblasti 1); 10 – Naměřená nebo zjištěna největší hodnota (musí ležet v oblasti 2; příklad: kdy leží i v oblasti 4 a shoduje se s hodnotou 8).

Extrémních hodnoty lze kartograficky vyjádřit následovně:

A. Oblast minimálních (obr. 6, 7 – legenda: 1), resp. maximálních (obr. 6, 7 – legenda: 2) naměřených nebo zjištěných hodnot

Oblast minimálních, resp. maximálních naměřených nebo zjištěných hodnot se doposud vymezuje převážně empirickými nebo i elementárními statistickými postupy. Vždy se jedná o vytvoření takové intervalové stupnice, ve které se vymezí mimo jiné i intervaly s minimálními a maximálními hodnotami. Postup ve vytváření takové intervalové stupnice je buď empirický, nebo s použitím základních rozptylových statistických metod.

Nad naměřenými daty se například vytváří stupnice intervalová s pravidelnou, s pravidelně rostoucí nebo s pravidelně klesající šířkou intervalu. Pokud v určité části spektra dat hodnoty chybějí, vytváří se intervalová stupnice skoková s hiátem stejným způsobem. V lepším případě se zkoumá rozdělení četností hodnot a pak se šířka intervalů volí nepravidelně, obvykle podle nejlépe přilehajícího teoretického rozdělení četností (normální, exponenciální, Pearsonova křivka III. typu apod.).

Volí se například meze intervalů podle velikosti směrodatné odchylky (s) ve vztahu k průměru ($x_{prům.}$): $(-\infty;x_{prům.} - s\rangle$; $(x_{prům.} - s;x_{prům.}\rangle$; $(x_{prům.;}\ x_{prům.} + s\rangle$ $(x_{prům.} + s\ ; + \infty)$, nebo méně častěji ve vztahu k jiné střední hodnotě, např. k mediánu, k průměrné odchylce od průměru apod. Meze intervalů se též určují pomocí kvartilů, pentilů, decilů. Pokud je rozdělení četností vícevrcholové, používají se ke stanovení šířky intervalů sedla a vznikají intervaly o nepravidelné šířce (Kaňok 1999b).

Nicméně k takto vytvářeným intervalům mohou být výhrady. Pokud se například vymezí intervaly ve vztahu ke střední hodnotě (např. u normálního rozdělení), rozdělí se oblast největší hustoty výskytu zkoumaného jevu. Pokud se vloží hranice intervalů právě do průměru, oblast homogenity se symetricky rozdělí. V tomto případě zpracovatel dat musí vybrat stupnici podle povahy řešeného problému a musí vědět, zda chce zobrazit homogenitu v oblasti průměru nebo rozložení podprůměrných, resp. nadprůměrných hodnot.

Kartografická znázornění minimálních, resp. maximálních naměřených nebo zjištěných hodnot se nejčastěji realizuje metodami pseudokartogramu a kartogramu (např. jednoduchý, kvalifikační, selektivní, tečkový, pseudokorelační atd.), metody kartodiagramu, metoda teček-topografický způsob (Kaňok 1999a).

B. Oblast minimálních (obr. 9 – legenda: 3), resp. maximálních (obr. 9 – legenda: 4) extrémních hodnot.

Na rozdíl od výše uvedených případů (viz A.) jsou oblasti minimálních, resp. maximálních extrémních hodnot vymezitelné mnohem přesněji, objektivněji – viz výše uvedený text o detekci a prokazatelnosti extremity. Předtím běžně používané metody byly založeny na použití výběrových charakteristik datového souboru.

Vytváření hranic intervalů:

a) $\langle\ x_{prům.} - s\ \ ;\ \ x_{prům.} + s\ \rangle$

b) $\langle\ x_{prům.} - 2s\ \ ;\ \ x_{prům.} + 2s\ \rangle$

Tyto intervaly slouží pro vymezení hranic v případě, že je pro zkoumaný jev podstatných 68% dat, resp. 99% dat, což odpovídá a), resp. b).

Ke znázornění se používají stejné kartografické metody (pseudokartogram a kartogram – jednoduchý, kvalifikační, selektivní, tečkový, pseudokorelační atd.; metodu kartodiagramu, metodu teček - topografický způsob). Na rozdíl od předcházející skupiny se kromě oblastí nejnižších a nejvyšších naměřených hodnot navíc barevně odlišují oblasti, kde se vyskytují hodnoty extrémní. Doporučuje se významněji zvýraznit odstíny použité barvy nebo je lépe doplnit stávající odstíny barev v příslušných intervalech rastrem, nejlépe liniovým. (Kaňok 1999a).

C. Nejnižší, resp. nejvyšší teoreticky dosažitelná hodnota (obr. 9 – legenda: 5, 6)

Křivka rozložení četností hodnot sledovaného jevu a následně zpracované teoretické rozdělení četností ukáže možnosti dosažitelnosti extrémních hodnot. U normálního rozdělení to mohou být například hodnoty - ∞ ; + ∞.

V konečných výběrových souborech lze určit nejnižší i nejvyšší možnou hodnotu. V některých sledovaných jevech určily nejnižší nebo nejvyšší teoretickou hodnotu výsledky bádání jiných věd (např. fyzika: minimální teplota -273,15 °C). V kartografickém vyjadřování se upřednostňují výrazné barvy, např. purpurové, které jasně upozorní na minimální, resp. maximální, teoreticky dosažitelnou, hodnotu jevu.

D. Naměřená nebo zjištěná extrémní minimální, resp. extrémní maximální hodnota (obr. 9 – legenda: 7, 8)

Reálně naměřené nebo zjištěné extrémní hodnoty musí ležet za teoretickými mezemi, které extrémy vymezují (viz B.). Je však nutno vnímat dvě teoretické meze, a to jednak pro vymezení minimálních extrémních hodnot a jednak pro vymezení maximálních extrémních hodnot. Teprve jedna z nich je nejmenší v oblasti minimálních extrémních hodnot a jedna z nich je největší v oblasti maximálních extrémních hodnot. Jsou to obvykle jiné hodnoty než hodnoty v předchozích případech (viz C.).

Při kartografickém zpracování reálně naměřených extrémních hodnot se upřednostňují méně výrazné barvy než pro označení minimální, resp. maximální teoretické hodnoty sledovaného jevu. Do kartogramu se pak, do příslušné dílčí územní jednotky, bodovým znakem vyznačí, kde se vyskytly reálně naměřené nebo zjištěné extrémní hodnoty (minimální i maximální).

E. Naměřená nebo zjištěná nejmenší, resp. největší hodnota přičemž to nemusí být extrémy (obr. 8, 9 – legenda: 9, 10)

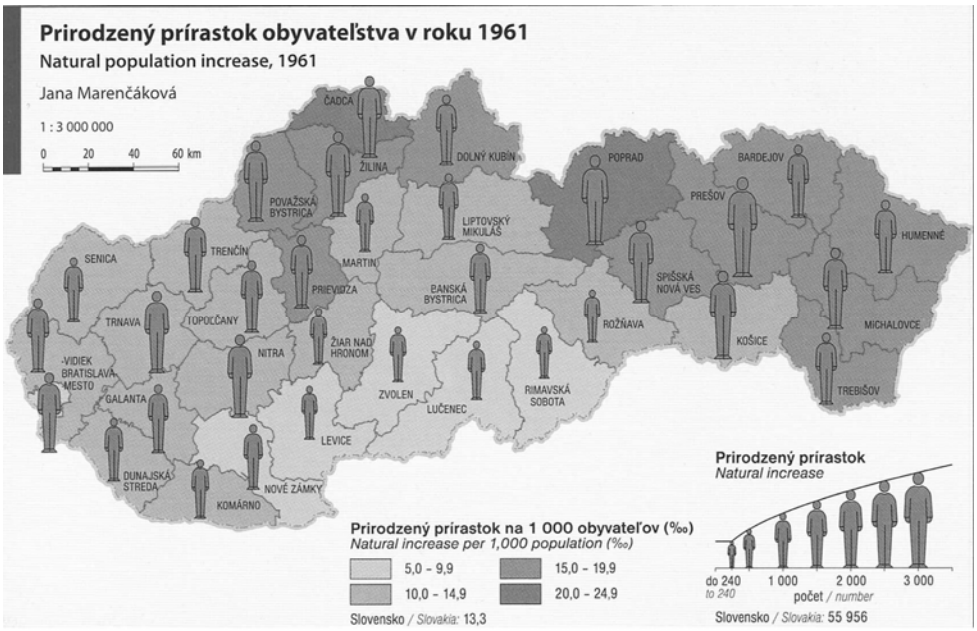
Hodnoty vyskytující se za mezemi extrémů se stávají hodnotami extrémními. Avšak za těmito mezemi se mohou, ale také nemusí vyskytovat reálně naměřené nebo zjištěné hodnoty. Tyto hodnoty se do intervalové stupnice označí ještě méně výraznou barvou než v předchozím případě. Do kartogramu se pak, do příslušné dílčí územní jednotky, bodovým znakem vyznačí, kde se vyskytly reálně naměřené nebo zjištěné nejnižší či nejvyšší hodnota.

Pro znázornění hodnot uvedených v C, D, E, lze též použít grafy a diagramy rozličných druhů. Znázornění či barevné zvýraznění nejmenší a největší naměřené či zjištěné hodnoty nebo znázornění nejmenšího a největšího extrému by neměl být problém.

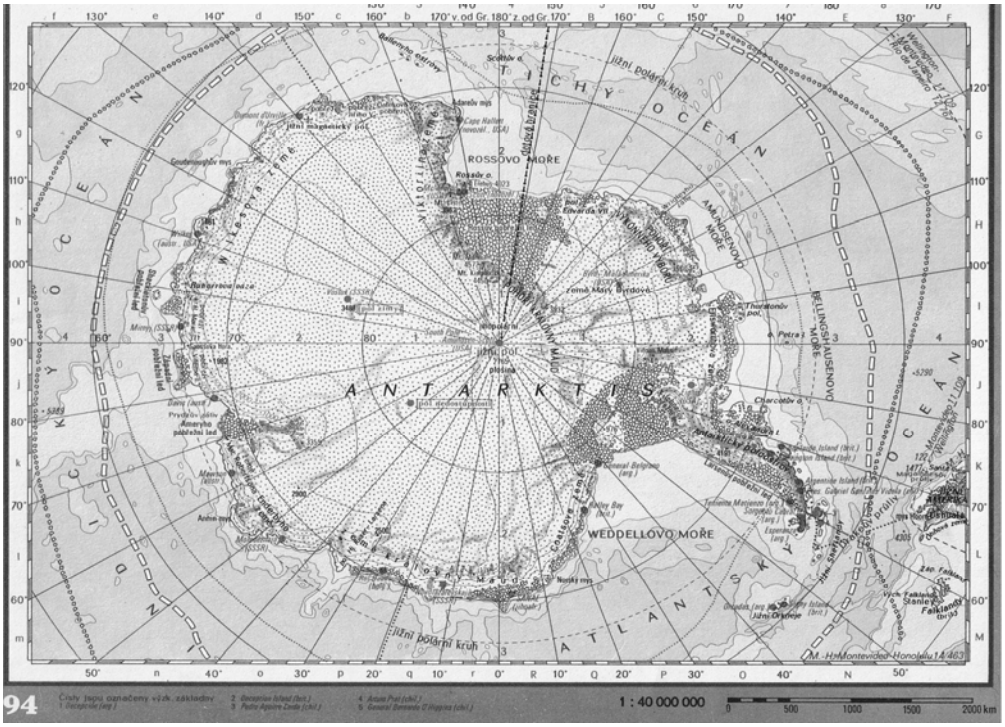
Doporučujeme provádět kontrolu správnosti metodiky získávání dat a hned v počátku vylučovat data, která metodicky správně zjištěna nebyla. S touto kontrolou je spojeno vyloučení nápadně vybočujících dat (outlier). Teprve poté lze přikročit ke statistickému a následně ke kartografickému vyjádření dat.

Pokud se řeší znázornění hodnot ve skupině A, musí mu předcházet statistická zpracování, vymezení oblastí extrémů a následné stanovení šířky intervalů stupnic. Podobným způsobem při zpracování dat (detekce extrémů, vytváření stupnic, kartografické znázornění) se postupuje i při vytváření kartodiagramu. Je nutno dodat, že jednotlivé datové hodnoty všech typů se v současných kartografických výstupech bohužel běžně neuvádějí. (obr. 8).

Bohužel se stává, že na některých mapách se v prostorovém vyjádření uvádějí „extrémy“, které buď nejsou definovány vůbec, nebo jsou definovány vágně. Příkladem může být „Pól zimy“ a „Pól nedostupnosti“ na mapě Antarktidy (obr. 9).



Obr. 8 Příklad pseudokartogramu a kartodiagramu, kde nebyly zkoumány a následně graficky vyjádřeny extrémní hodnoty. (Atlas obyvatelstva Slovenska 2006)



Obr. 9 Příklad znázornění neurčitě definovaných extrémních hodnot (Pól zimy. Pól nedostupnosti). (Atlas světa 1970)

Závěr

Nalezení optimálních prostředků vhodných pro detekci, prokázání, interpretaci a vizualizaci extrémů v datových sadách je náročnou úlohou. Výše uvedenou metodikou se opírají autoři o statistický přístup vycházející z předzpracování dat, ve kterém se nejprve detekují extrémy (a odlišují od tzv. outlier) a pak se prokáže jejich pravdivost, která vyúsťuje v signifikantní vizualizace. Kartografický přístup se opírá současné metody tematické kartografie, které dále rozvíjí. Navržený přístup je založen na nalezení hranic, které budou prokazatelně určovat meze výskytu extrémů, za kterými se budou již extrémy nacházet. Autoři vyčleňují dvě kategorie extrémů, a to extrémy datové a frekvenční. Způsoby jejich rozpoznání a následné vizualizace jsou podrobně popsány a graficky dokumentovány. V neposlední řadě je zmíněna míra entropie v mapách, která může rovněž posloužit k lepší kartografické vizualizaci extrémů.

Příspěvek je součástí výstupů projektu GA ČR 205/06/0965 „Vizualizace, interpretace a percepce prostorových informací v tematických mapách.

Literatura

Atlas krajiny Slovenskej republiky. (2002). Bratislava (Ministerstvo životného prostredia SR a Esprit).

Atlas obyvateľstva Slovenska. (2006). Eds. Mládek, J. et al. Bratislava (Univerzita Komenského).

Atlas světa. (1970). Eds. Koláčný, A. et al. Praha (Kartografické nakladatelství).

BRÁZDIL, R., KIRCHNER, K. a kol. (2007). *Vybrané přírodní extrémy a jejich dopady na Moravě a ve Slezsku*. Brno (Masarykova univerzita).

FILZMOSER. P. (2005). Identification of Multivariate Outliers: A Performance Study". *Austrian Journal of Statistics*, 34, 2, s. 127-138.

FILZMOSER. P. (2004). *A multivariate outlier detection method*; "Proceedings of the Seventh International Conference on Computer Data Analysis and Modeling", S. Aivazian, P. Filzmoser, Y. Kharin (ed.); Belarusian State University, s. 18- 22.

FOTHERINGHAM, A.S., BRUNSDON, CH., CHARLTON, M. (2000). *Quantitative Geography*. *Perspectives on Spatial Data Anaylsis*. London (Sage Publications).

GLADE, T., DIKAU, R. (2001). Gravitative Massenbewegung – vom Naturereignis zur Naturkatastrophe. *Petermanns Geographische Mitteilungen*, 145, 6, s. 42-53.

JARNÍK V. (1984a). *Diferenciální počet I*. Praha (Academia).

JARNÍK V. (1984b). *Diferenciální počet II*. Praha (Academia).

KAŇOK, J. (1999a). *Tematická kartografie.* Vročeno 2000. Ostrava (Ostravská univerzita v Ostravě).

KAŇOK, J. (1999b). Klasifikace stupnic a zásady jejich tvorby pro kartogram a kartodiagram. *Kartografické listy*, 7, s. 75-86.

PRAVDA, J. (2006). *Metódy mapového vyjadrovania. Klasifikácia a ukážky.* Geographia Slovaca, 21. Bratislava (Slovenská akadémia vied, Geografický ústav).

ROUSSEEUW, P.J., LEROY, A.M. (1987). *Robust Regression and Outlier Detection.* Series in Applied Probability and Statistics, New York (Wiley-Interscience).

ROUSEEUW, P.J., VAN DRIESSEN, K. (1998). *Computing LTS Regression for Large Data Sets*, Technical Report, Antverp (University of Antverp).

VOŽENÍLEK, V. (2001). *Aplikovaná kartografie I. Tematické mapy.* 2. vydání. Olomouc (Univerzita Palackého v Olomouci).

VOŽENÍLEK, V. (2005). *Cartography for GIS. Geovizualization and Map Communication.*.Olomouc (Univerzita Palackého v Olomouci).

S u m m a r y

Extreme detection, provableness and visualization of statistical data files

The aim of the article was to show, that the extreme values need to be visualized with the special attention. It is not necessary to call the highest value of the measured data as a extreme value. We have designed the methodology to recognize the extreme values in the frequency and data area of the dataset with the help of the statistical data processing. The fundamentals of this methodology are based on the estimation of the significant boundary for the extreme values. As we have said before, the extreme values were classified into two categories. The first category is defined as a extreme values in the frequency area and the second area is defined as a extreme values in the data area of the dataset. The end of the statistical processing of the dataset is dedicated to the computation of the entropy. This method could serve for the computation of the significant

number of intervals for the drawing of the map. The end of the article deals with the interpretation of the obtained results. Some examples are also shown.

Lektoroval:

Prof. Ing. Bohuslav VEVERKA, DrSc.,
České vysoké učení technické, Fakulta stavební, Praha, Česká republika