Univerzita Palackého v Olomouci
Přírodovědecká fakulta
Katedra geoinformatiky

PROSTOROVÉ ANALÝZY VEŘEJNĚ DOSTUPNÝCH DAT EKONOMICKÝCH SUBJEKTŮ V ČESKÉ REPUBLICE

Magisterská práce

Vojtěch CÍCHA

Vedoucí práce Mgr. Pavel TUČEK, Ph.D

Olomouc 2015
Geoinformatika

ANOTACE

Tato diplomová práce se zabývá tématem českých veřejně dostupných dat. Zaměřuje se konkrétně na oblast ekonomiky. Nejdříve provádí průzkum aktuální situace, prochází běžné i méně známé zdroje a sestavuje z nich přehled. Následně se věnuje podrobněji Registru ekonomických subjektů, popisuje jeho atributy a způsob přípravy těchto dat v databázovém prostředí PostgreSQL do podoby vhodné pro analýzy. Jedním z kroků je i vlastní návrh operace geokódování, kdy jsou k prvkům přiřazovány prostorové souřadnice porovnáváním s databází RÚIAN.

Analytická část je rozdělena do tří případových studií. V první je využito potenciálu dlouhého časového rozsahu databáze RES a zkoumá se vývoj ekonomického prostředí v České republice od roku 1990 do současnosti, nejčastěji formou vizualizace kumulativních součtů aktivních ekonomických subjektů. Současně je na datech provedena analýza přežití. V druhé případové studii se práce soustředí na popis oblastí České republiky z pohledu koncentrace různých cílových skupin ekonomických subjektů. Koncentrace je stanovená pro obce v přepočtu na počet obyvatel, a opět je brán v potaz jak aktuální stav, tak minulý, nebo rozdíl mezi nimi. V poslední případové studii je provedeno vícerozměrné shlukování s cílem vytvořit nové prostorové rozdělení České republiky z pohledu ekonomických aktivit.

KLÍČOVÁ SLOVA

veřejně dostupná data; ekonomická data; zpracování dat; prostorové analýzy;

Počet stran práce: 64
Počet příloh: 12 (z toho 2 volné a 10 elektronické)

ANOTATION

This diploma thesis is concentrating on topic of Czech publicly available data, especially economic. Firstly it makes survey of current situation, walks through more and also less known sources and compiles them into an overview. It focuses more on database of Czech Business Register (Register of Economic Subjects - RES), describing its attributes and methods of data preparations to form suitable for analyses in database environment PostgreSQL. One of these methods is also new concept of algorithm for geocoding, where are spatial coordinates assigned to elements via comparing of address with database RÚIAN.

Analytical part is divided into three case studies. In first one, there is use of advantage of long time range of database RES in examining of evolution of economic environment in the Czech Republic, from 1990 to present. The most common way to do that is by visualising cumulative totals of active business entities. There is also made survival analysis of these data. In second case study, there are regions of Czech Republic described by concentration of different target groups of business entities. The concentration is calculated for municipalities in relation with number of inhabitants and there is evaluated both current and past states, also with a difference between them. In the last case study is performed multidimensional clustering to create a new spatial division of the Czech Republic in terms of economic activities.

KEYWORDS

publicly available data; economic data; data preparation; spatial analysis;

Number of pages: 64

Number of appendixes: 12

Prohlašuji, že

- diplomovou práci včetně příloh, jsem vypracoval samostatně a uvedl jsem všechny použité podklady a literaturu.

- jsem si vědom, že na moji diplomovou práci se plně vztahuje zákon č.121/2000 Sb. - autorský zákon, zejména § 35 – využití díla v rámci občanských a náboženských obřadů, v rámci školních představení a využití díla školního a § 60 – školní dílo,

- beru na vědomí, že Univerzita Palackého v Olomouci (dále UP Olomouc) má právo nevýdělečně, ke své vnitřní potřebě, diplomovou práci užívat (§ 35 odst. 3),

- souhlasím, aby jeden výtisk diplomové práce byl uložen v Knihovně UP k prezenčnímu nahlédnutí,

- souhlasím, že údaje o mé diplomové práci budou zveřejněny ve Studijním informačním systému UP,

- v případě zájmu UP Olomouc uzavřu licenční smlouvu s oprávněním užít výsledky a výstupy mé diplomové práce v rozsahu § 12 odst. 4 autorského zákona,

- použít výsledky a výstupy mé diplomové práce nebo poskytnout licenci k jejímu využití mohu jen se souhlasem UP Olomouc, která je oprávněna v takovém případě ode mne požadovat přiměřený příspěvek na úhradu nákladů, které byly UP Olomouc na vytvoření díla vynaloženy (až do jejich skutečné výše).

V Olomouci dne Bc. Vojtěch Cícha

Děkuji Lukáši Markovi za cenné podněty a konzultace, Radku Janoštíkovi za IT rady, Františku Kuchařovi za ekonomický pohled. Za poskytnutá data děkuji Krajské správě ČSÚ v Olomouci.

Dále děkuji ženě a přátelům za podporu.

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

OBSAH

ÚVOD

O dostupnosti a otevřenosti dat se v poslední době živě mluví. Přibývá institucí, které svá data dávají k dispozici pro veřejnost, ať už za určitý úplatek nebo zdarma. S tím souvisí tendence tato data podrobovat výzkumům, hledat skryté souvislosti a vynášet na povrch zajímavé informace, čemuž se obecně říká data mining. Příznivý vliv veřejně dostupných a otevřených ve společnosti byl již několikrát ověřen (opendata.cz, 2014).

Stejným směrem se vydává tato diplomová práce. Sumarizuje současný stav v oblasti českých veřejně dostupných dat s ekonomickou tematikou a vybírá si vhodná data pro další zpracování. Práce s rozsáhlým datovým souborem přináší spoustu výzev, se všemi je snaha demonstrativně se vypořádat tak, aby mohla sloužit jako příkladný popis postupu pro případné zájemce při podobné činnosti. K tomu může být užitečný podrobně komentovaný programový kód, který všechny v práci popisované kroky převádí do konkrétních příkazů v databázovém prostředí PostgreSQL nebo statistickém programu R. Velký potenciál dalšího využití má geokódování pomocí databáze RÚIAN. V práci bude navržen způsob zprovoznění této vlastní geokódovací služby, nezatížené licenčními omezeními jako běžné současné webové služby. Návrh algoritmu geokódování může být následně otestován na datech, čímž bude umožněno hodnocení kvality a úspěšnosti tohoto způsobu geokódování.

Další fáze práce je složena z analyzování připravených dat pomocí tří případových studií. Motivací je v nich extrahovat co největší množství uchopitelných informací a tyto informace představit formou srozumitelných vizualizací v podobě grafů, animací nebo map. Zároveň bude snahou interpretovat nalezené jevy a pokusit se o zdůvodnění příčin, které k těmto situacím vedly. Případové studie jsou koncipovány tak, aby bylo na data aplikováno více různých pohledů, skrze různé analýzy. Je možné, že některé analýzy nepřinesou nové nebo nějakým způsobem zajímavé informace, stále však mohou sloužit jako inspirace pro případné budoucí podobné práce například na jiných datech.

1 CÍLE PRÁCE

Tato magisterská práce je demonstrativním pokusem o odkrytí informačního potenciálu českých veřejně dostupných dat s ekonomickou tematikou. To je hlavní myšlenkou, v rámci níž vyvstávají dílčí cíle práce. Prvním z nich je nalézt vhodná data. Slovo „vhodná“ v sobě obsahuje několik požadavků, plynoucích ze zamýšlených analýz. Pro potřeby časových řad bude nezbytné, aby data obsahovala časové vymezení. Pro potřeby prostorových analýz je naprosto samozřejmá schopnost jednoznačné identifikace umístění, aby data byla vztažitelná k prostorové entitě (jako např. obec nebo adresní bod). Vzhledem k záměru následné regionalizace se zvyšuje nárok na datovou sadu požadavkem existence záznamů pro podrobné oblasti ale zároveň pro každou z nich - např. pro všechny obce ČR. Některé analýzy potom mohou být náročné na počet atributů. Záměrem je propojení ekonomických dat se socioekonomickými a prostorovými daty. Taková data je snad nemožné přímo získat, nedílnou součástí práce je tudíž příprava dat. Její podrobný popis, specifický z pohledu práce s enormně velkou datovou sadou, který zároveň může sloužit jako inspirace pro podobné analýzy, je druhým dílčím cílem.

Po přípravě dat přijde na řadu další cíl, analytická část. Budou provedeny vybrané statistické, vícerozměrné, časově zaměřené a prostorové analýzy. Ty budou vybrány na základě doporučení vedoucího práce a inspirace jinými studiemi tak, aby jednak dávalo smysl jejich použití v dané situaci a zároveň aby zapadaly do koncepce tohoto díla v rozsahu odpovídajícím úrovni magisterské práce. Výsledky analýz je potřeba co nejlépe vizualizovat a srozumitelně interpretovat, což se stává dalším cílem práce.

Výsledky práce mohou případným zájemcům pomoci v hledání vhodného datového souboru mezi dostupnými ekonomickými daty, při snahách o jejich zpracování, nebo i při provádění shodných analýz. Všem čtenářům práce může být užitečná z pohledu zjištěných informacích z ekonomického prostředí v České republice za posledních bezmála 25 let.

2 POSTUP ZPRACOVÁNÍ

V této kapitole je nastíněna základní kontinuální linie práce. Ta je koncipována tak, jak logicky na sebe navazují části běžné geoinformatické úlohy.

1. Příprava dat

Základní stavební kámen kvalitní práce jsou kvalitní data. Nejprve byl proveden průzkum současné situace veřejně dostupných dat v České republice, jehož výstupem je soupis poskytovatelů a jejich datových sad se základními charakteristikami. Z nich jsou vybrány pro tuto práci vhodné sady. Ty jsou dále patřičně podrobněji popsány a uloženy do databáze. Následuje proces zpracování atributů dat, jejich úpravy, přepočty, vyzkoušení procesu geokódování a propojování s dalšími daty tak, aby výsledná data byla celistvá a přímo použitelná pro nadcházející analýzy.

2. Provedení případových studií

Byly vybrány tři případové studie. Jedná se v podstatě o tři skupiny analýz, které by měly dohromady využít a obsáhnout celé spektrum informací dosažitelné z připravených dat. Postup u jednotlivých analýz je vzájemně podobný. Nejprve proběhne výběr konkrétních prvků z celku dle atributů/polohy, případně je tento výběr agregován do plošných jednotek, proběhne výpočet, připraví se výstup a vše se náležitě okomentuje.

První studie: Vizualizace časových řad: Vznik a zánik ekonomických subjektů

V první případové studii je nejdůležitějším atributem čas. Bylo provedeno časové porovnání mezi různými obdobími, vykreslování kumulativních součtů, vytvořeny statické i animované vizualizace časového vývoje.

Druhá studie: Mapování hlavních sídel ekonomických subjektů a jejich vzory v prostoru a čase

V této studii je aplikován prostorový pohled na analyzovaná data. Jednotlivé údaje a jejich charakteristiky jsou prostorově agregovány na úroveň obcí ČR a následně jsou mezi sebou porovnávány. Opět je provedeno časové porovnání.

Třetí studie: Klasifikace zón ekonomické aktivity

Poslední studie přichází se záměrem nové regionalizace České republiky z pohledu ekonomické aktivity. Jako hlavní metoda analýzy bylo zvoleno vícerozměrné shlukování, jehož cílem je hledání na první pohled neviditelných souvislostí a testování chování dat. Jeho výsledkem je zjištění, zda některé regiony vykazují podobné charakteristiky či chování, které může být v nějakém ohledu označeno jako typické pro celou skupinu. Identifikované skupiny by měly být odlišné od ostatních oblastí. Současně je úkolem analýzy zjistit, zda se jevy v prostoru republiky vyskytují náhodně bez prokazatelných pravidelností či existuje jejich prostorový vzor.

Pro korektní provedení analýz stejně jako pro dosažení specifikovaných cílů práce je nezbytné dobře znát svá data. To platí i o jejich atributech s ekonomickými charakteristikami. Autor se však nepovažuje za dostatečně zběhlého v ekonomickém prostředí; pro správné pochopení ekonomických dat stejně jako s výstupy analýz, v nichž

se objeví nějaký ekonomický jev, je snahou autora dopomoci se ke správné interpretaci využitím konzultací s lidmi ekonomiky znalejšími.

3 ZDROJE INFORMACÍ A INSPIRACE

Oblast přípravy a zpracování dat

Protože bude v práci zacházeno s velkými datovými soubory (řádově miliony záznamů), odpadá možnost použít pro přípravu dat běžné tabulkové procesory. Období získávání prvních zkušeností s databázovým prostředím je však relativně náročné, zejména na čas. Vzhledem k tomu, že se autoru nepodařilo nalézt dílo, které by podrobně popisovalo způsob přípravy dat a samotné práce s enormně velkou datovou sadou, bylo nutno začít od začátku a průběžně hledat řešení na všechny problémy, na něž se při takové práci narazí.

Hlavními zdroji informací se kromě konzultací s odborníkem IT stala oficiální dokumentace PostgreSQL (obsahující tutoriály, informace o jazyku SQL o jeho struktuře a možných funkcích, možnosti nastavení databáze apod.) a nezastupitelnou roli hrají uživatelská fóra, jmenovitě například Stack Overflow (stackoverflow.com), fungující na principu konkrétních otázek a jejich řešení. Obecně je vysoká pravděpodobnost, že problém, na něž člověk v praxi narazí, už někdo před ním řešil, takže nejspíš bude existovat problémová otázka i s řešením. Čím uživatelsky větší fórum, tím větší pravděpodobnost existence řešení. Potom stačí jen se umět správně zeptat, resp. najít správnou otázku (pochopitelně v angličtině).

Prostorové analýzy ekonomických dat

Data s ekonomickou tematikou tvoří vděčný základ prostorovým analýzám. Lidstvo se totiž přirozeně zajímá, kde se druzí mají lépe (či hůř) a ptá se po důvodech. Právě prostorové analýzy nabízí nástroje pro vymezení a kvantifikaci (socio)ekonomických jevů v různě velkém prostoru s různou podrobností, od místní povahy (rozlišení v rámci města a jeho nejbližším okolí) k mnohem větším oblastem (kraje, státy). Následující studie byly zdrojem inspirace této diplomové práce, představují příklady využití prostorových analýz pro socioekonomickou klasifikaci území.

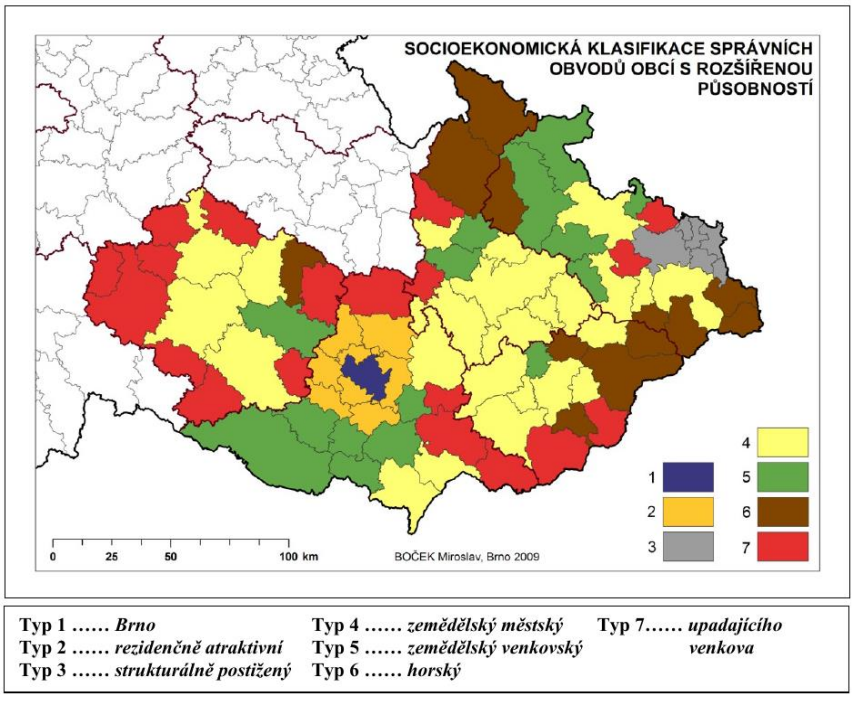
Představitelem lokální prostorové studie může být článek *Geographic clustering of firms and urban form: a multivariate clustering* (Maoh, Kanaroglou, 2007), kde na území kanadského města Hamilton a jeho nejbližšího okolí (75 km jihozápadně od Toronta, plocha odpovídající průměrnému okresu ČR) byly srovnány hustotní mapy firem z roku 2001 a 2006. Výsledkem bylo potvrzení výskytu největších shluků firem v centrální oblasti ale zároveň trendu obecné decentralizace firem (stěhování z centra do okrajových oblastí) s výjimkou odvětví ubytování, stravování a vzdělávání. Dále byla testována prostorová autokorelace jednotlivých oblastí působení firem se závěrečným konstatováním její existence téměř ve všech sektorech kromě vzdělávání a administrativní složky (Maoh, Kanaroglou, 2007).

V českém prostředí od roku 2001 na Institutu Geoinformatiky VŠB-TUO probíhaly v rámci programů dotovaných Grantovou agenturou ČR projekty na prostorové analýzy nezaměstnanosti a trhu práce pro lokální oblasti[1]. Jedním z cílů bylo dopomoci pracovníkům úřadu práce k znalostem a nástrojům pro zpracování prostorových analýz pracovního trhu v praxi na základě různých ukazatelů za jednotlivé obce, včetně kartografické vizualizace výstupů (Institut Geionformatiky VŠB-TUO, 2005). Probíhala

[1] pozn. http://gis.vsb.cz/pan-old/index.htm, http://gis.vsb.cz/pan/index.php

různá školení a byly vytvořeny příkladové studie (Horák a kol., 2004, Horák a Šimek, 2000) pro některé okresy Moravskoslezského kraje. Navíc v rámci školících dokumentů byl vytvořen *Úvod do analýzy časových řad* (Hančlová, Tvrdý 2003), popisující základy práce s časovými řadami jako například vysvětlení statistických charakteristik nebo způsoby vizualizací, což se ukázalo jako užitečné pro časově analytickou případovou studii této magisterské práce.

O něco větší plochu analyzoval ve své diplomové práci z roku 2009 M. Boček. Soustředil se na východní část Česka, konkrétně Vysočinu, Zlínský, Jihomoravský, Olomoucký a Moravskoslezský kraj, v podrobnosti na obvod obce s rozšířenou působností (ORP). Pro každou ORP sesbíral dostupné sociologické (hustota zalidnění, stupeň urbanizace, počet rozvodů, podíl cizinců, index stáří, …), socioekonomické (podíl osob zaměstnaných v priméru, cena pozemků, míra nezaměstnanosti, průměrná mzda, …) i fyzickogeografické (podíl zemědělské půdy, ekologická stabilita krajiny) atributy. Celkově se jednalo o 20 vybraných atributů, většina z nich byla platná pro rok 2007, některé však byly starší (zejména ukazatele pocházející ze Sčítání lidu, domů a bytů - SLDB) a to až z roku 2001. Analytická část spočívala v provedení faktorové analýzy, čímž bylo dosaženo zhuštění informace do pěti faktorů, a následného shlukování (Boček, 2009). Velmi podrobně ve své práci popsal proces přípravy dat, jednotlivé nastavení analýz a interpretaci výsledků včetně kritického pohledu. Výstupní kartogram shlukování viz obr. 1. V závěru navíc nabízí porovnání jeho studie s prací P. Votruby, který prováděl již v roce 2003 shlukování s podobným cílem, a to hodnocení socioekonomického vývoje regionů, pro celou Českou republiku na úrovni okresů na datech ze SLDB 1991 a 2001 (Votruba, 2003). Zajímavé je, že přes poměrně významný rozdíl ve vstupních atributech se obě práce ve třech faktorech z pěti významově téměř shodly; ve výsledné klasifikaci se potom objevily jak stejně vyhodnocené regiony (např. zázemí Brna, Ostravsko a Karvinsko) tak i diametrálně odlišné (např. Vsetínsko a Frýdecko-Místecko) (Boček, 2009).



Obr. 1 Socioekonomická typologie správních obvodů ORP východní části České republiky (zdroj: Boček, 2009)

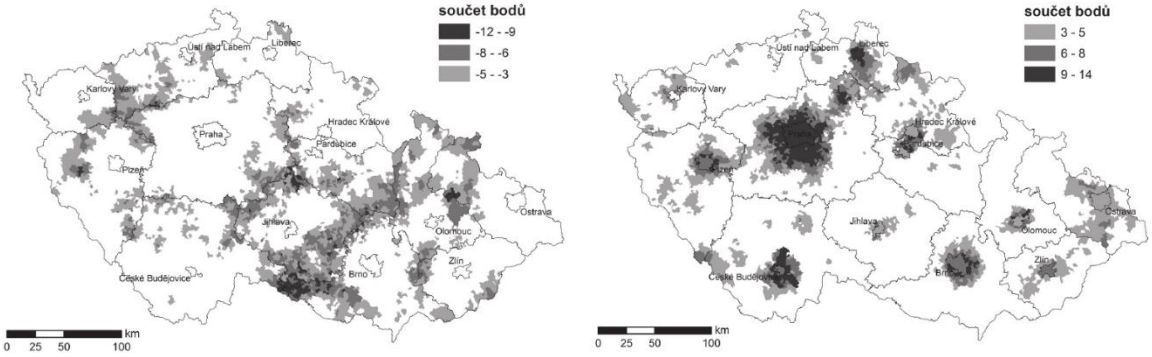
Další inspirativní studie provedená pro oblast celé ČR je popsána v článku *Prostorové vzorce sociálně-ekonomické diferenciace obcí v České republice* (Novák, Netrdová, 2011), jež vyšel v Českém Sociologickém časopise. Autoři kromě hlavního cíle provést jmenovanou analýzu chtějí dále zodpovědět následující klíčové otázky (Novák, Netrdová 2011, s. 718): „Jakým způsobem a do jakých regionálních celků je území diferencováno? Které diferenciační dimenze jsou pro formování těchto celků určující? Je vůbec možné je identifikovat pomocí dostupné datové základny, a případně jaké analytické postupy jsou pro jejich vymezení nejvhodnější? Jaké jsou prostorové průměty často sledovaných dichotomií typu rozvojový × zaostávající, centrální × periferní, atraktivní × neatraktivní? Jedná se z vývojového hlediska pouze o posílení tradičních prostorových vzorců konzervovaných v období komunismu, nebo se vytváří zcela nová forma sociálně prostorové diferenciace?“

V první části příspěvku se zabývají teoretickým zázemím studie, zejména přístupy k diferenciaci území a její metodologii. Mimo jiné zmiňují rizika využití faktorové a shlukové analýzy v aplikaci pro prostorová data, a to „... nevhodnost použití faktorové analýzy pro analýzu vnitřně heterogenních území a necitlivost obou metod k prostorovým souvislostem mezi studovanými jednotkami. Výsledky jsou při prostorové vizualizaci často velmi nepřehledné a necharakterizují přirozené územní celky.“ (Novák, Netrdová 2011, s. 724). Za vhodnější volí metody prostorové autokorelace, konkrétně jmenují Moranovo I kritérium a analýzu LISA (local indicators od spatial association) - lokální míry prostorové autokorelace (Novák, Netrdová 2011), kterých posléze využili v popisované studii. Vstupním souborem dat se stala databáze údajů pro obce ČR vzniklá při přípravě *Atlasu sociálně-prostorové diferenciace České republiky* (Ouředníček, Temelová, Pospíšilová, 2011), zúžená po korelační analýze na celkový počet 25 ukazatelů. Data pochází z veřejně dostupných zdrojů, časová platnost je různá mezi léty 2001 – 2010, s tím, že někdy jsou to konkrétní hodnoty a někdy se jedná o roční průměry (Novák, Netrdová 2011). Absolutní hodnoty a počty jsou relativizovány vztažením na počet obyvatel (příp. počet ekonomicky aktivních obyvatel). Ukazatele jsou rozděleny tematicky do čtyř skupin (Novák, Netrdová, 2011):

1. migrace a mobilita obyvatelstva (migrační saldo a jeho změna, přirozený přírůstek, charakteristiky dojížďky, …)
2. sociální struktura a charakteristiky obyvatelstva (index stáří, podíl vysokoškolsky vzdělaných, podíl domácností pobírajících příspěvek na bydlení, volební účast, …)
3. ekonomické prostředí (míra nezaměstnanosti a její změna, počet podnikajících FO, počet samostatných rolníků, podíl zaměstnaných v priméru a sekundéru)
4. charakteristika obcí (celkové a investiční výdaje obce, podíl počtu pracovních příležitostí, intenzita bytové výstavby, počet spojů veřejné dopravy, …)

Po přípravě atributů byla provedena analýza LISA pro každý z nich zvlášť. Výsledky poté bylo zapotřebí sloučit. Autoři zvolili dvě metody. Tou první byla „bodovací metoda“. Ze všech ukazatelů byly vybrány pouze ty, o nichž lze jednoznačně prohlásit, že jejich nízká nebo vysoká hodnota je vnímána pozitivně/negativně z hlediska rozvojového potenciálu území. Za každý pozitivní shluk byl obci do něj náležící přičten jeden bod, za příslušnost k negativnímu shluku přičten bod záporný. Váhy všech ukazatelů zůstaly stejné. Po sečtení bodů za všechny ukazatele došlo k výsledné klasifikaci; za rozvojové byly považovány obce dosahující minimálně dvou bodů, za problémové obce dosahující maximálně minus dvou bodů (Novák, Netrdová, 2011). Výsledek je možné vidět na obr. 2.

Zjednodušeně řečeno tento způsob ukázal problémové oblasti na hranicích krajů (zejména Vysočiny) nebo v některých horských oblastech, a rozvojové oblasti jako zázemí krajských měst nebo dobře viditelná osa Plzeň – Praha – Mladá Boleslav – Liberec.



Obr. 2 Problémové a rozvojové oblasti (zdroj: Novák, Netrdová 2011)

Druhou metodou syntézy výsledků analýzy LISA byla hierarchická shluková analýza. Tím, že do shlukování vstupují výsledky prostorově založené LISA, dosahuje se prostorově mnohem souvislejších výstupů než při shlukování konkrétních ukazatelů. Výhodou tohoto řešení (oproti předchozímu dichotomickému rozdělení rozvojový x problémový) byla citlivější kategorizace lépe reflektující vnitřní různorodost územních jednotek (Novák, Netrdová 2011). Ve výsledných shlucích se nakonec vyprofilovaly jádrové oblasti (opět kolem většiny krajských měst, s velkým spádovým územím kolem Prahy, zabírajícím větší polovinu středočeského kraje), nebo území Ostravska (dosahujících často problémových charakteristik ale například v míře nezaměstnanosti významné zlepšení oproti celorepublikovému vývoji), nebo také severní Čechy a Vysočina (jako oblasti strukturálně postižené, se stagnující či dokonce zhoršující se tendencí).

Tento způsob dle autorů ve velké míře potvrdil a doplnil dřívější výzkumy jádrových a problémových oblastí (Novák, Netrdová, 2011).

4 DATOVÉ SADY:

Jedním z dílčích cílů této diplomové práce je poskytnout seznam a popis možností veřejně dostupných datových sad s ekonomickou tematikou pro české prostředí. Slovní spojení „veřejně dostupná“ může obnášet jak bezplatně přístupná data, tak i datové sady, jež zájemce dostane za úplatu. Úhel pohledu na popisované datové soubory by se dal pojmenovat jako „geoinformatický“, tzn. za důležitější jsou považovány sady s prostorovou složkou. Níže uvedený výčet se zaměřuje spíše na poskytovatele a významné datové sady, pokud má poskytovatel v nabídce sad mnoho, jsou uváděny pouze příklady. Zároveň je potřeba brát v úvahu, že se jedná o poměrně živou oblast, kde může často docházet ke změnám. Platnost popisovaných informací je k dubnu 2015.

Otevřená data a situace v ČR

Otevřená data jsou náplní jednoho ze specifických cílů Akčního plánu Strategie rozvoje infrastruktury pro prostorové informace v ČR do roku 2020 (GeoInfoStrategie), schváleného českou vládou dne 8. července 2015. Zde se definují opatření jako vytvoření základní metodiky nebo publikování klíčových sad právě cestou otevřených dat (Ministerstvo vnitra, 2015). S pracovní skupinou pracující na tomto cílu spolupracuje Odborná skupina OS25 - OPEN SOURCE A OPEN DATA České asociace pro geoinformace (CAGI). Ta má v základních cílech a aktivitách definováno “podporovat tvorbu, sdílení a publikování volně dostupných a otevřených geodat a geoinformací” (CAGI, 2013). Směrem k větší dostupnosti dat směřuje i evropská směrnice INSPIRE, v rámci níž má každý stát povinnost publikovat určená prostorová data. Nicméně tyto tendence se začínají objevovat až v současnosti a několik let zpátky, do té doby byla situace v Česku nevalná.

To, co je běžné například v USA, že data veřejné správy jsou volně přístupná komukoliv, zdaleka nebylo a stále není u nás tak rozšířeno. Poslední dobou se však i v České republice objevují konkrétní iniciativy, které se snaží „otevírat dat“ veřejnosti. Zářným příkladem může být Institut plánování a rozvoje hl. m. Prahy, který k 1. dubnu 2015 zpřístupnil širé veřejnosti celkově 86 datových sad, zahrnující m.j. ortofotomapu, digitální model terénu a reliéfu, vrstvy budov, vrstevnic, územních plánů apod., to vše ve vysoké přesnosti a aktuálnosti, ve strojově čitelných formátech a se svobodnou licencí.

Výhody otevřenosti dat podle opendata.cz jsou následující:

Otevřená data přispívají větší transparentnosti veřejné správy, odborná veřejnost získává data pro výzkumnou činnost vedoucí k tlaku na zefektivnění veřejné správy a pro tvorbu aplikací zprostředkovávající tato data širé veřejnosti ve srozumitelné formě. Dále dochází k šetření veřejných financí, protože se zvýší přehled o datech již volně dostupných, snáze se vyhne duplicitnímu pořizování dat a není třeba vynakládat prostředky pro prezentování dat, protože o to se postarají aplikace třetích stran (opendata.cz, 2014).

Nicméně proces otevírání dat probíhá velmi pozvolna. Hlavní propagátoři této myšlenky nabízející zároveň potřebnou infrastrukturu jsou dvě iniciativy, otevrenadata.cz a opendata.cz.

První jmenovaná nabízí na svých webových stránkách informace o otevřených datech obecně, včetně studií dokazujících jejich přínos. Dále se zde nachází soupis významných zdrojů otevřených dat zejména z českého veřejně samosprávního prostředí.

Druhá iniciativa opendata.cz na svých webových stránkách poskytuje seznam datasetů, které jsou vybudovány na principech tzv. Linked Data, (způsob sémantického webu pomocí standardu RDF a identifikátorů URI) umožňující lepší propojování

informací. Obě tyto iniciativy odkazují na asi největší český katalog otevřených datových sad, fungující na open source software CKAN, dostupným na stránkách cz.ckan.net.

Autor doporučuje případnému zájemci o veřejně dostupná data začít vyhledávání právě na stránkách těchto iniciativ. I zde následující výčet zdrojů veřejně dostupných datových sad s ekonomickou tematikou vznikl z velké míry procházením tamních seznamů.

Ministerstvo financí ČR

Při snaze najít datové soubory s ekonomickou tematikou člověk logicky zavítá na Ministerstvo financí. To patří mezi pokrokovější ministerstva z pohledu otevřenosti dat, svá data významně uvolnilo v lednu 2015. Na webových stránkách[2] je k nalezení přímo sekce s daty. Prohlížet datové sady se dá buď v rámci sedmi skupin podle tematiky, celkového seznamu všech sad nebo možností je též vyhledávat pomocí filtrů. Datové sady jsou dostupné nejčastěji ve formátech xls(x) a csv, u každé sady je její krátký slovní popis, podmínky užití (většinou *volné dílo*) a katalogizační záznam s metadatovými údaji. V době psaní tohoto textu je již 22 sad k dispozici a 13 by mělo být zpřístupněno v brzké době (resp. některé měly být zpřístupněny v průběhu února 2015, u jiných je status „V současné chvíli probíhá analýza možností exportu a zveřejňování“) (Ministerstvo financí ČR, 2015).

Příklady zpřístupněných vrstev:

- Seznam smluv MF ČR (všechny smlouvy/dodatky MFČR včetně peněžních částek)
- Seznam faktur MF ČR (Přehled uhrazených faktur MF ČR a seznam partnerů MF ČR)
- Daňová statistika (rozsáhlé přehledy o zaplacených daní získané agregací všech daňových přiznání)
- Státní rozpočet (Číselné přílohy zákona č. 345/2014 Sb. o státním rozpočtu České republiky na rok 2015)
- Fiskální výhled ČR (Časové řady hodnot příjmů, výdajů, salda, zdrojů, užití a dluhu veřejných rozpočtů a vládních institucí ČR)
- Přehled povolených technických herních zařízení (Seznam povolených technických herních zařízení jako např. výherních automatů, včetně jejich lokalizace

V plánu ministerstva je uvolnit i souhrnné datové sady jako je Obchodní rejstřík, Rejstřík ekonomických subjektů, Rejstřík plátců spotřební daně, Zaplacené daně právnických osob, Pohledávky FÚ po splatnosti a Nespolehliví plátci DPH (Ministerstvo financí ČR, 2015). Není zatím jasné v jaké formě a šíři budou zveřejněny, nicméně tyto vrstvy v sobě skýtají velký informační potenciál a pokud opravdu dojde k jejich publikaci, mohou se stát cenným datovým základem pro různé prostorové analýzy.

Souběžně s výše popsanými otevřenými datovými vrstvami (myšleno konkrétními soubory ke stáhnutí) nabízí MF ČR možnost přístupu do Administrativního registru ekonomických subjektů (ARES). Pomocí XML rozhraní a standardizovaných dotazů lze získávat veřejné informace ze všech registrů, rejstříků, seznamů a evidencí ekonomických subjektů, kterými disponuje státní správa. Příkladem můžou být obchodní rejstřík, živnostenský rejstřík, Registr ekonomických subjektů, Evidence zemědělského

[2] http://data.mfcr.cz/cs

podnikatele, Registr zdravotnických zařízení, Seznam politických stran a hnutí nebo insolvenční rejstřík apod. (Ministerstvo financí ČR, 2012)… ARES tedy vystupuje jen jako zprostředkovatel těchto údajů, správci zodpovědní za aktuálnost a správnost jsou odpovídající ministerstva či útvary státní správy mající daný rejstřík dle zákona na starost.

Na stránkách ARES[3] jsou rozepsány možnosti dotazování, parametry, struktura a maximální počet dotazů za den/noc. Přes rozsáhlost tohoto informačního systému je však obtížně využitelný pro analýzy, nenabízí totiž příliš vhodné možnosti hromadných exportů záznamů a získat kompletní databázi ekonomických subjektů pomocí ARES není možné (zatím) vůbec. Využitelné to teoreticky je pro konkrétní dotazy na ekonomické subjekty, které nenajdou více odpovídajících záznamů než je limit 1000 (tzn. dotaz na všechny subjekty v Olomouci pro vysoký počet neprojde, ale pro malou obec či jednu olomouckou ulici už nejspíš ano). Zároveň nelze vybírat subjekty pro větší oblasti než obec. Jedinou možností je postupné dotazování podle např. seznamu IČ, který si však potenciální zájemce musí opatřit jinde, a zároveň nesmí překročit maximální povolený limit dotazů 1000/den a 5000/noc.

Český statistický úřad

Jedním z největších producentů a zprostředkovatelů volně dostupných datových sad ve veřejnosprávní sféře je bezesporu Český statistický úřad (ČSÚ). Nabízí jak konkrétní datové vrstvy s hodnotami jednotlivých ukazatelů pro různé časové období a územní celky, tak kompletní databáze či periodicky se opakující publikace se souhrny různých ukazatelů i s případnými interpretacemi hodnot.

ČSÚ nabízí dva způsoby, jak vyhledat požadovanou datovou vrstvu. Prvním je použitím Katalogu produktů dostupného na internetových stránkách[4]. Ten zprostředkovává možnost specifikování výběru pomocí jednoduchého slovního vyhledávače či filtru. Za roky 2010–2015 je zde k dispozici přes 4000 vydaných produktů. Dokonce je zde možnost podívat se výhledově až 12 měsíců do budoucnosti, který produkt kdy bude publikován. Toto je možné díky tomu, že se jedná o opakující se produkty s měsíční, kvartální, půlroční nebo roční periodicitou. Formát produktu se odvíjí od jeho povahy, pro tabulkové vrstvy je to nejčastěji xls(x) nebo pdf, pro textové publikace pdf a doc(x). Po výběru produktu se zobrazí podrobnější informace včetně odkazů na archivní dokumenty (předchozích časových období) shodného produktu. Zájemce o starší publikace (2004–2009) má možnost vybrat si požadovanou vrstvu v ročním soupisu publikací nazvaném Ediční plán a zažádat si o ni emailem nebo objednávkovým formulářem (dostupným současně s Edičními plány na stránce Katalogu produktů). Tato služba je však zpoplatněna.

Příklady produktů s ekonomickou tematikou dostupné v Katalogu produktů ČSÚ:

- Hlavní makroekonomické ukazatele (datová vrstva pro celou ČR, časová platnost od 1993 – současnost, ukazatele jako zaměstnanost, HDP, průměrné mzdy a ceny, …)
- Základní finanční ukazatele (sada datových vrstev pro celou ČR, rozděleno do konkrétních tabulek dle ekonomických odvětví, čtvrtletní periodicita, ukazatele jako počet zaměstnanců v oblasti, prům. hrubá mzda, …)

[3] http://wwwinfo.mfcr.cz/ares/ares.html.cz
[4] https://www.czso.cz/csu/czso/katalog-produktu

- Indexy cen a indexy tržeb (datové vrstvy pro celou ČR, pro různé oblasti ekonomiky, měsíční a kvartální periodicita, srovnání se stejným obdobím předchozího roku)
- Statistický bulletin (sada datových vrstev pro kraje vydávaná se čtvrtletní periodicitou, obsahuje mimo jiné některé ekonomické ukazatele či počty ekonomicky aktivních subjektů v odvětvích působení aj., nabízí porovnání s ostatními kraji)
- Statistická ročenka (sada datových vrstev široké škály statistických ukazatelů – ekonomických nevyjímaje, souhrnných či průměrovaných, vydávaná s roční periodicitou pro kraje nebo celou ČR)

Druhým způsobem, jak najít požadované datové vrstvy na stránkách ČSÚ, je Veřejná databáze[5] (VDB). Tento je navíc mnohem příznivější pro vyhledávání. Základní rozdělení je do devíti tematických kategorií, v rámci nich následuje další vnitřní hierarchické dělení. Ve výpisu konkrétních datových vrstev jsou zobrazeny kromě názvu a identifikátoru vrstvy taktéž přímo informace o časové platnosti a území, k němuž se statistický ukazatel vrstvy vztahuje. Při zobrazení konkrétní vrstvy prostředí VDB nabízí možnost upravovat parametry (například vybrat rok a konkrétní oblast, ke které chceme data zobrazit) nebo tabulku exportovat do formátu xls nebo xml.

Příklady datových vrstev s ekonomickou tematikou:

- Součty ekonomických subjektů pro kraje nebo ČR v různých letech a kategoriích působení
- vrstvy o hospodaření státu
- Příjmy a výdaje obcí/krajů v různých letech
- indexy cen
- míry nezaměstnanosti, míra ekonomické aktivity
- zahraniční obchod se zbožím

Nevýhodou VDB z autorova pohledu je absence hromadných výstupů. Případný zájemce o srovnání jednoho statistického ukazatele pro jedno konkrétní území v různých letech musí exportovat tabulku pro každý rok zvlášť. Další nevýhodou oproti vyhledávání v Katalogu produktů je dle autorovy zkušenosti slabší aktuálnost, data zveřejňovaná ve VDB mají několikaměsíční zpoždění od jejich publikování například v čtvrtletních statistických bulletinech.

Speciální datovou sadou, která je rovněž k dispozici skrze VDB, je Sčítání lidu, domů a bytů 2011. To má sice svoji vlastní webovou stránku[6], ale struktura je na standardní VDB velmi podobná. Opět je zde možnost výběru statistického ukazatele z tematických skupin, navíc dovoluje výběr konkrétnější oblasti (stát, kraj, okres, obec), u níž se zobrazí částečně volitelná skupina ukazatelů naráz. Samozřejmostí je možnost exportu do formátu xls či xml. V datových vrstvách jsou opět k nalezení některé ekonomické ukazatele jako průměrná mzda, velikost pracovní síly, nezaměstnanost apod. Předchozí sčítání takto propracovanou prezentaci výsledků nemá, z roku 2001 jsou k dispozici některé publikace vzniklé na tomto podkladě, nebo základní informace pro stát/kraj/okres/obec (jako například počty ekonomicky aktivního obyvatelstva).

[5] https://vdb.czso.cz/vdbvo/uvod.jsp
[6] http://vdb.czso.cz/sldbvo/

ČSÚ je správcem některých registrů a databází. Z ekonomického hlediska je zajímavá Databáze zahraničního obchodu. Při výběru této databáze na stránkách ČSÚ[7] se nabídnou možnosti specifikace datové vrstvy, a to jednak časového parametru (od r. 1999 až do současnosti), dále výběr druhu zboží (např. „bavlna“, „korkové výrobky“, „zbraně a střelivo“, …), směr obchodu (dovoz, vývoz, bilance, obrat) a výběr jedné nebo více zemí nebo kontinentů, pro které chceme zjistit hodnoty. Navíc je zde možnost porovnat více států v rámci kontinentu nebo uskupení zemí. Výslednou tabulku je možno exportovat do xls formátu. Tato databáze je udržována na dobré úrovni z hlediska aktuálnosti, příkladem je fakt, že v polovině dubna 2015 lze prohlížet hodnoty za únor 2015.

Za zmínku stojí možnost dostat se k tabulkám Eurostatu právě přes stránky ČSÚ. Tabulky je možné vyhledávat v kategorizovaném seznamu témat nebo ve skupině průřezových statistik či tabulek dle EU politik. Výsledné vrstvy nabízí srovnání mezi evropskými státy za roky od 1990 do současnosti (různě pro odlišné ukazatele). Porovnat si tak případný zájemce může například HDP, nezaměstnanost, průměrný a minimální plat, indexy spotřebitelských, průmyslových nebo nemovitostních cen, výše produkce v různých odvětvích nebo státní výdaje v různých oblastech aj.

Všechny výše popsané zdroje dat od ČSÚ mají společné to, že případný zájemce si musí vyhledat konkrétní datový soubor, který se dá následně exportovat pro vlastní využití. Odlišně funguje další databáze spravovaná ČSÚ, a to Registr ekonomických subjektů (RES). Na jeho webových stránkách je sice veřejně dostupný vyhledávací formulář, kde se na základě názvu subjektu či jeho IČO dají zobrazit základní údaje o subjektu, ale není zde možnost exportu. Protože jsou údaje o ekonomickém subjektu (ES) ze zákona (§20 odst. 3 písm. a) až j) zákona č.89/1995 Sb., o státní statistické službě) veřejné, je možné je poskytnout každému, kdo si o ně zažádá (Český statistický úřad, 2014). K dispozici je buď celá databáze i s číselníky a prohlížecím programem nebo zájemcem definované výběry. Výběr je možno zúžit jednak co do počtu atributů a taktéž co do počtu ES splňujících libovolné podmínky (plošné, atributové). Poskytování údajů z RES je zpoplatněno, výše poplatku se odvíjí podle počtu vybraných ES a požadovaných atributů. Základní cena celé databáze byla v dubnu 2015 20 000 Kč, v případě výběru atributů pro všechny ES potom 1500 Kč za atribut (Český statistický úřad, 2015). Sleva pro studenty je možná, ale oficiálně není uvedena. RES je aktualizován měsíčně a v atributech jsou u každého ES informace o názvu, datu vzniku (a příp. zániku se způsobem zániku), právní formě, počtu zaměstnanců, převažující činnosti a adrese. Tato databáze se stala zejména pro svoji kompletnost základem této diplomové práce, podrobněji bude popsána později.

Ministerstvo práce a sociálních věcí ČR

MPSV ČR monitoruje ukazatele týkající se práce, a následně je veřejně publikuje na svých stránkách[8] v sekci statistiky.

Příklady datových vrstev využitelných pro další analýzy:

- nezaměstnanost (pro okresy a kraje ČR, od r. 1997, aktualizováno měsíčně vždy 6. pracovní den měsíce následujícího, jedná se o komprimovaný soubor složený

[7] http://apl.czso.cz/pll/stazo/STAZO.STAZO
[8] http://portal.mpsv.cz/sz/stat

z několika tabulek formátu xlsx a v novějších datech i textové průvodní dokumentace s mapou)

- Regionální statistika ceny práce (pro jednotlivé kraje nebo porovnání všech dohromady, od roku 2001, aktualizace čtvrtletně, zvlášť podnikatelská a nepodnikatelská sféra, možnost zobrazení nebo stažení ve formátech pdf a xls/xlsx)
- absolventi škol (nezaměstnaní absolventi dle oborů nebo škol a okresů trvalého bydliště, od r. 2002 s půlročními aktualizacemi, formát xls/xlsx)
- nabídka a poptávka na trhu práce (různé výsledky analýz pro okresy, aktualizované čtvrtletně nebo měsíčně, až od roku 1991, zobrazitelné v prohlížeči nebo stažitelné v xml)
- zaměstnávání cizích státních příslušníků (pro celou ČR, rozepsáno podle státních příslušností, aktualizováno po měsících v letech 2005–2011, zobrazitelné v prohlížeči nebo stažitelné v xml)

Další instituce a projekty

Ministerstvo zemědělství

I na webových stránkách dalších ministerstev se dají nalézt souhrnné datové soubory, které mohou souviset s ekonomikou. Jmenovitě například Ministerstvo zemědělství nabízí Seznam osob podnikajících v ekologickém zemědělství[9], což jsou roční soupisy 2008–2013, s adresami, výměrami či výpisem chovu zvířat, ve formátu xls.

Policie ČR / Ministerstvo vnitra / Mapa kriminality

Na internetové adrese Police ČR[10] (resp. MVČR) jsou dohledatelné policejní statistiky kriminality (součty jednotlivých trestných činů a přestupků pro kraje a dohromady pro celou ČR, od roku 2000 až po současnost, v měsíčních kumulativních součtech pro každý rok, ve formátu xls). Podobně projekt mapakriminality.cz[11] nabízí tyto statistiky dokonce v úrovni až policejního obvodního oddělení (měsíční součty od ledna 2013 po současnost) a to buď přehledně zobrazené v mapě, exportovatelné do formátu csv pro jednotlivé trestné činy a ve volitelné časové a územní platnosti nebo pro strojové zpracování je k dispozici API přístup (pod licencí Creative Commons BY-NC-SA).

Česká obchodní inspekce

Česká obchodní inspekce (ČOI) úplně otevřela svá data v září 2013, navázala spolupráci s opendata.cz a pod otevřenou licencí publikuje svoji Databázi kontrol, sankcí a zákazů[12]. V šesti tabulkách ve formátech xlsx, ods nebo csv jsou informace o všech kontrolách provedených kontrolory ČOI od 1.1. 2012. Kromě samotných kontrol zprostředkovávají seznam pokut, zakázaných výrobků nebo zabavených padělků. Aktualizace probíhá jednou za tři měsíce doplněním nových kontrol do současných tabulek (Česká obchodní inspekce, 2015). Atributy tabulek jsou na webu kvalitně popsány a ke každé tabulce jsou dostupné metadata.

---

9 http://eagri.cz/public/web/mze/zivotni-prostredi/ekologicke-zemedelstvi/seznamy-podnikatelu/celkovy-seznam-podnikatelu/

10 http://www.policie.cz/statistiky-kriminalita.aspx

11 http://www.mapakriminality.cz/

12 http://www.coi.cz/cz/spotrebitel/open-data-databaze-kontrol-sankci-a-zakazu/

Rozpočet obce

Stránky projektu Rozpočet obce[13] nabízí přehledně uspořádané rozpočty všech obcí ČR za roky 2000–2013. Systém webových stránek nabídne rozklikávací nabídky, grafické vizualizace, souhrny kategorií položek rozpočtu i porovnání v čase pro danou obec či mezi vybranými obcemi navzájem. Možnosti exportu jsou dvojí. Buď po výběru statistik pro jednu obec a jejich zobrazení v prohlížeči, výsledný soubor formátu csv bude pro danou obec ve všech letech, nebo v sekci *data* jsou pro zájemce připraveny kompletní seznamy položek rozpočtů nebo roční součty všech obcí ČR dohromady pro daný rok. Tyto vrstvy jsou dostupné pro roky 2000–2012 a to buď ve formátu csv nebo gft (google fusion tables). Samozřejmostí je vysvětlení jednotlivých atributů. Zdrojem těchto dat jsou oficiální data zveřejňovaná MF ČR v systémech ARIS a ÚFIS (Rozklikávací rozpočet obce, 2014).

Vsechnyzakazky.cz

Portál vsechnyzakazky.cz[14] shromažďuje data o veřejných zakázkách. Na úvod se zobrazí vyhledávací formuláře, a to buď z pohledu zakázky, dodavatele anebo zadavatele. Tyto tři typy subjektů (entity) jsou navzájem propojené, tzn. při výběru konkrétní zakázky je možno vidět dodavatele a zadavatele, při přejití na detail zadavatele se zobrazí všechny zakázky a dodavatelé apod. Detail zakázky obsahuje informace jako popis, zadavatel, odvětví, datum zveřejnění, dodavatel, cena a počet nabídek (konkurence ve výběrovém řízení). Detaily zadavatelů nebo dodavatelů zobrazují základní informace o subjektu (dostupné z veřejných rejstříků) a následně statistiky pro všechny zakázky dohromady platné za specifikovaný čas (za rok od 2007 do současnosti). Zdrojem dat je zejména věstník veřejných zakázek (resp. jeho informační systém ISVZ) spravovaný Ministerstvem pro místní rozvoj ČR, aktualizace probíhají průběžně, neměly by být starší než jeden měsíc (Vsechnyzakazky.cz, 2015).

Pro export do formátu xml nebo json slouží API, jehož struktura je na stránkách dobře popsaná. Případně je zde ještě možnost exportovat do formátu xls zjednodušené výpisy entit vzniknuvší při vyhledávání.

Vybraná data

Jak již bylo nastíněno výše, datovým základem této diplomové práce se stal RES od ČSÚ. Důvodem byla zejména časová, atributová i prostorová komplexnost této databáze. Díky smlouvě o vzájemné spolupráci Katedry geoinformatiky Univerzity Palackého v Olomouci a Krajské správy ČSÚ v Olomouci byla data na základě žádosti o poskytnutí dat (viz příloha 12) vydána zdarma pouze pro účely diplomové práce výměnou za poskytnutí výstupů této práce Krajské správě ČSÚ. Dalším vybraným datovým souborem se stal rozpočet obcí, zejména pro svoji použitelnost díky hromadnému exportu pro všechny obce najednou.

[13] http://www.rozpocetobce.cz/
[14] http://vsechnyzakazky.cz/

5 DATA, HARDWARE A SOFTWARE

V následující podkapitolách jsou podrobněji popsány data, které budou využity, a programové a technické vybavení stojící za výpočty.

5.1 Popis a struktura dat

RES

Při zájmu o RES je od ČSÚ dodána kromě samotných dat i programová prohlížečka, umožňující procházení dat a libovolné exporty. S daty se samozřejmě dá pracovat přímo v libovolných aplikacích. Veškeré informace ale jsou v zájmu zachování integrity databáze v co největší míře zakódovány do bezvýznamových číselných kódů, k získání konkrétních hodnot je potom potřeba číselníků. K samotným datovým souborům RES je tedy dodávaných celkem jedenáct číselníků.

Celkový RES se dělí na dvě části, podle aktivity ekonomického subjektu (ES) na "živá" (ES stále aktivní) a "mrtvá" (ES které již ukončily svou činnost). Obě části jsou dále rozčleněny do tří tabulek *a, b, c.* Základní tabulka *a* (s názvem Tabulka ekonomických subjektů) obsahuje celkem 15 atributů, jejich výpis je v tabulce 1. Základním identifikátorem je IČO číslo (atribut ICOF), který je současně primárním klíčem tabulky, a přes něj jsou identifikovány další atributy v ostatních tabulkách. Datové typy atributů (kromě uvedených DATE) jsou textové (resp. character), a to i v případě číselných kódů z číselníků. Po seskupení "živých" a "mrtvých" ES bylo k 31. 1. 2014 celkem 4 069 188 záznamů.

tab. 1 Struktura tabulky ekonomických subjektů (zdroj: ČSÚ, 2013, upraveno)

název atributu	**číselník**	**akronym**	**znaků**
Identifikační číslo	ne	ICOF	8
Datum vzniku (struktura DATE)	ne	DDATVZN	8
Datum zániku (struktura DATE)	ne	DDATZAN	8
Způsob zániku	ano	ZPZANF	2
Datum aktualizace (struktura DATE)	ne	DDATPAKT	8
Právní forma (statistická)	ano	FORMAF	3
Převažující činnost (statistická)	ano	NACEF	5
Kategorie dle počtu pracovníků	ano	KATPOF	3
Identifikační číslo základní územní jednotky sídla organizace	ano	ICZUJF	6
Kód okresu sídla	ano	OKRESLAUF	6
Firma, název (jméno)	ne	FIRMA	254
Právní forma (registr osob)	ano	ROSFORMAF	3
Kód institucionálního sektoru (ESA95)	ano	ISEKTORF	5
Kód institucionálního sektoru (ESA2010)	ano	CISS2010F	5
Kód adresního místa dle ISUI (ČÚZK)	ne	KODADM	9

Tabulka *b* (tabulka právních forem a činností ekonomických subjektů) potom obsahuje výčet všech činností (v kategorizaci CZ-NACE) a právních forem vedených

u subjektu v rejstřících, veřejných seznamech úřadů, zjištěných statistickými šetřeními apod. U každého záznamu je kromě identifikátoru IČO kódové označení zdroje informace, oblast informace (právní forma/činnost ES) a samotná číselníková hodnota označující právní formu nebo činnost. IČO je k ostatním informacím ve vztahu 1:N, tzn. u každého ES představovaného IČO identifikátorem může být libovolně mnoho údajů o právní formě nebo činnosti. Pro představu o velikosti - celá tabulka *b* pouze v části živých ES sestává z 9 013 870 záznamů, z toho drtivá většina jsou informace o činnostech ES (zvláště u novějších záznamů). Běžně se stává, že má IČO číslo u sebe uvedeno více činností, v těchto případech se z něj stanoví statisticky převažující činnost, a ta je poté uvedena přímo v základní tabulce *a* (atribut NACEF).

Tabulka *c* (tabulka adres) obsahuje lokalizační informace k ES. Vztah k identifikátoru IČO je 1:1, každý ES představovaný IČO má tedy vedenou právě jednu adresu. Způsob reprezentace adresy v tabulce je dvojí, starší záznamy obsahuji pouze běžný textový zápis adresy, nijak specificky strukturovaný, novější již mají adresu rozepsanou do šesti atributů - PSČ, obec, část obce, ulice, číslo domovní a číslo orientační.

V podobě, v jaké byly data obdrženy od ČSÚ, byl datový formát tabulek dbf, s kódováním znaků Windows 1250. Číselníky byly ve formátech dbf nebo txt a popis číselníků a tabulek ve formátu doc. Aplikace s funkcí prohlížečky, běžně s daty RES dodávané, nebylo vyžadováno, proto ani nebyla od ČSÚ poskytnuta.

rozpočet obcí

Ze stránek projektu lze stáhnout jeden zip archiv s rozpočty všech obcí ČR v letech 2000–2012. Po rozbalení je vytvořeno 12 tabulek formátu csv (resp. gft - podle volby exportu na stránkách projektu) pro každý rok. V jednotlivých tabulkách jsou vypsány všechny položky všech rozpočtů (řádek = jedna položka rozpočtu konkrétní obce). Každá položka obsahuje informace o obci (název, kód dle ČSÚ, IČO číslem úřadu), okrese, kraji, druhové třídění rozpočtové skladby (hierarchicky dělené do tříd, skupin, podskupin), název položky, odvětvové třídění rozpočtové skladby (kategorie, sekce, subsekce a paragraf) a konečně peněžní hodnotu položky v českých korunách. Většina parametrů je ohodnocena jednak kódovým označením a zároveň textovým popisem. Rozsah tabulky je pochopitelně pro každý rok různý, například rok 2011 obsahuje celkově 1 020 699 záznamů, rok 2001 o 70 000 méně. Rozpočet nejmenších obcí je tvořen kolem padesáti položek, u většiny obcí je to v řádu stovek a rozpočet pražského magistrátu obsahuje bezmála 2700 položek (v roce 2011). Kódování znaků je v UTF-8.

RÚIAN

Registr územní identifikace, adres a nemovitostí (zkráceně RÚIAN) je jedním ze základních registrů státní správy České republiky, obsahuje referenční prostorové údaje, na něž se odkazují zbylé registry. Editory RÚIAN jsou obce, stavební úřady, ČSÚ a zejména ČÚZK (Formánek, 2014).

Prvky (vybrané):
- bonitované díly parcel, parcely, stavební objekty, adresní místa,
- ulice, katastrální území, městské obvody nebo městské části, části obce, volební okrsky, základní sídelní jednotky
- obce, POU, ORP, okresy, kraje, vyšší územně samosprávné celky, regiony soudržnosti, stát

RÚIAN je veřejným seznamem, pro bezplatný přístup slouží aplikace Veřejného dálkového přístupu[15] (VDP). Tato aplikace nabízí i možnost exportu skrze výměnný formát RÚIAN (VFR), založený na XML (resp. GML). Ve výstupním formuláři si uživatel vybere časovou platnost (přírůstky od určitého data nebo kompletní stav k určitému datu - lze zvolit aktuální nebo i historický údaj), rozlišení územních prvků, výběr z údajů a případné územní omezení. Výsledkem je výpis nalezených souborů ve výše zmíněném výměnném formátu. V případě zájmu o celou ČR k jednomu datu (aktualizováno vždy ke konci měsíce) v podrobnějším rozlišení na prvky velikosti obce a menší (kat. území, ulice, adresní místa, ...) je výsledkem vyhledávání jeden soubor pro každou obec. Tyto soubory lze stahovat buďto po jednom nebo lze exportovat textový seznam přímých odkazů na všechny soubory a ty stáhnout pomocí externí stahovací aplikace. Soubory jsou fyzicky ve formátu xml, komprimované ve formátu gz. Stažená data je poté nezbytné pomocí specializovaných nástrojů buďto nahrát do databáze (postup podrobně popsán později) nebo do prostředí GIS programu. Pro tuto variantu lze jmenovat například nástroj VFR import firmy ARCDATA PRAHA, s.r.o., jehož jednodušší verze umí nahrát stažené soubory do geodatabáze software ArcGIS, pokročilejší verze nástroje dokáže přímo v ArcMap on-line vyhledávat soubory RÚIAN, stahovat nebo i udržovat RÚIAN data aktualizované (ARCDATA PRAHA, 2015).

Kódování znaků je v UTF-8, souřadnice v souřadnicovém systému S-JTSK (EPSG 5514). Podrobnější popis atributů nebo formátu VFR viz specifikace na stránkách RÚIAN na webu ČÚZK[16]. Data RÚIAN se dají mimo jiné použít pro geokódování - porovnáváním hledaných adres s adresami uvedenými v RÚIAN a v případě správného nálezu přiřazením uvedených souřadnic. K tomuto účelu byla data využita i v této diplomové práci.

ArcČR 500

Databáze ArcČR 500 je produkt vzniklý spoluprácí firmy ARCDATA PRAHA a ČÚZK. Data jsou distribuována zdarma pro libovolné použití za podmínky uvedení autorských práv zmíněných subjektů. Skládá se ze dvou samostatných geodatabází (formátu firmy esri) - první část obsahuje topografické vrstvy, druhá administrativní členění ČR doplněné o statistické údaje (zdroj ČSÚ). Databáze je vektorová (kromě vrstev modelu reliéfu), používá souřadnicový systém S-JTSK a podrobnost odpovídá měřítku 1 : 500 000 (ARCDATA PRAHA, 2014). Podrobnější popis viz webové stránky databáze[17]. Pro tuto diplomovou práci byla využita jako prostorová podkladová data s výhodou již přiložených atributů obsahujících údaje o počtu obyvatel ze SLDB 1991, 2001 a 2011.

5.2 Použitý hardware a software

Hardware

Hlavní část práce vznikala na osobním počítači se čtyřjádrovým procesorem Intel Core i3 M 350 2,27 GHz a 4 GB paměti RAM. Nicméně primárně kvůli databázovým procesům (trvajícím řádově i několik dní) byl využit přístup na studentský server Katedry

[15] http://vdp.cuzk.cz/

[16] http://www.cuzk.cz/Uvod/Produkty-a-sluzby/RUIAN/2-Poskytovani-udaju-RUIAN-ISUI-VDP/Vymenny-format-RUIAN/Vymenny-format-RUIAN-(VFR).aspx

[17] http://www.arcdata.cz/produkty-a-sluzby/geograficka-data/arccr-500/

informatiky Přírodovědecké fakulty Univerzity Palackého v Olomouci. Ten byl původně sestaven v konfiguraci DELL R310 (čtyřjádrový procesor Intel Xeon X3470 2.93 GHz s funkcí hyperthreading, 2 x 4 GB ECC RAM, systém Gentoo Linux), ale v průběhu práce došlo k jeho nedobrovolnému přesunu (z důvodu destrukce) na virtualizovaný stroj se čtyřjádrovým procesorem Intel Xeon X5460 3,16 GHz s přidělením 6GB RAM.

Software

V této části magisterské práce se nachází pouze výčet použitých programů s krátkým vysvětlením funkcionality. Podrobněji budou popsány později v textu práce logicky tak, jak přijdou na řadu v kontinuálním pracovním procesu.

Na výše popsaném osobním počítači byly souběžně nainstalovány dva operační systémy. Prvním je linuxová distribuce Kubuntu ve verzi 14.04, která byla mnohem využívanější než druhý systém, Windows 7 od firmy Microsoft. Operačnímu systému odpovídají i použité programy. V systému Windows bylo využito pouze programu ArcMap 10.3 z balíku ArcGIS for Desktop od firmy esri[18], v jeho 60denní trial licenci (pro výpočty shlukování a vizualizace mapových výstupů) a pro finální stylizaci textu program Microsoft Word z balíku Microsoft Office 365[19] ve studentské licenci.

Vůbec časově největší podíl z celé práce probíhal v databázovém prostředí PostgreSQL[20] 9.3.6 s prostorovým rozšířením PostGIS[21] 2.1.2. Obnášelo to kompletní přípravu dat, zejména čištění, geokódování a výběry pro analýzy. Jako uživatelské rozhraní bylo využito zejména konzolového nástroje psql, ale taktéž grafického programu pgAdmin[22] nebo webové aplikace phpPgAdmin.

Dalším významným programem je R[23] 3.1.3. (resp. později 3.2) a jeho funkcionalitu rozšiřující balíčky, v němž byly vypracovány statistické analýzy včetně tvorby velké části grafických (neprostorových) výstupů. Ostatní grafické výstupy byly vytvořeny v rastrovém GIMP[24] nebo vektorovém Inkscape[25]. Velmi užitečným programem se ukázal QGIS[26] zejména pro export/import geodat do databáze PostgreSQL či rychlou prezentaci geodat v ní uložených. Pro tvorbu PHP skriptů vývojářský program NetBeans[27] IDE 7.4 a v neposlední řadě pro skládání textu či práci s menšími tabulkami posloužil kancelářský balík LibreOffice[28].

Všechny zde jmenované programy z linuxového prostředí mají společnou volnou licenci; jedná se o open source software. Obrovskou výhodou je jejich vzájemná spolupráce. Např. R a QGIS mají výborně fungující přímé napojení na PostgreSQL (tzn. práce s daty uloženými v PostgreSQL), v R se lehce tvoří výstupy v otevřeném vektorovém formátu svg editovatelném v Inkscape, nebo z nástrojů LibreOffice lze přímo vytvářet tabulky pro DB PostgreSQL. Pravdou je, že všechny tyto programy nabízí i své verze pro

[18] http://www.esri.com/software/arcgis/arcgis-for-desktop
[19] https://products.office.com/cs-cz/home
[20] http://www.postgresql.org/
[21] http://postgis.net/
[22] http://www.pgadmin.org/
[23] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/
[24] http://www.gimp.org/
[25] https://inkscape.org/cs/
[26] http://www.qgis.org/en/site/
[27] https://netbeans.org/
[28] https://cs.libreoffice.org/

Windows. Za volbou linuxového prostředí stojí hlavně možnost efektivní práce přes konzoli, podpora silné (zejména zahraniční) linuxové komunity uživatelů, existence některých užitečných drobných utilit pro příkazový řádek a další subjektivní autorovy pozitivní zkušenosti s Linuxem. Nicméně teoreticky by práce mohla proběhnout celá v kterémkoliv z těchto operačních systémů a v popisu práce je většinou snaha nabízet alternativy pro druhý systém.

6 PŘÍPRAVA DAT

Běžné tabulkové procesory jako MS Excel nebo LibreOffice Calc, které ve většině běžných geoinformatických úloh k přípravě dat dostačují, nejsou uzpůsobeny pro práci s velkým objemem dat. Dokáží zobrazit maximálně okolo jednoho milionu záznamů, a stejně v tomto množství je práce v nich velmi zdlouhavá a neoptimální. Mnohem lépe si s velkou datovou sadou poradí databázové systémy. Mezi světově nejznámější zdarma dostupné systémy řízení báze dat (angl. database management system DBMS) patří MySQL nebo PostgreSQL. Běžný uživatel funkcionálně mezi nimi rozdílu nenajde, oba používají jen mírně odlišný dialekt jazyka SQL. Pro tuto práci byl vybrán PostgreSQL (dále PG), ale bez problémů by vše mělo být proveditelné i v MySQL. Zde popisované kroky přímo související s programy (ať už R nebo PG) jsou přepsány do zdrojového kódu, který je k dispozici na CD a to zejména v souboru priprava_gc.sql.

6.1 Příprava databáze

PostgreSQL

Na internetu je spousta návodů, jak zprovoznit vlastní instanci DBMS PostgreSQL na desktopu i serveru a pro různé operační systémy. Neliší se to nijak zvlášť od instalací běžných aplikací. Při instalaci bude uživatel vyzván, aby zvolil heslo k účtu "postgres". To je důležité, neboť se jedná o hlavního správce PG.

Přístup k PG je skrze uživatelské rozhraní, kterých je na výběr více. Autor má nejlepší zkušenosti s desktopovou aplikací pgAdmin III, jeho největší výhodou je propracovaná kontextová nabídka, tzv. "klikací", je vhodná především pro uživatele - začátečníky. Spoustu operací (např. tvorba tabulek, databází, uživatelů nebo import dat do tabulkyz txt nebo csv souboru) lze zadat pomocí výběru možností, bez faktické znalosti dané oblasti SQL jazyka, pgAdmin to sám do SQL přeloží. Podobně tomu je u webová aplikace phpPgAdmin, pro její běh je však nutný server. Pokročilejší uživatelé však nedají dopustit na konzolovou aplikaci psql, výhodou je zejména minimální náročnost na operační paměť pro samotnou aplikaci. Všechny operace se zde provádí čistě zadáním SQL příkazu.

Prostorová data v PG

Prostorovým rozšířením PG je PostGIS. Definuje datové typy pro uložení geodat v relační DB a přidává funkce a procedury pro práci s prostorovými daty. Dodržuje specifikaci OGC Simple Features a rozšiřuje ji i pro třídimenzionální data. Při instalaci tohoto důležitého doplňku se většinou vytvoří prostorová databáze, případně jakákoliv dříve vytvořená neprostorová se dá jednoduše na prostorovou upravit (skrze příkaz `CREATE EXTENSION postgis`). Viditelný rozdíl oproti běžné DB je ten, že obsahuje navíc schéma topology, dávku funkcí pro prostorové data a již jednu vytvořenou tabulku spatial_ref_sys obsahující definice souřadnicových systémů. Je využito knihovny PROJ. Uživatel z českého prostředí by měl do této tabulky přidat definici v Česku běžně používaného systému S-JTSK. Důležité části této definice jsou EPSG (resp. esri) kód, parametry s. s. ve formátu WKT a transformační rovnice do systému WGS 84 ve formátu proj. Přesné parametry jsou k dispozici například na WIKI portále FreeGIS Fakulty Stavební ČVUT v Praze[29], konkrétní SQL příkaz je uveden ve zdrojovém kódu.

[29] http://freegis.fsv.cvut.cz/gwiki/S-JTSK

Spojení PG a R, základy práce v R

Prostředí R se skrze balíky sqldf[30] a RPostgreSQL[31] umí přímo připojit k PG databázi a pomocí SQL dotazu uložit požadovaná data do svého pracovního prostředí. Pro vizualizaci bylo využito balíku ggplot2[32]. Tento nabízí poněkud odlišný způsob vizualizací než běžné grafické výstupy z R, výhodou je přímá použitelnost (není nutná další grafická úprava, resp. případně tyto úpravy jsou minimální a zároveň ulehčené možností exportu ve vektorovém formátu svg) ale za cenu vyšší náročnosti na zkušenosti uživatele. Balík ggplot2 ve své bakalářské práci blíže popisoval Cícha (2013), případný zájemce zde nalezne podrobněji vysvětlené základy fungování tohoto užitečného nástroje. Cenným pomocníkem při vybírání barev jsou potom nástroje balíčku RColorBrewer[33].

6.2 Import dat do PG

import RES do PG

Před importem je nejprve nezbytné připravit si v PG tabulku, a to buď přímo SQL příkazem CREATE TABLE s vyjmenováním všech atributů, včetně jejich datových typů a délkových omezení. Důležité je, aby to přesně odpovídalo importovaným datům, jak v pořadí atributů, tak v datových typech.

PG má pro import dat do tabulky (a zároveň export) příkaz `COPY`. Ten umí pracovat s jednoduchými textovými formáty jako txt nebo csv. Data RES, vlastněné ve formátu dbf je tedy nutné nejprve konvertovat. Pro rozsáhlou velikost dat je potřeba využít sofistikovaných programů (nelze jen např. „přeuložit“ v tabulkovém procesoru). V prostředí Windows existuje volný program DBF Viewer Plus[34], který obsahuje funkci exportu. V Linuxovém prostředí je k dispozici například jednoduchý konzolový nástroj dbf_dump využívající balík libdbd-xbase-perl[35].

Jak je popsáno výše, RES se skládá ze dvou celků, aktivních IČO a těch s již ukončenou činností. Z pohledu dat ale mají stejnou strukturu, rozdíl je jen ten, že “mrtvé” ES mají vyplněné atributy datum zániku a způsob zániku, které mají “živé” ES prázdné. Je tedy užitečné mít tyto soubory v tabulce dohromady. Řešením může být buďto oba vytvořené csv soubory spojit do jednoho a ten nahrát do PG, nebo importovat do tabulky v PG obě části RES za sebou. V obou případech je potřeba dát pozor na hlavičky původních tabulek. Při spojování nesmí nastat situace, kdy je hlavička původní druhé tabulky zapsaná hned za první, často totiž nadpisy atributů neodpovídají datovým typům atributu a import by neproběhl. Podobně při importování je třeba ze stejného důvodu dbát na to, aby hlavička nebyla do tabulky PG nahrána spolu s daty. Příkaz `COPY` na toto pamatuje v parametru HEADER (boolean). Pochopitelně pro správně importovaná data je dále třeba korektně nastavit kódování znaků, znak pro oddělovač, případně volby uvození

---

[30] G. Grothendieck (2014). sqldf: Perform SQL Selects on R Data Frames. R package version 0.4-10.
http://CRAN.R-project.org/package=sqldf

[31] Joe Conway, Dirk Eddelbuettel, Tomoaki Nishiyama, Sameer Kumar Prayaga and Neil Tiffin (2013). RPostgreSQL: R interface to the PostgreSQL database system. R package version 0.4. http://CRAN.R-project.org/package=RPostgreSQL

[32] H. Wickham. ggplot2: elegant graphics for data analysis. Springer New York, 2009.
https://cran.r-project.org/web/packages/ggplot2/index.html

[33] Erich Neuwirth (2014). RColorBrewer: ColorBrewer Palettes. R package version 1.1-2.
http://CRAN.R-project.org/package=RColorBrewer

[34] http://www.alexnolan.net/software/dbf.htm

[35] http://manpages.ubuntu.com/manpages/natty/man1/dbf_dump.1p.html

textu a speciálních znaků a reprezentaci hodnoty NULL (buďto přes parametry SQL příkazu COPY nebo v nabídkách importovacího dialogového okna pgAdminu). Tímto způsobem byly nahrány[36] tabulky *a*, *c* do PG, tabulka *b* nebude pro zamýšlené analýzy potřebná.

import RÚIAN do PG

Postup zpracování dat RÚIAN do GIS již byl nastíněn dříve. Zde je uvažována již pouze varianta přímého nahrání do PG. Pro tuto operaci důležitým faktem je, že knihovna GDAL od verze 1.11 podporuje formát VFR (Landa, 2014). Díky tomu je možné stažené soubory RÚIAN snadno zpracovat pomocí několika nástrojů. Například všeobecně známý univerzální ogr2ogr[37] dokáže VFR konvertovat do libovnolného jiného formátu geodat implementovaného do GDAL, nebo i přímo importovat do PostGIS. Dalšími možnostmi jsou specifičtější konverzní skripty, které jsou zaměřené na VFR a nejčastější požadované formáty, mimo jiné i PostGIS (jmenovitě např. skript vfr2pg) vytvořené v pythonu na Katedře geomatiky ČVUT v Praze[38], nebo konzolový java program ruian2pgsql[39], který byl využit v rámci této diplomové práce a proto zde bude popsán podrobněji. Nejprve je potřeba si jej stáhnout[40] ze serveru GitHub.com. Dále pro svůj běh potřebuje v lokálním adresáři postahovaná data z VDP RÚIAN, která jsou zamýšlená pro import do PG (nelze tedy data nahrávat přímo online přes VDP), funkční java (JRE nebo JDK 7+) a jdbc driver k PG (jar soubor stažitelný z oficiálních stránek PostgreSQL). V samotném příkazu vyžaduje zadání názvu staženého souboru ruian2pgsql a jdbc driveru – obojí s kompletní adresářovou adresou, dále adresu ve formátu jdbc k databázi PG (ve tvaru `jdbc:postgresql://localhost/NazevDB?user=username&password=heslo`) a adresu adresáře se soubory VFR. V rámci parametrů dále umožňuje nastavit konverzi z GML do E-WKT, linearizovat křivkovité linie, automaticky vytvořit potřebné tabulky v PG (příp. smazat původní a vytvořit nové) a další… (Šulc, 2014). Použitý kód je přiložen ve zdrojovém kódu. Výhodou je, že se žádný prvek do PG nezkopíruje dvakrát, program si umí tedy poradit s aktualizací kompletních souborů (myšleno ne pouze změnových aktualizací RÚIAN ale i plné verze) nebo v případě potřeby zvládne do PG nahrát vše vyjma prostorových atributů dat (Šulc, 2014).

Při importu se vyskytl problém u VFR souboru obce Praha (s názvem datumvydani_OB_554782_UKSH.xml.gz). Jeden z obrovského množství řetězců souřadnic uvnitř nebyl správně ukončen, a celý soubor se kvůli tomu nedařilo nahrát. Tato chyba byla ČÚZK hlášena dle diskuse na GitHub stránkách ruian2pgsql (issue #28) už na jaře 2014, rok poté se v daném souboru (měsíčně aktualizovaném) vyskytuje stále. Řešením je manuální zásah do daného souboru VFR dle návodu na výše zmíněné diskusi. Zajímavostí je že např. nástroj VFR import v ArcMap se zpracováním tohoto souboru nemá problém.

---

[36] Při importování RES byl zjištěn výskyt celkem osmi ES, které měly název firmy v tabulce a delší než deklarovaných 254 znaků, import vždy kvůli nim selhal. Příkladem je třeba politické uskupení Pravý blok, který má ve svém názvu celý politický program. Vzhledem k tomu, že název firem není u analýz důležitý, byly všechny ručně zkráceny.

[37] http://www.gdal.org/ogr2ogr.html

[38] https://github.com/ctu-osgeorel/gdal-vfr

[39] https://github.com/fordfrog/ruian2pgsql

[40] Na stránkách GitHub je k dispozici zkompilovaný ruian2pgsql ve verzi 1.3, nicméně v binární podobě je k dispozici již ve verzi 1.6, a mezi danými verzemi došlo k poměrně dost významným změnám, autor tedy doporučuje stáhnutí nejnovější verze a jeho vlastní následné zkompilování např. pomocí Apache Maven - http://maven.apache.org/

Po úspěšném importu je v PG vytvořeno celkem 23 tabulek, ty mají nastavené primární klíče a prostorové indexy. Vyplněných tabulek s daty je však méně, výše popsaným způsobem zisku dat RÚIAN dojde k uložení informací v rozlišení na obec a přesněji (tedy městské části, katastrální území, ulice, parcely atd.), tabulky vyšších územních celků jako okresy a kraje jsou prázdné.

Bližší seznámení se s daty RES

Předtím, než začnou probíhat jakékoliv úpravy, čištění dat, příprava do požadované podoby apod., je třeba důkladně se seznámit s daty. Zde bude uveden přehled základních informacích o důležitých atributech:

- IČO číslo ES - ICOF

  Základní unikátní identifikátor ES, osmimístný bezvýznamový číselný kód. Může obsahovat v předních částech nuly, po převodu na datový typ integer tyto nuly ztrácí, výhodou je poté rychlejší vyhledávání. Celkový počet ES je 4 069 188, z toho 66 % aktivních a 34 % již neaktivních.

- Datum vzniku - DDATVZN

  Datum zápisu ES do rejstříku v číselném zápisu 'DD.MM.YYYY'. Po importu je vhodné jej z textové podoby převést na datový typ date funkcí `to_date(zdroj, formát)`. Nejstarší zápis je z 25. února 1842, kdy oficiálně vzniklo Právovárečné měšťanstvo v Plzni, další zápisy jsou již z dvacátého a jednadvacátého století. Celkově 99,5 % ES má datum zápisu starší než 1. 1. 1990, nejmladší zápisy jsou z 29. ledna 2014. Nutno podotknout, že vlastněná verze RES je aktuální k 31. 1. 2014.

- Datum zániku - DDATZAN

  Datum oficiálního ukončení činnosti ES. Aktivní ES mají tento atribut prázdný, neaktivní vyplněný v zápisu stejného tvaru jako datum vzniku, pro jeho použití je opět vhodné konvertovat do struktury date. První zánik zapsaný v RES se udál 1. 7. 1973. Co se týče nejmladšího uvedeného zániku, v RES se nachází přes deset ES u kterých je zapsán zánik “v budoucnosti” např. v roce 2035 - zde se jedná téměř jistě o chybu v zápisu. Nicméně oficiální zánik může být znán mírně dopředu, nejnovější zániky tedy odpovídají zhruba datu platnosti RES s tolerancí v řádu jednotek měsíců.

- Způsob zániku - ZPZANF

  Jedná se o číselníkový atribut platný u zaniklých ES. Udává, jakým způsobem k zániku došlo (např. oznámení o ukončení, likvidace, úmrtí fyzické osoby, odstěhování z okresu apod.). Je uveden celkově u necelých 617 000 ES, což je přibližně 44,6 % z počtu zaniklých ES.

- Datum aktualizace - DDATPAKT

  Datum poslední aktualizace ES v registru.

- Právní forma (statistická) - FORMAF

Číselníkový údaj označující formu ES z pohledu statistického (fyzická podnikající osoba, s.r.o., a.s., v.o.s., státní podnik, fond, pojišťovna (druh), škola (druh), státní administrativní jednotka, ...), definován u 100 % subjektů.

- Převažující činnost (statistická) - NACEF

Jedná se o číselníkový údaj označující oblast působení ES dle kategorizace CZ-NACE. Tato kategorizace je pětiúrovňová, v nejobecnější řadě je 22 základních kategorií. Atribut je definován u 79 % ES.

- Kategorie dle počtu pracovníků - KATPOF

Číselníkový údaj, obsahuje informaci o velikosti ES dle počtu zaměstnanců. Zlomové body intervalů kategorií jsou následující: 0, 1, 6, 10, 20, 25, 50, 100, 200, 250, 500, 1000, 1500, 2000, 2500, 3000, 4000, 5000, 10000. Atribut definován u 46,8 % ES.

- Identifikační číslo základní územní jednotky sídla organizace - ICZUJF

Šestimístný kód označující základní územní jednotku sídla ES. Odpovídá menším obcím nebo městským částem větších obcí. Není definován pouze u 359 subjektů.

- Kód okresu sídla - OKRESLAUF

Identifikátor okresu sídla ES ve tvaru CZ0123, definován u 85% ES.

- Firma, název - FIRMA

Textový řetězec s názvem firmy nebo jménem živnostníka, zadán u všech subjektů.

- Právní forma (registr osob) - ROSFORMAF

Číselníkový údaj označující formu ES z pohledu státního registru osob. Oproti statistické verzi je nepatrně chudší (89 kategorií oproti 120) a je definována pouze u 67 % ES.

- Kód adresního místa dle ISUI (ČÚZK) - KODADM

Devítimístný identifikátor adresního místa z RÚIAN, definován u 65 % (cca 2 635 tis.) ES.

- Text adresy - TEXTADR

Neformalizovaný textový zápis adresy, často nekompletní až útržkovitý, zkratkovitý, velmi těžce strojově zpracovatelný. Je zadán v případech kdy chybí údaj KODADM (= 35 %, zhruba 1,43 mil. záznamů), často ale v nepoužitelné podobě (např. údaj “nedefinováno” - 31 tis. záznamů).

- atributy adresy dle ISÚI - PSC, OBEC, COBCE, ULICE, TYPCDOM, CDOM, COR

Jedná se o rozepsané atributy adresy - PSČ, název obce, název části obce, název ulice (v případě její existence), typ domovního čísla (popisné/evidenční/náhradní), samotné domovní a popřípadě orientační číslo.

6.3 Geokódování

Původním záměrem bylo provádět některé prostorové analýzy v rozlišení na katastrální území. Pro tento cíl by bylo potřebné mít data lokalizovaná v přesnosti alespoň na ulici, aby se dala následně agregovat do kat. území. Jak je uvedeno výše, 65 % dat mělo u sebe přímo identifikátor adresního místa z RÚIAN, zde tedy agregaci nic nebrání. Zbylých 35 % však bylo nutné geokódovat (= přiřadit k entitám prostorové souřadnice pomocí nepřímé prostorové identifikace - adresy). Způsobů geokódování je mnoho. Poskytovatele webových geokódovacích služeb v českém prostředí popisuje Cícha (2013). Princip těchto služeb je následující (ve zjednodušené podobě): klient odešle na server požadavek obsahující adresní prvky v určité syntaxi, serverová aplikace porovná adresu se záznamy v databázi adres a vrátí klientovi výsledek - v úspěšném případě souřadnice, případně chybovou hlášku. Úspěšnost tohoto procesu je závislá na několika faktorech: kvalitě databáze poskytovatele (přesnost, rozsáhlost), porovnávacím procesu (aproximační algoritmy v případech drobných nepřesností), serverovém nebo licenčním omezení (povolený počet dotazů/sekundu nebo den) a pochopitelně na kvalitě adresy ve vlastních datech.

Při bližším zkoumání vlastností webových geokódovacích aplikací však bylo usouzeno, že žádná není příliš vhodná pro takové množství dat, jaké je v rámci této diplomové práce třeba provést. Např. Google Maps Geocoding API v bezplatné verzi povoluje pouze 2500 dotazů za den, Mapy.cz API v legálním procesu povoluje přibližně 1 dotaz/vteřinu (Cícha, 2013). To znamená, oněch 35 % ES, které je potřeba geokódovat by se přes Google Maps Geocoding API zpracovávalo 573 dní, přes Mapy.cz API potom necelých 17 dní.

Další možností geokódování v českém prostředí je porovnávat adresy objektů s adresami databáze RÚIAN. Rychlost výsledku bude daná pouze výkonností hardware, kde bude DB (, příp. kvalitou optimalizace porovnávacích algoritmů). Dalšími výhodami bude úplná kontrola nad probíhajícími porovnávacími procesy a jakákoliv absence vnějších limitů. Nevýhodou této metody je nejistota výsledku, zda bude proces reálně opravdu výhodnější než API Mapy.cz. Nicméně i z důvodů snahy o hledání inovativních cest a co nejširší využitelnost této práce bylo rozhodnuto vytvořit návrh geokódovacího algoritmu a otestovat jej na datech RES. Navíc díky existenci identifikátoru adresního místa u některých záznamů se nabízí ideální příležitost hodnocení kvality procesu.

Samotný průběh geokódování byl rozdělen na dvě části - extrakce adresních informací z textového adresního pole záznamů do odpovídajících jednotlivých adresních atributů (1) a proces porovnávání adresy s databází s přiřazováním souřadnic (2). Hlavní funkcionalitu obstarávají databázové operace porovnávání řetězců, konkrétně

```
replace(zdrojový řetězec, 'původníznaky', 'novéznaky')
```

nahrazující konkrétní řetězec jiným konkrétním řetězcem,

```
regexp_replace(zdrojový řetězec, výraz POSIX hledaného řetězce, výraz POSIX nahrazení)
```

umožňující při nahrazování používat zobecňující výrazy formátu POSIX a dále `regexp_matches(zdrojový řetězec, výraz POSIX hledaného řetězce)`, která vrací konkrétní řetězec (nebo pole řetězců) nalezený pomocí výrazu POSIX. Standard regulérních výrazů POSIX (angl. regular expression - zkr. regexpr) definuje sadu tzv. metaznaků, díky nimž lze specifikovat masku pro výběr znaků (PostgreSQL, 2015). Příkladem může být výraz '\d' značící libovolnou jednu číslici, v podobě '\d{3}_\c{2}' už obnáší 'tři číslice za sebou, podtržítko, dva libovolné alfanumerické znaky' apod.

Extrakce adresních informací

Pomocí výše zmíněných textových operací bylo posléze prováděno vytahování informací o obci a ulici z textového řetězce. Postup sestával z několika kroků, v některých případech s logickou návazností:

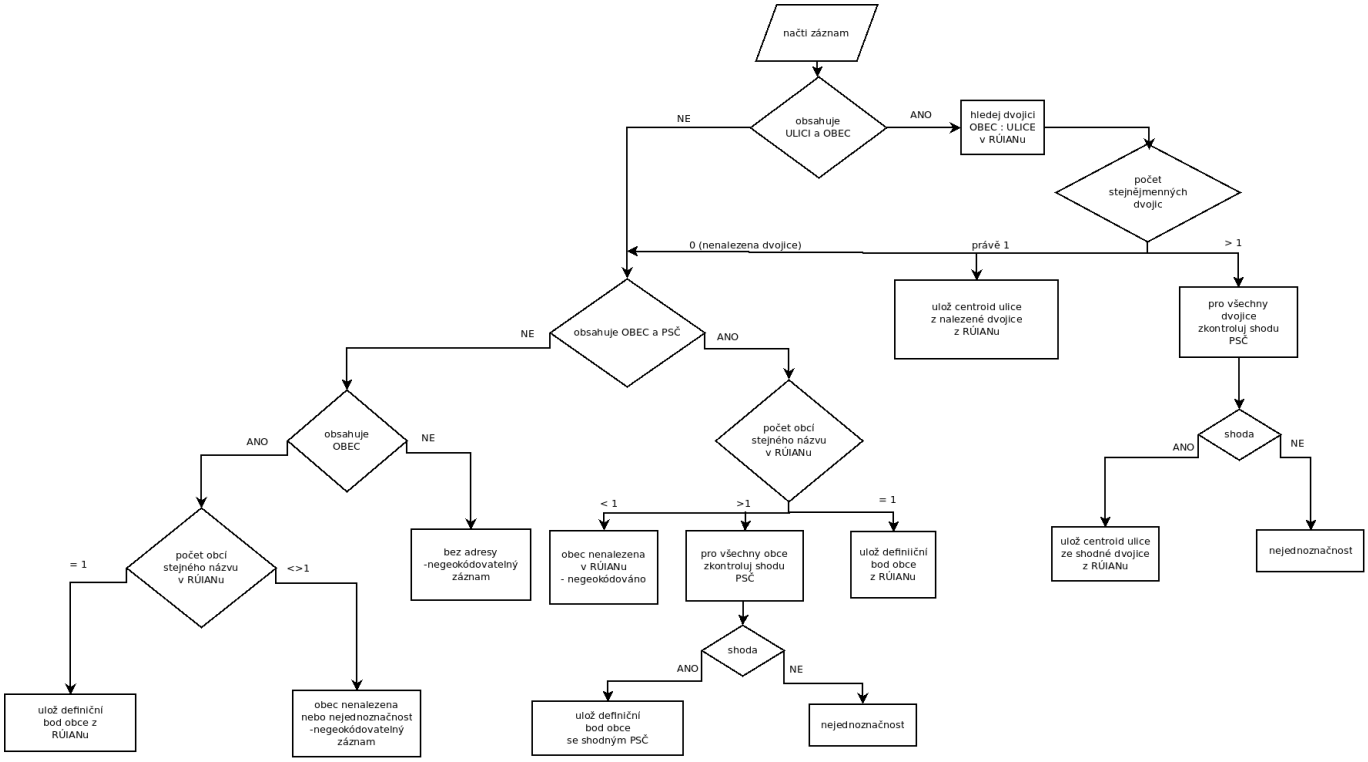
- pět číslic v řadě za sebou (=PSČ) - vypsáno do atributu PSČ
- tři číslice, mezera, dvě číslice (=PSČ) - vypsáno do atributu PSČ
- Vyzkoušena rovnost celého řetězce v TEXTADR na shodu se seznamem obcí v RÚIAN (zjištěno celkem 549 ES s adresou určenou pouze názvem obce)
- smazání řetězce "okres název_okresu" (aby nedošlo k záměně názvu okresního města s názvem obce sídla ES) - včetně různých kombinací zkratek (okr.) nebo verzálek
- převedení celého řetězce na minusky
- úprava názvu ulic na jednotný styl ("tř." na "třída", "nám." za "náměstí") - a to i v tabulce RÚIAN z důvodu větší šance na shodu (víc než na správném názvu ulice aktuálně záleží na její poloze)
- byla vytvořena dočasná tabulka obsahující všechny reálné kombinace obcí a ulic (výpis všech ulic s názvy obcí) z RÚIAN
- porovnání řetězce textové adresy se všemi dvojicemi z dočasné tabulky, při nalezení shody zapsány obě části do odpovídajících atributů (obce/ulice) u ES
- (u ES kde nedošlo k nalezení dvojice obec+ulice) porovnání všech částí řetězce textové adresy se všemi názvy obcí z RÚIAN, seřazeno podle délky názvu od nejdelšího (- aby tak například Bystřice pod Hostýnem byla v textu vyhledána a k ES zapsána dříve než samotná Bystřice, u níž by došlo taktéž k nalezení shody a tudíž nesprávnému přiřazení k ES) a zapsání názvu obce do odpovídajícího atributu.

Výstupem této části byla tabulka adres obohacená o PSČ a názvy obce a ulice, a to vždy pouze u těch záznamů, u nichž se tyto informace podařilo vydolovat. Například název obce se nepodařilo zjistit pouze u 2 760 ES, PSČ potom u 11 031 ES (nepočítaje záznamy s adresou ve tvaru "nedefinováno"). Z důvodu existence duplicit v názvu obcí ČR (existuje řada obcí stejného názvu v Česku, např. Lhoty) a dokonce i duplicit v kombinaci obec + ulice (např. existují tři obce Hranice a dvě z nich mají ulici Nádražní), nelze v těchto porovnáváních zároveň přiřazovat souřadnice, k tomu bude zapotřebí jednoznačná identifikace obcí např. pomocí PSČ.

Porovnávací algoritmus

Princip algoritmu je stavěn tak, aby se snažil nalézt co nejlepší prostorovou informaci. Prochází postupně všechny geokódované záznamy. Při existenci obce i ulice u ES hledá stejnou dvojici v RÚIAN. Pokud nalezne jednu shodu, přiřadí rovnou centroid dané ulice s poznámkou úspěšného geokódování na unikátní dvojici ulice + město. Pokud je shod

více (viz příklad zmiňovaných obcí Hranice s ulicemi Nádražní), zkusí otestovat shodu na PSČ. Pokud není k dispozici údaj PSČ nebo nebyla nalezena shoda, nelze jednoznačně daný ES identifikovat, přiřadí se tedy k němu poznámka nejednoznačnosti výsledku. V případě kdy rozhodovací algoritmus nenalezne žádnou shodnou dvojici v RÚIAN, posune se dále a snaží se geokódovat pouze na definiční bod obce. Porovná tedy název obce s obcemi v RÚIAN, v případě jednoznačné shody přiřadí definiční bod. Pokud je shod více a u ES je udáno PSČ, snaží se opět identifikovat správnou obec pomocí PSČ. Když se jednoznačná identifikace nepodaří, nelze zaručit správnost výsledku a tak se k ES připíše poznámka nejednoznačnosti výsledku. V případě, že se nenalezne shoda obce s RÚIAN, nebo chybí informace o obci u ES, opět rozhodovací proces přiřadí poznámku neúspěchu. Algoritmus je znázorněn ve vývojovém diagramu na obr. 3 a přepsaný v PHP skriptu v příloze. Všechny záznamy, u nichž se nepodařilo ověřit jednoznačnost výsledku, byly vyhodnoceny jako neúspěšně geokódované a pokud to dovolila existence hodnoty atributu ICZUJF (identifikátor ZÚJ), byly později k daným subjektům přiřazeny souřadnice centroidu těchto územních jednotek. Pokud ani tento krok k určení sídla nepomohl, dané subjekty nebyly použity pro prostorové analýzy.

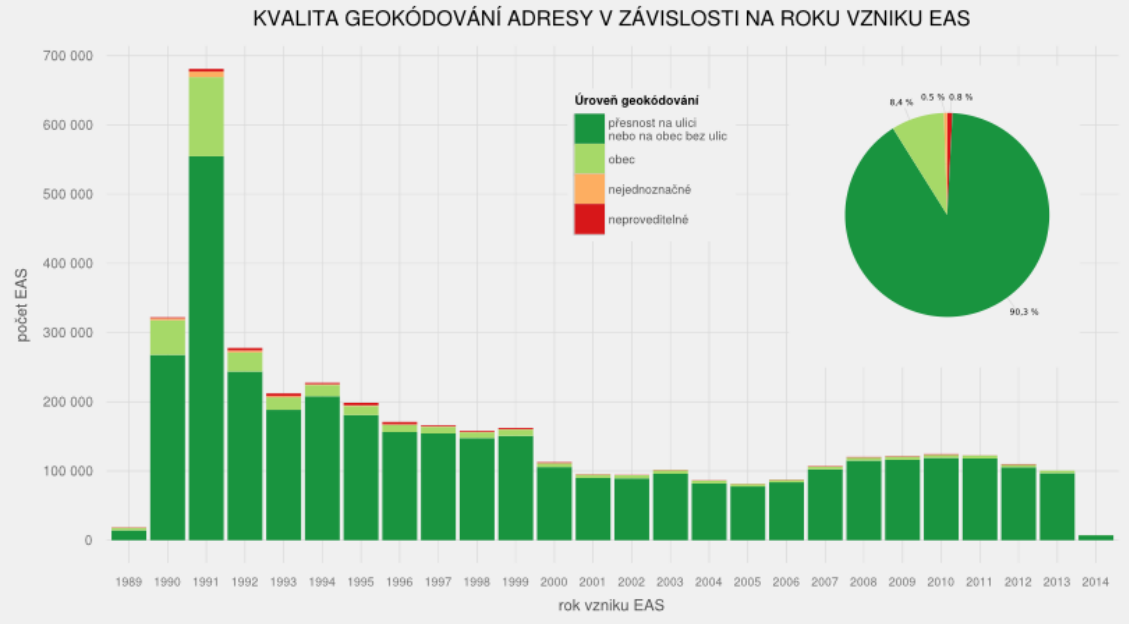


Obr. 3 Princip vlastního geokódování přes RÚIAN

Vyhodnocení geokódování

Navržený geokódovací algoritmus byl spuštěn na datech RES. Pro testování rychlosti a kvality byl proveden na všech datech, i na subjektech určených konkrétním adresním místem z RÚIAN. Rychlost při provádění na serveru dosahovala průměrně kolem 20 000 dotazů za hodinu. Oproti testům na desktopu byla tato rychlost až dvojnásobná. Z toho plyne nepřekvapivý závěr, že popsaný způsob geokódování je silně závislý na výpočetních schopnostech hardwaru. Výhodou potom může být, že s optimalizací procesu a se zvýšením výpočetní kapacity nebude problém zajistit i vyšší rychlosti geokódování.

Celková doba procesu není přesně známá, výpočty probíhaly po dávkách (většinou po několika statisících až půl milionech záznamů), s občasnými výpadky na straně serveru. Vše však bylo provedeno během dvou týdnů s tím, že samotný výpočet trval přibližně 8 dní. Výsledky kvality geokódování v závislosti na roku vzniku ES jsou na obr. 4.



Obr. 4 Výsledky kvality geokódování podle data vzniku ES

Kruhový diagram ukazuje podíly počtu geokódovaných subjektů podle 4 úrovní kvality geokódování. Nejlepší úroveň znamená plnou úspěšnost geokódování, u subjektu byla identifikována jedinečná dvojice ulice + obec nebo jedinečná obec, u které se ulice neuvádí (obce, kterých se to týká, byly zjištěny dotazem v RÚIAN v tabulce ulic). Takto identifikovaných záznamů bylo jednoznačně nejvíce, konkrétně přes 3 673 000.

Druhá úroveň obnáší přesnost na úrovni obce, tentokráte však takové, u níž nebyla určena ulice, přestože podle RÚIAN by se tam ulice nacházet měly. Počet je zatížen určitou nejistotou, neboť v případech nemála obcí mohlo dojít k vymezení ulic až po době vzniku ES. Tato nejistota může být relativně značná vzhledem k tomu, že největší část subjektů této kategorie vzniklo již v roce 1991. Důležitá je však alespoň správně a jedinečně určená obec.

Úroveň tři značí nejednoznačné identifikování. U subjektů této kategorie se nepodařilo určit unikátní umístění ani pomocí ověření kódu PSČ. Celkově se jednalo o 21 432 záznamů.

Poslední kategorie je složená ze záznamů, které geokódovat nelze. Buď chybí zápis adresy úplně, nebo je ve tvaru např. “nedefinováno”. Takových subjektů bylo dohromady téměř 34 tisíc.

Uvedené hodnocení kvality výsledků neříká tolik o geokódovacím algoritmu jako spíše o samotných datech, tzn. z chybných či nekompletních záznamů nelze zjistit správnou polohu ekonomického subjektu. Zjednodušeně řečeno, na chybných nebo nekompletních adresách zkrátka nelze správnou pozici najít. Vliv na správnost má také datum vzniku záznamu (viz obr. 4), v dobách rozvoje soukromého podnikání po revoluci se podle toho, co ukazují data, příliš na kvalitu zápisu nedbalo. V současnosti je to již až na ojedinělé výjimky podchyceno. Dalším významným vlivem na kvalitu je způsob ověřování jednoznačností podle PSČ. To se totiž v průběhu doby občas změnilo, buď byla zrušena

pošta, nebo se měnily dodávací územní obvody pošt. Například pro Olomouc se v současné době využívá kolem dvou desítek konkrétních PSČ, přičemž v RES jich je u všech záznamů dohromady uvedeno 185, pro ilustraci se však jen 40 z nich v záznamech vyskytuje více než třikrát.

Kvalita algoritmu se však dá dobře hodnotit podle teoreticky správných údajů adres. Předpokládá se, že záznamy, u nichž jsou jmenovitě v odpovídajících atributech udány adresní informace včetně kódu administrativní jednotky, mají tuto adresu zapsanou ve správném tvaru. Tyto subjekty byly taktéž podrobeny geokódování a neúspěšně (tzn. kvalitativní úrovně třetí nebo čtvrté) z nich vzešlo všehovšudy 1420 z 2 635 000 možných, což dělá chybovost algoritmu celkových 0,05 %.

Popisovaný princip geokódování je značně nedokonalý, při bližším přihlédnutím do záznamů se objeví předem těžce předvídané logické chyby. Například některé záznamy obsahovaly v adresním textu i pražskou městskou část “Holešovice”. Pokud u nich nedošlo k nalezení shody města a ulice, testovala se shoda adresního řetězce s názvy obcí. Tyto záznamy mohly obsahovat i název obce - “Praha”, ale algoritmus jdoucí podle názvu obcí od nejdelšího nalezl první shodu s obcí “Holešov” a tu zapsal do atributu. A jelikož Holešov je jediný v Česku, nebyla třeba dále potvrzovat adresní jednoznačnost a tyto záznamy byly nesprávně přiřazeny do tohoto moravského města a vypadaly jako správně geokódované. Druhým příkladem může být obec Vinařice, ve které se mimo jiné nachází ulice pojmenované římskými čísly - I. ulice, II. ulice, … až po IX. ulici. Algoritmus první, druhou, třetí, šestou, sedmou a osmou uložil shodně jako “I. ulice” protože to byla první shoda, jež nalezl. Dá se očekávat, že podobných logických chyb bude více, velmi těžce se však odhalují a v takovém množství záznamů se jedná o zanedbatelnou část. Současně v rámci této práce nebyl prostor zabývat se jejich vychytáváním, propracovanější princip geokódování s optimalizovaným průběhem by však mohl být námětem na samostatnou diplomovou práci, data z RES by k tomu mohla sloužit jako vhodná testovací data.

7 PŘÍPADOVÉ STUDIE

Hlavním výsledkem přípravné fáze byla v PG jedna kompletní tabulka RES, obsahující kromě všech dostupných původních atributů i souřadnice centroidu obce nebo ulice v závislosti na kvalitě geokódování. Charakteristickou vlastností RES je, že většina atributů není dostupná pro všechny záznamy, a zároveň pro některé analýzy bude třeba jen určité množiny z celku podle požadovaných hodnot atributů. Výsledná tabulka tedy tvořila základ pro specifické konkrétní výběry vhodných ES pro zamýšlené analýzy. Popisované kroky jsou opět k dispozici v příloze v souboru se zdrojovým kódem doplněným poznámkami a vysvětlivkami, který je k dispozici na CD v souboru pojmenovaném po případové studii. Tyto jsou v obyčejném textovém formátu z důvodu zachování kompaktnosti případové studie i přes časté střídání struktury kódu (podle stylu SQL a R). Rozdílná struktura je viditelná zejména způsobem značení komentářů.

7.1 Analýzy a vizualizace časových řad: Vznik a zánik ekonomických subjektů

Cílem první případové studie bylo vizualizovat vývoj počtu ekonomických subjektů se zaměřením na kategorie CZ-NACE, na tomto základě vytvořit animaci a data vyhodnotit pomocí analýzy přežití. Nejdůležitějšími atributy pro první případovou studii byla data vzniku a zániku ES. Na začátek je vhodné vysvětlit, co je myšleno pod těmito pojmy. Vznik ES z pohledu dat RES obnáší datum zápisu do živnostenského rejstříku v případě nové fyzické osoby anebo datum zápisu do obchodního rejstříku v případě právnické osoby. Těmto úkonům předchází určitá doba vyřizování potřebných dokumentů, ale hned v nejbližší době po zápisu se předpokládá, že ES začne vykonávat svou obchodní činnost. Vypovídací hodnota data vzniku je tedy na velmi dobré úrovni - datum vzniku reprezentuje reálný začátek funkčnosti subjektu. Oproti tomu zánik ES je zde sice opět představován jednoduchým datem, kdy byla oficiálně ukončena jeho činnost, tomuto momentu ale předchází určitá doba, kdy už tuto činnost nevykonával. Tato doba se může velmi lišit. V nejkratších případech (obnášející například úmrtí aktivní fyzické osoby nebo fúzi dvou společností) může být činnost daného subjektu ukončena prakticky ze dne na den, častěji však samotnému ukončení předchází doba útlumu činnosti, kdy se například obchodům moc nedaří, ale stále je nějaká činnost vykazována, pak se teprve činnost reálně zastaví, ale ke zrušení právní existence ES ještě může být potřeba například vyřídit insolvenční záležitosti. Současně nejsou vzácné případy, kdy člověk má vyřízené živnostenské oprávnění, ale tuto činnost vykonává jako vedlejší a je zároveň zaměstnancem. Tyto případy nelze monitorovat nebo hodnotit, míra aktivity může být různá a než dojde ke zrušení živnostenského oprávnění, opět tam může nastat relativně dlouhá doba nečinnosti. Vypovídací hodnota data zániku je tedy znatelně nižší a přidává do výpočtů určitou dávku nejistoty, kterou je třeba brát v úvahu zejména při interpretaci výsledků.

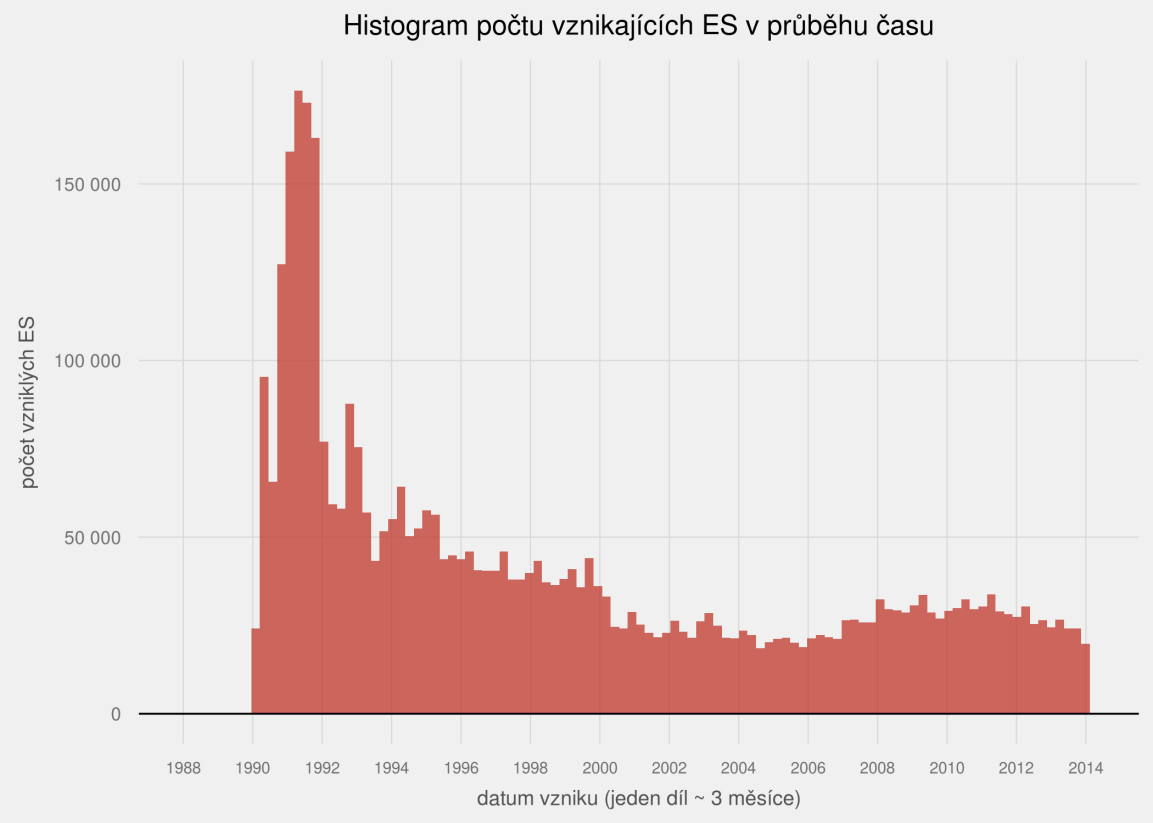
Úpravy časových atributů

Aby se dalo s datem pracovat, například porovnávat co je starší/mladší, je nutné konvertovat tento textový atribut do datového typu date. K tomu slouží funkce `to_date(atribut, 'DD.MM.YYYY')` kde první parametr je původní datumový atribut a druhý parametr je jeho tvar (tak aby počítač pochopil původní strukturu a správně si ji převedl). Dále všem "živým" ES bylo nastaveno datum zániku na 31. 3. 2014. Původně

toto datum u “živých” ES nebylo vyplněné, což je pro porovnávání nevhodné, výpočet by si nevěděl rady s dotazem např. *vyber všechny ES které zanikly po 1. 1. 2000*, a zároveň i pro vizualizace bude vhodné mít i u všech “živých” konkrétní datum na konci časové linie datového souboru. Vzhledem k tomu, že pouze zanedbatelné množství (necelé 0,5 %) ze všech ES vzniklo před rokem 1990, bude vhodné časový rozsah datového souboru zkrátit i zleva - posunout datum vzniků a zániků starších záznamů (raději do nového atributu) na 31. 12. 1989. Před touto změnou by byla případná časová osa ve vizualizacích zbytečně dlouhá a pro roky před rokem 1990 by nebylo pro malý počet záznamů oproti pozdějším rokům prakticky nic rozeznatelného. Popisované kroky jsou přepsány v konkrétní příkazy ve zdrojovém kódu v příloze na CD v souboru #1_time_series.txt.

Vývoj počtu ES

Existence dat vzniků a zániků nahrávají první vizualizaci, a to frekvenci vzniků/zániků. Pochopitelně nemá smysl sčítat a vizualizovat vzniky pro konkrétní dny, v časovém rozmezí 1. 1. 1990 - 31. 1. 2014 jsou i měsíce stále hodně nepřehledné, byl proto zvolen interval tří měsíců. Pro vizualizaci bylo využito spolupráce PG a R. Jednotlivé kroky přepsané do kódu jsou v příloze se zdrojovým kódem k první analýze. Znázornění počtů vznikajících ES je zobrazeno na obr. 5.



Obr. 5 Počty vznikajících ES

V grafu je vidět významný vrchol během prvních dvou let po revoluci následovaný pomalým poklesem, které se zastavilo až s příchodem roku 2007. Roky 2008–2011 se nesou ve znamení mírného vzestupu, po roce 2012 již opět vznikajících ES ubylo. Zajímavým úkazem je fakt, že světová ekonomická krize, probíhající po dobu několik let

zhruba od roku 2008, se vůbec na vznicích nových subjektů negativně nepodepsala, ba naopak, přinesla mírné oživení.

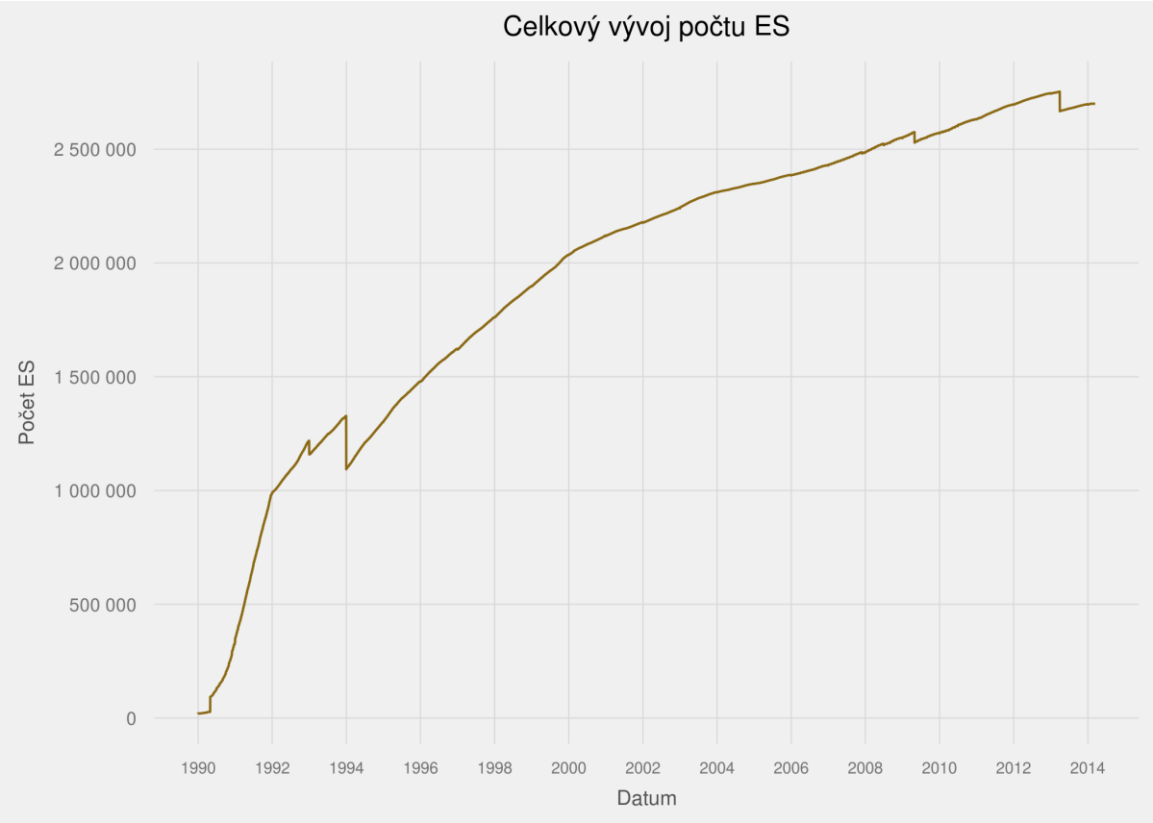
V případě zániků ES (viz příloha 3) se vzhledem k nejistotě ovlivňující data nedá příliš dobře hodnotit ekonomické prostředí. Za zmínku však stojí na první pohled viditelné čtyři tříměsíční úseky, které svými počty zániků výrazně převyšují dlouhodobé průměry. Za těmito extrémy stojí vždy konkrétní dny, jmenovitě se jedná o tato data:

- 31. 12. 1993, počet zanikajících ES: 235 001
- 31. 3. 2013, počet zanikajících ES: 86 358
- 31. 12. 1992, počet zanikajících ES: 52 331
- 30. 4. 2009, počet zanikajících ES: 46 996

Byly podniknuty kroky k hledání odůvodnění tohoto jevu, včetně dotazu na patřičná místa u správce rejstříku, nicméně s nevalnou úspěšností. K 30. 4. 2009 spadal konec platnosti Osvědčení o zápisu do evidence samostatně hospodařících rolníků, který nutil dotčené živnostníky se přeregistrovat, kdo tak neučinil, byl z rejstříku smazán - takových bylo tedy celkem necelých 47 tisíc. Důvodem poklesu počtu jednotek v roce 1994 je dokončení promítnutí přeregistrace podnikatelů podle živnostenského zákona. Lze předpokládat, že i za ostatními velkými "čistkami" rejstříků budou stát nějaké změny či úpravy zákona nebo vyhlášek týkajících se rejstříků, jejich existence se však nepodařila potvrdit.

Kumulativní počet ES

Když se pro každý den sečtou všechny vzniky, odečtou zániky a výsledek se přičte k předchozímu dni, vznikne tzv. kumulativní součet aktivních ES. Pro tento postup nebyla v R nalezena samostatná funkce, která by takovou funkcionalitou disponovala, a byla tedy vytvořena. Nejprve byla v PG do nové tabulky vypsána všechna data vzniků nebo zániků od nejstaršího. K nim byly přidány počty vzniků a počty zániků, do dalšího atributu potom jejich rozdíl, znamenající denní součty. Následně v R pomocí jednoduché funkce `cumsum` (dostupné v základním statistickém balíku) byl vytvořen požadovaný kumulativní součet. Tato funkce jednoduše sečetla hodnoty v předchozích řádcích tabulky, předpokladem správného výsledku bylo správné seřazení dnů. Alternativní variantou by mohl být databázový skript opakující pro každý den výběr ve znění "vyber počet všech ES, u nichž platí podmínky data vzniku menší než současné datum a data zániku větší než současné datum". Ve výstupu (obr. 6) jsou opět patrné čtyři "čistící" dny v rejstřících - v těchto dnech došlo k výraznému poklesu aktivních ES. Jinak se křivka představující počet aktivních ES dá rozdělit na tři fáze, první je výrazný růst, probíhající v letech 1990–1994. To vcelku odpovídá ekonomické situaci v té době v česku, kdy probíhal velmi výrazný rozvoj podnikání. V druhé fázi se tento růst mírně zpomalil, v letech 1994–2000 sice stále značně ES přibývalo, ne však tolik jako dříve. Ve třetí fázi (od roku 2000 dále) došlo k ještě výraznějšímu zpomalení růstu, nicméně k jeho úplnému zastavení nedošlo. Na vývoj počtu ES působí v různé míře mnohem více faktorů, konkrétních, jako prostředky státní podpory podnikání nebo výše daní, ale i těžce specifikovatelných a abstraktních, jako například nálada ve společnosti nebo změny trhu s příchodem EU. Pro jejich vysledování by však bylo zapotřebí mnohem podrobnějšího se zaměření na toto téma, což není úkolem této diplomové práce.



Obr. 6 Celkový (kumulativní) vývoj počtu aktivních ES

Je vhodné nyní opět zmínit ekonomickou krizi. Jejím projevem v českém prostředí bylo mírné zvýšení počtu vzniků a současně i zániků, takže celkový počet zůstal stále mírně rostoucí, jak je dobře viditelné na kumulativní křivce. Vysvětlení se nabízí takové, že náročnější ekonomická situace přinesla zvýšení nezaměstnanosti a současně častější zániky ES. Vzhledem k tomu, že nabídka pracovních pozic v takových obdobích nebývá pro nezaměstnané přívětivá, začínají se objevovat tendence zkusit soukromé podnikání. Stejná situace může nastat u neúspěšných podnikatelů, kteří můžou chtít zkoušet začít znovu. Potvrzují to záznamy RES, v roce 2008 bylo oproti roku 2006 o 31 % více vznikajících živnostenských oprávnění, a stejné čísla se držely po následující tři roky, než došlo opět k mírnému snížení (viz tabulka 1).

Tabulka 1: Počet vznikajících živnostenských oprávnění pro FO po roce 2000

rok	počet FO
2000	75482
2001	63615
2002	62667
2003	62581
2004	52038
2005	43768
2006	46470

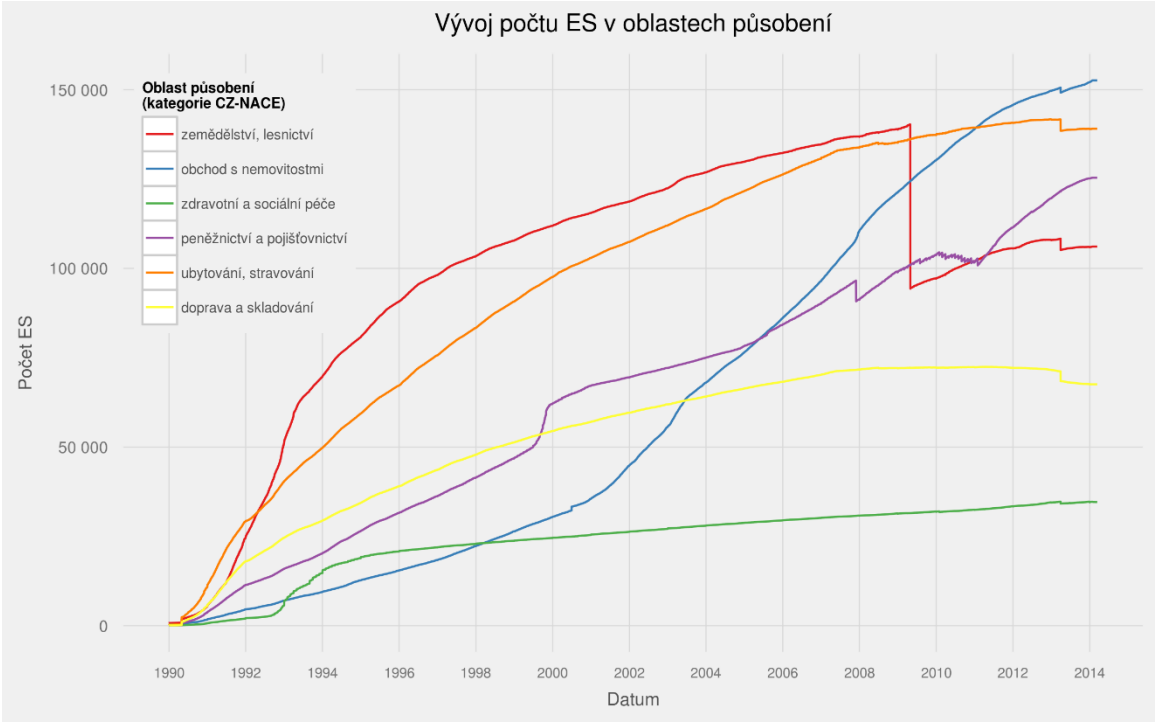
rok	počet FO
2007	54085
2008	61060
2009	60745
2010	63774
2011	66177
2012	58842
2013	54633

Vývoj počtu ES dle odvětví CZ-NACE

Tak jako pro celou sadu ES probíhal výpočet kumulativního součtu, je možné využít některých kategorických atributů dat a provést výpočet pro jednotlivé kategorie. Byl vybrán atribut NACEF, obsahující oblasti působení subjektů v kategorizaci CZ-NACE. Ta se dělí do různých úrovní podrobnosti, od nejobecnější třídy (značené písmenem), přes druhou úroveň (dvouciferné číslo)a podobně dále až po pátou úroveň (pěticiferné číslo), představující konkrétní činnost. Každý subjekt má v RES v atributu NACEF jednu hodnotu (statisticky převažující činnosti) z těchto možných úrovní, vždy tu co nejkonkrétnější. Pokud subjekt například provádí tři různé činnosti, hledá se poslední společná úroveň. Často u velkých subjektů je to první úroveň, malé subjekty mívají specifikaci čtvrté nebo páté úrovně. Pro sumarizační účely je však potřeba, aby byly kromě podrobných údajů známy i hodnoty obecnějších úrovní. Řešením je rozšíření (join) tabulky RES o připravenou tabulku, ve které jsou rozkódovány NACE hodnoty do všech úrovní. Při procesu napojování je přidáno ke každému záznamu pět atributů - pro každou úroveň jeden. V případě, že hodnota v RES je páté úrovně, doplní se všechny úrovně nižší, pokud je v RES obecná první úroveň, všechny ostatní úrovně zůstanou nulové. Výpis hlavních kategorií hlavní úrovně CZ-NACE současně s četností subjektů do nich patřících je v tabulce v příloze číslo 5.

Pro urychlení výpočtu (který se cyklicky opakuje pro každou kategorii) byl vytvořen PHP skript, který vytvoří do nové tabulky denní součty pro všechny kategorie. (K dispozici v příloze na CD mezi programovými kódy v souboru dailysum.php.) Princip je však stejný, jako tomu bylo u výpočtu celku. Nejprve se vytvoří tabulka všech dní, kdy probíhaly změny v počtech subjektů, následně pro každou kategorii se spočítají k jednotlivým dnům vzniky, zániky a posléze jejich rozdíl (= denní změna). Tabulka s denními změnami se poté v R pomocí funkce `cumsum` přepočítá na kumulativní součty.

Z celkových možných 22 základních kategorií první úrovně (resp. 23 včetně kategorie "neurčeno") bylo nejprve vybráno 12 kategorií, vhodných pro vizualizaci. Jde převážně o kategorie, pod nimiž si lze snadno představit, o jaké subjekty se jedná. Výsledný graf kumulativních součtů je v příloze pod číslem 4. Z těchto 12 kategorií bylo dále vybráno 6 (viz obr. 7), které mají nějakým způsobem zajímavý vývoj, hodný větší pozornosti a interpretace. Samotná interpretace je založena na tezi, že počet aktivních ES přímo souvisí s trhem v daných oblastech. Když je vysoká poptávka po poskytovaných službách, směřuje to k růstu počtů subjektů, a naopak, když je trh v dané oblasti nasycený, vývoj počtu ES stagnuje či klesá.



Obr. 7 Vývoj počtu aktivních ES ve vybraných oblastech působení

První kategorie, zemědělství a lesnictví, má oproti ostatním vybraným odvětvím početně nejsilnější nástup, který se zvolňuje kolem roku 1997, ale až do roku 2009 počet aktivních subjektů stále poměrně jistě roste. V dubnu 2009 s koncem platnosti Osvědčení o zápisu do evidence samostatně hospodařících rolníků (diskutovaným v oddíle s názvem Vývoj počtu ES) dochází k diskrétnímu významnému úbytku, nicméně tento je následován opět stabilním růstem. Křivky kategorií ubytování a stravování (oranžová barva) a doprava a skladování (žlutá barva) mají podobný tvar, s tím rozdílem, že ubytování a stravování dosahuje přibližně dvojnásobných hodnot než druhá jmenovaná kategorie. Obě ale mají protáhlý parabolický tvar, s tím že kategorie doprava a skladování dosáhla nejvyšších hodnot přibližně kolem roku 2009, od té doby začíná opět klesat; křivka ubytování a stravování vrcholu dosáhla v roce 2013 a průběh křivky napovídá následovné přinejmenším zastavení růstu, ne-li přímo klesání. Průběh počtu aktivních ES v oblasti zdravotní a sociální péče svědčí o relativní nasycenosti trhu v tomto odvětví již v roce 1995, přičemž stačily pouhé dva roky k tomu, aby bylo dosaženo 55 % prozatím konečné hodnoty. Od tohoto bodu následoval již jen velmi mírný, avšak stabilní růst. Téměř lineární podoby (až na výkyvy) dosahuje křivka kategorie peněžnictví a pojišťovnictví. Odchylkou od linearity je například vzestup před rokem 2000 nebo značná nestabilita v době ekonomické krize, ze které se však toto odvětví úspěšně vzpamatovalo v roce 2011 a opět se nastartovalo k silnému růstu. Otázkou je, jak dlouho bude takový vývoj udržitelný. Stejná hrozba platí pro poslední zobrazenou kategorii - obchodu s nemovitostmi, které se začalo dařit až po přelomu tisíciletí. Od té doby ale má nejsilnější růst, který nezastavila ani hospodářská krize.

Podrobnější činnosti

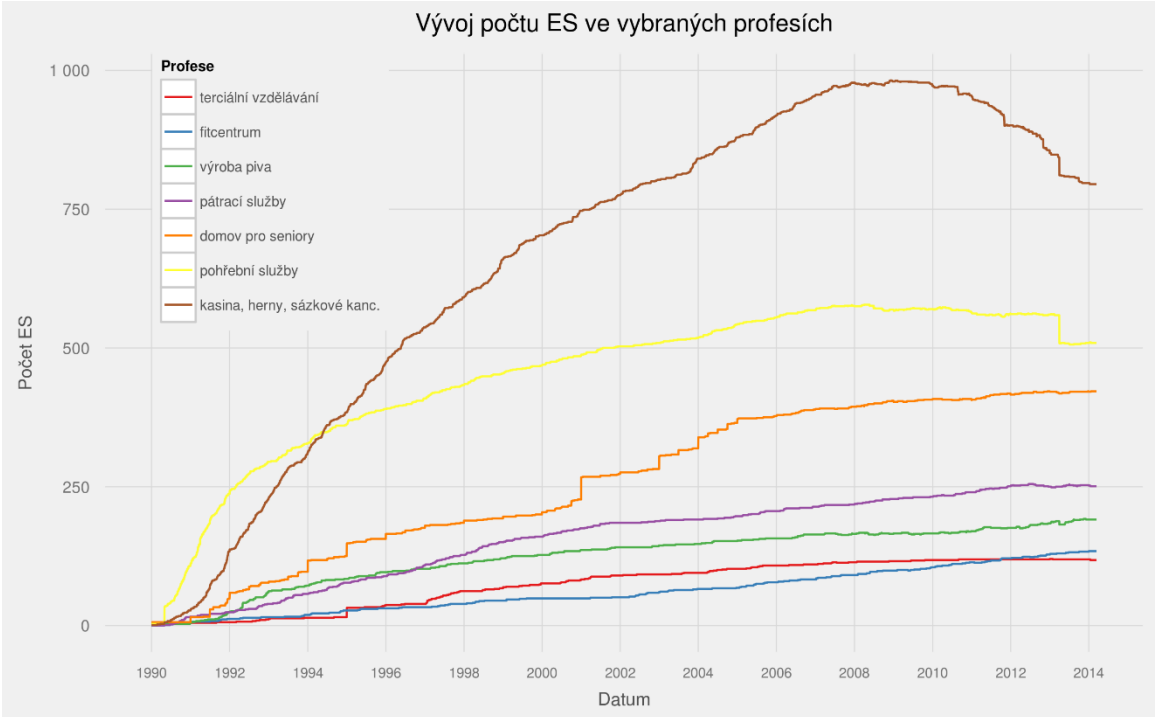
Kategorizace CZ-NACE nabízí možnost blíže se podívat na konkrétní činnosti. Byly tedy vytipovány některé kategorie zejména třetích a čtvrtých úrovní. Výběr nebyl řešen tematicky, ale podle počtu aktivních subjektů tak, aby vybrané kategorie tvořily tři

početně podobné skupiny a ve výsledných vizualizacích mohly být společně srovnávány. Seznam vybraných kategorií:

- výroba piva, fitcentra, domovy pro seniory, pohřební služby, pátrací služby, terciální vzdělávání, herny, kasina a sázkové kanceláře
- programování, zubní péče, umělecká tvorba, organizace pro děti a mládež, fotografické služby, těžba dřeva, zeměměřičství a kartografie
- kadeřnictví a kosmetika, reklamní služby, živočišná výroba, silniční doprava, instalatérské práce, nábytkářství

Způsob, jímž je v RES ukládána informace o oblasti působení, s sebou přináší významnou nevýhodu. Protože je u každého záznamu uváděna pouze jedna hodnota, schyluje se to k obecnějším hodnotám NACE, a když potom dochází k výběru konkrétních činností, jsou vybírány pouze ty subjekty, u nichž je požadovaná činnost uváděna přímo, nejedná se tedy o všechny subjekty, které tuto činnost reálně provozují. Jako příklad může sloužit první jmenovaná konkrétní činnost, výroba piva. Velké pivovary často kromě piva produkují i jiné nápoje (např. nealkoholické), což ale je jiná konkrétní kategorie, a tyto subjekty mají uvedenu hodnotu NACE obecnější úrovně. Naopak to ovšem nefunguje, nelze vybrat obecnější kategorii pouze z důvodu existence možnosti, že onu konkrétnější činnost budou provozovat též. V uváděné kategorii výroby piva jsou tedy vybrány pouze subjekty, u nichž je toto hlavní (a dá se říct jedinou) činností, zejména tedy minipivovary, a velké nápojové a pivovarské koncerny mezi nimi většinou chybí.

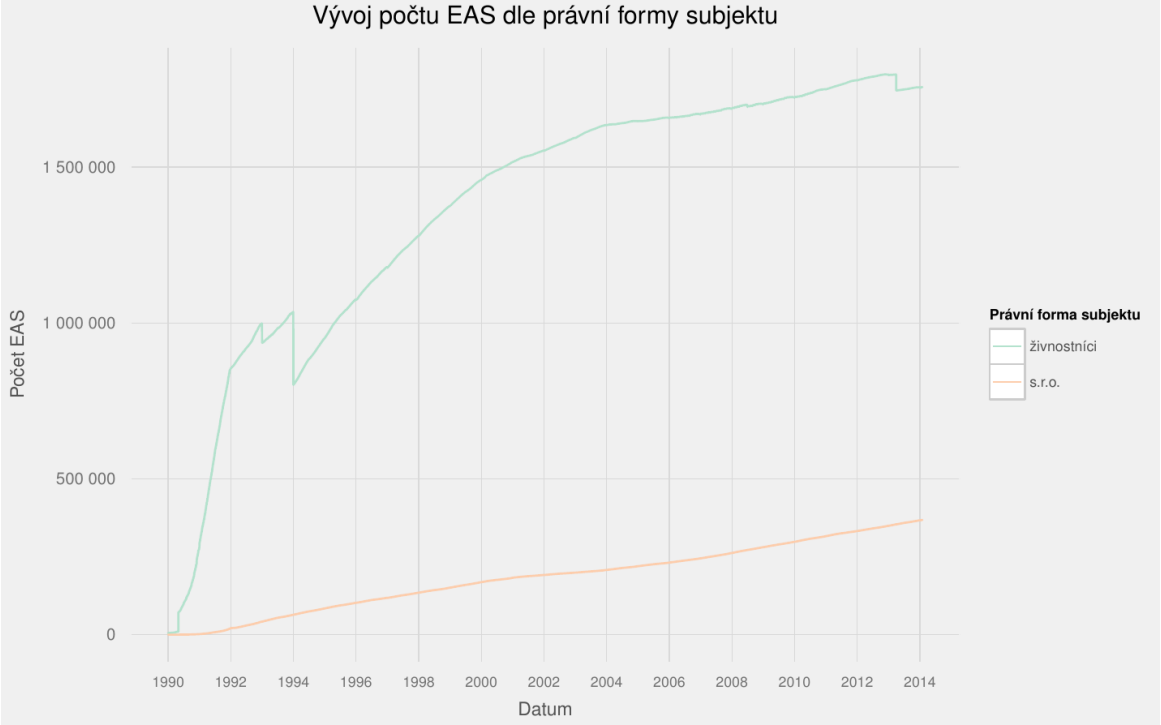
Výsledné grafy kumulativních součtů aktivních ES vybraných činností jsou na obr. 8 a v přílohách 6 a 7. Jejich přínos je však čistě zajímavostně - informativní, nejsou zde proto ani podrobeny interpretaci.



obr. 8 Skupina vybraných činností a jejich kumulativní součty

Vývoj počtu aktivních ES podle právních forem

Při této analýze byly vybrány dvě nejčetnější právní formy subjektů, a to fyzické osoby podnikající dle živnostenského zákona (=živnostníci) a společnost s ručením omezeným - s.r.o. Z výsledku (viz obr. č. 9) je pozorovatelné, že křivka živnostníků, kteří tvoří tři pětiny z celého množství ES, v podstatě přímo odpovídá křivce celého souboru (obr. 6) včetně jednorázových poklesů nebo fází vývoje. Až k závěru (od r. 2004 dále) je viditelný nepatrně menší přírůstek živnostníků oproti celku, dá se předpokládat, že důvodem je zvýšení podílu ostatních právních forem subjektů. Křivka subjektů s.r.o. potom ukazuje téměř rovnoměrný růst, bez nějakých významných výkyvů.



Obr. 9 Kumulativní křivky ES dle nejčastějších právních forem

Animovaný vývoj

Se znalostí časových značek vzniků/zániků a zároveň prostorové lokalizace záznamů se nabízí možnost vytvoření grafické animace. Relativně jednoduché řešení je v R pomocí balíků sp[41], spacetime[42] a plotKML[43] (pozn. pod čarou - odkazy/citace). Vybere se cílová skupina z dat - například zmínění výrobci piva, kromě atributu oblasti působení bude potřeba časových údajů (vznik a zánik) a prostorové určení (souřadnice x a y). Po načtení těchto informací do prostředí R se z této běžné datové matice vytvoří prostorový objekt definicí souřadnic (x a y) a odpovídajícího souřadnicového systému, tento objekt se následně transformuje do třídy STIDF (časoprostorový data frame), současně s definicí atributů pro začáteční a koncový čas platnosti záznamů. Z této třídy je možnost přímého exportu do formátu KML, který je implicitním formátem volně dostupného programu Google Earth. Při samotném zobrazování je v nabídce programu automatické animované přehrávání s možností libovolné změny pozorovatelova umístění nad trojrozměrným

[41] https://cran.r-project.org/web/packages/sp/index.html
[42] https://cran.r-project.org/web/packages/spacetime/index.htmldrg
[43] http://plotkml.r-forge.r-project.org/

povrchem. Navíc si zde uživatel může poupravit grafické parametry zobrazovaných objektů. Zdrojový kód postupu vytváření animace je v souboru #1_time_series.txt, výsledný kml soubor potom komprimovaný na CD jako pivovary.kmz.

Analýza přežití

Analýza přežití je označení pro skupinu statistických metod využívaných k analyzování dat se známým časovým intervalem, tedy s jasně definovaným počátkem a koncem. Nejznámější využití je v medicíně, kdy se například monitorují účinky různých léků při léčení vážných chorob. Počátek intervalu je čas začátku léčení, konec intervalu přichází s úmrtím infikovaného, a zkoumá se vliv léku, resp. pravděpodobnost nastání události (= úmrtí) při použití daného léčiva. Odtud pochází termín "doba přežití" užívaný pro časový interval (Uher, 2011). Analýza přežití má ovšem své využití i v jiných oblastech, jako strojírenství (například životnost strojů), sociologii a psychologii (délka studia na vysoké škole, ...), nebo i v ekonomii. V této práci je dobou přežití myšlena doba mezi vznikem a zánikem ES a cílem je zjistit, v jakém odvětví dle kategorií CZ-NACE je vyšší pravděpodobnost neúspěchu podnikání. Mohou však nastat také případy, kdy subjekt skončí svou existenci z jiných důvodů, než jsou pro analýzu podstatné, nebo doba jeho života je delší, než zkoumaná. Tyto (a další) jevy mají souhrnný název cenzorování. Vhodnost dat zaznamenaných v souboru RES pro účely analýzy přežití roste s atributem ZPZANF, který udává způsob zániku ES. Nabízí tedy možnost brát v úvahu pouze žádané důvody zániku, ostatní cenzorovat. Cílem bylo tyto důvody omezit pouze na ty, které s nejvyšší pravděpodobností představovaly důsledek vlivu trhu a ekonomického prostředí. Brány v potaz byly tyto důvody:

- Zrušení právnických osob likvidací
- Zrušení právnické osoby bez likvidace a bez nástupců
- Oznámení fyzické osoby o ukončení činnosti
- Rozhodnutím z důvodu nepřeregistrace

Další důvody (jako úmrtí fyzické osoby nebo její odstěhování z okresu, ukončení duplicitního IČO, zrušení právnické osoby bez likvidace s nástupci - rozdělením nebo fúzí, nebo ukončení duplicitního IČO atd.) byly vyhodnoceny jako pro analýzu nežádoucí.

Z datového souboru RES byly vybrány ty, které vznikly v roce 1990 (společný začátek) a je u nich znám způsob zániku a CZ-NACE kategorie. Výpočty v prostředí R provádí balík survival[44]. Na vstupu je třeba mít v tabulce čas zániku jako numerickou proměnnou, dále v jedné dichotomické proměnné informaci, zda u subjektu pozorovaná událost nastala (subjekt zanikl z vybraných důvodů) nebo nenastala (cenzorování) a v poslední proměnné obsaženou kategorii (CZ-NACE).

Nejprve se testuje (metodou Log-Rank), zda je mezi křivkami přežití rozdíl, skrze funkci `survdiff`.

Ve výpisu tohoto testování (viz příloha č. 8) vychází najevo, že je mezi jednotlivými křivkami významný rozdíl, což lze poznat z hodnoty p = 0, jež je menší než námi zvolená hladina významnosti 5% (tj. $p < 0,05$). Zároveň se ukazuje značný rozdíl mezi počty subjektů v jednotlivých kategoriích, zatímco nejmenší skupina *těžba a dobývání* je zastoupena 28 subjekty, v nejčetnější skupině - *profese* je zástupů 21 532. Tento fakt má vliv zejména na různou míru statistické významnosti.

[44] https://cran.r-project.org/web/packages/survival/index.html

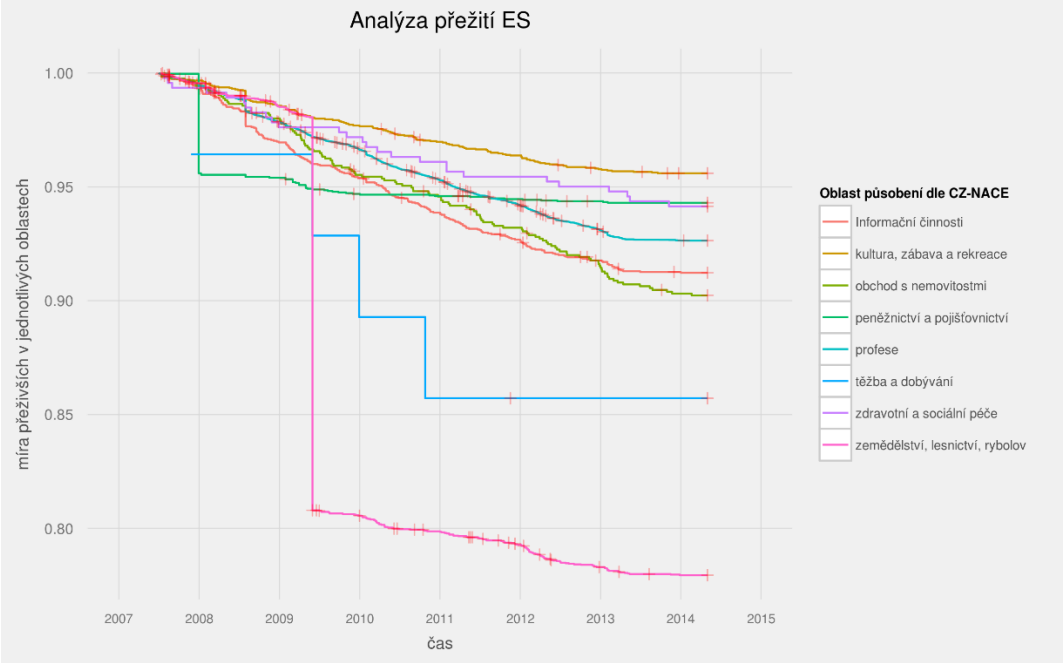
Další fází výpočtu je nalezení odhadu parametrů do Coxova regresního modelu a určení jejich statistické významnosti (viz obr. 10). Lehce se tím zjistí, zda jsou všechny kategorie CZ-NACE důležité.

```
> coxph(Surv(df_rok2$zaniknr, df_rok2$event) ~ df_rok2$jmenovka)
Call:
coxph(formula = Surv(df_rok2$zaniknr, df_rok2$event) ~ df_rok2$jmenovka)

                                                 coef exp(coef) se(coef)      z       p
df_rok2$jmenovkakultura, zábava a rekreace     -0.713     0.490   0.1033 -6.900 5.2e-12
df_rok2$jmenovkaobchod s nemovitostmi           0.102     1.107   0.1098  0.928 3.5e-01
df_rok2$jmenovkapeněžnictví a pojišťovnictví   -0.427     0.653   0.0977 -4.367 1.3e-05
df_rok2$jmenovkaprofese                        -0.187     0.830   0.0670 -2.784 5.4e-03
df_rok2$jmenovkatěžba a dobývání                0.532     1.702   0.5038  1.056 2.9e-01
df_rok2$jmenovkazdravotní a sociální péče      -0.421     0.656   0.2022 -2.083 3.7e-02
df_rok2$jmenovkazemědělství, lesnictví, rybolov 1.007     2.736   0.0703 14.328 0.0e+00

Likelihood ratio test=892  on 7 df, p=0  n= 36910, number of events= 3249
```

Obr. 10 Odhad parametrů do Coxova modelu

Pro každou kategorii jsou spočítány odhady v parametru „z“ coxova modelu „coef” s hladinou signifikance v parametru „p“. Ty jsou ve všech případech velmi blízko nule, odhady se tedy dají považovat za statisticky významné při zvolené hladině významnosti 5%.

Nejznámější vizualizací analýzy přežití jsou odhady funkce přežití, někdy nazývané jako Kaplan-Meierovy křivky (viz obr. 11).



Obr. 11 Funkce přežití pro různá odvětví (Kaplan - Maierovy křivky)

Vizualizace potvrzuje, že mezi kategoriemi CZ-NACE je významný rozdíl z pohledu funkce přežití. Většina kategorií ale končí s podobnou mírou přeživších (v rozmezí cca 5 %), výjimku tvoří kategorie *těžba a dobývání* a dále kategorie *zemědělství, lesnictví a rybolov*, na kterém se nejvíce podepsal již zmíněný den 30. dubna 2009, kde hodně

subjektů bylo z živnostenského rejstříku vymazáno z důvodu nepřeregistrace, což zapříčiňuje nejmenší pravděpodobost přežití mezi kategoriemi. Naopak nejlépe si stojí kategorie *kultura, zábava a rekreace*, z čehož se dá s nadsázkou usuzovat, že podnikatelé začínající byznys v tomto odvětví v roce 1990 měli lepší šanci na úspěch oproti ostatním odvětvím. Nejhůře na tom byli zemědělci, lesníci a lidé zabývající se rybolovem. O trochu lepší situaci měli podnikatelé v oblasti těžby a dobývání. Ostatní odvětví na tom byly víceméně podobně.

7.2 Mapování hlavních sídel ekonomických subjektů a jejich vzory v prostoru a čase

Charakteristickým rysem druhé případové studie je zaměření na obce. Záměrem je porovnat ekonomické aktivity subjektů a počet obyvatel obce či výši daňových příjmů obce. Popisované postupy jsou přepsány do zdrojového kódu a náležitě okomentovány na CD v souboru #2_sidla.sql.

Porovnání počtu ES s počtem obyvatel

Prvním krokem je příprava dat. Databáze ArcČR 500 obsahuje v geodatabázi administrativního členění datovou vrstvu obcí, která může sloužit jako vhodný základ, protože obsahuje kromě identifikátoru obce shodného s RÚIAN zároveň počty obyvatel ve sčítacích letech 1991, 2001, 2011 a další statistické údaje, které lze využít. Alternativou by mohla být vrstva obcí RÚIAN, která již je importovaná v PG, k ní by ale ještě byla potřeba připojit statistická data.

Způsobů, jak nahrát tabulku obcí z formátu ESRI geodatabase (gdb) do PG je mnoho. Knihovna GDAL od verze 1.11 již podporuje nativně formát gdb (GDAL, 2014), její implementace se projevila v QGIS od verze 2.10, kde lze číst geodata těchto formátů. V předešlých verzích byla třeba pro zpracování gdb doinstalovat externí potřebné drivery, nebo například provést export do shapefile v programu ArcGIS. Otevření této vrstvy v QGIS přináší výhodu jednoduchého kopírování do PG skrze Správce databází s přívětivým uživatelským rozhraním.

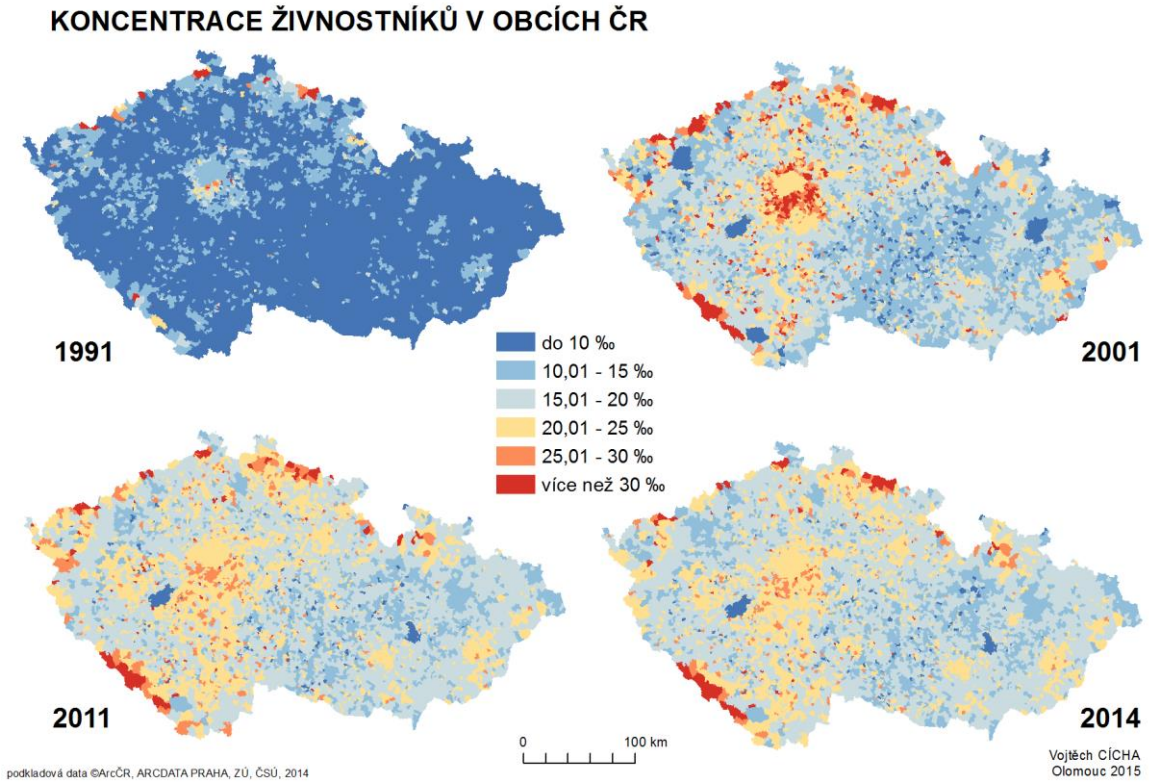
Pomocí identifikátoru IDOB mohou být do obcí agregovány počty aktivních ES. Bylo rozhodnuto, že tyto počty se stanoví právě ke sčítacím rokům, pro snadné porovnání. Při samotném sčítání (v prostředí PG) bylo využito SQL funkcí `distinct` (vracející "od každého jen jeden") a běžné `count`, vždy v kombinaci s časovou podmínkou. Pro rok 1991 kupříkladu měla podmínka tvar "vznik menší než 1. 1. 1992 a zánik větší než 31. 12. 1991". Tyto součty byly vytvořeny do nové dočasné tabulky, odkud se pomocí join připojily k tabulce obcí. Následné zkoumání porovnání součtů ES s počty obyvatel (= koncentrace ES) ukázalo jen slabou pravidelnost v chování, a to jak v grafických náhledech[45] tak tabelárních výstupech (viz tabulka obcí s nejvyššími koncentracemi ES v příloze č. 9). Obecně se sice dala identifikovat tendence obcí s koncentrovanějším výskytem ES shlukovat se buď v horských turistických a z pohledu obyvatelstva malých lokalitách, nebo v zázemí velkých měst, ale v každém z těchto tří časových úseků se stav výrazně měnil a zároveň se vyskytovalo množství obcí, které tyto snahy o pravidelnosti

[45] Protože následující koncentrace pouze živnostníků prokazovala velmi podobné chování jako celá skupina všech ES, ale mnohem lépe viditelné, grafické výstupy pro koncentraci ES nejsou vůbec publikovány.

porušovaly. Mezi obce, které ve všech třech obdobích dosahovaly nejvyšších koncentrací, patří Hřensko, Strážné, Loučná pod Klínovcem nebo Český Jiřetín; jedná se o malé obce (do 300 obyvatel) z turisticky významných oblastí Krušnohoří a Krkonoš. Zajímavé dále bylo zjištění existence celkově tří obcí, v níž bylo identifikováno větší množství ES než počet obyvatel s trvalým bydlištěm. Například v obci Květnice u Prahy bylo v roce 2001 registrováno 212 aktivních subjektů, zatímco obyvatel s trvalým bydlištěm pouze 120. Nicméně nutno dodat, že zde probíhal výrazný suburbanizační proces, v roce 2011 bylo obyvatel již desetinásobné množství, a zároveň podnikání (ani OSVČ) nemusí být vázáno na trvalé bydliště, stačí prokázání vlastnického či užívacího práva k objektu/prostorám, kde je hlášeno podnikání.

Počet živnostníků v obcích

Nyní po obecném součtu všech ES přichází na řadu specifické výběry, prvním z nich je skupina živnostníků, tedy fyzických osob (FO) podnikajících podle živnostenského nebo jiného zákona, samostatně hospodařící rolníci nebo podnikající osoby v zemědělství. Vzhledem k faktu, že se jedná o silně majoritní skupinu ES, předpokládá se, že její chování bude podobné jako chování celku. Stanovení koncentrací FO proběhlo obdobně pouze s přidáním podmínky k atributu právnické formy subjektu. Součet FO v obci byl stanoven i pro rok 2014 (tzn. Subjekty, které nezanikly nebo zanikly v průběhu ledna 2014). Koncentrace za všechny čtyři roky byly poskládány v jednu vizualizaci (viz série nepravých kartogramů na obr. 12).



Obr. 12 Srovnání koncentrace živnostníků v obcích v letech 1991, 2001, 2011 a 2014

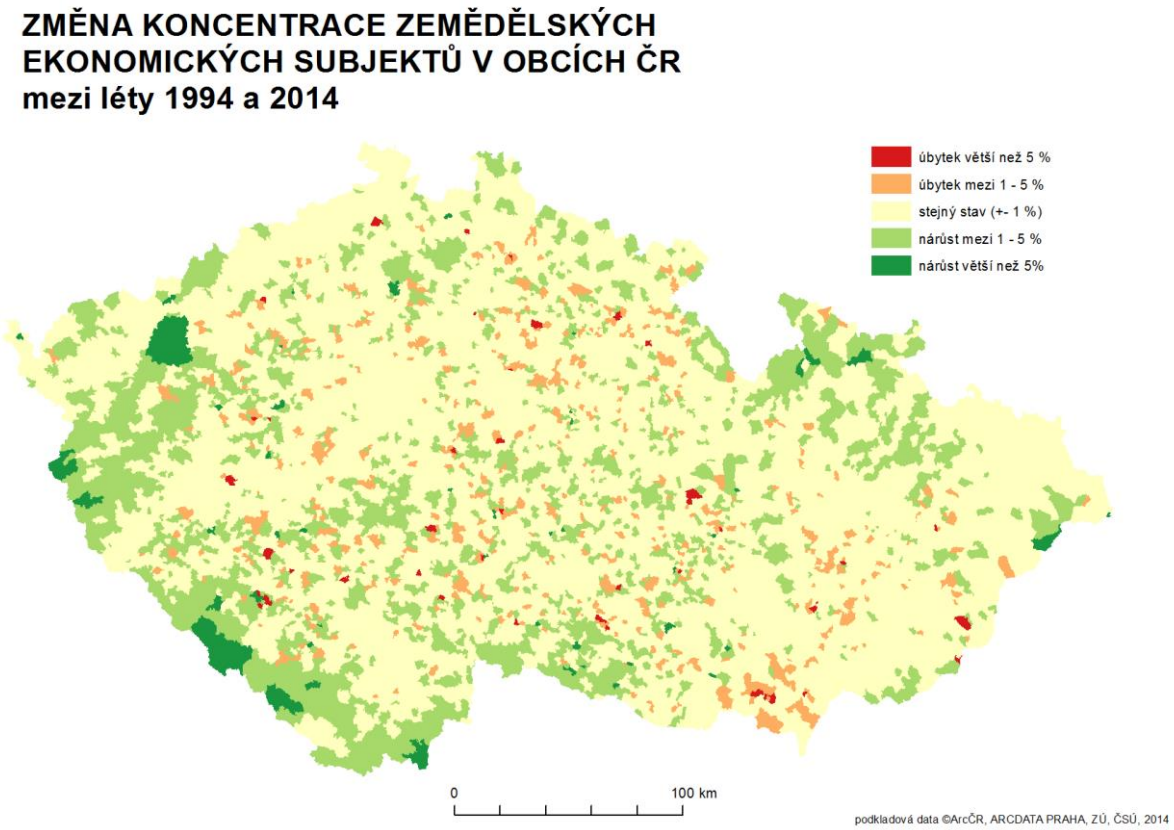
V roce 1991 ještě bylo živnostníků relativně málo. Už však v tomto roce dosahovaly některé pohraniční horské obce vysokých hodnot koncentrace. Je to zcela jistě důsledek turistického ruchu a zároveň malého počtu obyvatel s trvalým bydlištěm v těchto obcích.

Za následujících deset let došlo k viditelnému růstu koncentrace ve většině obcí ČR, nejvýrazněji v oblasti kolem Prahy a také v horských turisticky lákavých lokalitách, zejména v Krkonoších, Krušnohoří a na Šumavě. Naopak nižších hodnot se stále drží obce na Vysočině. Do roku 2011 poté pozoruhodně poklesla hodnota koncentrace v pražském zázemí. V době před rokem 2011 probíhaly významné ekonomické procesy a změny trhu související s ekonomickou krizí, a je více než pravděpodobné, že zde pozorovaný jev s tím souvisí. Mezi roky 2011 a 2014 už poté nedošlo k téměř žádným pozorovatelným změnám. Popisovaná vizualizace celkově potvrzuje vyšší ekonomickou vyspělost Čech oproti Moravě a Slezsku, v Čechách je na první pohled mnohem více obcí s vyššími hodnotami koncentrace.

Počet zemědělců

Pomocí výběru subjektů z odvětví zemědělství a porovnání dřívějšího a současného stavu se nabízí možnost určení, ve kterých oblastech se zemědělství vyvíjelo (přibývaly subjekty) nebo upadalo. Nárůst subjektů dále může znamenat i nadále probíhající decentralizaci trhu, rozpad velkých subjektů na drobnější, a úbytek potom slučování subjektů, prodej hospodářství větším firmám, nebo taktéž k úbytku koncentrace dojde se zachováním počtu subjektů při růstu obyvatel obce. Metoda výpočtu koncentrací je srovnatelná s předchozími, kdy je sesčítán počet aktivních subjektů v obcích - tentokrát z odvětví CZ-NACE druhé úrovně s názvem *Rostlinná a živočišná výroba, myslivost a související činnosti*, a na tomto základě je stanovena koncentrace poměrem na počet obyvatel. Bylo rozhodnuto, že v tomto případě se počty vytvoří k začátku roku 1994 a 2014. Důvodem je již potvrzený fakt (z obr. 5), že z pohledu počtu vznikajících subjektů se stav v roce 1994 již dá po předcházejících letech výrazné komercionalizace a přechodu ze socialistického plánovaného hospodářství na tržní považovat za stabilizovaný. Protože pro rok 1994 není dostupný údaj o počtu obyvatel v obcích,[46] poslouží k porovnání údaj z roku 1991. Výsledek dvacetiletého vývoje trhu z pohledu počtu ES v odvětví zemědělství je vidět na obr. 13.

[46] Tento údaj je dohledatelný na webu ČSÚ pro každou obec zvlášť, zisk pro všechny obce ČR by byl velmi pracný, pro účely analýzy bude stačit porovnání k roku 1991



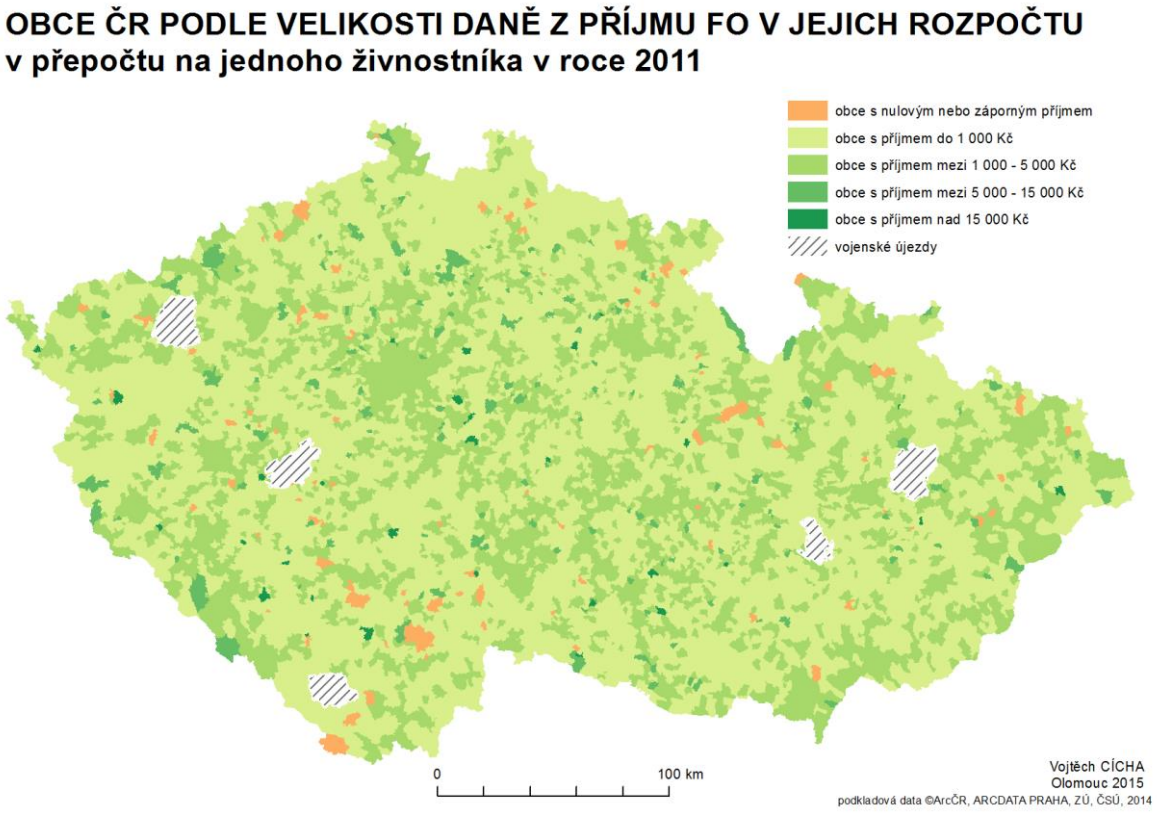
Obr. 13 Změna koncentrace zemědělců v obcích mezi léty 1994 a 2014

Z mapy lze vyvodit závěr, že převážná část obcí České republiky za 20 let zůstala s koncentrací zemědělců na stejné úrovni. K výraznějšímu růstu počtu subjektů došlo v hornatějších oblastech, zejména v jižních a západních Čechách, v Jeseníkách a nepravidelně na Vysočině. Důvodem může být v dnešní době čím dál častější návrat k tradičnímu zemědělství. K úpadku potom velmi roztroušeně docházelo po celé ploše Republiky, nejvýrazněji potom na jihu Moravy. Obecně se dá říci, že větší změny nejsou příliš četné, dohromady jen u 17 obcí došlo ke změně stavu přes deset procent, absolutně největší úbytek hlásí obec Hradiště (okres Plzeň - jih) s hodnotou -24 %.

Srovnání daňového příjmu obce a počtu ES

Poslední analýzou v rámci druhé případové studie je porovnání počtu ES s daňovým příjmem v rozpočtech obcí. Rozpočet obce je relativně komplexní záležitost, sestává se z velikého množství položek (jak je naznačeno v kapitole 5.1 v oddíle Rozpočet obcí). Pro tuto analýzu však není důležité znát a rozumět celému rozpočtu. Jedny z příjmových položek jsou příjmy ze zisků fyzických a právnických osob. U právnických osob v RES je náročné až téměř neproveditelné vyfiltrovat skutečně ty subjekty, které v reálu platí daně z příjmu, jež se potom podepíší na výši daňového příjmu konkrétní obce (jsou zde různé bezpříjmové subjekty, spolky a sdružení, státní a samosprávní subjekty nebo subjekty působící na různě velkém území atd. ...) U fyzických osob je to mnohem jednodušší, předpokládá se, že každá fyzická osoba samostatně výdělečně činná (OSVČ) si vede živnostenské oprávnění pro legální vykonávání určité činnosti za úplatu. Z hlediska zákona je v pořádku mít živnostenské oprávnění a nevykonávat činnost, ale za cenu samostatného každoročního vytvoření a podání daňového přiznání. Z dlouhodobějšího

hlediska je to rozhodně na obtíž a osoby, které živnost delší dobu nepoužívají, si běžně živnostenské oprávnění ruší. Základní premisa, z níž vychází následná analýza je následující: čím více si fyzická osoba vydělá (čistý zisk), tím více odevzdá státu na dani z příjmu. Jsou zde tzv. úlevy na dani (pro rodiče s dětmi, dárce krve, studenty, …) které do této rovnice mírně zasahují, neruší ale její obecnou platnost. Každá obec potom do svého rozpočtu obdrží určité procentuální množství (podle aktuálně platného zákonu o rozpočtovém určení daní) daně z příjmů fyzických osob, které mají na území dané obce bydliště. Když se to vezme obrácenou logikou, čím větší je daňový příjem ze zisků FO v obci po přepočtení na jednu FO, tím větších průměrných zisků živnostníci v dané obci dosahují. To je cílem této analýzy, odpovědět na otázku, kde se nachází oblasti s větším průměrným ziskem živnostníků, resp. jestli je možné tyto oblasti identifikovat a prostorově vymezit. Nutno dodat, že do této analýzy nemohou být zahrnuty vojenské újezdy, u nichž se rozpočet vytváří jinak.



Obr. 14 Srovnání daňového příjmu obcí z příjmů FO na počet FO v obcích

Z výsledné mapy (obr. 14) je však jisté, že cíle analýzy dosáhnout nelze. Prostorová distribuce větších hodnot je na ploše celého státu dá se říci nahodilá, výjimkou jsou snad jen střední Čechy, kde se nepatrně více drží vyšší hodnoty. Další zobecňující závěry však z pohledu statistiky nemají význam.

Velký vliv na výsledek má neznámé množství FO, které sice disponují platným živnostenským oprávněním, ale aktivně jej nevyužívají (nebo při běžném zaměstnání jej využívají omezeně). Takovíto živnostníci značně zkreslují přesnost výsledku, s menším množstvím zaplacené daně z příjmu snižují průměr obce. Naopak se mohou taktéž vyskytnout z pohledu příjmů i velké subjekty, které jsou právně vedeny jako živnost jedné fyzické osoby (například zemědělské farmy nebo významní finanční poradci). Ty zase mohou průměr vychylovat na opačnou stranu. Popsané jevy se potom projevují zejména

v obcích s menším počtem obyvatel. Dále nutno brát v potaz podstatu rozpočtu obce; není zcela vypovídající jeden vytržený rok od jiných, neboť každý rozpočet nese různě velké pozůstatky předchozího hospodaření obce. Na druhou stranu výše státního příspěvku do obecní pokladny v položce dani z příjmu FO tímto faktem není tolik zatížena jako jiné obecní výdaje a příjmy. Problémem však můžou být zásahy finančního úřadu, který je oprávněn dorovnávat chybně dopočítané daně zpětně až po tři roky.

7.3 Klasifikace zón ekonomické aktivity

Třetí případová studie si klade za cíl pomocí multiparametrického shlukování vytvořit novou klasifikaci České republiky z pohledu ekonomických aktivit. Pro vybranou základní administrativní jednotku budou vytvořeny koncentrace aktivních subjektů podle odvětví CZ-NACE a podle tří časových období, do roku 1995 včetně, od 1996 do 2005, a od 2006 dále.

Shlukování

Shlukování jako metoda statistické analýzy se snaží z množiny prvků vytvořit skupiny (shluky, clustery) podobných prvků tak, aby si prvky uvnitř shluku byly co nejpodobnější a zároveň celý shluk byl od ostatních shluků co nejodlišnější (Kučera, 2008). Konkrétní metody se dělí podle přístupu na hierarchické a nehierarchické. Hierarchické jsou tvořeny posloupností rozkladů, kdy na jedné straně je jeden shluk tvořen všemi objekty a na straně druhé jednoprvkové shluky, a tyto metody se snaží najít optimální rozložení. Nehierarchické metody k celé množině nepřistupují podle hierarchie, ale rozdělují ji do podmnožin podle specifikovaného kritéria. Počet shluků může být buďto předem definovaný nebo v některých případech variabilní. Nejznámějším zástupcem této skupiny a nejspíš i obecně analýzy shlukování je metoda K-means. Více informací o konkrétních metodách shlukové analýzy, vysvětlení jejích principů a příklady použití viz například v (Kučera, 2008).

Prostorové shlukování může probíhat buď přidáním dvou atributů (souřadnic x a y) k běžné metodě, nebo lze nalézt nemálo metod, které s prostorovým faktorem umí přímo pracovat. Zejména v geoinformatických programech jsou takové metody dostupné, výhodou bývá jednoduchá volba míry důležitosti prostorového parametru, například podmínka sousedství s dalšími prvky ve shluku apod. Při této analýze bylo zvoleno metod, které nabízí program ArcMap ze série ArcGIS for Desktop od firmy esri. Jeho výhodou je mimo jiné rozsáhlý výstupní soubor charakteristik k vytvořeným shlukům.

Příprava dat

Původním záměrem bylo použít jako základní prostorovou jednotku katastrální území, avšak kvůli velkému počtu na tak podrobné úrovni nelokalizovaných subjektů by potenciálně mohlo dojít k významným nepřesnostem a určitému zkreslení. Základní jednotkou byla následně vybrána obec, ale po provedení série shlukových analýz (s různými nastaveními parametrů) bylo shledáno, že chování obcí je velmi rozdílné, charakteristiky shluků vykazovaly velmi nízkou úroveň vnitřní podobnosti a zároveň rozdíly mezi shluky navzájem byly nepatrné (viz obr. v příloze 10). Důkazem rozdílnosti byl i pokus, kdy bylo provedeno shlukování neprostorově, při zobrazení v mapě se prostorová distribuce jednotlivých prvků shluku dala hodnotit jako čistě náhodná (viz obr. v příloze 11) a stejně nebylo dosaženo lepších charakteristik shluků. Řešením mohlo být zkusit shlukování provést ve větších administrativních jednotkách, kterými jsou obce

s rozšířenou působností (ORP). Postup přípravy dat je však pro obě varianty až na agregační identifikátor stejný, bude popsán pro variantu s ORP.

Základ tabulky vytvořila opět vrstva z databáze ArcČR, byly však z ní pro budoucí přehlednost odstraněny všechny atributy kromě identifikátoru a názvu ORP a počtů obyvatel v letech 1991, 2001 a 2011. V tabulce RES byl ke každému lokalizovanému subjektu přidán identifikátor ORP (pomocí join tabulky obcí). Byla vytvořena kopie tabulky RES a v ní pomocí vymazávání probíhal výběr cílové skupiny vhodných subjektů. Odstraněny byly všechny nelokalizované subjekty a ty, které neměly specifikovanou základní úroveň kategorie CZ-NACE. Dále byly ponechány jen právní formy, které by se daly označit jako "soukromé a profitující" - fyzické osoby, společnosti s ručením omezeným, zahraniční osoby, akciové společnosti apod., odstraněny byly formy jako sdružení, obce, organizační složky státu, školy nebo církevní organizace atd. Následně byla třeba mírně upravit kategorizaci CZ-NACE první úrovně s cílem menšího počtu proměnných pro shlukovou analýzu. Kategorie *Ostatní činnosti* a *Profesní, vědecké a technické činnosti* byly sloučeny do nové kategorie nazvané *Služby občanům* a kategorie *Zásobování vodou; činnosti související s odpadními vodami, odpady a sanacemi* s kategorií *Výroba a rozvod elektřiny, plynu, tepla a klimatizovaného vzduchu* potom do jedné pojmenované *Technické zázemí*. Došlo k odstranění početně výrazně malých kategorií *Těžba a dobývání* a *Veřejná správa a obrana; povinné sociální zabezpečení*, které měly zastoupení jen v řádech stovek subjektů. Výsledných patnáct tříd i s počty subjektů jsou viditelné v tabulce 2.

Tabulka 2: Upravené základní odvětví pro shlukovou analýzu s počty ES

kód	název	počet ES
G	Velkoobchod a maloobchod; autoopravárenství	811506
X	Služby občanům	507178
F	Stavebnictví	375811
C	Zpracovatelský průmysl	348651
I	Ubytování, stravování a pohostinství	164080
A	Zemědělství, lesnictví, rybářství	158147
K	Peněžnictví a pojišťovnictví	153083
L	Činnosti v oblasti nemovitostí	109637
H	Doprava a skladování	82896
J	Informační a komunikační činnosti	62060
N	Administrativní a podpůrné činnosti	58696
Q	Zdravotní a sociální péče	36747
P	Vzdělávání	36631
R	Kulturní, zábavní a rekreační činnosti	35283
Y	Technické zázemí	30529

Po eliminaci všech prvků, které nemají být započítávány v této analýze, a po úpravě kategorií, je dalším krokem jejich agregace. Je třeba připravit si za každé časové období sloupec pro všechny kategorie, celkem tedy 45 sloupců. Standardním řešením je mnohonásobné nakopírování SQL příkazu pro přidání sloupce s jednotlivými úpravami, nebo elegantnějším a univerzálnějším řešením je využít do sebe vnořených cyklů procházejících automatizovaně všechny hodnoty časové a kategoriální složky (viz skript

na CD #3_sloupce.php). Pro samotnou sumarizaci subjektů byl vytvořen další skript (#3_soucty.php). Základem jsou tři do sebe vnořené cykly. První prochází tři podmínky časové platnosti subjektů podle vybraných období. Ve druhém probíhá výběr z databáze pro každé ORP v dané časové platnosti s příkazem `group by` pro atribut kategorií odvětví a s požadavkem součtu členů `count`. To znamená na výstupu výběru je dvourozměrné pole kategorií se součty. V posledním cyklu se prochází po řádcích výstupní proměnné z předešlého cyklu a pro každý proběhne uložení hodnoty do databáze. Výsledkem skriptu jsou tedy u každého ORP uvedené počty aktivních subjektů dle období a kategorie. Dále je třeba tyto absolutní hodnoty přetvořit na relativní vzhledem k počtu obyvatel v ORP tak, aby bylo umožněno vzájemné porovnávání. Tuto činnost řeší skript #3_koncentrace.php, který vznikl jen malou úpravou skriptu na vytvoření sloupců. Jeho specifikum je v naplněním čerstvě vzniklého sloupce podílem mezi počtem subjektů a počtem obyvatel daného období = koncentrací.

Korelace

Před provedením shlukové analýzy je vhodné prověřit, do jaké míry jsou si odlišné jednotlivé parametry (kategorie). Pokud by míra podobnosti (označována jako korelace) parametru s ostatními byla často silná, znamenalo by to, že mezi parametry existují silné asociace. Při prokázané korelaci atributů potom klesá množství informace, kterou atribut do souboru přináší a pro shlukovou analýzu je dá se říci zbytečný. Hodnoty koncentrací byly tedy podrobeny korelační analýze. V ní se jako nejvíce korelovaná ukázala proměnná X (Služby občanům, různé profese) s korelačním koeficientem parametru v polovině případů dosahující minimálně hodnoty 0,6. Bylo rozhodnuto, že tento parametr tedy nebude vstupovat do shlukové analýzy. Podobným případem byla kategorie G, zde koeficienty dosahovaly podobně vysokých hodnot, nicméně ve třetím období hodnoty koeficientu významně klesly, a tento parametr zůstal pro shlukovou analýzu zachován. Další hodnoty koeficientů viz soubory korelace95, 05, 14.txt na CD.

Analýza hlavních komponent PCA

Běžnou praxí ještě před shlukovou analýzou je zredukování dimenze analyzovaného souboru, například pomocí analýzy hlavních komponent (Principal Component Analysis - PCA). Ta lineární kombinací znaků (parametrů) do několika tzv. komponent zjednodušuje datový soubor za cenu co nejmenší ztráty informace (Přibylová, 2015). Když si představíme parametry souboru jako osy n-rozměrného prostoru a hodnoty jako souřadnice, PCA provede transformaci souřadnic do jiného systému, na základě míry variability = rozptylu. Osa hlavní komponenty povede ve směru největší variability, druhá a další komponenty jsou pak kolmé na předchozí a popisují maximální možné množství zbývající variability souboru (Hebák a kol., 2005).

Komponenty jsou potom charakterizovány množstvím variability z celkového souboru a popisují, kolik variability konkrétních znaků vysvětlují. Cílem je, aby každá komponenta vysvětlovala co nejvíce znaků, ale zároveň aby každý znak byl obsažen v co nejméně komponentách, nejlépe jediné (Hebák a kol., 2005). Výpočet se provádí nejčastěji z korelační matice, případně kovarianční nebo lze i z vlastních dat. Pro stanovení optimálního počtu se dá využít metod vlastních čísel korelačních matic (tzv. eigenvalues) nebo určitá stanovená hodnota rozptylu, např. běžně udávaná uspokojivá variabilita 70 % (Přibylová, 2015).

```
> loadings(pca)

Loadings:
      Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15
x05ob -0.348         0.105                       -0.140                -0.104  -0.122   0.212   0.104  -0.334   0.801
h05ob -0.267 -0.233         0.236 -0.157  0.113  0.434 -0.266 -0.545          0.415          -0.146   0.106
l05ob -0.316                0.220 -0.196  0.129 -0.110         0.361  0.469           0.207  -0.399   0.467
c05ob -0.119 -0.454  0.262         0.621 -0.271 -0.315        -0.181  0.288                  -0.133
j05ob -0.330         0.199  0.132         0.197 -0.150                                 0.486   0.134  -0.433  -0.556
i05ob -0.129 -0.189 -0.766                -0.399        -0.208         -0.110  -0.137   0.177  -0.233  -0.194
q05ob -0.245  0.166 -0.344 -0.256  0.116  0.254 -0.216  0.628 -0.413          0.174
f05ob -0.160 -0.499         0.240  0.178  0.107  0.345  0.495  0.403 -0.309
g05ob -0.327                                    0.198        -0.287         -0.717  -0.173   0.346   0.290
k05ob -0.197  0.182        -0.665  0.260         0.548 -0.146  0.228  0.135           0.110
n05ob -0.333                0.102                                       0.361   0.137  -0.684          -0.471
p05ob -0.308  0.144  0.112         0.117  0.177 -0.250 -0.249         -0.613  -0.113  -0.315  -0.403   0.162  -0.135
y05ob -0.107 -0.378  0.234 -0.499 -0.630 -0.280 -0.161  0.148
a05ob  0.125 -0.463 -0.275 -0.192         0.676 -0.168 -0.296         0.154          -0.152   0.161
r05ob -0.323        -0.152                -0.188 -0.187 -0.180  0.234 -0.106   0.445           0.632   0.304

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15
SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000   1.000   1.000   1.000   1.000   1.000   1.000
Proportion Var  0.067  0.067  0.067  0.067  0.067  0.067  0.067  0.067  0.067   0.067   0.067   0.067   0.067   0.067   0.067
Cumulative Var  0.067  0.133  0.200  0.267  0.333  0.400  0.467  0.533  0.600   0.667   0.733   0.800   0.867   0.933   1.000
```

Obr. 15 Výsledek PCA pro časové období 1996-2005

Při pohledu na výstup prováděné PCA (obr. 15) je zřejmé, že tato data pro PCA nejsou vhodná, respektive výsledek po provedení PCA není použitelný. Jednak znaky jsou obsaženy ve velkém počtu komponent a jednak celková variabilita s hlavními komponentami přibývá velmi pozvolna. Přesněji řečeno, hlavní komponenty ani nalezeny nebyly, poněvadž všechny nalezené komponenty dosahují stejné míry variability celku - 0,067 %. Záměrem bylo ve výsledku dostat přibližně k počtu kolem pěti komponent, což by ale v tomto případě vysvětlovalo pouze 33 % variability. Pomůckou bývá ještě metoda rotace komponentních os (např. varimax), ale ani ta nepřinesla lepší výsledek.

PCA tedy svůj cíl nesplnila, nastalá situace prozrazuje o datovém souboru obecně nízkou variabilitu, což není pro následující shlukovou analýzu dobré. Hrozí, že shluky budou těžce identifikovatelné a mezi sebou málo rozdílné. Nicméně bylo rozhodnuto přesto přistoupit k shlukové analýze.

Provedení shlukové analýzy

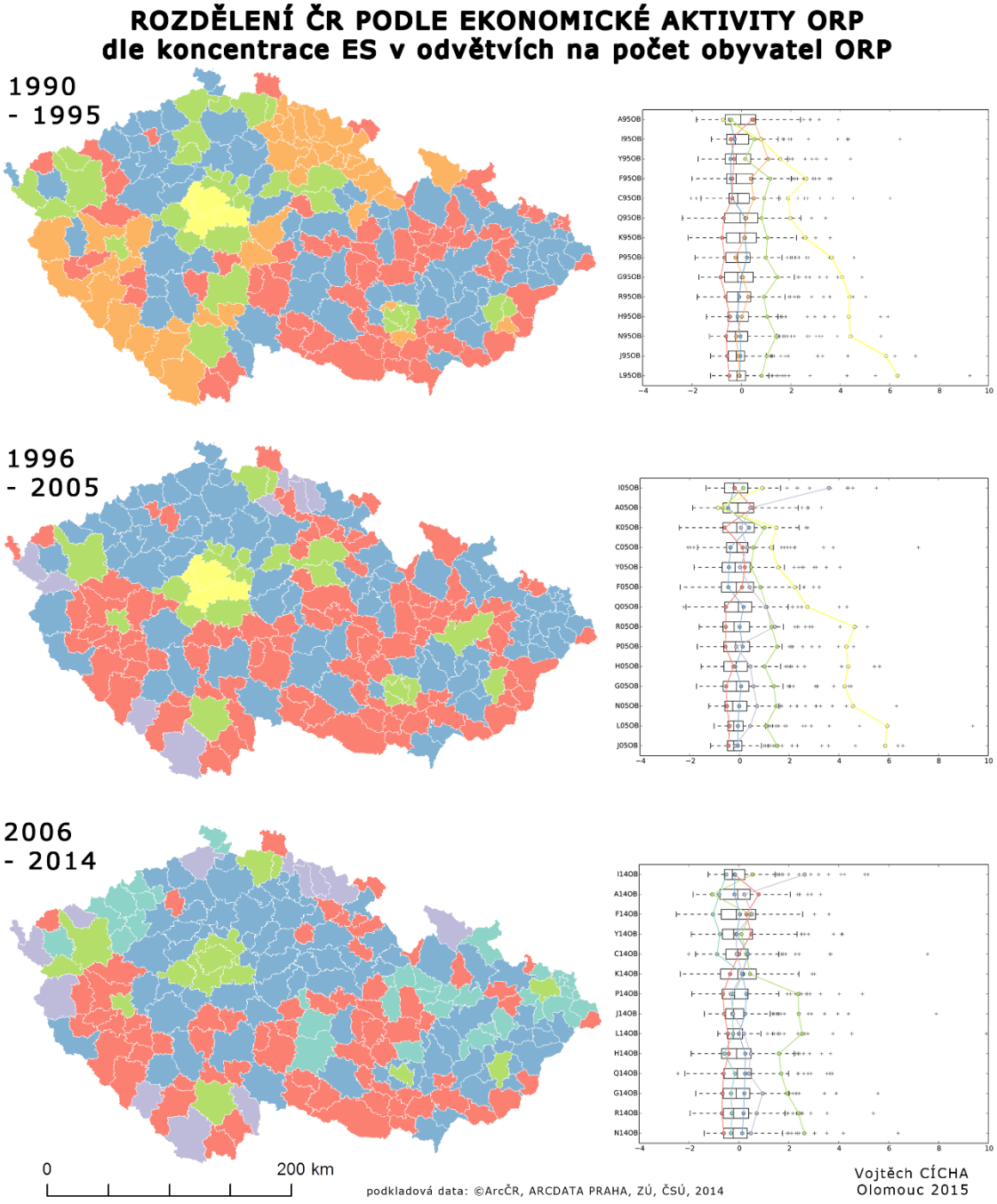
V programu ArcMap, v toolboxu Spatial Statistics Tools a jeho podčásti Mapping Clusters se nachází nástroj Grouping Analysis, který provádí shlukování. V jeho nastavení jsou kromě řadových věcí jako výběr prostorové vrstvy a atributů vstupujících do výpočtu nebo název výstupních souborů (prostorová vrstva se shluky a dokument s charakteristikami) i nastavitelné parametry, které výrazně ovlivňují výsledek. Jedním takovým je počet shluků. Program sice disponuje možností vypočítat optimální počet shluků při výpočtu analýzy, nicméně to je řízeno čistě statisticky, necitlivě k tématu nebo charakteru analýzy. Při testování této funkce byl nejčastěji navrhovaný počet shluků 2 a ve výsledku to znamenalo rozdělení Česka na Prahu a ostatní ORP. Vzhledem k tomu, že Praha s charakterem velkoměsta bude mít odlišné hodnoty než běžné české ORP, je toto rozdělení z pohledu dat sice pochopitelné, ale jako výstup nehodnotné. Nicméně je běžnou praxí nastavit tento počet subjektivně; zde bylo rozhodnuto vytvořit 5 shluků.

Dalším významným parametrem je výběr metody určující způsob, jak bude přistupováno k tvrobě shluků z prostorového hlediska. Možnosti jsou následující (ArcGIS help, 2014):

- CONTIGUITY_EDGES_ONLY - polygon může patřit do shluku pouze v případě, že sdílí minimálně jednu hranu s některým z dalších členů shluku

- CONTIGUITY_EDGES_CORNERS - polygon může patřit do shluku pouze v případě, že sdílí minimálně jeden hraniční bod s některým z dalších členů shluku
- DELAUNAY_TRIANGULATION - polygon (určen centroidem) nebo bod může patřit do shluku pouze v případě, že jiný člen shluku je jeho sousedem po Delaunay triangulaci
- K_NEAREST_NEIGHBORS - každý prvek shluku (bodový nebo polygonový) je minimálně k-tý nejbližší k dalšímu prvku shluku
- GET_SPATIAL_WEIGHTS_FROM_FILE - prostorové (a případně i temporální) vztahy prvků jsou definovány pomocí speciálně vytvořené matice vah
- NO_SPATIAL_CONSTRAINT - prostorové vztahy nejsou řešeny, je provedeno běžné K-means shlukování (konkrétně algoritmus NP-hard)

Experimentálně byly vyzkoušeny všechny metody a bylo zjištěno, že pokud je důraz kladen především na prostorovou souvislost výsledných shluků, rozdíly mezi charakteristikami shluků jsou potom výrazně menší. Byla tedy zvolena metoda bez prostorových vazeb, aby shluky byly co nejrozdílnější, s rizikem prostorových nesouvislostí. S tímto nastavením bylo shlukování provedeno ve všech třech obdobích. Vizualizace výsledku je na obr. 16. U každého časového období je mapa s klastry a paralelní boxplot s jejich charakteristikami. Osa X grafu je tvořena standardizovanými hodnotami proměnných (koncentrací), osa Y jednotlivými parametry - koncetracemi ES rozdělenými do upravených odvětví CZ-NACE. V názvech těchto atributů je pro interpretaci důležité první písmeno udávající kategorii, jejich vysvětlivky jsou k dispozici v tabulce 2 (výše). Barvy klastrů v mapě korespondují s barvami křivek v grafu.



Obr. 16 Vizualizace výsledků shlukování

V průniku všech tří období se podařilo identifikovat shluky vykazující podobné hodnoty koncentrace:

1. Praha a okolí (žlutá barva)
2. Regionální ekonomická centra (zelená barva)
3. Oblasti turistického ruchu (fialová barva)
4. ORP s hodnotami koncentrace poblíž mediánů (bledě modrá barva)
5. Zemědělské oblasti (červená barva)

Kromě těchto se však našly i dva shluky, ke kterým napříč obdobími nebylo nalezeno přímých ekvivalentů, a jsou zaznačeny tedy samostatně. V období do roku 1995 je jeden zabarven oranžově a další v období 2006-2014 tyrkysově.

Charakteristiky shluků

1. Praha a okolí

Oblast Prahy a jejího nejbližšího okolí dosahuje v obdobích do roku 1995 a 1995 - 2005 jednoznačně nejvyšších hodnot koncentrace téměř ve všech odvětvích s výjimkou kategorie A - *zemědělství*, a výsledek shlukování tímto vyzdvihuje dotčené oblasti jako ekonomicky nejvyspělejší. Pozoruhodné je, že v obou obdobích je skupina tvořena stejnými ORP, kromě Prahy jsou zahrnuty ještě Černošice a Říčany. V posledním období však ztrácí na významnosti a splývá s druhou nejvyspělejší skupinou.

2. Regionální centra

Z hlediska hodnot koncentrace druhá nejvyšší je skupina tvořená většinou krajskými městy, rozšířenými o vyspělejší ORP. Zejména v období do roku 1995 sem patří mnohé ORP například ze širšího okolí Prahy, západních nebo severních Čech. Na východě je členů skupiny méně, pravidelně pouze Zlínsko a Brněnsko, ve druhém období je skupina rozšířena o Olomouc a ve třetím o Ostravu. Z hlediska hodnot se tato skupina drží v nadprůměru, pouze v odvětvích F - *stavebnictví*, Y - *technické zázemí*, C - *zpracovatelský prům.* dosahuje průměrných hodnot a v odvětví A - *zemědělství* jsou koncentrace nejmenší. Obzvláště je to viditelné ve třetím období po sloučení skupiny s Pražskou oblastí.

3. Oblasti turistického ruchu

Třetí skupina není vymezena v prvním časovém období. Ve zbylých dvou však její členové, vyskytující se výhradně v pohraničí, dosahují většinou mírně nadprůměrných hodnot koncentrace. Neplatí to v případě kategorií P - *vzdělávání* a J - *informačních činností,* které patří mezi nejslabší stránky skupiny. Naopak výrazným rysem skupiny je vysoká koncentrace subjektů v odvětví I - *ubytování a stravování*, což je s nejvyšší pravděpodobností důsledek rozvinutého cestovního ruchu.

4. ORP s hodnotami koncentrace poblíž mediánů

Čtvrtá skupina je rozprostřena po celé ČR a je složena z ORP s průměrnými hodnotami ve většině kategorií. Charakteristickou vlastností pro ni je nízká hladina koncentrací v kategoriích A, F, C, Y, (*zemědělství, stavebnictví, zpracovatelský průmysl* a *technické zázemí*) kde se sice často drží mediánových hodnot, ale v porovnání s jinými shluky se jedná o hodnoty menší.

V prvním období by se sem mohla přiřadit i speciální skupina vyznačená oranžovou barvou, která většinou dosahuje hodnot okolo mediánu, zato v odvětvích A, F, C a Y a současně i I - *ubytování, stravování* jsou hodnoty významně nadprůměrné.

Ve třetím období k této mediánové skupině inklinuje speciální skupina vyznačená tyrkysovou barvou, která se většinou drží taktéž mediánu, ale v případech již zmíněných čtyřech skupin A, F, C a Y dosahuje ze všech skupin naopak jednoznačně nejmenších hodnot koncentrace.

5. Zemědělské oblasti

Poslední skupina ORP se dá charakterizovat ve všech obdobích největší koncentrací zemědělských ES. Mírného nadprůměru dosahuje ještě ve druhém a třetím období ve známých skupinách F, C a Y. Na druhou stranu v ostatních odvětvích jsou koncentrace na obecně nejslabších hodnotách.

Charakteristiky a vymezení shluků, zjištěné provedenou shlukovou analýzou, jsou nejdůležitějším výstupem třetí případové studie. Ukázalo se, že nejvíce rozdělující odvětví NACE jsou stavebnictví, technické zázemí (jakožto skupina vytvořená z odvětví CZ-NACE E - *Zásobování vodou; činnosti související s odpadními vodami, odpady a sanacemi* a CZ-NACE D - *Výroba a rozvod elektřiny, plynu, tepla a klimatizovaného vzduchu*), zpracovatelský průmysl, zemědělství a ubytování a stravování.

Česká republika je v této analýze hodnocena čistě podle koncentrací ES, pro větší vypovídací hodnotu analýzy a komplexnější hodnocení oblastí by bylo užitečné rozšířit parametry o další příbuzné témata, jako například informace o nezaměstnanosti nebo kupní síle apod.

8 DISKUZE

Jedním z cílů práce bylo vytvořit rešerši aktuální situace v oblasti dostupných dat s ekonomickou tematikou v ČR. Využitelnost tohoto výstupu práce však klesá s narůstající dobou, protože se situace mění, v některých případech poměrně rychle. Kromě neustálých aktualizací sad dochází i k publikování nových. Příkladem může být jmenované Ministerstvo Financí, které za dobu kompletování textu práce zveřejnilo 6 dalších datových vrstev.

Práce s tak rozsáhlým datovým souborem, jakým je RES, s sebou přináší svá specifika. Za výslednými postupy, které se většinou následně shrnou do několika konkrétních kroků, stojí často velké množství času, které je znásobeno právě vysokým počtem prvků. Kromě hledání co nejlepších řešení spoustu času zabere i samotné provedení příkazu. Kupříkladu operace aktualizace jednoho atributu na základě porovnání textových atributů pro čtyřmilionový datový soubor trvá na standardním počítači přibližně 15 minut. Zároveň výpočetní schopnosti počítače pro ostatní aplikace se tím velmi omezí a často nelze provádět jinou činnost. S tímto je třeba počítat při rozhodnutí se pro takovou činnost. Autor považuje za důležité zmínit tento fakt zejména pro lepší představu čtenáře, který s prací podobného rázu nemá zkušenosti.

Prostředí PostgreSQL je robustní, obsahující velké množství funkcí, rozšiřujících modulů, optimalizačních postupů apod. Zde navržené postupy však nemusí být nejoptimálnější. Za účelem zachování co nejlepší přehlednosti a pochopitelnosti jsou skládány tak, aby byly co nejjednodušší. Cílem bylo, aby se uvedené postupy daly lehce sledovat a provádět, například pro zisk nových zkušeností s prací v PostgreSQL. Citelným nedostatkem této snahy je fakt, že vzorová data RES nemohou být z licenčních důvodů přiložena na CD. Zde jsou pouze hotové prostorové vrstvy administrativních jednotek s agregovanými hodnotami z RES. Pokud by se tak opravdu našel zájemce o projití nastíněných postupů, je třeba se buďto pokusit o vlastní obstarání dat RES anebo použít data podobná s použitím potřebné míry představivosti. Nicméně některé postupy jsou bez obtíží aplikovatelné na libovolná data stejných atributových vlastností, například geokódování nebo skripty připravující agregace, kumulační součty či koncentrace.

Jedním z důležitých výstupů díla je algoritmus geokódování. V první řadě může vyvstat otázka, proč se vlastně tento způsob prováděl, když ve výsledku byly zjištěné souřadnice použity pouze v jedné vývojové animaci a pro ostatní prostorové analýzy stačil údaj o místní územní jednotce, obsažený přímo v jednom z atributů. Zvláště vzhledem k tomu, že příprava algoritmu, testování i samotný výpočet byl velmi náročný na čas. Jak však bylo zmíněno u třetí případové studie, v původním plánu bylo využít přesnějších souřadnic, kde by bylo třeba zmíněné geokódování. A po jeho vytvoření by byla škoda nenabídnout jej k dalšímu případnému vylepšení a využití. Nutno dodat, že pro další využití je vhodné jej upravit, například odstranit zjištěné nedostatky (viz konec kapitoly 6) a případně hledat další podobné. Užitečné by bylo i důkladnější testování výsledků, jmenovitě třeba cvičným geokódováním již lokalizovaných prostorových objektů s porovnáním vzdáleností (od zjištěného místa k správnému). Taková informace by mohla sloužit vyladění dalších logických chyb v postupu. Dále by byla využitelná funkce aproximace zadání, například při změně jednoho písmenka v adrese by pořád mohla být nalezena správná lokalita. Zavedení těchto pokročilejších funkcí by však vyžadovalo mnohem více práce, než bylo možné realizovat v rámci tohoto díla. Mohl by to však být zajímavý námět na samostatnou diplomovou práci.

V neposlední řadě je třeba zmínit slabší stránku některých provedených analýz - jejich vypovídací hodnotu. Netýká se to všech analýz a v dotčených analýzách je to při

interpretaci zmíněno. Například v analýze porovnání daňových příjmů obce s počtem FO by bylo hodnotnější započítat průměr položek v rozpočtu za několik posledních let. Zároveň zde je skrytá nejistota v podobě rozdílných možných příjmů fyzických osob výrazně ovlivňujících výsledek. Dále v závěru třetí případové studie, při klasifikaci ČR podle koncentrací ES pomocí shlukování by bylo pro přesnější hodnocení ekonomického prostředí užitečné rozšířit parametry o další ekonomické ukazatele z jiných zdrojů.

9 ZÁVĚR

Hlavní myšlenkou diplomové práce bylo pokusit se odkrýt informační potenciál českých veřejně dostupných dat. Rešerše aktuálního stavu ukázala, že veřejně dostupných dat s ekonomickou tematikou je mnoho a vzhledem k tomu, že podle současného trendu se množství neustále zvětšuje, roste s tím i zmiňovaný informační potenciál. Z nalezených datových sad byla vybrána databáze Registru ekonomických subjektů udržovaná Českým statistickým úřadem. Při zpracování dat takového rozsahu (více než 4 miliony záznamů) se osvědčilo databázové prostředí PostgreSQL s prostorovým rozšířením PostGIS. Ze začátku je sice náročné zvyknout si na způsob práce pomocí příkazu SQL, nicméně s rostoucími zkušenostmi začíná být využito efektivity nástrojů PostgreSQL jako jsou hromadné příkazy nebo skripty. V rámci práce bylo vytvořeno několik skriptů ve skriptovacím jazyku PHP. Všechny jsou skládány univerzálně, aby mohly být dále využity v situaci, kdy případný zájemce narazí na podobný problém. Stačí pouze upravit proměnné na začátku skriptu podle konkrétní situace.

Největší výzvou při přípravě dat bylo geokódování ekonomických subjektů. Bylo rozhodnuto nevyužít nabízených webových služeb, nýbrž pokusit se o vytvoření vlastního způsobu geokódování s přesností na ulici pomocí databáze RÚIAN. Byl navržen algoritmus, který porovnává textové řetězce adresy subjektů s názvy ulic a obcí z RÚIAN, snaží se ověřovat jedinečnost adresy podle kódu PSČ a zjištěné údaje doplňuje k tabulce subjektů. Algoritmus byl úspěšně vyzkoušen na všech datech RES, záznamů s korektně definovanou adresou se nepodařilo přesně lokalizovat 1400 z 2 636 000 možných (chybovost 0,05 %). Algoritmus sice obsahuje některé nedokonalosti, ale i tak společně s návodem na vlastní sestavení databáze RÚIAN může sloužit jako základ pro další využití na libovolných datech.

Snaha o odkrytí informačního potenciálu RES se uskutečňuje v rámci případových studií. V první z nich je využito časových a některých kategorických atributů, díky nimž se vizualizují vývojové křivky počtu aktivních ekonomických subjektů například dle odvětví. Výstupy přináší možnost nahlédnout do historického vývoje naší země od začátku devadesátých let minulého století do současnosti z pohledu ekonomických aktivit. Pozorovatelné jsou zejména výrazné růsty počtů subjektů v prvních letech a následné pozvolné stabilizování situace nebo i reakce českého trhu na světovou hospodářskou krizi v období kolem roku 2008. Aplikování analýzy přežití na datech RES dále poukázalo na odvětví, jež měla z pohledu dalšího vývoje od roku 1990 vyšší šanci na obchodní úspěch oproti jiným. Druhá studie přináší pohled na celou ČR pomocí koncentrací ekonomických subjektů v obcích. Ukázalo se například, jak se vyvíjel počet živnostníků v obcích, že počet subjektů v odvětví zemědělství se za posledních 20 let obecně příliš nezměnil, nebo které obce mají vysoký státní příspěvek do rozpočtu z daně z příjmu fyzických osob v přepočtu na jednoho živnostníka. Poslední studie pomocí metody shlukování rozděluje ČR ve třech časových obdobích do několika skupin podle míry koncentrací ekonomických subjektů v ORP, kde jsou vidět například regionální ekonomická centra, jaký je vliv hlavního města na nejbližší okolí nebo ve kterých oblastech je nejvýznamnějším odvětvím zemědělství či ubytování a stravování.

Provedenými případovými studiemi byla potvrzena vhodnost volby RES jako hlavního zdroje dat. Mezi nejdůležitější výhody se dá zařadit dlouhý časový rozsah, celorepubliková platnost a zároveň podrobnost na úrovni jednotlivých ekonomických subjektů. Analýz by mohlo být bez problému větší množství. Šlo by například propojit data RES ještě s dalšími datovými sadami, jako jsou například údaje o kriminalitě, hodnoty cen nemovitostí nebo další charakteristiky obyvatelstva. Zajímavé by mohly být i analýzy lokálnějšího rázu, kde

by se daly zkoumat prostorové vazby jednotlivých subjektů atd. Analýzy zahrnuté do tohoto díla však pro demonstrativní ukázku informačního potenciálu postačují.

Nicméně, když už se ve společnosti daří rozšiřovat množství dostupných dat, byla by škoda současně nevyužít rostoucího informačního potenciálu. Přáním autora je tak nejen obohatit čtenáře o zjištěné informace, ale nejlépe aby nastíněné postupy a použité metody analýz sloužily jako inspirace v dalších pracích odkrývajících informační potenciál, čímž je možné docílit většího užitku práce a podílet se tak na větším množství zjištěných informací, než by vůbec šlo zahrnout zde.

POUŽITÁ LITERATURA A INFORMAČNÍ ZDROJE

ArcČR 500. *ARCDATA PRAHA* [online]. 2014 [cit. 2015-06-23]. Dostupné z: http://www.arcdata.cz/produkty-a-sluzby/geograficka-data/arccr-500/

VFR Import. *ARCDATA PRAHA* [online]. 2015 [cit. 2015-06-23]. Dostupné z: http://www.arcdata.cz/produkty-a-sluzby/software/arcdata-praha/vfr-import

BOČEK, Miroslav. Sociální a ekonomická typologie území ČR (aplikace metod vícerozměrné analýzy, územní jednotky - správní obvody obcí s rozšířenou působností) [online]. 2009 [cit. 2015-04-24]. Diplomová práce. Masarykova univerzita, Přírodovědecká fakulta. Vedoucí práce Václav Toušek. Dostupné z: <http://theses.cz/id/1h31f3/>.

Odborná skupina pro OPEN SOURCE a OPEN DATA. *Česká asociace pro geoinformace* [online]. 2013 [cit. 2015-08-11]. Dostupné z: http://www.cagi.cz/os25-open-source-a-open-data

CÍCHA, Vojtěch. Správa, analýza a prezentace zdravotnických prostorových dat pomocí R. Olomouc, 2013. bakalářská práce (Bc.). UNIVERZITA PALACKÉHO V OLOMOUCI. Přírodovědecká fakulta. Vedoucí práce Lukáš Marek.

Struktura věty RES pro externí uživatele. *Český statistický úřad* [online]. 2013 [cit. 2015-06-21]. Dostupné z: https://www.czso.cz/csu/res/struktura_vety_res_pro_externi_uzivatele

O registru - RES. *Český statistický úřad* [online]. 2014, 11. 9. 2014 [cit. 2015-02-25]. Dostupné z: http://www.czso.cz/csu/redakce.nsf/i/o_registru_res

Ceník. *Český statistický úřad* [online]. 2015 [cit. 2015-04-13]. Dostupné z: https://www.czso.cz/csu/czso/cenik-informacnich-sluzeb-a-produktu-bwut?skupina=14

How Grouping Analysis works. ESRI. *ArcGIS Help 10.2* [online]. 2014 [cit. 2015-08-09]. Dostupné z: http://resources.arcgis.com/en/help/main/10.2/index.html#//005p0000004w000000

ESRI File Geodatabase. *GDAL: Geospatial Data Abstraction Library* [online]. 2014 [cit. 2015-08-04]. Dostupné z: http://www.gdal.org/drv_filegdb.html

FORMÁNEK, Jiří. RÚIAN - úvod do problematiky. In: ,. *Prezentace k RÚIAN/ISÚI/VDP* [online]. 2014 [cit. 2015-06-22]. Dostupné z: http://www.cuzk.cz/Uvod/Produkty-a-sluzby/RUIAN/7-Publicita-projektu/Prezentace-k-problematice-RUIAN-ISUI-VDP/1_RUIAN-uvod-do-problematiky.aspx

HANČLOVÁ, Jana a Lubor TVRDÝ. Úvod do analýzy časových řad. In: *Prostorová analýza nezaměstnanosti: učební texty ze školení* [online]. 2003 [cit. 2015-04-27]. Dostupné z: http://gis.vsb.cz/pan-old/Skoleni_Texty/TextySkoleni/AnalyzaCasRad.pdf

HEBÁK, Petr, Jiří HUSTOPECKÝ, Eva JAROŠOVÁ a Ivana MALÁ. *Vícerozměrné statistické metody*. Vyd. 1. Praha: Informatorium, 2005, 3 sv. ISBN 80-733-3025-3.

HORÁK, Jiří a Milan ŠIMEK. *Územní analýza nezaměstnanosti na příkladu okresu Nový Jičín*. Sborník vědeckých prací VŠB-TU Ostrava, řada HGF, ročník 46. 2000, s. 25-36. ISBN 80-7078-853-4.

HORÁK, Jiří, Milan ŠIMEK, Michal BOROŇ, Bronislava HORÁKOVÁ a Jana HANČLOVÁ. *Příklady použití multivariačního a multikriteriálního hodnocení nezaměstnanosti* [online]. 2004, 24 s. [cit. 2015-04-23]. Dostupné z: http://gis.vsb.cz/pan-old/Skoleni_Texty/PrikladyCviceni/MULTI.pdf

INSTITUT GEOINFORMATIKY VŠB-TU OSTRAVA. *Prostorová analýza nezaměstnanosti* [online]. 2005, 31. 8. 2007 [cit. 2015-04-23]. Dostupné z: http://gis.vsb.cz/pan-old/index.htm

KUČERA, Jiří. Metody kategorizace dat [online]. Brno, 2008 [cit. 2015-08-06]. Bakalářská práce. Masarykova univerzita, Fakulta informatiky. Vedoucí práce Matěj Štefaník. Dostupné z: <http://is.muni.cz/th/172767/fi_b/>.

LANDA, Martin. RUIAN / GDAL. *FreeGIS portál* [online]. 2014 [cit. 2015-06-26]. Dostupné z: http://freegis.fsv.cvut.cz/gwiki/RUIAN_/_GDAL

MAOH, Hanna a Pavlos KANAROGLOU. Geographic clustering of firms and urban form: a multivariate analysis. *Journal of Geographical Systems*. 2007-3-12, **9**(1): 29-52. DOI: 10.1007/s10109-006-0029-6. ISSN 1435-5930. Dostupné také z: http://link.springer.com/10.1007/s10109-006-0029-6

ARES - Popis. *Ministerstvo financí ČR: ARES - Administrativní Registr Ekonomických Subjektů* [online]. 2012 [cit. 2015-04-11]. Dostupné z: http://wwwinfo.mfcr.cz/ares/ares_popis.html.cz

Otevřená data Ministerstva financí. *Ministerstvo financí ČR* [online]. 2015 [cit. 2015-04-11]. Dostupné z: http://data.mfcr.cz/cs

GeoInfoStrategie. *Ministerstvo vnitra* [online]. 2015 [cit. 2015-08-11]. Dostupné z: http://www.mvcr.cz/clanek/geoinfostrategie.aspx?q=Y2hudW09OA%3D%3D

NOVÁK, Jakub a Pavlína NETRDOVÁ. Prostorové vzorce sociálně-ekonomické diferenciace obcí v České republice. *Český sociologický časopis / Czech Socilogical Review*. 47. 2011, 2011(4): 717-744. Dostupné také z:

http://sreview.soc.cas.cz/uploads/ebc816f4f0118b5ebd28a57282adf20b6a2fc2d7_Novak%20soccas2011-4.pdf

Opendata.cz: Iniciativa za otevřenou datovou infrastrukturu [online]. 2014 [cit. 2015-04-11]. Dostupné z: http://opendata.cz/cs

PostgreSQL: Documentation 9.4: Pattern Matching. POSTGRESQL GLOBAL DEVELOPMENT GROUP. *PostgreSQL* [online]. 2015 [cit. 2015-06-27]. Dostupné z: http://www.postgresql.org/docs/9.4/static/functions-matching.html

PŘIBYLOVÁ, Alexandra. Průzkumová analýza vícerozměrných dat [online]. Brno, 2015 [cit. 2015-08-11]. Diplomová práce. Masarykova univerzita, Přírodovědecká fakulta. Vedoucí práce RNDr. Marie Budíková, Dr.. Dostupné z: <http://theses.cz/id/2vad21/>.

Data. *Rozklikávací rozpočet obce* [online]. 2014 [cit. 2015-04-14]. Dostupné z: http://www.rozpocetobce.cz/zdroje-dat

Ruian2pgsql README. ŠULC, Miroslav. *GitHub* [online]. 2014 [cit. 2015-06-26]. Dostupné z: https://github.com/fordfrog/ruian2pgsql/blob/master/README.cs.md

UHER, Michal. Parametrické modely v analýze přežití [online]. 2011 [cit. 2015-08-02]. Bakalářská práce. Masarykova univerzita, Přírodovědecká fakulta. Vedoucí práce Tomáš Pavlík. Dostupné z: <http://theses.cz/id/suha77/>.

VOTRUBA, P.: Hodnocení regionální diferenciace sociálního a ekonomického vývoje v období 1991 – 2001 na příkladě okresů ČR s využitím metod vícerozměrné statistiky. 2003. Diplomová práce. Katedra geografie PřF MU, Brno. Vedoucí práce Václav Toušek.

Data. *Vsechnyzakazky.cz* [online]. 2015 [cit. 2015-04-15]. Dostupné z: http://vsechnyzakazky.cz/data/

PŘÍLOHY

SEZNAM PŘÍLOH

Volné přílohy:

Příloha 1 CD

Příloha 2 Poster

Vázané přílohy:

Příloha 3 Histogram počtu zanikajících ES

Příloha 4 Vývoj počtu ES v odvětvích CZ-NACE

Příloha 5 Tabulka četností ES v hlavních kategoriích odvětví CZ-NACE

Příloha 6 Vývoj počtu ES v druhé skupině konkrétních činností

Příloha 7 Vývoj počtu ES ve třetí skupině konkrétních činností

Příloha 8 Zjišťování rozdílnosti kategorií pomocí funkce `survdiff`

Příloha 9 Tabulka obcí s nejvyššími koncentracemi ES

Příloha 10 datový náhled výsledku nesprávného shlukování v obcích, metoda NO_SPATIAL_CONSTRAINT s pěti shluky

Příloha 11 Výstupní graf špatného shlukování v obcích, metoda K_NEAREST_NEIGHBORS pro 5 shluků

Příloha 12 Žádost o data pro krajskou správu ČSÚ v Olomouci

Popis struktury CD

Adresáře:

graficke_vystupy - všechny grafické výstupy vytvořené v R nebo ArcMap

metadata - XML metadatové záznamy informačního systému MICKA

poster - složka s digitální verzí přílohy č. 2 - posteru

shp - vrstvy obcí a ORP vzniklé během případových studií

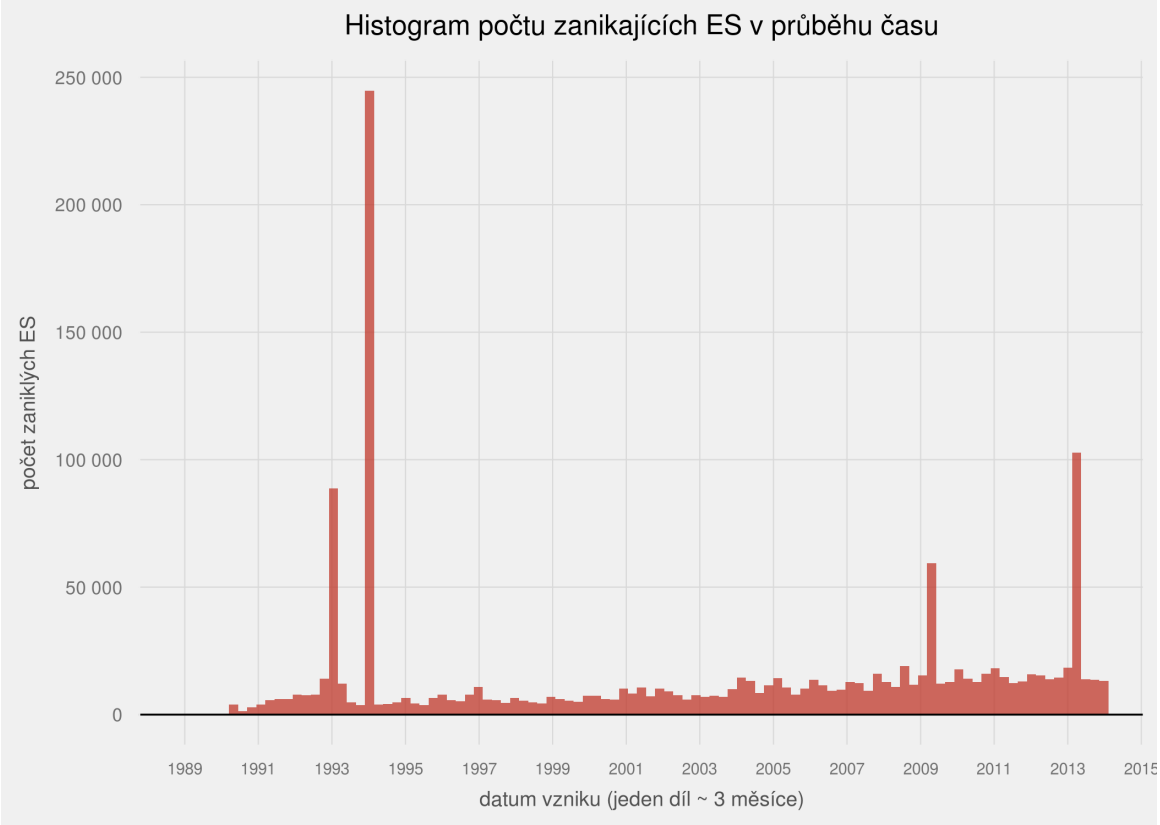
tabulky - tabulky pro import nebo join do PG, korelační matice

text_prace - složka s PDF souborem diplomové práce a použitými obrázky

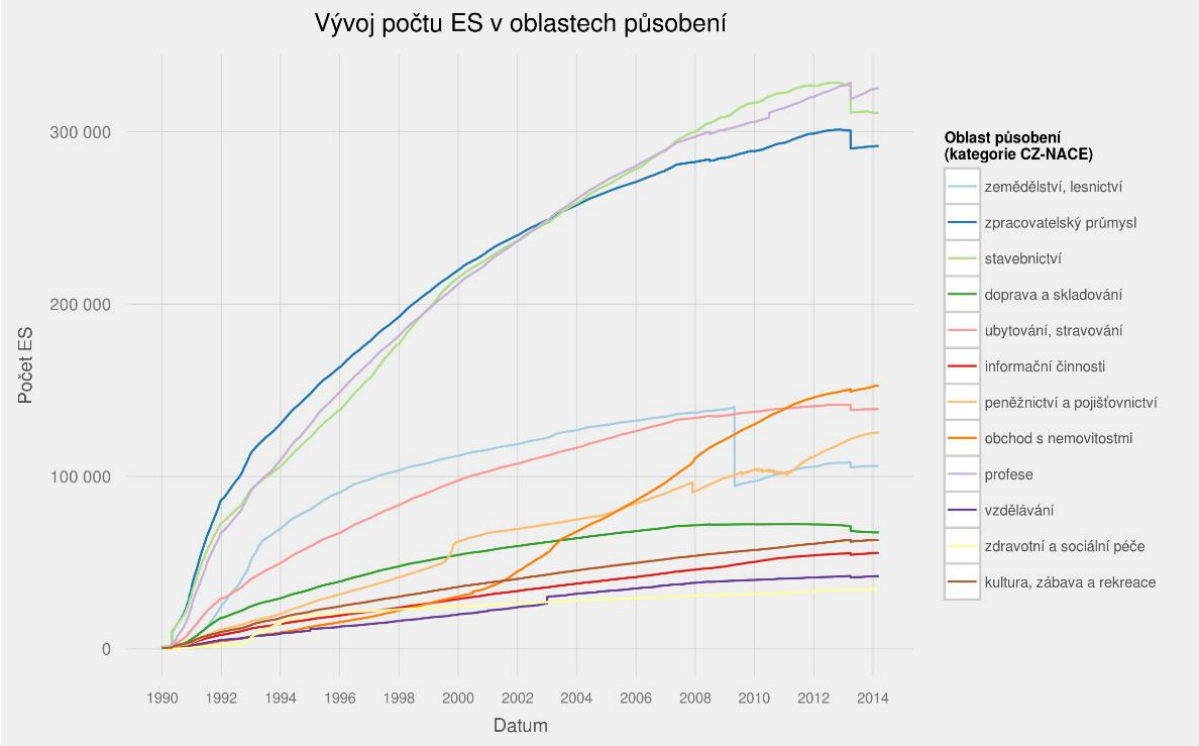
web - webové stránky vytvořené v rámci diplomové práce

zdrojovy_kod_skripty - soubory se zdrojovým kódem do PG nebo R, PHP skripty

Příloha 3 Histogram počtu zanikajících ES



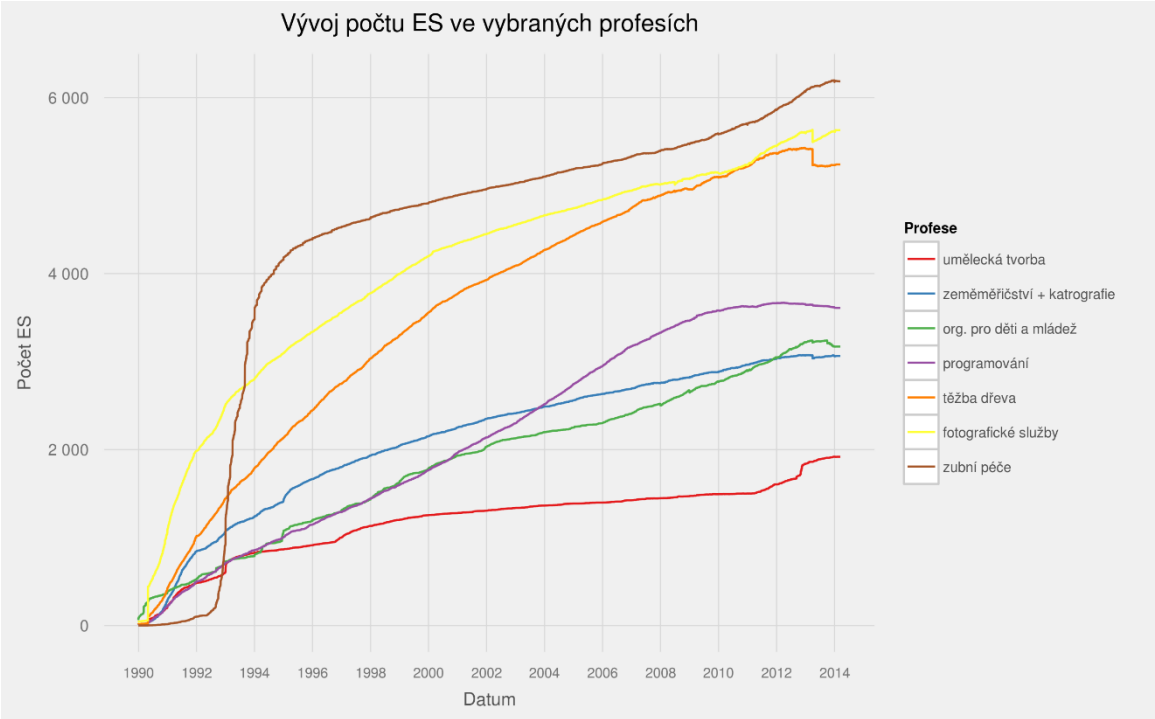
Příloha 4 Vývoj počtu ES v odvětvích CZ-NACE



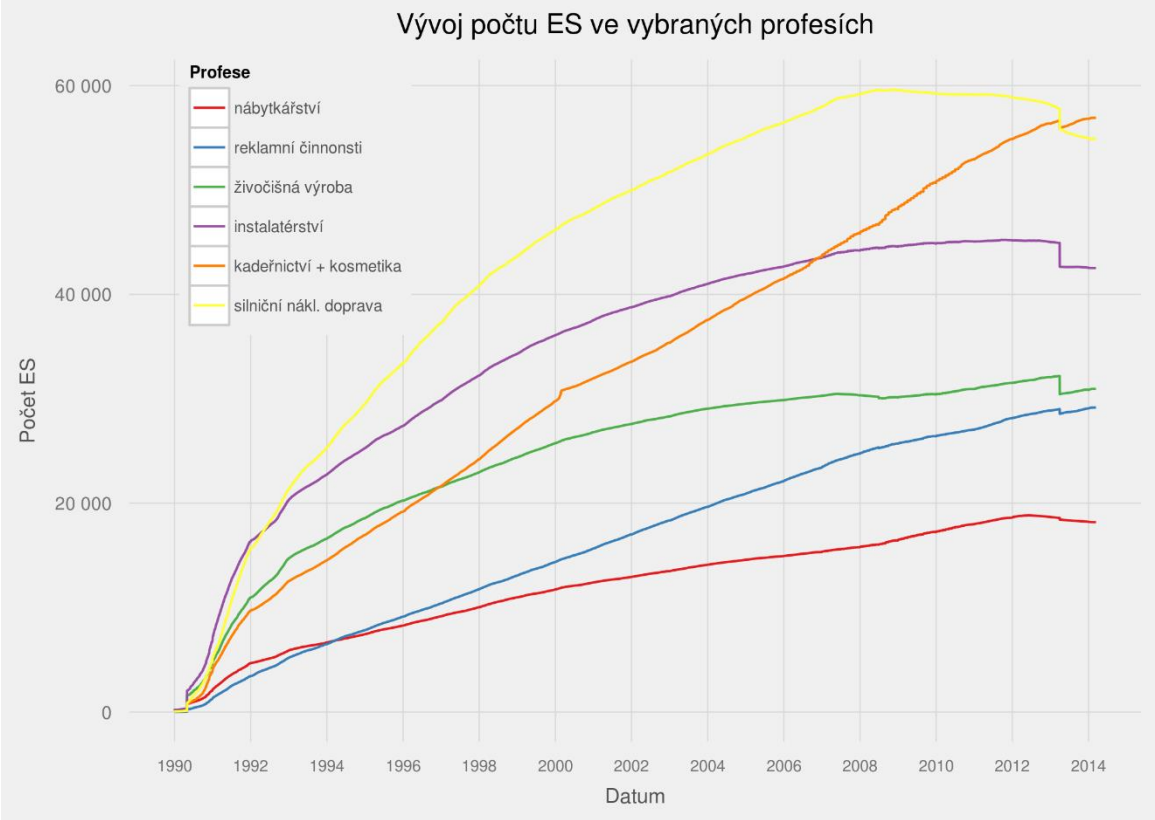
Příloha 5 Tabulka četností ES v hlavních kategoriích odvětví CZ-NACE

název kategorie	**počet ES**	**kód**
Velkoobchod a maloobchod; opravy a údržba motorových vozidel	811568	G
Stavebnictví	375888	F
Profesní, vědecké a technické činnosti	374895	M
Zpracovatelský průmysl	348678	C
Ostatní činnosti	222082	S
Zemědělství, lesnictví, rybářství	168571	A
Činnosti v oblasti nemovitostí	166399	L
Ubytování, stravování a pohostinství	164369	I
Peněžnictví a pojišťovnictví	153352	K
Doprava a skladování	82947	H
Kulturní, zábavní a rekreační činnosti	68686	R
Informační a komunikační činnosti	62233	J
Administrativní a podpůrné činnosti	58793	N
Vzdělávání	47196	P
Zdravotní a sociální péče	39438	Q
Výroba a rozvod elektřiny, plynu, tepla a klimatizovaného vzduchu	18231	D
Veřejná správa a obrana; povinné sociální zabezpečení	16081	O
Zásobování vodou; činnosti související s odpadními vodami, odpady a sanacemi	12494	E
Těžba a dobývání	823	B
Činnosti exteritoriálních organizací a orgánů	111	U
neurčeno	876353	Z

Příloha 6 Vývoj počtu ES v druhé skupině konkrétních činností



Příloha 7 Vývoj počtu ES ve třetí skupině konkrétních činností



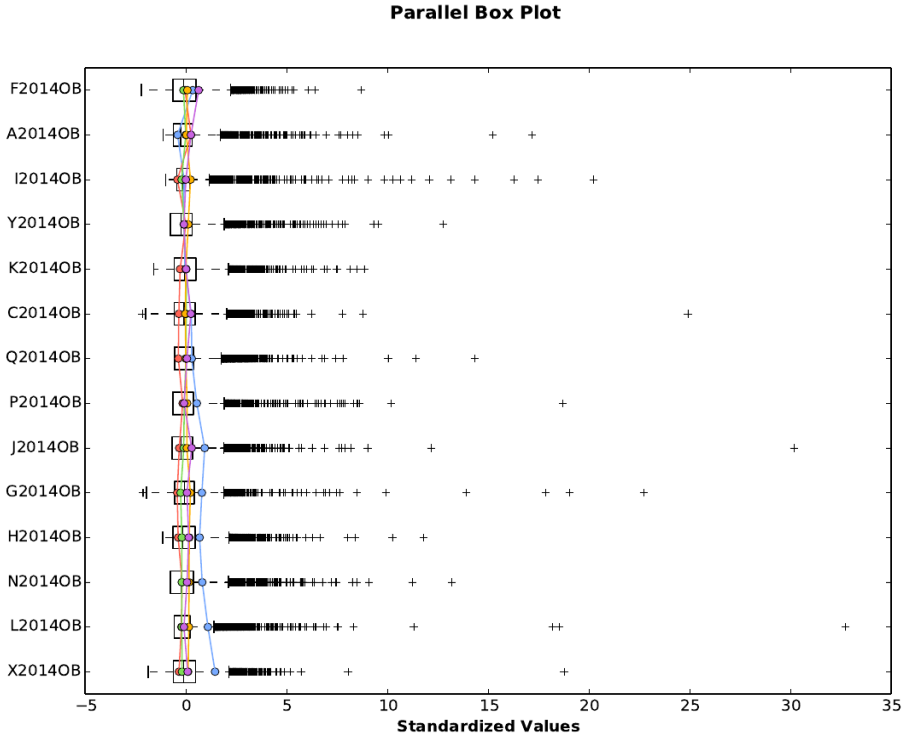
Příloha 8 Zjišťování rozdílnosti kategorií pomocí funkce `survdiff`

```
Call:
survdiff(formula = Surv(df_rok2$zaniknr, df_rok2$event) ~ df_rok2$jmenovka)

                                                N Observed Expected (O-E)^2/E (O-E)^2/V
df_rok2$jmenovka=Informační činnosti         2956      259   259.97    0.0036  3.93e-03
df_rok2$jmenovka=kultura, zábava a rekreace  3344      147   300.79   78.6294  8.71e+01
df_rok2$jmenovka=obchod s nemovitostmi       1255      122   110.58    1.1799  1.23e+00
df_rok2$jmenovka=peněžnictví a pojišťovnictví 3089     176   270.81   33.1910  3.64e+01
df_rok2$jmenovka=profese                    21532     1582  1913.44   57.4101  1.40e+02
df_rok2$jmenovka=těžba a dobývání              28        4     2.37    1.1227  1.13e+00
df_rok2$jmenovka=zdravotní a sociální péče    463       27    41.28    4.9421  5.03e+00
df_rok2$jmenovka=zemědělství, lesnictví, rybolov 4243   932   349.77  969.1829  1.09e+03

 Chisq= 1153  on 7 degrees of freedom, p= 0
```

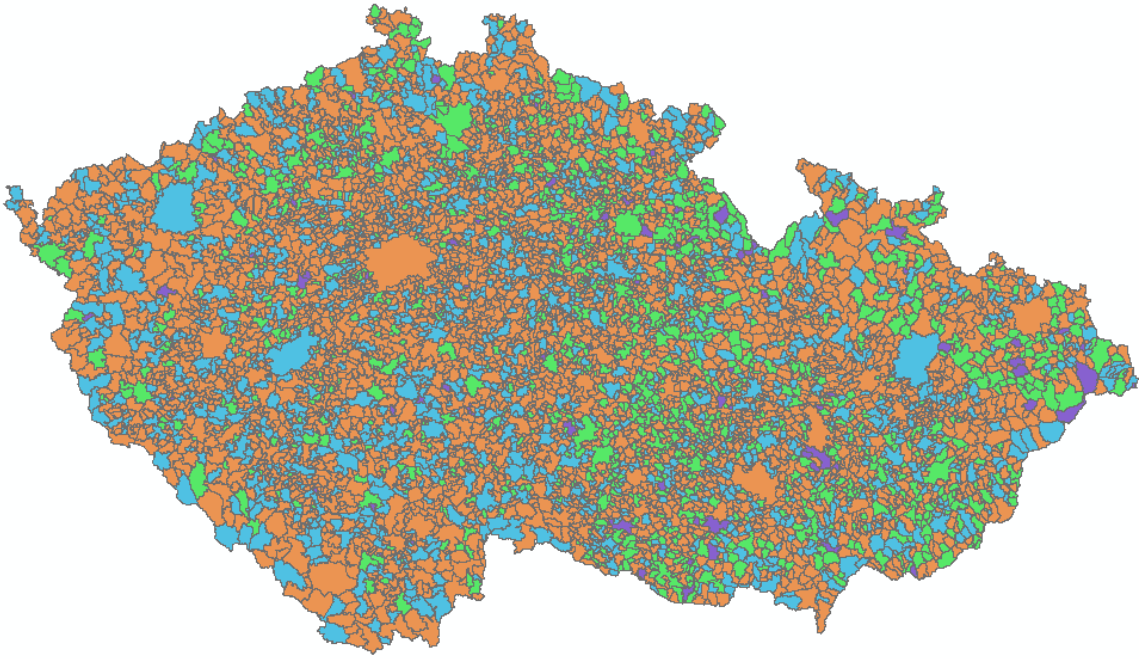
Příloha 9 Tabulka obcí s nejvyššími koncentracemi ES

pořadí	koncentrace 1991		koncentrace 2001		Koncentrace 2011	
1.	Loučná p. Klínovcem	0,70	Květnice	1,77	Vlkov	1,28
2.	Strážné	0,60	Nupaky	1,57	Modrava	0,78
3.	Malá Úpa	0,58	Herink	0,88	Pernink	0,75
4.	Boží Dar	0,49	Modrava	0,82	Hřensko	0,74
5.	Pec pod Sněžkou	0,48	Třebčice	0,79	Strážné	0,67
6.	Maršov u Úpice	0,47	Strážné	0,75	Horní Blatná	0,66
7.	Janov	0,47	Pohoří	0,70	Loučná p. Klínovcem	0,65
8.	Hřensko	0,44	Babice	0,70	Čilá	0,63
9.	Květnice	0,43	Okoř	0,69	Boží Dar	0,59
10.	Jetřichovice	0,42	Staňkovice	0,66	Šléglov	0,59
11.	Horská Kvilda	0,41	Vestec	0,65	Třebčice	0,57
12.	Kubova Huť	0,40	Hřensko	0,65	Český Jiřetín	0,56
13.	Český Jiřetín	0,36	Zlatá	0,64	Nové Hutě	0,56
14.	Bedřichov	0,35	Český Jiřetín	0,64	Moldava	0,56
15.	Ohrobec	0,35	Nasavrky	0,63	Těchařovice	0,55
16.	Nupaky	0,35	Čilá	0,63	Kvilda	0,55
17.	Hlincová Hora	0,34	Křimov	0,63	Stožec	0,54
18.	Okoř	0,34	Loučná p. Klínovcem	0,62	Prášily	0,53
19.	Pesvice	0,34	Březová-Oleško	0,62	Vysoká Lhota	0,53
20.	Babice	0,33	Nová Pláň	0,62	Šípy	0,52

Příloha 10 Výstupní graf špatného shlukování v obcích, metoda K_NEAREST_NEIGHBORS pro 5 shluků



Příloha 11 datový náhled výsledku nesprávného shlukování v obcích, metoda NO_SPATIAL_CONSTRAINT s pěti shluky



Příloha 12 Žádost o data pro ČSÚ v Olomouci

KATEDRA GEOINFORMATIKY
Univerzita Palackého v Olomouci

Věc: Žádost o poskytnutí dat pro diplomovou práci

Tímto potvrzuji, že diplomant Vojtěch Cícha je studentem Katedry geoinformatiky, Přírodovědecké fakulty Univerzity Palackého v Olomouci, a žádám o poskytnutí kompletní databáze **Registr ekonomických subjektů** pro účely diplomové práce s názvem Prostorové analýzy veřejně dostupných dat ekonomických subjektů v České republice, vypracovávané v akademických letech 2013-2014 a 2014-2015.

V Olomouci dne 13. 2. 2014

Mgr. Pavel Tuček, Ph.D.
vedoucí diplomové práce