Překryvné analýzy rastrových dat typu využití a pokryvu území

Markéta Hanzlová[1] , Jiří Horák[2], Lena Halounová[3], Dušan Židek[4], Jakub Heller[5]

[1]Institut geoinformatiky, Hornicko-geologická fakulta, VŠB-TU Ostrava, 17.listopadu 15, 70833, Ostrava-Poruba, Česká republika
marketa.hanzlova@vsb.cz

[2]Institut geoinformatiky, Hornicko-geologická fakulta, VŠB-TU Ostrava, 17.listopadu 15, 70833, Ostrava-Poruba, Česká republika
jiri.horak@vsb.cz

[3]Katedra mapování a kartografie, Fakulta stavební, ČVÚT, Thákurova 7, 16629, Praha 6, Česká republika
lena.halounova@fsv.cvut.cz

[4]ČHMÚ, pobočka Ostrava, K myslivně 3, 70800, Ostrava-Poruba, Česká republika
zidek@chmi.cz

[5]Cross Czech a.s., Václavské nám. 808/66, Praha 1, Česká republika
jakub.heller@crossczech.cz

Abstrakt. Příspěvek seznamuje s možnostmi využití koeficientu plošné korespondence (Coefficient of Areal Correspondence) a Chi-kvadrát metody pro hodnocení překryvu rastrových dat. Metody byly aplikovány v rámci srovnávací analýzy CORINE LC dat a klasifikovaných dat LANDSAT ETM+.

Klíčová slova: využití a pokryv krajiny, CORINE, LANDSAT, překryvné analýzy.

Abstract. The contribution meets the possibility of using Coefficient of Areal Correspondence and Chi-square methods for raster data overlay assessment. The methods were applied to CORINE LC and classified LANDSAT ETM+ data in the frame of the cross-reference (comparative) analysis.

Keywords: land use and land cover, CORINE, LANDSAT, overlay analysis.

1 Úvod

Hodnocení využití a pokryvu krajiny vychází z provedené klasifikace ať již mapových podkladů nebo dat DPZ. Výsledek klasifikace by měl být verifikován vůči skutečné situaci. K vlastnímu vyhodnocení přesnosti se používá výpočet proporcionální chyby, resp. přesnosti, kappa index [4], případně další metody, jejichž základem je zkoumání shody mezi 2 výběrovými vzorky, kde jeden je zjištěn z terénu (a musí být zjištěn nezávisle) a druhý je výsledkem klasifikace na stejných místech.
Jednotlivé klasifikační postupy se od sebe liší - počínaje zvoleným klasifikačním klíčem, metodou určování příslušnosti jednotlivých objektů či jevů na zemském

povrchu do těchto tříd, konkrétní implementací zvolené metody v různých, zpravidla programových prostředích.
Vedle zkoumání přesnosti klasifikace se proto často můžeme ptát, jaké jsou dopady rozdílné klasifikace (jako projev různého klíče, metody, postupu atd.)? Má nakonec význam použít složitější a náročnější metodu, liší-li se výsledek v území jen minimálně a s ohledem na aplikaci nemá prakticky žádný vliv? Jak takové rozdíly ve výsledku kvantifikovat?
Je evidentní, že ke zpracování 2 klasifikovaných vrstev lze dobře použít překryvných analýz [3]. Naším záměrem je posoudit, jaké konkrétní metody pro hodnocení překryvu volit a jak provést vyhodnocení rozdílů v provedené klasifikaci území na daném příkladě.

1.1 Popis dat

První soubor představuje CORINE Land Cover 2000 data (dále jen CLC), druhým zdrojem dat jsou klasifikované LANDSAT ETM+ snímky (dále jen ETM), jako výsledky per-pixel klasifikace. Podkladem pro vytvoření CLC byl také uvedený snímek ETM, proto rozdíly by měly být dány jen způsobem zpracování, nikoliv rozdílným stářím dat. Porovnáván byl stejný výřez dat nad oblastí povodí Bělé v Jeseníkách (vymezení viz např. [5]).
CORINE Land Cover data jsou vektorového charakteru; nomenklatura rozlišuje 44 tříd, které jsou seskupené do tříúrovňové hierarchie.
Kategorie hlavní úrovně (úroveň 1) podle [2] jsou:

(kód 1) - urbanizovaná území,
(kód 2) - zemědělské plochy,
(kód 3) - lesy a polopřírodní oblasti,
(kód 4) - humidní území,
(kód 5) - vodní plochy.

LANDSAT ETM+ snímky byly klasifikovány podle klasifikačního schématu navrženého v rámci grantového projektu GA 205/06/1037 „Využití geoinformačních technologií pro zpřesňování srážko-odtokových vztahů“. Schéma má 4 úrovně.
První úroveň obsahuje 4 třídy:

- Urbanizovaná území,
- Zemědělské plochy,
- Lesní a polopřírodní oblasti,
- Vodní plochy

které kopírují názvosloví CLC první úrovně. Třídy první úrovně lze využít pro globální mapování, kde lze využít družicových dat s nízkým a středním prostorovým rozlišením.
Druhá úroveň zahrnuje 8 tříd, které lze detekovat v družicových datech s vysokým prostorovým rozlišením (např. LANDSAT, ASTER), stejný zdroj dat lze doporučit také pro třetí úroveň (17 tříd), ale u některých tříd je třeba dalších podpůrných dat. Výše zmíněné kategorie družicových dat představují data multispektrální a

hyperspektrální, kde jejich klasifikace je založena především na spektrálních charakteristikách objektů na zemském povrchu.
Čtvrtá uroveň obsahuje 39 tříd, pro jejichž určení je třeba využít družicová data s velmi vysokým prostorovým rozlišením (více v příspěvku [5]).

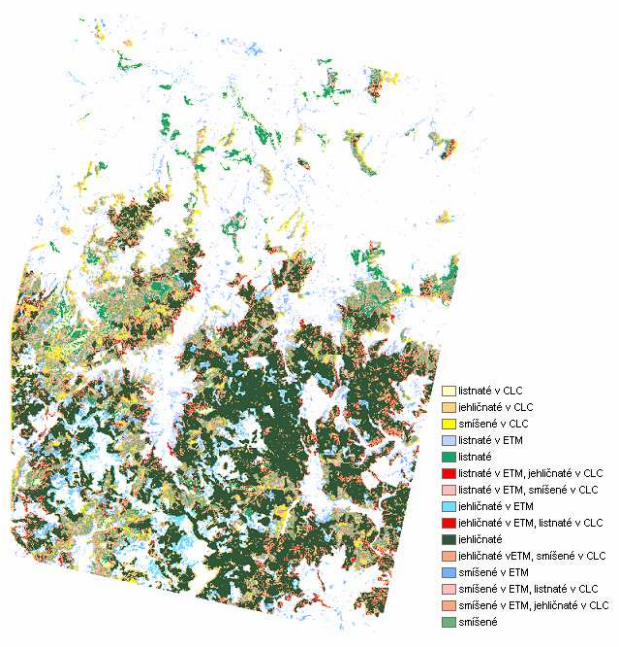
2 Metody řešení

Překryvné analýzy zahrnují operace nad jednou či více vrstvami rastrových či vektorových dat. Tyto analýzy můžeme rozdělit na překryvy aritmetické (součet, rozdíl, dělení, násobení) a překryvy logické (definování určité oblasti splňující soubor podmínek) [1]. Hodnoty v jedné datové vrstvě jsou porovnávány s hodnotami jiné datové vrstvy ve stejném místě (data jsou ve stejném souřadnicovém systému).
Vyžadujeme, aby vrstvy u rastrového modelu měly stejný rozměr buněk (stejné prostorové rozlišení) a stejný počátek a orientaci souřadnicového systému; potom nedochází k částečnému překryvu buněk.
U vektorového modelu dochází k vytváření nových polygonů vzhledem k tomu, že geoprvky v porovnávaných vrstvách nekoincidují stejně jako u rastru.
Příkladem implementace překryvných operací v prostředí ArcGIS jsou:

- funkce sjednocení UNION (výsledná vrstva obsahuje všechny kombinace geoprvků vstupních vrstev),
- funkce průniku INTERSECT (výsledná vrstva obsahuje jen kombinace překrývajících se geoprvků či částí geoprvků v obou vrstvách),
- funkce IDENTIFY (z kombinací vybere ty ležící uvnitř polygonu první vrstvy).

Překryvné analýzy prováděné nad vektorovým a rastrovým datovým modelem lze kombinovat s využitím konverze rastr-vektor a vektor-rastr.

Výsledek překryvné operace UNION pro klasikace v daném území je na obr.1.



Obr. 1. Porovnání dat pokryvu krajiny CLC a ETM - třídy lesa (jehličnaté, listnaté, smíšené)

Při překryvných operacích nás často zajímá, jaká je plocha společného překryvu a jak objekty v první vrstvě korespondují s objekty ve druhé vrstvě.
Ke kvantifikaci korespondence se běžně používají kontingenční tabulky, které ukazují zastoupení jednotlivých kombinací objektů při překryvu (viz tab.1).
Vedle této základní možnosti se nabízí využití dalších metod, jako je koeficient plošné korespondence CAC (z angl. Coefficient of Areal Correspondence) či využití X^2 (Chí-kvadrát) testu, který je základem různých měr asociace.

2.1 Hodnocení kontingenční tabulky

Kontingenční tabulka (tab. 1) ukazuje velikosti ploch (v tomto případu počty pixelů) pro kombinace jednotlivých tříd, vzniklé při překryvu.

Nejlepší shoda se ukazuje pro klasifikaci jehličnatého lesa, kde rozdíl mezi klasifikacemi dosahuje 15 % vůči ETM (počet odlišných pixelů dle CLC ku počtu pixelů jehličnatého lesa ETM) a 30 % vůči CLC (počet odlišných pixelů dle ETM ku počtu pixelů jehličnatého lesa CLC). Naopak v případě klasifikace listnatého lesa je na řádku ETM hodnota shody až 3. v pořadí, což je ohodnoceno velikostí rozdílu 81%.
Celkový rozdíl mezi klasifikacemi je 45 %, tedy shodu lze odhadnout na 55 %.

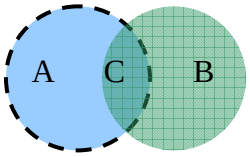
Tabulka 1. Kontingenční tabulka překryvu CLC a ETM ve sledovaném území

	CORINE LC 2000					
	Třída	listnaté	jehličnaté	smíšené	none	Suma
ETM 14.05.2000	listnaté	**48410**	28604	105707	83350	266071
	jehličnaté	1314	**491154**	43081	44928	580477
	smíšené	12343	121958	**111679**	41394	287374
	none	32428	62498	52720	**110021**	257667
	Suma	94495	704214	313187	279693	1391589
Rozdíl vůči	CLC	49%	30%	64%	61%	**45,3%**
	ETM	81%	15%	61%	57%	

2.2 Koeficient plošné korespondence CAC

Koeficient plošné korespondence CAC je založen na překryvné analýze, kde dvě distribuce stejného měřítka jsou porovnány superpozicí dvou vrstev. Jde o jednoduché stanovení rozsahu, kterému si dvě distribuce odpovídají. Metoda počítá s plochami koincidujících si geoprvků. "$Area_A$" představuje plochu nově utvořeného geoprvku z první vrstvy, "$Area_B$" představuje plochu nově utvořeného geoprvku z druhé vrstvy a "$Area_C$" plochu vzniklého průniku obou geoprvků [6].

$$CAC = \frac{Area_C}{Area_A + Area_B + Area_C}$$



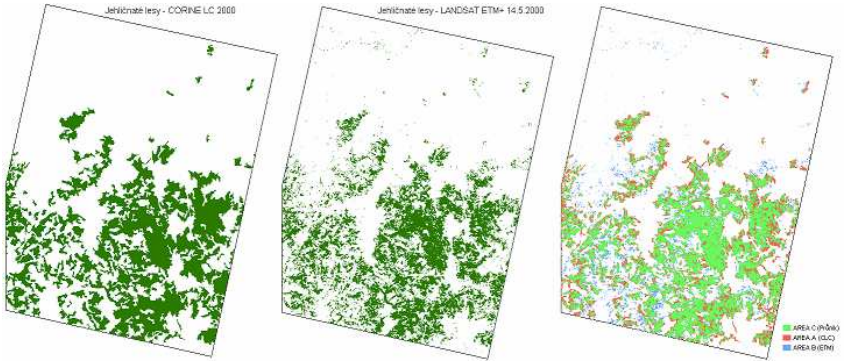
Ve výsledku CAC jednoduchým poměrem ploch popisuje korespondenci geoprvků dvou vrstev. Dva geoprvky spolu nekorespondují pokud CAC = 0, pokud CAC = 1 jde o úplnou korespondenci.
Je vhodné poznamenat, že CAC vlastně představuje plošnou obdobu statistického Russel-Raeova koeficientu asociace [7] pro 2 distribuce nominálních nebo binárních dat S_{RR}.

$$S_{RR} = \frac{d}{a+b+c+d}$$

kde d je počet pozitivních shodných případů pro obě distribuce (jev nastal v obou souborech), a počet případů negativní shody (jev nenastal ani v jednom souboru), b a c potom případy neshody s dvojí možnou polaritou.

Metodu CAC demonstrujeme na příkladu porovnáním tříd jehličnatý, listnatý a smíšený les dat pokryvu krajiny CORINE LC 2000 (CLC) a klasifikovaného snímku LANDSAT ETM+ (ETM). Data pro výpočet CAC byla upravena separováním tříd

jehličnatého (listnatého, smíšeného) lesa do tří vrstev a porovnání každé třídy zvlášť mezi daty pokryvu krajiny.
Využitím překryvné analýzy typu sjednocení (UNION) získáme kombinace tříd jehličnaté (listnaté, smíšené) lesy dvou typů dat pokryvu krajiny CLC a ETM. Vzniknou oblasti pokryvu (jehličnaté, listnaté, smíšené lesy) vyskytujících se jen v CLC (Area A), pokryv vyskytující se jen v ETM (Area B) a pokryv shodný v obou vrstvách, průnik dat (Area C).

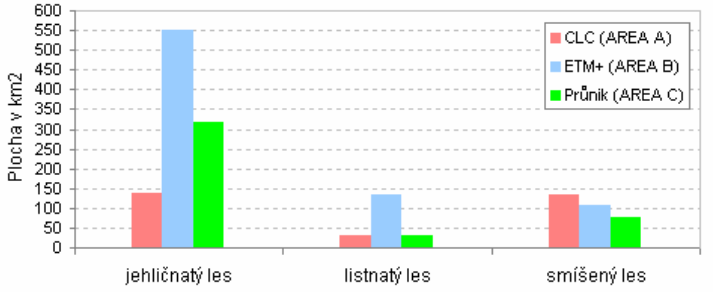


Obr. 2. Data pro výpočet CAC. (Zleva) CORINE LC 2000 – třída 312 jehličnaté lesy; klasifikovaná data LANDSAT ETM+ ze 14.5.2000 - třída jehličnaté lesy; porovnání plochy třídy jehličnaté lesy – data upravená pro výpočet CAC.

Tabulka 2. Koeficient plošné korespondence CAC pro CORINE 2000 a LANDSAT ETM+

Podtřídy lesního pokryvu	**Plocha [km^2]**			
	Nachází se pouze v CLC (AREA A)	**Nachází se pouze v ETM (AREA B)**	**Nachází se v CLC i ETM - průnik (AREA C)**	**CAC**
jehličnatý les	136,52	551,00	316,58	**0,32**
listnatý les	29,03	135,58	30,66	**0,16**
smíšený les	135,52	106,47	77,02	**0,24**

Zastoupení jednotlivých ploch je znázorněno na obr. 3.



Obr. 3. Porovnání ploch jednotlivých tříd

Je evidentní, že největší shoda mezi vrstvami (klasifikacemi) je ve třídě jehličnatý les, kde je plocha shodná v obou vrstvách a tvoří 32 %. Smíšený les a listnatý les vykazují menší shodu v překryvu (24 %, resp. 16 %). Rozdíly mezi vrstvami (klasifikací Corine LC a upravenou klasifikací ETM+) jsou tedy dostatečně zřetelné.

2.3 Chí-kvadrát test

X^2 (Chí-kvadrát) test se používá obecně pro prověření hypotéz shody zjištěných hodnot s předpokládanou distribucí. Lze ho ale uplatnit pro prověření existence asociace mezi 2 distribucemi. Z něho vycházejí některé míry asociace, které se dobře hodí pro sledování velikosti vazeb (míry asociace) mezi nominálními daty.
Vysvětlení aplikace Chí-kvadrát testování lze nalézt v řadě statistických textů, srozumitelné vysvětlení lze nalézt např. v [8].

Stejná data byla použita k testování a výpočtům. Byla vytvořena sada pozorování složená z dvojic, kde 1.člen dvojice představoval výsledek klasifikace ETM a druhý CLC.
Ke zpracování byl použit program SPSS. Protože program není vhodný pro zpracování příliš rozsáhlých dat, provedli jsme redukci počtu případů a nevkládali dvojice odpovídající jednotlivým pixelům, ale každé stovce pixelů (redukce z 1,391 milionu případů na 13915).

Tabulka 3. Upravená kontingenční tabulka v programu SPSS

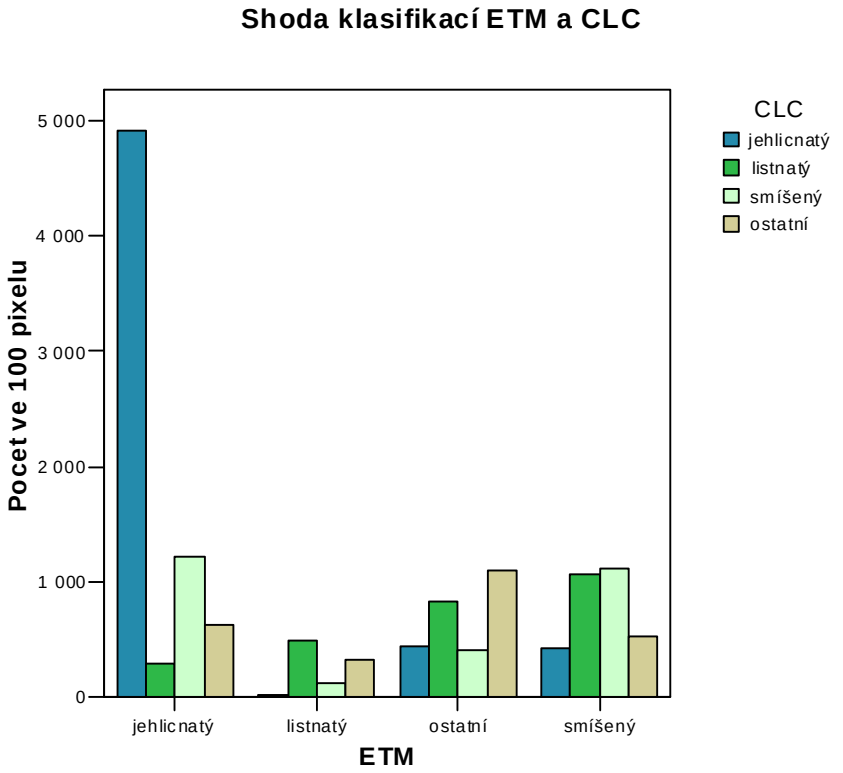
		CLC				Total
		jehličnatý	listnatý	ostatní	smíšený	
ETM	jehličnatý	4911	286	625	1220	7042
	listnatý	13	484	324	123	944
	ostatní	449	834	1100	414	2797
	smíšený	431	1057	527	1117	3132
Total		5804	2661	2576	2874	13915

Tabulka 4. Procentuální zastoupení vzhledem k ETM (SPSS)

		CLC				Total
		jehličnatý	listnatý	ostatní	smíšený	
ETM	jehličnatý	69.7%	4.1%	8.9%	17.3%	100.0%
	listnatý	1.4%	51.3%	34.3%	13.0%	100.0%
	ostatní	16.1%	29.8%	39.3%	14.8%	100.0%
	smíšený	13.8%	33.7%	16.8%	35.7%	100.0%
Total		41.7%	19.1%	18.5%	20.7%	100.0%

Tabulka 4 ukazuje procentuální zastoupení jednotlivých tříd CLC v rámci třídy pokryvu ETM. Opět je dobře vidět nejvyšší shodu v případě jehličnatého lesa. Jde

ovšem o pouze jednostranné posouzení (zastoupení CLC v rámci 1 třídy ETM). Výsledky je možné dokumentovat i v přehledném grafu na obr. 4.



Obr. 4. Srovnání zastoupení jednotlivých tříd pokryvu CLC a ETM při překryvu

Tabulka 5. Výsledek chí-kvadrát testu (SPSS)

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	6178.252	9	.000
Likelihood Ratio	6597.003	9	.000
N of Valid Cases	13915		

Výsledná hodnota chí-kvadrát testu je 6178 a to při daném počtu stupňů volnosti (9) potvrzuje nenáhodnost vazby mezi oběma vrstvami na nejvyšší hladině významnosti. Tato nenáhodnost je ovlivněna vysokým počtem pozorování (13915 pozorování). To je pochopitelně očekávaný výsledek a pouze nepotvrzení hypotézy by naznačovalo

zcela jiné poměry ve vrstvách. Tato hodnota však nic neříká o míře asociace. K tomu slouží až navazující testy.

Tabulka 6. Směrové míry asociace (SPSS)

			Value	Asymp. Std. Error(a)	Approx. T(b)	Approx. Sig.
Nominal by Nominal	Lambda	Symmetric	.204	.005	33.743	.000
		ETM Dependent	.181	.007	22.916	.000
		CLC Dependent	.223	.007	31.195	.000
	Goodman and Kruskal tau	ETM Dependent	.212	.004		.000
		CLC Dependent	.181	.004		.000
	Uncertainty Coefficient	Symmetric	.189	.004	47.596	.000
		ETM Dependent	.200	.004	47.596	.000
		CLC Dependent	.180	.004	47.596	.000

a Not assuming the null hypothesis.
b Using the asymptotic standard error assuming the null hypothesis.

Výsledky testů v tabulce 6 se pohybují kolem hodnoty 0.2, což lze interpretovat tak, že při znalosti 1 proměnné lze snížit asi o 20 % chybu předpovědi druhé proměnné. Obě klasifikace tedy není možné považovat za dostatečně shodné.

Tabulka 7. Symetrické míry asociace (SPSS)

		Value	Approx. Sig.
Nominal by Nominal	Phi	.666	.000
	Cramer's V	.385	.000
	Contingency Coefficient	.555	.000
N of Valid Cases		13915	

Phi test se nemůže použít, protože je platný pouze pro kontingenční tabulky s rozměry 2x2 buňky. Další symetrické míry asociace (Kramerův a kontingenční koeficient) vykazují významnou (viz významnost v posledním sloupci tabulky 7), ale spíše střední míru asociace. Kontingenční koeficient vychází podobně jako míra shody v kontingenční tabulce č.1. Obě jsou hodnoceny v intervalu 0 až 1, lépe interpretovatelná je zřejmě Kramerova míra.

3 Závěr

Pro zpracování nominálních dat při překryvných operacích a jejich kvantifikaci je možné použít řady metod, v příspěvku jsme dokumentovali vybrané z nich. Klasická kontingenční tabulka nám poskytuje podrobný přehled o velikosti překryvu mezi jednotlivými třídami. Není však jednoduché ji snadno interpretovat, proto je vhodné údaje dále zpracovat jinými metodami.
Koeficient plošné korespondence CAC představuje jednoduchý nástroj jak z hlediska provádění analýzy, tak především pro její interpretaci. Jeho nevýhodou je izolovanost posuzování jednotlivých tříd – změna v jedné třídě se neprojeví pouze v hodnocení plošné korespondence dané třídy, ale ovlivní i výsledky hodnocení plošné korespondence dalších tříd. Přesto ho lze považovat za efektivní nástroj popisu míry překryvu.
Ke statisticky více propracovaným metodám patří různé míry asociace spojené s chí-kvadrát testováním. Ty dovolují prověřit statistickou významnost nalezených vztahů i hodnotit míru shody. K doporučovaným mírám patří koeficient neurčitosti a Kramerova míra asociace.
Metody byly vyzkoušeny na datech CORINE Land Cover 2000 a klasifikaci LANDSAT ETM+, použitou v příspěvku [5] pro povodí Bělé v Jeseníkách s cílem posoudit a kvantifikovat rozdíly (resp. shodu) mezi oběma klasifikacemi. Dále budou zkoumány dopady rozdílů v klasifikaci na retenční schopnost krajiny a srážkoodtokové modelování.

4 Poděkování

Příspěvek vznikl na základě finanční podpory Grantové agentury České republiky v rámci projektu GA 205/06/1037 „Využití geoinformačních technologií pro zpřesňování srážko-odtokových vztahů“.

Reference

1. Aronoff, S. (1989): *Geographical information systems – A perspective management.* WDL Publications, Ottawa 1989.
2. Bossard, M., Feranec, J., Otahel, J. (2000): *Definice tříd CLC.* (připraveno jako metodická pomůcka pro řešení projektu VaV/250/1/01 - Aktualizace databáze CORINE Land Cover České republiky (I&CLC2000). Český překlad – M. Koželuh, EEA 2000
3. Burrough P., McDonnell A. (1998): *Principles of Geographical Information Systems*. Oxford University Press 1998.
4. Dobrovolný, P.: *Dálkový průzkum Země. Digitální zpracování obrazu.* Skripta, Brno 1998.

5. Hanzlová, M., Horák, J., Unucka, J., Halounová, L., Žídek, D., Heller, J. (2007): Klasifikace pokryvu území a jeho dopady na hodnocení srážko-odtokových poměrů. *Symposium GIS Ostrava 2007, 14.ročník*. Ostrava 2007.
6. Lembo, A.J., Jr.: Spatial Correspondence of Areal Distributions. http://www.css.cornell.edu/courses/620/lecture12.ppt
7. Meloun, M., Militký, J. (2002): *Kompendium statistického zpracování dat*. Praha, Academia, 2002, 764 s., ISBN 80-200-1008-4
8. Swoboda, H. (1977)*: Moderní statistika.* Nakladatelství Svoboda, Praha 1977