

Univerzita Palackého v Olomouci

Přírodovědecká fakulta

Katedra geoinformatiky

Bc. Michael HAVLÍK

**ČASOPROSTOROVÁ ANALÝZA VYBRANÝCH
ONEMOCNĚNÍ V OLOMOUCKÉM KRAJI**

Magisterská práce

Vedoucí práce: Ing. Zdena DOBEŠOVÁ, Ph.D.

Olomouc 2013

Čestné prohlášení

Prohlašuji, že jsem magisterskou práci magisterského studia oboru Geoinformatika vypracovala samostatně pod vedením Ing. Zdeny Dobešové Ph.D.

Všechny použité materiály a zdroje jsou citovány s ohledem na vědeckou etiku, autorská práva a zákony na ochranu duševního vlastnictví.

Všechna poskytnutá i vytvořená digitální data nebudu bez souhlasu školy poskytovat.

V Olomouci 18. 4 2013

podpis_____

Děkuji vedoucímu práce Ing. Zdeně Dobešové Ph.D. za podněty a připomínky při vypracování práce a konzultantu Mgr. Lukášovi Markovi za ochotné poskytnutí cenných rad.

Za poskytnutá data děkuji Krajské hygienické stanici v Olomouci.

Vložený originál **zadání** bakalářské/magisterské práce (s podpisy vedoucího katedry, vedoucího práce a razítkem katedry). Ve druhém výtisku práce je vevázána fotokopie zadání.

OBSAH

SEZNAM POUŽITÝCH ZKRATEK	6
ÚVOD	7
1 CÍLE PRÁCE.....	8
2 METODY A POSTUPY ZPRACOVÁNÍ	9
2.1 Postup zpracování	9
2.2 Použitá data	10
2.3 Použité programy	12
2.4 Výběr metody pro shlukovou analýzu	13
2.4.1 Metoda Hot Spot analýzy (Getis-Ord G_i^*).....	14
3 SOUČASNÝ STAV ŘEŠENÉ PROBLEMATIKY	16
3.1 Zahraniční studie poslední doby	17
3.2 Domácí studie poslední doby	19
3.3 Zhodnocení.....	20
4 VLASTNÍ ŘEŠENÍ	21
4.1 Oprava databáze EPIDAT.....	21
4.1.1 Stručný popis programu.....	22
4.1.2 Aktualizace dat	24
4.2 Popisná charakteristika vstupních dat	26
4.2.1 Analýza dat EPIDAT.....	26
4.2.2 Hepatitida A.....	30
4.3 Předzpracování dat	33
4.4 Testování náhodnosti rozmístění bodů výskytu.....	38
4.4.1 Metoda průměrného nejbližšího souseda.....	38
4.4.2 Ripleyova K-funkce.....	40
4.5 Shlukové analýzy	42
4.6 Vizualizace výstupů	44
4.7 Zhodnocení výsledků shlukových analýz	47
5 VÝSLEDKY	50
6 DISKUSE.....	52
7 ZÁVĚR	54
POUŽITÁ LITERATURA A INFORMAČNÍ ZDROJE	
SUMMARY	
PŘÍLOHY	

SEZNAM POUŽITÝCH ZKRATEK

Zkratka	Význam
EPIDAT	Epidemiologická databáze infekcí
KML	Keyhole Markup Language
RUIAN	Registr územní identifikace, adres a nemovitostí
ČSU	Český statistický úřad
ČUZK	Český úřad zeměměřický a katastrální
LISA	Local indicators of spatial association
KGI	Katedra geoinformatiky
ORP	Obec s rozšířenou působností
UP	Univerzita Palackého v Olomouci
SHP	Shapefile
HTML	HyperText Markup Language
GIS	Geografické informační systémy
XLS	MS Excel worksheet
k.ú.	Katastrální území
UP	Univerzita Palackého v Olomouci
KMZ	Keyhole Markup Language Zip file

ÚVOD

V dnešní době existuje mnoho forem ohrožení zdraví. Lidé na celém světě vymýšlí způsoby, jak tyto stavy vyléčit a předcházet jim. Jedním ze stavů, který dokáže narušit vnitřní prostředí lidského organismu je infekční onemocnění.

Míra závažnosti infekce je měřitelná. Každá infekce má jiné vlastnosti, průběh a inkubační dobu. Šíří se časem a prostorem. Zde se dostáváme k propojení s GIS (Geografický informační systém). Jakým směrem se infekce šíří? Má plynulý průběh v jednotlivých letech? Dají se infekce vizualizovat a vytváří shluky? Přesně tyto a další otázky nás jako geoinformatiky zajímají a bude jim věnována tato práce.

Díky KHS (Krajská hygienická stanice v Olomouci) můžeme znázornit přes 46 000 záznamů z EPIDAT (Epidemiologické databáze infekcí) za desetileté období 2002–2011. V mé předchozí bakalářské práci (Havlík, 2010), na kterou by měla tato práce plynule navazovat, byl řešen způsob opravy těchto dat a následné přiřazení prostorové složky.

V této práci tedy budou následovat procesy identifikace oblastí, testování náhodnosti, shlukování a vizualizace opravených dat. Stěžejní část bude věnována prostorovým analýzám, zejména procesu shlukování.

Jedná se o analýzu vícerozměrných dat k 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. Vyhledání a použití vhodné shlukovací metody bude popsáno v kapitolách níže.

Cílovým uživatelem budou především hygienici a epidemiologové, kteří si na základě výsledků této práce budou moci jasněji představit data, se kterými jsou v každodenním styku.

1 CÍLE PRÁCE

Cílem této diplomové práce je identifikace oblastí se zvýšeným výskytem vybraných onemocnění pomocí prostorových analýz.

V první části budou kompletně opravená data za rok 2002–2011 prostřednictvím poloautomatické opravy v databázovém programu MS Access. Zde bude využito programu vytvořeného v rámci předchozí bakalářské práce, na níž tato práce navazuje.

Práce bude obsahovat popisnou charakteristiku vstupních dat, kde bude mimo jiné provedena i statistická analýza infekcí. Na základě četnosti infekce a nepravidelných změn v jednotlivých letech bude podrobněji zkoumána vybraná infekce.

Při testování náhodnosti rozmístění bodů výskytu bude nad daty provedena statistická analýza pomocí metod nejbližšího souseda a K-funkce.

Hlavním cílem bude identifikace shlukování a případných shluků. Zde byla vybrána shlukovací metoda, která dokáže vhodně klasifikovat získaná data do tříd. Shlukování bylo provedeno jak na úrovni ulic. Zkoumaným územím byla ORP (obec s rozšířenou působností) Olomouc. U vybraných infekcí budou data rovněž zpracována s ohledem na různé skupiny pacientů (podle věku, zaměstnání, pohlaví apod.) Všechna data budou brána s ohledem na časový aspekt výskytu.

Výstupy budou prezentovány ve formě map, tabulek, grafů a souborů KML (Keyhole Markup Language). Tyto soubory budou vizualizovány prostřednictvím aplikace Google Earth.

Celá práce včetně všech výstupů bude odevzdána v digitální podobě na DVD a dále budou také vytvořeny webové stránky práce.

2 METODY A POSTUPY ZPRACOVÁNÍ

2.1 Postup zpracování

Postup diplomové práce je znázorněn vývojovým diagramem (viz obr. 1). Prvním krokem bylo prostudování literatury a seznámení se s problematikou shlukových analýz v geoinformaticce.

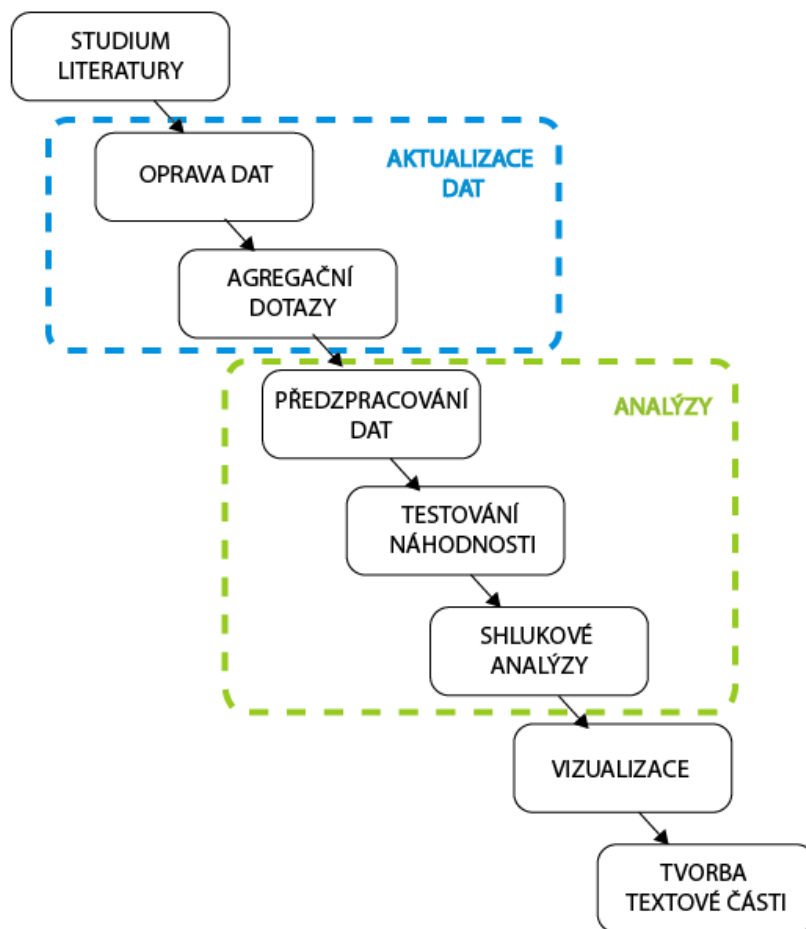
Následovala aktualizace dat databáze EPIDAT. Za pomoci programu pro poloautomatickou opravu byla opravena data za období 2002–2003 a 2009–2011. Výsledkem práce s přičtením dat opravených v rámci bakalářské práce byl ucelený soubor dat za desetileté období 2002–2011 v rozsahu přes 46 000 záznamů pro Olomoucký kraj. Výstupem z databáze byl soubor agregovaných tabulek na úroveň ulic a katastrálního území ORP Olomouc. Současně byla také vytvořena popisná charakteristika vstupních dat obsahující detailní výčet zastoupených infekcí v souboru dat a rozbor vybrané infekce.

V části předzpracování dat byla datům z databázového programu MS Access přidělena prostorová složka v programu ArcGIS od firmy ESRI. Následovalo vytvoření pravidelné hexagonální mřížky za pomoci extenze Repeat Shapes. Jednotlivé výskyty infekcí byly pomocí nástroje Create Random Points náhodně rozmístěny do plochy katastrálního území a silniční sítě. Prostřednictvím nástroje Identity byly standardizovány počty obyvatel v hexagonální mřížce a byl vypočítán výskyt infekce v hexagonu.

Současně bylo provedeno i testování náhodnosti rozmístění bodu výskytu, kde byla použita metoda nejbližšího souseda a K funkce. Prvně zmiňovaná metoda vypočítává vzdálenost mezi nejbližšími sousedními body. Druhá zkoumá, zda data vykazují statisticky významné shlukování nebo rozptyl.

Výsledné vrstvy vstupovaly do shlukové analýzy. Byla zvolena metoda Hot Spot analýza (Getis-Ord G_i^*), která vhodně reprezentuje lokální data a intenzitu jevu. Tato metoda klade důraz na hledání shluků v prostoru, nikoliv na hledání shluků na základě podobnosti. Tento nástroj se nachází v nástrojové sadě Spatial Statistics.

Vizualizace proběhla formou map a exportu souborů do formátu KML. Tyto soubory byly zobrazeny aplikací Google Earth. Výstupním datům byla přiřazena časová značka za účelem vytvoření časových animací ve zmíněném programu. Závěrem proběhlo slovní zhodnocení mapových výstupů.



Obr. 1: Vývojový diagram postupu zpracování.

2.2 Použitá data

V této části je seznam použitých dat, která vstupovala zejména do praktické části práce. Stěžejní data EPIDAT byla převzata od KHS. Další data byla zapůjčena KGI UP. Data převzatá a zapůjčená jsou detailněji rozepsaná níže.

Data, která vznikla během této práce, byly zejména vytvořené agregační tabulky z EPIDAT ve formátu XLS (MS Excel worksheet), dále výstupní vrstvy ze shlukových analýz ve formátu SHP (shapefile) a v neposlední řadě data ve formátu KML vstupující do programu Google Earth.

EPIDAT

Pro účely této práce byla poskytnuta KHS v Olomouci data z databáze EPIDAT za časové období 2002–2011. Data obsahovala přes 46 tis. záznamů deseti nejfrekventovanějších infekcí v rozsahu pro Olomoucký kraj.

UIR-ADR

Tento registr od 1.7.2012 spadající pod RUIAN (Registr územní identifikace, adres a nemovitostí), kde je nyní aktualizován, byl využit při opravě chybné databáze EPIDAT.

ČSU – demografická data

V souvislosti s analýzou EPIDAT a porovnáním věkových skupin této databáze byla použita data věkového složení obyvatelstva pro Olomoucký kraj.

ČUZK – katastrální území

Katedrou geoinformatiky v Olomouci byla převzata pro tuto práci data katastrálních území pro Olomoucký kraj. Data byla poskytnuta katedře Českým úřadem zeměměřickým a katastrálním.

CEDA – uliční síť

V rámci projektu POHOS realizovaného na Katedře geoinformatiky v Olomouci nám byla zapůjčena data uliční sítě pro oblast ORP Olomouc. Tato data byla vytvořena společností CEDA. Před samotným použitím a propojením s daty z EPIDAT bylo potřeba vrstvu road.shp obsahující úseky komunikací upravit.

ÚIR-ZSJ

Územně identifikační registr základních sídelních jednotek je soustava databázových číselníků jednotek územně správního, technického a sídelního členění státu až do úrovně části obcí, k.ú. (katastrální území) a ZSJ (základní sídelní jednotky) (Prohlížeč ÚIR-ZSJ, 2013). Z tohoto registru byla použita data pro ZSJ v Olomouckém kraji a vypočítány počty obyvatel jednotlivých katastrálních území v ORP Olomouc. Počet obyvatel k.ú. byl použit v předzpracování dat v kapitole 4.3, pro výpočet obyvatel v hexagonové síti.

2.3 Použité programy

Tato podkapitola uvádí výčet programů použitých pro praktickou část diplomové práce. Součástí popisu programu jsou také jejich funkce, s nimiž bylo v této práci pracováno.

MS Access

Nástroj pro správu relačních databází. Je součástí konferenčního balíku MS Office. Této souborové databázi bylo využito k opravě a aktualizaci vstupních dat databáze EPIDAT. Bylo zde užito několik desítek dotazů a formulářů. Pomocí této databáze byl vytvořen program na poloautomatickou opravu dat z databáze EPIDAT za pomoci celostátně platných dat z databáze UIR-ADR. Tento program byl vytvořen v předchozí bakalářské práci a nebude v této práci detailněji popisován (Havlík, 2010)

MS Excel

Tento tabulkový editor z balíku MS Office byl použit pro úpravu vstupních dat z databáze EPIDAT a pro výstupní soubory vstupující do programu ArcGIS. Nejvíce se programu využilo v popisné charakteristice vstupních dat při tvorbě tabulek a grafů (viz kapitola 4.2)

ESRI ArcGIS 10

Tento nejrozšířenější GIS nástroj byl stěžejním programem práce. Sloužil zejména pro předzpracování dat za použití nástrojů z nástrojové sady Data Management a Spatial Statistics. Dále zde byly hojně použity extenze Repeating shapes a Export to KML. Program také sloužil k závěrečné vizualizaci mapových výstupů.

Google Earth

Aplikace vytvářející trojrozměrný model skutečného světa, který kombinuje satelitní snímky, mapy a vrstvy. Uživatel tak může virtuálně navštívit libovolné místo na světě. V naší práci zde byly zobrazeny KML soubory shlukových analýz pro území ORP Olomouc. Vrstvy obsahující časovou značku zde byly znázorněny pomocí časových animací.

2.4 Výběr metody pro shlukovou analýzu

Při výběru vhodné shlukovací metody bylo více než důležité použít metodu, která dokáže vhodně znázornit a interpretovat použitá data této práce. Důraz byl kladen na hledání shluků v prostoru než na hledání shluků na základě podobnosti.

Jako vhodný nástroj byla použita nástrojová sada Spatial Statistics v programu ArcGIS. Tento prostorově statistický nástroj je často využíván v analýze kriminality, demografii, dopravní analýze, archeologii a v neposlední řadě také v epidemiologii. Nástroj Spatial Statistics umožňuje (ESRI, 2012):

- Shrnutí klíčových charakteristik distribuce prvků
- Identifikovat statisticky významné prostorové shluky (hot spot/cold spot) a prostorové outliery (naměřený jev s výrazně odlišnou hodnotou)
- Posoudit zda celkový vzorek vykazuje shlukování či rozptyl
- Modelovat prostorové vztahy

Dále tento nástroj obsahuje dvě metody pro shlukové analýzy v prostoru.

První metodou je LISA neboli lokální indikátor prostorové asociace. Dle Anselin (1995) se jedná o lokální ekvivalent Moranova I kritéria, který poskytuje pro každé pozorování údaj o rozsahu (velikosti) prostorově významných shluků podobných hodnot. Součet všech pozorování je přímo úměrný globálnímu ukazateli prostorové asociace. Ve výsledku ukazuje, zda se vyskytuje shluk prvků nebo prostorový outlier, přičemž se nezabývá velikostí hodnot. Následná kategorizace po vypočtení LISA je rozdělena do čtyř skupin podle typu prostorové autokorelace (viz obr. 2)

vážená hodnota proměnné v blízkých jednotkách	nízká – vysoká <i>negativní prostorová autokorelace</i>	vysoká – vysoká <i>pozitivní prostorová autokorelace</i>
	nízká – nízká <i>pozitivní prostorová autokorelace</i>	vysoká – nízká <i>negativní prostorová autokorelace</i>
	hodnota proměnné v prostorové jednotce	

Obr. 2: Moranův diagram. Pojmy „vysoká“ a „nízká“ znamenají u analýzy LISA nadprůměrnou a podprůměrnou hodnotu (Spurná, 2008).

Druhou metodou nástroje Spatial Statistics je metoda Hot Spot analýzy nazývaná také Getis-Ord G_i^* . Tato metoda vhodně znázorňuje intenzitu jevu. Identifikuje statisticky významné shluky vysokých (hot spots) a nízkých (cold spots) hodnot. Výsledná hodnota je vypočítána z velikosti hodnoty výskytu, a proto je vhodnější pro výslednou vizualizaci. Detailněji je metoda popsána v následující podkapitole.

2.4.1 Metoda Hot Spot analýzy (Getis-Ord G_i^*)

Metoda Hot Spot analýzy (dále jen Getis-Ord G_i^*) byla na základě konzultace s vedoucí práce použita při shlukových analýzách této práce. Obecný princip této metody je znázorněn na Obr. 3 a dále detailně popsán v textu.



Obr. 3: Princip metody Getis-Ord G_i^* (ESRI, 2012).

Vzorec pro výpočet metody je dán vztahem:

$$G_i^* = \frac{\sum_{j=1}^n w_{i,j} x_j - \bar{X} \sum_{j=1}^n w_{i,j}}{S \sqrt{\frac{n \sum_{j=1}^n w_{i,j}^2 - \left(\sum_{j=1}^n w_{i,j} \right)^2}{n-1}}} \quad (1)$$

kde x_j je hodnota atributu prvku j vstupující do analýzy, $w_{i,j}$ prostorová váha mezi prvky i a j , n je hodnota celkového počtu prvků a:

$$\bar{X} = \frac{\sum_{j=1}^n x_j}{n} \quad (2)$$

$$S = \sqrt{\frac{\sum_{j=1}^n x_j^2}{n} - \left(\bar{X}\right)^2} \quad (3)$$

Esri (2012) definuje, že výsledný GI* index je vypočten pro jednotlivé prvky. Hodnota je vyjádřena přímo hodnotou z-score. Pro testování statistické významnosti tedy nejsou potřeba další výpočty. Prvek s vysokou hodnotou však ještě nemusí být nutně statisticky významný prostorový shluk vysokých hodnot. Pro splnění funkce statisticky významného shluku vysokých hodnot musí prvek nejen obsahovat vysokou hodnotu, ale musí být také obklopen vysokými hodnotami sousedních prvků. Lokální součet hodnoty prvku a hodnot jeho okolí je proporcionálně srovnáván se součtem hodnot všech prvků v území. Pokud je lokální součet značně odlišný od očekávaného lokálního součtu, znamená to, že tento rozdíl nemůže vzniknout náhodně a jedná se tedy o statisticky významný výsledek (z-score).

Výstupní hodnoty z-score a p-value jsou vypočteny pro každý prvek a určují statistickou významnost prostorového shlukování. Z-score je test statistické významnosti, který pomáhá rozhodnout, zda přijmout nebo odmítnout nulovou hypotézu. Nulová hypotéza v případě hot Spot analýz nám říká, že hodnoty prvků jsou rozmístěny náhodně a neexistuje prostorové shlukování. P-value je pravděpodobnost, se kterou zavrhneme nulovou hypotézu (Esri, 2012).

Vysoká hodnota z-score a malá hodnota p-value (statisticky významná, tj. < 0,05) pro daný prvek vyjadřuje existenci prostorového shluku vysokých hodnot v okolí (hot spot). Nízká záporná hodnota z-score a malá hodnota p-value pro daný prvek vyjadřuje existenci prostorového shluku nízkých hodnot v okolí (cold spot). Čím je hodnota z-score větší nebo menší, tím je shlukování intenzivnější. Z-score blízké nule znamená, že se zde nevyskytuje žádný zjevný shluk.

U metody Getis-Ord Gi* se obecně doporučuje používat metodu konstantní vzdálenosti. Lze ale použít i metody založené na inverzní vzdálenosti. Vstupním datům musíme definovat souřadnicový systém a to především z důvodu práce s daty, v nichž je nutné definovat jednotky. Analýza pracuje pouze s body (centroidy) polygonů. Vyžaduje variabilitu vstupních hodnot a pracuje pouze s kladnými hodnotami, tudíž vstupní data nesmí obsahovat nulové hodnoty (Esri, 2012).

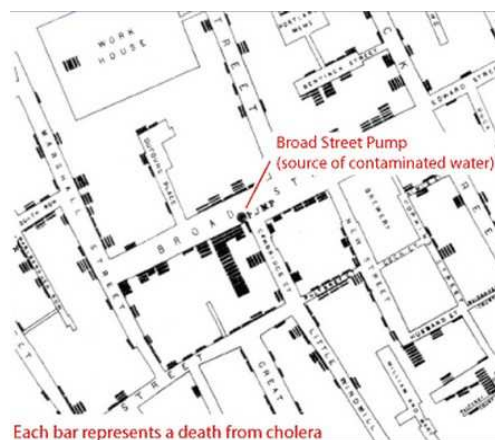
3 SOUČASNÝ STAV ŘEŠENÉ PROBLEMATIKY

Časoprostorová analýza infekčních onemocnění zahrnuje průřez minimálně třemi obvykle oddělenými disciplínami: epidemiologií, geografii a statistikou. Dnešní metody prostorové statistiky často odhalují zákonitosti prostorového výskytu nemocí v souvislosti s potenciálními rizikovými faktory a výskytem nemocí.

V současné době se velmi často používají metody disease mapping (DM), které slouží k vizualizaci a vyhodnocení geografických rozdílů ve výskytu nemoci (Štěpánová, 2008). Tyto metody jsou součástí prostorové epidemiologie (PM), která se zabývá analýzou uspořádání incidence nemoci v prostoru.

PM se dá rozdělit do tří skupin: disease mapping, disease clustering a prostorová regrese. V této práci je kladen důraz na druhou zmiňovanou metodu – disease clustering. Ta je použita při hledání výskytu shluků (Lawson, 2006).

V historii byla jednou z nejslavnějších map řešící výskyt nemoci mapa Johna Snowa z roku 1854 (obr. 4). Mapa znázorňovala výskyt cholery v Londýně. Z mapy bylo patrné, že nejvíce případů bylo seskupeno v okolí pumpy na Broad Street. Jeho mapa byla následně využita úřady a epidemie byla následkem toho zmírněna (Disease mapping and spatial epidemiology, 2008)



Obr. 4: Mapa Johna Snowa se shluky cholery poblíž vodní pumpy (Disease mapping and spatial epidemiology, 2008).

Tato práce klade důraz na lokální statistické metody, které jsou od 90. let hojně využívány především v oblastech zdravotnictví, kriminalistiky, životního prostředí a demografie. Velkou měrou se podíleli na rozvoji těchto statistických metod především L. Anselin (LISA), A. Getis a J. K. Ord (Hot Spot analysis).

V dnešní počítačové době už jsou výsledky shlukových analýz daleko propracovanější díky použití sofistikovaných software. V následující části práce budou představeny řešení konkrétních zahraničních i českých případových studií.

3.1 Zahraniční studie poslední doby

Většinu zahraničních studií vztahujících se k obsahu této práce a reflektující současný stav použití bylo možné získat ze stránek ESRI. Konkrétně v sekcích ArcGIS Resources a ArcUser Online. Dalšími zdroji byly US National Library of Medicine (<http://www.ncbi.nlm.nih.gov/>) a Web of Science.

Prostorové shlukování syndromu akutního respiračního selhání (SARS) v Hongkongu

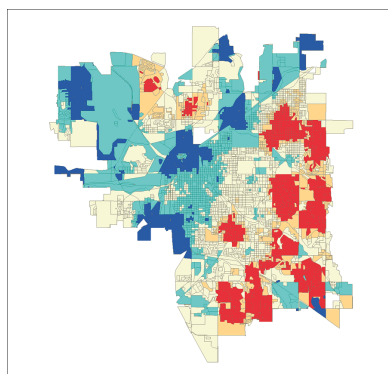
Skupina čínských vědců v roce 2003 použila geostatistické metody při analýze vzorků z šíření choroby SARS. Za pomoci GIS technologií zde byly analyzovány klinické a osobní údaje o 1755 pacientech za období průběhu choroby (15.2.–22.6.2003). Shluky vysokých a nízkých hodnot byly identifikovány pomocí Moranova I kritéria. Podle výsledků analýz byla první fáze (18.2.–6.3.) označena za fázi rostoucích příznaků epidemie a druhá fáze (15.5.–4.6.) jako pokles nákazy (Lai PC a kol. 2004).

Hodnocení rizik při analýze zastávek školních autobusů – Readlans, California

V roce 2005 se vědci z Californské univerzity pokusili provést analýzy, které měly identifikovat problematické zastávky školních autobusů. Město Readlans bylo dlouhodobě známé vysokým výskytem kriminality a tak se místní policie rozhodla zaměřit na bezpečnost dětí. Byla zde použita právě metoda Getis-Ord G_i^* . Do analýzy vstupovaly počty incidentů (trestných činů) v oblasti normalizované podle počtu obyvatelstva příslušné části. Výsledné analýzy vysokých a nízkých hodnot byly dále skloubeny s analýzou pomocí K funkce. Výsledné shluky byly klasifikovány do pěti kategorií (Claiborn, 2005).

Rozšířené analýzy trestné činnosti za pomoci ArcGIS Spatial Statistics Tools

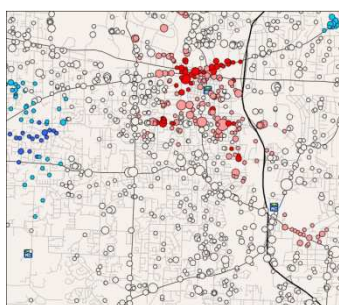
V roce 2012 byly v americkém státě Nebraska provedeny analýzy trestné činnosti. Pomocí prostorových analýz došlo k posouzení vzorů kriminality, optimalizaci a lokaci zdrojů a zlepšení tísňového volání. Byla zde využita metoda Getis-Ord G_i^* pro analýzu vandalství, která byla normalizována počtem trestných činů v oblasti. Bylo zjištěno, že cold spoty odrážely skutečnost environmentálních faktorů, které odrazovaly od trestné činnosti (obr. 5) (Scott a kol. 2012).



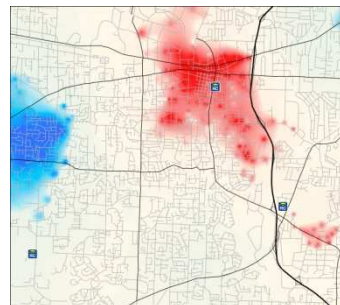
Obr. 5: Použití metody Hot Spot analýzy k analýze vandalismu v americkém státě Nebraska (ESRI, 2012).

Hot Spot analýza volání 911

Tato studie z roku 2007 ukázala, jak by mohla vypadat analýza bodových dat za pomoci nástroje Hot Spot analýzy k určení statistické významnosti. Cílem bylo zjištění, kde bylo provedeno více nebo méně hovorů, než se původně očekávalo. Byla zde provedena analýza dat volajících na linku 911. Shromážděné údaje byly nejprve seskupeny do vážených bodů. Metoda Getis-Ord G_i^* vyhodnotila údaje porovnáním střední průměrné hodnoty s globální průměrnou hodnotou. Výsledkem byla směrodatná odchylka (z-score) a pro jasnější představu byly tyto body interpolovány na povrchu (Frye, 2007).



Obr. 6: Bodová vizualizace metody Getis-Ord G_i^* povrch (ESRI, 2012).



Obr. 7: Interpolace Hot Spot analýzy na povrch (ESRI, 2012).

3.2 Domácí studie poslední doby

Z domácích studií byly vybrány především studie ze studentských prací a pracovníků KGI UP. Na KGI se využívají lokální prostorové analýzy nejvíce v praxi a tvoří částí diplomových či vědeckých prací. Dále byly zmapovány i domácí vědecké a odborné časopisy.

Prostorová autokorelace - všudypřítomný jev při analýze prostorových dat?

Spurná (2008) ve své práci publikované v sociologickém časopise popisuje možnosti a omezení prostorové autokorelace jako jedné z metod prostorové analýzy včetně ukázky některých aplikací. Jsou zde uvedeny přínosy a využití metod lokálních analýz. Zejména přínos statistické analýzy LISA a srovnání s globální statistikou. „Možnosti, které skýtá tento druh analýzy, jsou důležité zejména z hlediska identifikace prostorových odchylek a naopak oblasti podobného charakteru z hlediska zkoumaného jevu, například rozvojových os či problémových oblastí (Blažek, Spurná 2008, s. 785)“.

Shluková analýza funkčních ploch pro sledování urbánního vývoje Olomouckého regionu

Práce z roku 2011 využívala lokálních metod shlukové analýzy pro sledování výskytu statisticky významných shluků funkčních ploch v rámci města Olomouce a v rámci olomouckého regionu. Jako hlavní metoda zde byla zvolena metoda Getis-Ord G_i^* . Přímo v nastavení metody bylo využito metody konstantních vzdáleností, která je všeobecně doporučována. Dalším parametrem pro vzdálenost mezi prvky byla použita Euklidovská vzdálenost. Z výsledných shlukových analýz vybraných funkčních ploch na katastrálním území Olomouce bylo patrné, že Olomouc měla již v roce 1930 pevně danou sídelní i funkční prostorovou strukturu města, která se v pozdějších letech rozšiřovala. Byly poměrně přesně lokalizovány statisticky významné shluky bydlení, veřejného vybavení, výrobních ploch a zeleně v daných letech (Klícha, 2011).

Prostorové analýzy časových řad volebních výsledků

Cílem této práce bylo vytvořit soubor analýz a kartografických výstupů voleb v Olomouckém kraji od roku 2000. Jednou z použitých analýz je i metoda Getis-Ord G_i^* . Ta zde měla poskytnout jiný náhled na data a jejich volební jádra. Autor použil základní nastavení. Většina analýz zde však neodhalila významnější shluky nadprůměrných a podprůměrných hodnot. Výstupem bylo několik menších shluků, které při porovnání s volebními jádry určenou metodou územní volební podpory, byly ve většině případů totožné (Vlosinský, 2011).

Prostorové analýzy epidemiologických dat: Případová studie Olomouckého regionu

Tato práce vytvořená na KGI nejvíce koreluje s obsahem této práce. Je však více obecnější a nabízí ukázkou hned několika prostorových analýz řešených na datech EPIDAT pro území Olomouckého kraje. Prostorové analýzy jsou zde řešeny na úrovni k.ú. Jsou tu vypočteny hodnoty prostorové entropie (Shannonův index), které vypovídaly o prostorové variabilitě infekcí. Dále globální indexy prostorové autokorelace (Moranovo I pravidlo, Getis-Ord General G) a také lokální indikátory LISA a Getis-Ord G_i^* . Metoda LISA zde identifikovala oblasti shluků a jejich intenzitu. Výsledkem pak byly mapové výstupy pro Olomoucký kraj zobrazující shluky pro jednotlivé infekce a detailní popis jednotlivých shluků (Marek a kol. 2012).

3.3 Zhodnocení

Po prozkoumání zahraničních a domácích článků, publikací a prací byl učiněn závěr, že ačkoliv studie popisují použití lokálních metod, ne vždy popisují jednotlivé nastavení parametrů, které byly užity. Většina prací využívá metod konstantních vzdáleností, jež jsou obecně doporučovány. K měření vzdáleností jednotlivých sousedů používá Euklidovskou vzdálenost. Bylo by zajímavé vytvořit práci na užití všech možných metod a konstatovat, pro která data je možné uplatnit zvolené parametry.

4 VLASTNÍ ŘEŠENÍ

V této nejrozsáhlejší kapitole je podrobně popsán postup praktické části práce. Vlastní řešení obsahuje šest kapitol v pořadí, jak jsou postupně řešeny. Jsou zde vysvětleny jednotlivé metody, postupy a nastavení parametrů.

První kapitola bude popisovat aktualizaci dat databáze EPIDAT o nově získaná data. Bude zde i popsán stručně popis programu.

Následující kapitola se bude věnovat popisné charakteristice vstupních dat, kde bude provedena analýza dat EPIDAT a detailně prozkoumána Hepatitida A.

V kapitole předzpracování dat bude popsán postup práce s daty před vstupem do shlukové analýzy.

Dále bude provedeno testování náhodnosti rozmístění bodů výskytu, kde budou užity dvě metody pro zjištění, zda data vykazují shlukování nebo rozptyl.

V předposlední kapitole vlastního řešení bude popsán postup tvorby shlukové analýzy za pomoci metody Getis-Ord G_i^* .

Poslední část pojednává o vizualizaci analýz formou mapových výstupů a souborů KML použitých v aplikaci Google Earth.

4.1 Oprava databáze EPIDAT

Před samotným použitím dat pro shlukové analýzy byla provedena aktualizace dat EPIDAT pomocí poloautomatické opravy za použití databázového programu MS Access 2007. Samotná oprava a aktualizace všech chybných atributů, včetně výstupních souborů byla časově náročná. Bylo využito již vytvořeného programu v rámci předchozí bakalářské práce, na kterou tato práce navazovala. Tento program je řešen pomocí tabulek, dotazů, formulářů, sestav a maker. Po spuštění programu se otevře úvodní formulář s hlavní nabídkou (obr. 8).

Data byla opravována na základě celostátně platného registru UIR-ADR, který je od 1. 7. 2012 aktualizován a spravován celostátně platným registrem RUIAN.

V rámci předchozí práce byla opravena data za období 2004–2008. Tato práce aktualizovala data za období 2002–2003 a 2009–2011. Výsledný datový soubor tedy obsahoval 46 190 opravených záznamů za desetileté období. Výstupní data byla ve formátu xls.



Obr. 8: Úvodní formulář EPIDAT.

4.1.1 Stručný popis programu

Detailní popis programu je sepsán v bakalářské práci (Havlík, 2011), ovšem autor považuje za předmětné popsat stručný princip programu.

Program se skládá ze dvou částí, které na sebe navazují. První část je věnována opravě dat, druhá je průvodcem geokódování. V následujícím textu bude popsána první část.

V prvním kroku si uživatel nahraje chybná data, která bude následně opravovat. Druhý krok je rozdělen na tři části: průvodce opravou obcí, opravou okresů a opravou ulic. Všechny chybné záznamy se ukládají do tzv. chybovníku. Jedná se o formulář, kde uživatel pracuje s chybnými záznamy a postupně je poloautomaticky opravuje. Ve třetím kroku se všechny chybné záznamy (obce, okresy a ulice) zapíše do jedné tabulky pro následné vyhodnocení chyb. Ve čtvrtém kroku jsou přiřazeny všem opravovaným záznamům kódy, které zpřesňují význam prostorové informace. V poslední části je vytvořeno několik tabulek pomocí agregačních dotazů. Zde byly exportovány tabulky, které následně vstupovaly do části předzpracování dat shlukových analýz.

Stěžejní částí byla oprava ulic (obr. 9). Celkem bylo opraveno 13 863 záznamů ze 46 190 za období 2002-2011. Tato část zabrala nejvíce práce. V atributu ulice se mnohdy skrýval i název obce případně katastrálního území (viz tab. 1). Tato část programu nám tedy odhalila finální podobu celé opravy. Právě na atribut ULICE byl kladen velký důraz, protože následné shlukové analýzy byly provedeny právě na tuto úroveň dat.

CHYBNÉ ULICE	SPRAVNA OBEC	SPRAVNY OKRES	OPRAVA ULICE	OPRAVA OBEC	OPRAVA KU	OPRAVA OKRES	POZNAMKA
ezáče	OLOMOUC	OC	Řezáčova				oprava ulice
řmská	OLOMOUC	OC	Římská				oprava ulice
jezd u Uničova	UNICOV	OC		UJEZD	UJEZD U UNICOVA		oprava ku, obec
stín	USTIN	OC					název obce
zká	OLOMOUC	OC	Úzká				oprava ulice
široká	TOVACOV	PR	Široká				oprava ulice
šlechtitelů	OLOMOUC	OC	Šlechtitelů				oprava ulice
jit pánov	STEPANOV	OC					název obce
utarnov	STARNOV	OC					název obce
utřného	OLOMOUC	OC	Štřného				oprava ulice
umvald	SUMVALD	OC					název obce
uermova	HRANICE	PR	Švermova				oprava ulice
1. máje	HRANICE	PR	Tř. 1. máje				oprava ulice
1. května	PREROV	PR	1. května				oprava ulice
1. máje	NEZAMYSLICE	PV	1. máje				oprava ulice

Obr. 9: Průvodce opravou ulic.

Tab. 1: Ukázka opravy chyb u atributu ULICE (Havlík, 2010)

CHYBNÁ ULICE	OBEC	OKRES	OPRAVA ULICE	OPRAVA OBEC	OPRAVA KU	OPRAVA OKRES	VYSVĚTLIVKY
ezáče	OLOMOUC	OC	Řezáčova				nejčastější chyba, chybně zapsaná ulice spadající do příslušné obce a okresu
Jeremiáe	STEPANOV	OC	Jeremiášova	OLOMOUC			chybná ulice se nevyskytuje v příslušné obci, proto se zde opravuje i obec, do které ulice spadá
Štarnov	STEPANOV	OC		STARNOV			místo chybné ulice je zde napsaná obec, tudíž původní obec opravíme
Cakov	CHOLINA	OC		SENICE NA HANE	CAKOV		místo chybné ulice je zde napsaný název katastrálního území, zde se tudíž opravuje jako obec, tak katastrální území
Libuina	MORAVSKY BEROUN	BR		LIBINA		SU	zde je ukázána oprava obce i okresu, v chybné ulici je zapsaná obec, která nespadá do požadovaného okresu, tudíž musíme změnit i tento atribut
Legionarska	CHYBNY ZAZNAM	XX	Legionářská	OLOMOUC		OC	zde je ukázán názorný příklad, jak postupnou opravou přes obce, okresy a ulice můžeme dopátrat správné údaje
Lhota nad Moravou	LHOTA	XX		NAKLO	LHOTA NAD MORAVOU	OC	pomocí ulice zde zjišťujeme konkrétní katastrální území a k němu obec a okres

4.1.2 Aktualizace dat

Po opravě dat za období 2002-2003 a 2009-2011 bylo nutné tyto data seskupit s již opravenými daty a vytvořit ucelený soubor dat. Bylo také potřeba vytvořit řadu nových agregačních dotazů. Byly vybrány infekce, které budou vstupovat do prostorových analýz. Výběr byl většinou proveden na základě počtu dostupných dat a průběhu infekce (viz kapitola 5).

Výsledné agregační dotazy byly tedy vytvořeny na úroveň k.ú. a ulic. Důležitým mezníkem bylo rozhodnutí, k jakému území se data budou vztahovat. Na základě zapůjčených dat KGI s uliční sítí pro území ORP Olomouc byla po konzultaci s vedoucí práce mapována právě tato oblast. Data byla tedy upravena pro toto území. Počet záznamů pro území ORP Olomouc byl 10 560. Z tohoto počtu bylo celkem 7 402 údajů s uliční sítí.

Vzhledem k nízkému počtu dat katastrálních území byla vytvořena pouze jedna výstupní tabulka na tuto úroveň. Pokud by byly uvažovány k.ú. pro jednotlivé infekce, byla by velmi obtížná jejich následná vizualizace pomocí shlukových analýz, a data by byla nevyhovující. Proto byla většina agregačních dotazů brána na podrobnější úroveň ulic.

Bylo tedy vytvořeno deset výstupních tabulek na úroveň ulic vždy pro časové období deseti let.

Seznam všech výstupních tabulek:

- Celkem (za všechny infekce)
- Salmonela A02
- Varicella B01 (plané neštovice) - roky
- Varicella B01 (plané neštovice) - měsíce
- Parotitida A69.2
- Pohlaví celkem
- Zastoupení mužů
- Zastoupení žen
- Věková struktura (0-14, 15-59, 60+)
- Školská struktura (MŠ, ZŠ, SŠ, VŠ)
- Zaměstnaní – pro nízký počet dat bez shlukových analýz
- +
- Výskyty za k.ú. v rocích 2002-2011

Specifickou pozornost si vyžádala kategorie zaměstnání. Před tvorbou dotazu byly jednotlivé druhy zaměstnání, použité v této práci, klasifikovány do tříd podle Národní klasifikace zaměstnání. V atributu JAKEZAMEST bylo přes 300 udaných záznamů z 10 560 celkových rozklasifikováno do devíti tříd. Z těchto tříd byly vybrány čtyři nejčastější. Z důvodu pouze 3% zastoupení v tomto atributu nakonec nebyla kategorie zaměstnání zahrnuta do dalšího zpracování a tvorby shlukových analýz.

Nejčastěji zastoupená zaměstnání v EPIDAT byly v těchto sektorech:

1. Techničtí, zdravotničtí, pedagogičtí pracovníci a pracovníci v příbuzných oborech
2. Řemeslníci a kvalifikovaní výrobci, zpracovatelé, opraváři (kromě obsluhy strojů a zařízení)
3. Vědečtí a odborní duševní pracovníci
4. Obsluha strojů a zařízení

4.2 Popisná charakteristika vstupních dat

V této kapitole jsou podrobněji popsány charakteristiky vstupních dat. Ve velké míře zde budou analyzována vstupní data databáze EPIDAT. Tato kapitola byla vytvářena ještě před získáním dat za období 2002-2003 a 2011. Proto zde byly popsány data za období 2004-2010.

Popis jednotlivých atributů EPIDAT je zpracován v předchozí práci (Havlík, 2010). Každý záznam v tabulce má celkem 50 atributů. Ze sledovaných atributů byly nejdůležitější atributy ULICE, OBEC, OKRES, MISTONAKAZ a MISTOONEM. Ty byly opraveny. Další sledované atributy byly VEK, EPIDEMIE, KOLEKTIV a ZAMESTNANI.

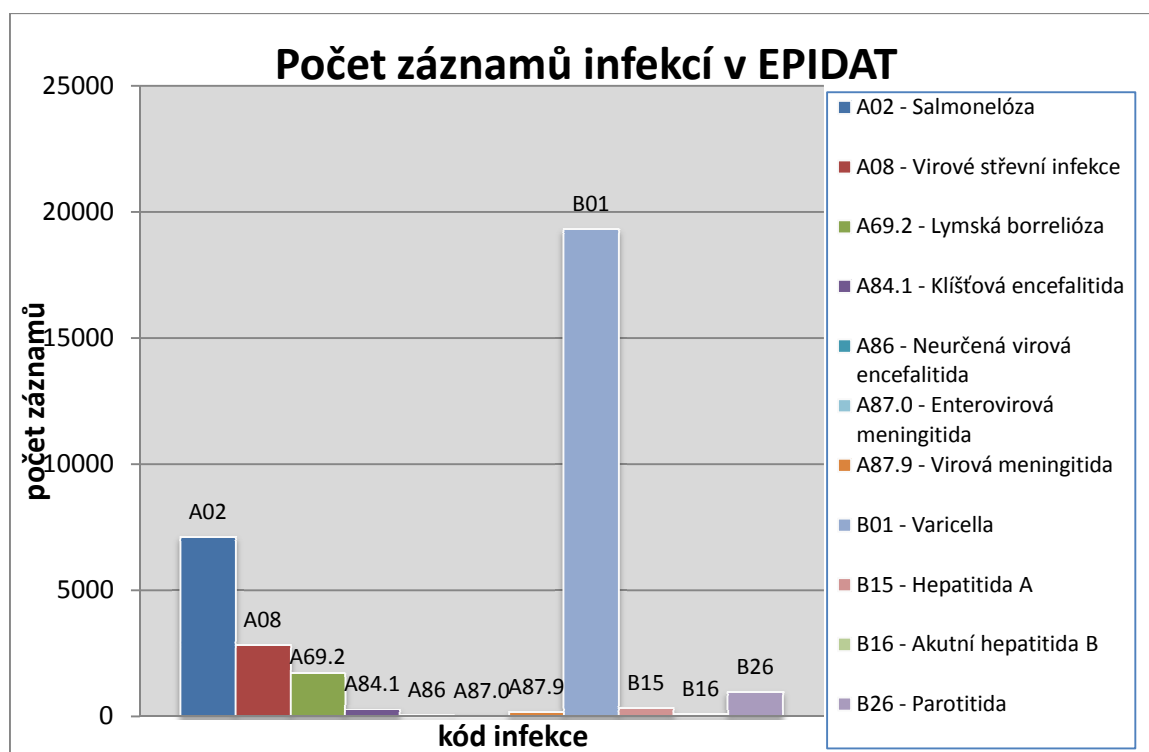
Součástí této kapitoly bude i porovnání vybrané infekce se statistickými daty pro Olomoucký kraj.

Dalšími důležitými vstupními daty pro tvorbu shlukových analýz byly data uliční sítě ORP Olomouc. Tyto data obsahovaly uliční segmenty, které bylo nutné před použitím dat sloučit z důvodu náhodného rozmístění výskytu podél této vrstvy. Důležitými atributy byly kód ulice, název a délka. Práce s těmito daty bude následně popsána v kapitole 4.3. Výčet všech použitých dat této práce byl sepsán v kapitole 2.2.

4.2.1 Analýza dat EPIDAT

Jak bylo uvedeno v úvodu této kapitoly, analýza byla provedena na datech za období 2004-2010. Data byla zkoumána za celé území Olomouckého kraje. Celkem bylo v datech pracováno s jedenácti infekcemi. Jelikož pro Enterovirovou meningitidu A87.0 byla dostupná data jen pro období 2009-2010, nebyla zahrnuta do dalšího zkoumání.

Největší zastoupení v datech mají plané neštovice (odborným názvem Varicella). Ta je zastoupena téměř ve dvou třetinách dat. Dalšími velmi častými infekcemi jsou salmonelóza, virová střevní infekce a lymfská borelióza. Přehledně jsou počty záznamů znázorněny na obr. 10.



Obr. 10: Graf počtu záznamů infekcí v EPIDAT.

Pro každou infekci bylo vypočteno její procentuální zastoupení v jednotlivých letech (viz tab. 3). Nejzajímavější pro další zkoumání byly neplynulé průběhy infekcí s výrazným vrcholem v rocích, případně s postupným nárůstem či úpadkem.

Tab. 3: Procentuální zastoupení infekcí v období 2004–2010

NÁZEV INFEKCE	ZKRATKA INFEKCE	2004	2005	2006	2007	2008	2009	2010	CELKEM
SALMONELÓZA	A02	1522	1706	1261	803	700	594	515	7101
		21,4%	24,0%	17,8%	11,3%	9,9%	8,4%	7,3%	100
VIROVÉ STŘEVNÍ INFEKCE	A08	198	197	581	476	535	396	431	2814
		7,0%	7,0%	20,6%	16,9%	19,0%	14,1%	15,3%	100
LYMSKÁ BORRELIÓZA	A69.2	180	229	319	225	274	220	237	1684
		10,7%	13,6%	18,9%	13,4%	16,3%	13,1%	14,1%	100
KLÍČTOVÁ ENCEFALITIDA	A84.1	33	28	58	32	24	46	38	259
		12,7%	10,8%	22,4%	12,4%	9,3%	17,8%	14,7%	100
NEURČENÁ VIROVÁ ENCEFALITIDA	A86	4	4	5	5	1	3	3	25
		16,0%	16,0%	20,0%	20,0%	4,0%	12,0%	12,0%	100
VIROVÁ MENINGITIDA	A87.9	15	31	16	26	16	29	25	158
		9,5%	19,6%	10,1%	16,5%	10,1%	18,4%	15,8%	100
VARICELLA	B01	3341	2865	1973	3529	1904	3057	2623	19292
		17,3%	14,9%	10,2%	18,3%	9,9%	15,8%	13,6%	100
HEPATITIDA A	B15	11	12	6	4	153	66	49	301
		3,7%	4,0%	2,0%	1,3%	50,8%	21,9%	16,3%	100
AKUTNÍ HEPATITIDA B	B16	22	20	14	6	18	5	4	89
		24,7%	22,5%	15,7%	6,7%	20,2%	5,6%	4,5%	100
PAROTITIDA	B26	1	18	356	126	115	120	227	963
		0,1%	1,9%	37,0%	13,1%	11,9%	12,5%	23,6%	100
CELKEM		5327	5110	4589	5234	3739	4537	4153	32689

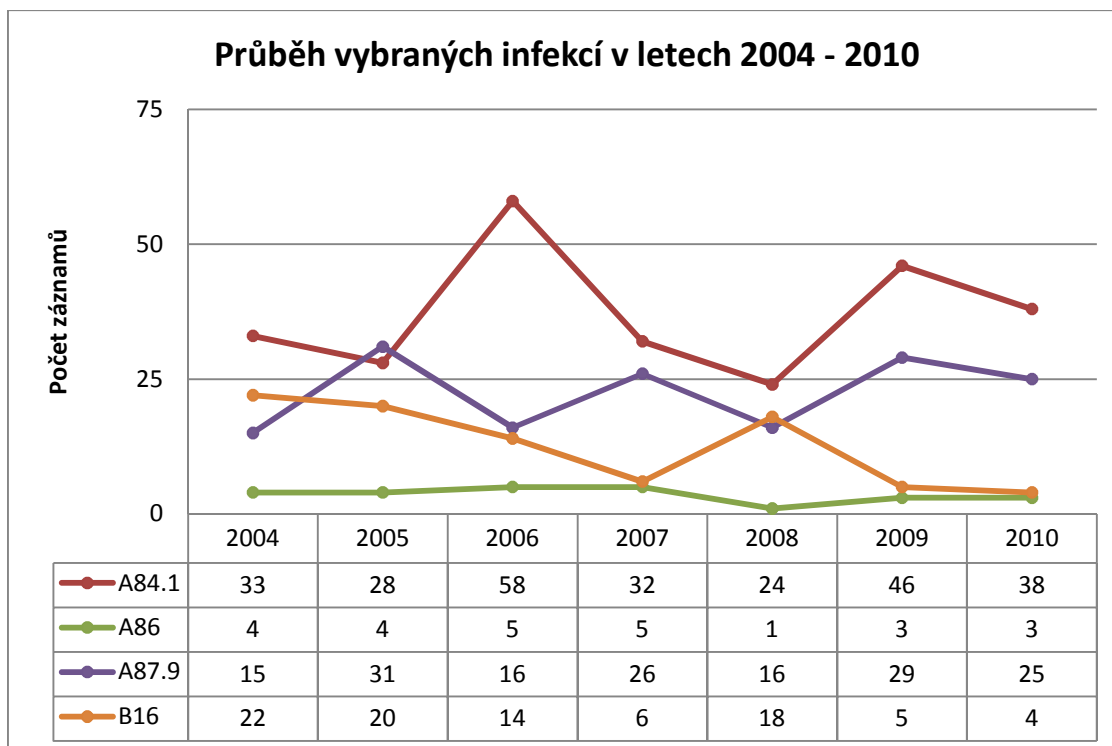
Do pomocné tabulky byl zapsán průběh infekce a dále bylo stanoveno, zda je infekce hodna dalšího zkoumání (viz tab. 4).

Nepravidelný průběh vykazovaly tyto infekce: HEPATITIDA A – B15, SALMONELÓZA – A02, PAROTITIDA – B26 a VARICELLA – B01, která vykazovala velmi neplynulý průběh v jednotlivých letech. Na obr. 11, 12 a 13 jsou zobrazeny grafy průběhu vybraných infekcí. Z důvodu velkého rozsahu dat byly infekce rozdělené na dva grafy.

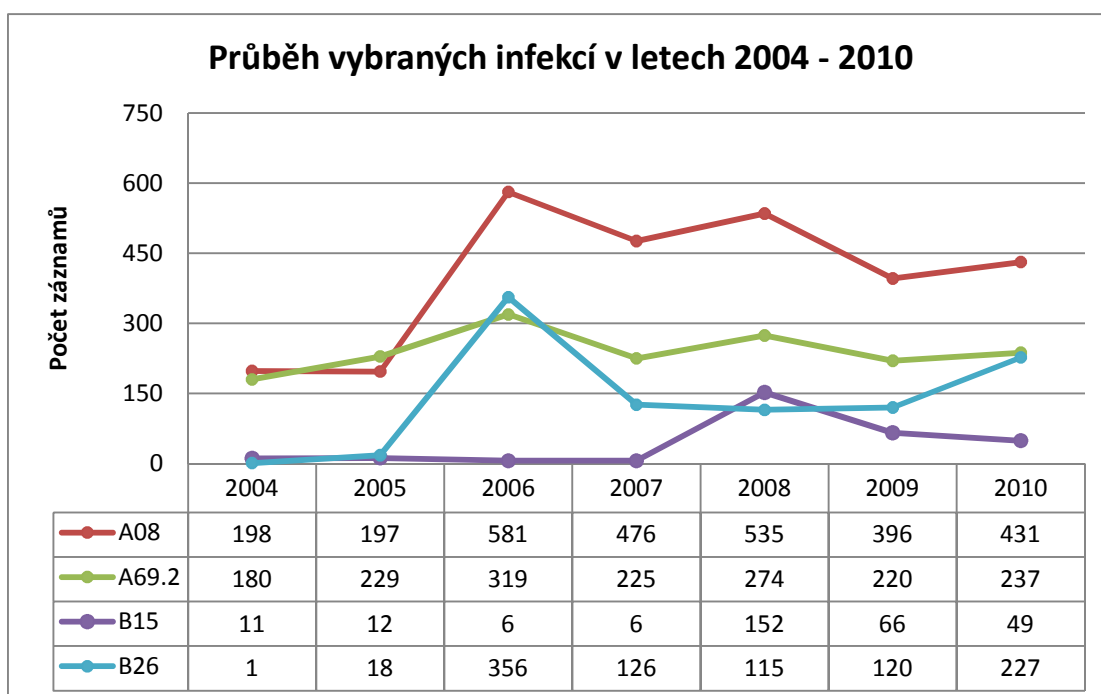
V následující kapitole byla provedena analýza vybrané infekce, která vykazovala nepravidelný průběh. Jednalo se o HEPATITIDU A.

Tab. 4: Pomocná tabulka s průběhem infekce

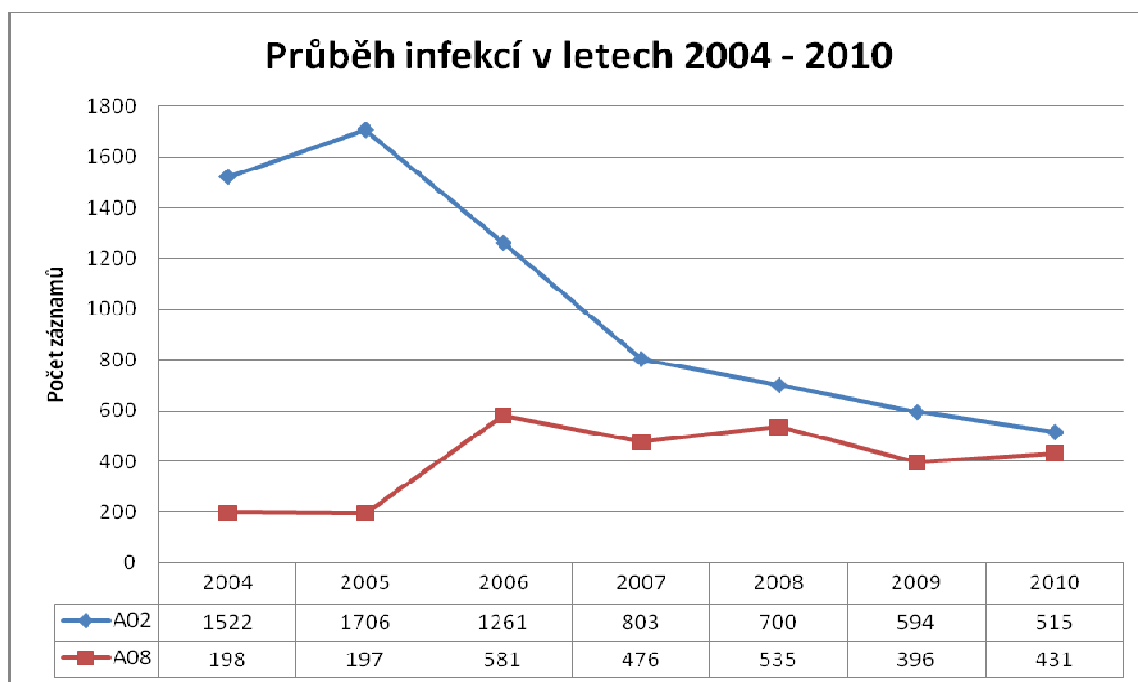
NÁZEV INFEKCE	ZKRATKA INFEKCE	Průběh infekce	POČET ZAZNAMU	VYRAZNA ODCHYLKA
SALMONELÓZA	A02	velmi strmá klesající tendence	7101	NE/ výrazný pokles
VIROVÉ STŘEVNÍ INFEKCE	A08	mírně stoupající tendence	2814	NE
LYMSKÁ BORRELIÓZA	A69.2	mírně stoupající tendence výskytu s vrcholem v roce 2006, žádný výrazný výkyv	1684	NE
KLÍŠŤOVÁ ENCEFALITIDA	A84.1	velmi mírné výkyvy v letech 2006 a 2009, téměř konstantní	259	NE
NEURČENÁ VIROVÁ ENCEFALITIDA	A86	téměř konstantní	25	NE
VIROVÁ MENINGITIDA	A87.9	téměř konstantní	158	NE
VARICELLA	B01	mírně klesající trend a velmi zajímavé výkyvy v letech 2006 a 2008	19292	NE/ výkyvy
HEPATITIDA A	B15	konstantní a velmi nízké hodnoty v letech 2004 - 2007, v roce 2008 výrazný nárůst, poté pokles a ustálení	301	ANO
AKUTNÍ HEPATITIDA B	B16	téměř konstantní	89	NE
PAROTITIDA	B26	stoupající tendence výskytu s výrazným vrcholem v roce 2006	963	ANO



Obr. 11: Průběh vybraných infekcí s nízkým počtem záznamů.



Obr. 12: Průběh vybraných infekcí s vyšším počtem záznamů.



Obr. 13: Průběh vybraných infekcí s vysokým počtem záznamů.

4.2.2 Hepatitida A

Hepatitida A je označována jako infekční žloutenka nebo "nemoc špinavých rukou". Vniká do těla nejčastěji ústy. Infikovaná osoba vylučuje viry stolicí. K šíření virové hepatitidy A významnou měrou přispívá zanedbaná osobní hygiena (Hepatitida A, 2013).

Hepatitida typu A je nejčastěji přenášena:

- znečištěnou stravou
- špinavými rukama

Onemocnění častěji postihuje příslušníky tzv. rizikových skupin:

- děti předškolního a školního věku
- mladí dospělí do 35 let
- cestovatelé

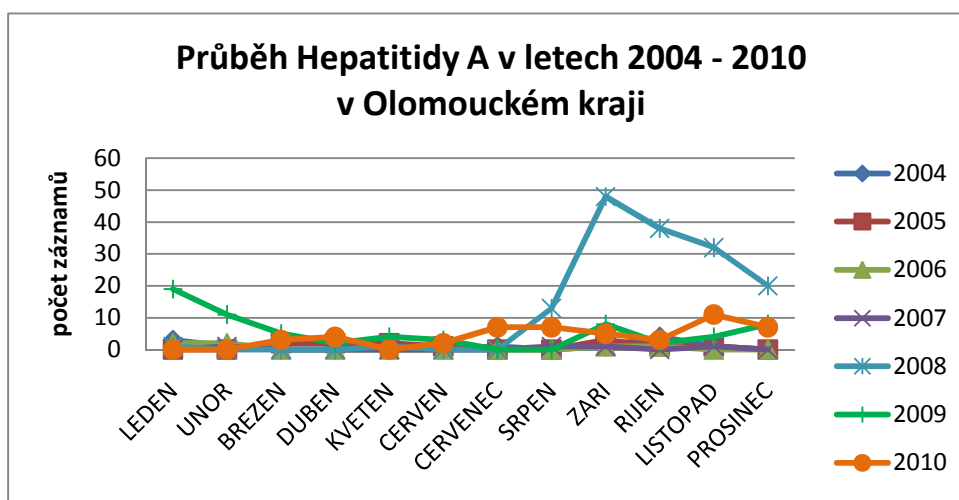
Ve zvýšeném riziku nákazy virem jsou také někteří lidé při výkonu svého povolání:

- zdravotníci
- vychovatelé a učitelé
- vojáci
- policisté
- celníci

- pracovníci soudů a vězeňské služby
- letecký personál
- průvodci cestovních kanceláří
- řidiči dálkové dopravy
- pracovníci s odpady

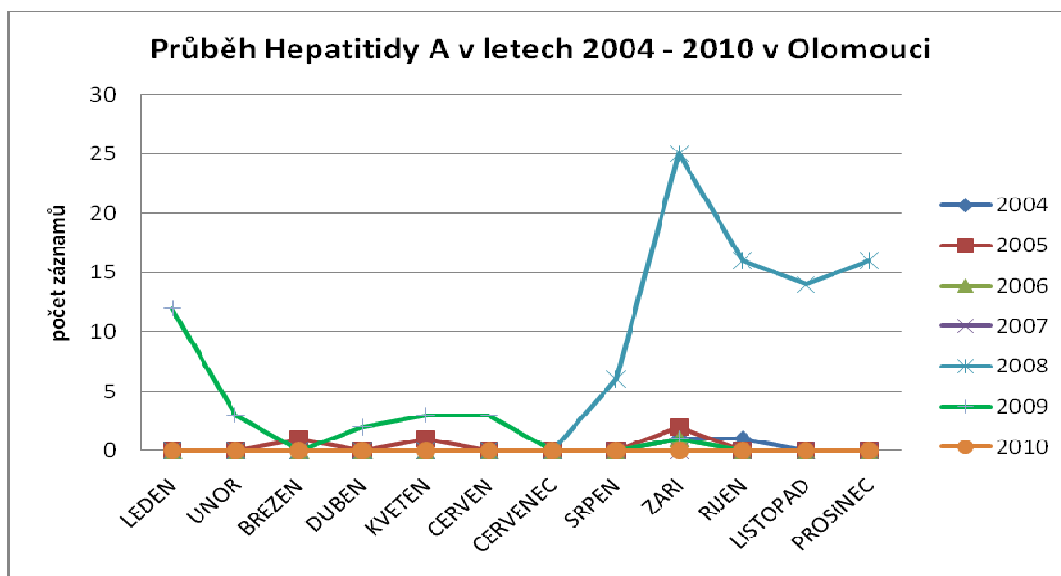
Před analýzou byla tedy zjištěna základní charakteristika této infekce a následně porovnána se zjištěnými údaji.

Hepatitis A byla vybrána na základě velmi zajímavého nárůstu. Od roku 2004 do června roku 2008 byl průběh infekce téměř konstantní. V červenci 2008 prudce stoupl výskyt této infekce ve všech obcích Olomouckého kraje (viz obr. 14). Ten měl vrchol v září 2008, a poté mírně klesal až do dubna 2009. Od té doby se více méně ustálil. U všech roků byl vrchol infekce v podzimním období, především v měsíci září.



Obr. 14: Průběh Hepatitidy A v letech 2004-2010 v Olomouckém kraji.

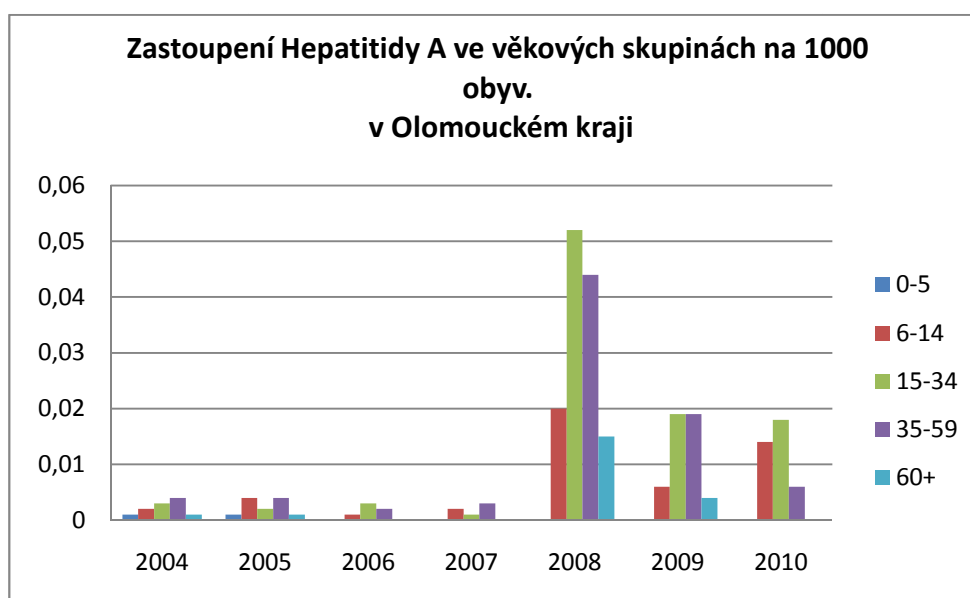
Zajímavé je také porovnání výskytu infekce v Olomouckém kraji s městem Olomouc (viz obr. 14), který má největší zastoupení této infekce ze všech měst. Zatímco, v Olomouckém kraji infekce po vrcholu v září 2008 klesala, v Olomouci ještě v prosinci vykazovala zvýšený výskyt. Zajímavé bylo zjištění, že v Olomouci nebyl v roce 2010 zjištěn ani jeden výskyt Hepatitidy A, zatímco v Olomouckém kraji měla infekce ještě mírně stoupající tendenci. Mohli bychom se domnívat, že lidé v Olomouci jsou více imunní oproti této infekci, nebo byli očkovaní.



Obr. 15: Průběh Hepatitidy A v letech 2004-2010 v Olomouci.

Celkově byl tedy určen jednoznačný průběh Hepatitidy A v jednotlivých měsících. Stoupající výskyt začínal v měsíci červenci s vrcholem v září. Poté nastal prudší pokles, který se ustálil až v měsíci únoru. Od března do června jsou hodnoty výskytu více méně konstantní. Rozhodně největší vliv na tuto křivku měl rok 2008.

Při porovnání s daty z ČSÚ bylo zjištěno zastoupení infekce ve věkových skupinách na 1000 obyvatel. Celkově byly nejvyšší hodnoty zjištěny u kategorie 15-34 let a naopak nejnižší u věkové skupiny 0-5 let. Při porovnání s obecným popisem infekce lze konstatovat, že výsledná analýza vybrané infekce odpovídá základnímu popisu.



Obr. 16 Zastoupení Hepatitidy A ve věkových skupinách na 1000 obyvatel v Olomouckém kraji.

4.3 Předzpracování dat

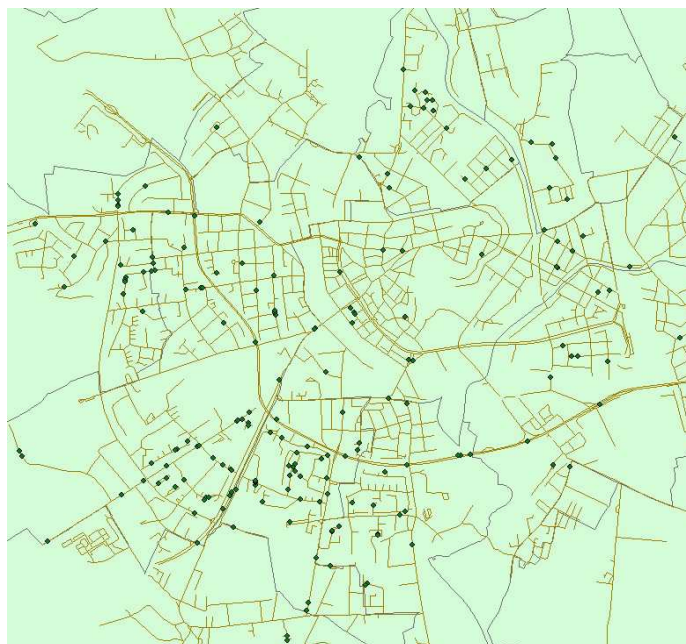
V této části práce bylo provedeno hned několik operací s daty před samotným vstupem do shlukové analýzy. Do shlukové analýzy Getis-Ord G_i^* vstupoval pouze jeden atribut. Tím byla určena intenzita výskytu jevu, konkrétněji počet výskytů na počet obyvatel v geometrické síti.

Na tuto část práce byl kladen velký důraz. Bylo potřeba využít metod hromadné práce s daty, jelikož většina sledovaných atributů byla zpracována za každý rok v období 2002-2011. U většiny nástrojů bylo důležité učinit správné kroky při rozhodování nastavení parametrů. Všechny operace byly provedeny v programu ArcGIS. Byly provedeny následující operace, které jsou dále v textu rozepsány:

- Přiřazení prostorové složky datům EPIDAT – Join Data
- Rozložení dat do k.ú. a uliční sítě – nástroj Create Random Points
- Vytvoření geometrické sítě nad daty – extenze Repeating Shapes for ArcGIS
- Standardizace dat (použití dat ZSJ) – nástroj Identity, Calculate Geometry

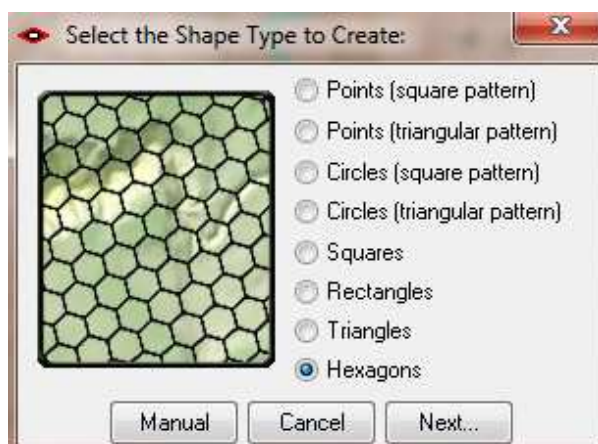
První operací bylo přiřadit tabulkovým datům s výskyty prostorovou složku. Původním záměrem bylo vytvořit adresový lokátor ulice a katastrálního území a využít nástroj Geocode Addresses z nástrojové sady Geocoding Tools. Bylo zjištěno, že výstupem jsou vždy bodová data. Zde byl shledán první problém. V následné operaci rozložení dat pomocí nástroje Create Random Points z nástrojové sady Data Management Tools nelze pracovat s bodovými daty pro parametr specififikující atribut s počtem bodů, které mají být náhodně rozmístěné podél zadané vrstvy. Byl tedy použit klasický nástroj na propojení tabulkových a prostorových dat Join Data a byla propojena data uliční sítě ORP Olomouc s výstupními tabulkami EPIDAT.

Druhou operací byla data rozložena do katastrálních území a uliční sítě. Před rozložením byly pomocí nástroje Merge (Multiple merging features in the same layer) sloučeny záznamy uličních segmentů se stejným jménem ve vrstvě ulice_ORP.shp. Z území ORP Olomouc bylo odstraněno 5 k.ú. (Město Libavá, Rudoltovice, Velká Střelná, Slavkov u Města Libavá a Čermná u Města Libavá). Bylo tak učiněno z důvodů velkého území (vojenský újezd Libavá) a velmi nízkého výskytu infekcí. Následně byl využit výše zmíněný nástroj Create Random Points. Z volitelných parametrů zde byla zvolena právě vytvořená vrstva z předchozího kroku pro zadání počtu bodů výskytu a přidána vrstva KU_ORP, která určovala rozsah dat. Ukázka náhodného rozmístění na uliční síti je vidět na obrázku 17. Samozřejmostí bylo použití nástroje ModelBuilder, kterým byl proces hromadného vytváření náhodných bodů usnadněn.



Obr. 17: Výřez náhodného umístění bodů na uliční síti Olomouce.

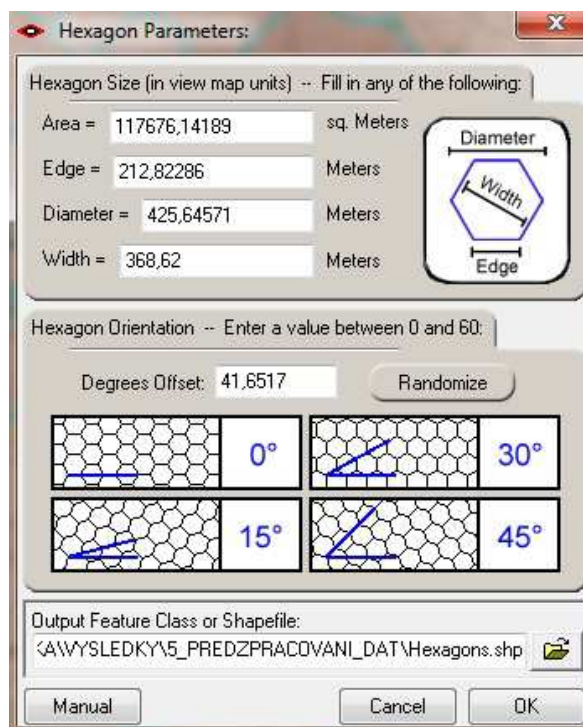
Při vytvoření geometrické sítě nad daty bylo nejprve pracováno s nástrojem Create Fishnet, který je dostupný v nástrojové sadě Data Management Tools. Tento nástroj vytváří čtvercové nebo obdélníkové geometrické sítě. Pro naši analýzu bylo vhodné najít tvar, který bude mít co nejvíce sousedů. Po konzultaci byla zvolena extenze Repeating Shapes, která je volně dostupná jako rozšíření pro ArcGIS. Ta nabízí mnoho různých geometrických sítí od bodů, čtverců, trojúhelníků až po hexagony (obr. 18).



Obr. 18: Výběr tvaru geometrické sítě v extenzi Repeating Shapes.

Pro námi zvolenou práci byla vybrána hexagonová síť, která má šest sousedních stran. Byla tedy vytvořena síť nad daty ORP Olomouce. V nastavení bylo také možné vybrat, zda geometrická síť bude vytvořena nad vybranými prvky vrstvy, nebo pro celou vrstvu.

V následujícím kroku bylo třeba nastavit parametry mřížky, např. délku hran, úhel natočení či plochu (obr. 19).

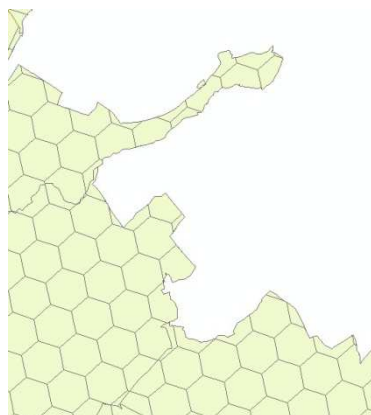


Obr. 19: Nastavení parametrů v extenzi Repeating Shapes.

Následně proběhla fáze rozhodování, jakou velikost mřížky zvolit. Byl vypočten průměr a medián uliční sítě pro ORP Olomouc. Hodnota průměru byla 554,52 m a mediánu 368,62 m. Po konzultaci byl zvolen medián, tedy střední hodnota. Úhel natočení byl vybrán náhodně a ostatní parametry se dopočítaly na základě zvolené velikosti hrany.

Následovala poslední fáze – standardizace dat. Počet nemocných se standardizoval celkovým počtem obyvatel. Po vypočtení plochy každého hexagonu v předchozím kroku bylo třeba tyto hexagony oříznout vrstvou ku_orp_ol.shp. Do této vrstvy byl přidán atribut OB, do něhož byly napsány počty obyvatel k.ú. Tyto údaje byly zjištěny prostřednictvím ÚIR-ZSJ, kde byly nejprve základní sídelní jednotky v rámci území sloučeny a následně byl vypočten celkový počet obyvatel. Pomocí nástroje Identity v nástrojové sadě Analysis Tools byla vrstva hexagonu rozdělena a vypočítána plocha každého hexagonu v k.ú. ORP Olomouc (viz obr. 20).

Do nově vzniklé vrstvy byl přidán atribut OB_mrizka, kde byl vypočítán pomocí nástroje Field Calculator počet obyvatel v mřížce. Bylo zde opět použito nástroje ModelBuilder pro urychlení a hromadné vytváření vrstev.

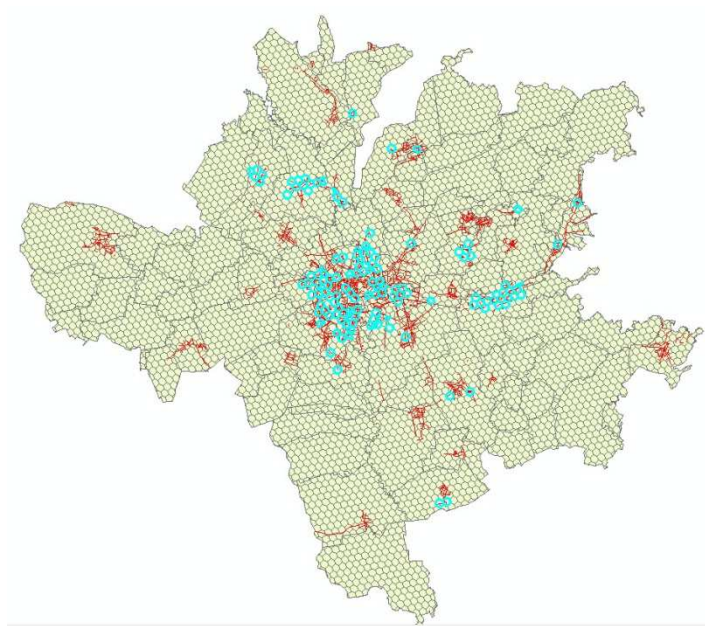


Obr. 20 Ukázka ořezání hexagonální sítě vrstvou k.ú. ORP Olomouc pomocí nástroje Identity.

Finální operací bylo propojení vrstvy identity.shp s vrstvou náhodného rozmístění bodů výskytu random.shp. Tato vrstva v sobě nesla počet výskytů. Následně byl přidán nový atribut s názvem VYSKYT (Obr. 21). Do něj byl opět pomocí nástroje Field Calculator dopočítán podíl výskytů a počtu obyvatel v mřížce. Tento atribut tedy vstupoval jako jediný parametr do shlukové analýzy Getis-Ord G_i^* . Celkem tedy bylo vytvořeno pro 10 + 1 sledovaných jevů dohromady 91 výstupních vrstev obsahující atribut VYSKYT. Přehled počtu vrstev a sledovaných jevů zobrazuje tabulka č. 2.

Tab. 2: Sledované jevy a počet vrstev z EPIDAT

POČET VRSTEV	NAZEV JEVU	PODSKUPINA
12	ZASTOUPENI B01 V MEŠÍČÍCH	LEDEN - PROSINEC
7x10	ZASTOUPENI - A02, A69.2, B01, MUZI, ZENY, CELKEM_ULICE, KU	2002 - 2011
4	SKOLSKA STRUKTURA	MŠ, ZŠ, SŠ, VŠ
3	VEKOVE ZASTOUPENI	0-14, 15-59, 60 +
2	POHLAVNI ZASTOUPENI	MUŽI, ŽENY



Obr. 21: Výskyt planých neštovic (Varicella B01) v roce 2002.

4.4 Testování náhodnosti rozmístění bodů výskytu

Tato kapitola se věnuje interferenčním statistickým testům pro body. Po provedení těchto testů lze určit, jaké chování vykazuje distribuce použitých bodů. Zda se shlukuje, rozptyluje nebo je náhodná.

Dle Horáka (2008) se k posouzení bodového vzorku používá řady technik. Mezi nejběžnější patří kvadrantové testy náhodnosti, metoda nejbližších vzdáleností a v poslední době často využívaná K-funkce.

Testování náhodnosti rozmístění bodů výskytu se často užívá v situacích, kdy body reprezentují určitou událost, např. výskyt environmentálního jevu, nemoci či kriminálního činu.

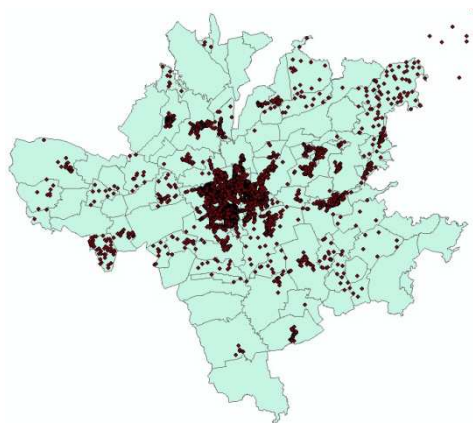
V následujícím textu budou popsány dvě interferenční metody, které byly použity pro testování našich dat. Obě tyto metody jsou vhodné pro prostorová data a jsou obsaženy v nástrojové sadě Spatial Statistics programu ArcGIS.

První metoda je Average Nearest Neighbour (Průměrný nejbližší sused) a Ripley's K-function (Ripleyova K-funkce). V následujícím textu bude použit jejich český ekvivalent. Po konzultaci s vedoucí práce a konzultantem byly tyto metody zvoleny jako dostačující, a tudíž nebylo použito žádných jiných funkcí a programů.

4.4.1 Metoda průměrného nejbližšího suseda

Metoda vypočítá index nejbližšího suseda založený na průměrné vzdálenosti jednotlivých prvků od nejbližšího susedního prvku. Výpočet je provedený z těžiště každého prvku. Poté je vypočtena průměrná vzdálenost všech nejbližších susedů a následně dále porovnávána.

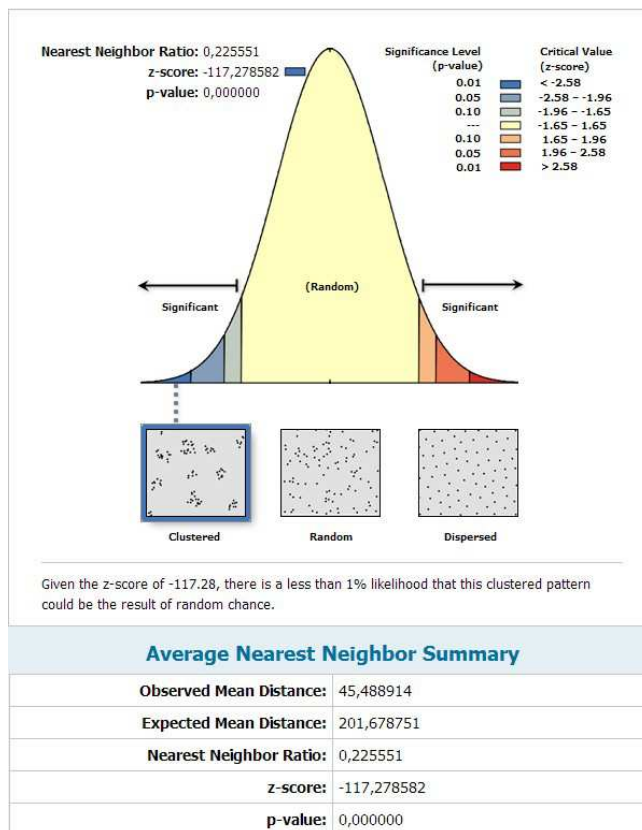
V případě, že průměrná vzdálenost je menší než hypotetická hodnota náhodné distribuce prvků, vykazují analyzované prvky shlukování. V opačném případě, pokud je průměrná hodnota větší než hypotetická, jsou data považována za rozptýlená. Nulová hypotéza udává, že vrstva nebo hodnoty ve vrstvě vykazují statisticky náhodný vzorek (ESRI, 2012).



Obr. 22: Bodový vzorek ulice_ku_merge.shp.

Poměr (ratio) průměrného nejbližšího souseda se vypočítá jako podíl pozorovaných průměrných vzdáleností s očekávanou průměrnou vzdáleností (ta je založena na hypotetickém náhodném rozdělení stejných hodnot prvků pokrývající stejnou plochu). V případě, že poměr je menší než 1, vzorek vykazuje shlukování, v opačném případě vykazuje disperzi (ESRI, 2012).

Do této analýzy vstupoval bodový vzorek ulice_ku_merge.shp (obr. 22). Byla tedy provedena analýza nad spojenými daty výskytu uliční sítě a katastrálního území. Zvolenými parametry byly Euklidovská vzdálenost prvků a volba generování zprávy (obr. 23). Zpráva byla generována ve formátu .html a zobrazena v prohlížeči.



Obr. 22: Výstup z metody průměrného nejbližšího souseda.

Výsledkem bylo pět hodnot: změřená střední vzdálenost, předpokládaná střední vzdálenost, poměr, z-score a p-value. Z výsledku je patrné, že poměr (Nearest Neighbor Ratio) je 0,22 tedy nižší než 1. Data tedy vykazují shlukování. z-score je standardní odchylka, kdy v tomto případě je silně záporná -117. Hodnota p-value udává pravděpodobnost výskytu.

4.4.2 Ripleyova K-funkce

Tato funkce je ideální pro zkoumání rozptylu či shlukování v různých vzdálenostech. Charakteristickým rysem této metody je shrnutí prostorových závislosti přes rozsah vzdáleností. Využívá možnosti nastavení prahových hodnot.

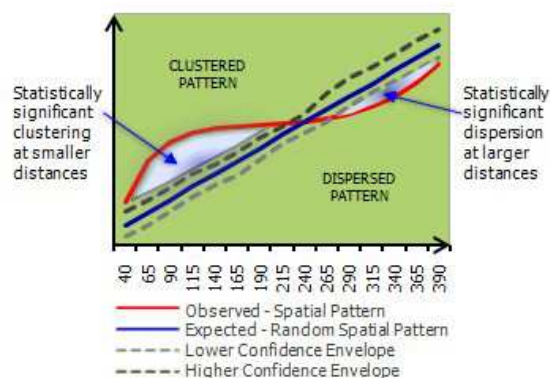
Při nastavení se zpravidla zadává počet vzdálenostních pásem a velikost přírůstku. Na základě těchto informací nástroj vypočítá průměrný počet sousedních prvků spojených s každým prvkem. Se zvyšujícím se počtem hodnotících vzdáleností bude mít každý prvek zpravidla více sousedů. V případě, že počet průměrných sousedů pro určitou vzdálenost je vyšší než průměrná koncentrace prvků v celé zkoumané oblasti, distribuce je považována za shlukování v této vzdálenosti (ESRI 2012).

Pokud je pozorovaná K hodnota (Observed – Spatial Pattern) větší než očekávaná (Expected – Random Spatial Pattern), distribuce prvků vykazuje shluky. Pokud je hodnota K menší než očekávaná, pak vykazuje rozptyl.

Kolem očekávané hodnoty je vypočítán interval spolehlivosti (Higher/Lower Confidence Envelope). Ten má stejný průběh jako očekávaná hodnota a vypočítává se na základě náhodného umístění bodů ve zkoumané oblasti. Počet bodů náhodně umístěných se rovná počtu bodů v třídě prvků. Každá sada náhodných umístěných bodů se nazývá permutace.

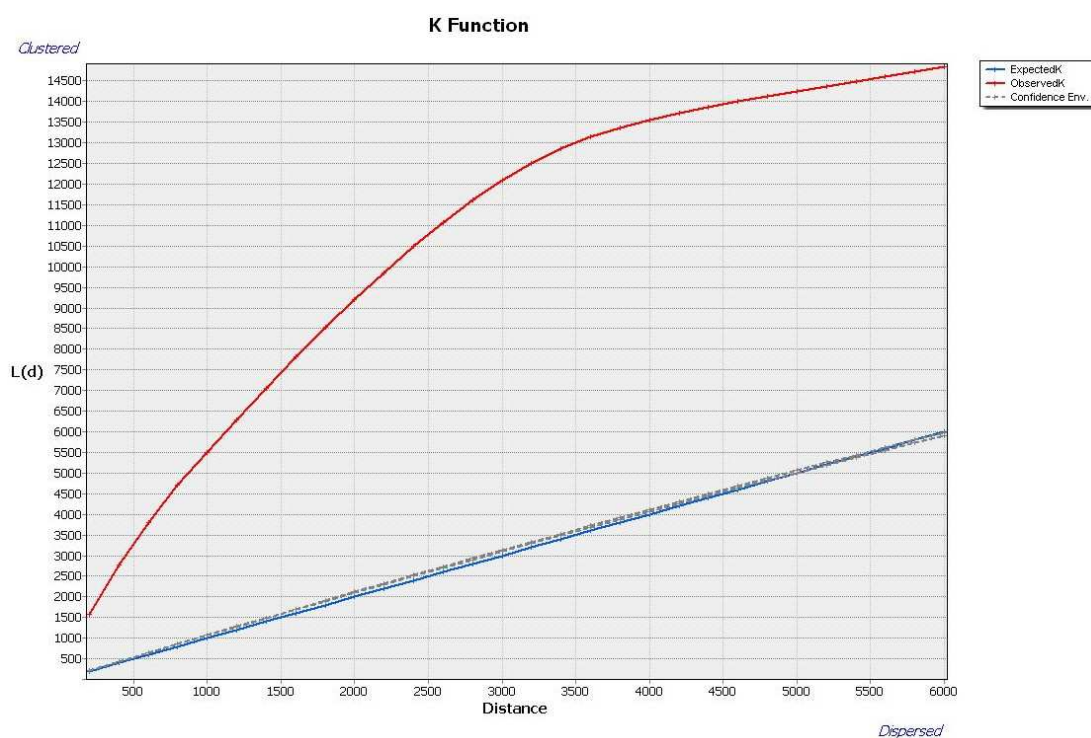
Počet permutací je další parametr, který lze zvolit při tvorbě této funkce. Udává tedy počet sad bodů, které budou náhodně umístěny, a na jejich základě je vypočítán interval spolehlivosti.

Na výběr máme z 0, 9, 99, 999 permutací. Tento parametr výrazně ovlivňuje rychlost výpočtu. Pokud je tedy pozorovaná hodnota K větší než vyšší hranice intervalu spolehlivosti (Higher Confidence Envelope), je prostorové shlukování pro tuto vzdálenost statisticky významné. Naopak pokud je pozorovaná hodnota K nižší než hodnota nižší hranice intervalu spolehlivosti (Lower Confidence Envelope), je prostorová disperze statisticky významná (ESRI, 2012). Vše je znázorněno na obrázku 23.



Obr. 23: Princip Ripley K-funkce.

V našem případě bylo použito: počet distančních pásem – 30, počet permutací – 9, počáteční vzdálenost – 0 a počet přírůstku – 200 m. Byla tedy vypočítaná vzdálenost 6 km od každého bodu. Z výsledků na obrázku 24 můžeme vidět, že data vykazují výrazné shlukování, které bylo potvrzeno u předchozí metody, a kdy se počet shluků ustaluje zhruba po třetím kilometru.



Obr. 24: Průběh K funkce pro naše data.

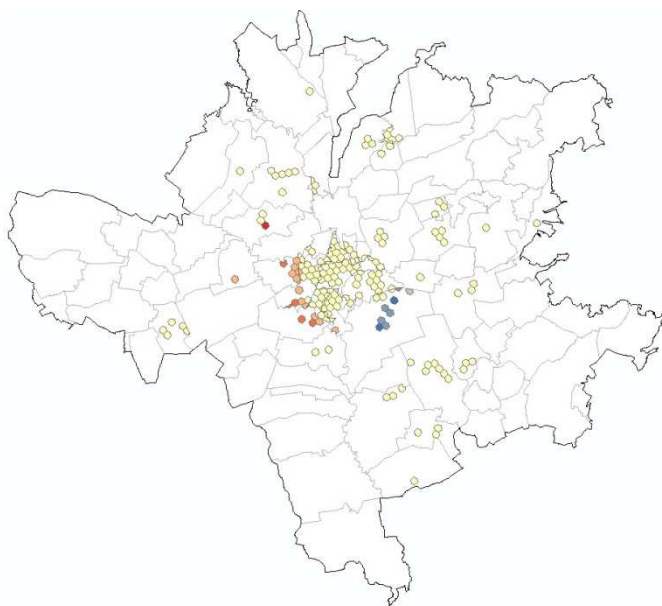
4.5 Shlukové analýzy

Použitou metodou pro shlukovou analýzu byla metoda Hot Spots analysis (Getis-Ord G_i^*) nacházející se v nástrojové sadě Spatial Statistics. Popis metody je popsán v kapitole 2.4.1.

Vstupními daty byl jediný parametr „výskyt“, který byl standardizován na počet obyvatel v mřížce. Analýzy byly provedeny na úroveň uliční sítě pro ORP Olomouc. Z důvodu nízkého počtu dat a špatné interpretace výsledků nebyly provedeny shlukové analýzy pro data výskytu infekcí v katastrálním území.

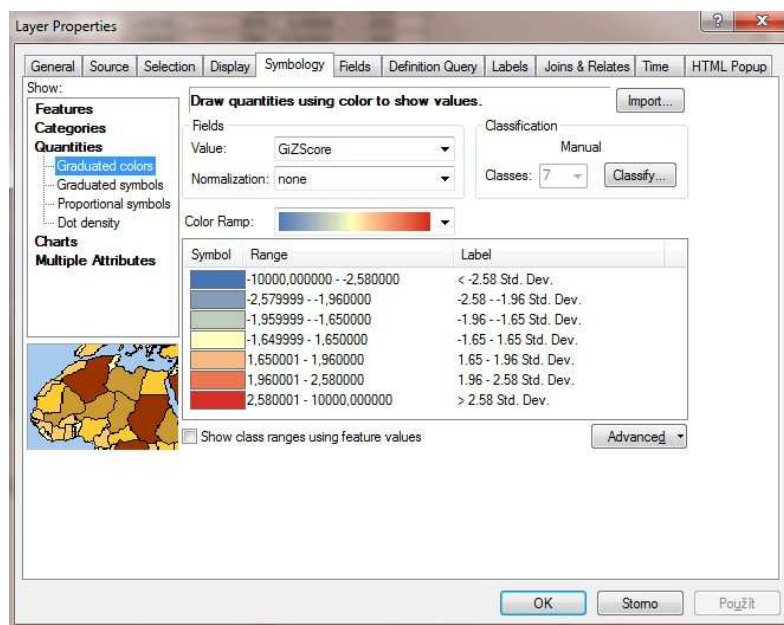
Metoda nabízela několik volitelných parametrů. Prvním parametrem, u kterého bylo třeba rozhodování, byl parametr specifikující prostorové vztahy mezi prvky. Po konzultaci byla zvolena metody konstantní vzdálenosti, která je i všeobecně doporučována jako vhodná metoda (Anselin, 1995). Tato metoda je založená na faktu, že všechny prvky jsou analyzovány v kontextu se sousedními prvky. Sousedící prvky, které se nachází v určité prahové vzdálenosti (další volitelný parametr), mají vliv na výpočet prvku. Hodnoty prvků za tímto prahem vliv nemají. V našem případě byla prahová hodnota vypočítána na základě velikosti vstupní třídy a počtu prvků (ESRI, 2012).

Další parametr určuje, jak jsou vypočítávány vzdálenosti z každého prvku k nejbližšímu prvku. Přímou vzdálenost mezi dvěma body určuje Euklidovská vzdálenost a vzdálenost mezi dvěma body měřenou podél osy v pravém úhlu určuje Manhattan vzdálenost. Byla zvolena přímá, tedy Euklidovská vzdálenost. Další parametry byly volitelné, a nebylo třeba je nastavovat. Výsledek metody Getis-Ord G_i^* je znázorněn na obrázku 25.



Obr. 25: Shluková analýza výskytu infekce B01 (plané neštovice) v roce 2002.

Byly vypočítány shlukové analýzy pro všech devět sledovaných jevů EPIDAT na úrovni uliční sítě. Výstupem byly hodnoty z-score a p-value přidáné do atributové tabulky. Hodnota z-score je přímo výstupem analýzy Getis-Ord G_i^* a její hodnota je znázorněná. Byla použita výchozí stupnice, která obsahovala sedm kategorií. Z důvodu velkého množství výstupních vrstev a přesnosti výsledků při změně kategorií nebyl shledán důvod, proč původní stupnici měnit.



Obr. 26: Klasifikace stupnice shlukové analýzy.



Obr. 27: Pojmenování legendy.

Výsledkem shlukové analýzy bylo 68 vrstev, které byly následně použity jako podklady pro mapové výstupy a vizualizovaný jako soubory ve formátu KML prostřednictvím aplikace Google Earth. Soubory SHP a projekty jsou přiloženy na DVD.

4.6 Vizualizace výstupů

Posledním bodem práce bylo znázornění získaných výstupů. Výsledky byly po konzultaci zobrazeny formou mapových výstupů a formou KML souborů. Byl brán zřetel na cílové uživatele KHS, a proto byla zvolena hlavně přehlednost a jednoduchost.

Celkem bylo vytvořeno šest mapových výstupů. Z toho čtyři formátu A2 a dva formátu A3. Formátu A2 bylo zvoleno při vizualizaci jevů, kde bylo zapotřebí na jednom mapovém listě zobrazit více menších map. V našem případě se jednalo o desetileté období nebo zobrazení 12 měsíců v roce. Formátu A3 bylo zvoleno pro věkovou strukturu ve třech kategoriích a školskou strukturu ve čtyřech kategoriích. Všechny výstupy byly zpracovány v programu ArcGIS.

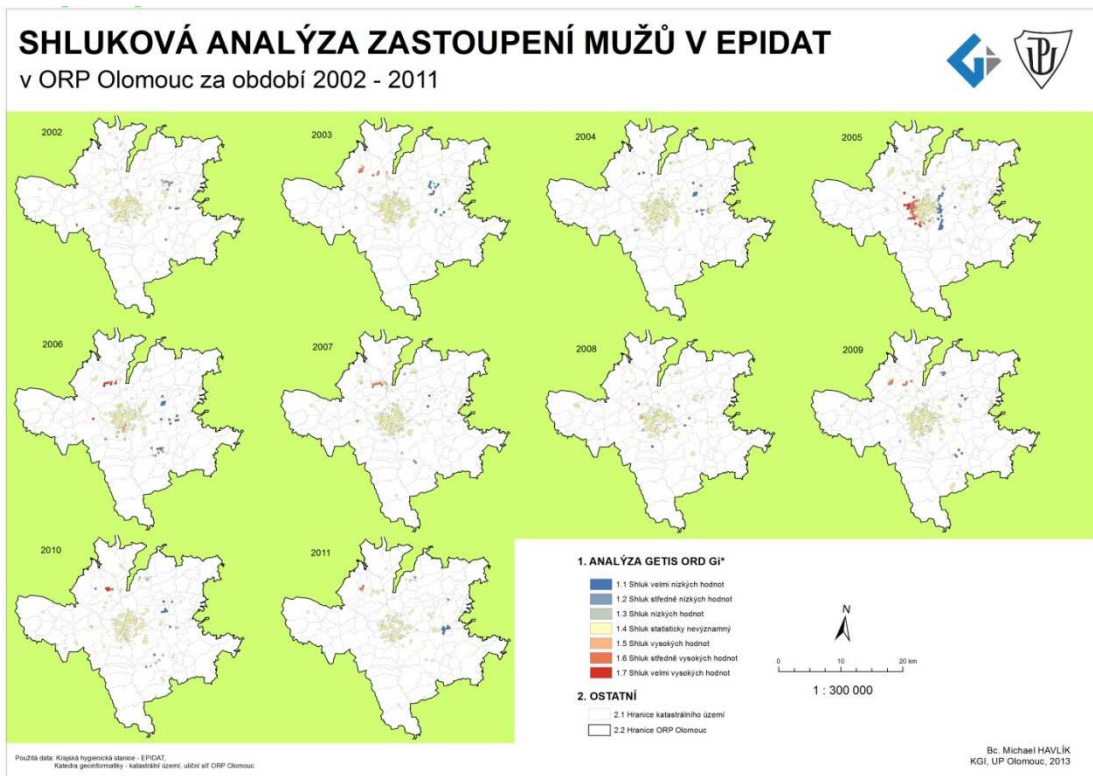
Výstupy formátu A2 jsou:

- Shluková analýza zastoupení mužů v EPIDAT v ORP Olomouc za období 2002–2011
- Shluková analýza zastoupení žen v EPIDAT v ORP Olomouc za období 2002–2011
- Shluková analýza planých neštovic v EPIDAT v ORP Olomouc za období 2002–2011
- Shluková analýza planých neštovic v měsících v ORP Olomouc za období 2002–2011

Výstupy formátu A3 jsou

- Shluková analýza věkové struktury v EPIDAT v ORP Olomouc za období 2002–2011
- Shluková analýza školské struktury v EPIDAT v ORP Olomouc za období 2002–2011

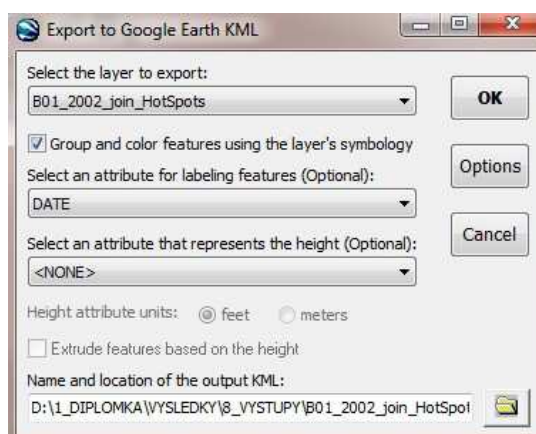
Všechny mapové výstupy byly zvoleny v jednotném měřítku 1 : 300 000. Jako podkladová data byla použita hranice ORP Olomouc a hranice k.ú. Součástí map jsou i loga UP a KGI. Pod hlavním mapovým polem jsou poskytovatelé vstupních dat. Výstupy jsou volnou přílohou práce a jsou přiloženy na DVD. Ukázka mapového výstupu je na obrázku 28.



Obr. 28: Mapový výstup formátu A2.

Výsledným vrstvám shlukových analýz byl přidán atribut s časovou značkou výskytu. Pro převod dat do formátu KML byla použita extenze Export to Google Earth KML (obr. 29). Původně byl použit nástroj Layer to KML v nástrojové sadě Conversion Tools. Ten exportoval data do formátu KMZ a při importu těchto dat do aplikaci Google Earth nebyl program schopný identifikovat časovou značku. Extenze však tento problém vyřešila a data byla úspěšně importována i s možností zobrazení časové osy a animace.

V aplikaci Google Earth byla vrstvám nastavená průhlednost a zobrazeny pouze vhodné vrstvy tak, aby byly výsledky co nejvíce čitelné.



Obr. 29: Extenze Export to Google Earth KML.

Na obrázku 30 je zobrazená importovaná KML vrstva nad ortofotem v aplikaci Google Earth. Celkem bylo vytvořeno 71 KML souborů ze šesti sledovaných jevů. Tyto soubory jsou součástí DVD. Seznam všech KML souborů je patrný z tabulky č. 3.

Tab. 3: Seznam všech jevů s vytvořenými KML soubory

POČET KML	NAZEV JEVU	PODSKUPINA
12	ZASTOUPENI B01 V MESÍCÍCH	LEDEN - PROSINEC
5x10	ZASTOUPENI - A02, A69.2, B01, MUZI, ZENY	2002 - 2011
4	SKOLSKA STRUKTURA	MŠ, ZŠ, SŠ, VŠ
3	VEKOVE ZASTOUPENI	0-14, 15-59, 60 +
2	POHLAVNI ZASTOUPENI	MUŽI, ŽENY



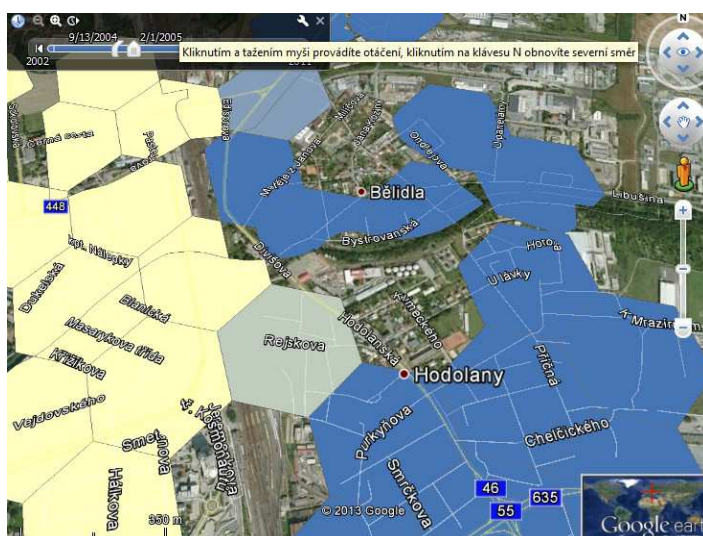
Obr. 30: KML soubor v aplikaci Google Earth

4.7 Zhodnocení výsledků shlukových analýz

Poslední kapitolou diplomové práce je slovní zhodnocení výsledků shlukových analýz. Zhodnocení v podstatě otevírá diskusi, proč jsou výsledky analýz v jednotlivých letech a měsících různé a kde byla příčina sníženého či zvýšeného výskytu. Autor práce se nezabýval podrobným zkoumáním příčin výskytů. Vzniká zde tedy prostor pro konzultaci s odborníky.

V následujícím textu bude popsán průběh šesti sledovaných jevů, které byly znázorněny na mapových výstupech práce.

Škála hodnocení byla zvolena velmi jemně a je tedy možné přesněji identifikovat oblasti. Pojmenování intenzity shluků bude podle klasifikace shluků na obrázku 27 v kapitole 4.5. K detekci shluků dobře posloužila aplikace Google Earth s využitím podkladových vrstev, ortofota a časové animace (obr. 31). Je třeba dodat, že velkou měrou se na datech projevila infekce planých neštovic (Varicella B01), která měla v datech dvoutřetinové zastoupení.



Obr. 31: Nalezení shluků velmi nízkých hodnot v aplikaci Google Earth

Shluková analýza zastoupení mužů a žen v EPIDAT v ORP Olomouc za období 2002–2011

V širším měřítku bylo zastoupení mužů a žen v datech v jednotlivých letech velmi podobné. Žádný rok nevykazoval nějakou výraznou změnu. Na výsledcích je patrné, že infekce planých neštovic měla opravdu velký vliv na výsledky. Pokud se detailně zaměříme na zastoupení, zjistíme, že v roce 2002 měly ženy výrazně větší podíl shluků vysokých hodnot než muži. V roce 2005, kdy bylo shlukování nejintenzivnější, měly ženy také větší zastoupení shluků velmi vysokých hodnot v západní části města Olomouce. V roce 2009 bylo vysokých i nízkých shlukování více v kategorii mužů. Obecně můžeme říci, že výraznějších hodnot bylo zjištěno v kategorii žen a že shluky

vysokých hodnot byly převážně v západní části území ORP Olomouc a shluky nízkých hodnot ve východní části.

Shluková analýza planých neštovic v EPIDAT v ORP Olomouc za období 2002–2011

V globálu za desetileté období převládaly u této infekce shluky vysokých hodnot v západní části území ORP Olomouc. Naopak shluky nízkých hodnot se vyskytovaly ve východní části. Nejvýznamnější byl rok 2005, kdy byl extrémní výskyt této infekce. Zejména v Olomouci byly velmi vysoké shluky v jižní a západní části města (oblasti Nové Sady, Nová Ulice, Neředín a také Hejčín). V těchto částech je velká koncentrace sídlišť, kde je nižší věková struktura než v jiných částech města. Výsledky tedy potvrdily předpoklad vyššího výskytu planých neštovic u dětí předškolního věku. Nízká shlukovost byla ve východní části města (oblast Hodolany, Bělidla a Pavlovičky). Roky 2003–2004 a 2006–2007 vykazovaly obdobné shlukování. Vysoká shlukovost byla v těchto letech v obci Skrbeň, naopak velmi nízká v obcích Velký Týnec a Samotišky. V letech 2008–2009 byly velmi vysoké shluky v obcích Nemilany a Slavonín, naopak nízké ve Chvalkovicích a Holicích. V letech 2002, 2008 a 2011 byly shluky statisticky nevýznamné.

K detekci shluků vhodně posloužila animace v aplikaci Google Earth.

Shluková analýza planých neštovic v měsících v ORP Olomouc za období 2002–2011

Z této shlukové analýzy bylo jednoduché odhadnout, v jakých měsících se infekce vyskytovala nejvíce. Měsíce únor až březen vykazovaly mnohem větší hodnoty shlukování než měsíce ostatní. Vůbec nejsilnější byl měsíc únor. Je zajímavé sledovat, jak se shluky šířily z centra do okolních obcí. V prvních měsících byly shluky vysokých a nízkých hodnot především ve městě Olomouc. Nejslabší měsíce byly měsíce srpen, září a říjen. Od listopadu do konce dubna se výskyt nepatrně zvyšovaly. Zajímavý byl měsíc červenec, kdy náhle shluky dosáhly vysokých hodnot v obcích Nemilany, Nedvězí a Topolany.

Shluková analýza věkové struktury v EPIDAT v ORP Olomouc za období 2002–2011

Při sledování věkové struktury byly sledovány tři skupiny (0–14 let, 15–59 let a 60 a více let).

První skupina 0–14 let vykazovala shluky velmi vysokých hodnot v západní části města (části města Nová Ulice a Neředín). Byly nalezeny spousty statisticky nevýznamných shluků jak v Olomouci, tak v přilehlých obcích. Shluky nízkých hodnot byly detekovány jen velmi zřídka.

Pro skupinu 15–59 byly výrazné shluky nalezeny mimo Olomouc. Velmi výrazné shluky vysokých hodnot byly v obci Lutín a Štěpánov, vysoké shlukování bylo také v obcích Grygov a Horka nad Moravou. Naopak velmi nízké hodnoty byly v obcích Hlubočky a Majetín.

V kategorii 60 a více let, tedy u lidí důchodového věku, bylo nalezeno méně shluků než v předchozích skupinách. Velmi vysoké shluky byly opět v obci Lutín a velmi nízké v obcích Hlubočky a Majetín. Statisticky nevýznamné shluky byly v městě Olomouci a ve Velkém Týnci.

Shluková analýza školské struktury v EPIDAT v ORP Olomouc za období 2002–2011

Byly sledovány čtyři skupiny. Skupina MŠ (mateřská školka) vykazovala největší zastoupení prostorových shluků velmi vysokých hodnot v západní části města Olomouce (oblasti Nová ulice, Povel, Neředín) a naopak velmi nízké shlukování ve východní části Olomouce (Hodolany, Bělidla).

Další skupina ZŠ (základní škola) už nevykazovala tak výrazné hodnoty. Na území Olomouce byly statisticky nevýznamné shluky. Vysoké hodnoty se objevily v oblastech obcí Nedvězí, Slavonín a Nemilany, naopak nízké hodnoty v oblasti Svatého Kopečku a Samotišek.

Skupina SŠ zaznamenala středně vysoké shluky v oblastech Neředín a Řepčín, naopak středně nízké shluky byly v Hodolanech a Holicích.

Kategorie VŠ (vysoká škola) už tolik shluků neobsahovala. Byly zjištěny jen opravdu malé výskyty zvýšených hodnot v obcích Horka nad Moravou a Topolany. Centrum Olomouce vykazovalo shluky nevýznamné.

Obecně lze říci, že čím starší populace, tím méně záznamů a menší výskyt shluků. Také se zde ukázalo, že čím starší obyvatelstvo, tím více shluků mimo centrum města Olomouce.

5 VÝSLEDKY

Teoretická část práce se zabývala současným stavem řešené problematiky. Bylo vybráno několik zahraničních a českých studií, které využily lokálních statistických metod v posledních letech. Většina studií byla z oblasti zdravotnictví, kriminalistiky a životního prostředí. Cílem teoretické části bylo zejména přiblížit použití lokálních statistických metod v praxi, což bylo splněno.

Důraz byl kladen především na praktické výsledky. Ty budou následně členěny do několika menších sekcí odsazených odstavci. Celkem bude popsáno pět hlavních výsledků, které jsou také hierarchicky popsány v kapitole 4.

Prvním dílčím výsledkem byla aktualizace dat databáze EPIDAT. Pomocí programu pro poloautomatickou opravu dat, který byl vytvořený v bakalářské práci, na niž tato práce navazovala, byla opravena data za období 2002–2003 a 2009–2011. Data byla přidána k původním datům a výsledkem byl ucelený soubor dat pro Olomoucký kraj obsahující 46 190 opravených záznamů za desetileté období 2002–2011. Pro úroveň shlukových analýz byla zvolena oblast ORP Olomouc. Před exportem dat byla data vyfiltrována pro tuto oblast. Následně byla data pomocí agregačních dotazů vyexportována do tabulek, které byly vstupními daty pro prostorové analýzy. Celkem bylo vyexportováno 11 tabulek ve formátu XLS.

Druhým výsledkem byla popisná charakteristika vstupních dat. Podrobně byla popsána vstupní data EPIDAT a vytvořena řada tabulek a grafů, které popisovaly průběh infekcí v jednotlivých letech a zastoupení v souboru dat. K podrobnému popisu byla vybrána Hepatitida A. Dále byly vytvořeny grafy výskytu infekce v jednotlivých měsících pro Olomoucký kraj i pro zastoupení v Olomouci. Zastoupení Hepatitidy A bylo také porovnáno s daty věkových skupin z ČSÚ.

Třetím výsledkem bylo vytvoření obecné metodiky pro předzpracování dat shlukových analýz. Pomocí nástrojů a extenzí v programu ArcGIS bylo provedeno několik operací. V prvním kroku byla přiřazena prostorová složka vstupním datům EPIDAT pomocí propojení dat s uliční sítí ORP Olomouc. Následně byla data rozložena do k.ú. a uliční síť pomocí nástroje *Create Random Points*. Pro vytvoření geometrické sítě nad daty byla použita extenze *Repeating Shapes for ArcGIS*. Za pomoci dat ZSJ pro k.ú. byla data následně standardizovaná s využitím nástroje *Identity*. Celkem bylo vytvořeno 91 výstupních vrstev formátu SHP pro 11 sledovaných jevů. Tyto data sloužily jako vstupní data pro shlukové analýzy.

Před provedením shlukových analýz bylo provedeno testování náhodnosti rozmístění bodů výskytu. Pomocí interferenční statistiky byly vybrány dvě metody nástrojové sady *Spatial Statistics* v programu ArcGIS. Tyto metody rozhodly, zda distribuce použitých výskytů vykazuje prostorové shlukování, rozptyl nebo zda jsou uspořádána náhodně. Obě zvolené metody vykazaly prostorové shlukování v datech, ačkoli jsou každá založená na jiných principech. První metoda *Average Nearest Neighbour* (Průměrný nejbližší soused) vypočítala poměr (ratio) nejbližšího souseda založený na průměrné vzdálenosti jednotlivých prvků od nejbližšího sousedního prvku. V případě, že poměr je menší než 1,

vzorek vykazuje shlukování, v opačném případě vykazuje disperzi (ESRI, 2012). Poměr vyšel 0,22, tudíž data vykazovala shlukování. Druhá metoda Ripleyova K-funkce byla ideální pro zkoumání rozptylu či shlukování v různých vzdálenostech. Zkoumala se vzdálenost 6 km od každého bodu. Přírůstky byly nastaveny po 200 m a dále byl vypočítán interval spolehlivosti. Data vykazovala výrazné shlukování, které bylo potvrzeno i u předchozí metody. Také se zjistilo, že zhruba po třetím kilometru se počet shluků ustaluje.

Posledním výsledkem byly výstupy ze shlukových analýz a následná vizualizace. Pomocí metody Getis-Ord G_i^* v nástrojové sadě Spatial Statistics programu ArcGIS se vypočítala shluková analýza devíti sledovaných jevů. Dva jevy nebyly pro nízký počet výskytů znázorněny. Výstupem byly hodnoty z-score a p-value přidané do atributové tabulky. Výsledkem bylo 68 vrstev, které byly následně použity jako podklady pro mapové výstupy a dále také byly znázorněny ve formátu KML prostřednictvím aplikace Google Earth. Celkem bylo vytvořeno šest mapových výstupů formátu A2 a A3, které jsou volnou přílohou práce. Výsledným datům byla přidána datová značka a pomocí extenze Export to Google Earth bylo vytvořeno 68 souborů formátu KML. V aplikaci Google Earth je možné data zobrazit nad ortofotem a vytvořit časovou animaci.

Závěrem bylo provedeno slovní zhodnocení výsledků analýz, kde byl popsán průběh všech sledovaných jevů.

6 DISKUSE

V této práci byla řešena spousta problémů a rozhodování. Některé problémy byly vyřešeny bez potíží, jiné byly vyřešeny jiným, mnohdy pracnějším postupem.

Začneme-li od vstupních dat, lze říci, že není možnost jejich 100% opravy. Při poloautomatických opravách je třeba někdy rozhodovat podle vlastního uvážení, např. pokud obec spadá do více okresů a nelze ji specifikovat ani přidáním ulic či jiným atributem. Zde se můžeme domnívat, že pravděpodobně spadá do okresu v Olomouckém kraji. Není zde vyloučena i lidská chyba, byť bylo celkem opraveno přes několik tisíc ulic. Samozřejmě chyba je na straně zadavatelů dat a provozovateli programu EPIDAT, který by mohl vytvořit omezení pro zadávání vstupních atributů.

V kapitole předzpracování dat bylo potřeba často rozhodovat o použitých metodách a parametrech. Původně bylo v plánu použít nástroje pro geokódování k přiřazení prostorové složky. Zjistilo se, že výstupem nástroje Geocode Addresses je vždy jen bodová vrstva. Ta následně nešla použít jako vstupní parametr o počtu výskytů do nástroje Create Random Points. Ten požadoval pouze vrstvu liniovou a polygonovou. Byl tedy použit klasický nástroj pro propojení tabulek Join Data.

Výstupní data nástroje Create Random Point neměla definována souřadnicový systém. Nástroj nedokázal přebrat systém vstupních dat a musel být přidán ručně. Taktéž při standardizaci dat bylo zapotřebí ručního přidání atributů pro výpočet počtu obyvatel v mřížce. V co největší míře byl použit nástroj ModelBuilder, případně volba Batch u zvolené metody. Nešlo tedy vytvořit kompletní model pro vstupní data obsahující všechny použité nástroje. Není vyloučeno, že neexistovaly v jednotlivých krocích efektivnější nástroje.

Při tvorbě geometrické sítě byly vybrány z celé řady tvarů právě hexagony. Tento tvar byl zvolen jako vhodný, protože umožňoval přístup k co největšímu počtu sousedů. Jelikož jsou lokální metody shlukových analýz založeny na výpočtu vzdáleností mezi body, byla volba hexagonu nejlepší možnou variantou výběru. Úhel natočení mřížky byl vybrán náhodně a velikost se vypočítala z mediánu velikosti uliční sítě v ORP Olomouc. Myslím, že zde bylo vše zvoleno správně a výsledky jsou tedy v té nejvyšší možné míře vypovídající.

Pro popisnou charakteristiku vstupních dat byly použity metody programu ArcGIS. Nebyly využity funkce dalších známých statistických software – R project, STATISTICA. Program ArcGIS nabízel vhodné znázornění výsledků těchto analýz a použité metody nejbližšího souseda a výpočtu K-funkce byly dostačující.

Použitá metoda Getis-Ord G_i^* v programu ArcGIS je hojně používaná pro lokalizaci shluků zdravotnických dat, proto nebylo třeba hledat nové nástroje v jiných statistických softwarech. Pro shlukové analýzy lze například použít software GeoDA, který byl vyvinutý pod vedením Luca Anselina, tvůrce lokální metody LISA. Program ArcGIS není statisticky založený program, ovšem pro tuto práci byl zvolen jako vhodný.

Oblast ORP Olomouc byla na základě získání dat uliční sítě zvolena oblastí působení shlukových analýz. Do budoucna tedy lze vytvořit shlukové analýzy pro území Olomouckého kraje.

Výsledky je možné samozřejmě využít i v budoucnu. Určitý námět na pokračování by mohl být ve formě vizualizace získaných dat. Velmi zajímavé zpracování formou statických, dynamických, interaktivních a google maps výstupů bylo shledáno na adrese <http://zdravi.geogr.muni.cz/index.php>. Jedná se o webový portál VisualHealth, který si klade za cíl ukázat možnosti použití kartografických metod a znázornění zdravotnických dat.

Možným rozšířením je i tvorba korelací s věkovou strukturou obyvatelstva. Zde je možné použít věkového složení obyvatelstva pro jednotlivé roky na úroveň okresů poskytnuté ČSÚ.

7 ZÁVĚR

Tato magisterská práce si kladla za hlavní cíl identifikovat oblasti se zvýšeným výskytem vybraných infekcí pomocí prostorových analýz.

Pro splnění hlavního cíle byly postupně řešeny cíle dílčí. Byla provedena aktualizace dat databáze EPIDAT a vytvořen ucelený soubor dat za období 2002-2011. Z tohoto souboru byly vytvořeny pomocí agregačních dotazů tabulky, které dále vstupovaly do prostorových analýz.

Byla sepsána popisná charakteristika vstupních dat, zejména vstupních dat EPIDAT. Výsledky se znázornily především formou tabulek a grafů. K detailnímu popisu byla vybrána Hepatitida A, která vykazovala nepravidelný průběh. Proběhlo zde porovnání infekce s daty ČSÚ.

Před vstupem dat do shlukových analýz se provedla série operací pomocí nástrojů a extenzí v programu ArcGIS. V jednotlivých krocích byla vytvořena obecná metodika pro předzpracování dat.

Pro testování náhodnosti rozmístění bodů výskytu byly použity dvě metody interferenční statistiky, a to Average Nearest Neighbour (Průměrný nejbližší soused) a Ripleyova K-funkce. Výsledky těchto metod nám objasnily fakt, že distribuce použitých dat vykazují prostorové shlukování.

K identifikaci shluků bylo využito lokální statistické metody Getis-Ord G_i^* v nástrojové sadě Spatial Statistics programu ArcGIS. Provedla se shluková analýza devíti sledovaných jevů na úroveň uliční sítě. Výsledkem bylo 68 vrstev, které byly následně použity jako podklad pro mapové výstupy.

Výsledné vrstvy byly exportovány do formátu KML a znázorněny v aplikaci Google Earth, kde byly vytvořeny časové animace.

Závěrem bylo provedeno slovní zhodnocení výsledků analýz s popsáním průběhů sledovaných jevů.

POUŽITÁ LITERATURA A INFORMAČNÍ ZDROJE

Knihy a jiné zdroje:

1. ANSELIN, L. *Local indicators of spatial association – LISA*, In: Geographical Analysis, Ohio State University Press, 1995, 27, s. 93-115.
2. CLAIBORN L. *Quantitative School Bus Stop Risk Assessment of Redlands, California*. [online]. 2005 [cit. 2013-04-10]. Dostupné z: http://gis.cityofredlands.org/Claiborn_Phillips_Quantitative_School_Bus.pdf
3. Disease mapping and spatial epidemiology: *Introduction*. ZevRoss *Spatial Analysis* [online]. 2008 [cit. 2013-04-09]. Dostupné z: http://www.zevross.com/special/special_intro.html
4. ESRI (2012) *Webová nápověda pro ArcGIS Desktop* [online]. 2012. vyd. [cit. 2013-03-10]. Dostupné z: http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#
5. FRYE CH. *Hot Spot Analysis of 911 Calls map*. [online]. 2007 [cit. 2013-04-10]. Dostupné z: <http://mappingcenter.esri.com/index.cfm?fa=maps.hotSpot911>
6. GETIS, A., ORD J. K. *The Analysis of Spatial Association by Use of Distance Statistics*, In: Geographical Analysis, Ohio State University Press, 1992 24, s. 286-306.
7. Hepatitida A. *Virová-hepatitida.cz* [online]. 2013 [cit. 2013-04-12]. Dostupné z: <http://www.virova-hepatitida.cz/hepatitida-a>
8. HAVLÍK, Michael. *Průvodce geokódováním zdravotnických dat databáze EPIDAT*. Olomouc, 2010. Bakalářská práce. Univerzita Palackého Olomouc. Dostupné z: <http://www.geoinformatics.upol.cz/dprace/bakalarske/havlik10/>.
9. HORÁK, J. (2008): *Prostorové analýzy dat*. VŠB-TU Ostrava, HGF, Institut geoinformatiky, 2. vydání, Ostrava.
10. KLÍCHA, Martin. *Shluková analýza funkčních ploch pro sledování urbánního vývoje Olomouckého regionu*. Olomouc, 2011. Bakalářská práce. UP Olomouc. Dostupné z: <http://www.geoinformatics.upol.cz/dprace/bakalarske/klichal1/>.

11. LAWSON, B. *Statistical Methods in Spatial Epidemiology*. WILEY, 2006. ISBN 978-0-470-01484-4.
12. Lai PC et al. *Understanding the Spatial Clustering of Severe Acute Respiratory Syndrome (SARS) in Hong Kong*. [online]. 2004 [cit. 2013-04-10]. Dostupné z: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1247620/>
13. MAREK, L., PASZTO, V., & TUČEK, P. (2012). *Spatial Analyses of Epidemiological Data: Case Study In Olomouc Region*. 12th International Multidisciplinary Scientific GeoConference SGEM: SGEM 2012, Proceedings Volume II (pp. 1155 – 1162). Sofia, Bulgaria: STEF92 Technology Ltd. doi:10.5593/sgem2012, ISSN: 1341-2704
14. Prohlížeč ÚIR-ZSJ [online]. 2013 [cit. 2013-03-24]. Dostupné z: http://www.czso.cz/csu/rso.nsf/i/prohlizec_uir_zsj
15. SCOTT L., WARMERDAM N. *Extend Crime Analysis with ArcGIS Spatial Statistics Tools*. [online]. 2012 [cit. 2013-04-10]. Dostupné z: http://www.esri.com/news/arcuser/0405/ss_crimestats1of2.html
16. SPURNÁ, P. *Prostorová autokorelace*. Sociologický časopis. 2008, č. 44, s. 767 - 787. Dostupné z: http://sreview.soc.cas.cz/uploads/6097969a40937f30519a0d976493521f0469993b_516_2008-4Spurna.pdf
17. ŠTĚPÁNOVÁ, Radka. *Disease mapping*. Olomouc, 2008. Bakalářská práce. Masarykova univerzita. Dostupné z: < http://is.muni.cz/th/184531/prif_b/Disease_mapping.pdf>.
18. VLOSINSKÝ, Jakub. *Prostorové analýzy časových řad volebních výsledků*. Olomouc, 2011. Diplomová práce. UP Olomouc. Dostupné z: < <http://www.geoinformatics.upol.cz/dprace/magisterske/vlosinsky11/>>.

SUMMARY

This master thesis aims to use spatial analysis to identify areas where increased occurrence of certain diseases has been identified.

In the theoretical part, several foreign and Czech studies that used local statistical methods in recent years were selected. Most of them were from health area, criminology and environment. The goal of the theoretical part was to demonstrate practical usage of local statistical methods.

The practical part included several partial results. In the first section input data were updated from the EPIDAT database. As a result, input data were completely corrected from the year 2002 to 2011 by semi-automatic corrections in the MS Access database program.

The second part described input data characteristics. Significant number of tables and graphs were created to describe the process of infections throughout the years. Hepatitis A was selected for the detailed description. Graphs of occurrence of the infections in each month were created for the Olomouc region and for representation in Olomouc city. The representation of Hepatitis A was compared with the age groups from the Czech Statistical Office.

It was followed by a general methodology for preprocessing of data that entered the cluster analysis. Substantial number of tools and extension of the ArcMap program have been used. The first step was to assign the spatial component to the input data of the EPIDAT. The data were then divided according to the cadastral districts and the street network by using *Create Random Points* tools. For creating geometric network of the data the Repeating Shapes extension for ArcGIS was used. The data were subsequently standardized by using the data of the ZSJ for cadastral districts with using the Identity tool. There were created altogether 91 output layers in the SHP format for 11 observed phenomena. These data were used as the input data for cluster analysis.

Another aim was to test the point distribution randomness. The interference statistics was used to selected two methods from the *Spatial Statistics* toolbox in the program ArcGIS (*Average Nearest Neighbour and Ripleys K-function*). These methods are based on different principles and they were used to decide if the point distribution of the used occurrences showed the spatial clustering, dispersion or random arrangement. Both methods confirmed significant clustering.

*Getis-Ord Gi** method was used for the cluster analysis in the Spatial Statistics toolbox of the program ArcGIS. This method gives a set of weighted features identifying statistically significant hot and cold spots. The output values z-score a p-value were added to the attribute table. As a result, there were 68 layers that were afterwards used as the basis for the map outputs. They were displayed in the KML format provided by Google Earth application.

In total, there were created 6 map outputs in the format A2 and A3 which are attached to the thesis. The date tags were added to the result data. Using the Export to Google Earth extension there were created 68 files in format KML. It is possible to display the data over the orthophoto and create an animation in the Google Earth application. Written evaluation of the map outputs concluded the thesis.

PŘÍLOHY

SEZNAM PŘÍLOH

Volné přílohy

Příloha 1–6 Shluková analýza dat databáze EPIDAT v ORP Olomouc za období 2002–2011

Příloha 1 Zastoupení mužů

Příloha 2 Zastoupení žen

Příloha 3 Plané neštovice

Příloha 4 Plané neštovice v měsících

Příloha 5 Věková struktura

Příloha 6 Školská struktura

Příloha 7 DVD

Popis struktury DVD

Adresáře:

DP – text diplomové práce ve formátu pdf

www_stranky – vytvořené webové stránky práce

vystupni_data – kml, mxd, shp, pdf

Poznámka: Původní data EPIDAT i opravená data se nenacházejí na DVD ROM, neboť jsou to data neveřejná, která byla poskytnuta pouze pro účely zpracování. Katedra geoinformatiky PřF UP se zavázala, že nebude data žádným způsobem šířit ani zveřejňovat, tudíž nemohou být umístěna v příloze bakalářské práce. Výsledná data byla předána vedoucímu práce.