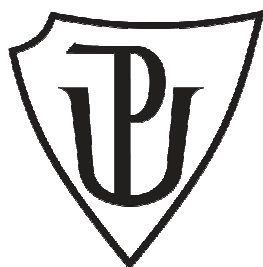


UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA GEOINFORMATIKY



Michael HAVLÍK

**PRŮVODCE GEOKÓDOVÁNÍM
ZDRAVOTNICKÝCH DAT DATABÁZE
EPIDAT**

Bakalářská práce

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

Olomouc 2010

Prohlašuji, že jsem zadanou práci řešil sám a že jsem uvedl všechnu použitou literaturu. Získaná vstupní data nebudu bez souhlasu Krajské hygienické služby Olomouc poskytovat třetí osobě.

Olomouc, 25. května.2010

.....
Podpis

Děkuji vedoucí práce Ing. Zdeně Dobešové, Ph.D. a konzultantovi Mgr. Janu Harbulovi za pomoc a věcné připomínky při řešení této práce.

OBSAH

ÚVOD	- 7 -
1. CÍL PRÁCE.....	- 8 -
2. ÚVODNÍ ZPRACOVÁNÍ	- 9 -
2.1. Použitá data.....	- 9 -
2.1.1. EPIDAT	- 9 -
2.1.2. UIR-ADR	- 10 -
2.1.3. Český statistický úřad	- 11 -
2.2. Pracovní prostředí	- 12 -
2.3. Epidemiologie	- 13 -
2.4. Informační technologie ve zdravotnictví.....	- 14 -
3. VÝSLEDKY.....	- 17 -
3.1. Předzpracování dat.....	- 17 -
3.1.1. Úprava dat	- 17 -
3.2. Analýza EPIDAT	- 19 -
3.3. Popis EPIDAT.....	- 21 -
3.4. Identifikace chyb.....	- 23 -
3.4.1. Chyby v EPIDAT	- 26 -
3.5. Tvorba a popis programu pro opravu chyb a geokódování....	- 29 -
3.5.1. Průvodce opravou EPIDAT	- 30 -
3.5.2. Průvodce geokódováním	- 33 -
3.5.3. Závěrečné shrnutí programu	- 34 -
3.6. Výpočet definičních bodů obcí pro následné geokódování....	- 35 -
3.6.1. Elipsa chyb.....	- 36 -
3.7. Geokódování	- 39 -
3.7.1. Tvorba průvodce geokódováním v ArcGIS Desktop	- 39 -
4. DISKUZE	- 41 -
5. ZÁVĚR.....	- 42 -
6. POUŽITÉ ZDROJE.....	- 43 -
SUMMARY.....	- 45 -
PŘÍLOHY	- 46 -

ÚVOD

V dnešní době 21. století si jen málokdo z nás dokáže představit život bez počítače. Tato věc, která se bezesporu stala historickým mezníkem v historii lidstva, nám dokáže usnadnit spoustu práce a času. Přeci jenom jsou i tací, kteří si odmítají usnadnit práci. Mluvím o starší populaci tedy lidech v důchodovém věku. Na jednu stranu je to určitě škoda, na druhou stranu je musíme respektovat.

Ve zdravotnictví se nachází mnoho doktorů ve starším věku, kteří odmítají zapisovat věci do digitální podoby a raději volí osvědčenou kartotéku. Tímto narážím na stěžejní část mé bakalářské práce, ve které se budu věnovat opravě chybné epidemiologické databáze, dále jen EPIDAT. Právě tato databáze obsahuje mimo jiné mnoho chyb zaviněných právě přepisem těchto spisů do počítače.

Díky vzájemné spolupráci s Krajskou hygienickou stanicí v Olomouci, máme k dispozici data z této databáze v rozsahu pro Olomoucký kraj. Existuje mnoho manuálních postupů opravy dat, ale pro řešení problému s 24 000 záznamy je automatický postup téměř nutností. Tento postup by měl být dále použitelný i na větší rozsah dat (pro ČR).

Za použití databázových programů a programovacích jazyků lze s využitím celostátně platné databáze Územně identifikačního registru adres (ÚIR - ADR) nalezené chyby vyhledat a nahradit je správnými daty.

V druhé části této práce je využito spojení s geografickými informačními systémy v procesu geokódování. Jedná se o proces, pomocí kterého se databázovým záznamům přiřadí zeměpisné souřadnice, pomocí kterých lze záznam lokalizovat. Z databáze tak vznikne vektorová (bodová) mapová vrstva, se kterou je již možné plnohodnotně pracovat[6]

1. CÍL PRÁCE

Cílem této bakalářské práce je oprava chybných adres databáze EPIDAT a následné přiřazení geometrické složky těmto datům pro potřeby provádění prostorových analýz.

Nejprve je nutné správně analyzovat data a identifikovat chyby. Za použití celostátně platného Územně identifikačního registru adres, dále jen ÚIR – ADR, porovnat tyto databáze a nahradit chybná data správnými. Vše za pomoci databázových programů a programovacího jazyka. Jedním z prvních cílů této práce je vytvoření automatizovaného postupu opravy případně nahrazení těchto dat.

Dalším krokem je v procesu geokódování propojit tyto data s GIS. V tomto procesu přiřadíme ulicím, obcím a okresům zeměpisné souřadnice. Přesná epidemiologická data s lokalizací až na adresní body nebyla v tomto případě poskytnuta.

Závěrem budou vytvořeny internetové stránky, které budou umístěny na serveru katedry.

2. ÚVODNÍ ZPRACOVÁNÍ

2.1. Použitá data

V následující části jsou popsány nejdůležitější zdroje dat této bakalářské práce. Stěžejní databáze EPIDAT poskytnutá Krajskou hygienickou stanicí v Olomouci obsahuje 24 000 záznamů v rozsahu pro Olomoucký kraj a časovém rozpětí v letech 2004 – 2008. Každý záznam obsahuje detailní zápis pacienta, který prodělal epidemiologické onemocnění. Záznam je poskytnut beze jména, příjmení a rodného čísla pacienta a obsahuje mimo jiné i ulici, obec, místní část, okres, datum hlášení, datum přijetí, týden onemocnění, diagnózu, místo nákazy, místo onemocnění, okres nákazy, místo izolace, kolektiv, zaměstnání a další. Stěžejními atributy jsou ulice, obec a okres. V uvedených attributech také nacházíme nejvíce chyb vlivem manuálního dopisování do databáze.

UIR-ADR můžeme považovat za kompletní a aktuální databázi. V mé práci jsou použita data z CD poskytnuté Ministerstvem práce a sociálních věcí, které by se mělo v případě potřeby aktualizovat změnovými soubory. Tyto soubory vychází každý týden na internetových stránkách Ministerstva práce a sociálních věcí. Pro moji práci je důležité tyto data propojit s chybnou databází EPIDAT. Cílem je tedy najít společný atribut databází, pomocí kterého mohu data doplnit a opravit. Detailní popis opravy chybné databáze EPIDAT pomocí UIR-ADR je v dalších kapitolách.

Posledním zdrojem dat této práce je Český statistický úřad, který společně s Ústavem zdravotnických informací a statistik ČR (ÚZIS) poskytuje data o problematice použití informačních technologií ve zdravotnictví. Poskytovatelé veřejné i soukromé zdravotní péče se dlouhodobě potýkají s problémem zajištění provozního zviditelnění informací a efektivní přístup informací pomocí informačních technologií je určitý krok kupředu v poskytování zdravotnických dat.

2.1.1. EPIDAT

K zajištění povinného hlášení, evidence a analýzy výskytu infekčních nemocí v České republice slouží databáze EPIDAT, který je celostátně používán Hygienickou službou ČR od 1.1.1993.

Program EPIDAT navazuje na ISPO (Informační systém přenosných onemocnění) z let 1982-1992. Hlášení infekčních nemocí je základem pro místní, regionální, národní a nadnárodní kontrolu šíření infekčních nemocí. Jeho

zákonným podkladem jsou závazné předpisy : Zákon č. 258/2000 Sb., o ochraně veřejného zdraví, ve znění pozdějších předpisů Vyhláška MZ ČR č. 195/2005 Sb., kterou se upravují podmínky předcházení vzniku a šíření infekčních onemocnění mezinárodní předpisy závazné pro členské země Světové zdravotnické organizace.

Úložiště dat slouží k bezpečné výměně aktuálních datových souborů o výskytu infekcí mezi jednotlivými pracovišti HS ČR, MZ ČR a Státním zdravotním ústavem v Praze.

Základní výstupy z databáze EPIDAT jsou zveřejňovány v časopise Zprávy epidemiologie a mikrobiologie, který je možný stáhnout z internetu.

Na stránkách státního zdravotního úřadu je možné sledovat s měsíční periodikou průběžný stav vybraných hlášených infekcí v ČR[7].

2.1.2. UIR-ADR

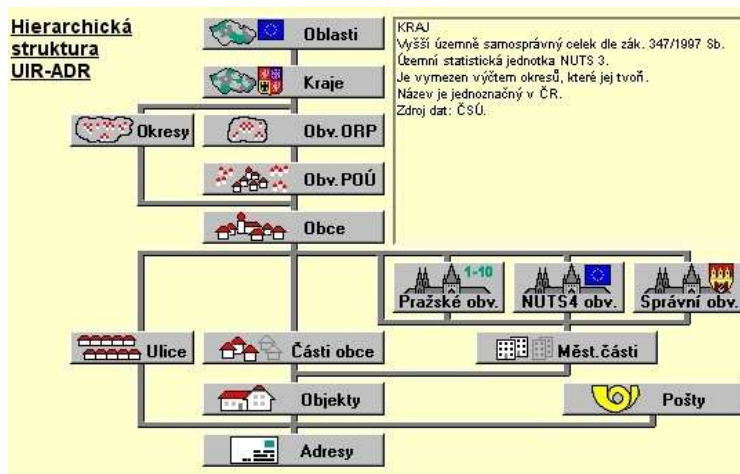
Územně identifikační registr adres byl vybudován v letech 1997-1999 za spolupráce obecních úřadů, Ministerstva pro místní rozvoj, Ministerstva vnitra, Českého úřadu zeměměřického a katastrálního, Českého statistického úřadu a České pošty.

Registr byl původně určen pro potřeby informačních systémů MPSV. Po svém dokončení byl uvolněn i pro ostatní zájemce a rychle se rozšířil mezi desítky uživatelů ze státní správy i mimo ni. Dnes dává MPSV tento registr k dispozici veřejnosti. Kromě zpřístupnění dat registru na www stránkách MPSV je možno získat zdarma CD-ROM s daty a programy pro prohlížení a periodickou aktualizaci těchto dat.

Registr prošel od začátku své existence kvalitativním vývojem od původní struktury 1.0 až k současné struktuře 4.2. Podrobnou dokumentaci k registru UIR-ADR struktury 4.2 i dřívějších struktur 3.0, 4.0 a 4.1 je možné stáhnout z internetu.

Současná struktura 4.2 obsahuje následující celostátní číselníky: číselníky oblastí, krajů, okresů, obvodů ORP, obvodů POÚ, obcí, pražských obvodů, NUTS4-obvodů a městských částí/městských obvodů jsou do UIR-ADR přebírány z Českého statistického úřadu, číselník správních obvodů z Magistrátu hl. m. Prahy, číselník částí obce z Ministerstva pro místní rozvoj, číselník adresních pošt z České pošty s.p. a číselníky ulic a veřejných prostranství, stavebních objektů a adresních míst jsou udržovány Ministerstvem práce a sociálních věcí na základě hlášení z obecních úřadů. Souřadnice adresních míst jsou do UIR-ADR poskytovány společností CEDA.

Používání UIR-ADR je možné hned několika způsoby: prohlížet pomocí internetu, ověřovat adresy v registru po internetu, udržovat si vlastní kopii registru, dotazovat se z vlastní aplikace on-line do registru na MPSV, zadávat reklamace na správnost dat v registru on-line způsobem[8].



Obrázek 1: Hierarchická struktura UIR-ADR

2.1.3. Český statistický úřad

Český statistický úřad (ČSÚ) je ústřední státní orgán České republiky, zajišťující sběr, zpracování a publikaci statistických údajů. Jedná se o hlavní orgán státní statistické služby, který koordinuje sběr a zpracování statistických údajů, které jsou prováděny jednotlivými ministerstvy.

Hlavní sídlo ČSÚ je v nové budově v Praze-Strašnicích. V jednotlivých krajích jsou zřízeny vnitřní organizační jednotky, tzv. krajské správy.

ČSÚ pravidelně zveřejňuje velké množství dat - nejnovější údaje o ekonomickém i sociálním vývoji ČR se zveřejňují formou tzv. Rychlých informací. Úřad také vydává množství publikací včetně Statistické ročenky ČR[9].

Hlavním zdrojem zdravotnických dat pro ČSÚ je Ústav zdravotnických informací a statistik ČR (ÚZIS). V části informační technologie ve zdravotnictví jsem uvedl aktuální stav zdravotnictví v oblasti využívání informačních technologií v ordinacích lékařů. Tímto jsem chtěl ukázat, kde vznikají časté chyby v datech, způsobené přepisy tištěných forem evidence pacienta do digitální.

2.2. Pracovní prostředí

Microsoft Excel

Microsoft Excel je tabulkový procesor, který je součástí kancelářského balíku Microsoft Office. Už delší dobu zaujímá dominantní postavení na trhu. V posledních letech mu zdatně konkuruje open source tabulkový procesor Calc, který se součástí OpenOffice.org.

Nástroje tohoto programu byly využity zejména v souvislosti s úpravou dat pro následné porovnání v databázovém programu Microsoft Access. Především možnost použití maker byla oceněna při úpravě databáze UIR-ADR.

Microsoft Access

Microsoft Access je nástroj pro správu relačních databází. Je součástí kancelářského balíku Microsoft Office. Jedná se o souborovou databázi s vhodným pracovním a grafickým prostředím.

Pro tvorbu programu pro opravu EPIDAT byla použita verze 2003. Volbu databázového programu Microsoft Office jistě ocení lidé ve zdravotnictví, kterým je výsledek této práce určen. Prostředí je uživatelsky známé a obsahuje i vhodné grafické rozhraní.

ArcGIS

ArcGIS Desktop 9.x je GIS software pro vizualizaci, vytváření, správu a analýzu prostorových dat. Základní softwarové produkty, které se liší úrovní funkcionality jsou: ArcView, ArcEditor, ArcInfo. Obsahuje sadu spolupracujících aplikací ArcMap, ArcCatalog a ArcToolbox.

ArcMap je centrální aplikací v ArcGIS Desktop, kterou užijeme při mapově orientovaných úlohách, prostorových analýzách a editaci dat. ArcCatalog obsahuje nástroje pro prohlížení a vyhledávání geografických informací, zaznamenávání a prohlížení metadat, k prohlížení datových sad a vytváření schématu struktury geografických vrstev. ArcToolbox obsahuje mnoho nástrojů pro prostorové operace[13]. V této práci je použit tento software pro geokódování nad opravenými daty databáze EPIDAT.

Corel Draw

Corel Draw patří již řadu let mezi přední světové grafické editory. Je součástí balíku programů CorelDRAW Graphics Suite. Poskytovatelem je americká firma Corel Corporation[14].

Tento grafický editor pracující především s vektory byl využit při tvorbě nápověd pro zpřehlednění opravy EPIDAT. Jeho funkcí bylo využíváno i při vytváření grafického rozhraní pro webové stránky.

2.3. Epidemiologie

Epidemiologie je lékařské odvětví zabývající se studiem faktorů ovlivňujících zdraví a nemocnost obyvatelstva a slouží také jako poklad k opodstatnění lékařských zásahů, provedených v zájmu veřejného zdravotnictví a preventivního lékařství. Epidemiologie je považována za základ metodologie výzkumu ve zdravotnictví a je vysoce ceněna v medicíně založené na důkazech, protože pomáhá rozpoznat rizikové faktory pro přenos nemocí a určuje optimální postup jejich léčby v klinické praxi.

Práce epidemiologů, zabývajících se ať už přenosnými nebo nepřenosnými chorobami, v sobě zahrnuje zkoumání vzniku nemoci, výběr vhodné studie, sběr a analýzu dat s ohledem na vývoj statistických modelů, sestavení hypotézy a sepsání závěrů do článku uveřejněném v odborném časopise. Mezi nejznámější patří časopis Epidemiologie, mikrobiologie, imunologie vydávaný 4 krát ročně. Epidemiologové však musí vycházet i z dalších vědeckých odvětví. Poznatky z biologie jsou potřeba k pochopení působení nemocí, zatímco společenské vědy jako sociologie a filozofie pomáhají vyhodnotit bezprostřední i méně aktuální rizikové faktory.

Epidemiologie se dělí se na epidemiologii obecnou, zabývající se metodologií práce a obecnými epidemiologickými zákonitostmi, a speciální epidemiologii konkrétních nemocí. Dále se ještě rozlišuje klinická epidemiologie, která se zabývá konkrétní klinickou aplikací poznatků obou předchozích a posuzováním výstupů klinických testů[10].

V poskytnuté epidemiologické databázi EPIDAT se vyskytuje 10 nejběžnějších onemocnění, kterým se přiřazují v lékařské terminologii použité zkratky v závorkách – salmonelóza (A02), virová střevní infekce (A08), lymerská borelióza (A69.2), klíšťová encefalitida (A84.1), neurčená virová encefalitida (A86), virová meningitida (A87.9), varicella (B01), hepatitida A (B15), akutní hepatitida B (B16), parotitida (B26).

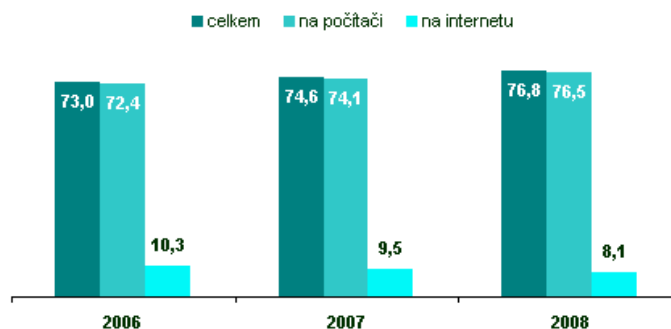
2.4. Informační technologie ve zdravotnictví

Poskytování informací, o informačních technologiích ve zdravotnictví, se zabývá od roku 2003 Český statistický úřad, který se snaží přehledně poskytnout dostatečné množství statistických údajů o rozvoji IT ve zdravotnictví. Mezi hlavní poskytovatele dat pro ČSU patří Ústav zdravotnických informací a statistik ČR (ÚZIS). Doplnkovými zdroji jsou webové stránky nemocni, šetření o využívání informačních a komunikačních technologií mezi jednotlivci a databáze systému IZIP (elektronická zdravotní knížka). Pro mezinárodní srovnání slouží průzkum vybavenosti a využívání IT v ordinacích praktických lékařů Evropské unie, realizován společností Empirica pod záštitou Evropské komise[11].

Následně jsou uvedeny vybrané poznatky za rok 2008 a první polovinu roku 2009. Aktuální informace za kompletní rok 2009 ještě nebyly zveřejněny.

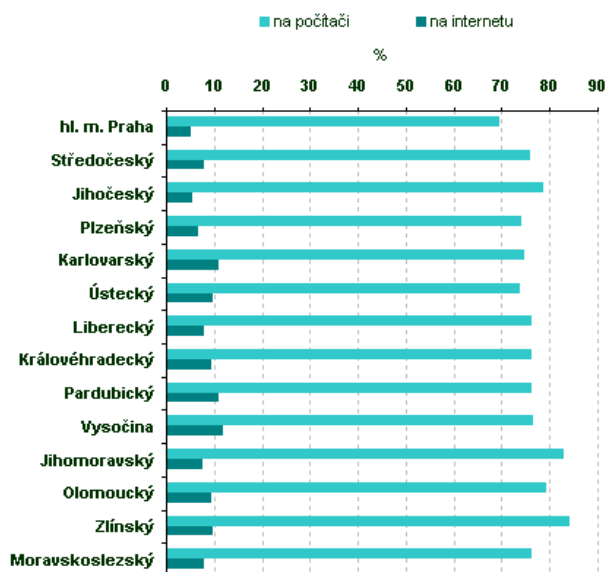
- v roce 2008 v ČR využívalo počítač při práci 96 % samostatných ordinací lékařů, připojení k internetu uvedlo 72 % z nich a vysokorychlostně bylo připojeno 44 % z nich
- vlastnictví webové stránky uvedlo 13 % samostatných ordinací lékařů
- elektronickou zdravotnickou dokumentaci o pacientech si na disku počítače vedlo 77 % samostatných ordinací lékařů, kdežto na internetu pouze 8 %
- objednat se on-line na vyšetření pomocí webového formuláře umožňovalo svým pacientům pouze 5 % samostatných ordinací lékařů
- průzkum webových stránek ukázal, že v polovině roku 2009 mělo funkční webovou stránku 95 % nemocnic ČR
- systému hrazení regulačních poplatků podávalo na svém webu informace 49 % těchto nemocnic
- možnost objednání se na vyšetření (nejčastěji na mamograf) pomocí webového formuláře nabízelo pouze 6 % těchto nemocnic
- z šetření o využívání ICT mezi jednotlivci za rok 2008 vyplynulo, že v rámci věkové skupiny 16-74 let použilo internet k vyhledávání informací o zdraví 14 % občanů ČR a 28 % občanů Evropské unie
- obyvatelé Finska vykazali za vyhledávání informací o zdraví na internetu nejvyšší podíl 51 %, naopak v Bulharsku se dané činnosti věnovalo pouze 7 % tamních obyvatel
- do systému IZIP se do konce první poloviny roku 2009 zaregistrovalo celkem 5 782 zdravotnických zařízení ČR a 1 059 758 občanů ČR

- nejvíce zaregistrovaných zdravotnických zařízení pochází ze Středočeského kraje (645), nejméně pak z kraje Karlovarského (237)
- klientů se zaregistrovalo v daném období nejvíce z kraje Ústeckého (143 265 osob) a nejméně z hl.m. Prahy (36 662 osob)



Graf 1: Grafické znázornění samostatných ordinací lékařů používající IT k vedení zdrav. dokumentace (% z celkového počtu samostatných ordinací lékařů)

Kartotéku se zdravotnickou dokumentací jednotlivých pacientů si v roce 2006 na svém počítači (ať už byla data uchovávána na disku či vkládána do zdravotní knížky na internet), vedlo 73 % samostatných ordinací lékařů (počítač v tomto roce v zaměstnání používalo 92 % ordinací). Zdravotnickou dokumentaci na počítačovém disku uchovávalo v tomto roce přes 72 % ordinací, na internet data vkládalo pouze cca 10 % ordinací (některé ordinace tak spravovaly dokumentaci jak v počítači, tak on-line). Zatímco vedení dokumentace v počítači vykázalo během let 2006 - 2008 mírný vzestup až na hodnotu 76,5 %, podíly za využívání elektronických zdravotnických knížek (např. systém vedení zdravotnických knížek IZIP) postupně klesají, podíl za rok 2008 je o přibližně 2 procentní body nižší než zmíněný podíl za rok 2006. Celkově za rok 2008 použilo počítač pro vedení zdravotnické dokumentace buď v počítači či na internetu necelých 77 % ordinací (počítač v tomto roce používalo téměř 96 % ordinací).



Graf 2: Grafické znázornění samostatných ordinací lékaře používající IT k vedení zdravotnické dokumentace v krajích ČR; 2008 (% z celkového počtu samostatných ordinací lékaře v daném kraji)

Mezi jednotlivými kraji České republiky jsou znatelné rozdíly ve využívání IT k vedení zdravotnické dokumentace především za uchovávání dat na počítačovém disku. Rozdíl mezi krajem s nejvyšším a nejnižším podílem (nejvyšší podíl Zlínský kraj přes 84 % - nejnižší podíl překvapivě hl.m. Praha necelých 70 %) činí cca 15 procentních bodů. V ostatních krajích se podíly pohybují mezi 70 - 80 %, přes 80 % se dostal ještě kraj Jihomoravský.

Formu vedení zdravotnické dokumentace na internetu volily v roce 2008 nejvyšším podílem samostatné ordinace lékaře v kraji Vysočina - cca 12 %, na opačném konci stojí opět hl.m. Praha s podílem 5 %[11].

3. VÝSLEDKY

3.1. Předzpracování dat

V této kapitole je popsána příprava dat pro samotné porovnání a tvorbu programu pro opravu dat. Poskytnutá data je důležité upravit do jednotného formátu pro lepší porovnávání.

Krajská hygienická stanice poskytla data databáze EPIDAT ve formátu xls. Tento tabulkový formát je vhodný pro konvertování do databázového programu. Databáze ÚIR-ADR, která byla zaslána na instalačním CD Ministerstvem práce a sociálních věcí má data v následujících databázových formátech: dBase, FoxPro, Oracle, MS SQL Server, SQL_92, CSV.

3.1.1. Úprava dat

Před samotnou úpravou dat je důležité si uvědomit, že data databáze EPIDAT, by se neměla v žádném případě upravovat. Je zde důležité zachovat původní atributy pro následné porovnání s opravenými daty. Formát textu sledovaných atributů v EPIDAT je velkými písmeny bez diakritiky (kromě ulice). Všechny atributy v EPIDAT, které jsou vybírány z číselníku, mají tento formát. Je zde, ale možnost i ručního dopisování dat uživatelem, což se projevilo nejvíce na chybovosti dat.

ULICE	OBEC	MISTNICAST	OKRES
	OLESNICE		OC
	UJEZDEC		PR
	ZELEC		PV
Tř.Svornosti	OLOMOUC		OC
Osmek	PREROV		PR
	HVOZD		PV

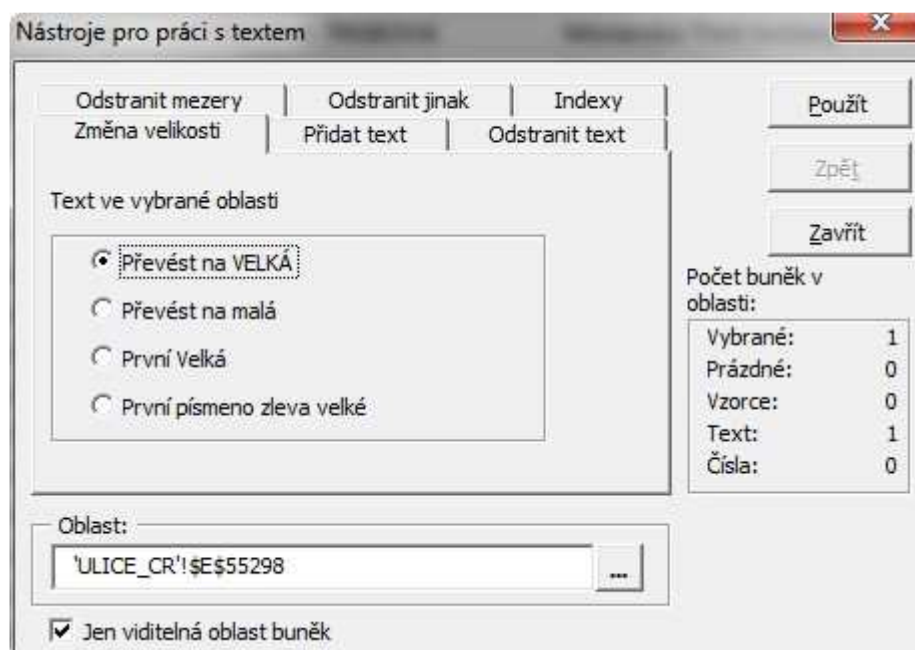
Obrázek 2: Ukázka formátu dat v EPIDAT

V databázovém programu MS Access 2003 byla načtena data obcí a ulic ve formátu dBase z databáze ÚIR-ADR. Pomocí dotazů se k datům připojily

zkratky okresů a rozdělil data na obce a ulice v Olomouckém kraji a České republice.

Dalším krokem bylo odstranění diakritiky a převod na velká písmena. K této úpravě byl použit nástroj RJ Tools, který je volně stažitelný z internetu[15]. Jedná se o soubor předdefinovaných maker obsažených v nástrojích, která jsou importována do programu MS Excel.

Jelikož data ulic obsahovala přes 70 000 záznamů, bylo nutné použít novější verzi MS Excel 2007. Starší verze MS Excel 2003 může pracovat maximálně s 65 536 záznamy.

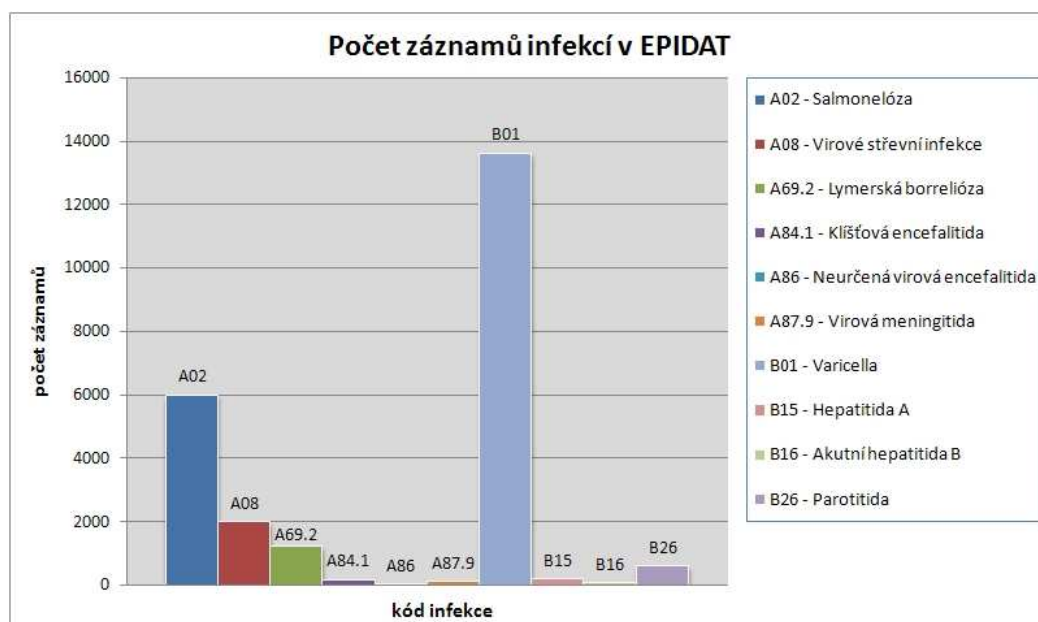


Obrázek 3: Ukázka nástroje pro převod na velká písmena

3.2. Analýza EPIDAT

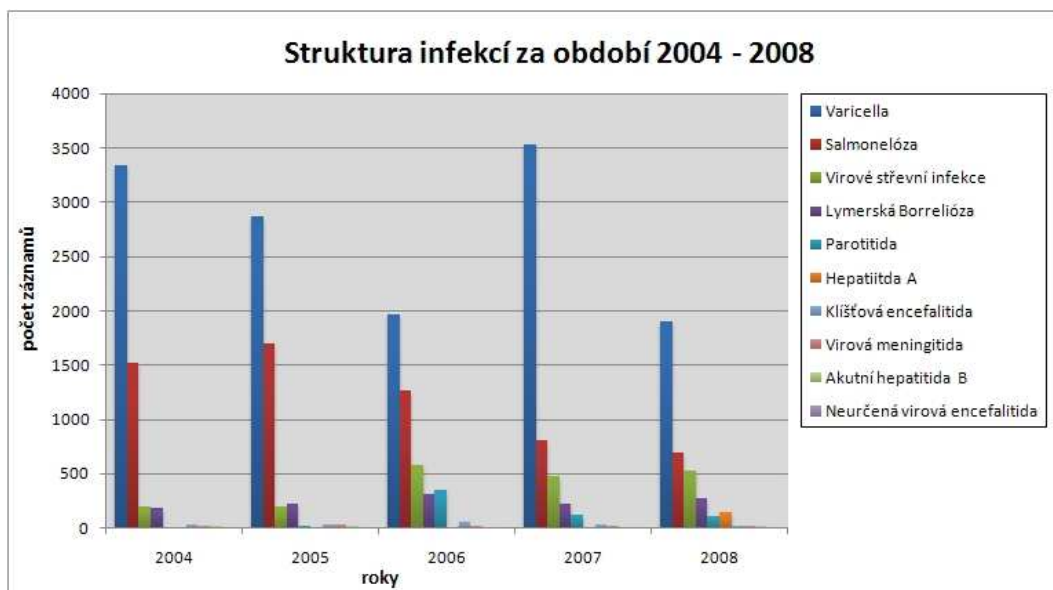
V této části byl analyzován EPIDAT z hlediska struktury infekcí obsažených v poskytnutých datech. Ty obsahovaly 23 999 záznamů za období 5 let. Analýza dat byla prováděna formou dotazů v programu MS Access.

Rozdělení EPIDAT podle infekcí můžeme vidět v grafu č. 3. Největší část z celkového počtu dat (13612 záznamů) je spojena s infekcí B01 – Varicella u nás známá jako plané neštovice. Následuje A02 – Salmonelóza (5992 záznamů), A08 – Virová střevní infekce (1987 záznamů), A69.2 – Lymerská borelióza (1227 záznamů). Ostatní infekce mají jen velmi nízký podíl na celkovém počtu infekcí.



Graf 3: Počet záznamů infekcí v EPIDAT

Strukturu infekcí za období 2004 – 2008 můžeme vidět v grafu č. 4. V grafu můžeme sledovat vývoj jednotlivých infekcí. Nejsilnějším rokem z hlediska součtu záznamů všech infekcí je rok 2007, naopak nejslabším je rok 2008. Nedá se konstatovat, že by vývoj nějaké infekce prudce stoupal nebo klesal. Výrazné výkyvy oproti normálu prokazuje zejména parotitida v roce 2006 a hepatitida A v roce 2008. Obecně lze říci, že počet záznamů u salmonelózy každým rokem pozvolna klesá a počet záznamů u virové střevní infekce pozvolna stoupá. Pro lepší představivost můžeme vidět v tabulce č. 1 jednotlivé infekce v číslech.



Graf 4: Struktura infekcí v EPIDAT za období 2004 – 2008

Tabulka 1: Infekce v EPIDAT za období 2004 - 2008

NÁZEV INFEKCE	ZKRATKA INFEKCE	2004	2005	2006	2007	2008	CELKEM
SALMONELÓZA	A02	1522	1706	1261	803	700	5992
VIROVÉ STŘEVNÍ INFEKCE	A08	198	197	581	476	535	1987
LYMESKÁ BORRELIÓZA	A69.2	180	229	319	225	274	1227
KLÍČŤOVÁ ENCEFALITIDA	A84.1	33	28	58	32	24	175
NEURČENÁ VIROVÁ ENCEFALITIDA	A86	4	4	5	5	1	19
VIROVÁ MENINGITIDA	A87.9	15	31	16	26	16	104
VARICELLA	B01	3341	2865	1973	3529	1904	13612
HEPATITIDA A	B15	11	12	6	6	152	187
AKUTNÍ HEPATITIDA B	B16	22	20	14	6	18	80
PAROTITIDA	B26	1	18	356	126	115	616
CELKEM		5327	5110	4589	5234	3739	23999

3.3. Popis EPIDAT

Část popisu dat je neméně důležitá z hlediska pochopení souvislostí mezi jednotlivými atributy v databázi EPIDAT. V každém záznamu je obsaženo 50 atributů, které můžeme vidět v tabulce č. 2. Pro naší práci jsou nejdůležitějšími atributy ULICE, OBEC, OKRES, MISTONAKAZ a MISTOONEM. Atribut, který byl použit z číselníku se vyznačoval v datech velkými písmeny. Po prozkoumání dat můžeme určit, že i v datech z číselníku se vyskytly chyby dané možnostmi dopisování do těchto atributů.

Tabulka 2: Popis atributů v EPIDAT

Název atributu	Popis atributu	Příklad atributu	Číselník
N1	nespecifikováno	3	
KHS	číslo krajské hygienické stanice	71 (KHS Olomouc)	A
OHS	zkratka okresní hygienické stanice	JE (OHS Jeseník)	A
KOD	nespecifikováno	4	
SKUPINA	nespecifikováno	L	
SMAZ	chybný záznam	-	
PRIJMENI	příjmení pacienta	neudáno	
JMENO	jméno pacienta	neudáno	
POHLAVI	pohlaví pacienta	Z (žena)	
NAROZEN	datum narození	neudáno	
RODNECISLO	rodné číslo	neudáno	
VEK	věk pacienta	neudáno	
ULICE	název ulice	Dukelská	
OBEC	název obce	PREROV	A
MISTNICAST	název místní části, pokud se vyskytuje v obci či městě	Němčičky	
OKRES	zkratka okresu	OC (Olomouc)	A
DAT1PRI	datum prvního přijetí	6.1.2004	A
TYDONEM	týden onemocnění - pořadové číslo týdne, ve kterém pacient onemocněl	2	
DG	kód diagnózy	B01 (VARICELLA)	A
DATHLAS	datum hlášení - nahlášení pacienta hygienické stanici nebo do EPIDAT	8.1.2004	
TYDVYK	týden vykazání - pořadové číslo vykázaných týdnů v EPIDAT	2	
IMPORT	import infekce z jiného státu do ČR nebo import záznamu pacienta do ČR	A (ano)	A
IMPODKUD	import z jakého státu	Španělsko	A
IMPKYM	kdo importoval infekci (onemocnění)	cizinec	A
IMPKYM1	kód importu	U3 (cizinec)	A
MISTONAKAZ	místo, kde se pacient nakazil	JESENÍK	A
OKRESNAK	okres nákazy pacienta	UO (Ústí nad Orlicí)	

MISTOONEM	místo, kde pacient onemocněl (kde byl prohlášen za nemocného)	PROSTEJOV	A
OKRESONEM	okres onemocnění	OC (Olomouc)	
PRACOVIST	pracoviště, kde pacient onemocněl		
ZAMESTNANI	zaměstnání pacienta	ZACI,STUDENTI	A
JAKEZAMEST	druh zaměstnání pacienta	ekonom	
KOLEKTIV	prostředí, ve kterém se pacient nyní vyskytuje	ZAKLADNI SKOLA	A
KOLEKTIV1	kód kolektivu	4 (základní škola)	A
INTERNATNI	bydlící v ubytovacím zařízení	A (ano)	A
SOUVISLOST	souvislost s nakažením	KOLEKTIV	A
SOUVISLOST1	kód souvislosti	K (kolektiv)	A
ETNIKA	etnická skupina	ROMOVE	A
ETNIKA1	kód etnické skupiny	B (romové)	A
MISTOISOL	místo izolace pacienta	INFEKCNI ODD	A
MISTOISOL1	kód místa izolace	2 (infekční oddělení)	A
DATUMISOLA	datum izolace, může být odlišný od DAT1PRI A DATHLAS	7.1.2004	A
UMRTI	umrtí pacienta	N (ne)	A
DGUMRTI	diagnóza umrtí	B16 (akutní hepatitida B)	
DATUMUMRTI	datum umrtí	5.5.2006	
LEKAR	název lékaře	neuvedeno	
EPIDEMIE	epidemie	A (ano)	A
JAKAEPID	jaká epidemie - může to být místo nebo konkretizace epidemie	PV interna (přerovská interna)	
POZNAMKA	poznámka	neuvedeno	
ROK	rok zápisu záznamu, podle DATHLAS	2004	

3.4. Identifikace chyb

Identifikace chyb v datech je důležitou fází, která předchází tvorbě vlastního programu na odstranění chyb. Je nutné nalézt všechny možné varianty chyb. Jedná se zejména o chyby v diakritice, prázdné řádky, použití zkratk či vynechání písmene v názvu. K porovnání dat jsou zde užity, jak data z UIR-ADR, tak data z ISKN obsahujících katastrální území. Právě získání těchto dat nám pomohlo k přesné opravě obcí, kde se právě často katastrální území vyskytovaly. Data byly srovnávány v rámci Olomouckého kraje a České republiky. Poskytnutá databáze EPIDAT je v rozsahu pro Olomoucký kraj, ovšem mnoho záznamů obsahuje právě obce mimo tento kraj.

Srovnávacími atributy jsou OBEC, ULICE, MISTONAKAZ, MISTOONEM. Poslední jmenované se od sebe liší. Zatímco místo nákazy (MISTONAKAZ) je místo, kde se pacient nakazil, místo onemocnění (MISTOONEM) je místo, kde pacient nahlásil podezření z infekce.

Tabulka 3: Porovnání sledovaných atributu v EPIDAT s UIR-ADR v rámci Olomouckého kraje

Název atributu	dobré záznamy	prázdné řádky	chybné záznamy		celkem
			jedinečná hodnota	celkem	
Obec	22 255	0	255	1 744	23 999
Ulice	9 067	6 529	2 675	8 403	23 999
Místo nákazy	21 903	118	348	1 978	23 999
Místo onemocnění	22 044	6	293	1 949	23 999

Tabulka 4: Porovnání sledovaných atributu v EPIDAT s UIR-ADR v rámci České republiky

Název atributu	dobré záznamy	prázdné řádky	chybné záznamy		celkem
			jedinečná hodnota	celkem	
Obec	22 328	0	210	1 671	23 999
Ulice	10 078	6 529	2 439	7 392	23 999
Místo nákazy	21 919	118	336	1 962	23 999
Místo onemocnění	22 047	6	290	1 946	23 999

Výsledky porovnání můžeme vidět v tabulkách 3 a 4. Identifikace chyb jsem prováděl v databázovém programu MS Office za pomoci SQL dotazů. Ačkoliv poskytnutá databáze EPIDAT je pro Olomoucký kraj, srovnání s Českou republikou nám odhalilo mnohdy nemalé rozdíly.

Tabulka 5: Procentuální zastoupení záznamů v rámci Olomouckého kraje a České republiky

Název atributu	Olomoucký kraj			Česká republika		
	dobré záznamy	prázdné řádky	chybné záznamy	dobré záznamy	prázdné řádky	chybné záznamy
OBEC	92,73%	0,00%	7,27%	93,04%	0,00%	6,96%
ULICE	37,78%	27,21%	35,01%	41,99%	27,21%	30,80%
MÍSTO NÁKAZY	91,27%	0,49%	8,24%	91,33%	0,49%	8,18%
MÍSTO ONEMOCNĚNÍ	91,85%	0,03%	8,12%	91,87%	0,03%	8,11%

V tabulce 5 můžeme vidět procentuální zastoupení záznamů EPIDAT v porovnání s UIR-ADR. Největší počet záznamů, které byly správně propojeny, jsou u atributu OBEC. Spolu s atributy MÍSTO NÁKAZY A MÍSTO ONEMOCNĚNÍ mají možnost vybírat záznamy v číselnících. Je zde, ale i možnost dopisování dat a zejména tímto vznikají chyby. Nejvíce prázdných řádků a chybných záznamů u atributu ULICE je způsobené zejména nedostupností číselníku ulic. Mnoho prázdných řádků značí, že obce nemají ulice.

OBEC_EPIDAT	OBEC_UIRADR	OKRES_EPIDA	OKRES_UIRADR
BRODEK U PREROVA		OC	
BRODEK U PROSTEJ.		PV	
BUKOVA		JE	
BUKOVICE		JE	
BUKOVICE		SU	
BYSTRICE		PR	
BYSTRICE P. HOSTYN		KM	
ČAKOV		OC	
ČECHY POD KOSIREM		OC	
ČELECHOVICE		OC	
ČELECHOVICE		PV	

Obrázek 4: Ukázka chybných obcí EPIDAT v MS Access

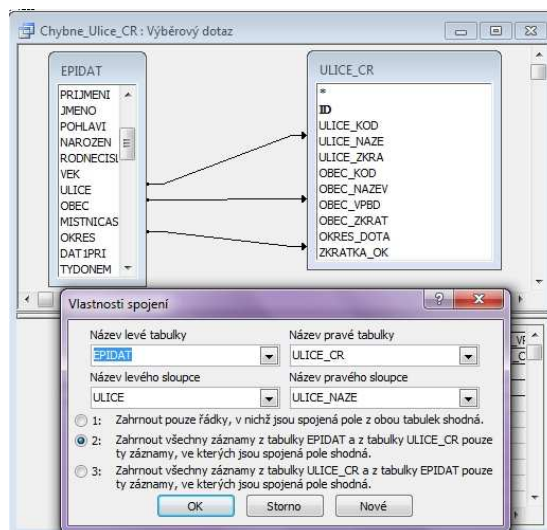
U chybných obcí může vznikat mnoho chyb. Na obrázku č. 4 můžeme vidět příklady chyb. Např. u obcí Brodek u Prostějova a Bystřice pod Hostýnem

můžeme vidět použití zkráceniny. Obce Čechy pod Kosířem a Čelechovice spadají naopak do jiného okresu.

ULICE_EPIDAT	ULICE_UIRADR
Alova	
ALŠOVA	
amanovská	
Ambroze	
Americká	
Annenská	
Annín	
Ant. Slavička	
Arinova	
árka	
Arm.sboru	
Arnoltice	

Obrázek 5: Ukázka chybných ulic EPIDAT v MS Access

U atributu ulice se v databázi EPIDAT vyskytovalo nejvíce chyb. Mnoho záznamů nebylo vyplněno, a proto bude velmi těžké tyto chyby napravit. Na obrázku č. 5 vidíme ukázkou chybných ulic v databázi EPIDAT. Tento atribut nebyl vybírán z číselníku, proto není napsán velkými písmeny. Chyby jsou zde nejčastěji v použití velkých písmen, zkráceninách, chybějících písmen, názvu obcí a katastrálních území

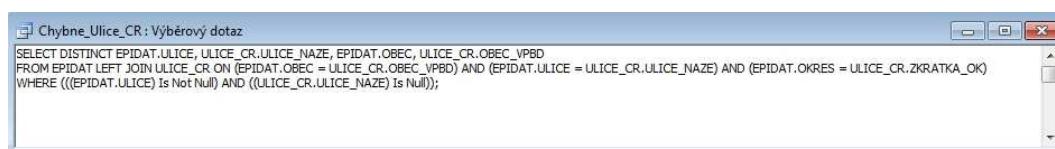


Obrázek 6: Ukázka relací ulic v EPIDAT v MS Access

Abychom mohli přesně detekovat vzniklé chyby, je nutné správně zvolit relace mezi atributy. Jen tak můžeme jednoznačně určit správné a chybné

záznamy. Na obrázku č. 6 vidíme relace pro odhalení chybných ulic. Relace jsou nastaveny, jak pro obce, ulice a okresy. Důležité je nastavit vlastnosti spojení mezi relacemi. Tím nám vykrystalizují opravdu chybné záznamy, které je nutno opravit.

Každý dotaz si můžeme zobrazit v podobě SQL dotazu a upravovat tak funkčnost. Na obrázku č. 7 vidíme prostředí pro zápis v SQL formě.



Obrázek 7: Prostředí pro úpravu a tvorbu SQL dotazů

3.4.1. Chyby v EPIDAT

V následující kapitole jsou detailněji rozvedeny chyby vyskytující se v databázi EPIDAT. Chyby se vyskytovali, ve všech sledovaných atributech – OBEC, OKRES, ULICE. U každého atributu jsou uvedeny nejčastější chyby opravy. V jednotlivých tabulkách jsou vždy ukázány červeně původní dat a modře data opravená.

1. OBEC (obec, místo nákazy, místo onemocnění)

Tabulka 6: Ukázka opravy chyb u atributu OBEC

CHYBNÁ OBEC	SPRÁVNÁ OBEC	VYSVĚTLENÍ
CELECHOVICE NA H.	CELECHOVICE NA HANE	zkrácenina
Němetice	NEMETICE	obec je napsána s diakritikou
oc	OLOMOUC	chybná zkratka okresu v názvu obce
TERNBERK	STERNBERK	chybějící počáteční písmenko - nejčastěji písmenko s háčkem, může chybět i uprostřed obce - P EROV (PŘEROV)
Vyškovsko	VYSKOV	název regionu v poli pro obec
Čechovice	PROSTEJOV	část obce, nelze dohledat ani správnou obec, ku - nutný jiný zdroj (internet)
Doloplazy - Tršice	TRSICE	specifikace obce, zde je dána přednost katastrálnímu území Tršice
FRANCIE		obce mimo ČR případně názvy států se neopravují, pouze se запиše do poznámky - MIMO CR
MIMO UZEMI CR	CHYBNY ZAZNAM	v případě, že se vyskytuje jiný název než název obce či státu, je záznam opraven jako chybný
SKOKY	SKOKY U STAMERIC	obec je jen z části napsána správně
FN OLOMOUC	OLOMOUC	zkratka FN (fakultní nemocnice) se do opravené obce nepřidává

2. OKRES (okres, okres nákazy, okres onemocnění)

Okresy byly kontrolovány až po opravě obcí. Vyskytovalo se zde méně variant chyb než u OBCE. Důležitým zjištěním je, že OKRES mnohdy opravil i obec do finální podoby.

Tabulka 7: Ukázka opravy chyb u atributu OKRES

CHYBNÁ OBEC	CHYBNÝ OKRES	SPRÁVNÁ OBEC	SPRÁVNÉ KU	SPRÁVNÝ OKRES	VYSVĚTLIVKY
ALBRECHTICE	BR	ALBRECHTICE		XX	největší chyba při opravě okresů, nelze opravit správný okres, jelikož obec se vyskytuje ve 2 okresech a ani jeden z nich není v Olomouckém kraji
BRODEK U PREROVA	OC	BRODEK U PREROVA		PR	oprava okresu
BILOVICE	PV	BILOVICE-LUTOTÍN		PV	oprava obce, zde dávám přednost správnému okresu z Olomouckého kraje a opravuji obec
KOBERICE	PV	HRADCANY-KOBERICE	KOBERICE	PV	oprava obce na základě katastrálního území, do výsledného EPIDAT je pak přidána obec i KU
KOSTELEČ	OC	KOSTELEČ NA HANÉ		PV	oprava obce i okresu, obec KOSTELEČ se nevyskytuje v okrese OC ani v Olomouckém kraji, proto zde opravuji na základě správné obce i okres
PRAHA	11	PRAHA		A	chybný okres obsahuje číslíci
OLMOUC		OLMOUC		OC	chybný okres je nevyplněný, dále se vychází z chybné obce

3. ULICE (ulice)

Poslední opravou byla oprava ulic, ve které se vyskytovalo nejvíce chyb. Pozitivní věc na této opravě byla, že se nám zde odhalila finální podoba opravené obce, katastrálního území nebo okresu.

Tabulka 8: Ukázka opravy chyb u atributu ULICE

CHYBNÁ ULICE	OBEC	OKRES	OPRAVA ULICE	OPRAVA OBEC	OPRAVA KU	OPRAVA OKRES	VYSVĚTLIVKY
ezáče	OLMOUC	OC	Řezáčova				nejčastější chyba, chybně zapsaná ulice spadající do příslušné obce a okresu
Jeremiáe	STEPANOV	OC	Jeremiášova	OLMOUC			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
Čakov	CHOLINA	OC		SENICE NA HANÉ	ČAKOV		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			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 MORAV OU	OC	pomocí ulice zde zjišťujeme konkrétní katastrální území a k němu obec a okres

3.5. Tvorba a popis programu pro opravu chyb a geokódování

Program pro opravu chybných dat databáze EPIDAT je stěžejní částí bakalářské práce. Je vytvořen v programu Microsoft Access, který pracuje s relačními databázemi. Program je rozdělen na 2 hlavní části. V první uživatel postupně opravuje pomocí průvodců obce, okresy a ulice. V této části také uživatel získá potřebné výstupní tabulky pro následné geokódování. V druhé části je popsán postup pro geokódování opravených dat v programu ArcGIS Desktop.

Program je řešen pomocí tabulek, dotazů, formulářů, sestav a maker. Celkem vstupují do programu 4 tabulky: EPIDAT, OBCE_CR, KU, ULICE_CR. Poslední 3 zmiňované byly vyexportovány z UIR-ADR a upraveny pro opravu.

PRŮVODCE OPRAVOU EPIDAT

1. NAHRAJ DATA PRO OPRAVU

2. PRŮVODCE OPRAVOU OBČÍ PRŮVODCE OPRAVOU OKRESŮ PRŮVODCE OPRAVOU ULIC

3. CHYBY V EPIDAT

4. PRŮVODCE KÓDOVÁNÍM

5. PRŮVODCE AGREGAČNÍMI DOTAZY

PRŮVODCE GEOKÓDOVÁNÍM

1 2 3 4

KONEC PROGRAMU

Michael HAVLÍK
Bakalářská práce: Průvodce geokódováním zdravotnických dat databáze EPIDAT
Vedoucí práce: Ing. Zdena Dobešová, Ph.D.
Konzultant: Mgr. Jan Harbula
2010

O PRŮVODCI 1. 2. 3. 4. 5.

Průvodce opravou EPIDAT vznikl, jako součást bakalářské práce Průvodce geokódováním zdravotnických dat databáze EPIDAT.

Autor zde pracuje s epidemiologickou databází EPIDAT, poskytnutou Krajskou hygienickou stanicí Olomouc a celostátní platnou databází UIR-ADR (Územně identifikační registr adres). Pomocí této databáze, jsou opravována chybná data.

Všechny dotazy a relace jsou nastaveny tak, aby bylo možné pracovat i s novými, aktuálnějšími daty.

Průvodce je rozdělen do několika částí podle čísl. Je nutné postupovat hierarchicky v pořadí od 1, aby se vytvářeli postupně tabulky a dotazy.

V chybné databázi EPIDAT se opravuje celkem 7 sloupců, které jsou potřebné pro následné geokódování v mapě. Jedná se o ULICI, OBEC, OKRES, MÍSTO NÁKAZY, OKRES NÁKAZY, MÍSTO ONEMOCNĚNÍ A OKRES ONEMOCNĚNÍ.

Obrázek 8: Úvodní formulář opravy EPIDAT

Po spuštění programu se otevře úvodní formulář s hlavní nabídkou viz. obrázek č. 8. Uživateli je nabídnuto hned několik tlačítek nabízející různé funkce. Součástí je i 5 obrázků s logy, kde se po kliknutí na obrázek může uživatel připojený i internetu dostat na příslušnou webovou adresu. Dole pod čarou je možné po kliknutí na jméno autora práce, získat jeho e-mail. V pravé části je popsáno pomocí nápověd vše podstatné pro následnou opravu.

3.5.1. Průvodce opravou EPIDAT

Průvodce opravou EPIDAT je rozdělen do 5 kroků. Je důležité zde postupovat hierarchicky od 1. V prvním si uživatel nahraje chybná data, které se budou v následujících krocích opravovat. Po stisknutí tlačítka je zobrazena tabulka s předem definovanými atributy. Je zde důležité, aby se po přidání dat vyskytovala data v příslušných sloupcích. Data zde můžeme přidávat i mazat.

Druhý krok je rozdělen na 3 části: průvodce opravou obcí, průvodce opravou okresů a průvodce opravou ulic. Na obrázku č. 9 je zobrazena část průvodce opravou obcí. Jsou zde opravovány obce, místa nákazy a místa onemocnění. V každé části je obsaženo několik tlačítek pro opravu. Všechny chybné obce jsou ukládány do tzv. chybovníku, se kterým uživatel dále pracuje. Chybovník je samostatná tabulka, která se uživateli jeví jako formulář, kde si uživatel nahraje pomocí tlačítek chybné záznamy (zde chybné obce). Po přidání chybných obcí do chybovníku je možné vybírat správnou obec pomocí číselníku obcí z UIR-ADR. Uživatel postupně píše písmena obce a z číselníku jsou mu nabízeny možné obce pro opravu. Dále zde uživatel píše poznámky k datům, kde uvádí popis opravy (např. oprava okres, chybný záznam). V poslední fázi se pomocí tlačítka doplní opravené obce do nově vzniklé tabulky, která obsahuje kromě původních dat i naše opravená data. Společně se s opravenými záznamy přidává i další atribut, který definuje druh opravy (např. OPRAVA OBEC, OPRAVA MÍSTA NÁKAZY atd.). Jedná se o poznámku k provedené opravě, která ve výsledné tabulce obsahující původní data i opravená data, stojí vždy u opraveného atributu.

CHYBNÉ OBCE	SPRÁVNÉ OBCE
BRODEK U PROSTĚJ.	BRODEK U PROSTĚJOVA
BYSTRICE P. HOSTÝN	BYSTRICE POD HOSTÝNEM
CELECHOVICE NA H.	CELECHOVICE NA HANĚ
DOMASOV N. BYSTRIC	DOMASOV NAD BYSTRICÍ
DOMASOV U STERNB.	DOMASOV U STERNBERKA
DVUR KRALOVE N. L.	DVUR KRALOVE NAD LABEM
HRADEC N. MORAVICI	HRADEC NAD MORAVICÍ
HUSTOPECE N. BECV.	HUSTOPECE NAD BECVOU
KOVALOVICE-OSICA	KOVALOVICE-OSICANY
LITOMYSLAV	LITOMYSLAV

Obrázek 9: Průvodce opravou obcí

Průvodce opravou okresu a průvodce opravou ulic pracují na stejném principu, který je popsán výše. Nutno podotknout, že opravená obec nemusí být

ve výsledku správná. Každá oprava nám více specifikuje výslednou podobu opraveného záznamu. Například opravíme obec, ale při opravě okresu zjistíme, že se daná obec nevyskytuje v okrese, který je jí přiřazen. Zde se uživateli nabízí hned několik možností opravy. Může název obce změnit na název katastrálního území, které spadá pod požadovaný okres. V atributu obec se totiž neřeší, zda se jedná o obec nebo katastrální území. Další možnost nastává, když se obec vyskytuje ve více okresech najednou. Zde je dána přednost okresům v Olomouckém kraji, jelikož data jsou především z Olomouckého kraje a je zde největší pravděpodobnost, že obec tímto opravíme správně. Pokud se obec vyskytuje ve více okresech a žádný není z Olomouckého kraje, pak není jiná možnost než přidat do poznámek, že obec nelze správně opravit.

Důležité zjištění nastalo při opravě ulic. Uživatelé EPIDAT, zde nevyužívali tento atribut vždy k doplnění správné ulice, ale zapisovali zde mnohdy i správný název místa onemocnění. Vyskytoval se zde často název katastrálního území případně název obce. U atributu ulice neměli uživatelé EPIDAT výběr z číselníku. Je zcela zřejmé, že to byl největší důvod, proč se zde vyskytovalo nejvíce chyb. Na obrázku č. 10 je zobrazen průvodce opravou ulic, ve kterém bylo opravováno nejvíce záznamů.

CHYBNÉ ULICE	SPRÁVNÁ OBEC	SPRÁVNÝ OKRES	OPRAVA ULICE	OPRAVA OBEC	OPRAVA KU	OPRAVA OKRES	POZNÁMKA
bzačce	OLOMOUC	OC	Rezáčova				oprava ulice
IMSKÁ	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
puroká	TOVACOV	PR	Široká				oprava ulice
plechitělů	OLOMOUC	OC	Šlechtitělů				oprava ulice
st párov	STEPANOV	OC					název obce
starnov	STARNOV	OC					název obce
střtného	OLOMOUC	OC	Štítného				oprava ulice
sumvald	SUMVALD	OC					název obce
švermova	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

Obrázek 10: Průvodce opravou ulic v EPIDAT

Ve třetím kroku opravy dat bylo zapotřebí všechny chybné záznamy (obce, okresy, ulice) zapsat do jedné tabulky pro následné vyhodnocení chyb. Všechny chybné záznamy se tedy zapisovaly do tzv. velkého chybovníku. Po vytvoření struktury tabulky jsou pomocí tlačítek přidávány chybné záznamy. Výslednou tabulku je možné pomocí tlačítka vyexportovat do tabulkového formátu xls.

Obrázek 11: Velký chybovník z EPIDAT

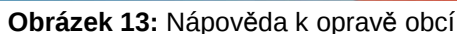
Čtvrtý krok je důležitou pomůckou pro následné geokódování a prostorové analýzy. Zde byly přiřazeny všem ulicím, obcím, katastrálním územím a okresům kódy, které zpřesnily význam prostorové informace. Kódy byly přiřazeny pomocí dotazů z databáze UIR-ADR.

Obrázek 12: Kódování v EPIDAT

V posledním kroku opravy dat bylo vytvořeno několik tabulek pomocí agregačních dotazů. Zde byla důležitá konzultace s uživatelem EPIDAT, který bude těchto dotazů využívat pro následné prostorové analýzy. Byly exportovány tabulky, které obsahovali rozdělení podle pohlaví (muži, ženy), kolektivu (bez kolektivu, pracoviště, mateřská škola, základní škola, střední škola, vysoká škola), věkových skupin (0-6, 7-14, 15-59, 60+), zaměstnání (dítě, žáci a studenti, důchodci, jiné) a podle data prvního přijetí (zde rozděleno podle roků za období 2004-2008).

Druhá část programu je věnována postupu pro následné geokódování v programu ArcGIS Desktop. Postup zde byl rozdělen na 4 části (načtení toolbox, tvorba adresového lokátoru, vlastní geokódování a kontrola chyb). Pomocí tlačítek se uživateli zobrazí nápovědy. Postup geokódování je popsán v následující kapitole této práce.

Kromě hlavní nápovědy se vyskytuje v programu dalších 7 nápověd, které doplňují jednotlivé kroky opravy dat. Po rozkliknutí tlačítka se vždy tyto nápovědy nacházejí v horní části průvodce vedle nadpisu a jsou označeny otazníkem. Obsah nápovědy můžeme vidět na obrázku č. 13.



3.5.3. Závěrečné shrnutí programu

Celkem program obsahuje:

- 1) 22 tabulek
- 2) 142 dotazů
- 3) 14 formulářů
- 4) 11 sestav
- 5) 14 maker

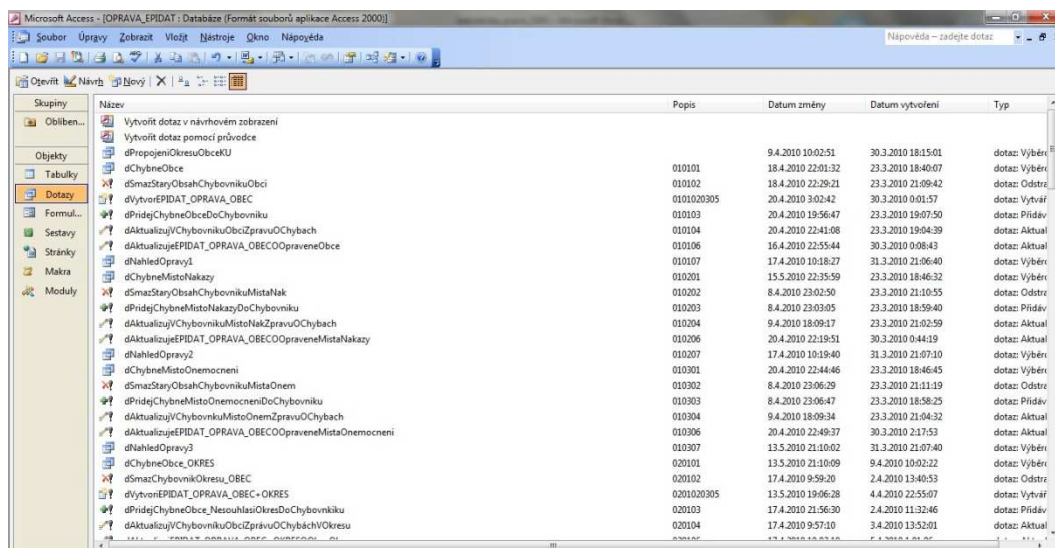
Ad 1) z celkového počtu 22 tabulek jsou 4 tabulky vstupní, 4 tabulky postupné opravy, 8 tabulek s chybovníky a 6 tabulek s výstupními agregačními dotazy

Ad 2) z celkového počtu 143 dotazů je 19 dotazů určených pro opravu obcí, 20 pro opravu okresů a 10 pro opravu ulic. 9 dotazů je použito k vytvoření velkého chybovníku, 8 pro kódování dat a 76 pro tvorbu agregačních dotazů

Ad 3) z celkového počtu 14 formulářů jsou 4 pro opravu obcí, 4 pro opravu okresů a 2 pro opravu ulic. Po jednom formuláři – úvodní nabídka, kódování, velký chybovník a agregační dotazy. Ukázka dotazů je na obrázku č. 14.

Ad 4) z celkového počtu 11 sestav náleží 7 sestav k opravě EPIDAT a 4 sestavy, jako průvodce ke geokódování

Ad 5) z celkového počtu 14 maker je 8 použito pro export tabulek a 6 pro otevření tabulek



Název	Popis	Datum změny	Datum vytvoření	Typ
Vytvořit dotaz v návrhovém zobrazení				
Vytvořit dotaz pomocí průvodce				
dPropojeniOkresuObceKU				
dChybneObce	010101	9.4.2010 10:02:51	30.3.2010 18:15:01	dotaz: Výběr
dSmazStaryObsahChybovníkuObci	010102	18.4.2010 22:01:32	23.3.2010 18:40:07	dotaz: Výběr
dVytvorEPIDAT_OPRAVA_OBEC	010102	18.4.2010 22:29:21	23.3.2010 21:09:42	dotaz: Odstr
dPridajChybneObceDoChybovníku	010103	20.4.2010 3:02:42	30.3.2010 0:01:57	dotaz: Vytvř
dAktualizujVChybovníkuObciZpravuOChybach	010103	20.4.2010 19:56:47	23.3.2010 19:07:50	dotaz: Přidáv
dAktualizujeEPIDAT_OPRAVA_OBECOPravenaObce	010104	20.4.2010 22:41:08	23.3.2010 19:04:39	dotaz: Aktual
dNahledOpravy1	010106	16.4.2010 22:55:44	30.3.2010 0:08:43	dotaz: Aktual
dChybneMistoNakazy	010107	17.4.2010 10:18:27	31.3.2010 21:06:40	dotaz: Výběr
dSmazStaryObsahChybovníkuMistaNak	010201	15.5.2010 22:35:59	23.3.2010 18:46:32	dotaz: Výběr
dPridajChybneMistoNakazyDoChybovníku	010202	8.4.2010 23:02:50	23.3.2010 21:10:55	dotaz: Odstr
dAktualizujVChybovníkuMistoNakZpravuOChybach	010203	8.4.2010 23:03:05	23.3.2010 18:59:40	dotaz: Přidáv
dAktualizujeEPIDAT_OPRAVA_OBECOPravenaMistaNakazy	010204	9.4.2010 18:09:17	23.3.2010 21:02:59	dotaz: Aktual
dNahledOpravy2	010206	20.4.2010 22:19:51	30.3.2010 0:44:19	dotaz: Aktual
dChybneMistoOnemocneni	010207	17.4.2010 10:19:40	31.3.2010 21:07:10	dotaz: Výběr
dSmazStaryObsahChybovníkuMistaOnem	010301	20.4.2010 22:44:46	23.3.2010 18:46:45	dotaz: Výběr
dPridajChybneMistoOnemocneniDoChybovníku	010302	8.4.2010 23:06:29	23.3.2010 21:11:19	dotaz: Odstr
dAktualizujVChybovníkuMistoOnemZpravuOChybach	010303	8.4.2010 23:06:47	23.3.2010 18:58:25	dotaz: Přidáv
dAktualizujeEPIDAT_OPRAVA_OBECOPravenaMistaOnemocneni	010304	9.4.2010 18:09:34	23.3.2010 21:04:32	dotaz: Aktual
dNahledOpravy3	010306	20.4.2010 22:49:37	30.3.2010 2:17:53	dotaz: Aktual
dChybneObce_OKRES	010307	13.5.2010 21:10:02	31.3.2010 21:07:40	dotaz: Výběr
dSmazChybovníkOkresu_OBEC	020101	13.5.2010 21:10:09	9.4.2010 18:02:22	dotaz: Výběr
dVytvorEPIDAT_OPRAVA_OBEC-OKRES	020102	17.4.2010 9:49:20	2.4.2010 13:46:53	dotaz: Odstr
dPridajChybneObce_NesouhlasIOkresDoChybovníku	020103	13.5.2010 19:06:28	4.4.2010 22:55:07	dotaz: Vytvř
dAktualizujVChybovníkuObciZpravuOChybachVOkresu	020103	17.4.2010 21:56:30	2.4.2010 11:32:46	dotaz: Přidáv
	020104	17.4.2010 9:57:10	3.4.2010 13:52:01	dotaz: Aktual

Obrázek 14: Pracovní prostředí v MS Access

3.6. Výpočet definičních bodů obcí pro následné geokódování

Tato kapitola je věnována výpočtu definičních bodů obcí pomocí definičních bodů katastrálních území. Podmětem této operace bylo získání dat definičních bodů katastrálních území od Katastrálního úřadu pro Olomoucký kraj. Z ISKN (Informační systém katastru nemovitostí) byly vyexportovat souřadnice v následujících formátech: pdf, txt, xml, rtf a postskript. Všechny data jsou v rozsahu pro Olomoucký kraj a byla vyexportována po okresech.

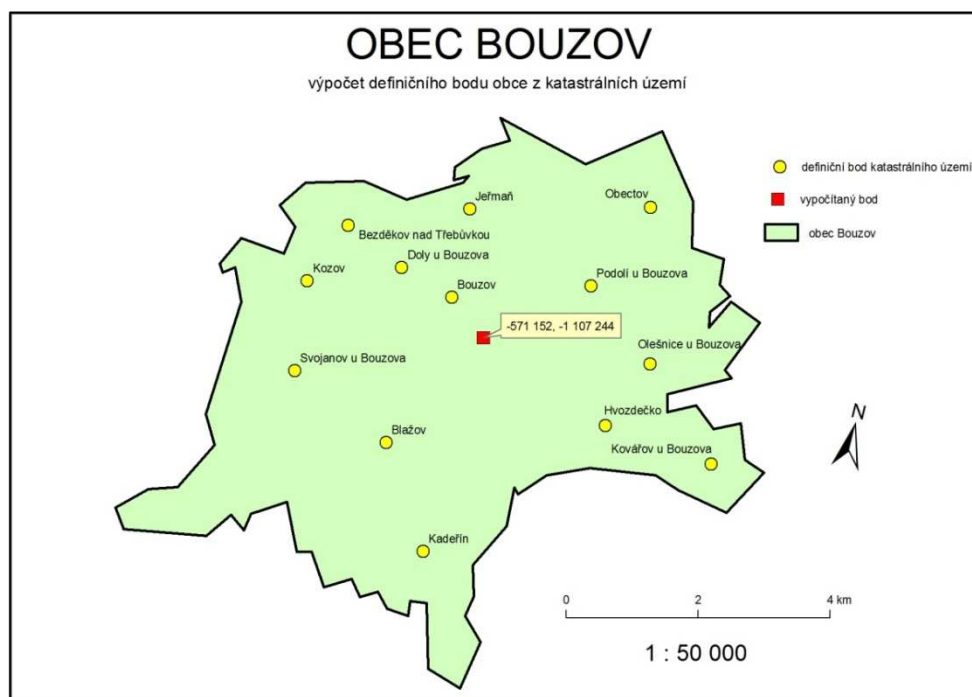
Obec 503941 Libavá			
Katastrální území / prac.č.		Souřadnice definičního bodu	
990078 /	219 Černná u Města Libavá	Y =	523283.00 X = 1116441.00
693421 /	115 Město Libavá	Y =	526439.00 X = 1109214.00
990191 /	234 Rudoltovice	Y =	518808.00 X = 1109752.00
990205 /	235 Slavkov u Města Libavá	Y =	524632.00 X = 1124741.00
990256 /	240 Velká Střelná	Y =	529263.00 X = 1114320.00

Obrázek 15: Ukázka dat z ISKN

Vlastní výpočet definičního bodu obce probíhal prostým výpočtem aritmetického průměru jednotlivých katastrálních území. Důvod, proč provádět tuto akci je takový, že v datech EPIDAT nemůžeme jednoznačně určit, zda záznam pacienta spadá do požadované obce nebo katastrálního území. Proto se vypočítá aritmetický průměr všech katastrálních území, který bude zatížen chybou. Z hlediska přesnosti dat se nejedná o zcela ideální metodu. Musíme však mít na mysli, že epidemie se šíří mnoha směry, lidé se pohybují, cestují a přenášejí epidemie do různých oblastí. Také pro prostorové analýzy je tento způsob přínosnější.

Výpočet lze také provést váženým aritmetickým průměrem, kde kromě souřadnic X, Y je potřeba znát počet obyvatel katastrálního území a výměru území. Tato metoda by byla jistě přesnější, vztahovala by se blíže k větším městům, ale data o počtu obyvatel a výměře nám nebyla na Katastrálním úřadě pro Olomoucký kraj poskytnuta. V ISKN je nebylo možné dohledat pro rozsah Olomouckého kraje.

Celkem je v Olomouckém kraji 765 katastrálních území spadajících pod 395 obcí. Obcí, které mají alespoň 2 katastrální území je 139. Výpočet aritmetického průměru byl proveden v programu MS Excel.



Obrázek 16: Ukázka výpočtu definičního bodu obce z katastrálních území

3.6.1. Elipsa chyb

Při výpočtu definičního bodu obce za pomoci aritmetického průměru z definičních bodů katastrálních území, nám vzniká chyba, kterou je výsledek zatížen. Za pomoci elipsy chyb vypočítáme těžiště, parametry střední elipsy chyb, stočení soustavy a výslednou chybu. Není zapotřebí počítat elipsu chyb pro každou obec, níže uvádím výpočet elipsy chyb pro obec Bouzov.

Tabulka 9: Souřadnice definičních bodů katastrálních území pro obec Bouzov

ID	CISLO_OBCE	NAZEV_OBCE	CISLO_KU	NAZEV_KU	SOURADNICE_Y	SOURADNICE_X
1	500861	BOUZOV	608688	BEZDEKOV NAD TREBUVKOU	573210	1105538
2	500861	BOUZOV	672017	BLAZOV	572637	1108834
3	500861	BOUZOV	608696	BOUZOV	571642	1106629
4	500861	BOUZOV	608700	DOLY U BOUZOVA	572403	1106176
5	500861	BOUZOV	608718	HVOZDECKO	569313	1108579
6	500861	BOUZOV	608726	JERMAN	571370	1105290
7	500861	BOUZOV	672025	KADERIN	572078	1110487
8	500861	BOUZOV	671487	KOVAROV U BOUZOVA	567701	1109160
9	500861	BOUZOV	672033	KOZOV	573833	1106381
10	500861	BOUZOV	724289	OBECTOV	568621	1105265
11	500861	BOUZOV	724297	OLESNICE U BOUZOVA	568628	1107638
12	500861	BOUZOV	724301	PODOLI U BOUZOVA	569524	1106459
13	500861	BOUZOV	672041	SVOJANOV U BOUZOVA	574025	1107743

Tabulka 10: Pracovní tabulka pro počítání jednotlivých hodnot

i	x_i	y_i	x'_i	y'_i	$x'x'$	$y'y'$	$x'y'$
1	1105538	573210	-1706.54	2057.31	2912273.52	4232514.941	-3510874.704
2	1108834	572637	1589.46	1484.31	2526387.98	2203169.325	2359249.988
3	1106629	571642	-615.54	489.31	378887.60	239422.0178	-301187.7041
4	1106176	572403	-1068.54	1250.31	1141774.44	1563269.325	-1336001.858
5	1108579	569313	1334.46	-1839.69	1780787.60	3384467.787	-2454998.627
6	1105290	571370	-1954.54	217.31	3820220.60	47222.63314	-424736.2426
7	1110487	572078	3242.46	925.31	10513556.83	856194.3254	3000274.604
8	1109160	567701	1915.46	-3451.69	3668992.91	11914179.79	-6611583.858
9	1106381	573833	-863.54	2680.31	745698.67	7184049.325	-2314548.781
10	1105265	568621	-1979.54	-2531.69	3918572.52	6409465.941	5011582.296
11	1107638	568628	393.46	-2524.69	154811.98	6374071.249	-993369.3195
12	1106459	569524	-785.54	-1628.69	617070.67	2652638.633	1279400.45
13	1107743	574025	498.46	2872.31	248463.91	8250151.479	1431734.911
Σ	14394179.00	7424985.00	0.00	0.00	32427499.23	55310816.77	-4865058.85

- 1) Vypočteme těžiště – vyrovnanou hodnotu:

$$x_T = \frac{[x]}{n} = 1107244.54$$

$$y_T = \frac{[y]}{n} = 571152.69$$

- 2) Těžiště T posuneme do počátku souřadnic:

$$x'_i = x_i - x_T = \frac{[x]}{n} = \text{výpočet probíhá v pracovní tabulce}$$

$$y'_i = y_i - y_T = \frac{[y]}{n} = \text{výpočet probíhá v pracovní tabulce}$$

Kontrolně vypočteme $\Sigma x'_i$ a $\Sigma y'_i$, má platit, že

$$[x'_i] = 0 \quad [y'_i] = 0$$

- 3) Vypočítáme čtverce empirických středních chyb ve směru os x,y a čtverec kovariance

$$m_x^2 = \frac{[x'x']}{n-1} = 2702291.60$$

$$m_y^2 = \frac{[y'y']}{n-1} = 9218469.46$$

$$(cov(xy)) = m_{xy} = \frac{[x'y']}{n-1} = -810843.14$$

$$(cov(xy))^2 = m_{xy}^2 = \left(\frac{[x'y']}{n-1}\right)^2 = 657466599348.33$$

- 4) Vypočteme parametry střední elipsy chyb a,b,ω

$$m_{max}^2 = \frac{m_x^2 + m_y^2}{2} + \sqrt{\frac{(m_x^2 - m_y^2)^2}{4} + m_{xy}^2} = 9317851.32$$

$$a = \sqrt{m_{max}^2} = 3052.52$$

$$m_{min}^2 = \frac{m_x^2 + m_y^2}{2} - \sqrt{\frac{(m_x^2 - m_y^2)^2}{4} + m_{xy}^2} = 2602909.75$$

$$b = \sqrt{m_{min}^2} = 1613.35$$

$$\omega = \frac{1}{2} \cdot \arctg \frac{2m_{xy}}{m_x^2 - m_y^2} = 6.99^\circ$$

- 5) Míru závislosti (těsnost vazby) vyjadřujeme pomocí Pearsonova korelačního koeficientu pro vyjádření korelace mezi x a y při stejných a různých vahách:

$$r = \frac{[x'y']}{\sqrt{[x'x'] \cdot [y'y']}} = -0.114875$$

- 6) Zápis výsledků vyrovnání

T [x _T , y _T]	→	1107244.54 m,	571152.69 m	
a, b, ω	→	3052.52 m	, 1613.35 m	, 6.99°
r	→	-0.114875		

Výsledná chyba na ose x je v našem případě 3052.52 m a na ose y 1613.35 m. Stočení soustavy je 6.99°. Touto chybou je tedy výsledek definičního bodu pro obec Bouzov zatížen.

3.7. Geokódování

Geokódování je proces, pomocí kterého se databázovým záznamům přiřadí zeměpisné souřadnice, pomocí kterých lze záznam lokalizovat. Z databáze tak vznikne vektorová (bodová) mapová vrstva, se kterou je již možné plnohodnotně pracovat. Databázové záznamy obsahují nejčastěji adresy, čísla popisné, PSČ, případně město, stát nebo různé kombinace všech těchto atributů. Reverzní geokódování je opačný proces. Na základě zeměpisných souřadnic získáváme souřadnice neprostorové. Nějakému místu tedy přiřadíme například nejbližší adresní bod, popisné číslo, město, stát apod.

Pro provádění obou procesů je vždy zapotřebí mít 2 skupiny dat. První skupinou dat jsou vstupní data, která zpracováváme. Druhou skupinu tvoří data, se kterými naše vstupní data konfrontujeme. Tou mohou být například polygony měst a obcí, polygony zemí, adresní body a podobně. Druhá sada musí být samozřejmě vybavena patřičnými atributy. V poslední řadě potřebujeme program, který nám umožní oba procesy provést[6].

Historie geokódování se datuje v roce 1960 ve Spojených státech amerických, kde bylo potřeba najít způsoby mapování dat shromážděných v rámci celé země. Začalo se používat při sčítání lidu, domů a bytů[16].

Úroveň (přesnost) geokódování

Nejpřesnější úrovní geokódování v České republice je geokódování s přesností na adresní bod. Tato úroveň vyžaduje shodu v názvu ulice a popisném, případně orientačním čísle domu.

Pokud se záznam nepodaří dohledat s přesností čísla domu, je umístěn na střed ulice. Jedná se o další úroveň geokódování, která je však stále většinou zcela dostačující pro velkou většinu geografických analýz. Pokud i tuto úroveň přesnosti není možné u některých záznamů dosáhnout, je záznam umístěn na střed části obce nebo na střed obce[5].

3.7.1. Tvorba průvodce geokódováním v ArcGIS Desktop

Po opravě dat databáze EPIDAT a vyexportování výstupních tabulek můžeme zahájit proces geokódování. Geokódování je prováděno v prostředí ArcGIS Desktop. Lze provést v každé z jeho aplikací – ArcMap, ArcCatalog a ArcToolbox, kde se liší jen grafickým provedením.

Součástí přílohy je i průvodce geokódováním, který v jednotlivých krocích popisuje průběh geokódování. Součástí DVD ROM je i toolbox, který si uživatel importuje do ArcToolbox. V nápovědě toolbox pro geokódování je popsán

samotný průběh. Na DVD ROM je také 10 výstupních geokódovaných vrstev ve formátu shp, které byly geokódovány s přesností na obec. Geokódování s přesností na ulice nebylo provedeno z důvodu nedostatku dat uliční sítě. V případě tohoto geokódování by se pomocí SQL dotazů vybraly záznamy v rámci jednoho města a provedlo se následné geokódování.

Průvodce je rozdělen na 4 části:

- 1) Načtení toolbox / Add Toolbox
- 2) Tvorba adresového lokátoru / Create Address Locator
- 3) Vlastní geokódování / Geocode Addresses
- 4) Kontrola a oprava chyb / Review/Rematch Addresses

Úroveň geokódování v této práci lze s přesností na ulice a obce. Adresní body nebyly poskytnuty.

4. DISKUZE

Po opravě celostátní databáze EPIDAT lze konstatovat mnoho věcí. Vysoká chybovost zkoumaných atributů je zde způsobena možností ručního dopisování dat do atributů, které mají být vybírány z číselníku. To je hlavní problém celé databáze. Z pohledu geoinformatika je to problém zcela zásadní, ale pohled lékaře na tuto věc je zřejmě jiný.

Vymyslet skript, který by zaručil správnost opravy je nadlidský výkon, proto je oprava řešena pomocí programu v prostředí databázového programu MS Access. Uživatel se často setkává s momentem rozhodnutí, jak danou ulici, obec případně okres opravit. Proto je důležité si ujednotit správnou opravu a tím předejít chybně opraveným datům.

Největší chybovost byla zjištěna u atributu ulice, kde uživatelé EPIDAT kromě chybně zapsané ulice měli možnost dopisovat i název obce případně katastrálního území. Zde by bylo vhodné založení nového pole v databázi, kde by se z číselníku vybírala katastrální území, které by vhodně doplňovali obec. Případně i založení atributu, kde by uživatelé dopisovali ty záznamy, které by nemohli najít v jakémkoliv číselníku. Dalším vhodným vylepšením by bylo doplnění kódu k ulicím, obcím a okresům. Obec, u které chybí správný okres a vyskytuje se ve více okresech v České republice, nemáme šanci správně opravit.

Při samotné opravě bylo v některých případech těžké rozhodnutí, jak správnou obci opravit. Bylo vycházeno z faktu, že poskytnutá data byla pro Olomoucký kraj a proto se obcím a okresům z tohoto kraje dávala větší váha při rozhodování o správné opravě. Některé případy nebylo možné přesně opravit, tudíž pomocí tohoto programu není možné dosáhnout 100% opravy. Všechny chybné záznamy v EPIDAT se ukládaly do tzv. velkého chybovníku, který lze využít při identifikaci chyb při příjmu nových dat. Oprava dat např. za rok 2010 bude tedy mnohem rychlejší, jelikož se dá předpokládat obdobný výskyt chyb.

Je důležité řešit otázku, zda uživatel, který dostane do ruky program pro opravu dat vytvořeným v této práci, dokáže program správně používat. Intuitivní prostředí programu MS Access, které je jednoduché, přehledně řešené a uživatelsky známé by snad nemělo být překážkou. Navíc je v programu použita celá řada nápověd, které vhodně doplňují každou část opravy.

Tvorba výstupních tabulek a návod na vytvoření geometrické složky k těmto datům, tedy proces geokódování je součástí programu. Obsah průvodce geokódováním v ArcGIS Desktop najdeme také v příloze. Při tvorbě a obsahu výstupních tabulek z EPIDAT byla důležitá konzultace s budoucím uživatelem, který bude nad opravenými daty provádět prostorové analýzy.

5. ZÁVĚR

Cílem této práce bylo navrhnout řešení pro opravu chybných adres v zdravotnické databázi EPIDAT podle celostátně platné databáze UIR-ADR. Dále pak navrhnout a vytvořit postup pro tvorbu geometrické složky dat k této databázi pro potřeby následně prováděných prostorových analýz. Pro obě řešení úpravy a tvorby dat byl navržen automatizovaný postup.

Před samotnou opravou bylo nutné převést obě databáze do jednotné podoby, aby bylo možné správně opravovat. Za pomoci maker v programu MS Excel byla upravena data z celostátně platné databáze UIR-ADR. Struktura opravované databáze EPIDAT, která nám byla poskytnuta Krajskou hygienickou stanicí pro Olomoucký kraj a obsahovala 23 999 záznamů, zůstala nezměněna. Následně byly provedeny úvodní analýzy nad daty, kde byly zjištěny četnosti jednotlivých infekcí v průběhu 5 let. Identifikace chyb byla také provedena před samotnou tvorbou programu a odhalila nám zastoupení chyb v námi sledovaných attributech ulice, obce, okresy, místo nákazy, okres nákazy, místo onemocnění a okres onemocnění. Byla vytvořena úvodní představa o chybovosti v porovnání s daty z UIR-ADR pro Olomoucký kraj a Českou republikou. Tato část již byla prováděna pomocí SQL dotazů v databázovém programu Microsoft Access.

Samotný program pro opravu EPIDAT, který je stěžejní částí této bakalářské práce, se skládá ze 2 průvodců. První průvodce opravou se skládá z 5 na sebe navazujících částí. Nejprve si uživatel nahraje data, poté pomocí průvodců pro opravu obcí, okresů a ulic data opraví a následně těmito opraveným datům přiřadí kódy. Výsledkem je tedy kompletně opravená databáze, která obsahuje kromě původních dat i opravené záznamy s kódy. V dalším kroku je vytvořen „velký chybovník“, do kterého se ukládají všechny nalezené chyby z EPIDAT. Posledním krokem opravy je tvorba agregačních dotazů a výstupních tabulek, které jsou dále užity při procesu geokódování.

V druhém průvodci je uživateli nabízen postup následné tvorby geometrické složky dat k opraveným datům v prostředí programu ArcGIS Desktop. Samotný průvodce geokódováním je také obsažen v příloze této práce. Součástí přílohy je i toolbox, který obsahuje nápovědu s průběhem geokódování a 10 výstupních geokódovaných vrstev ve formátu shp. Geokódování není prováděno pomocí webových služeb nýbrž v prostředí ArcGIS Desktop, kde se vytváří data pro následné prostorové analýzy.

6. POUŽITÉ ZDROJE

Tištěné zdroje:

- [1] Dobešová, Z.: *Prostorové databázové systémy*. Vydavatelství Univerzity Palackého, Olomouc, 2004, 76 s., ISBN 80-244-0891-0.
- [2] Voženílek, V.: *Diplomové práce z geoinformatiky*. Vydavatelství Univerzity Palackého, Olomouc, 2002, 31 s.
- [3] PÍSEK, Slavoj. *Access 2003 : snadno a rychle*. Praha : Grada Publishing a.s., 2004. 124 s. ISBN 80-247-0787-X
- [4] Geokódování a reverzní geokódování. *Geobusiness*. Leden - únor 2009, 01/2009, s. 38-39.
- [5] LACKO, Luboslav. *SQL : Hotová řešení*. Brno : Computer Press, 2003. 296 s. ISBN 80-7226-975-5.

Internetové zdroje:

- [6] *CS map* [online]. 2008 [cit. 2010-01-05]. Dostupný z WWW: <<http://www.csmmap.cz/optimalizace-dopravy/geokodovani.htm>>.
- [7] *Infekce v ČR - EPIDAT* [online]. 2010 [cit. 2010-01-06]. Dostupný z WWW: <<http://www.szu.cz/publikace/data/infekce-v-cr>>.
- [8] *UIR-ADR: Územně identifikační registr adres* [online]. 2005 [cit. 2010-02-05]. Dostupný z WWW: <<http://forms.mpsv.cz/uir/>>.
- [9] *Český statistický úřad* [online]. 2009 [cit. 2010-02-02]. Dostupný z WWW: <http://cs.wikipedia.org/wiki/%C4%8Cesk%C3%BD_statistick%C3%BD_%C3%BA%C5%99ad>.
- [10] *Epidemiologie* [online]. 2009 [cit. 2010-02-05]. Dostupný z WWW: <<http://cs.wikipedia.org/wiki/Epidemiologie>>.
- [11] *Informační technologie a jejich využití ve zdravotnictví* [online]. 2010 [cit. 2010-02-10]. Dostupný z WWW: <http://www.czso.cz/csu/redakce.nsf/i/informacni_technologie_ve_zdravotnictvi>.

- [12] *Ústav zdravotnických informací a statistik* [online]. 2007 [cit. 2010-03-19]. NZIS. Dostupné z WWW:
<http://www.uzis.cz/article.php?order=publikace&sort=asc&type=4&mnu_id=7200&mnu_action=select>.
- [13] *Seznamte se s ArcGIS* [online]. Praha : ARCDATA PRAHA, 2001 [cit. 2010-05-19]. Dostupné z WWW:
<http://gis.vsb.cz/GIS_Ostrava/GIS_Ova_2002/char_ArcGIS.pdf>.
- [14] *Corel* [online]. 2010 [cit. 2010-04-12]. About Corel. Dostupné z WWW:
<<http://www.corel.com/servlet/Satellite/eu/en/Content/1150905725000>>.
- [15] *Radek Jureček & Excel* [online]. 2010 [cit. 2010-05-18]. RJ Tools. Dostupné z WWW: <<http://rjurecek.ic.cz/rjtools.htm>>.
- [16] *NCJRS* [online]. 1999 [cit. 2010-05-18]. Spatially enabling the data: What is geocoding?. Dostupné z WWW:
<http://www.ncjrs.gov/html/nij/mapping/ch4_3.html>.
- [17] Wong, A., Crosier, S.: *Geocoding in ArcGIS Tutorial*, ESRI, Redlans, USA, 2008, <[Geocoding_in_ArcGIS_Tutorial.pdf](#)>

SUMMARY

The aim of this work was to propose resolution for correction of false addresses from health database EPIDAT according to statewide valid database UIR-ADR. Next aim was to propose and make process for the creation of geometric component to this database for needs consequential spatial analyses. For both resolutions was proposed appropriate automated process.

Before this correction was necessary to transfer both databases to unit form to would be correct rightly. With helping of macros in MS Excel were adjusted data from statewide valid database UIR-ADR. Structure of correcting database with 23 999 entries stayed the same. Consequently there were done preliminary analyses of data where was detected frequency of single infection during 5 years. Identification of mistakes was also done before creation of program and it revealed us number of mistakes in attributes streets, towns, districts, place of infection, district of infection, place of illness and district of illness. There was created preliminary conception of mistakes in comparsion with data from UIR-ADR for Olomoucký region and the Czech Republic. This part was already done with helping SQL inquiries in database program Microsoft Access.

The program for correction EPIDAT, which is the most important part of this bachelor thesis, is consist of 2 manuals. The first manual of correction is consist of 5 parts. Firstly user upload the data, then he correct data with helping of manuals for correction of towns, districts and streets and consequently he assign codes to these corrected data. As a result is completly corrected database which contains besides orginal data also corrected data with codes. In next step there is created „big mistaker“, in which are saved all found mistakes from EPIDAT. The last step of correction is creation of agregation inquiries and output tables, which are then used during process geocoding.

In the second manual there is offered to user process consequential creation of geometric component to corrected data in program ArcGIS Desktop. Single manual of geocoding is also contained in attachment of this bachelor thesis. The part of attachment is toolbox which contains help with course of geocoding and 10 output geocoded layers in format shp. Geocoding is not done with helping of web services but in program ArcGIS Desktop where is created data for consequential spatial analyses.

PŘÍLOHY

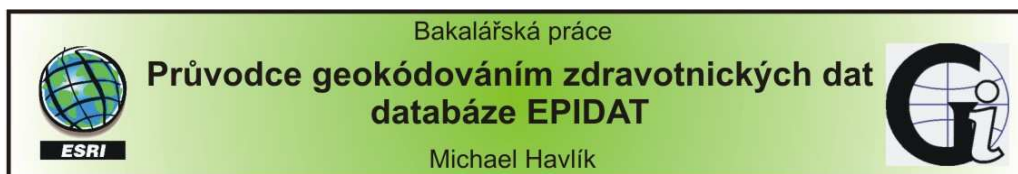
SEZNAM PŘÍLOH

1. Průvodce geokódováním

DVD ROM:

2. Program pro opravu chyb z EPIDAT
3. Výstupní tabulky z EPIDAT
4. Velký chybovník z EPIDAT
5. Toolbox pro geokódování

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.



Průvodce geokódováním

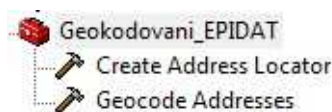
Po opravě dat databáze EPIDAT a vyexportování výstupních tabulek ve formátu dbf nebo xls můžeme zahájit proces geokódování. Jedná se o proces přiřazení zeměpisných souřadnic neprostorovým datům, zde v souřadnicovém systému S-JTSK. Geokódování je prováděno v prostředí software ArcGIS Destkop. Lze ho provést ve všech jeho aplikačních částí: ArcMap, ArcCatalog, ArcToolbox. Vlastní postup je ve všech aplikačních částech téměř stejný, liší se v některých případech pouze grafickým provedením.

V této práci je možné geokódovat s přesností na ulice a obce. Lze geokódovat i s přesností na adresní bod, ovšem ty nejsou obsahem naší databáze EPIDAT.

Níže je uvedeno geokódování v prostředí ArcToolbox, které je jednoduché na obsluhu uživatelem a byla zde vytvořena nápověda pro přehledné ovládání. V jednotlivých krocích je ukázáno geokódování obcí. Tento průvodce geokódováním je i součástí průvodce opravou EPIDAT.

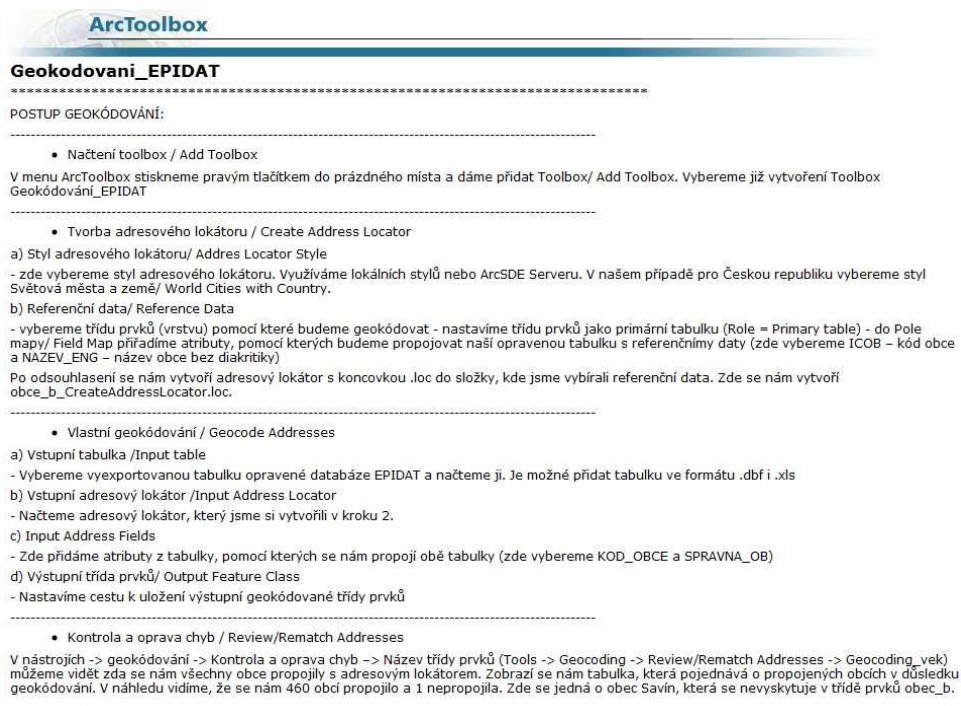
1. Načtení toolbox / Add Toolbox

V menu ArcToolbox stiskneme pravým tlačítkem do prázdného místa a dáme přidat toolbox/ Add Toolbox. Vybereme již vytvořený toolbox Geokódování_EPIDAT.



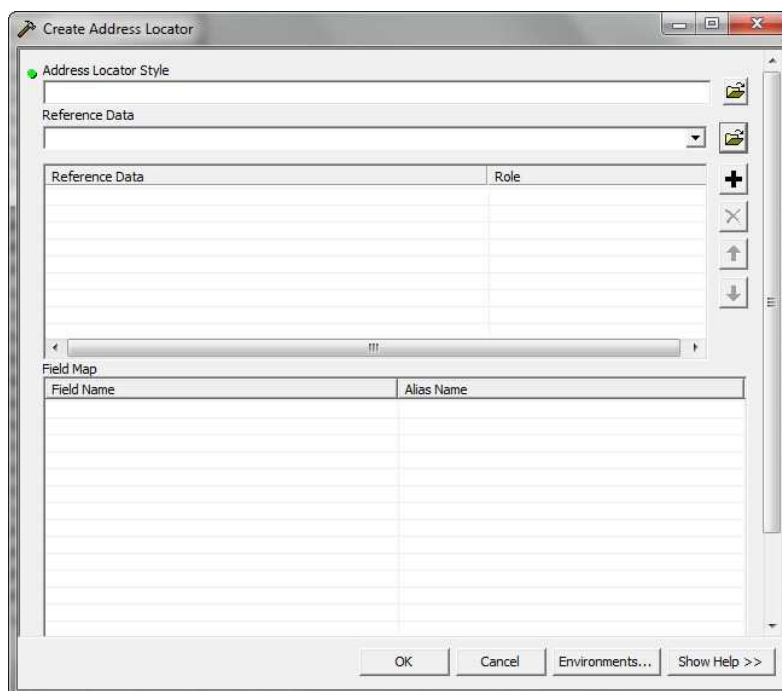
Obrázek 17: Toolbox Geokódování_EPIDAT

Součástí toolboxu je i nápověda s postupem pro geokódování viz. obrázek č. 18. Nápovědu otevřeme pravým tlačítkem na toolbox -> Help.



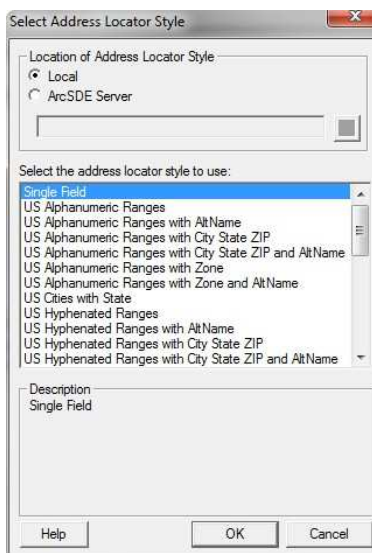
Obrázek 18: Nápopvěda v prostředí ArcToolbox

2. Tvorba adresového lokátoru / Create Address Locator



Obrázek 19: Tvorba adresového lokátoru

a) Styl adresového lokátoru/ Address Locator Style



Obrázek 20: Styl adresového lokátoru

- zde vybereme styl adresového lokátoru. Využíváme lokálních stylů nebo ArcSDE Serveru. V našem případě pro Českou republiku vybereme styl Světová města a země/ World Cities with Country.

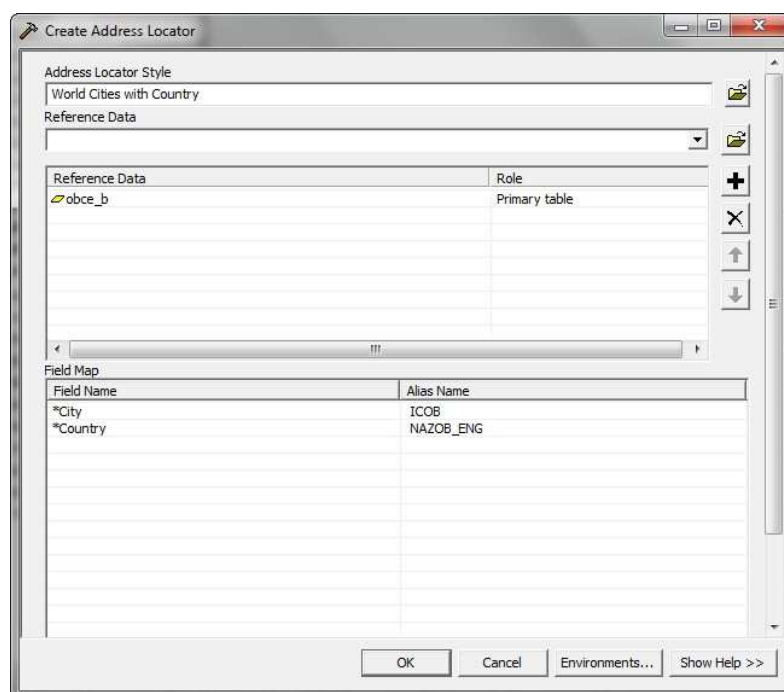
b) Referenční data/ Reference Data

- vybereme třídu prvků (vrstvu) pomocí které budeme geokódovat (zde použitá bodová vrstva obcí pro ČR)
- nastavíme třídu prvků jako primární tabulku (Role = Primary table)

c) Pole mapy/ Field Map

- přiřadíme atributy, pomocí kterých budeme propojovat naši opravenou tabulku s referenčními daty (zde vybereme ICOB – kód obce a NAZEV_ENG – název obce bez diakritiky)

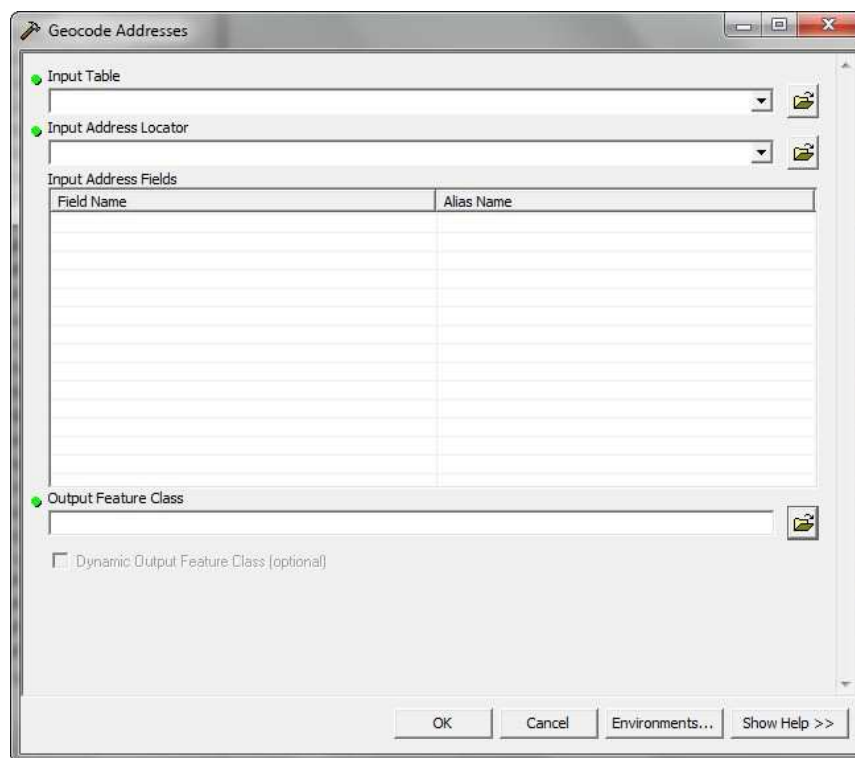
Vyplněná tabulka bude vypadat takto:



Obrázek 21: Výsledná tabulka adresového lokátoru

Po odsouhlasení se nám vytvoří adresový lokátor s koncovkou .loc do složky, kde jsme vybírali referenční data. Zde se nám vytvoří **obce_b_CreateAddressLocator.loc**.

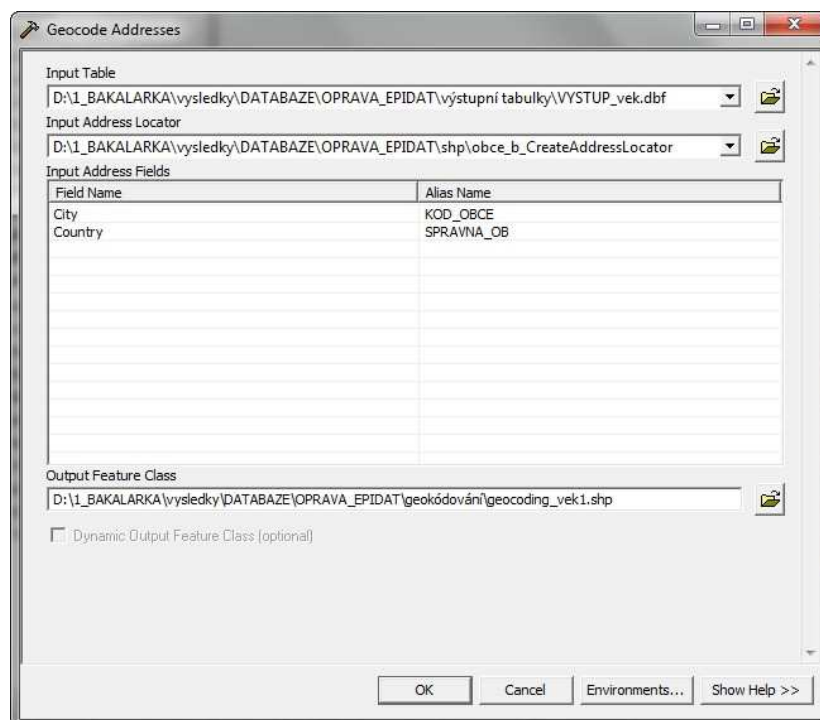
3. Vlastní geokódování / Geocode Addresses



Obrázek 22: Geokódování adres

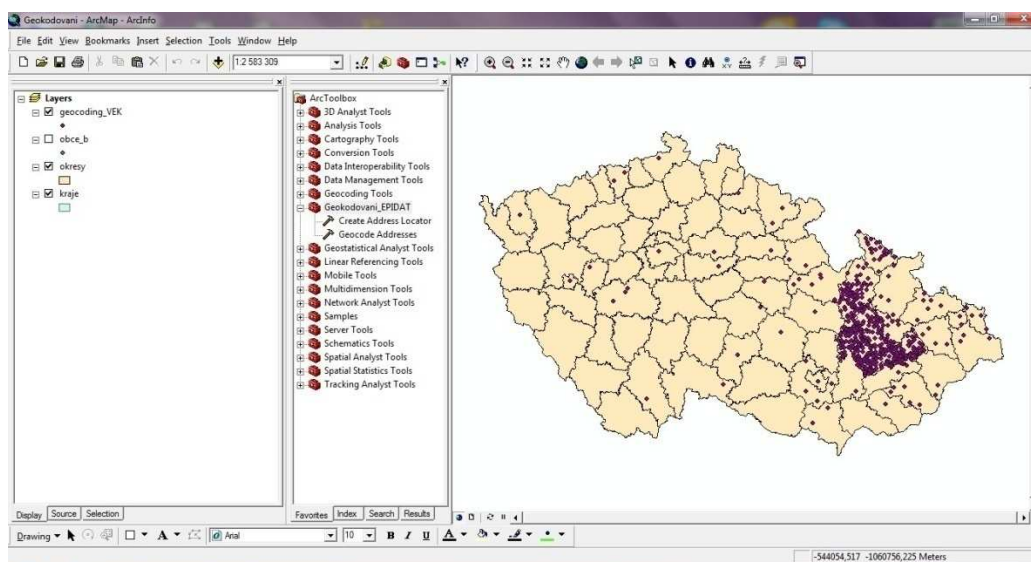
- a) Vstupní tabulka / Input table
 - Vybereme vyexportovanou tabulku opravené databáze EPIDAT a načteme ji. Je možné přidat tabulky ve formátu dbf, xls, csv, třídy prvků i data ze souborové a osobní geodatabáze.
- b) Vstupní adresový lokátor / Input Address Locator
 - Načteme adresový lokátor, který jsme si vytvořili v kroku 2.
- c) Vstupní adresní pole/ Input Address Fields
 - Zde přidáme atributy z tabulky, pomocí kterých se nám propojí obě tabulky (zde vybereme KOD_OBCE a SPRAVNA_OB)
- d) Výstupní třída prvků/ Output Feature Class
 - Nastavíme cestu k uložení výstupní geokódované třídy prvků

Vyplněná tabulka bude vypadat takto:



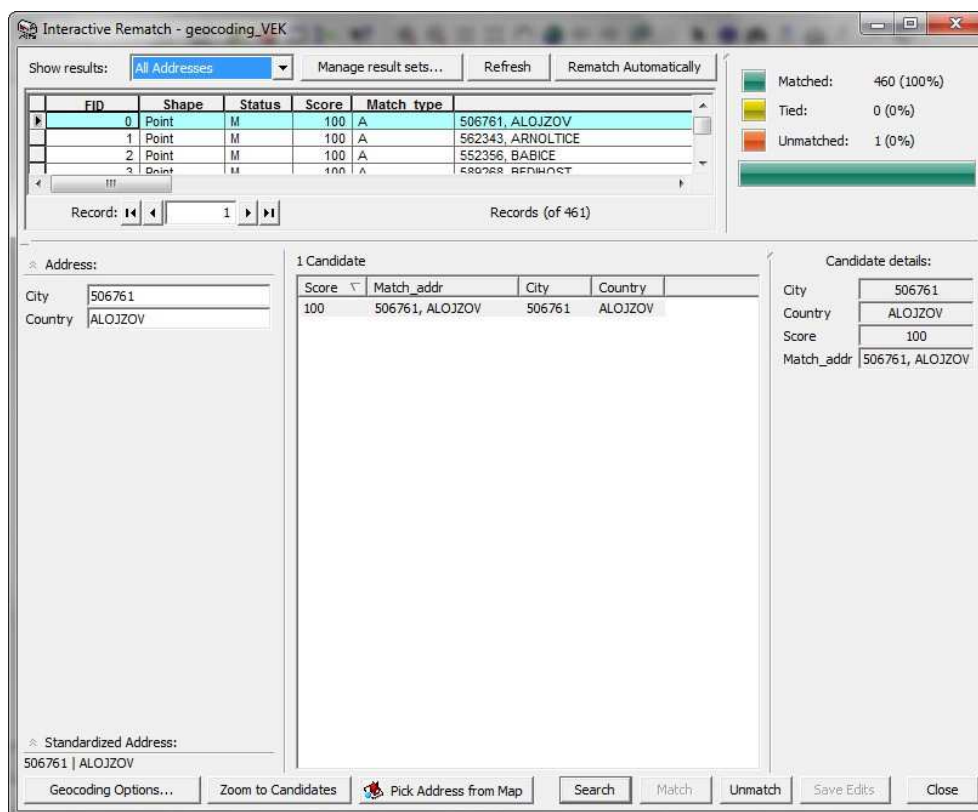
Obrázek 23: Výsledná tabulka geokódování adres

Po dokončení geokódování se vytvoří nová vrstva se všemi atributy z tabulky a z adresového lokátoru. Nyní můžeme provádět různé vizualizace a prostorové analýzy. Výsledek můžeme vidět na obrázku č. 24.



Obrázek 24: Výsledek geokódování

4. Kontrola a oprava chyb / Review/Rematch Addresses



Obrázek 25: Kontrola a oprava chyb

V nástrojích -> geokódování -> Kontrola a oprava chyb -> Název třídy prvků (Tools -> Geocoding -> Review/Rematch Addresses -> Geocoding_vek) můžeme vidět zda se nám všechny obce propojily s adresovým lokátorem. Zobrazí se nám tabulka, která pojednává o propojených obcích v důsledku geokódování. V náhledu vidíme, že se nám 460 obcí propojilo a 1 nepropojila. Zde se jedná o obec Savín, která se nevyskytuje v třídě prvků obec_b.