PROSTOROVÉ HIERARCHICKÉ SHLUKOVÁNÍ

Jiří HORÁK[1], Igor IVAN[1], Tomáš INSPEKTOR[1]

[1] Institut geoinformatiky, Hornicko-geologická fakulta, VŠB-TUO, 17. listopadu 15/2172, 708 33, Ostrava-Poruba, ČR

jiri.horak@vsb.cz, igor.ivan@vsb.cz, tomas.inspektor@vsb.cz

Abstrakt

Jeden z potenciálně významných nástrojů pro studium homogenity dat a jejich kategorizaci představuje shlukování jako jedna z multivariačních metod analýzy dat. Klasické metody shlukování nevyužívají prostorové vlastnosti, a proto vytvářejí shluky (a kategorie) objektů, které nerespektují prostorové vztahy (zejména požadavek kontinuity výsledného shluku).

Při vývoji prostorového rozšíření se střetáváme s problémy vymezení sousedství, definice vhodných vah apod. Současně je to výzva k hledání nových možností organizace dat. Rovněž není jednoduché tvorba přehledných a srozumitelných výstupů, zohledňující atributové vzdálenosti mezi objekty a pozici objektů v hierarchii shluku.

Vyvíjená databázová aplikace používá běžné míry vzdálenosti, jako jsou Euklidovská, Euklidovská s druhou mocninou, Manhattanská (L1norm) a Mahalanobisova. K dispozici jsou hlavní hierarchické agregační metody používané v různých aplikacích, konkrétně úplné propojení (complete linkage), jednoduché propojení (single linkage), průměrné propojení (average linkage), Wardova metoda (Wardova-Wishartova metoda), metoda těžiště a mediánová metoda.

K testování jsou používány 2 aplikační oblasti – adresní body a areály obcí. U adresních bodů se testuje sousedství definované Thiessenovými polygony a Euklidovskou vzdáleností. Množství vznikajících variant se prověřuje s cílem zjištění stabilních struktur. Dosavadní výsledky naznačují výhodné chování a robustnost zejména metody úplného propojení ve shodě s některými jinými pracemi.

Metoda by mohla sloužit pro vhodnou kategorizaci území s respektováním požadovaných prostorových vazeb a priorit.

Abstract

Cluster analyses represent one of the important tools for study of data homogeneity and its classification. Traditional methods do not use spatial aspects of the data (spatial features) and create clusters which do not respect spatial relationships (usually the contiguity of the cluster is required). During the development of spatial variants of clustering we meet many issues concerning definitions of neighbouring, using appropriate weights etc. It can be seen as a big challenge for new possibilities of data organisation. Other issues are linked with creating easy and understandable outputs respecting attribute distances and a position of objects in the hierarchy of clusters.

A new database application is under development. Following distance measures are applied: Euklidean, Euklidean Square, Manhattan (L1norm) a Mahalanobis. Concerning agglomeration methods our aim is to implement all main methods including complete linkage, single linkage, average linkage, Ward's method, centroid and median methods.

We selected two pilot application sets – polygonal scheme of municipalities and address points. For address points we define neighbouring by Thiessen polygons or thresholding of Euklidean distances. A comparison with a traditional hierachical agglomeration method is provided. Results for complete linkage are demonstrated.

Klíčová slova: prostorové hierarchické shlukování; shlukovací analýza.

Keywords: spatial hierarchical clustering; cluster analysis.

1. ÚVOD

Klasifikace objektů do tříd je využívána pro řadu prostorových aplikací, jako jsou např. generalizace mapových výstupů, regionalizace území či klasifikace objektů v dálkovém průzkumu Země. K vytvoření klasifikace se přirozeně nabízí využití shlukovacích multivariačních statistických metod, které však a priori neobsahují potřebná prostorová omezení. Vhodná implementace prostorového rozšíření shlukovacích metod by umožnila lépe studovat prostorové vazby, vytvářet přirozené třídy prostorových objektů a charakterizovat je. Jednou z možností je prostorové hierarchické shlukování.

2. SHLUKOVÁ ANALÝZA

Shluková analýza označuje skupinu metod, jejichž cílem je na základě analýzy vícerozměrných dat provést roztřídění množiny objektů do několika relativně homogenních podsouborů, označených jako shluky (clustery). Objekty uvnitř shluků mají být co nejvíce podobné a objekty patřících do různých shluků co nejvíce rozdílné. Základní kritériem pro tvorbu shluků objektů je podobnost mezi objekty. Měření podobnosti lze provádět pomocí vhodné míry korelace, míry vzdálenosti nebo míry asociace. Korelační a vzdálenostní míry se používají především pro poměrová (ratio) data, zatímco asociační míry jsou určeny spíše pro výčtová (nominální) data.

K základní korelačním mírám patří běžně používané korelační koeficienty, jako jsou Pearsonův nebo Spearmanův korelační koeficient.

Míry vzdálenosti představují nejvíce používané míry (Meloun et al. 2005). Vzdálenosti jsou měřeny v prostoru, jehož souřadnice jsou ale představovány hodnotami měřených znaků objektů, nikoliv klasickými souřadnicemi. Nezbytnou podmínkou je standardizace těchto znaků (sjednocení měřítek), jinak dochází ke značně odchýleným výsledkům. Největší problémy se pak vyskytují u čtverce Euklidovské vzdálenosti. K běžným mírám vzdálenosti patří Euklidovská vzdálenost, čtverec Euklidovské vzdálenosti, Manhattanská vzdálenost (Hammingova metrika), zobecněná Minkovského metrika, tětivová vzdálenost a Mahalanobisova metrika. Pro lepší rozlišení je budeme označovat jako lexikální vzdálenosti na rozdíl od geografické vzdálenosti mezi objekty.

Základní koeficienty podobnosti pro míry asociace vyjmenovává např. Meloun et al. (2005). Vedle výběru vhodné míry podobnosti se musí zvolit rovněž vhodná shlukovací procedura. K běžným shlukovacím procedurám patří:

- metoda nejbližšího souseda – pár pro shlukování se vybere podle nejmenší vzdálenosti
- metoda nejvzdálenějšího souseda - pár pro shlukování se vybere podle největší vzdálenosti
- metoda průměrné vzdálenosti – vychází se z průměrné vzdálenosti všech objektů v 1.shluku ke všem objektům ve 2.shluku.
- Wardova metoda – kritériem je minimalizace heterogenity shluků. V každém kroku se spočítá přírůstek součtu čtverců odchylek, vzniklý sloučením shluků. Spojí se ty shluky, které mají minimální hodnotu přírůstku (Meloun et al. 2005).

$$VSS = \sum_{j=1}^{m} \sum_{i=1}^{k} (x_{ij} - \overline{x_j})^2$$

- metoda těžiště
- metoda mediánová

Jejich podrobnější popis a algoritmizaci lze nalézt v (Lukasová, Šarmanová 1985) či Meloun et al. (2005).

Na základě shlukové analýzy získáme hierarchickou posloupnost objektů setříděných podle podobnosti od nejméně po nejvíce podobné. Těmito objekty jsou buď původní objekty (obsažené v datové sadě) nebo nově

vzniklé objekty v procesu shlukování (de facto třídy objektů). Tuto hierarchickou posloupnost je možné vyjádřit dendrogramem (obr.2).

Výstupem bývá klasifikace vytvořených shluků. Vymezení shluků může být stanoveno zpravidla na základě počtu požadovaných výsledných shluků nebo na základě maximální přípustné nepodobnosti mezi objekty.

Základní nevýhodou pro uplatnění shlukovacích metod na prostorová data je skutečnost, nevyužívají prostorové vazby mezi objekty a vytvářejí shluky z objektů, které nemají prostorové vazby (typicky nejsou sousedy) (viz členové shluků na obr. 3). Pokud chceme respektovat prostorová omezení, musí se upravit shlukovací algoritmus.

Metoda hierarchického prostorového shlukování je popsána např. v Carvalho et al. (2009) nebo (Horák, 2011). Algoritmus zahrnuje následující kroky:

1) Určení sousedství jednotlivých areálů – zpravidla se používají topologická sousedství typu královna.

2) Spočítá se vektor lexikálních vzdáleností mezi všemi páry tvořenými sousedními areály a vybuduje se matice blízkosti (symetrická).

3) Nalezne se pár sousedů, které mají nejmenší vzájemnou vzdálenost. Tento pár se seskupí do jednoho shluku.

4) V nejjednodušším případě je pro definici nového shluku nutné kombinovat seznamy sousedů. Proto bude nový seznam sousedů vytvořen spojením seznamu sousedů města A a seznamu sousedů města B (provedeme sjednocení obou relací).

5) Pro nových N-1 shluků musí být aktualizována matice blízkosti. Aktualizace matice blízkosti (nebo vzdálenosti) závisí na metodě shlukování. Například pro metodu nejbližšího souseda je vzdálenost mezi dvěma shluky I a J minimální vektor vzdáleností mezi všemi dvojicemi vektorů proměnné ve dvou shlucích. Na druhou stranu pro metodu nejvzdálenějšího souseda je vzdálenost mezi dvěma shluky maximální vektor vzdáleností mezi všemi páry vektorů.

6) Opakujeme kroky 3 až 5, dokud zbude jen jeden shluk, který bude obsahovat všech areály.

Výsledky vykazují značné odchylky od tradičních shlukovacích algoritmů. Nemusí např. platit, že následující hodnota podobnosti je menší než u předchozího vytvořeného shluku.

Metoda hierarchického prostorového shlukování je implementována v databázové aplikaci, kterou vyvíjíme. Tato aplikace načítá hodnoty geografické vzdálenosti mezi objekty a používá ji ke tvorbě prostorových omezení při výběru párů pro shlukování. Nejjednodušší mírou geografické vzdálenosti je topologické sousedství typu královna, je možné ale aplikovat i jiné koncepty, např. euklidovskou vzdálenost (vhodné pro bodové objekty). Aplikace postupně seskupuje páry objektů, eviduje jejich podobnosti a postavení v hierarchické posloupnosti. Rovněž vypočítá prostorové souřadnice těžiště a průměrnou hodnotu sledovaných znaků (proměnných) v nově vytvářených objektech (seskupeních). Výsledky práce je možné demonstrovat na 2 příkladech.

3. PROSTOROVÉ HIERARCHICKÉ SHLUKOVÁNÍ OBCÍ OKRESU KARVINÁ

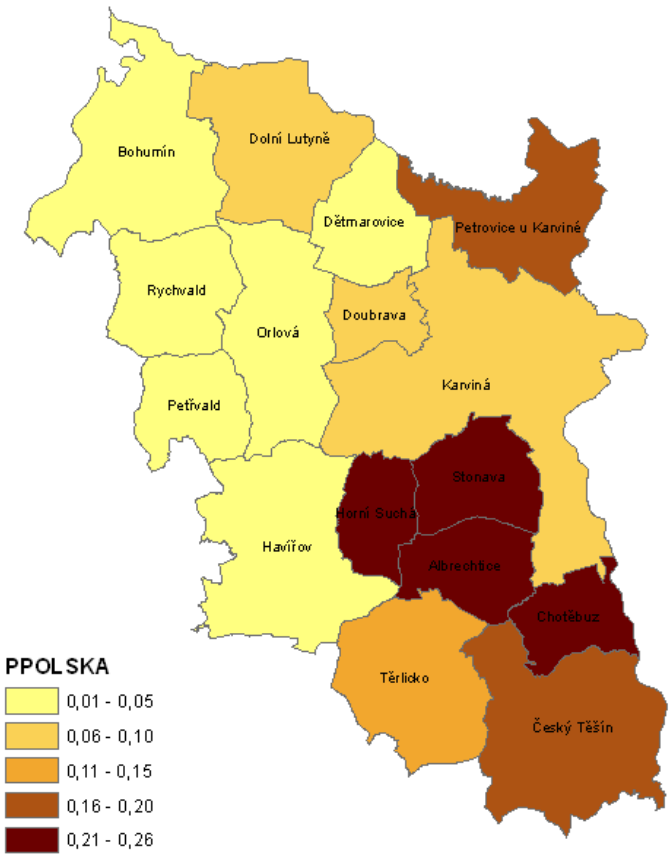
Obce okresu Karviná byly klasifikovány podle podílu obyvatel polské národnosti (data SLDB 2001) (obr. 1, stejná šířka tříd).

Nejdříve byla aplikována klasická metoda hierarchického shlukování s parametry standardizace proměnné (Z skóre), euklidovská vzdálenost (jako míra nepodobnosti podílu Poláků v obcích), shlukovací metoda nejvzdálenějšího souseda (furthest neighbor, resp. complete linkage). Výpočty byly provedeny v prostředí SPSS, verze 17. Dendrogram je uveden na obr. 2. Zařazení obcí v hierarchické posloupnosti shlukování bylo transformováno pomocí číselné klasifikace. Kořen dostává hodnotu 1, podřízený prvek stromu dostává další desetinné místo ke stávající klasifikaci, přitom levý prvek hodnotu 1 a pravý prvek hodnotu 2. Každá obec tak dostane číselný kód, který určuje její postavení v hierarchii (kód identifikuje všechny nadřazené prvky na cestě ke kořenu). Tedy např. kód 1,221112 pro Havířov ukazuje, že nejpříbuznější je situace v Bohumíně (sesterský „levý“ kód 1,221111), dále příbuzná je Orlová (v pozici „tety“) a jejich společný prapředek se jmenoval 1.22 (a jeho předek 1.2).

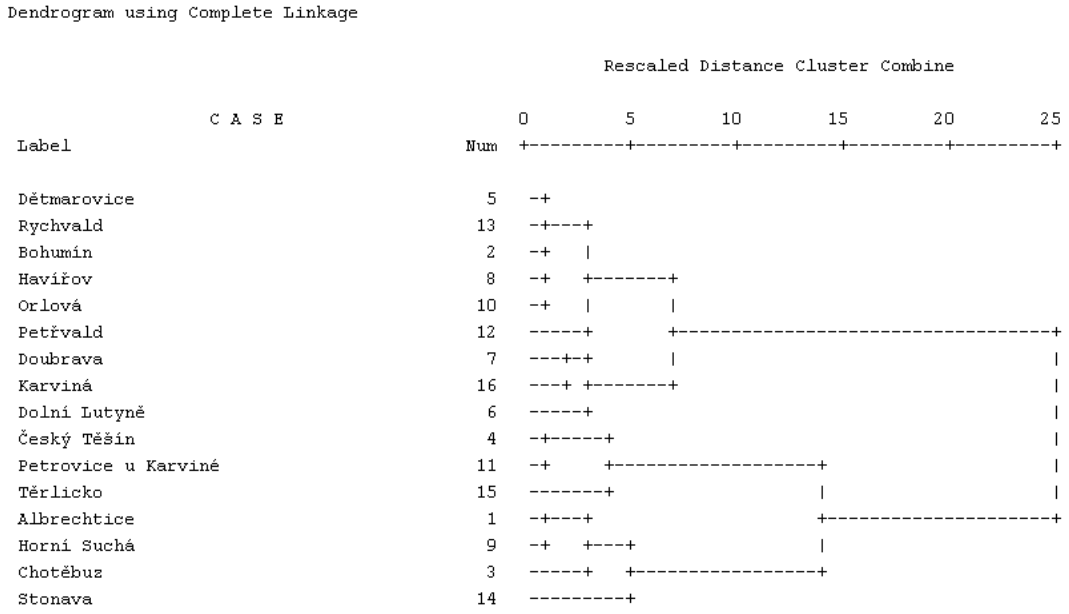
Pokud seskupíme klasifikace podle prvních 2 desetinných míst, dostaneme celkem 4 shluky (obr. 3). Je zřejmé, že pouze třídy č. 1 a 4 jsou spojité, třídy 2. a 3. (dle označení v legendě mapy) obsahují obce, které spolu nesousedí, ale mají podobné zastoupení polské národnosti.

Dále byla využita metoda prostorového hierarchického shlukování s podobným nastavením, tj. standardizace hodnot, euklidovská vzdálenost a metoda nejvzdálenějšího souseda. Prostorová podmínka byla stanovena na topologické sousedství typu královna. Tím je zajištěna spojitost vymezených shluků.

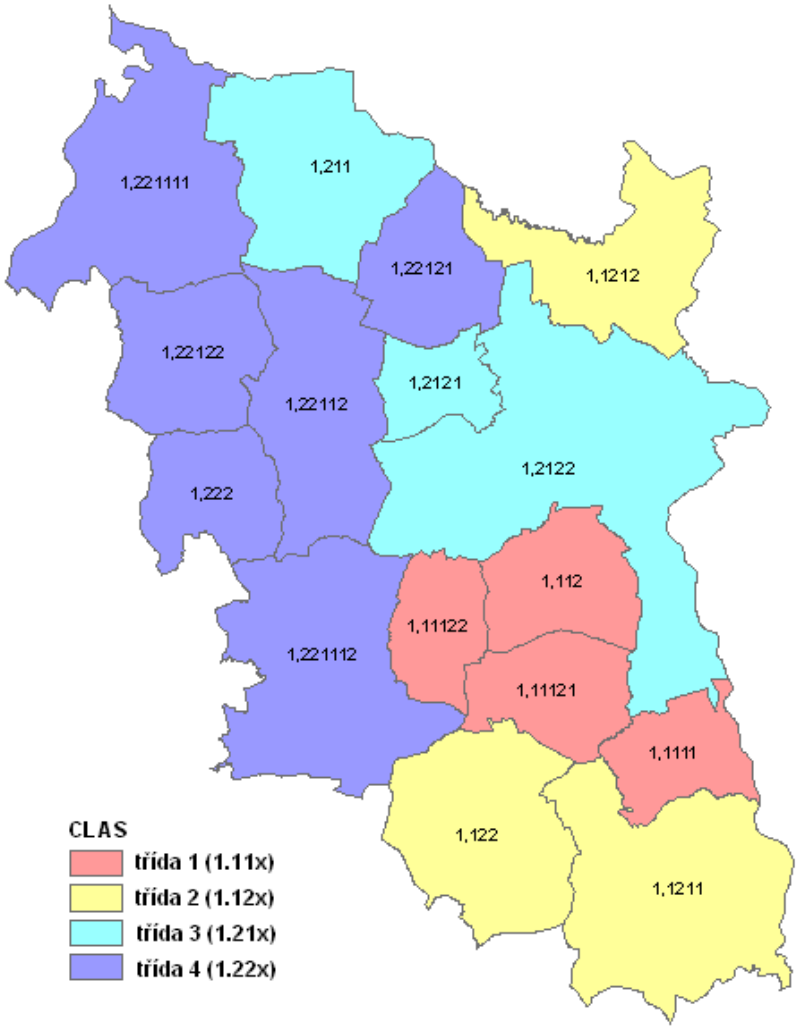
Následně bylo provedeno několik reklasifikací výsledku s cílem ukázat na možnosti vymezení různého počtu shluků a sledování vazeb v území.



Obr. 1. Podíl polské národnosti v okrese Karviná (podíly jedné).



Obr. 2. Dendrogram pro hierarchické shlukování bez prostorového omezení



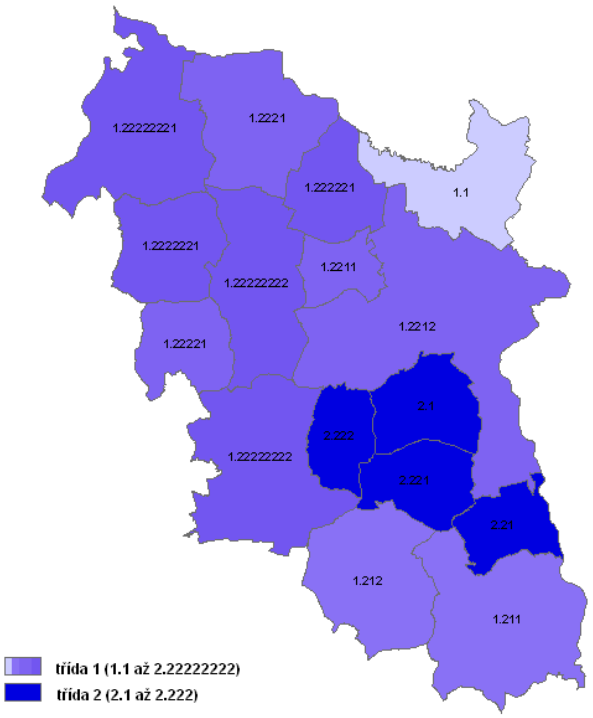
Obr. 3. Výsledek klasického hierarchického shlukování (CL, Euklid.v., standardizace).

Obrázek 4 ukazuje výsledek reklasifikace po provedení řezu těsně u kořene. Celý strom se rozpadá na 2 shluky, kdy 2. třídu tvoří 4 menší sídla mezi Karvinou a Českým Těšínem. Všechny ostatní obce lze klasifikovat jako součást druhé větve. Vnitřní rozdíly jsou vyjádřeny opět číselným kódem a sytostí modré. Nejvíce se odchylují Petrovice u Karviné.

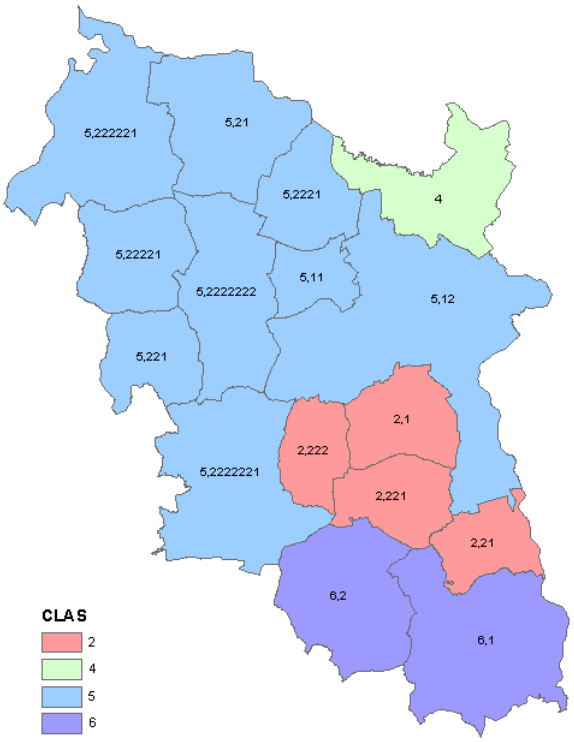
Dále byl proveden řez na úrovni 10% rozdílu průměrných hodnot skupin (obr. 5). Vznikají 4 shluky s různým počtem členů.

Následující obrázek (obr. 6) přidává informaci o lexikální vzdálenosti jednotlivých obcí. Hodnota vzdálenosti je vyjádřena stupněm šedé. Z obrázku je zřejmé, které obce jsou si nejpodobnější (jasnější barvy) a které se nejvíce odchylují (výrazně šedé).

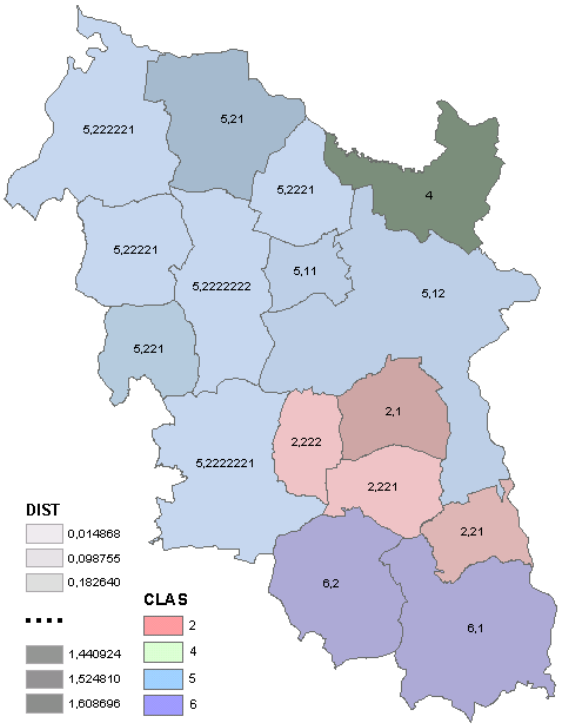
Nakonec byla provedena i reklasifikace hierarchie na základě řezu na úrovni 5% rozdílu průměrného zastoupení polské národnosti (obr. 7). Vzniká 5 shluků.



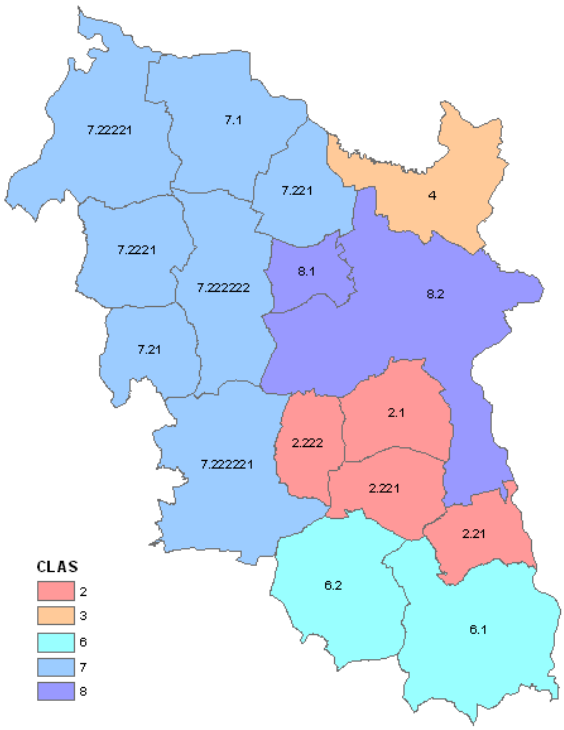
Obr. 4. Dva shluky pro podíl polské národnosti v obcích okresu Karviná (geografická blízkost na základě sousedství typu královna, euklidovské lexikální vzdálenosti, shlukování metodou nejvzdálenějšího souseda, řez dendrogramu těsně u kořene), zvýraznění hierarchie sdružování



Obr. 6. Čtyři shluky pro podíl polské národnosti v obcích okresu Karviná (geografická blízkost na základě sousedství typu královna, euklidovské lexikální vzdálenosti, shlukování metodou nejvzdálenějšího souseda, řez dendrogramu pro rozdíly větší než 10%), bez zvýraznění hierarchie sdružování



Obr. 6. Shlukování pro podíl polské národnosti v obcích okresu Karviná, stejné nastavení jako obr. 5, hierarchie zvýrazněna pomocí vzdálenosti párů vyjádřené stupněm šedi



Obr. 7. Pět shluků pro podíl polské národnosti v obcích okresu Karviná (geografická blízkost na základě sousedství typu královna, euklidovské lexikální vzdálenosti, shlukování metodou nejvzdálenějšího souseda, řez dendrogramu pro rozdíly větší než 5%), bez zvýraznění hierarchie sdružování

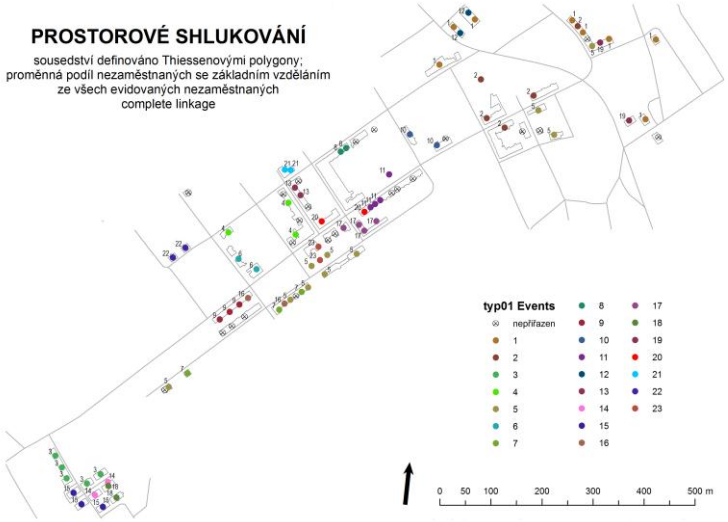
Vedle tohoto základního nastavení byly testovány i jiné parametry volby lexikální vzdálenosti (Manhattonská a čtverec euklidovské vzdálenosti) a shlukovací metody (nejbližší soused a průměrné sousedství).

4. PROSTOROVÉ HIERARCHICKÉ SHLUKOVÁNÍ ADRESNÍCH BODŮ

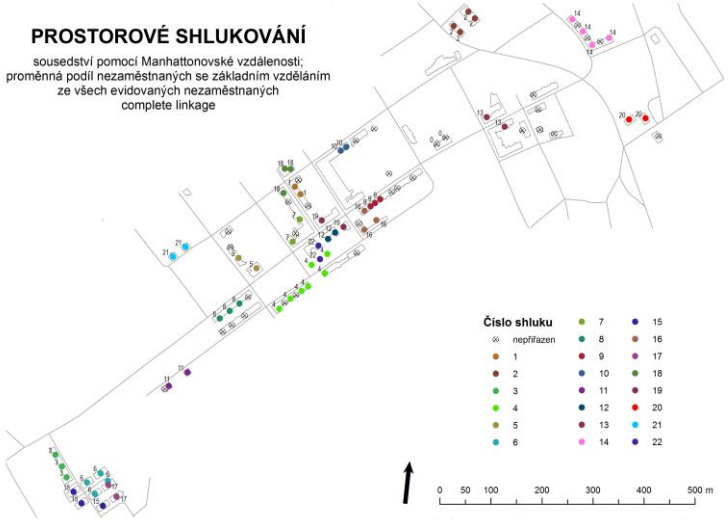
Druhou testovací sadu tvoří adresní body v části Přívozu (Ostrava). Porovnávaným znakem byl podíl nezaměstnaných s nízkým vzděláním k 30. 11. 2010. Pro rozlišení geograficky blízkých adresních bodů byly použity Thiessenovy polygony jako analogie topologického sousedství a výpočet euklidovské vzdálenosti s limitem 50 m (všechny body do 50 m byly považovány za sousedy).

Jednotlivé barvy a čísla ukazují příslušnost k primárním shlukům (jádra). V obou případech byla použita hraniční hodnota vzdálenosti 15%. Thissenovy polygony v tomto omezeném prostoru s nerovnoměrnou distribucí adresních bodů neposkytují ideální výsledky, pokud jde o sousedství bodů – často se vytváří třískovité polygony, které ve výsledku umožní „sousedit“ i adresním bodů značně vzdáleným či odděleným jinou skupinou bodů.

Lepší výsledky poskytuje využití euklidovské vzdálenosti bodů. Nastavený limit 50 m odpovídá představě nejbližšího sousedství bodů. Výhodou je, že nemusíme řešit komplexní prostorové vazby dané konfigurací příslušné části prostoru, silně ovlivněné individuální rozmístěním budov a jejich vchodů. Zvolený přístup považujeme za vhodný pro praktické využití.



Obr. 8. Shlukování v Přívoze s omezením na Thiessenovy polygony (geografická blízkost na základě Thiessenových polygonů, znak – podíl uchazečů s nízkým vzděláním k 30. 11. 2010, euklidovské lexikální vzdálenosti, shlukování metodou nejvzdálenějšího souseda) (Horák et al. 2011)



Obr. 9. Shlukování v Přívoze s omezením do 50 m (geografická blízkost určena euklidovskou vzdáleností do 50m, znak – podíl uchazečů s nízkým vzděláním k 30. 11. 2010, euklidovské lexikální vzdálenosti, shlukování metodou nejvzdálenějšího souseda) (Horák et al. 2011)

5. ZÁVĚR

Metoda prostorového hierarchického shlukování umožňuje klasifikovat objekty podle míry podobnosti jejich znaků a současně respektovat požadovaná prostorová omezení. Ukazují i zajímavé možnosti volit klasifikaci podle požadovaných shluků či podle nastavení hraničních rozdílů (či podobnosti) mezi skupinami.

Vyvíjená aplikace umožní používat různé varianty prostorových omezení či preferencí a lépe tak řídit požadované výsledky klasifikace.

6. LITERATURA

Carvalho, A. X. Y., Albuquerque, P. H. M., Almeida Junior, G. R. de Guimarães, Dantas, R. (2009) Clusterização espacial hierárquica. São Paulo. roč. 27. č .3. 412 – 443 s.

Horák J. (2011) Prostorová analýza dat. 127 stran. Ostrava. 3. vydání

Horák J., Inspektor T., Soukup P., Ivan I. (2011) Methods of Spatial Clustering in a City. Referát Brno, Geografie a geoinformatika - výzva pro praxi a vzdělávání, 8 – 9. 9. 2011. 12 stran.

Hendl J. (2006) Přehled statistických metod zpracování dat: analýza a metaanalýza dat. 583 s. ISBN 80-7367-123-9

Meloun, M., Militký, J., Hill, M. (2005) Počítačová analýza vícerozměrných dat v příkladech. 1. vyd. Praha: Nakladatelství Akademie věd České republiky. 449 s. ISBN 80-200-1335-0.

Lukasová, A., Šarmanová, M. (1985) Metody shlukové analýzy. Praha: SNTL.